POLYNOMIAL METHODS IN STATISTICAL INFERENCE: THEORY AND PRACTICE

BY

PENGKUN YANG

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2018

Urbana, Illinois

Doctoral Committee:

Associate Professor Maxim Raginsky, Chair
Assistant Professor Yihong Wu, Director of Research
Professor Bruce Hajek
Professor Rayadurgam Srikant
Assistant Professor Sewoong Oh

ABSTRACT

Recent advances in genetics, computer vision, and text mining are accompanied by analyzing data coming from a large domain, where the domain size is comparable or larger than the number of samples. In this dissertation, we apply the polynomial methods to several statistical questions with rich history and wide applications. The goal is to understand the fundamental limits of the problems in the large domain regime, and to design sample optimal and time efficient algorithms with provable guarantees.

The first part investigates the problem of property estimation. Consider the problem of estimating the Shannon entropy of a distribution over k elements from n independent samples. We obtain the minimax mean-square error within universal multiplicative constant factors if n exceeds a constant factor of $k/\log(k)$; otherwise there exists no consistent estimator. This refines the recent result on the minimal sample size for consistent entropy estimation. The apparatus of best polynomial approximation plays a key role in both the construction of optimal estimators and, via a duality argument, the minimax lower bound.

We also consider the problem of estimating the support size of a discrete distribution whose minimum non-zero mass is at least $\frac{1}{k}$. Under the independent sampling model, we show that the sample complexity, i.e., the minimal sample size to achieve an additive error of ϵk with probability at least 0.1 is within universal constant factors of $\frac{k}{\log k}\log^2\frac{1}{\epsilon}$, which improves the state-of-the-art result of $\frac{k}{\epsilon^2\log k}$. Similar characterization of the minimax risk is also obtained. Our procedure is a linear estimator based on the Chebyshev polynomial and its approximation-theoretic properties, which can be evaluated in $O(n+\log^2 k)$ time and attains the sample complexity within constant factors. The superiority of the proposed estimator in terms of accuracy, computational efficiency and scalability is demonstrated in a variety of synthetic and real datasets.

When the distribution is supported on a discrete set, estimating the support size is also known as the distinct elements problem, where the goal is to estimate the number of distinct colors in an urn containing k balls based on n samples drawn with replacements. Based on discrete polynomial approximation and interpolation, we propose an estimator with additive error guarantee that achieves the optimal sample complexity within $O(\log \log k)$ factors, and in fact within constant factors for most cases. The estimator can be computed in $O(n)$ time for an accurate estimation. The result also applies to sampling without replacement provided the sample size is a vanishing fraction of the urn size. One of the key auxiliary results is a sharp bound on the minimum singular values of a real rectangular Vandermonde matrix, which might be of independent interest.

The second part studies the problem of learning Gaussian mixtures. The method of moments is one of the most widely used methods in statistics for parameter estimation, by means of solving the system of equations that match the population and estimated moments. However, in practice and especially for the important case of mixture models, one frequently needs to contend with the difficulties of non-existence or non-uniqueness of statistically meaningful solutions, as well as the high computational cost of solving large polynomial systems. Moreover, theoretical analysis of the method of moments are mainly confined to asymptotic normality style of results established under strong assumptions.

We consider estimating a k-component Gaussian location mixture with a common (possibly unknown) variance parameter. To overcome the aforementioned theoretic and algorithmic hurdles, a crucial step is to denoise the moment estimates by projecting to the truncated moment space (via semidefinite programming) before solving the method of moments equations. Not only does this regularization ensures existence and uniqueness of solutions, it also yields fast solvers by means of Gauss quadrature. Furthermore, by proving new moment comparison theorems in the Wasserstein distance via polynomial interpolation and majorization techniques, we establish the statistical guarantees and adaptive optimality of the proposed procedure, as well as oracle inequality in misspecified models. These results can also be viewed as provable algorithms for generalized method of moments which involves non-convex optimization and lacks theoretical guarantees.

To my parents and my wife, for their love and support.

ACKNOWLEDGMENTS

I am grateful to my adviser, Professor Yihong Wu, for introducing me to the statistical theory in communication engineering. When I worked on communication systems earlier I was always frightened by the deep theory in the literature; I used many statistical tools but never truly understood why they worked so well, and where are the limitations. Through the years, I learned from Professor Wu how to think deeply and clearly, and how to translate the mathematics into science and engineering. He is always passionate to share his thoughts and courageous to tackle difficult problems with me. Every piece of progress in this dissertation comes from many hours of discussion with him.

I thank Professor Maxim Raginsky for serving as my doctoral committee chair in both preliminary and final exams. I thank Professor Bruce Hajek, Professor R. Srikant, and Professor Sewoong Oh for being on my committee.

I would like to thank the faculties at the University of Illinois who created so many wonderful courses. I learned from them the rigorous measure theory, probability theory, statistical theory, and learning theory. Those further help me grasp the ideas in many aspects of communication engineering and computer science. They built the foundation of my research.

I am also thankful to my friends and the graduate students in the Coordinated Science Laboratory for their support. I was fortunate to have interacted with them in both research and extracurricular life. In particular, I thank Jiaming, Lili, and Taposh for their help during my transition to the new environment, Shengmei for the training for a full marathon, and Shiyan, Wenyu and Xinke for conquering the Rocky Mountains together!

I am indebted to my parents for the constant support and appreciate that I can study abroad without worries. Finally, none of this could be done without the companionship of my beloved wife Dandan during our journey and adventure of life.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

CDF	Cumulative Distribution Function
CLT	Central Limit Theorem
DMM	Denoised Method of Moments
EM	Expectation-Maximization
GMM	Generalized Method of Moments
i.i.d.	independently and identically distributed
KL	Kullback-Leibler
KS	Kolmogorov-Smirnov distance
LP	Linear Programming
MLE	Maximum Likelihood Estimate
MM	Method of Moments
NPMLE	Non-Parametric Maximum Likelihood Estimate
MSE	Mean Squared Error
RMSE	Root Mean Squared Error

LIST OF SYMBOLS

$\|\cdot\|_p$	the vector ℓ_p-norm, for $1 \le p \le \infty$
$[x \pm a]$	the interval $[x-a, x+a]$
$\langle x, y\rangle$	the L_2 inner product $\sum_i x_i y_i$
Φ	fingerprint of the samples
σ-subgaussian	$\mathbb{E}[e^{tX}] \le \exp(t^2\sigma^2/2)$ for all $t \in \mathbb{R}$
$\chi^2(P\|Q)$	χ^2-divergence between probability measures P and Q
$a \vee b$	maximum between a and b
$a \wedge b$	minimum between a and b
$a_n \lesssim b_n$	$a_n \le cb_n$ for some absolute positive constant c
$a_n \gtrsim b_n$	$a_n \ge cb_n$ for some absolute positive constant c
$a_n \asymp b_n$	$a_n \gtrsim b_n$ and $b_n \gtrsim a_n$
$a_n = O(b_n)$	$a_n \lesssim b_n$
$a_n = \Omega(b_n)$	$a_n \gtrsim b_n$
$a_n = \Theta(b_n)$	$a_n \asymp b_n$
$a_n = o(b_n)$	$a_n \ll b_n$, i.e., $\lim a_n/b_n = 0$
$a_n = \omega(b_n)$	$b_n = o(a_n)$
$A^\top$	the transpose of a matrix A
$A \succeq B$	$A - B$ being positive semidefinite
Bern(p)	the Bernoulli distribution with mean p
binomial(n, p)	the binomial distribution with n trials and success probability p

$D(P\|\|Q)$	the Kullback-Leibler (KL) divergence between probability measures P and Q
$E_L(g, [a, b])$	best uniform approximation error of function g on $[a, b]$ by a polynomial of degree at most L
$\mathbb{E}_\pi[P_\theta]$	the mixture of a parametrized family of distributions $\{P_\theta\}$ under the prior π
$\mathbb{E}_n[f(X)]$	the empirical mean of f from n samples
$f[x_1, \ldots, x_n]$	divided difference of function f on $x_1, \ldots, x_n$
h	histogram of N, also known as profile or fingerprint
$H(P)$	Shannon entropy of a discrete distribution P
$[k]$	a set of integers $\{1, 2, \ldots, k\}$
log	all logarithms are with respect to the natural base and the entropy is measured in nats
$\mathbf{m}_n(\mu)$	n^{th} moment vector of μ
$\mathcal{M}_n(K)$	n^{th} moment space on K
N	histogram of original samples
$o_\delta(1)$	convergence to zero that is uniform in all other parameters as $\delta \to 0$
$\hat{P}$	empirical distribution
$P^{\otimes n}$	n-fold product of a given distribution P
$\mathcal{P}_L$	the space of all polynomials of degree no greater than L
$\text{Poi}(\lambda)$	the Poisson distribution with mean λ whose probability mass function is $\text{poi}(\lambda, j) \triangleq \frac{\lambda^j e^{-\lambda}}{j!}, j \in \mathbb{Z}_+$
$\mathsf{TV}(P, Q)$	the total variation between probability measures P and Q

CHAPTER 1

INTRODUCTION

Communication and information theory are traditionally based on statistical inference on probabilistic models. For instance, the transmission channel is modeled as a conditional distribution, and the transmitted signal is detected or estimated from the received noisy signal. In modern science and engineering, statistical inference is one of the most fundamental problems under various models such as the Markov model for natural languages, latent Dirichlet allocation for document topics, Gaussian mixtures for biometric systems, etc. One central question is how to design an *accurate* and *efficient* algorithm to infer properties or draw conclusions of the unknown distribution by analysis of data.

Over the last century, statisticians, information theorists, and computer scientists have extensively studied the asymptotic regime, where the population is *fixed* and large amounts of data are available. Nowadays, however, modern datasets such as genetics data and social media are accompanied by an underlying distribution on a far larger domain. As the domain enlarges, the dimension and the volume of the parameter space increase so fast that the limited amounts of data are too sparse to reach statistically sound conclusions, known as "curse of dimensionality" [1]. For example, in machine learning, we measure more and more features and also use the conjunction of different features in pursuing a finer description of an object [2]; in natural language processing, besides a large vocabulary, bigram, trigram, etc., are frequently used in practice [3]; in computer vision, each image can have hundreds to thousands of different features, and the image is understood from a combination of those features [4]. Similar situation arises in areas like speech recognition, spam detection, and alignment of biological data. There are improved algorithms based on classical statistical procedures, but applying them on large domains typically yields unsatisfactory results, and one is therefore urged to develop new theory and algorithms for the classical

problems of modern datasets.

Recently, several challenging problems have been successfully solved using the theory of polynomials. This approach is called the polynomial method. In this dissertation, polynomial methods will be investigated and applied to statistical inference problems on large domains. They provide useful tools not only in the design of estimators in practice, but also in establishing the theoretical limits of the inference problems.

1.1 Review and examples

Polynomial methods bring one mathematical element to the frontier of science ranging from physics and chemistry to economics and social science. Many challenging problems in coding theory, harmonic analysis, combinatorics, econometrics, etc. were successfully solved from the perspective of polynomials. In various fields of applications spanning speech processing, control theory, finance, and game theory, optimization is often formulated naturally with polynomial constraints; examples include Markov models, optimal control, option pricing, and Nash equilibria. Polynomial methods provide a natural tool to either exactly or approximately solve these diverse problems in both theory and practice and are increasingly in evidence. In this dissertation, we study the following polynomial methods.

Polynomial approximation. Approximation is one philosophy of science that simplifies complicated theories, reveals the underlying structure and aids the deep understanding of complicated objects, and is also naturally required in practice. Polynomial approximation is one such subject and is one of the most well-understood methods. For instance, Taylor's polynomials characterize the local behavior of a function, and are widely used in modern solutions like gradient descent [2]; trigonometric polynomials represent functions in the frequency domain, known as Fourier analysis in signal processing, and help remove irrelevant noises and make the wireless communication possible [5, 6]. In statistical inference, a good polynomial is a natural proxy for the property of interest, which is represented as a function of the underlying distribution, a data generating model that is unknown or partially unknown. The property itself might be difficult to estimate, and classical methods yield

poor accuracy or require a colossal number of samples. Nevertheless, we can first find a good approximant for the original function as a proxy property that is easy to estimate, and then focus on this approximant. The degree of approximation aims for a tradeoff between approximation error and estimation error, which is a balance between bias and variance from a statistical viewpoint.

Moments and positive polynomials. The theory of moments plays a key role in the developments of analysis, probability, statistics, and optimization. See the classics [7, 8] and the recent monographs [9, 10] for a detailed treatment. In statistical inference, the method of moments was originally introduced by Pearson [11] and its extensions have been widely applied in practice, for instance, to analyze economic and financial data [12]. The method of moments estimates are obtained by solving polynomial equations. They are useful for their simplicity, especially in models without the complete specification of the joint distribution of data, and also in cases when data might be contaminated. Moments of a distribution satisfy many geometric properties such as the Cauchy-Schwarz and Hölder inequalities, and a complete description can be phrased in terms of positive polynomials (Riesz-Haviland representation theorem). Positive polynomials are further closely related to sums of squares, which is equivalent to positive semidefiniteness of the representing matrix (see Section 2.3).

1.2 Polynomial methods in practice

We will apply the above two polynomial methods in the following two types of statistical inference problems, respectively.

Estimating distributional properties on large domains. Given samples drawn from an unknown distribution, the goal is to estimate a specific property of that distribution, such as information measures and the support size. This falls under the category of *functional estimation* [13], where we are not interested in directly estimating the high-dimensional parameter (the distribution P) per se, but rather a function thereof. Estimating a distributional functional has been intensively studied in nonparametric statistics,

e.g., estimate a scalar function of a regression function such as linear functional [14, 15], quadratic functional [16], L_q norm [17], etc. To estimate a function, perhaps the most natural idea is the "plug-in" approach, namely, first estimate the parameter and then substitute into the function. As frequently observed in functional estimation problems, the plug-in estimator can suffer from severe bias (see [18, 19] and the references therein). Indeed, although plug-in estimate is typically asymptotically efficient and minimax (cf., e.g., [20, Sections 8.7 and 8.9]), in the fixed alphabet regime, it can be highly suboptimal in high dimensions, where, due to the a large alphabet size and resource constraints, we are constantly contending with the difficulty of *undersampling* in applications such as

- corpus linguistics: about half of the words in the Shakespearean canon only appeared once [21];
- network traffic analysis: many customers or website users are only seen a small number of times [22];
- analyzing neural spike trains: natural stimuli generate neural responses of high timing precision resulting in a massive space of meaningful responses [23, 24, 25].

In this dissertation, we focus on estimating some classical properties of interest including Shannon entropy and the number of unseen. Those properties can be easily estimated if the number of samples far exceeds the size of the underlying distribution, but how can it be done if the samples are relatively scarce, such as only a sublinear number of samples are available? It turns out the best polynomial approximation provides a principled approach to design an optimal estimator.

Learning Gaussian mixtures. A sample from a mixture model can be viewed as being generated by a two-step process: first draw a parameter θ from the unknown mixing distribution; then draw a sample from P_θ. If we are only given unlabeled data from the mixture model, can we reconstruct the parameters in each components efficiently? In the special case that each P_θ is a Gaussian distribution, this is the problem of learning Gaussian mixtures. Learning Gaussian mixtures has a long history dating back to the work of Pearson [11], where the method of moments was first introduced. Despite its

long history, it is still one part of the core machine learning toolkit, such as the popular scikit-learn package in Python [26], Google's Tensorflow [27], and Spark's MLlib [28], but very few provable guarantees are available. It is until recently proved in [29, 30] that a mixture of constant number of components can be learned in polynomial time using a polynomial number of samples. The sharp rate error rate for learning a mixture of two univariate Gaussians is proved more recently in [31]. What is a systematic way to obtain the sharp error rates, and what is the best way to learn a Gaussian mixture? We will investigate the moment methods for optimal estimation of Gaussian mixtures.

1.3 Polynomial methods in the theoretical limits

In this dissertation, we also study the fundamental limits of statistical inference. While the use of polynomial methods on the constructive side is admittedly natural, the fact that it also arises in the optimal lower bound is perhaps surprising. It turns out the optimality can be established via duality in the optimization sense that will be elaborated in this section.

To give a precise definition of the fundamental limits, we begin with an account of the general framework for statistical inference. We assume that an observation X is generated from an unknown distribution P from a space of distributions $\mathcal{P}$. The goal is to estimate some properties of that distribution $T(P)$. See an illustration of this framework in Figure 1.1. In the problems introduced above, we consider the following two types of properties:

- *Estimating distributional properties:* $T(P)$ is a functional of the unknown distribution P, such as the Shannon entropy $H(P) = \sum_i p_i \log \frac{1}{p_i}$ and the support size $S(P) = \sum_i \mathbf{1}_{\{p_i>0\}}$ of a discrete distribution $P = (p_1, p_2, \ldots)$.

- *Learning Gaussian mixtures:* $T(P)$ represents the parameters, including the mean, variance, and the mixing weights, of each Gaussian component. Equivalently, $T(P)$ can be viewed as the mixing distribution of the mixture model (see Chapter 6).

For a loss function $\ell(\hat{T}, T(P))$ that penalizes the output of an estimator $\hat{T}$,

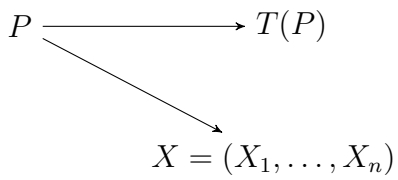


Figure 1.1: Illustration of information flow in statistical inference.

the decision-theoretic fundamental limit is defined as the minimiax risk

$$R_n^* \triangleq \inf_{\hat{T}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[\ell(\hat{T}, T(P))], \tag{1.1}$$

where $\hat{T}$ is an estimator measurable with respect to n independent samples $X_1, \ldots, X_n \sim P$. Examples of the loss function include quadratic loss $\ell(x, y) = \|x - y\|_2^2$ and zero-one loss $\ell(x, y) = \mathbf{1}_{\{\|x-y\|_2 > \epsilon\}}$ for a desired accuracy ϵ. For the zero-one loss, we also consider the sample complexity.

Definition 1.1. For a desired accuracy ϵ and confidence $1 - \delta$, the sample complexity is the minimal sample size n such that there exists an estimator $\hat{T}$ based on n samples drawn independently from a distribution P such that $\mathbb{P}[|\hat{T} - T(P)| < \epsilon] \geq 1 - \delta$ for any $P \in \mathcal{P}$.

In this dissertation, the fundamental limits of statistical inference refer to a characterization of the minimax risk (1.1) or the sample complexity in Definition 1.1. These involve an upper bound given by the statistical guarantees of estimators and a minimax lower bound.

A general idea for obtaining lower bounds is based on a reduction of estimation to testing (Le Cam's method). If there are two distributions P and Q that cannot be reliably distinguished based on a given number of samples, while $T(P)$ and $T(Q)$ are different, then any estimate suffers a maximum risk at least proportional to the distance between $T(P)$ and $T(Q)$. The above two distributions can be generalized to two priors on the space of all distributions (also known as fuzzy hypothesis testing in [32]). Here the polynomial methods enter the scene again: statistical closeness between two distributions can be established by comparing their moments, which is exactly the basis for the moment methods!

To implement the above lower bound program, the strategy is to choose two priors with matching moments up to a certain degree, which ensures that the induced distributions of data are impossible to test. The minimax lower

bound is then given by the maximal separation in the expected functional values subject to the moment matching condition. It turns out this optimization problem is the *dual* problem of the best polynomial approximation. This approach yields the optimal minimax lower bounds in the statistical inference problems investigated in this dissertation.

1.4 Dissertation organization

A plot of this dissertation is shown in Figure 1.2.

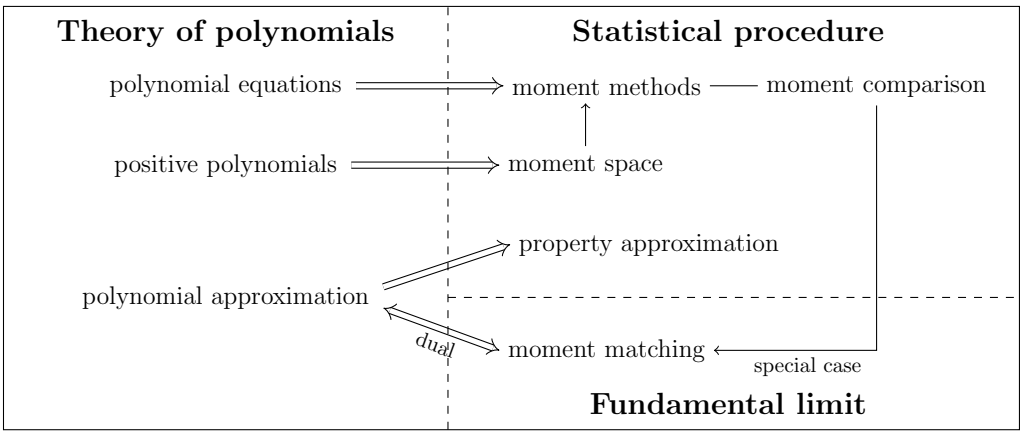


Figure 1.2: A diagram of topics in the dissertation.

A background on the theory of polynomials is briefly introduced in Chapter 2, including polynomial approximation, interpolation, theory of moments and positive polynomials, and orthogonal polynomials.

Part I is devoted to property estimation. In Chapter 3, common techniques are introduced, including Poisson sampling, approximation-theoretical techniques for constructing the statistical procedures, and moment matching for the minimax lower bounds. The problem of entropy estimation is studied in detail in Chapter 4. In Chapter 5, we studied the estimation of the unseen, including support size estimation and the distinct elements problem.

Learning Gaussian mixtures in Part II relies on the moment methods. A general framework is established in Chapter 6, and moment comparison theorems are developed in Chapter 7. These results are independent of properties of Gaussians and are applicable to general mixture models. Estimators and their statistical guarantees are given in Chapter 8.

CHAPTER 2

BACKGROUND ON THE THEORY OF POLYNOMIALS

We begin with background on polynomials that are useful in statistical inference. For a comprehensive survey on the theory of polynomials see the monographs by Prasolov [33] and Timan [34]. Our focus is on the algebraic (ordinary) polynomials in one variable. Extensions to trigonometric polynomials and multivariate polynomials are briefly introduced. See [34] for more details.

The simplest polynomials are functions of one variable x of the form

$$p_n(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n,$$

for some $n \in \mathbb{N}$, where $a_0, a_1, \ldots, a_n$ are arbitrary real or complex coefficients. The degree of a polynomial is the highest power in x with a non-zero coefficient. The set of all polynomials is a vector space $\mathcal{P}$ over the field of coefficients with countably infinite dimensions; if one restricts to polynomials of degree at most n, then it is a vector space $\mathcal{P}_n$ of $n+1$ dimensions.

The canonical basis for the polynomials space is the monomial basis, with coordinates being the coefficients of polynomials. Any set of $n+1$ polynomials $\{p_0, p_1, \ldots, p_n\}$ such that each p_m has degree m can serve as a basis for the polynomials space $\mathcal{P}_n$, and every polynomial of degree at most n can be uniquely represented by a linear combination of these polynomials via a change of basis.

Trigonometric polynomials are functions in θ of the form

$$p_n(\theta) = \sum_{k=0}^{n} (a_k \cos k\theta + b_k \sin k\theta),$$

with coefficients a_k and b_k. The degree of a trigonometric polynomial is the largest k such that a_k and b_k are not both zero. The functions $\cos k\theta$ and $\sin(k+1)\theta / \sin\theta$ are ordinary polynomials in $\cos\theta$, named Chebyshev

polynomials of the first and second kind, respectively:

$$\cos k\theta = T_k(\cos\theta), \quad \frac{\sin(k+1)\theta}{\sin\theta} = U_k(\cos\theta). \tag{2.1}$$

Multivariate polynomials of variable $x = (x_1, \ldots, x_n)$ are finite linear combinations of monomials $x^\alpha = x_1^{\alpha_1} \ldots x_n^{\alpha_n}$ with $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}^n$ of the form

$$p(x) = \sum_\alpha c_\alpha x^\alpha,$$

with coefficients c_α. The degree of a multivariate polynomial is the largest k such that there exists non-zero c_α with $\alpha_1 + \cdots + \alpha_n = k$.

2.1 Uniform approximation and moment matching

One fundamental theorem in the approximation theory was discovered by Weierstrass.

Theorem 2.1 (Weierstrass)**.** *Given a function f that is continuous on the interval $[a, b]$, and any $\epsilon > 0$, there exists a polynomial p such that*

$$|f(x) - p(x)| < \epsilon, \quad \forall x \in [a, b].$$

If f is continuous and has the period 2π, then there exists a trigonometric polynomial q such that

$$|f(x) - q(x)| < \epsilon, \quad \forall x.$$

This theorem has been proved in a great variety of ways, and can be generalized to the approximation of multivariate continuous functions in a closed bounded region. For more information on this theorem, refer to [34, Chapter 1]. In the first case of the theorem, one very elegant construction is the Bernstein polynomial to approximate continuous functions on $[0, 1]$:

$$B_n(x) = \sum_{k=0}^{n} f(k/n) \binom{n}{k} x^k (1-x)^{n-k}. \tag{2.2}$$

The approximation of a function f using Bernstein polynomials can be char-

acterized in terms of the modulus of continuity

$$\omega(\delta) = \sup\{f(x) - f(y) : |x - y| \le \delta\}.$$

Clearly $\omega(\delta)$ vanishes with δ if f is a continuous function.

Theorem 2.2 (T. Popoviciu)**.** *Given a continuous function f on $[0, 1]$,*

$$\sup_{0\le x\le 1} |f(x) - B_n(x)| \le \frac{5}{4}\omega(n^{-\frac{1}{2}}).$$

Proof. Note that the Bernstein polynomials can be expressed using the binomial distributions as

$$B_n(x) = \mathbb{E}[f(N/n)],$$

where $N \sim \text{binomial}(n, x)$. For any $\delta > 0$,

$$|f(x) - B_n(x)| \le \mathbb{E}|f(x) - f(N/n)|\mathbf{1}_{\left\{|x-\frac{N}{n}|\le\delta\right\}}| + \mathbb{E}|f(x) - f(N/n)|\mathbf{1}_{\left\{|x-\frac{N}{n}|>\delta\right\}}.$$

To prove an upper bound of the right-hand side, we note that $|f(x)-f(y)| \le 1 + \lfloor\frac{|x-y|}{\delta}\rfloor\omega(\delta)$. Then we have

$$|f(x) - B_n(x)| \le \omega(\delta) + \frac{\omega(\delta)}{\delta}\mathbb{E}\left|x - \frac{N}{n}\right|\mathbf{1}_{\left\{|x-\frac{N}{n}|>\delta\right\}}.$$

The variance of the binomial distribution gives

$$\mathbb{E}\left|x - \frac{N}{n}\right|\mathbf{1}_{\left\{|x-\frac{N}{n}|>\delta\right\}} < \frac{1}{\delta}\mathbb{E}\left|x - \frac{N}{n}\right|^2 = \frac{x(1-x)}{n\delta} \le \frac{1}{4n\delta}.$$

The statement is established by choosing $\delta = n^{-1/2}$. □

The approximation using the Bernstein polynomial is in general not as good as other polynomials. Using the Bernstein polynomial, the rate $\omega(n^{-\frac{1}{2}})$ in Theorem 2.2 cannot be replaced by other functions decreasing more rapidly. This can be shown by considering the approximation of $f(x) = |x - x_0|^\alpha$ for some fixed $0 < x_0 < 1$ and $0 < \alpha \le 1$. However, a continuous function can be better approximated by polynomials indicated by Jackson [35].

Theorem 2.3 (Jackson)**.** *Given a continuous function f on $[0, 1]$, there exists*

a polynomial P_n of degree at most n such that

$$\sup_{0\le x\le 1} |f(x) - P_n(x)| \le 3\omega(n^{-1}).$$

For information concerning this theorem we refer to [35]. Generalizations and extensions, called Jackson-type theorems, provide degree of approximation in terms of various notions of modulus of continuity. The uniform approximation of the logarithm function is studied in Section 4.3.4 for entropy estimation. See [34, Chapter V] and [36, Chapter 7] for more constructive approximations.

2.1.1 Best uniform approximation

Given a continuous function f on an interval $[a, b]$, its best uniform approximation by $\mathcal{P}_n$ is $P_n \in \mathcal{P}_n$ such that

$$\sup_{x\in[a,b]} |f(x) - P_n(x)| = \inf_{P\in\mathcal{P}_n} \sup_{x\in[a,b]} |f(x) - P(x)|. \tag{2.3}$$

The concept of uniform approximation is introduced by Chebyshev. The best polynomial always exists and is unique (see, e.g., [36, Chapter 3]), with the following remarkable characterization.

Theorem 2.4 (Chebyshev)**.** *A polynomial $P_n \in \mathcal{P}_n$ is the best uniform approximation of a continuous function f on $[a, b]$ by $\mathcal{P}_n$ if and only if there exists $n + 2$ points x_j, $a \le x_0 < \cdots < x_{n+1} \le b$ such that $f(x_j) - P_n(x_j) = \pm\sup_{x\in[a,b]} |f(x) - P(x)|$ with successive changes of sign, i.e., $f(x_{j+1}) - P_n(x_{j+1}) = -(f(x_j) - P_n(x_j))$ for $j = 0, \ldots, n$.*

This statement holds in general besides polynomials for any Haar system, such as the ordinary polynomials on the complex plane and the trigonometric polynomials. See [36, Section 3.3 – 3.5] for a proof of this theorem and more information.

Finding the exact magnitude of the best approximation and the explicit formula of the best polynomial is of special interest in each concrete case. See [34, Section 2.11] for examples with explicit solutions. We shall give one example investigated by Chebyshev, where the polynomial will be used in Chapter 5 for estimating the unseen.

Theorem 2.5. *For $n \in \mathbb{N}$, the polynomial of degree n with unitary leading coefficient of the least deviation from zero over $[-1, 1]$ is $\frac{1}{2^{n-1}} T_n(x)$, where T_n is the Chebyshev polynomial of the first kind of degree n given by (2.1). The least deviation is $\frac{1}{2^{n-1}}$.*

Proof. We observe that the problem is equivalent to finding the best polynomial of degree $n-1$ to approximate $x \mapsto x^n$ over $[-1, 1]$:

$$\inf_{a_0,\ldots,a_{n-1}} \sup_{x\in[-1,1]} |x^n - a_{n-1}x^{n-1} - \cdots - a_1 x_1 - a_0|.$$

The polynomial $\frac{1}{2^{n-1}} T_n(x)$ is monic (i.e., with unitary leading coefficient) with maximum magnitude $\frac{1}{2^{n-1}}$. The Chebyshev polynomial T_n successively attains 1 or -1 at $\cos(k\pi/n)$ for $k = 0, \ldots, n$. The optimality of $\frac{1}{2^{n-1}} T_n(x)$ follows from Theorem 2.4. □

The best polynomial of degree n can be obtained by solving the following linear program with $n + 2$ decision variables and infinite constraints:

$$\begin{aligned} \min \quad & t, \\ \text{s.t.} \quad & a_0 + a_1 x + \cdots + a_n x^n - t \le f(x), \quad x \in [a, b], \\ & a_0 + a_1 x + \cdots + a_n x^n + t \ge f(x), \quad x \in [a, b]. \end{aligned} \tag{2.4}$$

In practical computing of the best polynomial, the solution can be approximated by seeking the optimum over a discrete set of constraints replacing the continuum. Rather than discrete analogues, the original problem can be solved by the Algorithm 2.1 (Remez algorithm).

2.1.2 Dual of best uniform approximation

The dual (infinite dimensional) linear program of (2.4) is a moment matching problem

$$\begin{aligned} \max \quad & \int f \mathrm{d}\mu - \int f \mathrm{d}\nu, \\ \text{s.t.} \quad & \int x^j \mathrm{d}\mu = \int x^j \mathrm{d}\nu, \quad j = 0, \ldots, n, \\ & \int \mathrm{d}\mu + \int \mathrm{d}\nu = 1, \\ & \mu, \nu \text{ are positive measures on } [a, b]. \end{aligned} \tag{2.5}$$

Algorithm 2.1 Remez algorithm.

Input: a continuous function f, an interval $[a, b]$, a degree n.
Output: a polynomial P of degree at most n.

1: Initialize $n+2$ points $a \le x_0 < x_1 < \cdots < x_{n+1} \le b$.
2: **repeat**
3: Solve the system of linear equations

$$f(x_j) - Q_n(x_j) = (-1)^j \delta, \quad j = 0, \ldots, n+1,$$

where $Q_n(x) = \sum_{i=0}^{n} a_i x^i$, with respect to unknowns $\delta, a_0, \ldots, a_n$.
4: Find ξ and d such that

$$|f(\xi) - Q_n(\xi)| = \max_{x \in [a,b]} |f(x) - Q_n(x)| = d.$$

5: Update the sequence $x_0 < \cdots < x_{n+1}$ by replacing one x_j by ξ such that $f - Q_n$ successively changes sign.
6: **until** stopping criterion is satisfied.
7: Report Q_n.

The constraint imposes that the measures μ and ν matching moments up to degree n. In this problem, strong duality can be shown by verifying sufficient conditions in general convex optimization (see [37, pp. 48–50]). In the primal problem, as a consequence of Chebyshev's characterization in Theorem 2.4, only $n+2$ constraints are binding in the optimal solution. Consequently, in the dual problem, the optimal μ^* and ν^* are supported on $n+2$ atoms by complementary slackness. The dual solution can be obtained accordingly from the primal solution.

Theorem 2.6. *Suppose the maximum deviation of the best polynomial P^* in (2.3) is attained at $a \le x_0 < \cdots < x_{n+1} \le b$. The dual optimal solution of (2.5) is*

$$\mu^*\{x_i\} = \frac{w_i}{w_0 + w_1 + \cdots + w_{n+1}}, \quad f(x_i) > P^*(x_i),$$
$$\nu^*\{x_i\} = \frac{w_i}{w_0 + w_1 + \cdots + w_{n+1}}, \quad f(x_i) < P^*(x_i),$$

where $w_i = (\prod_{j \ne i} |x_i - x_j|)^{-1}$.

Proof. By Theorem 2.4, $f(x_i) - P^*(x_i)$ successively changes sign. Hence, μ^* is supported on either $\{x_0, x_2, \ldots\}$ or $\{x_1, x_3, \ldots\}$ and ν^* is supported on

the rest. Denote the maximum deviation of P^* by ϵ. Then $f - P^*$ is almost surely ϵ and $-\epsilon$ under μ^* and ν^*, respectively.

We first verify the feasibility. For each $m \in \{0, 1, \ldots, n\}$, consider a polynomial $P(x) = \sum_i x_i^m \frac{\prod_{j\neq i}(x-x_j)}{\prod_{j\neq i}(x_i-x_j)}$ of degree at most $n+1$. $P(x)$ coincides with x^m on $n+2$ distinct points $x_0, \ldots, x_{n+1}$. Hence $P(x) = x^m$ and $\sum_i \frac{x_i^m}{\prod_{j\neq i}(x_i-x_j)} = 0$.

For optimality it suffices to show a zero duality gap:

$$\int f \mathrm{d}\mu^* - \int f \mathrm{d}\nu^* = \int (f - P^*)\mathrm{d}\mu^* - \int (f - P^*)\mathrm{d}\nu^* = \epsilon.$$

The first equality is due to moment matching constraints. □

Remark 2.1. Alternatively, the achievability part can be argued from an optimization perspective (zero duality gap, see [38, Exercise 8.8.7, p. 236]), or using the Riesz representation of linear operators as in [36], which has been used in [17] and [39].

2.2 Polynomial interpolation

Interpolation is a method of estimating values within the range of given data points. Given a discrete set of data points (x_i, f_i) for $i = 0, \ldots, n$, the interpolation problem consists of finding a simple function Φ such that

$$\Phi(x_i) = f_i, \quad i = 0, \ldots, n.$$

Examples of the simple function Φ include an ordinary polynomial and a trigonometric polynomial of the lowest degree. Interpolation is also a basic tool for the approximation. For a comprehensive survey on related topics, see [40].

2.2.1 Interpolation formulas of Lagrange and Newton

Theorem 2.7. *Given $n+1$ distinct data points (x_i, f_i) for $i = 0, \ldots, n$, there exists a unique interpolating polynomial P of degree at most n such that*

$$P(x_i) = f_i, \quad i = 0, \ldots, n.$$

Proof. Given two interpolating polynomials P and P' of degree at most n, the polynomial $Q = P - P'$ is of degree at most n satisfying $Q(x_i) = 0$ for $i = 0, \ldots, n$. $Q \equiv 0$ and the uniqueness follows. Existence is given by Lagrange or Newton formula discussed next. □

The interpolating polynomial P can be explicitly constructed with the help of Lagrange basis:

$$L_i(x) = \prod_{j \neq i} \frac{x - x_j}{x_i - x_j} = \begin{cases} 1, & x = x_i, \\ 0, & x = x_j, j \neq i. \end{cases}$$

Applying the above basis, we obtain the Lagrange formula

$$P(x) = \sum_{i=0}^{n} f_i L_i(x). \tag{2.6}$$

Alternatively, the Newton formula for the interpolating polynomial is of the form

$$\begin{aligned} P(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \ldots \\ + a_n(x - x_0) \cdots (x - x_{n-1}). \end{aligned} \tag{2.7}$$

The coefficients can be successively calculated by

$$\begin{aligned} f_0 &= P(x_0) = a_0, \\ f_1 &= P(x_1) = a_0 + a_1(x_1 - x_0), \\ &\ldots \end{aligned}$$

In general, they coincide with the divided differences $a_k = f_{0\ldots k}$ that are recursively defined as

$$f_{i_0 i_1 \ldots i_k} = \frac{f_{i_1 \ldots i_k} - f_{i_0 \ldots i_{k-1}}}{x_{i_k} - x_{i_0}}. \tag{2.8}$$

The above recursion can be calculated on the *Neville's diagram* (cf. [41, Section 2.1.2]) shown in Figure 2.1. In Neville's diagram, the k^{th} order divided differences are computed in the k^{th} column, and are determined by the previous column and the interpolation nodes $x_0, \ldots, x_n$. The coefficients in (2.7)

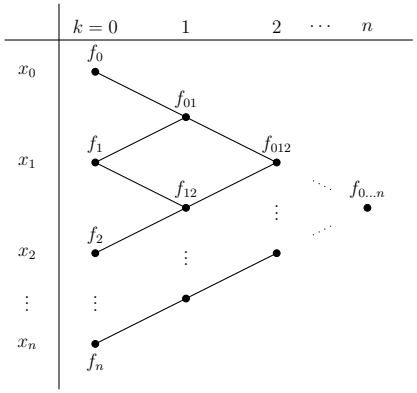


Figure 2.1: Illustration of Neville's digram.

are found in the top diagonal. In Chapter 7 Neville's diagram will be used to bound the coefficients in Newton formula (2.7).

If the data points are with values $f_i = f(x_i)$ of a given function f on a set of distinct points (commonly referred to as *nodes*) $\{x_0, \ldots, x_n\}$, the divided difference $f_{i_0 \ldots i_k}$ can be considered as a multivariate function of the nodes $x_{i_0}, \ldots, x_{i_k}$ and is denoted by $f[x_{i_0}, \ldots, x_{i_k}]$. If f is k^{th} order differentiable, then its k^{th} order divided differences are connected to its derivatives by the Genocchi-Hermite formula

$$f[x_0, \ldots, x_k] = \int_0^1 \int_0^{s_1} \cdots \int_0^{s_{k-1}} f^{(k)}((1-s_1)x_0 + \ldots \\ +(s_{k-1} - s_k)x_{k-1} + s_k x_k) \mathrm{d}s_k \ldots \mathrm{d}s_1. \tag{2.9}$$

Concerning information on this formula and other representations of the divided differences, see [42].

Equation (2.9) can generalize the divided differences by allowing repeated x_i:

$$f[x_0, \ldots, x_k] = \frac{f^{(k)}(x_0)}{k!}, \quad x_0 = x_1 = \cdots = x_k. \tag{2.10}$$

The above generalization will be very useful when incorporating information on derivatives, called Hermite interpolation discussed in Section 2.2.2.

The remainder in the polynomial interpolation can be conveniently expressed in terms of the divided differences

$$R(x) \triangleq f(x) - P(x) = f[x_0, \ldots, x_n, x] \prod_{i=0}^{n} (x - x_i). \tag{2.11}$$

If the function f is $(n+1)^{\text{th}}$ order differentiable, then the remainder term can be represented using (2.9) by

$$R(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{i=0}^{n} (x - x_i), \tag{2.12}$$

for some ξ in the convex hull of $\{x_0, \ldots, x_n, x\}$.

Equation (2.12) can be applied to analyze the approximation error of the interpolation polynomials on a given set of nodes. A special case is on the Chebyshev nodes, which on $[a, b]$ is given by

$$x_i = \frac{b+a}{2} + \frac{b-a}{2} \cos\left(\frac{2k+1}{2n+2}\pi\right), \quad k = 0, \ldots, n,$$

which yields that (see [43, Eq. (4.7.28)])

$$|R(x)| \leq \frac{\max_{x \in [a,b]} |f^{(n+1)}(x)|}{2^n (n+1)!} \left(\frac{b-a}{2}\right)^{n+1}, \quad x \in [a, b]. \tag{2.13}$$

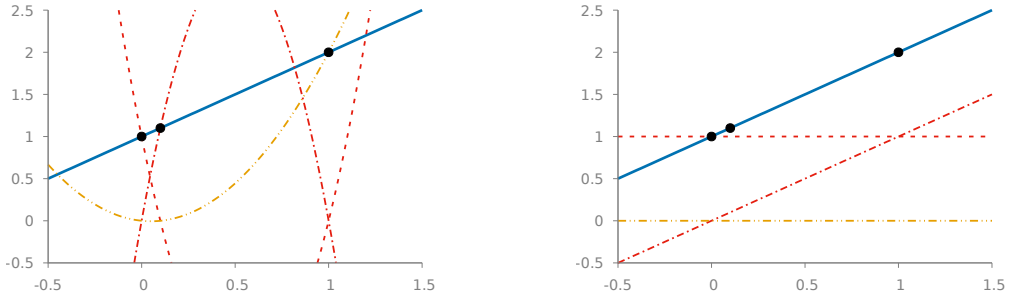
Interpolating polynomials are the main tool to prove moment comparison theorems in Chapter 7. Specifically, we will interpolate step functions by polynomials in order to bound the difference of two CDFs via their moment difference. Therefore, it is critical to have good control over the coefficients of the interpolating polynomial. To this end, it turns out the Newton form is more convenient to use than the Lagrange form because the former takes into account the cancellation between each terms in the polynomial. Indeed, in the Lagrange form (2.6), if two nodes are very close, then each term can be arbitrarily large, even if f itself is a smooth function. In contrast, each term of (2.7) is stable when f is smooth since divided differences are closely related to derivatives. The following example and Figure 2.2 illustrate this point.

Example 2.1 (Lagrange versus Newton form)**.** Given three points $x_1 = 0, x_2 = \epsilon, x_3 = 1$ with $f(x_1) = 1, f(x_2) = 1 + \epsilon, f(x_3) = 2$, the interpolating polynomial is $P(x) = x + 1$. The next equation gives the interpolating

polynomial in Lagrange's and Newton's form respectively.

$$\text{Lagrange: } P(x) = \frac{(x-\epsilon)(x-1)}{\epsilon} + (1+\epsilon)\frac{x(x-1)}{\epsilon(\epsilon-1)} + 2\frac{x(x-\epsilon)}{1-\epsilon};$$

$$\text{Newton: } P(x) = 1 + x + 0.$$



(a) Lagrange formula

(b) Newton formula

Figure 2.2: Interpolation on three data points $(0, 1)$, $(0.1, 1.1)$, and $(1, 2)$ in black dots. (a) Illustration of three terms in Lagrange formula in dashed lines, and the interpolating polynomial in the solid line as a summation of three terms. (b) Illustration of three terms in Newton formula in dashed lines, and the same interpolating polynomial.

Although we will mainly use Newton formula in Part II, this is not to say Lagrange formula has no advantage. Lagrange formula is theoretically important in the development of numerical analysis. While it was rarely used in practice for many years, a variant of this formula is recently found to be practically advantageous and is widely implemented, especially in situations where the interpolating nodes x_i are fixed. See [44].

2.2.2 Hermite interpolation

Polynomial interpolation can be generalized to interpolate the value of derivatives, known as the *Hermite interpolation*.

Theorem 2.8. *Given $n+1$ distinct real numbers $x_0 < x_1 < \cdots < x_n$, and values $f_i^{(k)}$ for $i = 0, \ldots, n$ and $k = 0, \ldots, m_i$, there exists a unique polynomial of degree at most $N = n + \sum_i m_i$ such that*

$$P^{(k)}(x_i) = f_i^{(k)}, \quad i = 0, \ldots, n, \ k = 0, \ldots, m_i.$$

Proof. Given two interpolating polynomials P and P' of degree at most N, the polynomial $Q = P - P'$ is of degree at most N and satisfies $Q^{(k)}(x_i) = 0$ for $i = 0, \ldots, n$ and $k = 0, \ldots, m_i$. Therefore, each x_i is a root of Q of multiplicities $m_i + 1$. Since $\sum_i (m_i + 1) > N$, $Q \equiv 0$, and the uniqueness follows. The existence is given by the generalized Lagrange or Newton formula introduced next. □

Analogous to the Lagrange formula (2.6), the interpolating polynomial can be explicitly constructed with the help of the generalized Lagrange polynomials $L_{i,k}$ satisfying

$$L_{i,k}^{(k')}(x_{i'}) = \begin{cases} 1, & i = i', k = k', \\ 0, & \text{otherwise.} \end{cases}$$

For an explicit formula of the generalized Lagrange polynomials, see [41, pp. 52–53]. The Hermite interpolating polynomial can be simply written as

$$P(x) = \sum_{i,k} f_i^{(k)} L_{i,k}(x).$$

The procedure to evaluate the generalized Lagrange polynomials is tedious even for a small number of data points. The Newton formula (2.7) can also be extended by using generalized divided differences, which, for repeated nodes, is defined as the value of the derivative:

$$f_{i \ldots i+r} = \frac{f^{(r)}(x_0)}{r!}, \quad x_i = x_{i+1} = \ldots = x_{i+r}. \tag{2.14}$$

To this end, we define an expanded sequence by replacing each x_i for k_i times:

$$\underbrace{x_0 = \ldots = x_0}_{k_0} < \underbrace{x_1 = \ldots = x_1}_{k_1} < \ldots < \underbrace{x_m = \ldots = x_m}_{k_m}. \tag{2.15}$$

The Hermite interpolating polynomial is obtained by (2.7) using this new sequence and generalized divided differences, which can also be calculated from the Neville's diagram by replacing differences by derivatives whenever encountering repeated nodes. When the data points are from a given function f, the remainder equations (2.11) and (2.12) remain valid. Below we give an example using Hermite interpolation to construct polynomial majorization,

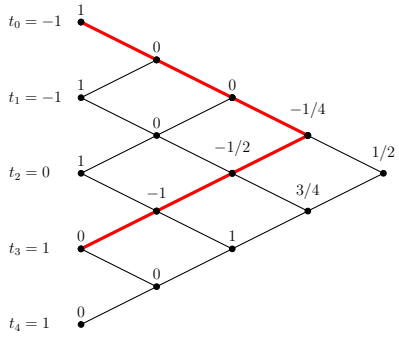
Table 2.1: Interpolation values of f.

x	-1	0	1
$P(x)$	1	1	0
$P'(x)$	0	any	0

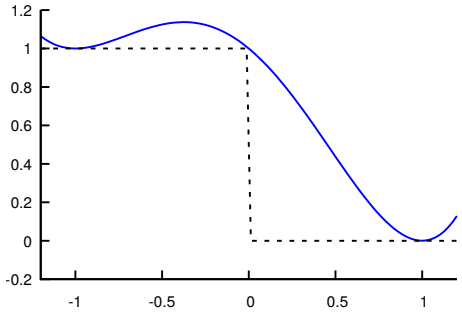
which will be used to prove moment comparison theorems in Chapter 7.

Example 2.2 (Hermite interpolation as polynomial majorization)**.** Let $f(x) = \mathbf{1}_{\{x\le 0\}}$. We want to find a polynomial majorization $P \ge f$ such that $P(x) = f(x)$ on $x = \pm 1$. To this end we interpolate f on $\{-1, 0, 1\}$ with values in Table 2.1. The Hermite interpolation P is of degree four and majorizes f [45, p. 65]. To see this, we note that $P'(\xi) = 0$ for some $\xi \in (-1, 0)$ by Rolle's theorem. Since $P'(-1) = P'(1) = 0$, P has no other stationary point than $-1, \xi, 1$, and thus decreases monotonically in $(\xi, 1)$. Hence, $-1, 1$ are the only local minimum points of P, and thus $P \ge f$ everywhere. The polynomial P is shown in Figure 2.3b.

To explicitly construct the polynomial, we have an expanded sequence $-1, -1, 0, 1, 1$ by (2.15). Applying Newton formula (2.7) with generalized divided differences from the Neville's diagram Figure 2.3a, we obtain that $P(x) = 1 - \frac{1}{4}x(x+1)^2 + \frac{1}{2}x(x+1)^2(x-1)$.



(a) Neville's diagram.



(b) Hermite interpolation.

Figure 2.3: Neville's diagram and Hermite interpolation. In (a), values are recursively calculated from left to right. For example, the red thick line shows that $f[-1, -1, 0, 1]$ is calculated by $\frac{-1/2-0}{1-(-1)} = -1/4$.

2.3 Moments and positive polynomials

The n^{th} moment vector of a distribution μ is an n-tuple

$$\mathbf{m}_n(\mu) = (m_1(\mu), \ldots, m_n(\mu)).$$

The n^{th} moment space on $K \subseteq \mathbb{R}$ is defined as

$$\mathcal{M}_n(K) = \{\mathbf{m}_n(\mu) : \mu \text{ is supported on } K\},$$

which is the convex hull of $\{(x, x^2, \ldots, x^r) : x \in K\}$. This convex set satisfies many geometric constraints such as the Cauchy-Schwarz and Hölder inequalities. A complete description can be phrased in terms of positive polynomials by the next theorem. Note that a sequence of numbers $(m_1, m_2, \ldots)$ can be viewed as values of a linear functional L such that $L(x^j) = m_j$. It is in the full moment space, i.e., the first n numbers is in the n^{th} moment space for every n, if there exists a representation measure μ such that $L(p) = \int p \mathrm{d}\mu$ for every polynomial p. Apparently, if the sequence is in the moment space, then for every positive polynomial $p \geq 0$ we have $L(p) \geq 0$. The next theorem shows that the converse also holds.

Theorem 2.9 (Riesz-Haviland)**.** *Let $K \subseteq \mathbb{R}$ be closed. If L is a linear functional such that $L(p) \geq 0$ for every positive polynomial $p \geq 0$ on K, then there exists a representing measure μ for L, i.e., $L(p) = \int p \mathrm{d}\mu$ for every polynomial p.*

A truncated sequence $(m_1, \ldots, m_n)$ can be similarly viewed as values of a linear functional on $\mathcal{P}_n$, the set of all polynomials of degree at most n, such that $L(x^j) = m_j$. The truncated moment space can also be characterized in terms of positive polynomials by Theorem 2.10.

Theorem 2.10. *Let $K \subseteq \mathbb{R}$ be compact. If L is a linear functional on $\mathcal{P}_n$ such that $L(p) \geq 0$ for every $p \geq 0$ on K, then there exists a representing measure μ for L, i.e., $L(p) = \int p \mathrm{d}\mu$ for every $p \in \mathcal{P}_n$.*

The above theorems can be generalized to multiple dimensions. For the proofs of these theorems, see [10, pp. 17–18]. However, an efficient characterization of positive polynomials is not known in general in multiple dimensions. On the real line, positive polynomials have representations using sum

of squares in the next proposition ([10, Propositions 3.1–3.3]). We denote the set of finite sum of squares p^2, where $p \in \mathcal{P}_n$, by $\mathcal{S}_n^2$.

Proposition 2.1.

- $p \geq 0$ *on* $\mathbb{R}$, $\deg(p) = 2n \Rightarrow p(x) = f(x)^2 + g(x)^2$, $f, g \in \mathcal{P}_n$.
- $p \geq 0$ *on* $[0, \infty)$, $\deg(p) = 2n \Rightarrow p(x) = f(x) + xg(x)$, $f \in \mathcal{S}_n^2, g \in \mathcal{S}_{n-1}^2$.
- $p \geq 0$ *on* $[0, \infty)$, $\deg(p) = 2n+1 \Rightarrow p(x) = f(x) + xg(x)$, $f, g \in \mathcal{S}_n^2$.
- $p \geq 0$ *on* $[a, b]$, $\deg(p) = 2n \Rightarrow p(x) = f(x) + (b-x)(x-a)g(x)$, $f \in \mathcal{S}_n^2, g \in \mathcal{S}_{n-1}^2$.
- $p \geq 0$ *on* $[a, b]$, $\deg(p) = 2n+1 \Rightarrow p(x) = (b-x)f(x) + (x-a)g(x)$, $f, g \in \mathcal{S}_n^2$.

Using the above results, next we derive the characterization of the truncated moment space on a compact interval $K = [a, b]$, namely $\mathcal{M}_n([a, b])$ that was obtained in [7, Theorem 3.1]. Other cases can be obtained analogously (see [10, Parts II–III] and [9, Chapter 3]). To state the result we abbreviate the Hankel matrix with entries $m_i, m_{i+1}, \ldots, m_j$ by

$$\mathbf{M}_{i,j} = \begin{bmatrix} m_i & m_{i+1} & \cdots & m_{\frac{i+j}{2}} \\ m_{i+1} & m_{i+2} & \cdots & m_{\frac{i+j}{2}+1} \\ \vdots & \vdots & \ddots & \vdots \\ m_{\frac{i+j}{2}} & m_{\frac{i+j}{2}+1} & \cdots & m_j \end{bmatrix}.$$

The matrix $\mathbf{M}_n = \mathbf{M}_{0,n}$ is also referred to as the moment matrix of order n, a Hankel matrix of size $(n+1) \times (n+1)$.

Theorem 2.11. *A vector* $\mathbf{m}_n = (m_1, \ldots, m_n)$ *is in the moment space* $\mathcal{M}_n([a, b])$ *if and only if*

$$\begin{cases} \mathbf{M}_n \succeq 0, \quad (a+b)\mathbf{M}_{1,n-1} \succeq ab\mathbf{M}_{n-2} + \mathbf{M}_{2,n}, & n \text{ even}, \\ b\mathbf{M}_{n-1} \succeq \mathbf{M}_{1,n} \succeq a\mathbf{M}_{n-1}, & n \text{ odd}. \end{cases} \tag{2.16}$$

Proof. If n is even, by Theorem 2.10 and Proposition 2.1, $\mathbf{m}_n \in \mathcal{M}_n([a, b])$ if and only if $L(p^2) \geq 0$ for every $p \in \mathcal{P}_n$ and $L((b-x)(a-x)q^2(x)) \geq 0$

for every $q \in \mathcal{P}_{n-1}$. These are equivalent to $\mathbf{M}_{0,n} \succeq 0$ and $(a+b)\mathbf{M}_{1,n-1} \succeq ab\mathbf{M}_{0,n-2} + \mathbf{M}_{2,n}$, respectively.

If n is odd, then $\mathbf{m}_n \in \mathcal{M}_n([a,b])$ if and only if $L((x-a)p^2(x)) \geq 0$ and $L((b-x)p^2(x)) \geq 0$ for every $p \in \mathcal{P}_n$. These are equivalent to $b\mathbf{M}_{0,n-1} \succeq \mathbf{M}_{1,n} \succeq a\mathbf{M}_{0,n-1}$. □

Remark 2.2. Alternatively, the above characterization of the moment space can be obtained from the recursive properties of Hankel matrices. See [46].

Example 2.3 (Moment spaces on $[0,1]$)**.** $\mathcal{M}_2([0,1])$ is simply described by $m_1 \geq m_2 \geq 0$ and $m_2 \geq m_1^2$. $\mathcal{M}_3([0,1])$ is described by

$$\begin{bmatrix} 1 & m_1 \\ m_1 & m_2 \end{bmatrix} \succeq \begin{bmatrix} m_1 & m_2 \\ m_2 & m_3 \end{bmatrix} \succeq 0.$$

Using Sylvester's criterion (see [47, Theorem 7.2.5]), they are equivalent to

$$\begin{aligned} &0 \leq m_1 \leq 1, \quad m_2 \geq m_3 \geq 0, \\ &m_1 m_3 \geq m_2^2, \quad (1-m_1)(m_2-m_3) \geq (m_1-m_2)^2. \end{aligned}$$

The necessity of the above inequalities is apparent: the first two follow from the range $[0,1]$, and the last two follow from the Cauchy-Schwarz inequality. It turns out that they are also sufficient.

Moment matrices of discrete distributions satisfy more structural properties. For instances, the moment matrix of a k-atomic distribution of any order is of rank at most k, and is a deterministic function of $\mathbf{m}_{2k-1}$; the number of atoms can be characterized using the determinants of moment matrices (see [48, p. 362] or [49, Theorem 2A]) by Theorem 2.12.

Theorem 2.12. *A sequence $m_1, \ldots, m_{2r}$ is the moments of a distribution with exactly r points of support if and only if* $\det(\mathbf{M}_{r-1}) > 0$ *and* $\det(\mathbf{M}_r) = 0$.

2.4 Orthogonal polynomials and Gauss quadrature

The theory of orthogonal polynomials is another classical topic in the theory of polynomials. The trigonometric polynomials used in Fourier analysis is one

set of orthogonal polynomials on the unit circle. In general the orthogonality of functions is defined as follows.

Definition 2.1. A set of functions $\{f_1, \ldots, f_n\}$ is orthogonal under the positive measure μ if

$$\mathbb{E}_\mu[f_i f_j] = \int f_i f_j \mathrm{d}\mu = 0, \quad i \neq j.$$

Given a set of linear independent functions, an orthogonal set can be obtained by the Gram-Schmidt orthogonalization process. In Section 2.4.1 we will review some classical orthogonal polynomials under commonly used measures. Here we present the Gauss quadrature, an algorithm to find a representing measure for a given vector of moments, that is based on the general theory of orthogonal polynomials.

Gauss quadrature is a discrete approximation for a given distribution in the sense of moments and plays an important role in the execution of our Gaussian mixture estimator in Chapter 8. Given μ supported on $K \subseteq \mathbb{R}$, a k-point Gauss quadrature is a k-atomic distribution $\mu_k = \sum_{i=1}^k w_i \delta_{x_i}$, also supported on K, such that, for any polynomial P of degree at most $2k-1$,

$$\mathbb{E}_\mu P = \mathbb{E}_{\mu_k} P = \sum_{i=1}^k w_i P(x_i). \tag{2.17}$$

Gauss quadrature is known to always exist and is uniquely determined by $\mathbf{m}_{2k-1}(\mu)$ (cf. e.g. [41, Section 3.6]), which shows that any valid moment vector of order $2k-1$ can be realized by a unique k-atomic distribution. A basic algorithm to compute Gauss quadrature is Algorithm 2.2 [50] and many algorithms with improved computational efficiency and numerical stability have been proposed; cf. [51, Chapter 3].

We briefly show the correctness of Algorithm 2.2. Note that in (2.18), Φ is a polynomial of degree at most k. If $(m_1, \ldots, m_{2k-1})$ is the moments of a distribution μ, then Φ is orthogonal to all polynomial $P \in \mathcal{P}_{k-1}$ under μ (by expanding the determinant with respect to the last row of (2.18) and taking

Algorithm 2.2 Quadrature rule.

Input: a vector of $2k-1$ moments $(m_1, \ldots, m_{2k-1})$.
Output: nodes $x = (x_1, \ldots, x_k)$ and weights $w = (w_1, \ldots, w_k)$.

1: Define the following degree-k polynomial Φ

$$\Phi(x) = \det \begin{bmatrix} 1 & m_1 & \cdots & m_k \\ \vdots & \vdots & \ddots & \vdots \\ m_{k-1} & m_k & \cdots & m_{2k-1} \\ 1 & x & \cdots & x^k \end{bmatrix}. \tag{2.18}$$

2: Let the nodes $(x_1, \ldots, x_k)$ be the roots of the polynomial Φ.
3: Let the weights $w = (w_1, \ldots, w_k)$ be

$$w = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_k \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{k-1} & x_2^{k-1} & \cdots & x_k^{k-1} \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ m_1 \\ \vdots \\ m_{k-1} \end{bmatrix}.$$

expectations):

$$\mathbb{E}[\Phi(X)X^j] = \det \begin{bmatrix} 1 & m_1 & \cdots & m_k \\ \vdots & \vdots & \ddots & \vdots \\ m_{k-1} & m_k & \cdots & m_{2k-1} \\ m_j & m_{j+1} & \cdots & m_{j+k} \end{bmatrix} = 0, \quad j \le k-1.$$

For any polynomial P of degree $2k-1$, we have

$$P(x) = \Phi(x)Q(x) + R(x),$$

where Q, R are polynomials of degree at most $k-1$. The polynomial R can be expressed by the Lagrangian interpolation formula

$$R(x) = \sum_{i=1}^{k} R(x_i) \frac{\prod_{j \neq i}(x - x_j)}{\prod_{j \neq i}(x_i - x_j)} = \sum_{i=1}^{k} P(x_i) \frac{\prod_{j \neq i}(x - x_j)}{\prod_{j \neq i}(x_i - x_j)}.$$

The following quadrature rule follows immediately from orthogonality:

$$\mathbb{E}[P(X)] = \sum_{i=1}^{k} w_i P(x_i), \quad w_i \triangleq \frac{\mathbb{E}\prod_{j \neq i}(X - x_j)}{\prod_{j \neq i}(x_i - x_j)}.$$

It is necessary that $w_i \geq 0$. Consider the squared Lagrange basis $P_i(x) = \frac{\prod_{j\neq i}(x-x_j)^2}{\prod_{j\neq i}(x_i-x_j)^2}$, which is a non-negative polynomial of degree $2k-2$. Then, by the quadrature rule,

$$0 \leq \mathbb{E}[P(X)] = \sum_{j=1}^{k} w_j P_i(x_j) = w_i.$$

2.4.1 Classical orthogonal polynomials

In this subsection, we present some classical orthogonal polynomials along with some properties that will be used in this dissertation.

Chebyshev polynomials Chebyshev polynomial T_n of degree n is defined as

$$T_n(x) = \cos(n \arccos x) = (z^L + z^{-L})/2, \tag{2.19}$$

where z is the solution of the quadratic equation $z + z^{-1} = 2x$. They are orthogonal with respect to the weight function $(1-x^2)^{-1/2}$:

$$\begin{aligned}\int_{-1}^{1} T_n(x)T_m(x)(1-x^2)^{-1/2}\mathrm{d}x &= \int_0^{\pi} \cos(n\theta)\cos(m\theta)\mathrm{d}\theta \\ &= \begin{cases} 0, & n \neq m, \\ \pi, & n = m = 0, \\ \pi/2, & n = m \neq 0. \end{cases}\end{aligned}$$

They have the following algebraic formula:

$$T_n(x) = \frac{n}{2}\sum_{k=0}^{\lfloor n/2 \rfloor} (-1)^k \frac{(n-k-1)!}{k!(n-2k)!}(2x)^{n-2k}.$$

Hermite polynomials Denote the Hermite polynomial of degree n by H_n. They are orthogonal under the standard normal distribution, i.e., for $Z \sim N(0,1)$,

$$\mathbb{E}[H_n(Z)H_n(Z)] = \int H_n(x)H_m(x)\phi(x)\mathrm{d}x = \begin{cases} n! & n = m, \\ 0 & n \neq m, \end{cases} \tag{2.20}$$

where $\phi(x) \triangleq \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ denote the standard normal density, and H_n has the following formula

$$H_n(x) = \mathbb{E}(x + \mathbf{i}Z)^n = n! \sum_{j=0}^{\lfloor n/2 \rfloor} \frac{(-1/2)^n}{n!(n-2j)!} x^{n-2j}, \tag{2.21}$$

where $\mathbf{i} = \sqrt{-1}$. Hermite polynomials are the unique unbiased estimate of the normal mean:

$$\mathbb{E}[H_n(\mu + Z)] = \mu^n.$$

The exponential generating function of Hermite polynomials is [52, 22.9.17]

$$\sum_{j \geq 0} H_j(x) \frac{u^j}{j!} = \frac{\phi(x-u)}{\phi(x)} = e^{-\frac{u^2}{2} + xu}.$$

Laguerre polynomials The Laguerre polynomials are orthogonal under the exponential distribution (i.e., with respect to the weight function e^{-x}) with the following close-form formula:

$$\mathcal{L}_n(x) = \sum_{k=0}^{n} \binom{n}{k} \frac{(-x)^k}{k!}. \tag{2.22}$$

Denote the degree-n generalized Laguerre polynomial by $\mathcal{L}_n^{(k)}$ that can be obtained from Rodrigues representation:

$$\mathcal{L}_n^{(k)}(x) = \frac{x^{-k} e^x}{n!} \frac{\mathrm{d}^n}{\mathrm{d}x^n} (e^{-x} x^{n+k}) = (-1)^k \frac{\mathrm{d}^x}{\mathrm{d}k^x} L_{n+k}(x), \quad k \in \mathbb{N}. \tag{2.23}$$

Then the simple Laguerre polynomials are $\mathcal{L}_n(x) = \mathcal{L}_n^{(0)}$. The orthogonality is given by

$$\int_0^\infty x^k e^{-x} \mathcal{L}_n^{(k)}(x) \mathcal{L}_m^{(k)}(x) = \begin{cases} \frac{\Gamma(n+k+1)}{n!}, & n = m, \\ 0, & n \neq m. \end{cases}$$

The Laguerre polynomials have the following upper bound [52, 22.14.13]

$$|L_n^{(k)}(x)| \leq \binom{n+k}{n} e^{x/2}, \quad x \geq 0,\ k \in \mathbb{N}. \tag{2.24}$$

Laguerre polynomials also appear in the second moments of factorial mo-

ments under Poisson distribution in Chapter 4. The factorial moment is defined as

$$(x)_m \triangleq \frac{x!}{(x-m)!},$$

which gives an unbiased estimator for the monomials of the Poisson mean: $\mathbb{E}[(X)_m] = \lambda^m$ where $X \sim \text{Poi}(\lambda)$. The second moments of $(X)_m$ can be expressed in terms of Laguerre polynomials. Using the probability mass function of the Poisson distribution, we can explicitly compute $\mathbb{E}(X)_m^2$:

$$\mathbb{E}(X)_m^2 = \sum_{j=m}^{\infty} \frac{e^{-\lambda}\lambda^j}{j!} \frac{j!^2}{(j-m)!^2} = \lambda^m m! \sum_{j=0}^{\infty} \frac{e^{-\lambda}\lambda^j}{j!} \frac{(j+m)!}{j!m!}.$$

The summation in the right-hand side can be expressed as an expectation of a binomial coefficient

$$\mathbb{E}\binom{X+m}{X} = \sum_{k=0}^{m} \binom{m}{k} \mathbb{E}\binom{X}{X-k} = \sum_{k=0}^{m} \binom{m}{k} \frac{\mathbb{E}(X)_k}{k!}.$$

Again using $\mathbb{E}(X)_k = \lambda^k$, we obtain that

$$\mathbb{E}(X)_m^2 = \lambda^m m! \sum_{k=0}^{m} \binom{m}{k} \frac{\lambda^k}{k!} = \lambda^m m! \mathcal{L}_m(-\lambda). \tag{2.25}$$

Discrete Chebyshev polynomials The discrete Chebyshev polynomials, denoted by $\{t_0, \ldots, t_{M-1}\}$, are orthogonal with respect to the counting measure over the discrete set $\{0, 1, \ldots, M-1\}$ with the following formula [53, Sec. 2.8]: for $x = 0, 1, \ldots, M-1$,

$$t_m(x) \triangleq \frac{1}{m!}\Delta^m p_m(x) = \frac{1}{m!} \sum_{j=0}^{m} (-1)^j \binom{m}{j} p_m(x+m-j), \quad 0 \le m \le M-1, \tag{2.26}$$

where

$$p_m(x) \triangleq x(x-1)\cdots(x-m+1)(x-M)(x-M-1)\cdots(x-M-m+1), \tag{2.27}$$

and Δ^m denotes the m-th order forward difference. The orthogonality is given by (cf. [53, Sec. 2.8.2, 2.8.3]):

$$\sum_{x=0}^{M-1} t_m(x)t_\ell(x) = 0, \quad m \neq \ell,$$
$$\sum_{x=0}^{M-1} t_m^2(x) = \frac{M(M^2-1^2)(M^2-2^2)\cdots(M^2-m^2)}{2m+1} \triangleq c(M,m).$$

2.4.2 Gauss quadrature of the standard normal distribution

In this subsection we present a few properties of the Gauss quadrature of the standard normal distribution that will be used in Chapter 8.

Lemma 2.1. *Let g_k be the k-point Gauss quadrature of $N(0,\sigma^2)$. For $j \geq 2k$, we have $m_j(g_k) \leq m_j(N(0,\sigma^2))$ when j is even, and $m_j(g_k) = m_j(N(0,\sigma^2)) = 0$ otherwise. In particular, g_k is σ-subgaussian.*

Proof. By scaling it suffices to consider $\sigma = 1$. Let $\nu = N(0,1)$. If j is odd, $m_j(g_k) = m_j(\nu) = 0$ by symmetry. If $j \geq 2k$ and j is even, the conclusion follows from the integral representation of the error term of Gauss quadrature (see, e.g., [41, Theorem 3.6.24]):

$$m_j(\nu) - m_j(g_k) = \frac{f^{(2k)}(\xi)}{(2k)!}\int \pi_k^2(x)\mathrm{d}\nu(x),$$

for some $\xi \in \mathbb{R}$; here $f(x) = x^j$, $\{x_1,\ldots,x_k\}$ is the support of g_k, and $\pi_k(x) \triangleq \prod_i (x - x_i)$. Consequently, g_k is 1-subgaussian [54, Lemma 2]. □

Lemma 2.2. *Let g_k be the k-point Gauss quadrature of $N(0,1)$. Then*

$$\mathbb{E}_{g_k}|X| \geq (4k+2)^{-1/2}, \quad k \geq 2.$$

Proof. Let $G_k \sim g_k$. Note that $|G_k| \leq \sqrt{4k+2}$ using the bound on the zeros of Hermite polynomials [53, p. 129]. The conclusion follows from $1 = \mathbb{E}[G_k^2] \leq \mathbb{E}|G_k|\sqrt{4k+2}$. □

Lemma 2.3. *Let g_k be the k-point Gauss quadrature of $N(0,1)$. Then $\mathbb{E}_{g_k}[H_j] = 0$ for $j = 1,\ldots,2k-1$, and $\mathbb{E}_{g_k}[H_{2k}] = -k!$, where H_j is the Hermite polynomial of degree j (see (2.21)).*

Proof. Let $Z \sim N(0,1)$ and $G_k \sim g_k$. By orthogonality of Hermite polynomials (2.20) we have $\mathbb{E}[H_j(Z)] = 0$ for all $j \geq 1$ and thus $\mathbb{E}[H_j(G_k)] = 0$ for $j = 1, \ldots, 2k-1$. Expand $H_k^2(x)$ as

$$H_k^2(x) = H_{2k}(x) + a_{2k-1}H_{2k-1}(x) + \cdots + a_1 H_1(x) + a_0.$$

Since G_k is supported on the zeros of H_k, we have $0 = \mathbb{E}[H_k^2(G_k)] = \mathbb{E}[H_{2k}(G_k)] + a_0$. The conclusion follows from $k! = \mathbb{E}[H_k^2(Z)] = a_0$ (see (2.20)). $\square$

Part I

Property Estimation

CHAPTER 3

POLYNOMIAL APPROXIMATION IN STATISTICAL INFERENCE

In this part, we apply the polynomial approximation method in the estimation of scalar properties $T(P)$ of a distribution P, including the Shannon entropy and the support size. In this chapter, we begin with an investigation on the common techniques that will be used in Part I.

3.1 Poisson sampling

Let P be a distribution over an alphabet of cardinality k. Let $X_1, \ldots, X_n$ be i.i.d. samples drawn from P. Without loss of generality, we shall assume that the alphabet is $[k] \triangleq \{1, \ldots, k\}$. To perform statistical inference on the unknown distribution P or any functional thereof, a sufficient statistic is the histogram $N \triangleq (N_1, \ldots, N_k)$, where

$$N_j = \sum_{i=1}^{n} \mathbf{1}_{\{X_i = j\}} \tag{3.1}$$

records the number of occurrences of $j \in [k]$ in the sample. Then $N \sim$ multinomial(n, P). To investigate the decision-theoretic fundamental limit (1.1), we consider the minimax quadratic risk:

$$R^*(k, n) \triangleq \inf_{\hat{T}} \sup_{P \in \mathcal{M}_k} \mathbb{E}(\hat{T} - T(P))^2,$$

where $\hat{T}$ is an estimator measurable with respect to n independent samples, and $\mathcal{M}_k$ denotes the set of probability distributions on $[k]$.

The multinomial distribution of the sufficient statistic $N = (N_1, \ldots, N_k)$ is difficult to analyze because of the dependency. A commonly used technique is the so-called *Poisson sampling* where we relax the sample size n from being deterministic to a Poisson random variable n' with mean n. Under

this model, we first draw the sample size $n' \sim \mathrm{Poi}(n)$, then draw n' i.i.d. samples from the distribution P. The main benefit is that now the sufficient statistics $N_i \overset{\text{ind}}{\sim} \mathrm{Poi}(np_i)$ are independent, which significantly simplifies the analysis.

Analogous to the minimax risk under multinomial sampling, we define its counterpart under the Poisson sampling model:

$$\tilde{R}^*(k,n) \triangleq \inf_{\hat{T}} \sup_{P \in \mathcal{M}_k} \mathbb{E}(\hat{T} - T(P))^2, \tag{3.2}$$

where $\hat{T}$ is an estimator measurable with respect to $N_i \overset{\text{ind}}{\sim} \mathrm{Poi}(np_i)$ for $i = 1, \ldots, k$. In view of the exponential tail of Poisson distributions, the Poissonized sample size is concentrated near its mean n with high probability, which guarantees that the minimax risk under Poisson sampling is provably close to that with fixed sample size. Indeed, the next theorem allows us to focus on the risk of the Poisson model.

Theorem 3.1. *Let* $R_k^* = R^*(k,0) = (\frac{\sup T(P) - \inf T(P)}{2})^2$ *where* $P \in \mathcal{M}_k$. *For any* $\alpha > 0$ *and* $0 < \beta < 1$,

$$\tilde{R}^*(k,(1+\alpha)n) - R_k^* e^{-n\alpha^2/4} \le R^*(k,n) \le \frac{\tilde{R}^*(k,(1-\beta)n)}{1 - \exp(-n\beta^2/2)}. \tag{3.3}$$

Proof. We first prove the right inequality. We follow the same idea as in [55, Appendix A] using the Bayesian risk as a lower bound of the minimax risk with a more refined application of the Chernoff bound. We express the risk under the Poisson sampling as a function of the original samples that

$$\tilde{R}^*(k,(1-\beta)n) = \inf_{\{\hat{T}_m\}} \sup_{P \in \mathcal{M}_k} \mathbb{E}[\ell(\hat{T}_{n'}, T(P))],$$

where $\{\hat{T}_m\}$ is a sequence of estimators, $n' \sim \mathrm{Poi}((1-\beta)n)$ and $\ell(x,y) \triangleq (x-y)^2$ is the loss function. The Bayesian risk is a lower bound of the minimax risk:

$$\tilde{R}^*(k,(1-\beta)n) \ge \sup_{\pi} \inf_{\{\hat{T}_m\}} \mathbb{E}[\ell(\hat{T}_{n'}, T(P))], \tag{3.4}$$

where π is a prior over the parameter space $\mathcal{M}_k$. For any sequence of esti-

mators $\{\hat{T}_m\}$,

$$\mathbb{E}[\ell(\hat{T}_{n'}, T)] = \sum_{m\geq 0} \mathbb{E}[\ell(\hat{T}_m, T)]\mathbb{P}[n' = m] \geq \sum_{m=0}^{n} \mathbb{E}[\ell(\hat{T}_m, T)]\mathbb{P}[n' = m].$$

Taking infimum of both sides, we obtain

$$\begin{aligned}\inf_{\{\hat{T}_m\}} \mathbb{E}[\ell(\hat{T}_{n'}, T)] &\geq \inf_{\{\hat{T}_m\}} \sum_{m=0}^{n} \mathbb{E}[\ell(\hat{T}_m, T)]\mathbb{P}[n' = m] \\ &= \sum_{m=0}^{n} \inf_{\hat{T}_m} \mathbb{E}[\ell(\hat{T}_m, T)]\mathbb{P}[n' = m].\end{aligned}$$

Note that for any fixed prior π, the function $m \mapsto \inf_{\hat{T}_m} \mathbb{E}[\ell(\hat{T}_m, T)]$ is decreasing. Therefore

$$\begin{aligned}\inf_{\{\hat{T}_m\}} \mathbb{E}[\ell(\hat{T}_{n'}, T)] &\geq \inf_{\hat{T}_n} \mathbb{E}[\ell(\hat{T}_n, T)]\mathbb{P}[n' \leq n] \\ &\geq \inf_{\hat{T}_n} \mathbb{E}[\ell(\hat{T}_n, T)](1 - \exp(n(\beta + \log(1-\beta)))) \\ &\geq \inf_{\hat{T}_n} \mathbb{E}[\ell(\hat{T}_n, T)](1 - \exp(-n\beta^2/2)), \qquad (3.5)\end{aligned}$$

where we used the Chernoff bound (see, e.g., [56, Theorem 5.4]) and the fact that $\log(1-x) \leq -x - x^2/2$ for $x > 0$. Taking supremum over π on both sides of (3.5), the conclusion follows from (3.4) and the minimax theorem (cf. e.g. [57, Theorem 46.5]).

Next we prove the left inequality of (3.3). Recall that $0 \leq R^*(k,m) \leq R^*(k,0)$ and $m \mapsto R^*(k,m)$ is decreasing. Therefore,

$$\begin{aligned}\tilde{R}^*(k, (1+\alpha)n) &\leq \sum_{m>n} R^*(k,m)\mathbb{P}[n' = m] + \sum_{0\leq m\leq n} R^*(k,m)\mathbb{P}[n' = m] \\ &\leq R^*(k,n) + R^*(k,0)\mathbb{P}[n' \leq n] \\ &\leq R^*(k,n) + R^*(k,0)\exp(-n(\alpha - \log(1+\alpha))) \\ &\leq R^*(k,n) + R_k^*\exp(-n\alpha^2/4),\end{aligned}$$

where $n' \sim \mathrm{Poi}((1+\alpha)n)$ and we used the Chernoff bound and the fact that $\log(1+x) \leq x - x^2/4$ for $0 < x < 1$. □

3.2 Functional estimation on large alphabets via polynomial approximation

Functional estimation is a common task in statistical inference. As shown in Figure 1.1, given data from an unknown distribution, the quantity of interest is a function of that distribution rather than the high-dimensional parameters or the entire density. For instance, in operations management the optimal inventory level is a function of the distribution of the random demand in the future. To estimate a function of a distribution, one natural idea is a two-step approach: first estimate the distribution and then substitute into the function, called the *plug-in* approach. However, this approach often suffers from large bias [18, 19]. It is natural to expect that estimating a functional is simpler than the entire distribution in the sense of lower sample complexity. In this section, rather than reducing to a more complicated problem, we describe the polynomial approximation methods to directly estimate a functional.

Functional estimation on large alphabets with insufficient samples has a rich history in information theory, statistics and computer science, with early contributions dating back to Fisher [58], Good and Turing [59], Efron and Thisted [21] and recent renewed interest in compression, prediction, classification and estimation aspects for large-alphabet sources [60, 61, 62, 63, 64]. However, none of the current results allow a general understanding of the fundamental limits of functional estimation on large alphabets. The particularly interesting case is when the sample size scales *sublinearly* with the alphabet size.

In Part I, the design of optimal estimator and the proof of a matching minimax lower bound both rely on the apparatus of *best polynomial approximation*. We will discuss the design of estimators in this section and the minimax lower bound in the next section. Our inspiration comes from previous work on functional estimation in Gaussian mean models [17, 39]. Nemirovski (credited in [65]) pioneered the use of polynomial approximation in functional estimation and showed that unbiased estimators for the truncated Taylor series of the smooth functionals is asymptotically efficient. This strategy is generalized to non-smooth functionals in [17] using best polynomial approximation and in [39] for estimating the ℓ_1-norm in Gaussian mean model.

On the constructive side, the main idea is to trade bias with variance. Under the i.i.d. sampling model, it is easy to show (see, e.g., [66, Proposition 8]) that to estimate a functional $T(P)$ using n samples, an unbiased estimator exists if and only if $T(P)$ is a polynomial in P of degree at most n. Similarly, under Poisson sample model, $T(P)$ admits an unbiased estimator if and only if T is real analytic. Consequently, there exists no unbiased entropy estimator or the support size with or without Poissonized sampling. Therefore, a natural idea is to approximate the functional by polynomials which enjoy unbiased estimation, and reduce the bias to at most the uniform approximation error. The choice of the degree aims to strike a good bias-variance balance. In fact, the use of polynomial approximation in functional estimation is not new. In [67], the authors considered a truncated Taylor expansion of $\log x$ at $x = 1$ which admits an unbiased estimator, and proposed to estimate the remainder term using Bayesian techniques; however, no risk bound is given for this scheme. Paninski also studied how to use approximation by Bernstein polynomials to reduce the bias of the plug-in estimators [66], which forms the basis for proving the existence of consistent estimators with sublinear sample complexity in [68].

This idea is also used by [69] in the upper bound of estimating Shannon entropy and power sums with a slightly different estimator which also achieves the minimax rate. For more recent results on estimating Shannon entropy, support size, Rényi entropy and other distributional functionals on large alphabets, see [70, 71, 72, 73, 74].

Next we present more details of the above recipe. Let the set of functions that can be estimated with zero bias using n independent samples be $\mathcal{F}_n = \{f_i : i \in \mathcal{I}_n\}$, and the estimator for f_i be $\hat{f}_i$ with variance at most σ_i^2 for each i. We need to devise a good approximation of T by $\sum_i a_i f_i$ that is estimated by $\hat{T} = \sum_i a_i \hat{f}_i$ with small $|a_i|$:

- the bias of $\hat{T}$ is the approximation error $\sum_i a_i \hat{f}_i - T$;
- the standard deviation of $\hat{T}$ is at most $\sum_i |a_i| \sigma_i$.

The choice of coefficient magnitudes aims to strike a good balance of bias and variance.

The same approximation idea can be applied on a smaller family of functions as a subset of $\mathcal{F}_n$. One special case is when each f_i can be estimated

by an additive function $\hat{f}_i(X_1, \ldots, X_n) = \sum_j \hat{f}_{ij}(X_j)$. The variance of each $\hat{f}_{ij}$ is at most σ_{ij}^2. Then the variance of $\hat{T}$ is

$$\mathsf{var}[\hat{T}] = \sum_{j=1}^{n} \mathsf{var}\left[\sum_i a_i g_{ij}(X_j)\right] \le \sum_{j=1}^{n} \left(\sum_i |a_i|\sigma_{ij}\right)^2. \tag{3.6}$$

Under the multinomial sampling model, to estimate any monomial p_i^m using $N_i \sim \text{binomial}(n, p_i)$, there exists an unbiased estimator given by

$$\frac{N_i(N_i-1)\ldots(N_i-m+1)}{n(n-1)\ldots(n-m+1)},$$

where N_i counts the occurrences of symbol i. Under the Poisson sampling model, the monomial p_i^n is estimated using $N_i \sim \text{Poi}(np_i)$ by

$$\frac{N_i(N_i-1)\ldots(N_i-m+1)}{n^m}.$$

3.3 Lower bounds from moment matching

While the use of best polynomial approximation on the constructive side is admittedly natural, the fact that it also arises in the optimal lower bound is perhaps surprising. As carried out in [17, 39], the strategy is to choose two priors with matching moments up to a certain degree, which ensures the impossibility to test. The minimax lower bound is then given by the maximal separation in the expected functional values subject to the moment matching condition. This problem is the *dual* of best polynomial approximation in the optimization sense. Using moment matching techniques, we obtain the optimal minimax lower bounds for the estimation problems investigated in Part I.

A general idea for obtaining lower bounds is based on a reduction of estimation to testing. Consider the estimation of some functional $T_\mu = T(\mu)$ with a distance metric[1] $\rho(\hat{T}, T_\mu)$ as the loss function, where μ belongs to a family of distributions $\mathcal{M}$. If two hypotheses

$$H_0 : X \sim \mu, \quad H_1 : X \sim \mu',$$

[1]The reduction is similar if ρ is not a distance but satisfies triangle inequality within a constant factor. See [32, Chapter 2].

cannot be reliably distinguished from the samples, while the functional values T_μ and $T_{\mu'}$ are different, then any estimate suffers a maximum risk at least proportional to $\rho(T_\mu, T_{\mu'})$.

Theorem 3.2. *For any estimate $\hat{T}$, and any two distributions $\mu, \mu' \in \mathcal{M}$, we have*

$$\sup_{\mu} \mathbb{E}\rho(T_\mu, \hat{T}) \geq \frac{1}{2}\rho(T_\mu, T_{\mu'})(1 - \mathsf{TV}(\mu, \mu')).$$

Proof. We will use the average risk as a lower bound of the maximum risk. Consider an uniform prior π on $\{\mu, \mu'\}$. Then

$$r_\pi(\hat{T}) = \frac{1}{2}\int \rho(T_\mu, \hat{T})\mathrm{d}\mu + \frac{1}{2}\int \rho(T_{\mu'}, \hat{T})\mathrm{d}\mu'.$$

Since ρ is non-negative, the right-hand side is at least

$$\frac{1}{2}\int (\rho(T_\mu, \hat{T}) + \rho(T_{\mu'}, \hat{T}))\min\{\mathrm{d}\mu, \mathrm{d}\mu'\}.$$

Applying the triangle inequality yields that

$$r_\pi(\hat{T}) \geq \frac{1}{2}\rho(T_\mu, T_{\mu'})\int \min\{\mathrm{d}\mu, \mathrm{d}\mu'\}.$$

The integral in the last inequality is precisely $1 - \mathsf{TV}(\mu, \mu')$ [75]. □

This is also known as Le Cam's *two-point method*. It can be generalized by introducing two composite hypotheses (also known as fuzzy hypotheses in [32]):

$$H_0 : \mu \in \mathcal{M}_0, \quad H_1 : \mu \in \mathcal{M}_1,$$

where $\mathcal{M}_0, \mathcal{M}_1 \subseteq \mathcal{M}$, such that $\rho(T_\mu, T_{\mu'}) \geq d$ for any $\mu \in \mathcal{M}_0$ and $\mu' \in \mathcal{M}_1$. Similarly, if no test can distinguish the above two hypotheses reliably, then any estimate suffers a maximum risk at least proportional to d. Denote the mixture distribution by

$$\pi_\nu = \int P\mathrm{d}\nu(P), \tag{3.7}$$

where ν is the mixing (prior) distribution on $\mathcal{M}$. We obtain Theorem 3.3 by a proof similar to that of Theorem 3.2.

Theorem 3.3. *Let ν and ν' be distributions on $\mathcal{M}_0$ and $\mathcal{M}_1$, respectively.*

For any estimate $\hat{T}$, we have

$$\sup_{\mu} \mathbb{E}\rho(T_\mu, \hat{T}) \geq \frac{1}{2} \inf_{\substack{\mu \in \mathcal{M}_0 \\ \mu' \in \mathcal{M}_1}} \rho(T_\mu, T_{\mu'})(1 - \mathsf{TV}(\pi_\nu, \pi_{\nu'})).$$

In order to apply the above result to obtain a minimax lower bound, we must find two appropriate priors on $\mathcal{M}$. In parametric models, $\mathcal{M} = \{p_\theta : \theta \in \Theta\}$ and we need to find priors on Θ. There are two main ingredients in Le Cam's method: (1) functional values separation; (2) indistinguishabiblity, i.e., statistical closeness between distributions.

It turns out these two goals can be simultaneously accomplished by the dual of uniform approximation (2.5), which enables us to construct two (discrete) distributions μ and μ' supported on a closed interval $[a, b]$ such that

$$\mathbb{E}_\mu[f] - \mathbb{E}_{\mu'}[f] = 2 \inf_{P \in \mathcal{P}_n} \max_{x \in [a,b]} |P(x) - f(x)|, \tag{3.8}$$

and that μ and μ' match their first n moments:

$$\mathbb{E}_\mu[X^j] = \mathbb{E}_{\mu'}[X^j], \quad j = 0, \ldots, n. \tag{3.9}$$

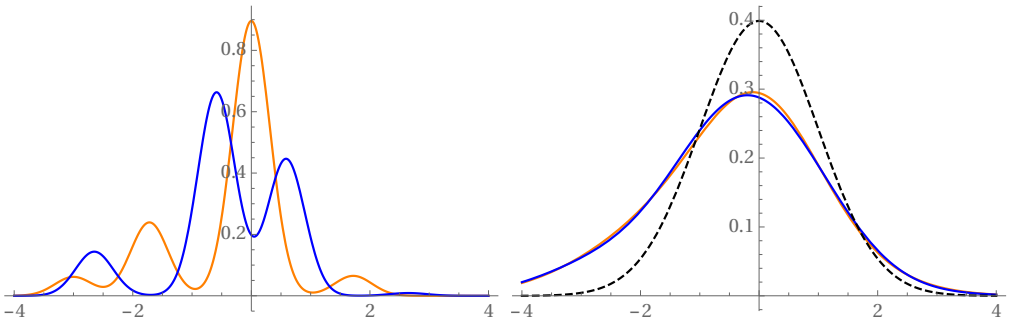
Statistical closeness between two mixture distributions of the form (3.7) can be established through moment matching (3.9). The results are developed for Gaussian mixtures and Poisson mixtures in this subsection. The lower bounds using (3.8) and (3.9) in specific problems will be established in Chapters 4, 5, and 8.

Gaussian mixtures. In Gaussian mixtures, the distribution is of the form

$$\pi_\nu = \int N(\theta, 1) \mathrm{d}\nu(\theta) = \nu * N(0, 1).$$

The statistical closeness is demonstrated in Figure 3.1, and is made precise in Theorem 3.4. Statistical closeness via moment matching has been established, for instance, by orthogonal expansion [76, 39], by Taylor expansion [31, 55], and by the best polynomial approximation [72]. Similar results to this lemma were previously obtained in [76, 39, 31].

Theorem 3.4. *Suppose ν and ν' are centered distributions such that $\mathbf{m}_\ell(\nu) = \mathbf{m}_\ell(\nu')$.*



(a) Mixing distributions (b) Mixture distributions

Figure 3.1: Statistical closeness via moment matching. In (a), two different mixing distributions coincide on their first six moments; in (b), the mixing distributions are convolved with the standard normal distribution (the black dashed line), and the Gaussian mixtures are statistically close.

- *If ν and ν' are ϵ-subgaussian for $\epsilon < 1$, then*

$$\chi^2(\nu * N(0,1) \| \nu' * N(0,1)) \leq O\left(\frac{1}{\sqrt{\ell}} \frac{\epsilon^{2\ell+2}}{1-\epsilon^2}\right). \tag{3.10}$$

- *If ν and ν' are supported on $[-\epsilon, \epsilon]$ for $\epsilon < 1$, then*

$$\chi^2(\nu * N(0,1) \| \nu' * N(0,1)) \leq O\left(\left(\frac{e\epsilon^2}{\ell+1}\right)^{\ell+1}\right). \tag{3.11}$$

Proof. This is a special case of the moment comparison result Lemma 7.5 in Chapter 7. Let $U \sim \nu$ and $U' \sim \nu'$. If ν and ν' are ϵ-subgaussian, then $\mathsf{var}[U'] \leq \epsilon^2$, and $\mathbb{E}|U|^p, \mathbb{E}|U'|^p \leq 2(\epsilon\sqrt{p/e})^p$ [54]. Applying the χ^2 upper bound from moment difference in Lemma 7.5 yields that

$$\chi^2(\nu * N(0,1) \| \nu' * N(0,1)) \leq e^{\epsilon^2/2} \sum_{j \geq \ell+1} \frac{16\epsilon^{2j}}{\sqrt{2\pi j}},$$

where we used Stirling's approximation $n! > \sqrt{2\pi n}(n/e)^n$. If ν and ν' are supported on $[-\epsilon, \epsilon]$, the conclusion is obtained similarly by using $\mathbb{E}|U|^p, \mathbb{E}|U'|^p \leq \epsilon^p$. □

Remark 3.1 (Tightness of Theorem 3.4)**.** When ℓ is odd, there exists a pair of ϵ-subgaussian distributions ν and ν' such that $\mathbf{m}_\ell(\nu) = \mathbf{m}_\ell(\nu')$, while $\chi^2(\nu * N(0,1) \| \nu' * N(0,1)) \geq \Omega_\ell(\epsilon^{2\ell+2})$. They can be constructed using

Gauss quadrature introduced in Section 2.3. To this end, let $\ell = 2k-1$ and we set $\nu = N(0,\epsilon^2)$ and $\tilde{g}_k$ to be its k-point Gauss quadrature. Then $\mathbf{m}_{2k-1}(\nu) = \mathbf{m}_{2k-1}(\tilde{g}_k)$, and $\tilde{g}_k$ is also ϵ-subgaussian (see Lemma 2.1). It is shown in [76, (54)] that

$$\chi^2(\tilde{g}_k * N(0,1) \| \nu * N(0,1)) = \sum_{j \geq 2k} \frac{1}{j!} \left(\frac{\epsilon^2}{1+\epsilon^2}\right)^j |\mathbb{E}_{g_k}[H_j]|^2,$$

where g_k is the k-point Gauss quadrature of the standard normal distribution, and H_k is the degree-k Hermite polynomial defined in (2.21). Since $\mathbb{E}_{g_k}[H_{2k}] = -k!$ (see Lemma 2.3), for any $\epsilon < 1$, we have

$$\chi^2(\tilde{g}_k * N(0,1) \| \nu * N(0,1)) \geq \frac{(k!)^2}{(2k)!} \left(\frac{\epsilon^2}{1+\epsilon^2}\right)^{2k} \geq (\Omega(\epsilon))^{4k}.$$

Poisson mixtures. Now we show the result for Poisson mixtures

$$\pi_\nu = \int \mathrm{Poi}(\lambda) \mathrm{d}\nu(\lambda).$$

Poisson mixtures are discrete distributions supported on $\mathbb{N}$. The following result gives a sufficient condition for Poisson mixtures to be indistinguishable in terms of moment matching. Analogous results for Gaussian mixtures have been obtained in [17, Section 4.3] using Taylor expansion of the KL divergence and orthogonal basis expansion of χ^2-divergence in [39, Proof of Theorem 3]. For Poisson mixtures we directly deal with the total variation as the ℓ_1-distance between the mixture probability mass functions.

Theorem 3.5 (Poisson mixtures)**.** *Suppose ν and ν' are supported on $[0,\Lambda]$ and match the first ℓ moments such that $2\Lambda \leq \ell + 1$. Denote the mixture distributions by μ and μ' with mixing distributions ν and ν', respectively. Then*

$$\mathsf{TV}(\mu,\mu') \leq \frac{1}{2}\left(\frac{2e\Lambda}{\ell+1}\right)^{\ell+1}.$$

Proof. Denote the probability mass functions of μ and μ' by p and p', respectively. Then

$$p(i) = \mathbb{E}\left[e^{-U}\frac{U^i}{i!}\right], \quad p'(i) = \mathbb{E}\left[e^{-U'}\frac{U'^i}{i!}\right],$$

where $U \sim \nu$ and $U' \sim \nu'$. Applying Taylor's expansion to $x \mapsto e^{-x}$ yields

that

$$|p(i) - p'(i)| = \frac{1}{i!}\left|\sum_{j\geq 0}\frac{(-1)^j}{j!}\Delta m_{i+j}\right|,$$

where $\Delta m_{i+j} = \mathbb{E}[U^{i+j}] - \mathbb{E}[U'^{i+j}]$. When ν and ν' match the first ℓ moments and are supported on $[0,\Lambda]$, we have $\Delta m_{i+j} = 0$ when $j \leq \ell - i$, and $|\Delta m_{i+j}| \leq \Lambda^{i+j}$ when $j \geq \ell - i + 1$. Then,

$$|p(i) - p'(i)| \leq \sum_{j\geq \ell - i + 1}\frac{\Lambda^{i+j}}{i!j!}.$$

The total variation distance can be expressed as (see [32, Lemma 2.1])

$$\mathsf{TV}(\mu,\mu') = \frac{1}{2}\sum_{i\geq 0}|p(i) - p'(i)|.$$

Then we obtain that

$$\mathsf{TV}(\mu,\mu') \leq \frac{1}{2}\sum_{i+j\geq \ell+1}\frac{\Lambda^{i+j}}{i!j!} = \frac{1}{2}\sum_{j\geq \ell+1}\frac{(2\Lambda)^j}{j!} \leq \frac{1}{2}\left(\frac{2e\Lambda}{\ell+1}\right)^{\ell+1},$$

where Chernoff bound is used in the last inequality. □

Remark 3.2. In an earlier version of [55],[2] the following weaker total variation bound

$$\mathsf{TV}(\mathbb{E}\left[\mathrm{Poi}(V)\right], \mathbb{E}\left[\mathrm{Poi}(V')\right]) \leq 2\exp\left(-\left(\frac{L}{2}\log\frac{L}{2eM} - M\right)\right) \wedge 1, \quad (3.12)$$

was proved by truncating the summation in the total variation. This bound suffices for our purpose; in fact, the same proof techniques have been subsequently used in [69, Lemma 11] for minimax lower bound of estimating other functionals. Nevertheless, (3.13) provides a strict improvement over (3.12), whose proof is even simpler and involves no truncation argument. What remains open is the optimal number of matching moments to ensure indistinguishability of the Poisson mixtures. The above result implies that as soon as L/M exceeds $2e$ the total variation decays exponentially; it is unclear whether L needs to grow linearly with M in order to drive the total variation to zero.

[2] See Lemma 3 in `http://arxiv.org/pdf/1407.0381v2.pdf`.

The above result is simple to prove. The next result is an improvement in terms of constants. This is crucial for the purpose of obtaining good constants for the sample complexity bounds in Chapter 5.

Theorem 3.6. *Let V and V' be random variables taking values on $[0, \Lambda]$. If $\mathbb{E}[V^j] = \mathbb{E}[V'^j],\ j = 1, \ldots, L$, then*

$$\mathsf{TV}(\mathbb{E}[\mathrm{Poi}(V)], \mathbb{E}[\mathrm{Poi}(V')]) \leq \frac{(\Lambda/2)^{L+1}}{(L+1)!} \left(2 + 2^{\Lambda/2-L} + 2^{\Lambda/(2\log 2)-L}\right). \quad (3.13)$$

In particular, $\mathsf{TV}(\mathbb{E}[\mathrm{Poi}(V)], \mathbb{E}[\mathrm{Poi}(V')]) \leq (\frac{e\Lambda}{2L})^L$. *Moreover, if* $L > \frac{e}{2}\Lambda$, *then*

$$\mathsf{TV}(\mathbb{E}[\mathrm{Poi}(V)], \mathbb{E}[\mathrm{Poi}(V')]) \leq \frac{2(\Lambda/2)^{L+1}}{(L+1)!}(1 + o(1)), \quad \Lambda \to \infty.$$

Proof. Denote the best degree-L polynomial approximation error of a function f on an interval I by

$$E_L(f, I) = \inf_{p \in \mathcal{P}_L} \sup_{x \in I} |f(x) - p(x)|.$$

Let

$$f_j(x) \triangleq \frac{e^{-x}x^j}{j!}. \quad (3.14)$$

Let $P^*_{L,j}$ be the best polynomial of degree L that uniformly approximates f_j over the interval $[0, \Lambda]$ and the corresponding approximation error by $E_L(f_j, [0, \Lambda]) = \max_{x \in [0,\Lambda]} |f_j(x) - P^*_{L,j}(x)|$. Then $\mathbb{E}P^*_{L,j}(V) = \mathbb{E}P^*_{L,j}(V')$ and hence

$$\begin{aligned}
&\mathsf{TV}(\mathbb{E}[\mathrm{Poi}(V)], \mathbb{E}[\mathrm{Poi}(V')]) = \frac{1}{2}\sum_{j=0}^{\infty} |\mathbb{E}f_j(V) - \mathbb{E}f_j(V')| \\
&\leq \frac{1}{2}\sum_{j=0}^{\infty} |\mathbb{E}(f_j(V) - P^*_{L,j}(V))| + |\mathbb{E}(f_j(V') - P^*_{L,j}(V'))| \\
&\leq \sum_{j=0}^{\infty} E_L(f_j, [0, \Lambda]). \quad (3.15)
\end{aligned}$$

A useful upper bound on the degree-L best polynomial approximation error of a function f is via the Chebyshev interpolation polynomial, whose uniform approximation error can be bounded using the L^{th} derivative of f.

Specifically, we have (see (2.13))

$$E_L(f,[0,\Lambda]) \le \frac{1}{2^L(L+1)!}\left(\frac{\Lambda}{2}\right)^{L+1} \max_{x\in[0,\Lambda]} \left|f^{(L+1)}(x)\right|. \tag{3.16}$$

To apply (3.16) to $f = f_j$ defined in (3.14), note that $f_j^{(L+1)}$ can be conveniently expressed in terms of Laguerre polynomials $\mathcal{L}_n^{(k)}$ in (2.23).

If $j \le L+1$,

$$f_j^{(L+1)}(x) = \frac{\mathrm{d}^{L+1-j}}{\mathrm{d}x^{L+1-j}}\left(\frac{\mathrm{d}^j}{\mathrm{d}x^j}\frac{e^{-x}x^j}{j!}\right) = \frac{\mathrm{d}^{L+1-j}}{\mathrm{d}x^{L+1-j}}(\mathcal{L}_j(x)e^{-x}).$$

Note that $\mathcal{L}_j$ is a degree-j polynomial, whose derivative of order higher than j is zero. Applying general Leibniz rule for derivatives yields that

$$\begin{aligned} f_j^{(L+1)}(x) &= \sum_{m=0}^{(L+1-j)\wedge j} \binom{L+1-j}{m} \frac{\mathrm{d}^m \mathcal{L}_j(x)}{\mathrm{d}x^m} e^{-x}(-1)^{L+1-j-m} \\ &= (-1)^{L+1-j}e^{-x} \sum_{m=0}^{(L+1-j)\wedge j} \binom{L+1-j}{m} \mathcal{L}_{j-m}^{(m)}(x). \end{aligned} \tag{3.17}$$

Applying (2.24) yields that

$$\left|f_j^{(L+1)}(x)\right| \le e^{-x} \sum_{m=0}^{(L+1-j)\wedge j} \binom{L+1-j}{m}\binom{j}{j-m} e^{x/2} = e^{-x/2}\binom{L+1}{j}.$$

Therefore $\max_{x\in[0,\Lambda]} |f_j^{(L+1)}(x)| \le \binom{L+1}{j}$ when $j \le L+1$.[3] Then, applying (3.16), we have

$$\sum_{j=0}^{L+1} E_L(f_j,[0,\Lambda]) \le \sum_{j=0}^{L+1} \frac{\binom{L+1}{j}(\Lambda/2)^{L+1}}{2^L(L+1)!} = \frac{2(\Lambda/2)^{L+1}}{(L+1)!}. \tag{3.18}$$

If $j \ge L+2$, the derivatives of f_j are related to the Laguerre polynomial by

$$f_j^{(L+1)}(x) = \frac{(L+1)!}{j!}x^{j-L-1}e^{-x}\mathcal{L}_{L+1}^{(j-L-1)}(x).$$

[3]This is in fact an equality. In view of (3.17) and the fact that $L_{j-m}^{(m)}(0) = \binom{j}{j-m}$ [52, 22.3], we have $|f_j^{(L+1)}(0)| = \sum_m \binom{L+1-j}{m}\binom{j}{j-m} = \binom{L+1}{j}$.

Again applying (2.24) when $x \geq 0$ and $k \in \mathbb{N}$, we obtain

$$\left| f_j^{(L+1)}(x) \right| \leq \frac{(L+1)!}{j!} x^{j-L-1} e^{-x} \binom{j}{L+1} e^{x/2} = \frac{1}{(j-L-1)!} e^{-x/2} x^{j-L-1},$$

where the maximum of right-hand side on $[0, \Lambda]$ occurs at $x = (2(j-L-1)) \wedge \Lambda$. Therefore

$$\max_{x \in [0,\Lambda]} |f_j^{(L+1)}(x)| \leq \begin{cases} \frac{1}{(j-L-1)!} \left(\frac{2(j-L-1)}{e} \right)^{j-L-1}, & L+1 \leq j \leq L+1+\Lambda/2, \\ \frac{1}{(j-L-1)!} e^{-\Lambda/2} \Lambda^{j-L-1}, & j \geq L+1+\Lambda/2. \end{cases}$$

Then, applying (3.16) and Stirling's approximation that $(\frac{j-L-1}{e})^{j-L-1} < \frac{(j-L-1)!}{\sqrt{2\pi(j-L-1)}}$, we have

$$\begin{aligned} \sum_{\substack{j \geq L+2 \\ j < L+1+\Lambda/2}} E_L(f_j, [0,\Lambda]) &\leq \frac{(\Lambda/2)^{L+1}}{2^L (L+1)!} \sum_{\substack{j \geq L+2 \\ j < L+1+\Lambda/2}} \frac{2^{j-L-1}}{\sqrt{2\pi(j-L-1)}} \\ &\leq \frac{(\Lambda/2)^{L+1} 2^{\Lambda/2}}{2^L (L+1)!}, \end{aligned} \tag{3.19}$$

$$\begin{aligned} \sum_{j \geq L+1+\Lambda/2} E_L(f_j, [0,\Lambda]) &\leq \frac{(\Lambda/2)^{L+1} e^{-\Lambda/2}}{2^L (L+1)!} \sum_{j \geq L+1+\Lambda/2} \frac{\Lambda^{j-L-1}}{(j-L-1)!} \\ &\leq \frac{(\Lambda/2)^{L+1} e^{\Lambda/2}}{2^L (L+1)!}. \end{aligned} \tag{3.20}$$

Assembling the three ranges of summations in (3.18)-(3.20) in the total variation bound (3.15), we obtain

$$\mathsf{TV}(\mathbb{E}[\mathrm{Poi}(V)], \mathbb{E}[\mathrm{Poi}(V')]) \leq \frac{(\Lambda/2)^{L+1}}{(L+1)!} \left(2 + 2^{\Lambda/2-L} + 2^{\Lambda/(2\log 2)-L} \right).$$

Finally, applying Stirling's approximation $(L+1)! > \sqrt{2\pi(L+1)} (\frac{L+1}{e})^{L+1}$, we conclude that $\mathsf{TV}(\mathbb{E}[\mathrm{Poi}(V)], \mathbb{E}[\mathrm{Poi}(V')]) \leq (\frac{e\Lambda}{2L})^L$. If $L > \frac{e}{2}\Lambda > \frac{\Lambda}{2\log 2} > \frac{\Lambda}{2}$, then $2^{\Lambda/2-L} + 2^{\Lambda/(2\log 2)-L} = o(1)$. □

CHAPTER 4

ENTROPY ESTIMATION

In this chapter, we begin the application of polynomial approximation method in entropy estimation. The Shannon entropy [77] of a discrete distribution P is defined as

$$H(P) = \sum_i p_i \log \frac{1}{p_i}.$$

Entropy estimation has found numerous applications across various fields, such as psychology [78], neuroscience [79], physics [67], telecommunication [80], biomedical research [81], etc. Furthermore, it serves as the building block for estimating other information measures expressible in terms of entropy, such as mutual information and directed information, which are instrumental in machine learning applications such as learning graphical models [82, 83, 84, 85]. However, the definition of Shannon entropy uses the complete distribution of the data source, and the domain size can be quite large, which makes the estimation task difficult, especially when a limited amount of samples are obtainable due to resource constraints.

We first discuss the maximum likelihood estimate, which is also known as the *empirical entropy*. As introduced in Section 3.2, this is the plug-in approach in functional estimation, for which we substitute the estimated distribution into the function. This approach suffers from large bias with insufficient samples, and can be highly suboptimal when we are dealing with high-dimensional data.

We then describe the polynomial approximation method to reduce the bias applying the polynomial approximation method in Chapter 3. To investigate the decision-theoretic fundamental limit, we consider the minimax quadratic risk of entropy estimation:

$$R_H^*(k,n) \triangleq \inf_{\hat{H}} \sup_{P \in \mathcal{M}_k} \mathbb{E}_P[(\hat{H} - H(P))^2], \tag{4.1}$$

where $\mathcal{M}_k$ denotes the set of probability distributions on $[k] \triangleq \{1, \ldots, k\}$, and $\hat{H}$ is an estimator measurable with respect to n independent samples from P. In this chapter, we will discuss

- a constant-factor approximation of the minimax risk $R_H^*(k, n)$;
- a linear-time estimator that provably attains $R_H^*(k, n)$ within universal constant factors.

We present a preview of the fundamental limits in this chapter and briefly discuss the impact of large domains. A constant-factor approximation of the minimax risk $R_H^*(k, n)$ is given by Theorem 4.1.

Theorem 4.1. *If $n \gtrsim \frac{k}{\log k}$, then*

$$R_H^*(k, n) \asymp \left(\frac{k}{n \log k}\right)^2 + \frac{\log^2 k}{n}. \tag{4.2}$$

If $n \lesssim \frac{k}{\log k}$, there exists no consistent estimators, i.e., $R_H^*(k, n) \gtrsim 1$.

To interpret the minimax rate (4.2), we note that the second term corresponds to the classical "parametric" term inversely proportional to $\frac{1}{n}$, which is governed by the variance and the central limit theorem (CLT). The first term corresponds to the squared bias, which is the main culprit in the regime of insufficient samples. Note that $R_H^*(k, n) \asymp (\frac{k}{n \log k})^2$ if and only if $n \lesssim \frac{k^2}{\log^4 k}$, where the bias dominates. As a consequence, the minimax rate implies that to estimate the entropy within ϵ bits with probability, say 0.9, the minimal sample size is given by

$$n \asymp \frac{\log^2 k}{\epsilon^2} \vee \frac{k}{\epsilon \log k}. \tag{4.3}$$

The worst-case mean-square error of the empirical entropy, denoted by $H(\hat{P}_n)$, is given by Theorem 4.2.

Theorem 4.2. *If $n \gtrsim k$, then*

$$\sup_{P: S(P) \le k} \mathbb{E}(H - H(\hat{P}_n))^2 \asymp \left(\frac{k}{n}\right)^2 + \frac{\log^2 k}{n}. \tag{4.4}$$

If $n \lesssim k$, there exists no consistent estimators, i.e., the left-hand side of (4.4) is $\Omega(1)$.

Note that the first and second terms in the risk again correspond to the squared bias and variance respectively. Comparing (4.2) and (4.4), we reach the following verdict on the plug-in estimator: Empirical entropy is rate-optimal, i.e., achieving a constant factor of the minimax risk, if and only if we are in the "data-rich" regime $n = \Omega(\frac{k^2}{\log^2 k})$. In the "data-starved" regime of $n = o\big(\frac{k^2}{\log^2 k}\big)$, empirical entropy is strictly rate-suboptimal. The comparison between the optimal estimator and the empirical entropy is demonstrated in Figure 4.1.

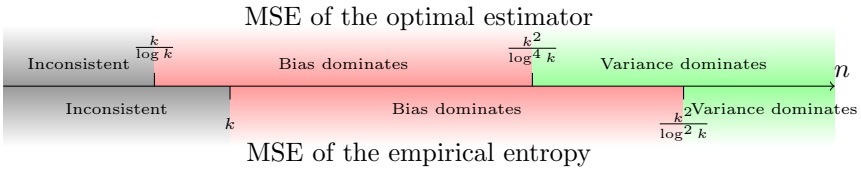


Figure 4.1: Classification and comparison of regimes between optimal entropy estimator and the empirical entropy.

4.1 Empirical entropy and Bernstein polynomials

Given n independent samples $X_1, \ldots, X_n$ from a discrete distribution $P = (p_1, \ldots, p_k)$, the maximum likelihood estimate of the distribution is the empirical distribution

$$\hat{P}_n = (\hat{p}_1, \ldots, \hat{p}_k),$$

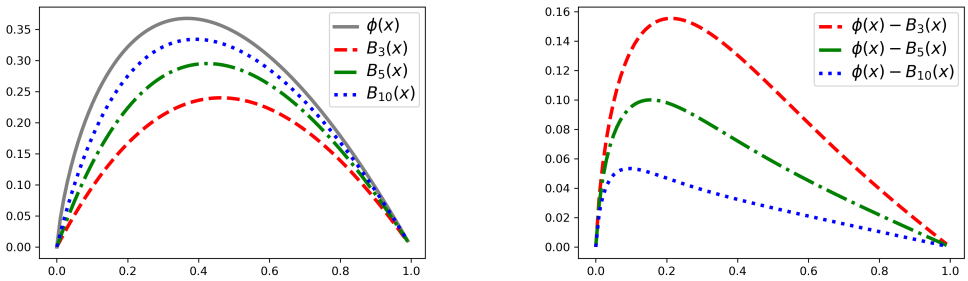
with $\hat{p}_i = N_i/n$, where N_i records the number of occurrences of samples with label i and is the sufficient statistics referred to as the histogram. Then the empirical entropy is

$$H(\hat{P}_n) = \sum_i \hat{p}_i \log \frac{1}{\hat{p}_i}. \tag{4.5}$$

Let $\phi(x) = x \log \frac{1}{x}$. Then the bias of empirical entropy is

$$\begin{aligned} \mathbb{E}[H(\hat{P}_n)] - H(P) &= \sum_i \left(\sum_{j=0}^{n} \phi(j/n) \binom{n}{j} p_i^j (1-p_i)^{n-j} - \phi(p_i) \right) \\ &= \sum_i (B_n(p_i) - \phi(p_i)), \end{aligned} \tag{4.6}$$

where B_n is the Bernstein polynomial of degree n to approximate ϕ using the equation (2.2). See an illustration of Bernstein approximation in Figure 4.2. We shall next derive several results on the bias of the empirical entropy using the Bernstein approximation.



(a) Bernstein polynomials (b) Approximation errors

Figure 4.2: Illustration of Bernstein polynomial approximation of ϕ of degree $3, 5$ and 10. (a) shows the actual Bernstein polynomials, and (b) shows the errors of approximation.

Lemma 4.1. *If f is convex on $[0,1]$, then the Bernstein polynomials approximation (2.2) satisfies the following inequalities:*

$$B_n(x) \geq B_{n+1}(x) \geq f(x).$$

The inequalities are strict if f is strictly convex.

Proof. Applying the formula of Bernstein polynomials (2.2), we can calculate that (see [36, pp. 309–310])

$$B_n(x) - B_{n+1}(x) = \frac{x(1-x)}{n(n+1)} \sum_{k=0}^{n-1} f\left[\frac{k}{n}, \frac{k+1}{n+1}, \frac{k+1}{n}\right] \binom{n-1}{k} x^k (1-x)^{n-k},$$

where $f[\frac{k}{n}, \frac{k+1}{n+1}, \frac{k+1}{n}]$ is the divided difference that can be evaluated using (7.22). This divided difference is non-negative when f is convex (see (2.9)). □

Note that ϕ is strictly concave on $[0,1]$. In this case we have

$$B_n(x) < B_{n+1}(x) < \phi(x), \quad 0 < x < 1. \tag{4.7}$$

See Figure 4.2 for an illustration. We conclude from (4.6) that the empirical entropy is always underbiased, and the bias is strictly detcreasing in magnitude as the number of samples increases [66].

Bernstein approximation has the following asymptotic formula.

Lemma 4.2. *Fix $x \in [0,1]$. If f is bounded, differentiable in a neighborhood of x, and $f''(x)$ exists, then*

$$\lim_{n\to\infty} n(B_n(x) - f(x)) = \frac{x(1-x)}{2} f''(x). \tag{4.8}$$

Proof. By Taylor's expansion,

$$f(t) = f(x) + f'(x)(t-x) + \frac{f''(x)}{2}(t-x)^2 + h(t-x)(t-x)^2,$$

where $h(y)$ is bounded and vanishes with y. Note that $B_n(x) = \mathbb{E}f(\hat{p})$ where $\hat{p} = N/n$ and $N \sim \text{binomial}(n, x)$. Then

$$B_n(x) - f(x) = \frac{x(1-x)}{2n} f''(x) + \mathbb{E}[h(\hat{p}-x)(\hat{p}-x)^2].$$

The last term is $o(1/n)$ by the continuity of h and the concentration of binomial distributions (see [36, pp. 304–308]). □

In entropy estimation, $\phi''(x) = -1/x$. By using (4.6), for a *fixed* distribution P, the asymptotic bias of the empirical entropy as n diverges is given by

$$\mathbb{E}[H(\hat{P}_n)] - H(P) = \sum_i \frac{p_i - 1}{2n}(1 + o(1)) = \frac{1 - S(P)}{2n}(1 + o(1)), \tag{4.9}$$

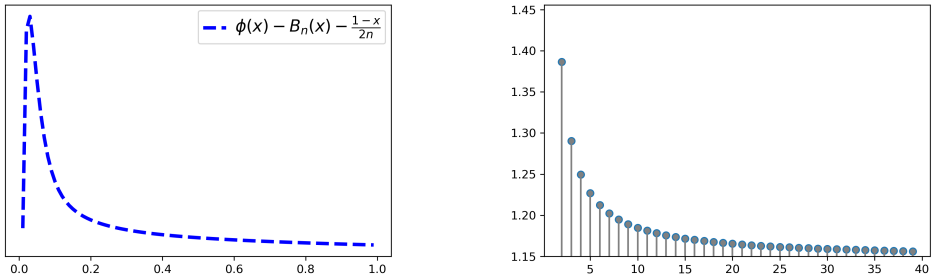
where $S(P)$ denotes the support size of P. This asymptotic formula inspires the well-known bias reduction to the empirical entropy, named the Miller-Madow estimator [86]:

$$\hat{H}_{\text{MM}} = \hat{H}_{\text{plug}} + \frac{\hat{S} - 1}{2n}, \tag{4.10}$$

where $\hat{S}$ is the number of observed distinct symbols. For higher-order asymptotic expansions of the bias, as well as various types of bias reduction, see [87]. This formula also holds when the fixed distribution assumption is relaxed to

$n \min_i p_i \to \infty$ [66, Theorem 5].

However, the asymptotic estimate (4.8) is not uniform over $[0, 1]$ (see (4.11), and also an illustration in Figure 4.3). When $S(P)$ is comparable



(a) Approximation error (b) Non-uniform convergence

Figure 4.3: Illustration of the non-uniform convergence of (4.8). (a) shows $\phi(x) - (B_n(x) + \frac{1-x}{2n})$ for $n = 50$. (b) further shows $\frac{2n}{1-x}(\phi(x) - B_n(x))$ at $x = 1/n$ for different n. The sequence of values is not converging to one.

or far exceeds the number of samples, this asymptotic estimate of the bias in (4.9) is no longer true. Applying (4.5) yields that

$$H(P) - \mathbb{E}[H(\hat{P}_n)] = \mathbb{E}[D(\hat{P}_n \| P)],$$

where $D(\cdot\|\cdot)$ denotes the Kullback-Leibler (KL) divergence. We obtain the following upper bound of the bias [66, Proposition 1].

Proposition 4.1.

$$0 \le H(P) - \mathbb{E}[H(\hat{P}_n)] \le \log\left(1 + \frac{S(P) - 1}{n}\right).$$

Proof. The KL divergence is related to the χ^2-divergence by [75]

$$D(\hat{P}_n \| P) \le \log(1 + \chi^2(\hat{P}_n \| P)).$$

Since log is a concave function, we obtain from the Jensen's inequality that

$$\mathbb{E}[D(\hat{P}_n \| P)] \le \log(1 + \mathbb{E}[\chi^2(\hat{P}_n \| P)]).$$

The expectation in the right-hand side of the above inequality is

$$\mathbb{E}[\chi^2(\hat{P}_n\|P)] = \sum_i \frac{\mathbb{E}(\hat{p}_i - p_i)^2}{p_i} = \sum_i \frac{1-p_i}{n} = \frac{S(P)-1}{n}. \qquad \square$$

We next discuss the tightness of the previous bias analysis of the empirical entropy using the Bernstein polynomial (4.6) again. We first state a lower bound on the Bernstein approximation obtained in [88, Theorem 5].

Lemma 4.3. *For $x \geq 15/n$,*

$$\begin{aligned} |B_n(x) - \phi(x)| &\geq \frac{1-x}{2n} + \frac{1}{12n^2x} - \frac{x}{12n^2} - \frac{1}{2n^3x^2} \\ &\geq \frac{1-x}{2n} + \frac{1}{20n^2x} - \frac{x}{12n^2}. \end{aligned} \tag{4.11}$$

Consequently, using (4.6) and (4.7), for a distribution P with $p_i \geq 15/n$, we have[1]

$$|H(\hat{P}_n) - H(P)| \geq \frac{S(P)-1}{2n} + \frac{1}{20n^2}\left(\sum_i \frac{1}{p_i}\right) - \frac{1}{12n^2}.$$

From the above lower bound and the monotonicity in (4.7), for a uniform distribution over k elements, the bias of the empirical entropy is at least $\Omega(\min\{\frac{k}{n}, 1\})$.

Now we evaluate the variance of the empirical entropy. Note that empirical entropy is a linear estimate

$$H(\hat{P}) = \sum_i g(N_i) = \sum_j \Phi_j g(j), \tag{4.12}$$

where $g(j) = \phi(j/n)$ and Φ_j denotes the number of elements that appeared exactly j times (also known as histogram order statistics [66], fingerprint [89], or profile [60]). A variance upper bound can be obtained by the Efron-

[1]For a fixed distribution, as n diverges, it is obtained in [87, (14)] that

$$H(P) - H(\hat{P}_n) = \frac{S(P)-1}{2n} + \frac{1}{12n^2}\left(\sum_i \frac{1}{p_i} - 1\right) + O(n^{-3}).$$

Stein-Steele inequality [90]:

$$\mathsf{var}H(\hat{P}_n) \le \frac{n}{2}\mathbb{E}(\Delta g(\tilde{N}_{X_1}) - \Delta g(\tilde{N}_{X_1'}))^2,$$

where X_1' is another independent sample from P, $\tilde{N}_i$ counts the occurrences of symbol i in $X_2, \ldots, X_n$, and $\Delta g(j)$ denotes the difference $g(j+1) - g(j)$. Applying the triangle inequality yields that

$$\mathsf{var}H(\hat{P}_n) \le n\mathbb{E}(\Delta g(\tilde{N}_{X_1}))^2. \tag{4.13}$$

Another way of writing the above upper bound is

$$\mathsf{var}H(\hat{P}_n) \le n\sum_i p_i\mathbb{E}(\Delta g(\tilde{N}_i))^2 = \sum_i \mathbb{E}N_i(\Delta g(N_i - 1))^2, \tag{4.14}$$

where $N_i \sim \text{binomial}(n, p_i)$ and $g(j) = 0$ for $j < 0$. We have the following result on the variance of empirical entropy.

Proposition 4.2.

$$\mathsf{var}H(\hat{P}_n) \le \frac{\log^2(\min\{n, eS(P)\})}{n}.$$

Proof. Let $g(j) = \phi(j/n)$. The difference $\Delta g(j)$ can be uniformly upper bounded by $\frac{\log n}{n}$ in magnitude for every $j = 0, \ldots, n-1$, and thus by (4.13) we obtain that

$$\mathsf{var}H(\hat{P}_n) \le \frac{\log^2 n}{n}.$$

The derivative of ϕ over $[\frac{j}{n}, \frac{j+1}{n}]$ is at most $\max\{|\log\frac{ej}{n}|, |\log\frac{e(j+1)}{n}|\}$ in magnitude. This yields a refined upper bound for $j = 1, \ldots, n-1$:

$$|\Delta g(j)| \le \frac{\max\{|\log(ej/n)|, 1\}}{n}.$$

Combining with the uniform upper bound $\frac{\log n}{n}$, we get

$$|\Delta g(j)| \le \frac{1}{n}\log\frac{en}{j+1}, \quad j = 0, \ldots n-1.$$

Applying (4.14) yields that

$$\mathsf{var} H(\hat{P}_n) \le \frac{1}{n^2} \sum_i \mathbb{E} N_i \log^2(en/N_i).$$

Note that $x \mapsto x \log^2(e/x)$ is concave on $[0,1]$, and $x \mapsto \log^2(ex)$ is concave on $[1,\infty)$. We obtain that

$$\sum_i \mathbb{E}[\log^2(en/N_i) N_i/n] \le \sum_i p_i \log^2(e/p_i) \le \log^2(eS(P)),$$

according to Jensen's inequality. □

We obtain in Section 4.3 that, when the distribution is supported on k elements, the MSE of any estimate using n independent samples is $\Omega(\frac{\log^2 k}{n})$ in the worst case (see Proposition 4.2). This lower bound also applies to the empirical entropy. The results of this section prove the worst-case MSE of the empirical entropy (4.4).

4.2 Optimal entropy estimation on large domains

From the analysis in Section 4.1, the empirical entropy is asymptotically optimal for distributions on a fixed alphabet as n diverges. Specifically, using (4.4), the mean squared error of the empirical entropy is $O(\frac{\log^2 k}{n})$ when $n \ge \frac{k^2}{\log^2 k}$, which is the optimal rate. However, the empirical entropy suffers from large bias using linear or sublinear number of samples, i.e., $n = O(k)$. In this section, we describe the design of the minimax rate-optimal estimator.

4.2.1 Previous results

We begin with a review of previous results on entropy estimation on large domain. It is well known that to estimate the distribution P itself, say, with total variation loss at most a small constant, we need at least $\Theta(k)$ samples (see, e.g., [91]). However, to estimate the entropy $H(P)$ which is a scalar function, it is unclear from first principles whether $n = \Theta(k)$ is necessary. This intuition and the inadequacy of plug-in estimator have already been noted by Dobrushin [92, p. 429], who wrote:

...This method (empirical entropy) is very laborious if m, the number of values of the random variable is large, since in this case most of the probabilities p_i are small and to determine each of them we need a large sample of length N, which leads to a lot of work. However, it is natural to expect that in principle the problem of calculating the single characteristic H of the distribution $(p_1, \ldots, p_m)$ is simpler than calculating the m-dimensional vector $(p_1, \ldots, p_m)$, and that therefore one ought to seek a solution of the problem by a method which does not require reducing the first and simpler problem to the second and more complicated problem.

Using non-constructive arguments, Paninski first proved that it is possible to consistently estimate the entropy using *sublinear* sample size, i.e., there exists $n_k = o(k)$, such that $R^*(k, n_k) \to 0$ as $k \to \infty$ [68]. Valiant proved that no consistent estimator exists, i.e., $R^*(k, n_k) \gtrsim 1$ if $n \lesssim \frac{k}{\exp(\sqrt{\log k})}$ [93]. The sharp scaling of the minimal sample size of consistent estimation is shown to be $\frac{k}{\log k}$ in the breakthrough results of Valiant and Valiant [89, 94]. However, the optimal sample size as a function of alphabet size k and estimation error ϵ has not been completely resolved. Indeed, an estimator based on linear programming is shown to achieve an additive error of ϵ using $\frac{k}{\epsilon^2 \log k}$ samples [64, Theorem 1], while $\frac{k}{\epsilon \log k}$ samples are shown to be necessary [89, Corollary 10]. This gap is partially amended in [95] by a different estimator, which requires $\frac{k}{\epsilon \log k}$ samples but only valid when $\epsilon > k^{-0.03}$. We obtain (4.2) that generalizes their result by characterizing the full minimax rate and the sharp sample complexity is given by (4.3).

We briefly discuss the difference between the lower bound strategy of [89] and ours. Since the entropy is a permutation-invariant functional of the distribution, a sufficient statistic for entropy estimation is the histogram of the histogram N:

$$\Phi_i = \sum_{j=1}^{k} \mathbf{1}_{\{N_j = i\}}, \quad i \in [n], \tag{4.15}$$

also known as *histogram order statistics* [66], *profile* [60], or *fingerprint* [89], which is the number of symbols that appear exactly i times in the sample. A canonical approach to obtain minimax lower bounds for functional estimation is Le Cam's two-point argument [96, Chapter 2], i.e., finding two distributions which have very different entropy but induce almost the same distri-

bution for the sufficient statistics, in this case, the histogram $N_1, \ldots, N_k$ or the fingerprints $\Phi_1, \ldots, \Phi_n$, both of which have non-product distributions. A frequently used technique to reduce dependence is *Poisson sampling* (see Section 3.1), where we relax the fixed sample size to a Poisson random variable with mean n. This does not change the statistical nature of the problem due to the exponential concentration of the Poisson distribution near its mean. Under the Poisson sampling model, the sufficient statistics $N_1, \ldots, N_k$ are independent Poissons with mean np_i; however, the entries of the fingerprint remain highly dependent. To contend with the difficulty of computing statistical distance between high-dimensional distributions with dependent entries, the major tool in [89] is a new CLT for approximating the fingerprint distribution by quantized Gaussian distribution, which is parameterized by the mean and covariance matrices and hence more tractable. This turns out to improve the lower bound in [93] obtained using Poisson approximation.

In contrast, we shall not deal with the fingerprint directly, but rather use the original sufficient statistics $N_1, \ldots, N_k$ due to their independence endowed by the Poissonized sampling. Our lower bound relies on choosing two random distributions (priors) with almost i.i.d. entries which effectively reduces the problem to one dimension, thus circumventing the hurdle of dealing with high-dimensional non-product distributions. The main intuition is that a random vector with i.i.d. entries drawn from a positive unit-mean distribution is not exactly but *sufficiently close* to a probability vector due to the law of large numbers, so that effectively it can be used as a prior in the minimax lower bound.

While we focus on estimating the entropy under the additive error criterion, approximating the entropy multiplicatively has been considered in [97]. It is clear that in general approximating the entropy within a constant factor is impossible with any finite sample size (consider Bernoulli distributions with parameter 1 and $1 - 2^{-n}$, which are not distinguishable with n samples); nevertheless, when the entropy is large enough, i.e., $H(P) \gtrsim \gamma/\eta$, it is possible to approximate the entropy within a multiplicative factor of γ using $n \lesssim k^{(1+\eta)/\gamma^2} \log k$ number of samples ([97, Theorem 2]).

4.2.2 Optimal estimator via best polynomial approximation

The major difficulty of entropy estimation lies in the bias due to insufficient samples. Recall that the entropy is given by $H(P) = \sum \phi(p_i)$, where $\phi(x) = x \log \frac{1}{x}$. It is easy to see that the expectation of any estimator $T : [k]^n \to \mathbb{R}_+$ is a polynomial of the underlying distribution P and, consequently, no unbiased estimator for the entropy exists (see [66, Proposition 8]). This observation inspired us to approximate ϕ by a polynomial of degree L, say g_L, for which we pay a price in bias as the approximation error but yield the benefit of zero bias. While the approximation error clearly decreases with the degree L, it is not unexpected that the variance of the unbiased estimator for $g_L(p_i)$ increases with L as well as the corresponding mass p_i. Therefore we only apply the polynomial approximation scheme to small p_i and directly use the plug-in estimator for large p_i, since the signal-to-noise ratio is sufficiently large.

Next we describe the estimator in detail. In view of the relationship between the risks with fixed and Poisson sample size in Section 3.1, we shall assume the Poisson sampling model to simplify the analysis, where we first draw $n' \sim \mathrm{Poi}(2n)$ and then draw n' i.i.d. samples $X = (X_1, \ldots, X_{n'})$ from P. We split the samples equally and use the first half for selecting to use either the polynomial estimator or the plug-in estimator and the second half for estimation. Specifically, for each sample X_i we draw an independent fair coin $B_i \overset{\text{i.i.d.}}{\sim} \mathrm{Bern}\left(\frac{1}{2}\right)$. We split the samples X according to the value of B into two sets and count the samples in each set separately. That is, we define $N = (N_1, \ldots, N_k)$ and $N' = (N'_1, \ldots, N'_k)$ by

$$N_i = \sum_{j=1}^{n'} \mathbf{1}_{\{X_j = i\}} \mathbf{1}_{\{B_j = 0\}}, \quad N'_i = \sum_{j=1}^{n'} \mathbf{1}_{\{X_j = i\}} \mathbf{1}_{\{B_j = 1\}}.$$

Then N and N' are independent, where $N_i, N'_i \overset{\text{i.i.d.}}{\sim} \mathrm{Poi}\,(np_i)$.

Let $c_0, c_1, c_2 > 0$ be constants to be specified. Let $L = \lfloor c_0 \log k \rfloor$. Denote the best polynomial of degree L to uniformly approximate $x \log \frac{1}{x}$ on $[0, 1]$ by

$$p_L(x) = \sum_{m=0}^{L} a_m x^m. \tag{4.16}$$

Through a change of variables, we see that the best polynomial of degree L

to approximate $x \log \frac{1}{x}$ on $[0, \beta]$, where $\beta = \frac{c_1 \log k}{n}$, is

$$P_L(x) \triangleq \sum_{m=0}^{L} a_m \beta^{1-m} x^m - x \log \beta.$$

Define the factorial moment by $(x)_m \triangleq \frac{x!}{(x-m)!}$, which gives an unbiased estimator for the monomials of the Poisson mean: $\mathbb{E}[(X)_m] = \lambda^m$ where $X \sim \mathrm{Poi}(\lambda)$. Consequently, the polynomial of degree L,

$$g_L(N_i) \triangleq \frac{1}{n} \left(\sum_{m=0}^{L} \frac{a_m}{(c_1 \log k)^{m-1}} (N_i)_m - N_i \log \beta \right), \tag{4.17}$$

is an unbiased estimator for $P_L(p_i)$.

Define a preliminary estimator of entropy $H(P) = \sum_{i=1}^{k} \phi(p_i)$ by

$$\tilde{H} \triangleq \sum_{i=1}^{k} \left(g_L(N_i) \mathbf{1}_{\{N_i' \le T\}} + g(N_i) \mathbf{1}_{\{N_i' > T\}} \right), \tag{4.18}$$

where $T = c_2 \log k$, $g(j) = \phi(j/n) + \frac{1}{2n}$, and we apply the estimator from polynomial approximation if $N_i' \le T$ or the bias-corrected plug-in estimator otherwise (cf. the asymptotic expansion (4.9) of the bias under the original sampling model). In view of the fact that $0 \le H(P) \le \log k$ for any distribution P with alphabet size k, we define our final estimator by

$$\hat{H} = (\tilde{H} \vee 0) \wedge \log k.$$

The next result gives an upper bound on the above estimator under the Poisson sampling model, which, in view of the right inequality in (3.3) and (4.4), implies the upper bound on the minimax risk $R^*(n, k)$ in (4.2).

Proposition 4.1. *Assume that $\log n \le C \log k$ for some constant $C > 0$. Then there exists c_0, c_1, c_2 depending on C only, such that*

$$\sup_{P \in \mathcal{M}_k} \mathbb{E}[(H(P) - \hat{H}(N))^2] \lesssim \left(\frac{k}{n \log k} \right)^2 + \frac{\log^2 k}{n},$$

where $N = (N_1, \ldots, N_k) \overset{ind}{\sim} \mathrm{Poi}(np_i)$.

Before proving the above statistical guarantee, we make a few comments

on the optimal estimator.

Computation complexity. The estimate $\tilde{H}$ in (4.18) can be expressed in terms of a linear combination of the fingerprints (see (4.12)) of the second half of samples. The coefficients $\{a_m\}$ can be *pre-computed* using fast best polynomial approximation algorithms (e.g., Algorithm 2.1 due to Remez), it is clear that the estimator $\hat{H}$ can be computed in linear time in n, which is sublinear in the alphabet size.

Difficulty in entropy estimation. The estimator in this section uses the polynomial approximation of $x \mapsto x \log \frac{1}{x}$ for those masses below $\frac{\log k}{n}$ and the bias-reduced plug-in estimator otherwise. This suggests that the main difficulty of entropy estimation lies in those probabilities in the interval $[0, \frac{\log k}{n}]$, which are individually small but collectively contribute significantly to the entropy. In Section 4.2.3, to prove a minimax lower bound, the pair of unfavorable priors consists of randomized distributions whose masses are below $\frac{\log k}{n}$ (except for possibly a fixed large mass at the last element). See Remark 4.4 and the proof of Proposition 4.3 for details.

Bias reduction from polynomial approximation. To show the effect of bias reduction using the best polynomial approximation, we illustrate $\phi(p) - \mathbb{E}[\tilde{g}(N)]$ as a function of p, where $N \sim \text{binomial}(n, p)$ and

$$\tilde{g}(j) = \begin{cases} g_L(j), & j \le T, \\ \phi(j/n) + \frac{1-(j/n)}{2n}, & j > T. \end{cases}$$

Here g_L is obtained by (4.17) using the best polynomial approximation. We also compare with that of the Miller-Madow estimate where $\tilde{g}'(j) = \phi(j/n) + \frac{1-(j/n)}{2n}$ for every j. In Figure 4.4, we take a sample size $n = 100$; $g_L(j)$ is obtained using the best polynomial of degree four to approximation ϕ on $[0, 0.06]$, and is applied with $T = 3$. We can clearly see the improvement on the bias as compared to the Miller-Madow estimate when p is small.

Adaptivity. The estimator in (4.18) depends on the alphabet size k only through its logarithm; therefore the dependence on the alphabet size is rather insensitive. In many applications such as neuroscience the discrete data are

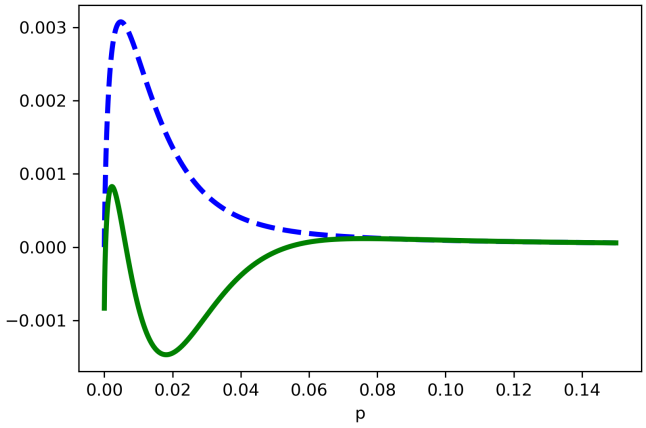


Figure 4.4: Comparison of the bias of estimators for $\phi(p)$ using $N \sim \text{binomial}(n, p)$. The green solid line is the bias of the polynomial estimator $\tilde{g}(N)$ as a function of p; the blue dashed line shows the bias of the Miller-Madow estimator $\tilde{g}'(N)$.

obtained from quantizing an analog source and k is naturally determined by the quantization level [25]. Nevertheless, it is also desirable to obtain an optimal estimator that is adaptive to k. To this end, we can replace all $\log k$ by $\log n$ and define the final estimator by $\tilde{H} \vee 0$. Moreover, we need to set $g_L(0) = 0$ since the number of unseen symbols is unknown. Following [69], we can simply let the constant term a_0 of the approximating polynomial (4.16) go to zero and obtain the corresponding unbiased estimator (4.17) through factorial moments, which satisfies $g_L(0) = 0$ by construction.[2] The bias upper bound becomes $\sum_i (P_L(p_i) - \phi(p_i) - P_L(0))$ which is at most twice the original upper bound since $P_L(0) \le \|P_L - \phi\|_\infty$. The minimax rate in Proposition 4.1 continues to hold in the regime of $\frac{k}{\log k} \lesssim n \lesssim \frac{k^2}{\log^2 k}$, where the plug-in estimator fails to attain the minimax rate. In fact, $P_L(0)$ is always strictly positive and coincides with the uniform approximation error (see Remark 4.1 for a short proof). Therefore, removing the constant term leads to $g_L(N_i)$ which is always underbiased as shown in Figure 4.5. A better choice for adaptive estimation is to find the best polynomial satisfying $p_L(0) = 0$ that uniformly approximates ϕ.

[2]Alternatively, we can directly set $g_L(0) = 0$ and use the original $g_L(j)$ in (4.17) when $j \ge 1$. Then the bias becomes $\sum_i (P_L(p_i) - \phi(p_i) - \mathbb{P}[N_i = 0] P_L(0))$. In sublinear regime that $n = o(k)$, we have $\sum_i \mathbb{P}[N_i = 0] = \Theta(k)$; therefore this modified estimator also achieves the minimax rate.

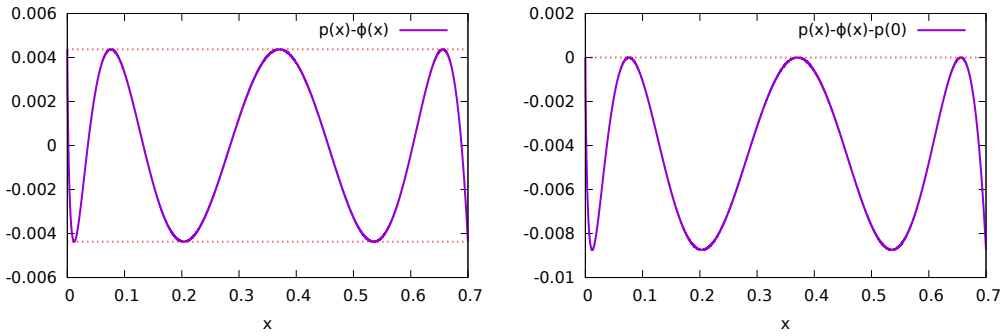


Figure 4.5: Bias of the degree-6 polynomial estimator with and without the constant term.

Sample splitting. The benefit of sample splitting is that we can first condition on the realization of N' and treat the indicators in (4.18) as deterministic, which has also been used in the entropy estimator in [69]. Although not ideal operationally or aesthetically, this is a frequently used idea in statistics and learning to simplify the analysis (also known as sample cloning in the Gaussian model [98, 39]) at the price of losing half of the sample thereby inflating the risk by a constant factor. It remains to be shown whether the optimality result in Proposition 4.1 continues to hold if we can use the same sample in (4.18) for both selection and estimation.

Note that the estimator (4.18) is *linear* in the fingerprint of the second half of the sample. We also note that for estimating other distribution functionals, e.g., support size [72], it is possible to circumvent sample splitting by directly using a linear estimator obtained from best polynomial approximation.

4.2.3 Statistical guarantees of the optimal estimator

Given that N_i' is above (resp. below) the threshold T, we can conclude with high confidence that p_i is above (resp. below) a constant factor of T using the Chernoff bound for Poissons ([56, Theorem 5.4]): if $N \sim \mathrm{Poi}(np)$, then

$$\mathbb{P}[N \geq T] \leq \exp(-T(\alpha_1 - 1 - \log \alpha_1)), \quad p < \alpha_1 T/n, \tag{4.19}$$

$$\mathbb{P}[N \leq T] \leq \exp(-T(\alpha_2 - 1 - \log \alpha_2)), \quad p > \alpha_2 T/n, \tag{4.20}$$

where $\alpha_1 < 1 < \alpha_2$. We apply the estimator from bias-correct plug-in estimator if $N_i' > T$ and the polynomial estimator otherwise. Next we analyze

the two cases separately.

Bias-corrected plug-in estimator. When p is large, (4.8) provides a precise estimate of the bias of empirical entropy (see Figure 4.3). In this regime, using bias reduction similar to the Miller-Madow estimate (4.10) to estimate $\phi(p)$ by $g(N)$ given in (4.18). The bias and variance of $g(N)$ are analyzed in the following.

Lemma 4.4.

$$-\frac{1}{6n^2p} \le \mathbb{E}[\phi(p) - g(N)] \le \frac{1}{3n^3p^2} + \frac{5}{6n^2p}, \tag{4.21}$$

$$\mathsf{var}[g(N)] \le \frac{2p\log^2(ep)}{n} + \frac{2(1+3np)}{n^3p}, \tag{4.22}$$

where $N \sim \mathrm{Poi}(np)$.

Polynomial estimator. When p is small, the function $\phi(p)$ is approximated by $P_L(p)$, which can be estimated by $g_L(N)$ in (4.17), which is an unbiased estimate for $P_L(p)$. Therefore the bias of $g_L(N)$ as an estimate for $\phi(p)$ is at most the approximation error, which is obtained in [34, Section 7.5.4]:

$$\sup_{x\in[0,\beta]} |P_L(x) - \phi(x)| \le O(\beta/L^2). \tag{4.23}$$

We have the following upper bound on the standard deviation of $g_L(N)$.

Lemma 4.5. *Let* $\sigma(g_L(N))$ *denote the standard deviation of* $g_L(N)$. *We have*

$$\sigma(g_L(N)) \le \beta \sum_{m=0}^{L} |a_m| \left(\frac{mp}{n\beta^2}\right)^{m/2} (2e)^{\sqrt{mnp}} + \sqrt{p/n}\log\beta, \tag{4.24}$$

where $N \sim \mathrm{Poi}(np)$.

Combining these two regimes, we now prove Proposition 4.1.

Proof. With the threshold $T = c_2 \log k$, by (4.19) and (4.19), with probability at least $1-\delta$ such that $\delta = k^{1-c_2(\frac{c_1}{c_2}-\log\frac{ec_1}{c_2})} + k^{1-c_2(\frac{c_3}{c_2}-\log\frac{ec_3}{c_2})}$, we have

$$N_i' \ge T \Rightarrow p_i > c_3 \log k/n, \quad N_i' \le T \Rightarrow p_i < c_1 \log k/n, \quad \forall\ i.$$

The above implications fail with probability at most δ. In this case, we have $|H - \hat{H}| \leq \log k$. Define two set of indices $I_1 = \{i : N_i' > T\}$ and $I_2 = \{i : N_i' \leq T\}$. In the remaining proof the above high probability event is assumed to have occurred. Hence, we have

$$p_i > c_3 \log k/n, \ \forall \ i \in I_1; \quad p_i < c_1 \log k/n, \ \forall \ i \in I_2.$$

We first consider I_1. Denote the error by

$$\mathcal{E}_1 = \sum_{i \in I_1} \phi(p_i) - g(N_i).$$

The mean and variance of $\mathcal{E}_1$ is upper bounded using (4.21) and (4.22), respectively, by

$$\begin{aligned} |\mathbb{E}[\mathcal{E}_1]| &\leq \frac{k}{3n(c_3 \log k)^2} + \frac{5k}{6nc_3 \log k} \lesssim \frac{k}{n \log k}, \\ \mathsf{var}[\mathcal{E}_1] &\leq \frac{2}{n} \sum_{i \in I_1} p_i \log^2(ep_i) + \frac{2k(1 + 3c_3 \log k)}{n^2 c_3 \log k} \lesssim \frac{\log^2 k}{n} + \frac{k}{n^2}, \end{aligned}$$

where in the variance bound we used the concavity of $x \mapsto \log^2(ex)$ on $[1, \infty)$. These upper bounds yield that

$$\mathbb{E}(\mathcal{E}_1)^2 \lesssim \left(\frac{k}{n \log k} \right)^2 + \frac{\log^2 k}{n}.$$

Now we consider I_2. Denote the error similarly by

$$\mathcal{E}_2 = \sum_{i \in I_2} \phi(p_i) - g(N_i).$$

The bias is upper bounded by the uniform approximation error (4.23) by

$$|\mathbb{E}[\mathcal{E}_2]| \leq \sum_{i \in I_2} \frac{c_1 \log k}{n} O(1/L^2) \leq O\left(\frac{k}{n \log k} \right).$$

If we choose the polynomial degree L such that $L \leq c_1 \log k$, then in (4.24) we have $\frac{mp}{n\beta^2} \leq 1$ for $p \leq \beta$. Then the variance of $\mathcal{E}_2$ is upper bounded by

$$\mathsf{var}[\mathcal{E}_2] \leq k \left(\frac{c_1 \log k L \max_m |a_m|}{n} (2e)^{\sqrt{c_1 \log kL}} \right)^2.$$

We additionally need an upper bound on the magnitude of the coefficients a_m. Recall that $\sum_m a_m x^m$ is the best polynomial to approximate ϕ on $[0,1]$, which is bounded by e^{-1}. Then the approximation error is at most e^{-1} and thus p_L is bounded by $2e^{-1}$. For a bounded polynomial the coefficients are at most $2e^{-1}2^{3L}$. Since $L = \lfloor c_0 \log k \rfloor$, we have

$$\mathsf{var}[\mathcal{E}_2] \lesssim \frac{(\log k)^4}{n^2} k^{1+2(c_0 \log 8 + \sqrt{c_0 c_1}\log(2e))}.$$

The above variance upper bound is $O(\frac{k}{n \log k})^2$ as long as $c_0 \log 8 + \sqrt{c_0 c_1}\log(2e) < 1/4$.

To conclude the proof, we specify all the constants. By assumption, $\log n \le C \log k$ for some constant C. Choose $c_1 > c_2 > c_3 > 0$ such that $c_2(\frac{c_1}{c_2} - \log \frac{e c_1}{c_2}) - 1 > C$ and $c_2(\frac{c_3}{c_2} - \log \frac{e c_3}{c_2}) - 1 > C$ hold simultaneously, e.g., $c_2 = C + 1$, $c_1 = 4c_2$, $c_3 = 0.1 c_2$. Choosing $c_0 = \frac{1}{300 c_1} \wedge c_1 \wedge 0.01$ completes the proof. □

Remark 4.1 (Approximation error at the end points)**.** By Chebyshev alternating theorem [99, Theorem 1.6], the error function $g(x) \triangleq P_L(x) - \phi(x)$ attains uniform approximation error (namely, $\pm E_L(\phi)$) on at least $L+2$ points with alternative change of signs; moreover, these points must be stationary points or endpoints. Taking derivatives, $g'(x) = P_L'(x) + \log(ex)$ and $g''(x) = \frac{x P_L''(x) + 1}{x}$. Since g'' has at most $L-1$ roots in $(0,1)$ and hence g' has at most $L-1$ stationary points, the number of roots of g' and hence the number of stationary points of g in $(0,1)$ are at most L. Therefore the error at the ends points must be maximal, i.e., $|g(0)| = |g(1)| = E_L(\phi)$. To determine the sign, note that $g'(0) = -\infty$ then $g(0)$ must be positive for otherwise the value of g at the first stationary point is below $-E_L(\phi)$ which is a contradiction. Hence $a_0 = g(0) = E_L(\phi)$.

4.2.4 Proof of lemmas

Proof of Lemma 4.4. Let $\hat{p}$ denote the ratio N/n. We first analyze the bias which can be expressed as $\mathbb{E}[\phi(p) - \phi(\hat{p}) - \frac{1}{2n}]$, and prove (4.21). Applying Taylor's expansion of ϕ yields that

$$\phi(\hat{p}) = \phi(p) - \log(ep)(\hat{p} - p) - \frac{1}{2p}(\hat{p} - p)^2 + \frac{1}{6p^2}(\hat{p} - p)^3 - R_3(\hat{p}),$$

where $R_3(\hat{p})$ is the remainder and can be expressed using Taylor's theorem for $\hat{p} > 0$ as

$$R_3(\hat{p}) = \frac{1}{3}\int_p^{\hat{p}} \left(\frac{\hat{p}}{t} - 1\right)^3 \mathrm{d}t.$$

If $\hat{p} \geq p$, then the integrand is non-negative and is at most $(\frac{\hat{p}}{p} - 1)^3$. Hence, we obtain that

$$0 \leq R_3(\hat{p}) \leq \frac{(\hat{p} - p)^4}{3p^3};$$

if $0 < \hat{p} < p$, the integral can be rewritten as $\int_{\hat{p}}^{p}(1 - \frac{\hat{p}}{t})^3 \mathrm{d}t$, and the same inequalities are obtained; the above inequalities obviously hold for $\hat{p} = 0$. Using the central moments Poisson distribution:

$$\mathbb{E}(X - \lambda)^2 = \lambda,\ \mathbb{E}(X - \lambda)^3 = \lambda,\ \mathbb{E}(X - \lambda)^4 = \lambda(1 + 3\lambda), \quad X \sim \mathrm{Poi}(\lambda),$$

we obtain the following:

$$-\frac{1}{6n^2p} \leq \mathbb{E}[\phi(p) - g(N)] \leq \frac{1 + 3np}{3n^3p^2} - \frac{1}{6n^2p}.$$

Now we analyze the variance and prove (4.22). The variance can be upper bounded by the mean square error $\mathbb{E}(\phi(p) - \phi(\hat{p}))^2$. Applying Taylor's expansion of ϕ again yields that

$$\phi(\hat{p}) = \phi(p) - \log(ep)(\hat{p} - p) - R_1(\hat{p}),$$

where the remainder $R_1(\hat{p})$ can be expressed using Taylor's theorem for $\hat{p} > 0$ as

$$R_1(\hat{p}) = \int_p^{\hat{p}} \left(\frac{\hat{p}}{t} - 1\right) \mathrm{d}t.$$

Analogous to the previous inequalities for R_3, we obtain that

$$0 \leq R_1(\hat{p}) \leq \frac{(\hat{p} - p)^2}{p}, \quad \hat{p} \geq 0.$$

Applying the triangle inequality yields that

$$\mathbb{E}(\phi(p) - \phi(\hat{p}))^2 \leq \frac{2p\log^2(ep)}{n} + \frac{2(1 + 3np)}{n^3p}. \qquad \square$$

Proof of Lemma 4.5. The standard deviation of sum of random variables is

at most the sum of individual standard deviations. Let $\sigma(X)$ denote the standard deviation of a random variable X. Then

$$\sigma(g_L(N)) \le \beta \sum_{m=0}^{L} |a_m| \frac{\sigma((N)_m)}{(n\beta)^m} + \frac{\sigma(N)}{n} \log \beta. \tag{4.25}$$

The variance of $(N)_m$ is analyzed in the following.

Lemma 4.6. *Let $X \sim \mathrm{Poi}(\lambda)$. Then*

$$\mathsf{var}(X)_m = \lambda^m m! \sum_{k=0}^{m-1} \binom{m}{k} \frac{\lambda^k}{k!} \le (\lambda m)^m (2e)^{2\sqrt{\lambda m}}. \tag{4.26}$$

Proof. The equality part follows from (2.25). We prove the inequality part. Using $\binom{m}{k} \le \frac{m^k}{k!}$, we have

$$\mathsf{var}(X)_m \le \lambda^m m! \sum_{k=0}^{m-1} \frac{(\lambda m)^k}{(k!)^2}.$$

The maximal term in the summation is attained at $k^* = \lfloor \sqrt{\lambda m} \rfloor$. Therefore we obtain that

$$\mathsf{var}(X)_m \le \lambda^m m! m \frac{(\lambda m)^{k^*}}{(k^*!)^2} \le (\lambda m)^m \frac{(\lambda m)^{k^*}}{(k^*!)^2}.$$

If $\lambda m < 1$ then $k^* = 0$ and $\frac{(\lambda m)^{k^*}}{(k^*!)^2} = 1$; otherwise $\lambda m \ge 1$ and hence $\frac{\sqrt{\lambda m}}{2} < k^* \le \sqrt{\lambda m}$. Applying $k^*! > \left(\frac{k^*}{e}\right)^{k^*}$ yields that

$$\frac{(\lambda m)^{k^*}}{(k^*!)^2} \le \frac{(\lambda m)^{k^*}}{\left(\frac{\lambda m}{4e^2}\right)^{k^*}} \le (2e)^{2\sqrt{\lambda m}}. \qquad \square$$

Remark 4.2. Note that the formula of $\mathbb{E}(X)_m^2$ obtained above coincides with $\lambda^m m! \mathcal{L}_m(-\lambda)$, where $\mathcal{L}_m$ denotes the Laguerre polynomial of degree m (see (2.22)). The term $e^{\sqrt{\lambda m}}$ agrees with the sharp asymptotics of the Laguerre polynomial on the negative axis [53, Theorem 8.22.3].

In the last term of (4.25), $\sigma(N)$ can be explicitly evaluated to be $\sqrt{np}$.

Using (4.26), the summation in (4.25) is upper bounded by

$$\sum_{m=0}^{L} |a_m| \left(\frac{mp}{n\beta^2}\right)^{m/2} (2e)^{\sqrt{mnp}}. \qquad \square$$

4.2.5 Numerical experiments

In this subsection, we compare the performance of our entropy estimator to other estimators using synthetic data.[3] Note that the coefficients of best polynomial to approximate ϕ on $[0, 1]$ are independent of data so they can be pre-computed and tabulated to facilitate the computation in our estimation. It is very efficient to apply the Remez algorithm which provably has linear convergence for all continuous functions to obtain those coefficients (see, e.g., [99, Theorem 1.10]). Considering that the choice of the polynomial degree is logarithmic in the alphabet size, we pre-compute the coefficients up to degree 400 which suffices for practically all purposes. In the implementation of our estimator we replace N_i' by N_i in (4.18) without conducting sample splitting. Though in the proof of theorems we are conservative about the constant parameters c_0, c_1, c_2, in experiments we observe that the performance of our estimator is in fact not sensitive to their value within the reasonable range. In the subsequent experiments the parameters are fixed to be $c_0 = c_2 = 1.6, c_1 = 3.5$.

We generate data from four types of distributions over an alphabet of $k = 10^5$ elements, namely, the uniform distribution with $p_i = \frac{1}{k}$, Zipf distributions with $p_i \propto i^{-\alpha}$ and α being either 1 or 0.5, and an "even mixture" of geometric distribution and Zipf distribution where for the first half of the alphabet $p_i \propto 1/i$ and for the second half $p_{i+k/2} \propto (1 - \frac{2}{k})^{i-1}$, $1 \le i \le \frac{k}{2}$. Using parameters mentioned above, the approximating polynomial has degree 18, the parameter determining the approximation interval is $c_1 \log k = 40$, and the threshold to decide which estimator to use in (4.18) is 18; namely, we apply the polynomial estimator g_L if a symbol appeared at most 18 times and the bias-corrected plug-in estimator otherwise. After obtaining the preliminary estimate $\tilde{H}$ in (4.18), our final output is $\tilde{H} \vee 0$.[4] Since the plug-in

[3] The C++ and Python implementation of our estimator is available at `https://github.com/Albuso0/entropy`.

[4] We can, as in Proposition 4.1, output $(\tilde{H} \vee 0) \wedge \log k$, which yields a better performance. We elect not to do so for a stricter comparison.

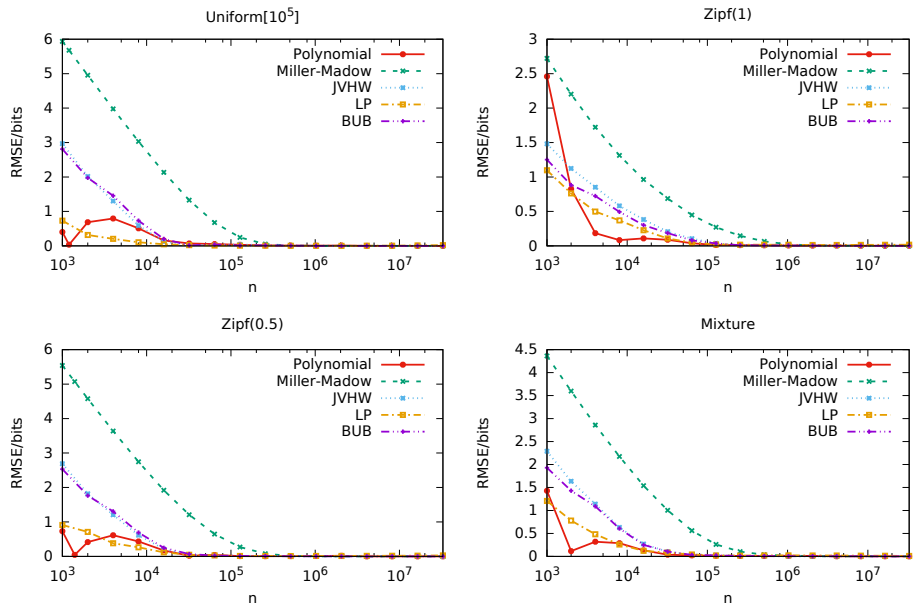


Figure 4.6: Performance comparison with sample size n ranging from 10^3 to 3×10^7.

estimator suffers from severe bias when samples are scarce, we forgo the comparison with it to save space in the figures and instead compare with its bias-corrected version, i.e., the Miller-Madow estimator (4.10). We also compare the performance with the linear programming estimator in [64], the best upper bound (BUB) estimator [66], and the estimator based on similar polynomial approximation techniques[5] proposed by [69] using their implementations with default parameters. Our estimator is implemented in C++ which is much faster than those from [64, 69, 66] implemented in MATLAB so the running time comparison is ignored. We notice that the linear programming in [64] is much slower than the polynomial estimator in [69], especially when the sample size becomes larger.

We compute the root mean squared error (RMSE) for each estimator over 50 trials. The full performance comparison is shown in Figure 4.6 where the sample size ranges from one percent to 300 folds of the alphabet size. In Figure 4.7 we further zoom into the more interesting regime of fewer samples with the sample size ranging from one to five percent of the alphabet size. In this regime our estimator, as well as those from [64, 69, 66], outperforms the

[5]The estimator in [69] uses a smooth cutoff function in lieu of the indicator function in (4.18); this seems to improve neither the theoretical error bound nor the empirical performance.

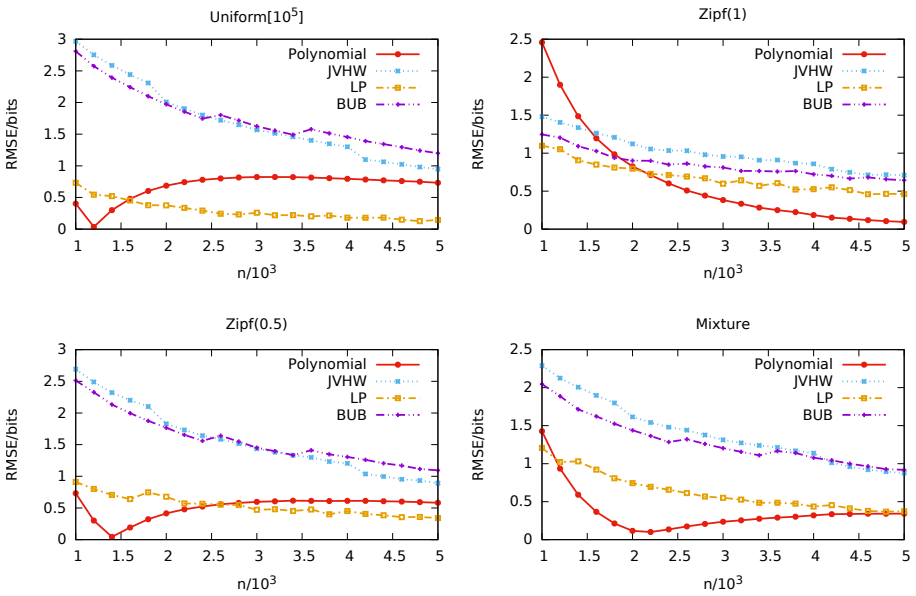


Figure 4.7: Performance comparison when sample size n ranges from 1000 to 5000.

classical Miller-Madow estimator significantly; furthermore, our estimator performs better than those in [69, 66] in most cases tested and comparably with that in [64]. When the samples are abundant all estimators achieve very small error; however, it has been empirically observed in [69] that the performance of linear programming starts to deteriorate when the sample size is very large, which is also observed in our experiments (see [100]). The specific figures of that regime are ignored since the absolute errors are very small and even the plug-in estimator without bias correction is accurate. By (4.18), for large sample size our estimator tends to the Miller-Madow estimator when every symbol is observed many times.

4.3 Fundamental limits of entropy estimation

Thus far, we have described the empirical entropy and the construction of an estimator using the polynomial approximation method such that the bias is smaller than the empirical entropy. The worst-case MSE of both estimators are analyzed. To establish a constant-factor approximation of the fundamental limit of entropy estimation (4.1), we need a matching minimax lower bound. This is the goal of the present section.

To obtain the lower bound part of (4.2), it suffices to show that the minimax risk is lower bounded by the two terms in (4.2) separately. This follows from combining Propositions 4.2 and 4.3.

Proposition 4.2. *For all $k, n \in \mathbb{N}$,*

$$R_H^*(k,n) \gtrsim \frac{\log^2 k}{n}. \tag{4.27}$$

Proposition 4.3. *For all $k, n \in \mathbb{N}$,*

$$R_H^*(k,n) \gtrsim \left(\frac{k}{n \log k}\right)^2 \vee 1. \tag{4.28}$$

Proposition 4.2 follows from a simple application of Le Cam's *two-point method*: If two input distributions P and Q are sufficiently close such that it is impossible to reliably distinguish between them using n samples with error probability less than, say, $\frac{1}{2}$, then any estimator suffers a quadratic risk proportional to the separation of the functional values $|H(P) - H(Q)|^2$.

Proof. For any pair of distributions P and Q, Le Cam's two-point method (see, e.g., [32, Section 2.4.2]) yields

$$R_H^*(k,n) \geq \frac{1}{4}(H(P) - H(Q))^2 \exp(-nD(P\|Q)). \tag{4.29}$$

Therefore it boils down to solving the optimization problem:

$$\sup\{H(P) - H(Q) : D(P\|Q) \leq 1/n\}. \tag{4.30}$$

Without loss of generality, assume that $k \geq 2$. Fix an $\epsilon \in (0,1)$ to be specified. Let

$$P = \left(\frac{1}{3k'}, \ldots, \frac{1}{3k'}, \frac{2}{3}\right), \quad Q = \left(\frac{1+\epsilon}{3k'}, \ldots, \frac{1+\epsilon}{3k'}, \frac{2-\epsilon}{3}\right), \tag{4.31}$$

where $k' = k-1$. Direct computation yields $D(P\|Q) = \frac{2}{3}\log\frac{2}{2-\epsilon} + \frac{1}{3}\log\frac{1}{\epsilon+1} \leq \epsilon^2$ and $H(Q) - H(P) = \frac{1}{3}(\epsilon \log k' + \log 4 + (2-\epsilon)\log\frac{1}{2-\epsilon} + (1+\epsilon)\log\frac{1}{\epsilon+1}) \geq \frac{1}{3}\log(2k')\epsilon - \epsilon^2$. Choosing $\epsilon = \frac{1}{\sqrt{n}}$ and applying (4.29), we obtain the desired (4.27). □

Remark 4.3. In view of the Pinsker inequality $D(P\|Q) \geq 2\mathsf{TV}^2(P,Q)$ [101, p. 58] as well as the continuity property of entropy with respect to the total variation distance, $|H(P) - H(Q)| \leq \mathsf{TV}(P,Q) \log \frac{k}{\mathsf{TV}(P,Q)}$ for $\mathsf{TV}(P,Q) \leq \frac{1}{4}$ [101, Lemma 2.7], we conclude that the best lower bound given by the two-point method, i.e., the supremum in (4.30), is on the order of $\frac{\log k}{\sqrt{n}}$. Therefore the choice of the pair (4.31) is optimal.

The remainder of this section is devoted to proving Proposition 4.3. Since the best lower bound provided by the two-point method is $\frac{\log^2 k}{n}$ (see Remark 4.3), proving (4.28) requires more powerful techniques. To this end, we use a generalized version of Le Cam's method involving two *composite* hypotheses (also known as fuzzy hypothesis testing in [32]):

$$H_0 : H(P) \leq t \quad \text{versus} \quad H_1 : H(P) \geq t + d, \tag{4.32}$$

which is more general than the two-point argument using only simple hypothesis testing. Similarly, if we can establish that no test can distinguish (4.32) reliably, then we obtain a lower bound for the quadratic risk on the order of d^2. By the minimax theorem, the optimal probability of error for the composite hypotheses test is given by the Bayesian version with respect to the least favorable priors. For (4.32) we need to choose a pair of priors, which, in this case, are distributions on the probability simplex $\mathcal{M}_k$, to ensure that the entropy values are separated.

4.3.1 Construction of the priors

The main idea for constructing the priors is as follows: First, the symmetry of the entropy functional implies that the least favorable prior must be permutation-invariant. This inspires us to use the following *i.i.d. construction*. For conciseness, we focus on the case of $n \asymp \frac{k}{\log k}$ for now and our goal is to obtain an $\Omega(1)$ lower bound. Let U be a $\mathbb{R}_+$-valued random variable with unit mean. Consider the random vector

$$\mathsf{P} = \frac{1}{k}(U_1, \ldots, U_k),$$

consisting of i.i.d. copies of U. Note that P itself is *not* a probability distribution; however, the key observation is that, since $\mathbb{E}[U] = 1$, as long as the

variance of U is not too large, the weak law of large numbers ensures that P is *approximately* a probability vector.

Example 4.1. A deterministic $U = 1$ generates

$$\mathsf{P} = \left(\frac{1}{k}, \ldots, \frac{1}{k}\right),$$

which is a uniform distribution over k elements. A binary $U \sim \frac{1}{2}(\delta_0 + \delta_2)$ generates

$$\mathsf{P} = \left(\frac{U_1}{k}, \ldots, \frac{U_k}{k}\right),$$

where roughly half of U_i is two and others are zero. This is approximately an uniform distribution over $k/2$ elements with the support set uniformly chosen at random.

From this viewpoint, the CDF of the random variable $\frac{U}{k}$ plays the role of the *histogram of the distribution* P, which is the central object in the Valiant-Valiant lower bound construction (see [89, Definition 3]). Using a conditioning argument we can show that the distribution of P can effectively serve as a prior.

Next we outline the main ingredients in implementing Le Cam's method:

1. *Functional value separation*: Define $\phi(x) \triangleq x \log \frac{1}{x}$. Note that

$$H(\mathsf{P}) = \sum_{i=1}^{k} \phi\left(\frac{U_i}{k}\right) = \frac{1}{k}\sum_{i=1}^{k} \phi(U_i) + \frac{\log k}{k}\sum_{i=1}^{k} U_i,$$

   which concentrates near its mean $\mathbb{E}[H(\mathsf{P})] = \mathbb{E}[\phi(U)] + \mathbb{E}[U]\log k$ by the law of large numbers. Therefore, given another random variable U' with unit mean, we can obtain P' similarly using i.i.d. copies of U'. Then with high probability, $H(\mathsf{P})$ and $H(\mathsf{P}')$ are separated by the difference of their mean values, namely,

$$\mathbb{E}[H(\mathsf{P})] - \mathbb{E}[H(\mathsf{P}')] = \mathbb{E}[\phi(U)] - \mathbb{E}[\phi(U')],$$

   which we aim to maximize.

2. *Indistinguishability*: Note that given a distribution $P = (p_1, \ldots, p_k)$, the sufficient statistics satisfy $N_i \overset{\text{ind}}{\sim} \text{Poi}(np_i)$. Therefore, if P is drawn

from the distribution of P, then $N = (N_1, \ldots, N_k)$ are i.i.d. distributed according the *Poisson mixture* $\mathbb{E}[\mathrm{Poi}(\frac{n}{k}U)]$. Similarly, if P is drawn from the prior of P', then N is distributed according to $(\mathbb{E}[\mathrm{Poi}(\frac{n}{k}U')])^{\otimes k}$. To establish the impossibility of testing, we need the total variation distance between the two k-fold product distributions to be strictly bounded away from one, for which a sufficient condition is

$$\mathsf{TV}(\mathbb{E}[\mathrm{Poi}(nU/k)], \mathbb{E}[\mathrm{Poi}(nU'/k)]) \le c/k \tag{4.33}$$

for some $c < 1$.

To conclude, we see that the i.i.d. construction fully exploits the independence blessed by the Poisson sampling, thereby reducing the problem to *one dimension*. This allows us to sidestep the difficulty encountered in [89] when dealing with fingerprints which are high-dimensional random vectors with dependent entries.

What remains is the following scalar problem: choose U, U' to maximize $|\mathbb{E}[\phi(U)] - \mathbb{E}[\phi(U')]|$ subject to the constraint (4.33). A commonly used proxy for bounding the total variation distance is *moment matching*, i.e., $\mathbb{E}[U^j] = \mathbb{E}[U'^j]$ for all $j = 1, \ldots, L$. Together with L_∞-norm constraints, a sufficiently large degree L ensures the total variation bound (4.33). Combining the above steps, our lower bound is proportional to the value of the following convex optimization problem (in fact, infinite-dimensional linear programming over probability measures):

$$\begin{aligned} \mathcal{F}_L(\lambda) \triangleq \sup\ & \mathbb{E}[\phi(U)] - \mathbb{E}[\phi(U')], \\ \text{s.t. } & \mathbb{E}[U] = \mathbb{E}[U'] = 1, \\ & \mathbb{E}\left[U^j\right] = \mathbb{E}\left[U'^j\right], \quad j = 1, \ldots, L, \\ & U, U' \in [0, \lambda], \end{aligned} \tag{4.34}$$

for some appropriately chosen $L \in \mathbb{N}$ and $\lambda > 1$ depending on n and k.

Finally, we connect the optimization problem (4.34) to the machinery of *best polynomial approximation*. Denote by $\mathcal{P}_L$ the set of polynomials of degree L and

$$E_L(f, I) \triangleq \inf_{p \in \mathcal{P}_L} \sup_{x \in I} |f(x) - p(x)|, \tag{4.35}$$

which is the best uniform approximation error of a function f over a finite

interval I by polynomials of degree L. We prove that

$$\mathcal{F}_L(\lambda) \geq 2E_L(\log, [1/\lambda, 1]). \tag{4.36}$$

Due to the singularity of the logarithm at zero, the approximation error can be made bounded away from zero if λ grows *quadratically* with the degree L (see (4.55)). Choosing $L \asymp \log k$ and $\lambda \asymp \log^2 k$ leads to the impossibility of consistent estimation for $n \asymp \frac{k}{\log k}$. For $n \gg \frac{k}{\log k}$, the lower bound for the quadratic risk follows from relaxing the unit-mean constraint in (4.34) to $\mathbb{E}[U] = \mathbb{E}[U'] \leq 1$ and a simple scaling argument. Analogous construction of priors and proof techniques have subsequently been used in [69] to obtain sharp minimax lower bound for estimating the power sum in which case the $\log p$ function is replaced by p^α.

4.3.2 Minimax lower bound from two composite hypotheses

For $0 < \epsilon < 1$, define the set of *approximate* probability vectors by

$$\mathcal{M}_k(\epsilon) \triangleq \left\{ P \in \mathbb{R}_+^k : \left| \sum_{i=1}^k p_i - 1 \right| \leq \epsilon \right\}, \tag{4.37}$$

which reduces to the probability simplex $\mathcal{M}_k$ if $\epsilon = 0$. Generalizing the minimax quadratic risk (3.2) for Poisson sampling, we define

$$\tilde{R}^*(k, n, \epsilon) \triangleq \inf_{\hat{H}'} \sup_{P \in \mathcal{M}_k(\epsilon)} \mathbb{E}(\hat{H}'(N) - H(P))^2, \tag{4.38}$$

where $N = (N_1, \ldots, N_k)$ and $N_i \overset{\text{ind}}{\sim} \text{Poi}(np_i)$ for $i = 1, \ldots, k$. Since P is not necessarily normalized, $H(P)$ may not carry the meaning of entropy. Nevertheless, H is still valid a functional. The risk defined above is connected to the risk (4.1) for multinomial sampling by Lemma 4.7.

Lemma 4.7. *For any $k, n \in \mathbb{N}$ and $\epsilon < 1/3$,*

$$R^*(k, n/2) \geq \frac{1}{3}\tilde{R}^*(k, n, \epsilon) - \log^2 k(\epsilon^2 + e^{-n/50}) - \phi^2(1 + \epsilon).$$

To establish a lower bound of $\tilde{R}^*(k, n, \epsilon)$, we apply generalized Le Cam's method involving two composite hypotheses as in (4.32), which entails choos-

ing two priors such that the entropy values are separated with probability one. It turns out that this can be relaxed to separation *on average*, if we can show that the entropy values are concentrated at their respective means. This step is made precise in Lemma 4.8.

Lemma 4.8. *Let U and U' be random variables such that $U, U' \in [0, \lambda]$ and $\mathbb{E}[U] = \mathbb{E}[U'] \leq 1$ and $|\mathbb{E}[\phi(U)] - \mathbb{E}[\phi(U')]| \geq d$, where $\lambda < k/e$. Let $\epsilon = \frac{4\lambda}{\sqrt{k}}$. Then*

$$\tilde{R}^*(k, n, \epsilon) \geq \frac{d^2}{16}\left(\frac{7}{8} - k\mathsf{TV}(\mathbb{E}\left[\mathrm{Poi}\left(nU/k\right)\right], \mathbb{E}\left[\mathrm{Poi}\left(nU'/k\right)\right]) - \frac{32\lambda^2 \log^2 \frac{k}{\lambda}}{kd^2}\right). \tag{4.39}$$

The statistical closeness between two Poisson mixtures is established in Section 3.3. To apply Lemma 4.8 we need to construct two random variables, namely U and U', that have matching moments of order $1, \ldots, L$, and large discrepancy in the mean functional value $|\mathbb{E}[\phi(U)] - \mathbb{E}[\phi(U')]|$, as described in Section 4.3.1 and formulated in (4.34). As shown in Section 2.1.2, we can obtain U, U' with matching moments from the dual of the best polynomial approximation of ϕ, namely (4.35); however, we have little control over the value of the common mean $\mathbb{E}[U] = \mathbb{E}[U']$ and it is unclear whether it is less than one as required by Lemma 4.8. Of course we can normalize U, U' by their common mean which preserves moments matching; however, the mean value separation $|\mathbb{E}[\phi(U)] - \mathbb{E}[\phi(U')]|$ also shrinks by the same factor, which results in a suboptimal lower bound.

To circumvent this issue, we first consider auxiliary random variables X, X' supported on a interval bounded away from 0; leveraging the property that their "zeroth moments" are one, we then construct the desired random variables U, U' via a change of measure. To be precise, given $\eta \in (0, 1)$ and any random variables $X, X' \in [\eta, 1]$ that have matching moments up to the L^{th} order, we can construct U, U' from X, X' with the following distributions

$$\begin{aligned} P_U(\mathrm{d}u) &= \left(1 - \mathbb{E}\left[\frac{\eta}{X}\right]\right)\delta_0(\mathrm{d}u) + \frac{\alpha}{u}P_{\alpha X/\eta}(\mathrm{d}u), \\ P_{U'}(\mathrm{d}u) &= \left(1 - \mathbb{E}\left[\frac{\eta}{X'}\right]\right)\delta_0(\mathrm{d}u) + \frac{\alpha}{u}P_{\alpha X'/\eta}(\mathrm{d}u), \end{aligned} \tag{4.40}$$

for some fixed $\alpha \in (0, 1)$. Since $X, X' \in [\eta, 1]$ and thus $\mathbb{E}\left[\frac{\eta}{X}\right], \mathbb{E}\left[\frac{\eta}{X'}\right] \leq 1$,

these distributions are well-defined and supported on $[0, \alpha\eta^{-1}]$. Furthermore,

$$\mathbb{E}[U] = \mathbb{E}[U'] = \alpha, \tag{4.41}$$

$$\mathbb{E}\left[U^j\right] = \mathbb{E}\left[U'^j\right], \ j = 1, \ldots, L+1, \tag{4.42}$$

$$\mathbb{E}[\phi(U)] - \mathbb{E}[\phi(U')] = \alpha(\mathbb{E}[\log(1/X)] - \mathbb{E}[\log(1/X')]). \tag{4.43}$$

To choose the best X, X', we consider the following auxiliary optimization problem over random variables X and X' (or equivalently, the distributions thereof):

$$\begin{aligned} \mathcal{E}^* = \max \ & \mathbb{E}[\log(1/X)] - \mathbb{E}[\log(1/X')], \\ \text{s.t. } & \mathbb{E}[X^j] = \mathbb{E}[X'^j], \quad j = 1, \ldots, L, \\ & X, X' \in [\eta, 1], \end{aligned} \tag{4.44}$$

where $0 < \eta < 1$. Note that (4.44) is an infinite-dimensional linear programming problem with finitely many constraints. Therefore it is natural to turn to its dual. In Section 2.1.2 we show that the maximum $\mathcal{E}^*$ exists and coincides with twice the best L_∞ approximation error of the log over the interval $[\eta, 1]$ by polynomials of degree L:

$$\mathcal{E}^* = 2E_L(\log, [\eta, 1]). \tag{4.45}$$

By definition, this approximation error is decreasing in the degree L when η is fixed; on the other hand, since the logarithm function blows up near zero, for fixed degree L the approximation error also diverges as η vanishes. As shown in Lemma 4.9, in order for the error to be bounded away from zero which is needed in the lower bound, it turns out that the necessary and sufficient condition is when η decays according to L^{-2}. See Lemma 4.9.

With the above preparations, we now prove the minimax lower bound in Proposition 4.3.

Proof. Let X and X' be the maximizer of (4.44). Now we construct U and U' from X and X' according to the recipe (4.40). By (4.41) – (4.43), the first $L+1$ moments of U and U' are matched with means equal to α which is less than one; moreover,

$$\mathbb{E}[\phi(U)] - \mathbb{E}[\phi(U')] = \alpha\mathcal{E}^*. \tag{4.46}$$

Recall the universal constants c and c' defined in Lemma 4.9. If $n \geq \frac{2k}{\log k}$, let $c_1 \leq 2$ be a constant satisfying $\frac{c}{2}\log\frac{c}{4ec_1} > 2$ and thus $c > 4ec_1$. Let $\eta = \log^{-2} k$, $L = \lfloor c \log k \rfloor \geq \frac{c\log k}{2}$, $\alpha = \frac{c_1 k}{n \log k}$ and $\lambda = \alpha\eta^{-1} = \frac{c_1 k \log k}{n}$. Therefore $\alpha \leq 1$. Using (4.40) and (4.46), we can construct two random variables $U, U' \in [0, \lambda]$ such that $\mathbb{E}[U] = \mathbb{E}[U'] = \alpha$, $\mathbb{E}[U^j] = \mathbb{E}[U'^j]$, for all $j \in [L]$, and $\mathbb{E}\left[\phi(U)\right] - \mathbb{E}\left[\phi(U')\right] = \alpha\mathcal{E}^*$. It follows from (4.45) and Lemma 4.9 that $\mathcal{E}^* \geq 2c'$ and thus $\left|\mathbb{E}\left[\phi(U)\right] - \mathbb{E}\left[\phi(U')\right]\right| \geq 2c'\alpha$. By the choice of c_1, applying Theorem 3.5 yields $\mathsf{TV}(\mathbb{E}\left[\mathrm{Poi}\left(nU/k\right)\right], \mathbb{E}\left[\mathrm{Poi}\left(nU'/k\right)\right]) \leq 2k^{-2}$. Finally, applying Lemma 4.7 and Lemma 4.8 with $d = 2c'\alpha$ yields the desired lower bound $R^*(k, n/2) \gtrsim \alpha^2 \asymp (\frac{k}{n\log k})^2$. Consequently, $R_H^*(k, n) \gtrsim (\frac{k}{n\log k})^2$ when $n \geq \frac{k}{\log k}$. If $n \leq \frac{k}{\log k}$ by monotonicity, $R_H^*(k, n) \geq R^*(k, \frac{k}{\log k}) \gtrsim 1$. □

Remark 4.4 (Structure of the least favorable priors)**.** From Theorem 2.6, we conclude that X, X' are in fact discrete random variables each of which has $L + 2 \asymp \log k$ atoms, and their support sets are disjoint. Therefore U, U' are also finitely valued; however, our proof does not rely on this fact. Nevertheless, it is instructive to discuss the structure of the prior. Except for possibly a fixed large mass, the masses of random distributions P and P' are drawn from the distribution U and U' respectively, which lie in the interval $[0, \frac{\log k}{n}]$. Therefore, although P and P' are distributions over k elements, they only have $\log k$ distinct masses and the locations are randomly permuted. Moreover, the entropy of P and P' constructed based on U and U' (see (4.48)) are concentrated near the respective mean values, both of which are close to $\log k$ but differ by a constant factor of $\frac{k}{n\log k}$.

4.3.3 Proof of lemmas

Proof of Lemma 4.7. Denote the left-hand side of the above equation be $R_{\hat{H}}$. For a fixed sample size m, there exists an estimator, e.g., the minimax estimator, denoted by $\hat{H}(\cdot, m)$, such that

$$\sup_{P \in \mathcal{M}_k} \mathbb{E}[(\hat{H}(N, m) - H(P))^2] \leq \begin{cases} R_{\hat{H}}, & m \geq n/2, \\ \log^2 k, & m < n/2. \end{cases} \tag{4.47}$$

Using the sequence $\{\hat{H}(\cdot, m) : m \in \mathbb{N}\}$, we construct an estimator for the functional $H(P)$, where $P = (p_1, \ldots, p_k) \in \mathcal{M}_k(\epsilon)$, using statistics $N =$

$(N_1, \ldots, N_k)$ with $N_i \overset{\text{ind}}{\sim} \text{Poi}(np_i)$. Denote the total number of samples by $n' = \sum_i N_i$. The estimator is defined by

$$\tilde{H}(N) = \hat{H}(N, n').$$

The functional $H(P)$ is related to entropy of the normalized P by

$$H(P) = \sum_{i=1}^{k} p_i \log \frac{1}{p_i} = \phi(s) + sH(\tilde{P}),$$

where s denotes the summation of all p_i, which differs from one by at most ϵ, and $\tilde{P} = P/s$ denotes the normalized distribution. Applying the triangle inequality yields that

$$\frac{1}{3}(\tilde{H}(N) - H(P))^2 \le (\tilde{H}(N) - H(\tilde{P}))^2 + ((1-s)H(\tilde{P}))^2 + \phi^2(s).$$

In the right-hand side of the above inequality, the second terms is at most $(\epsilon \log k)^2$, and the third term is at most $\phi^2(1+\epsilon)$ since ϕ is increasing on $[0, 1/3]$. For the first term, we observe that conditioned on $n' = m$, $N \sim$ multinomial$(m, \tilde{P})$. Hence, we have

$$\mathbb{E}(\tilde{H}(N) - H(\tilde{P}))^2 = \sum_{m=0}^{\infty} \mathbb{E}[(\hat{H}(N, m) - H(\tilde{P}))^2 | n' = m] \mathbb{P}[n' = m].$$

Using (4.47), we obtain that

$$\mathbb{E}(\tilde{H}(N) - H(\tilde{P}))^2 \le R_{\hat{H}} + \log^2 k \mathbb{P}[n' < n/2].$$

Combining the above inequalities yields a lower bound on $R_{\hat{H}}$. In the statement of lemma we applied $\mathbb{P}[n' < n/2] < e^{-n/50}$ by the Chernoff bound ([56, Theorem 5.4]). □

Proof of Lemma 4.8. Denote the common mean by $\alpha \triangleq \mathbb{E}[U] = \mathbb{E}[U'] \le 1$. Define two random vectors

$$\mathsf{P} = \left(\frac{U_1}{k}, \ldots, \frac{U_k}{k}, 1-\alpha\right), \quad \mathsf{P}' = \left(\frac{U_1'}{k}, \ldots, \frac{U_k'}{k}, 1-\alpha\right), \tag{4.48}$$

where U_i, U_i' are i.i.d. copies of U, U', respectively. Note that $\epsilon = \frac{4\lambda}{\sqrt{k}} \ge$

$4\sqrt{\frac{\mathsf{var}[U]\vee\mathsf{var}[U']}{k}}$. Define the following events indicating that U_i and $H(\mathsf{P})$ are concentrated near their respective mean values:

$$E \triangleq \left\{\left|\sum_i \frac{U_i}{k} - \alpha\right| \le \epsilon, |H(\mathsf{P}) - \mathbb{E}\left[H(\mathsf{P})\right]| \le \frac{d}{4}\right\},$$
$$E' \triangleq \left\{\left|\sum_i \frac{U_i'}{k} - \alpha\right| \le \epsilon, |H(\mathsf{P}') - \mathbb{E}\left[H(\mathsf{P}')\right]| \le \frac{d}{4}\right\}.$$

Using the independence of U_i, Chebyshev's inequality and union bound yield that

$$\begin{aligned}\mathbb{P}\left[E^c\right] &\le \mathbb{P}\left[\left|\sum_i \frac{U_i}{k} - \alpha\right| > \epsilon\right] + \mathbb{P}\left[|H(\mathsf{P}) - \mathbb{E}\left[H(\mathsf{P})\right]| > \frac{d}{4}\right] \\ &\le \frac{\mathsf{var}[U]}{k\epsilon^2} + \frac{16\sum_i \mathsf{var}[\phi(U_i/k)]}{d^2} \le \frac{1}{16} + \frac{16\lambda^2\log^2\frac{k}{\lambda}}{kd^2}, \end{aligned} \tag{4.49}$$

where the last inequality follows from the fact that $\mathsf{var}\left[\phi\left(\frac{U_i}{k}\right)\right] \le \mathbb{E}\left[\phi\left(\frac{U_i}{k}\right)\right]^2 \le \left(\phi\left(\frac{\lambda}{k}\right)\right)^2$ when $\lambda/k < e^{-1}$ by assumption. By the same reasoning,

$$\mathbb{P}\left[E'^c\right] \le \frac{1}{16} + \frac{16\lambda^2\log^2\frac{k}{\lambda}}{kd^2}. \tag{4.50}$$

Note that conditioning on E and E' the random vectors in (4.48) belong to $\mathcal{M}_k(\epsilon)$. Now we define two priors on the set $\mathcal{M}_k(\epsilon)$ using (4.48) with the following conditional distributions:

$$\pi = P_{\mathsf{P}|E}, \quad \pi' = P_{\mathsf{P}'|E'}.$$

It follows from $H(\mathsf{P}) = \frac{1}{k}\sum_i \phi(U_i) + \frac{\log k}{k}\sum_i U_i + \phi(1-\alpha)$ that $\mathbb{E}\left[H(\mathsf{P})\right] = \mathbb{E}\left[\phi(U)\right] + \mathbb{E}\left[U\right]\log k + \phi(1-\alpha)$. Similarly, $\mathbb{E}\left[H(\mathsf{P}')\right] = \mathbb{E}\left[\phi(U')\right] + \mathbb{E}\left[U'\right]\log k + \phi(1-\alpha)$. By assumption $|\mathbb{E}\left[H(\mathsf{P})\right] - \mathbb{E}\left[H(\mathsf{P}')\right]| = |\mathbb{E}\left[\phi(U)\right] - \mathbb{E}\left[\phi(U')\right]| \ge d$. By the definition of events E, E' and triangle inequality, we obtain that under π, π'

$$|H(\mathsf{P}) - H(\mathsf{P}')| \ge \frac{d}{2}. \tag{4.51}$$

Now we consider the total variation of the sufficient statistics $N = (N_i)$ under two priors. Note that conditioned on p_i, we have $N_i \sim \mathrm{Poi}(np_i)$. The

triangle inequality of total variation then yields

$$\begin{aligned}\mathsf{TV}\left(P_{N|E}, P_{N'|E'}\right) \leq & \mathsf{TV}\left(P_{N|E}, P_N\right) + \mathsf{TV}\left(P_N, P_{N'}\right) + \mathsf{TV}\left(P_{N'}, P_{N'|E'}\right) \\ = & \mathbb{P}\left[E^c\right] + \mathsf{TV}\left(P_N, P_{N'}\right) + \mathbb{P}\left[E'^c\right] \\ \leq & \mathsf{TV}\left(P_N, P_{N'}\right) + \frac{1}{8} + \frac{32\lambda^2 \log^2 \frac{k}{\lambda}}{kd^2}, \end{aligned} \tag{4.52}$$

where in the last inequality we have applied (4.49)–(4.50). Note that $P_N, P_{N'}$ are marginal distributions under priors $P_{\mathsf{P}}, P_{\mathsf{P}'}$ respectively. In view of the fact that the total variation between product distributions is at most the sum of total variations of pair of marginals, we obtain

$$\begin{aligned}\mathsf{TV}\left(P_N, P_{N'}\right) \leq & \sum_{i=1}^{k} \mathsf{TV}\left(P_{N_i}, P_{N_i'}\right) + \mathsf{TV}(\mathrm{Poi}(n(1-\alpha)), \mathrm{Poi}(n(1-\alpha))) \\ = & k\mathsf{TV}(\mathbb{E}\left[\mathrm{Poi}\left(nU/k\right)\right], \mathbb{E}\left[\mathrm{Poi}\left(nU'/k\right)\right]). \end{aligned} \tag{4.53}$$

Then it follows from (4.51)–(4.53) and Le Cam's lemma [96] that

$$\tilde{R}^*(k, n, \epsilon) \geq \frac{d^2}{16}\left(\frac{7}{8} - k\mathsf{TV}(\mathbb{E}\left[\mathrm{Poi}\left(nU/k\right)\right], \mathbb{E}\left[\mathrm{Poi}\left(nU'/k\right)\right]) - \frac{32\lambda^2 \log^2 \frac{k}{\lambda}}{kd^2}\right). \tag{4.54}$$

□

4.3.4 Best polynomial approximation of the logarithm function

Lemma 4.9. *There exist universal positive constants c, c', L_0 such that for any $L \geq L_0$,*

$$E_{\lfloor cL \rfloor}(\log, [L^{-2}, 1]) \geq c'. \tag{4.55}$$

Proof of Lemma 4.9. Recall the best uniform polynomial approximation error $E_m(f, I)$ defined in (4.35). Put $E_m(f) \triangleq E_m(f, [-1, 1])$. In the sequel we shall slightly abuse the notation by assuming that $cL \in \mathbb{N}$, for otherwise the desired statement holds with c replaced by $c/2$. Through simple linear transformation we see that $E_{cL}(\log, [L^{-2}, 1]) = E_{cL}(f_L)$ where

$$f_L(x) = -\log\left(\frac{1+x}{2} + \frac{1-x}{2L^2}\right).$$

The difficulty in proving the desired

$$E_{cL}(f_L) \gtrsim 1 \tag{4.56}$$

lies in the fact that the approximand f_L changes with the degree L. In fact, the following asymptotic result has been shown in [34, Section 7.5.3, p. 445]: $E_L(\log(a-x)) = \frac{1+o(1)}{L\sqrt{a^2-1}(a+\sqrt{a^2-1})^L}$ for *fixed* $a > 1$ and $L \to \infty$. In our case $E_{cL}(f_L) = E_{cL}(\log(a-x))$ with $a = \frac{1+L^{-2}}{1-L^{-2}}$. The desired (4.56) would follow if one substituted this a into the asymptotic expansion of the approximation error, which, of course, is not a rigorous approach. To prove (4.56), we need non-asymptotic lower and upper bounds on the approximation error. There exist many characterizations of approximation error, such as Jackson's theorem, in term of various moduli of continuity of the approximand. Let $\Delta_m(x) = \frac{1}{m}\sqrt{1-x^2} + \frac{1}{m^2}$ and define the following modulus of continuity for f (see, e.g., [99, Section 3.4]):

$$\tau_1(f, \Delta_m) = \sup\{|f(x) - f(y)| : x, y \in [-1, 1], |x - y| \le \Delta_m(x)\}.$$

We first state the following bounds on τ_1 for f_L.

Lemma 4.10 (Direct bound)**.**

$$\tau_1(f_L, \Delta_m) \le \log\left(\frac{2L^2}{m^2}\right), \ \forall m \le 0.1L. \tag{4.57}$$

Lemma 4.11 (Converse bound)**.**

$$\tau_1(f_L, \Delta_L) \ge 1, \forall L \ge 10. \tag{4.58}$$

From [99, Theorem 3.13, Lemma 3.1] we know that $E_m(f_L) \le 100\tau_1(f_L, \Delta_m)$. Therefore, for all $c \le 10^{-7} < 0.1$, the direct bound in Lemma 4.10 gives us

$$\begin{aligned}\frac{1}{L}\sum_{m=1}^{cL} E_m(f_L) &\le \frac{100}{L}\sum_{m=1}^{cL}\log\left(\frac{2L^2}{m^2}\right) = 100c\log 2 + \frac{200}{L}\log\frac{L^{cL}}{(cL)!} \\ &< \frac{1}{400} - \frac{100}{L}\log(2\pi cL),\end{aligned}$$

where the last inequality follows from Stirling's approximation $n! > \sqrt{2\pi n}$

$(n/e)^n$. We apply the converse result for approximation in [99, Theorem 3.14] that

$$\tau_1(f_L, \Delta_L) \le \frac{100}{L} \sum_{m=0}^{L} E_m(f_L), \tag{4.59}$$

where $E_0(f_L) = \log L$. Assembling (4.58)–(4.59), we obtain for all $c \le 10^{-7}$ and $L > 10 \vee \left(100 \times 400 \log \frac{1}{2\pi c}\right)$,

$$\begin{aligned} \frac{1}{L} \sum_{m=cL+1}^{L} E_m(f_L) &\ge \frac{1}{100} - \left(\frac{1}{L} E_0(f_L) + \frac{1}{L} \sum_{m=1}^{cL} E_m(f_L) \right) \\ &\ge \frac{1}{100} - \left(\frac{1}{400} + \frac{100 \log \frac{1}{2\pi c}}{L} \right) > \frac{1}{200}. \end{aligned}$$

By definition, the approximation error $E_m(f_L)$ is a decreasing function of the degree m. Therefore for all $c \le 10^{-7}$ and $L > 4 \times 10^4 \log \frac{1}{2\pi c}$,

$$E_{cL}(f_L) \ge \frac{1}{L - cL} \sum_{m=cL+1}^{L} E_m(f_L) \ge \frac{1}{L} \sum_{m=cL+1}^{L} E_m(f_L) \ge \frac{1}{200}. \qquad \square$$

Remark 4.5. From the direct bound Lemma 4.10 we know that $E_{cL}(\log, [1/L^2, 1]) \lesssim 1$. Therefore the bound (4.55) is in fact tight: $E_{cL}(\log, [1/L^2, 1]) \asymp 1$.

Proof of Lemmas 4.10 and 4.11. First we show (4.57). Note that

$$\tau_1(f_L, \Delta_m) = \sup_{x \in [-1,1]} \sup_{y : |x-y| \le \Delta_m(x)} |f_L(x) - f_L(y)|.$$

For fixed $x \in [-1, 1]$, to decide the optimal choice of y we need to consider whether $\xi_1(x) \triangleq x - \Delta_m(x) \ge -1$ and whether $\xi_2(x) \triangleq x + \Delta_m(x) \le 1$. Since ξ_1 is convex, $\xi_1(-1) < -1$ and $\xi_1(1) > -1$, then $\xi_1(x) > -1$ if and only if $x > x_m$, where x_m is the unique solution to $\xi_1(x) = -1$, given by

$$x_m = \frac{m^2 - m^4 + \sqrt{-m^2 + 3m^4}}{m^2 + m^4}. \tag{4.60}$$

Note that Δ_m is an even function and thus $\xi_2(x) = -\xi_1(-x)$. Then $\xi_2(x) < 1$ if and only if $x < -x_m$.

Since f_L is strictly decreasing and convex, for fixed x and $d > 0$ we have $f_L(x-d) - f_L(x) > f_L(x) - f_L(x+d) > 0$ as long as $-1 < x - d < x + d < 1$.

If $m \geq 2$ since $\xi_1(0) > -1$ then $x_m < 0$ and $-x_m > 0$. Therefore,

$$\tau_1(f_L, \Delta_m) = \sup_{x<x_m} \{f_L(x) - f_L(\xi_2(x))\} \vee \sup_{x<x_m} \{f_L(-1) - f_L(x)\} \vee \sup_{x\geq x_m} \{f_L(\xi_1(x)) - f_L(x)\}.$$

Note that the second term in the last inequality is dominated by the third term since $f_L(\xi_1(x_m)) - f_L(x_m) = f_L(-1) - f_L(x_m) > f_L(-1) - f_L(x)$ for any $x < x_m$. Hence,

$$\begin{aligned}\tau_1(f_L, \Delta_m) &= \sup_{x\in[-1,x_m)} \{f_L(x) - f_L(\xi_2(x))\} \vee \sup_{x\in[x_m,1]} \{f_L(\xi_1(x)) - f_L(x)\} \\ &= \sup_{x\in[-1,x_m)} \{\log(1+\beta_L(x))\} \vee \sup_{x\in[x_m,1]} \{-\log(1-\beta_L(x))\},\end{aligned} \tag{4.61}$$

where $\beta_L(x) \triangleq \frac{\Delta_m(x)}{x+\frac{L^2+1}{L^2-1}}$. If $m = 1$ we know that $x_1 > 0$ and $-x_1 < 0$ by (4.60), then

$$\tau_1(f_L, \Delta_m) = \sup_{x<x_m} \{f_L(x) - f_L(\xi_2(x) \wedge 1)\} \vee \sup_{x<x_m} \{f_L(-1) - f_L(x)\} \vee \sup_{x\geq x_m} \{f_L(\xi_1(x)) - f_L(x)\}.$$

Since $f_L(\xi_2(x)\wedge 1) \geq f_L(\xi_2(x))$, by the same argument, (4.61) remains a valid upper bound of $\tau_1(f_L, \Delta_1)$. Next we will show separately that the two terms in (4.61) both satisfy the desired upper bound.

For the first term in (4.61), note that

$$\beta_L(x) = \frac{\frac{1}{m}\sqrt{1-x^2}+\frac{1}{m^2}}{x+1+\frac{2}{L^2-1}} \leq \frac{1}{m^2}\frac{L\sqrt{1-x^2}+1}{(x+1)+\frac{2}{L^2}} = \frac{L^2}{m^2}\frac{\sqrt{1-x^2}+\frac{1}{L}}{L(x+1)+\frac{2}{L}}.$$

One can verify that $\frac{\sqrt{1-x^2}+\frac{1}{L}}{L(x+1)+\frac{2}{L}} \leq 1$ for any $x \in [-1, 1]$. Therefore,

$$\log(1+\beta_L(x)) \leq \log\left(1+\frac{L^2}{m^2}\right), \ \forall x \in [-1, 1],$$

and, consequently,

$$\sup_{x\in[-1,x_m)}\{\log(1+\beta_L(x))\}\le\log\left(\frac{2L^2}{m^2}\right),\ \forall m\le L. \tag{4.62}$$

For the second term in (4.61), it follows from the derivative of $\beta_L(x)$ that it is decreasing when $x>\frac{1-L^2}{1+L^2}$. From (4.60) we have $x_m>\frac{1-m^2}{1+m^2}$ and hence $x_m>\frac{1-L^2}{1+L^2}$ when $m\le L$. So the supremum is achieved exactly at the left end of $[x_m,1]$, that is:

$$\begin{aligned}\sup_{x\in[x_m,1]}\{-\log(1-\beta_L(x))\}&=-\log(1-\beta_L(x_m))\\&=\log\left(\frac{1+x_m}{2}L^2+\frac{1-x_m}{2}\right).\end{aligned}$$

From (4.60) we know that $x_m\ge-1$ and $x_m<-1+\frac{3.8}{m^2}$. Therefore $\frac{1-x_m}{2}\le1$ and $\frac{x_m+1}{2}<\frac{1.9}{m^2}$. For $m\le0.1L$, we have

$$\sup_{x\in[x_m,1]}\{-\log(1-\beta_L(x))\}\le\log\left(1+\frac{1.9m^2}{L^2}\right)\le\log\left(\frac{2m^2}{L^2}\right). \tag{4.63}$$

Plugging (4.62) and (4.63) into (4.61), we complete the proof of Lemma 4.10.

Next we prove (4.58). Recall that $x_L-\Delta_L(x_L)=-1$. By definition,

$$\tau_1(f_L,\Delta_L)\ge f_L(x_L-\Delta_L(x_L))-f_L(x_L)=\log\left(\frac{1+x_L}{2}L^2+\frac{1-x_L}{2}\right).$$

Using the close-form expression of x_L in (4.60) with $m=L$, we further obtain

$$\tau_1(f_L,\Delta_L)\ge\log\left(\frac{2L^2+\sqrt{-L^2+3L^4}}{2(L^2+1)}+\frac{2L^4-\sqrt{-L^2+3L^4}}{2(L^2+L^4)}\right)\ge1,$$

when $L\ge10$. □

CHAPTER 5

ESTIMATING THE UNSEEN

Estimating the support size of a distribution from data is a classical problem in statistics with widespread applications. For example, a major task for ecologists is to estimate the number of species [58] from field experiments; linguists are interested in estimating the vocabulary size of Shakespeare based on his complete works [102, 21, 103]; in population genetics it is of great interest to estimate the number of different alleles in a population [104]. Estimating the support size is equivalent to estimating the number of unseen symbols, which is particularly challenging when the sample size is relatively small compared to the total population size, since a significant portion of the population are never observed in the data. This chapter discusses two closely related problem: support size estimation and the distinct elements problem.

5.1 Definitions and previous work

We adopt the following statistical model [105, 106]. Let P be a discrete distribution over some countable alphabet. Without loss of generality, we assume the alphabet is $\mathbb{N}$ and denote $P = (p_1, p_2, \dots)$. Given n independent samples $X \triangleq (X_1, \dots, X_n)$ drawn from P, the goal is to estimate the support size

$$S = S(P) \triangleq \sum_i \mathbf{1}_{\{p_i > 0\}}. \tag{5.1}$$

Since support size is a symmetric function of the distribution, the histogram of samples (3.1) and the fingerprint (4.15) are both sufficient statistics for estimating $S(P)$.

It is clear that unless we impose further assumptions on the distribution P, it is impossible to estimate $S(P)$ within a given accuracy, for otherwise there can be arbitrarily many masses in the support of P that, with high

probability, are never sampled and the worst-case risk for estimating $S(P)$ is obviously infinite. To prevent the triviality, a conventional assumption [106] is to impose a lower bound on the non-zero probabilities. Therefore we restrict our attention to the parameter space $\mathcal{D}_k$, which consists of all probability distributions on $\mathbb{N}$ whose minimum non-zero mass is at least $\frac{1}{k}$; consequently $S(P) \leq k$ for any $P \in \mathcal{D}_k$. This is called the **Support Size** problem in this chapter. The decision-theoretic fundamental limit is given by the *minimax risk*:

$$R_S^*(k,n) \triangleq \inf_{\hat{S}} \sup_{P \in \mathcal{D}_k} \mathbb{E}(\hat{S} - S)^2, \tag{5.2}$$

where the loss function is the MSE and $\hat{S}$ is an integer-valued estimator measurable with respect to the samples $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P$.

We also investigate the *sample complexity* of the **Support Size** problem, which is defined as follows.

Definition 5.1. The sample complexity $n^*(k, \epsilon)$ is the minimal sample size n such that there exists an integer-valued estimator $\hat{S}$ based on n samples drawn independently from a distribution P such that $\mathbb{P}[|\hat{S} - S| \geq \epsilon k] \leq 0.1$ for any $P \in \mathcal{D}_k$.

Clearly, since $\hat{S} - S$ is an integer, the only interesting case is $\epsilon \geq \frac{1}{k}$, with $\epsilon = \frac{1}{k}$ corresponding to the exact estimation of the support size since $|\hat{S} - S| < 1$ is equivalent to $\hat{S} = S$. Furthermore, since $S(P)$ takes values in $[k]$, $n^*(k, \frac{1}{2}) = 0$ by definition.

Another common assumption on the support size estimation problem is that p_i has the special form $p_i = \frac{k_i}{k}$ with $k_i \in \mathbb{Z}_+$, which arises naturally from the **Distinct Elements** problem [107]:

> *Given n balls randomly drawn from an urn containing k colored balls, how to estimate the total number of distinct colors in the urn?*

Originating from ecology, numismatics, and linguistics, this problem is also known as the *species problem* in the statistics literature [108, 109]. Apart from the theoretical interests, it has a wide array of applications in various fields, such as estimating the number of species in a population of animals [58, 59], the number of dies used to mint an ancient coinage [110], and the

vocabulary size of an author [21]. In computer science, this problem frequently arises in large-scale databases, network monitoring, and data mining [106, 111, 107], where the objective is to estimate the types of database entries or IP addresses from limited observations, since it is typically impossible to have full access to the entire database or keep track of all the network traffic. The key challenge in the **Distinct Elements** problem is similar: given a small set of samples where most of the colors are not observed, how to accurately extrapolate the number of unseens? The **Distinct Elements** is a special case of the general support size estimation problem introduced above. We define the corresponding sample complexity as the smallest sample size needed to estimate the number of distinct colors with a prescribed accuracy and confidence level. A formal definition follows.

Definition 5.2. The sample complexity $n^*(k, \Delta)$ is the minimal sample size n such that there exists an integer-valued estimator $\hat{C}$ based on n balls drawn independently with replacements from the urn, such that $\mathbb{P}[|\hat{C} - C| \geq \Delta] \leq 0.1$ for any urn containing k balls with C different colors.[1]

5.1.1 Previous work on the **Support Size** problem

There is a vast amount of literature devoted to the support size estimation problem. In parametric settings, the data generating distribution is assumed to belong to certain parametric family such as uniform or Zipf [112, 102, 113] and traditional estimators, such as maximum likelihood estimator and minimum variance unbiased estimator, are frequently used [114, 115, 116, 21, 112, 104] – see the extensive surveys [109, 117]. When difficult to postulate or justify a suitable parametric assumption, various nonparametric approaches are adopted such as the Good-Turing estimator [59, 118] and variants due to Chao and Lee [119, 120], Jackknife estimator [105], empirical Bayes approach (e.g., Good-Toulmin estimator [121]), one-sided estimator [122]. Despite their practical popularity, little is known about the performance guarantee of these estimators, let alone their optimality. Next we discuss provable results assuming the independent sampling model.

For the naive plug-in estimator that counts the number of observed distinct

[1]Clearly, since $\hat{C} - C \in \mathbb{Z}$, we shall assume without loss of generality that $\Delta \in \mathbb{N}$, with $\Delta = 1$ corresponding to the exact estimation of the number of distinct elements.

symbols, it is easy to show that to estimate $S(P)$ within $\pm\epsilon k$ the minimal required number of samples is $\Theta(k \log \frac{1}{\epsilon})$, which scales logarithmically in $\frac{1}{\epsilon}$ but linearly in k, the same scaling for estimating the distribution P itself. Recently Valiant and Valiant [94] showed that the sample complexity is in fact sub-linear in k; however, the performance guarantee of the proposed estimators are still far from being optimal. Specifically, an estimator based on a linear program (LP) that is a modification of [21, Program 2] is proposed and shown to achieve $n^*(k, \epsilon) \lesssim \frac{k}{\epsilon^{2+\delta} \log k}$ for any arbitrary $\delta > 0$ [94, Corollary 11], which has subsequently been improved to $\frac{k}{\epsilon^2 \log k}$ in [64, Theorem 2, Fact 9]. The lower bound $n^*(k, \epsilon) \gtrsim \frac{k}{\log k}$ in [89, Corollary 9] is optimal in k but provides no dependence on ϵ. These results show that the optimal scaling in terms of k is $\frac{k}{\log k}$ but the dependence on the accuracy ϵ is $\frac{1}{\epsilon^2}$, which is even worse than the plug-in estimator. From Theorem 5.2 we see that the dependence on ϵ can be improved from polynomial to polylogarithmic $\log^2 \frac{1}{\epsilon}$, which turns out to be optimal. Furthermore, this can be attained by a linear estimator which is far more scalable than linear programming on massive datasets. Finally, we mention that a general framework of designing and analyzing linear estimators is given in [95] based on linear programming (as opposed to the approximation-theoretic approach in this chapter).

5.1.2 Previous work on the **Distinct Elements** problem

The **Distinct Elements** problem has been extensive studied by both statisticians and computer scientists.

Statistics literature The **Distinct Elements** problem is equivalent to estimating the number of species (or classes) in a finite population, which has been extensively studied in the statistics (see surveys [109, 123]) and the numismatics literature (see survey [110]). Motivated by various practical applications, a number of statistical models have been introduced for this problem, and the most popular four are (cf. [109, Figure 1]):

- *The multinomial model*: n samples are drawn uniformly at random with replacement;
- *The hypergeometric model*: n samples are drawn uniformly at random without replacement;

- *The Bernoulli model*: each individual is observed independently with some fixed probability, and thus the total number of samples is a binomial random variable;

- *The Poisson model*: the number of observed samples in each class is independent and Poisson distributed, and thus the total sample size is also a Poisson random variable.

These models are closely related: conditioned on the sample size, the Bernoulli model coincides with the hypergeometric one, and Poisson model coincides with the multinomial one; furthermore, hypergeometric model can simulate multinomial one and is hence more informative. The multinomial model is adopted as the main focus of this chapter and the sample complexity in Definition 5.2 refers to the number of samples with replacement. In the undersampling regime where the sample size is significantly smaller than the population size, all four models are approximately equivalent.

Under these models various estimators have been proposed such as unbiased estimators [124], Bayesian estimators [125], variants of Good-Turing estimators [120], etc. None of these methodologies, however, have a provable worst-case guarantee. Finally, we mention a closely related problem of estimating the number of connected components in a graph based on sampled induced subgraphs. In the special case where the underlying graph consists of disjoint cliques, the problem is exactly equivalent to the **Distinct Elements** problem [126].

Computer science literature The interests in the **Distinct Elements** problem also arise in the database literature, where various intuitive estimators [127, 128] have been proposed under simplifying assumptions such as uniformity, and few performance guarantees are available. More recent work in [107, 129] obtained the optimal sample complexity under the *multiplicative* error criterion, where the minimum sample size to estimate the number of distinct elements within a factor of α is shown to be $\Theta(k/\alpha^2)$. For this task, it turns out the least favorable scenario is to distinguish an urn with unitary color from one with *almost* unitary color, the impossibility of which implies large multiplicative error. However, the optimal estimator performs poorly compared with others on an urn with many distinct colors [107], the case where most estimators enjoy small multiplicative error. In

view of the limitation of multiplicative error, additive error is later considered by [106, 130]. To achieve an additive error of ck for a constant $c \in (0, \frac{1}{2})$, the result in [107] only implies an $\Omega(1/c)$ sample complexity lower bound, whereas a much stronger lower bound scales like $k^{1-O(\sqrt{\frac{\log\log k}{\log k}})}$ obtained in [106], which is almost linear. Determining the optimal sample complexity under additive error is the focus of the present chapter.

The **Distinct Elements** problem considered here is not to be confused with the formulation in the streaming literature, where the goal is to approximate the number of distinct elements in the observations with low space complexity, see, e.g., [131, 132]. There, the proposed algorithms aim to optimize the memory consumption, but still require a full pass of every ball in the urn. This is different from the setting in this chapter, where only random samples drawn from the urn are available.

To close this subsection, we mention the *Species extrapolation* problem whose recent resolution relies on results in this chapter. Given n independent samples drawn from an unknown distribution, the goal is to predict the number of hitherto unseen symbols that would be observed if m additional samples were collected from the same distribution. Originally formulated in [58] and further studied in [121, 21, 119], this problem reduces to support size estimation if $m = \infty$; in contrast, for finite m, this problem remains non-trivial even if no lower bound on the minimum non-zero probability is imposed on the underlying distribution, since very rare species will typically not appear in the new samples. The recent result [133] showed that the furthest range for accurate extrapolation is $m = o(n \log n)$ and obtained the minimax estimation error as a function of m, n for all distributions, where the lower bound is obtained via a reduction to support size estimation studied in this chapter.

5.2 Estimating the support size

5.2.1 Fundamental limits of the **Support Size** problem

Theorem 5.1. *For all $k, n \geq 2$,*

$$R_S^*(k,n) = k^2 \exp\left(-\Theta\left(\sqrt{\frac{n\log k}{k}} \vee \frac{n}{k} \vee 1\right)\right). \tag{5.3}$$

Furthermore, if $\frac{k}{\log k} \ll n \ll k\log k$, as $k \to \infty$,

$$k^2 \exp\left(-c_1\sqrt{\frac{n\log k}{k}}\right) \leq R_S^*(k,n) \leq k^2 \exp\left(-c_2\sqrt{\frac{n\log k}{k}}\right), \tag{5.4}$$

where $c_1 = \sqrt{2}e + o(1)$ and $c_2 = 1.579 + o(1)$.

To interpret the rate of convergence in (5.3), we consider three cases:

Simple regime $n \gtrsim k\log k$: we have $R_S^*(k,n) = k^2\exp(-\Theta(\frac{n}{k}))$ which can be achieved by the simple plug-in estimator

$$\hat{S}_{\text{seen}} \triangleq \sum_i \mathbf{1}_{\{N_i>0\}}, \tag{5.5}$$

that is, the number of observed symbols or the support size of the empirical distribution. Furthermore, if $\frac{n}{k\log k}$ exceeds a sufficiently large constant, all symbols are present in the data and $\hat{S}_{\text{seen}}$ is in fact exact with high probability, namely, $\mathbb{P}[\hat{S}_{\text{seen}} \neq S] \leq \mathbb{E}(\hat{S}_{\text{seen}} - S)^2 \to 0$. This can be understood as the classical coupon collector's problem (cf. e.g., [56]).

Non-trivial regime $\frac{k}{\log k} \ll n \ll k\log k$: in this case the samples are relatively scarce and the naive plug-in estimator grossly underestimate the true support size as many symbols are simply not observed. Nevertheless, accurate estimation is still possible and the optimal risk is given by $R_S^*(k,n) = k^2\exp(-\Theta(\sqrt{\frac{n\log k}{k}}))$. This can be achieved by a linear estimator based on the Chebyshev polynomial and its approximation-theoretic properties. Although more sophisticated than the plug-in estimator, this procedure can be evaluated in $O(n + \log^2 k)$ time.

Impossible regime $n \lesssim \frac{k}{\log k}$: any estimator suffers an error proportional to k in the worst case.

The next result characterizes the sample complexity within universal constant factors that are within a factor of six asymptotically.

Theorem 5.2. *Fix a constant $c_0 < \frac{1}{2}$. For all $\frac{1}{k} \le \epsilon \le c_0$,*

$$n^*(k, \epsilon) \asymp \frac{k}{\log k} \log^2 \frac{1}{\epsilon}. \tag{5.6}$$

Furthermore, if $\epsilon \to 0$ and $\epsilon = k^{-o(1)}$, as $k \to \infty$,

$$\frac{\tilde{c}_1 k}{\log k} \log^2 \frac{1}{\epsilon} \le n^*(k, \epsilon) \le \frac{\tilde{c}_2 k}{\log k} \log^2 \frac{1}{\epsilon}, \tag{5.7}$$

where $\tilde{c}_1 = \frac{1}{2e^2} + o(1)$ and $\tilde{c}_2 = \frac{1}{2.494} + o(1)$.

Compared to Theorem 5.1, the only difference is that here we are dealing with the zero-one loss $\mathbf{1}_{\{|S-\hat{S}| \ge \epsilon k\}}$ instead of the quadratic loss $(S-\hat{S})^2$. In the proof we shall obtain upper bound for the quadratic risk and lower bound for the zero-one loss, thereby proving both Theorem 5.1 and 5.2 simultaneously. Furthermore, the choice of 0.1 as the probability of error in the definition of the sample complexity is entirely arbitrary; replacing it by $1 - \delta$ for any constant $\delta \in (0, 1)$ only affect $n^*(k, \epsilon)$ up to constant factors.[2]

5.2.2 Optimal estimator via Chebyshev polynomials

In this section we prove the upper bound part of Theorem 5.1 and describe the rate-optimal support size estimator in the non-trivial regime. Following the same idea as in Section 3.1, we shall apply the Poissonization technique to simplify the analysis where the sample size is $\mathrm{Poi}(n)$ instead of a fixed number n and hence the sufficient statistics $N = (N_1, \ldots, N_k) \overset{\text{ind}}{\sim} \mathrm{Poi}(np_i)$. Analogous to (5.2), the minimax risk under the Poisson sampling is defined by

$$\tilde{R}^*(k, n) \triangleq \inf_{\hat{S}} \sup_{P \in \mathcal{D}_k} \mathbb{E}(\hat{S} - S)^2. \tag{5.8}$$

[2]Specifically, upgrading the confidence to $1 - \delta$ can be achieved by oversampling by merely a factor of $\log \frac{1}{\delta}$: Let $T = \log \frac{1}{\delta}$. With nT samples, divide them into T batches, apply the n-sample estimator to each batch and aggregate by taking the median. Then Hoeffding's inequality implies the desired confidence.

Due to the concentration of $\mathrm{Poi}(n)$ near its mean n, the minimax risk with fixed sample size is close to that under the Poisson sampling, as given by Theorem 3.1, which allows us to focus on the model using Poissonized sample size. In the next proposition, we first analyze the risk of the plug-in estimator $\hat{S}_{\mathrm{seen}}$, which yields the optimal upper bound of Theorem 5.1 in the regime of $n \gtrsim k \log k$. This is consistent with the coupon collection intuition.

Proposition 5.1. *For all $n, k \geq 1$,*

$$\sup_{P \in \mathcal{D}_k} \mathbb{E}(S(P) - \hat{S}_{\mathrm{seen}}(N))^2 \leq k^2 e^{-2n/k} + k e^{-n/k}, \tag{5.9}$$

where $N = (N_1, N_2, \dots)$ and $N_i \overset{ind}{\sim} \mathrm{Poi}(np_i)$.

Conversely, for P that is uniform over $[k]$, for any fixed $\delta \in (0,1)$, if $n \leq (1-\delta) k \log \frac{1}{\epsilon}$, then as $k \to \infty$,

$$\mathbb{P}[|S(P) - \hat{S}_{\mathrm{seen}}(N)| \leq \epsilon k] \leq e^{-\Omega(k^\delta)}. \tag{5.10}$$

In order to remedy the inaccuracy of the plug-in estimate $\hat{S}_{\mathrm{seen}}$ in the regime of $n \lesssim k \log k$, our proposed estimator adds a linear correction term:

$$\hat{S} = \hat{S}_{\mathrm{seen}} + \sum_{j \geq 1} u_j \Phi_j, \tag{5.11}$$

where the coefficients u_j's are to be specified. Equivalently, the estimator can be expressed in terms of the histogram as

$$\hat{S} = \sum_i g(N_i), \tag{5.12}$$

where $g : \mathbb{Z}_+ \to \mathbb{R}$ is defined as $g(j) = u_j + 1$ for $j \geq 1$ and $g(0) = 0$. Then the bias of $\hat{S}$ is

$$\mathbb{E}[\hat{S} - S] = \sum_{i: p_i > 0} e^{-np_i} \left(\sum_{j \geq 1} u_j \frac{(np_i)^j}{j!} - 1 \right) \triangleq \sum_{i: p_i > 0} e^{-np_i} P(p_i), \tag{5.13}$$

where $P(0) = -1$ by design. Therefore the bias of $\hat{S}$ is at most

$$S \max_{x \in [p_{\min}, 1]} |e^{-nx} P(x)|,$$

and the variance can be upper bounded by $2S\|g\|_\infty^2$ using the Efron-Stein inequality [90]. Next we choose the coefficients in order to balance the bias and variance. The construction is done using Chebyshev polynomials, which we first introduce.

Recall that the usual Chebyshev polynomial T_L (2.19). Note that T_L is bounded in magnitude by one over the interval $[-1, 1]$. The shifted and scaled Chebyshev polynomial over the interval $[l, r]$ is given by

$$P_L(x) = -\frac{T_L(\frac{2x-r-l}{r-l})}{T_L(\frac{-r-l}{r-l})} \triangleq \sum_{m=1}^{L} a_m x^m - 1, \tag{5.14}$$

the coefficients $a_1, \ldots, a_L$ can be obtained from those of the Chebyshev polynomial [34, 2.9.12] and the binomial expansion, or more directly,

$$a_j = \frac{P_L^{(j)}(0)}{j!} = -\left(\frac{2}{r-l}\right)^j \frac{1}{j!} \frac{T_L^{(j)}(-\frac{r+l}{r-l})}{T_L(-\frac{r+l}{r-l})}. \tag{5.15}$$

We now let

$$L \triangleq \lfloor c_0 \log k \rfloor, \quad r \triangleq \frac{c_1 \log k}{n}, \quad l \triangleq \frac{1}{k}, \tag{5.16}$$

where $c_0 < c_1$ are constants to be specified, and choose the coefficients of the estimator as

$$u_j = \begin{cases} \frac{a_j j!}{n^j}, & j = 1, \ldots, L, \\ 0, & \text{otherwise.} \end{cases} \tag{5.17}$$

The estimator $\hat{S}$ is defined according to (5.11).

We proceed to explain the reasoning behind the choice (5.17) and the role of the Chebyshev polynomial. The main intuition is that since $c_0 < c_1$, then with high probability, whenever $N_i \le L = \lfloor c_0 \log k \rfloor$ the corresponding mass must satisfy $p_i \le \frac{c_1 \log k}{n}$. That is, if $p_i > 0$ and $N_i \le L$ then $p_i \in [l, r]$ with high probability, and hence $P_L(p_i)$ is bounded by the sup-norm of P_L over the interval $[l, r]$, which controls the bias in view of (5.13). In view of the extremal property of Chebyshev polynomials [34, Ex. 2.13.14], (5.14) is the unique degree-L polynomial that passes through the point $(0, -1)$ and deviates the least from zero over the interval $[l, r]$. This explains the coefficients (5.12) which are chosen to minimize the bias. The degree of the polynomial is only logarithmic so that the variance is small.

The next proposition gives an upper bound of the quadratic risk of our

estimator (5.12).

Proposition 5.2. *Assume the Poissonized sampling model where the histograms are distributed as* $N = (N_1, N_2, \dots) \overset{ind}{\sim} \mathrm{Poi}(np_i)$. *Let* $c_0 = 0.558$ *and* $c_1 = 0.5$. *As* $\delta \triangleq \frac{n}{k \log k} \to 0$ *and* $k \to \infty$, *the bias and variance of* $\hat{S}$ *are upper bounded by*

$$|\mathbb{E}(\hat{S} - S)| \le 2S(1 + o_k(1)) \exp\left(-(1 + o_\delta(1))\sqrt{\kappa \frac{n \log k}{k}}\right),$$
$$\mathsf{var}[\hat{S}] \le O\left(Sk^c\right),$$

for some absolute constant $c < 1$, *and consequently,*

$$\sup_{P \in \mathcal{D}_k} \mathbb{E}(\hat{S}(N) - S(P))^2 \le 4k^2(1 + o_k(1)) \exp\left(-(2 + o_\delta(1))\sqrt{\kappa \frac{n \log k}{k}}\right), \tag{5.18}$$

where $\kappa = 2.494$.

The minimax upper bounds in Theorems 5.1 and 5.2 follow from combining Propositions 5.1 and 5.2.

Proof of upper bound of Theorem 5.1. Combining Theorem 3.1 and Proposition 5.2 yields the upper bound part of (5.4), which also implies the upper bound of (5.3) when $n \lesssim k \log k$. The upper bound part of (5.3) when $n \gtrsim k \log k$ follows from Proposition 5.1. □

Proof of upper bound of Theorem 5.2. By the Markov inequality,

$$R_S^*(k, n) \le 0.1k^2\epsilon^2 \Rightarrow n^*(k, \epsilon) \le n. \tag{5.19}$$

Therefore our upper bound is

$$n^*(k, \epsilon) \le \inf\{n : R_S^*(k, n) \le 0.1k^2\epsilon^2\}.$$

By the upper bound of $R_S^*(k, n)$ in (5.18), we obtain that

$$n^*(k, \epsilon) \le \frac{1 + o_{\delta'}(1) + o_\epsilon(1) + o_k(1)}{\kappa} \frac{k}{\log k} \log^2 \frac{1}{\epsilon},$$

as $\delta' \triangleq \frac{\log(1/\epsilon)}{\log k} \triangleq 0$, $\epsilon \to 0$, and $k \to \infty$. Consequently, we obtain the upper

bound part of (5.6) when $\frac{1}{k^c} \le \epsilon \le c_0$ for the fixed constant $c_0 < 1/2$, where c is some small constant.

The upper bound part of Theorem 5.2 when $\frac{1}{k} \le \epsilon \le \frac{1}{k^c}$ follows from the monotonicity of $\epsilon \mapsto n^*(k, \epsilon)$ that

$$n^*(k, \epsilon) \le n^*(k, 1/k) \le 3k \log k \asymp \frac{k}{\log k} \log^2 \frac{1}{\epsilon},$$

where the middle inequality follows from Proposition 5.1 and (5.19). □

Note that the optimal estimator (5.12) relies on the choice of parameters in (5.16), which, in turn, depends on the knowledge of $1/k$, the lower bound on the minimum non-zero probability $p_{\min}$. While k is readily obtainable in certain applications where the samples are uniformly drawn from a database or corpus of known size (see [111, 21] as well as the empirical results in Section 5.2.5), it is desirable to construct estimators that are agnostic to $p_{\min}$ and retains the same optimality guarantee. To this end, we provide the following alternative choice of parameters. Let $\tilde{S}$ be the linear estimator defined using the same coefficients in (5.17), with the approximation interval $[l, r]$ and the degree L in (5.16) replaced by

$$l = \frac{c_1}{c_0^2} \frac{\log^2(1/\epsilon)}{n \log n}, \quad r = \frac{c_1 \log n}{n}, \quad L = \lfloor c_0 \log n \rfloor. \tag{5.20}$$

Here ϵ is the desired accuracy and the constants c_0, c_1 are the same as in Proposition 5.2. Following the same analysis as in the proof of Proposition 5.2, the above choice of parameters leads to the following upper bound of the quadratic risk.

Proposition 5.3. *Let c_0, c_1, c be the same constants as Proposition 5.2. There exist constants C, C' such that, if $\epsilon > n^{-C}$, then*

$$\mathbb{E}(\tilde{S} - S)^2 \le C'(S^2 \epsilon^{2(1-\sqrt{\alpha})} + S n^c),$$

where $\alpha = \max\left(1 - \frac{c_0^2}{c_1} \frac{n \log n}{k \log^2(1/\epsilon)}, 0\right)$.

Therefore, whenever the sample size satisfies $n \ge (\frac{c_1}{c_0^2} + o_k(1)) \frac{k}{\log k} \log^2 \frac{1}{\epsilon}$ and $n \le (\epsilon^2 k)^{\frac{1}{c}}$, the upper bound is at most $O((\epsilon k)^2)$, recovering the optimal risk bound in Proposition 5.2. The new result here is that even when n is not that large the risk degrades gracefully.

We finish this subsection with a few remarks.

Remark 5.1. Combined with standard concentration inequalities, the mean-square error bound in Proposition 5.2 can be easily converted to a high-probability bound. In the regime of $n \lesssim k \log k$, for any distribution $P \in \mathcal{D}_k$, the bias of our estimate $\hat{S}$ is at most the uniform approximation error (see (5.40)):

$$|\mathbb{E}[\hat{S}] - S| \le S \exp\left(-\Theta\left(\sqrt{\frac{n \log k}{k}}\right)\right).$$

The standard deviation is significantly smaller than the bias. Indeed, the coefficients of the linear estimator (5.12) is uniformly bounded by $\|g\|_\infty^2 \le k^c$ for some absolute constant $c < 1$ (see (5.53) as well as Figure 5.1 for numerical results). Therefore, by Hoeffding's inequality, we have the following concentration bound:

$$\mathbb{P}[|\hat{S} - \mathbb{E}[\hat{S}]| \ge tk] \le 2 \exp\left(-\frac{t^2 k}{2\|g\|_\infty^2}\right) = \exp\left(-t^2 k^{\Omega(1)}\right).$$

Remark 5.2. The estimator (5.12) belong to the family of *linear estimators*:

$$\hat{S} = \sum_i f(N_i) = \sum_{j \ge 1} f(j) \Phi_j, \tag{5.21}$$

which is a linear combination of fingerprints Φ_j's defined in (4.15).

Other notable examples of linear estimators include:

- Plug-in estimator (5.5): $\hat{S}_{\text{seen}} = \Phi_1 + \Phi_2 + \dots$.
- Good-Toulmin estimator [121]: for some $t > 0$,

$$\hat{S}_{\text{GT}} = \hat{S}_{\text{seen}} + t\Phi_1 - t^2\Phi_2 + t^3\Phi_3 - t^4\Phi_4 + \dots \tag{5.22}$$

- Efron-Thisted estimator [21]: for some $t > 0$ and $J \in \mathbb{N}$,

$$\hat{S}_{\text{ET}} = \hat{S}_{\text{seen}} + \sum_{j=1}^{J} (-1)^{j+1} t^j b_j \Phi_j, \tag{5.23}$$

  where $b_j = \mathbb{P}[\text{binomial}(J, 1/(t+1)) \ge j]$.

By definition, our estimator (5.12) can be written as

$$\hat{S} = \sum_{j=1}^{L} g(j)\Phi_j + \sum_{j>L} \Phi_j. \tag{5.24}$$

By (5.14), P_L is also a polynomial of degree L, which is oscillating and results in coefficients with alternating signs (see Figure 5.1). Interestingly,

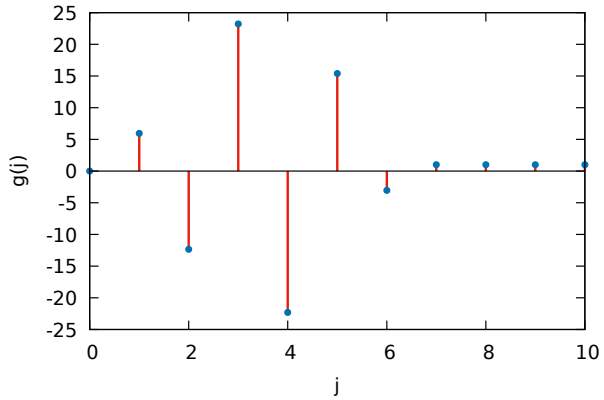


Figure 5.1: Coefficients of estimator $g(j)$ in (5.12) with $c_0 = 0.45, c_1 = 0.5$, $k = 10^6$ and $n = 2 \times 10^5$.

this behavior, although counterintuitive, coincides with many classical estimators, such as (5.22) and (5.23). The occurrence of negative coefficients can be explained as follows. Note that the rationale of linear estimator is to form a linear prediction the number of unseen Φ_0 using the observed fingerprints $\Phi_1, \Phi_2, \ldots$; this is possible because the fingerprints are correlated. Indeed, since the sum of all fingerprints coincides with the support size, i.e., $\sum_{j\geq 0} \Phi_j = S$, for each $j \geq 1$, the random variable Φ_j is negatively correlated with Φ_0 and hence some of the coefficients in the linear estimator are negative.

Remark 5.3 (Time complexity)**.** The evaluation of the estimator (5.21) consists of three parts:

1. Construction of the estimator: $O(L^2) = O(\log^2 k)$, which amounts to computing the coefficients $g(j)$ per (5.15);

2. Computing the histograms N_i and fingerprints Φ_j: $O(n)$;

3. Evaluating the linear combination: $O(n \wedge k)$, since the number of nonzero terms in the second summation of (5.21) is at most $n \wedge k$.

Therefore the total time complexity is $O(n + \log^2 k)$.

Remark 5.4. The technique of polynomial approximation has been previously used for estimating non-smooth functions (L_q-norms) in Gaussian models [65, 17, 39] and more recently for estimating information quantities (entropy and power sums) on large discrete alphabets [55, 69]. The design principle is to approximate the non-smooth function on a given interval using algebraic or trigonometric polynomials for which unbiased estimators exist; the degree is chosen to balance the bias (approximation error) and the variance (stochastic error). Note that in general uniform approximation by polynomials is only possible on a compact interval. Therefore, in many situations, the construction of the estimator is a two-stage procedure involving *sample splitting*: First, use half of the sample to test whether the corresponding parameter lies in the given interval; second, use the remaining samples to construct an unbiased estimator for the approximating polynomial if the parameter belongs to the interval or apply plug-in estimators otherwise (see, e.g., [55, 69] and [39, Section 5]).

While the benefit of sample splitting is to make the analysis tractable by capitalizing on the independence of the two subsamples, it also sacrifices the statistical accuracy since half of the samples are wasted. In this chapter, to estimate the support size, we forgo the sample splitting approach and directly design a linear estimator. Instead of using a polynomial as a proxy for the original function and then constructing its unbiased estimator, the best polynomial approximation of the indicator function arises as a natural step in controlling the bias (see (5.13)).

5.2.3 Suboptimality of the Good-Turing and Chao-1 estimators

In this subsection we show that unless the sample size n far exceeds k, the reciprocal of the minimal probability, both the Good-Turing estimator and its variant (the Chao-1 estimator) lead to non-vanishing normalized mean-square error for estimating the support size. Therefore, neither of them can operate in the sublinear regime.

The intuition is that although both estimators work well for uniform distributions, as soon as the probability masses take two or more values, they

become biased. To this end, consider a distribution p_n with n symbols with probability $\frac{1}{2n}$ and $2n$ symbols with probability $\frac{1}{4n}$. This distribution has $p_{\min} = \frac{1}{4n}$ and support size $S = 3n$. Given n samples drawn i.i.d. from p_n (similar arguments continue to hold under the Poisson sampling model), the expected values of the first few fingerprints are as follows:

$$\begin{aligned}
\mathbb{E}[\Phi_0] &= n\left(1-\frac{1}{2n}\right)^n + 2n\left(1-\frac{1}{4n}\right)^n \\
&= n\left(e^{-1/2} + 2e^{-1/4} + o(1)\right), \\
\mathbb{E}[\Phi_1] &= n\binom{n}{1}\frac{1}{2n}\left(1-\frac{1}{2n}\right)^{n-1} + 2n\binom{n}{1}\frac{1}{4n}\left(1-\frac{1}{4n}\right)^{n-1} \\
&= n\left(\frac{1}{2}e^{-1/2} + \frac{1}{2}e^{-1/4} + o(1)\right), \\
\mathbb{E}[\Phi_2] &= n\binom{n}{2}\frac{1}{(2n)^2}\left(1-\frac{1}{2n}\right)^{n-1} + 2n\binom{n}{2}\frac{1}{(4n)^2}\left(1-\frac{1}{4n}\right)^{n-2} \\
&= n\left(\frac{1}{8}e^{-1/2} + \frac{1}{16}e^{-1/4} + o(1)\right).
\end{aligned}$$

By the McDiarmid's inequality, $\Phi_j = \sum_i \mathbf{1}_{\{N_i=j\}}$, for $j = 0, 1, 2$, concentrates on the respective mean: $\Phi_j = \mathbb{E}[\Phi_j] + O_P(\sqrt{n})$. Therefore,

$$\begin{aligned}
\hat{S}_{\text{seen}} &= S - \Phi_0 = (3 - e^{-1/2} - 2e^{-1/4} + o_P(1))n, \\
\hat{S}_{\text{G-T}} &= \frac{\hat{S}_{\text{seen}}}{1-\Phi_1/n} = n\left(\frac{3 - e^{-1/2} - 2e^{-1/4}}{1 - \frac{1}{2}e^{-1/2} - \frac{1}{2}e^{-1/4}} + o_P(1)\right) \\
&\approx (2.72 + o_P(1))n, \\
\hat{S}_{\text{Chao1}} &= \hat{S}_{\text{seen}} + \frac{\Phi_1^2}{2\Phi_2} = n\left(3 - e^{-1/2} - 2e^{-1/4} + \frac{(e^{-1/2} + e^{-1/4})^2}{e^{-1/2} + \frac{1}{2}e^{-1/4}} + o_P(1)\right) \\
&\approx (2.76 + o_P(1))n,
\end{aligned}$$

as compared to the true support size $S = 3n$.

5.2.4 Correlation decay between fingerprints

Recall that the fingerprints are defined by $\Phi_j = \sum_i \mathbf{1}_{\{N_i=j\}}$, where N_i denotes the histogram of samples. The estimation of support size is equivalent to estimating the unseen, namely, Φ_0. In (5.17), we let the coefficients $u_j = 0$ when $j > L$, which is because higher-order fingerprints are *almost uncorrelated*

with Φ_0. In fact, the correlation between Φ_0 and Φ_j decays exponentially. Under the Poisson model, $N_i \overset{\text{ind}}{\sim} \text{Poi}(np_i)$. Then

$$\begin{aligned}\mathsf{cov}(\Phi_j, \Phi_{j'}) &= -\sum_i \mathbb{P}[N_i = j]\mathbb{P}[N_i = j'], \quad j \neq j', \\ \mathsf{var}[\Phi_j] &= \sum_i \mathbb{P}[N_i = j](1 - \mathbb{P}[N_i = j]).\end{aligned}$$

The correlation coefficient between Φ_0 and Φ_j follows immediately:

$$\begin{aligned}|\rho(\Phi_0, \Phi_j)| &= \sum_i \frac{\mathbb{P}[N_i = 0]\mathbb{P}[N_i = j]}{\sqrt{\sum_l \mathbb{P}[N_l = 0](1 - \mathbb{P}[N_l = 0]) \sum_l \mathbb{P}[N_l = j](1 - \mathbb{P}[N_l = j])}} \\ &\leq \sum_i \frac{\mathbb{P}[N_i = 0]\mathbb{P}[N_i = j]}{\sqrt{\mathbb{P}[N_i = 0](1 - \mathbb{P}[N_i = 0])\mathbb{P}[N_i = j](1 - \mathbb{P}[N_i = j])}} \\ &= \sum_i \sqrt{\frac{\mathbb{P}[N_i = 0]}{1 - \mathbb{P}[N_i = 0]} \frac{\mathbb{P}[N_i = j]}{1 - \mathbb{P}[N_i = j]}} = \sum_i \sqrt{\frac{e^{-\lambda_i}}{1 - e^{-\lambda_i}} \frac{\frac{e^{-\lambda_i}\lambda_i^j}{j!}}{1 - \frac{e^{-\lambda_i}\lambda_i^j}{j!}}},\end{aligned} \tag{5.25}$$

where $\lambda_i = np_i$. Note that $\max_{x>0} \frac{e^{-x}x^j}{j!} = \frac{e^{-j}j^j}{j!} \to 0$ as $j \to \infty$. Therefore, for any $x > 0$,

$$\frac{e^{-x}}{1 - e^{-x}} \frac{\frac{e^{-x}x^j}{j!}}{1 - \frac{e^{-x}x^j}{j!}} = \frac{1}{j!} \frac{e^{-2x}x^j}{1 - e^{-x}}(1 + o_j(1)), \tag{5.26}$$

where $o_j(1)$ is uniform as $j \to \infty$. Taking derivative, the function $x \mapsto \frac{e^{-2x}x^j}{1-e^{-x}}$ on $x > 0$ is increasing if and only if $x + e^x(j - 2x) - j > 0$, and the maximum is attained at $x = j/2 + o_j(1)$. Therefore, applying $j! > (j/e)^j$,

$$\frac{1}{j!} \frac{e^{-2x}x^j}{1 - e^{-x}} \leq (1 + o_j(1))2^{-j}. \tag{5.27}$$

Combining (5.25) – (5.27), we conclude that

$$|\rho(\Phi_0, \Phi_j)| \leq k2^{-j/2}(1 + o_j(1)).$$

5.2.5 Experiments

We evaluate the performance of our estimator on both synthetic and real datasets in comparison with popular existing procedures.[3] In the experiments we choose the constants $c_0 = 0.45, c_1 = 0.5$ in (5.16), instead of $c_0 = 0.558$ which is optimized to yield the best rate of convergence in Proposition 5.2 under the i.i.d. sample model. The reason for such a choice is that in the real-data experiments the samples are not necessarily generated independently and dependency leads to a higher variance. By choosing a smaller c_0, the Chebyshev polynomials have a slightly smaller degree, which results in smaller variance and more robustness to model mismatch. Each experiment is averaged over 50 independent trials and the standard deviations are shown as error bars.

Synthetic data We consider data independently sampled from the following distributions:

- the uniform distribution with $p_i = \frac{1}{k}$;
- Zipf distributions with $p_i \propto i^{-\alpha}$ and α being either 1 or 0.5;
- an even mixture of geometric distribution and Zipf distribution where for the first half of the alphabet $p_i \propto 1/i$ and for the second half $p_{i+k/2} \propto (1 - \frac{2}{k})^{i-1}$, $1 \leq i \leq \frac{k}{2}$.

The alphabet size k varies in each distribution so that the minimum nonzero mass is roughly 10^{-6}. Accordingly, a degree-6 Chebyshev polynomial is applied. Therefore, according to (5.24), we apply the polynomial estimator g to symbols appearing at most six times and the plug-in estimator otherwise. We compare our results with the Good-Turing estimator [59], the Chao 1 estimator [119, 134], the two estimators proposed by Chao and Lee [120], and the linear programming approach proposed by Valiant and Valiant [64]. Here the Good-Turing estimator refers to first estimate the total probability of seen symbols (sample coverage) by $\hat{C} = 1 - \frac{\Phi_1}{n}$ then estimate the support size by $\hat{S}_{\mathrm{G-T}} = \hat{S}_{\mathrm{seen}}/\hat{C}$; the Chao 1 estimator refers to the bias-corrected form $\hat{S}_{\mathrm{Chao1}} = \hat{S}_{\mathrm{seen}} + \frac{\Phi_1(\Phi_1-1)}{2(\Phi_2+1)}$. The plug-in estimator simply counts the number

[3] The implementation of our estimator is available at `https://github.com/Albuso0/support`.

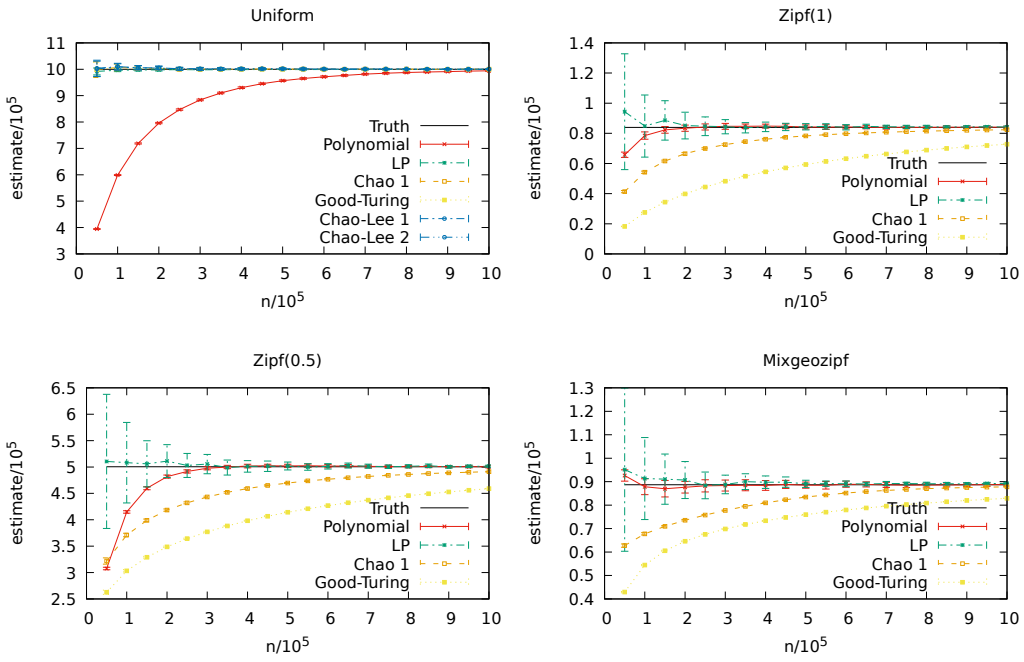


Figure 5.2: Performance comparison under four data-generating distributions.

of distinct elements observed, which is always outperformed by the Good-Turing estimator in our experiments and hence omitted in the comparison.

The results are shown in Figure 5.2. Good-Turing's estimate on sample coverage performs remarkably well in the special case of uniform distributions. This has been noticed and analyzed in [120, 113]. Chao-Lee's estimators are based on Good-Turing's estimate with further correction terms for non-uniform distributions. However, with limited number of samples, if no symbol appears more than once, the sample coverage estimate $\hat{C}$ is zero and consequently the Good-Turing estimator and Chao-Lee estimators are not even well defined. For Zipf and mixture distributions, the output of Chao-Lee's estimators is highly unstable and thus is omitted from the plots; the convergence rates of Good-Turing estimator and Chao 1 estimator are much slower than our estimator and the LP estimator, partly because they only use the information of how many symbols occurred exactly once and twice, namely the first two fingerprints Φ_1 and Φ_2, as opposed to the full spectrum of fingerprints $\{\Phi_j\}_{j\geq 1}$, and they suffer provably large bias under non-uniform distributions as simple as mixtures of two uniform distributions (see Section 5.2.3); the linear programming approach has similar convergence

rate to ours but suffers from large variance when samples are scarce.

Real data Next we evaluate our estimator by a real data experiment based on the text of *Hamlet*, which contains about $32,000$ words in total consisting of about $4,800$ distinct words. Here and below the definition of "distinct word" is any distinguishable arrangement of letters that are delimited by spaces, insensitive to cases, with punctuations removed. We randomly sample the text with replacement and generate the fingerprints for estimation. The minimum non-zero mass is naturally the reciprocal of the total number of words, $\frac{1}{32,000}$. In this experiment we use the degree-4 Chebyshev polynomial. We also compare our estimator with the one in [64]. The results are plotted in Figure 5.3, which shows that the estimator in [64] has similar convergence

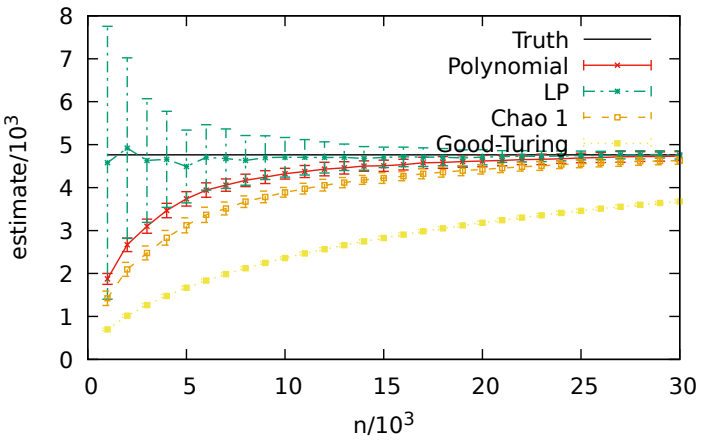


Figure 5.3: Comparison of various estimates of the total number of distinct words in *Hamlet*.

rate to ours; however, the variance is again much larger and the computational cost of linear programming is significantly higher than linear estimators, which amounts to computing linear combinations with pre-determined coefficients.

Next we conduct a larger-scale experiment using the *New York Times Corpus* from the years 1987 – 2007.[4] This corpus has a total of 25,020,626 paragraphs consisting of 996,640,544 words with 2,047,985 distinct words. We randomly sample 1% – 50% out of the all paragraphs with replacements and feed the fingerprint to our estimator. The minimum non-zero mass is

[4]Dataset available at `https://catalog.ldc.upenn.edu/LDC2008T19`.

also the reciprocal of the total number of words, $1/10^9$, and thus the degree-9 Chebyshev polynomial is applied. Using only 20% samples our estimator achieves a relative error of about 10%, which is a systematic error due to the model mismatch: the sampling here is paragraph by paragraph rather than word by word, which induces dependence across samples as opposed to the i.i.d. sampling model for which the estimator is designed; in comparison, the LP estimator[5] suffers a larger bias from this model mismatch. Furthermore, the proposed linear estimator is significantly faster than linear programming based methods: given the sampled data, the curve in Figure 5.4 corresponding to the LP estimator takes over 5 hours to compute; in contrast, the proposed linear estimator takes only 2 seconds on the same computer, which clearly demonstrate its computational advantage even if one takes into account the fact that our implementation is based on C++ while the LP estimator is in MATLAB.

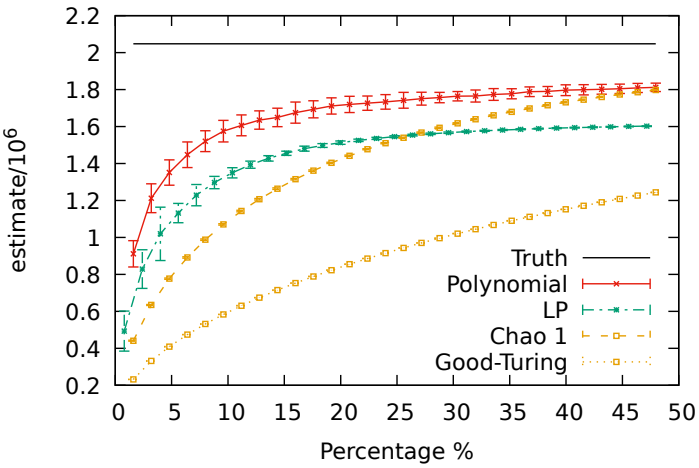


Figure 5.4: Performance comparison using *New York Times Corpus*.

Finally, we perform the classical experiment of "how many words did Shakespeare know". We feed the fingerprint of the entire Shakespearean canon (see [21, Table 1]), which contains 31,534 word types, to our estimator. We choose the minimum non-zero mass to be the reciprocal of the total number of English words, which, according to known estimates, is between 600,000 [136] to 1,000,000 [137], and obtain an estimate of 63,148 to 73,460

[5]In this large-scale experiment, the original MATLAB code of the linear programming estimator given in [64] is extremely slow; the result in Figure 5.4 is obtained using an optimized version provided by the author [135].

for Shakespeare's vocabulary size, as compared to 66,534 obtained by Efron-Thisted [21]. Using the alternative choice of parameters that are agnostic to k in Proposition 5.3, by setting the desired accuracy to be 0.05 and 0.1, we obtain an estimate of 62,355 to 72,454.

5.2.6 Minimax lower bound

The lower bound argument follows the idea in [17, 39, 55] and relies on the generalized Le Cam's method involving two composite hypothesis, also known as the method of fuzzy hypotheses [32]. The main idea is similar to Section 4.3. Specifically, suppose the following (composite) hypothesis testing problem,

$$H_0 : S(P) \le s, P \in \mathcal{D}_k \quad \text{versus} \quad H_1 : S(P) \ge s + \delta, P \in \mathcal{D}_k,$$

cannot be tested with vanishing probability of error on the basis of n samples, then the sample complexity of estimating $S(P)$ within δ with high probability must exceed n. In particular, the impossibility to test the above composite hypotheses is shown by constructing two priors (i.e., two random probability vectors) so that the induced distribution of the samples are close in total variation. Next we elaborate the main ingredients of Le Cam's method:

- construction of the two priors;
- separation between functional values;
- bound on the total variation.

Let $\lambda > 1$. Given unit-mean random variables U and U' that take values in $\{0\} \cup [1, \lambda]$, define the following random vectors

$$\mathsf{P} = \frac{1}{k}(U_1, \ldots, U_k), \quad \mathsf{P}' = \frac{1}{k}(U'_1, \ldots, U'_k), \tag{5.28}$$

where U_i and U'_i are i.i.d. copies of U and U', respectively. Although P and P' need not be probability distributions, as long as the standard deviations of U and U' are not too big, the law of large numbers ensures that with high probability P and P' lie in a small neighborhood near the probability simplex, which we refer as the set of *approximate* probability distributions.

Furthermore, the minimum non-zeros in P and P' are at least $\frac{1}{k}$. It can be shown that the minimax risk over approximate probability distributions is close to that over the original parameter space $\mathcal{D}_k$ of probability distributions. This allows us to use P and P' as priors and apply Le Cam's method. Note that both $S(\mathsf{P})$ and $S(\mathsf{P}')$ are binomially distributed, which, with high probability, differ by the difference in their mean values:

$$\mathbb{E}[S(\mathsf{P})] - \mathbb{E}[S(\mathsf{P}')] = k(\mathbb{P}[U > 0] - \mathbb{P}[U' > 0]) = k(\mathbb{P}[U' = 0] - \mathbb{P}[U = 0]).$$

If we can establish the impossibility of testing whether data are generated from P or P', the resulting lower bound is proportional to $k(\mathbb{P}[U' = 0] - \mathbb{P}[U = 0])$.

To simplify the argument we apply the Poissonization technique where the sample size is a $\mathrm{Poi}(n)$ random variable instead of a fixed number n. This provably does not change the statistical nature of the problem due to the concentration of $\mathrm{Poi}(n)$ around its mean n. Under Poisson sampling, the histograms (3.1) still constitute a sufficient statistic, which are distributed as $N_i \overset{\text{ind}}{\sim} \mathrm{Poi}(np_i)$, as opposed to multinomial distribution in the fixed-sample-size model. Therefore through the i.i.d. construction in (5.28), $N_i \overset{\text{i.i.d.}}{\sim} \mathbb{E}[\mathrm{Poi}(\frac{n}{k}U)]$ or $\mathbb{E}[\mathrm{Poi}(\frac{n}{k}U')]$. Then Le Cam's lemma is applicable if $\mathsf{TV}(\mathbb{E}[\mathrm{Poi}(\frac{n}{k}U)]^{\otimes k}, \mathbb{E}[\mathrm{Poi}(\frac{n}{k}U')]^{\otimes k})$ is strictly bounded away from one, for which it suffices to show

$$\mathsf{TV}(\mathbb{E}[\mathrm{Poi}(nU/k)], \mathbb{E}[\mathrm{Poi}(nU'/k)]) \le \frac{c}{k}, \tag{5.29}$$

for some constant $c < 1$.

The above construction provides a recipe for the lower bound. To optimize the ingredients it boils down to the following optimization problem (over one-dimensional probability distributions): Construct two priors U, U' with unit mean that maximize the difference $\mathbb{P}[U' = 0] - \mathbb{P}[U = 0]$ subject to the total variation distance constraint (5.29), which, in turn, can be guaranteed by *moment matching*, i.e., ensuring U and U' have identical first L moments for some large L, and the L_∞-norms U, U' are not too large. To summarize, our

lower bound entails solving the following optimization problem:

$$\begin{aligned} &\sup \mathbb{P}[U'=0]-\mathbb{P}[U=0], \\ &\text{s.t. } \mathbb{E}[U]=\mathbb{E}[U']=1, \\ &\qquad \mathbb{E}[U^j]=\mathbb{E}[U'^j], \quad j=1,\ldots,L \\ &\qquad U,U'\in\{0\}\cup[1,\lambda]. \end{aligned} \tag{5.30}$$

The final lower bound is obtained from (5.30) by choosing $L \asymp \log k$ and $\lambda \asymp \frac{k\log k}{n}$.

In order to evaluate the infinite-dimensional linear programming problem (5.30), we consider its dual program. It is well known that the problem of best polynomial and moment matching are dual to each other; however, unlike the standard moment matching problem which impose the equality of moments, the extra constraint in (5.30) is that the values of the first moment must equal to one. Therefore its dual is no longer the best polynomial approximation problem. Nevertheless, for the specific problem (5.30) which deals with the indicator function $x \mapsto \mathbf{1}_{\{x=0\}}$, via a change of variable we show in Section 5.2.7 that (5.30) coincides exactly with the best uniform approximation error of the function $x \mapsto \frac{1}{x}$ over the interval $[1,\lambda]$ by degree-$(L-1)$ polynomials:

$$\inf_{p\in\mathcal{P}_{L-1}} \sup_{x\in[1,\lambda]} \left|\frac{1}{x}-p(x)\right|,$$

where $\mathcal{P}_{L-1}$ denotes the set of polynomials of degree at most $L-1$. This best polynomial approximation problem has been well-studied, cf. [34, 138]; in particular, the exact formula for the best polynomial that approximates $x \mapsto \frac{1}{x}$ and the optimal approximation error have been obtained in [34, Sec. 2.11.1].

Applying the procedure described above, we obtain the following sample complexity lower bound.

Proposition 5.4. *Let* $\delta \triangleq \frac{\log\frac{1}{\epsilon}}{\log k}$ *and* $\tau \triangleq \frac{\sqrt{\log k}/k^{1/4}}{1-2\epsilon}$. *As* $k\to\infty$, $\delta\to 0$ *and* $\tau\to 0$,

$$n^*(k,\epsilon) \geq (1-o_\delta(1)-o_k(1)-o_\tau(1))\frac{k}{2e^2\log k}\log^2\frac{1}{2\epsilon}. \tag{5.31}$$

Consequently, if $\frac{1}{k^c} \leq \epsilon \leq \frac{1}{2}-c'\frac{\sqrt{\log k}}{k^{1/4}}$ *for some constants* c, c', *then* $n^*(k,\epsilon) \gtrsim \frac{k}{\log k}\log^2\frac{1}{2\epsilon}$.

The lower bounds announced in Theorems 5.1 and 5.2 follow from Proposition 5.4 combined with a simple two-point argument.

Proof of lower bound of Theorem 5.2. The lower bound part of (5.7) follows from Proposition 5.4. Consequently, we obtain the lower bound part of (5.6) for $\frac{1}{k^c} \le \epsilon \le c_0$ for the fixed constant $c_0 < 1/2$, where c is some small constant.

The lower bound part of (5.6) for $\frac{1}{k} \le \epsilon \le \frac{1}{k^c}$ simply follows from the fact that $\epsilon \mapsto n^*(k,\epsilon)$ is decreasing:

$$n^*(k,\epsilon) \ge n^*(k,1/k^c) \gtrsim k\log k \asymp \frac{k}{\log k}\log^2\frac{1}{\epsilon}. \qquad \square$$

Proof of lower bound of Theorem 5.1. By the Markov inequality,

$$n^*(k,\epsilon) > n \Rightarrow R_S^*(k,n) > 0.1k^2\epsilon^2.$$

Therefore, our lower bound is

$$R_S^*(k,n) \ge \sup\{0.1k^2\epsilon^2 : n^*(k,\epsilon) > n\} = 0.1k^2\epsilon_*^2,$$

where $\epsilon_* \triangleq \{\epsilon : n^*(k,\epsilon) > n\}$. By the lower bound of $n^*(k,\epsilon)$ in (5.31), we obtain that

$$\epsilon_* \ge \exp\left(-\left(\sqrt{2}e + o_\delta(1) + o_{\delta'}(1) + o_k(1)\right)\sqrt{\frac{n\log k}{k}}\right),$$

as $\delta \triangleq \frac{n}{k\log k} \to 0$, $\delta' \triangleq \frac{k}{n\log k} \to 0$, and $k \to \infty$. Then we conclude the lower bound part of (5.4), which implies the lower bound part of (5.3) when $n \lesssim k\log k$.

For the lower bound part of (5.3) when $n \gtrsim k\log k$, we apply Le Cam's two-point method [96] by considering two possible distributions, namely $P = \mathrm{Bern}(0)$ and $Q = \mathrm{Bern}(\frac{1}{k})$. Then

$$\begin{aligned} R_S^*(k,n) &\ge \frac{1}{4}(S(P) - S(Q))^2\exp(-nD(P\|Q)) \\ &= \frac{k^2}{4}\exp\left(n\log\left(1-\frac{1}{k}\right) - 2\log k\right). \end{aligned}$$

Since $n \gtrsim k\log k$, we have $n\log\left(1-\frac{1}{k}\right) - 2\log k \gtrsim -\frac{n}{k}$. $\square$

5.2.7 Dual program of (5.30)

Define the following infinite-dimensional linear program:

$$\begin{aligned}\mathcal{E}_1^* \triangleq \sup\ & \mathbb{P}\left[U'=0\right]-\mathbb{P}\left[U=0\right],\\ \text{s.t. } & \mathbb{E}\left[U\right]=\mathbb{E}\left[U'\right]=1,\\ & \mathbb{E}\left[U^j\right]=\mathbb{E}\left[U'^j\right],\quad j=1,\dots,L+1,\\ & U,U'\in\{0\}\cup I,\end{aligned}\tag{5.32}$$

where $I=[a,b]$ with $b>a\geq 1$ and the variables are probability measures on I (distributions of the random variables U,U'). Then (5.30) is a special case of (5.32) with $I=[1,\lambda]$.

Lemma 5.1. $\mathcal{E}_1^*=\inf_{p\in\mathcal{P}_L}\sup_{x\in I}\left|\frac{1}{x}-p(x)\right|$.

Proof. We first show that (5.30) coincides with the following optimization problem:

$$\begin{aligned}\mathcal{E}_2^* \triangleq \sup\ & \mathbb{E}\left[\frac{1}{X}\right]-\mathbb{E}\left[\frac{1}{X'}\right],\\ \text{s.t. } & \mathbb{E}\left[X^j\right]=\mathbb{E}\left[X'^j\right],\quad j=1,\dots,L,\\ & X,X'\in I.\end{aligned}\tag{5.33}$$

Given any feasible solution U,U' to (5.30), construct X,X' with the following distributions:

$$\begin{aligned}P_X(\mathrm{d}x)&=xP_U(\mathrm{d}x),\\ P_{X'}(\mathrm{d}x)&=xP_{U'}(\mathrm{d}x).\end{aligned}\tag{5.34}$$

It is straightforward to verify that X,X' are feasible for (5.33) and

$$\mathcal{E}_2^*\geq\mathbb{E}\left[\frac{1}{X}\right]-\mathbb{E}\left[\frac{1}{X'}\right]=\mathbb{P}\left[U'=0\right]-\mathbb{P}\left[U=0\right].$$

Therefore $\mathcal{E}_2^*\geq\mathcal{E}_1^*$.

On the other hand, given any feasible X,X' for (5.33), construct U,U' with the distributions:

$$\begin{aligned}P_U(\mathrm{d}u)&=\left(1-\mathbb{E}\left[\frac{1}{X}\right]\right)\delta_0(\mathrm{d}u)+\frac{1}{u}P_X(\mathrm{d}u),\\ P_{U'}(\mathrm{d}u)&=\left(1-\mathbb{E}\left[\frac{1}{X'}\right]\right)\delta_0(\mathrm{d}u)+\frac{1}{u}P_{X'}(\mathrm{d}u),\end{aligned}\tag{5.35}$$

which are well-defined since $X,X'\geq 1$ and hence $\mathbb{E}\left[\frac{1}{X}\right]\leq 1,\mathbb{E}\left[\frac{1}{X'}\right]\leq 1$.

Then U, U' are feasible for (5.30) and hence

$$\mathcal{E}_1^* \geq \mathbb{P}[U' = 0] - \mathbb{P}[U = 0] = \mathbb{E}\left[\frac{1}{X}\right] - \mathbb{E}\left[\frac{1}{X'}\right].$$

Therefore $\mathcal{E}_1^* \geq \mathcal{E}_2^*$. Finally, the dual of (5.33) is precisely the best polynomial approximation problem (see, e.g., [55, Appendix E]) and hence

$$\mathcal{E}_1^* = \mathcal{E}_2^* = \inf_{p \in \mathcal{P}_L} \sup_{x \in I} \left| \frac{1}{x} - p(x) \right|. \qquad \square$$

5.2.8 Proof of upper bounds

Proof of Proposition 5.1. First we consider the bias:

$$|\mathbb{E}(\hat{S}_{\text{seen}} - S)| = \sum_{i: p_i \geq \frac{1}{k}} (1 - \mathbb{P}(N_i \geq 1)) = \sum_{i: p_i \geq \frac{1}{k}} \exp(-np_i) \leq S \exp(-n/k).$$

The variance satisfies

$$\mathsf{var}[\hat{S}_{\text{seen}}] = \sum_{i: p_i \geq \frac{1}{k}} \mathsf{var}\mathbf{1}_{\{N_i > 0\}} \leq \sum_{i: p_i \geq \frac{1}{k}} \exp(-np_i) \leq S \exp(-n/k).$$

The conclusion follows.

For the negative result, under the Poissonized model and with the samples drawn from the uniform distribution, the plug-in estimator $\hat{S}_{\text{seen}}$ is distributed as binomial$(k, 1-e^{-n/k})$. If $n \leq (1-\delta)k \log \frac{1}{\epsilon} < k \log \frac{1}{\epsilon}$, then $1-e^{-n/k} < 1-\epsilon$. By the Chernoff bound,

$$\begin{aligned} \mathbb{P}[|\hat{S}_{\text{seen}} - S(P)| \leq \epsilon k] &= \mathbb{P}[\text{binomial}(k, 1 - e^{-n/k}) \geq (1-\epsilon)k] \\ &\leq e^{-kd(1-\epsilon \| 1-e^{-n/k})} = e^{-kd(\epsilon \| e^{-n/k})}, \end{aligned}$$

where $d(p\|q) \triangleq p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$ is the binary divergence function. Since $e^{-n/k} \geq \epsilon^{1-\delta} > \epsilon$,

$$d(\epsilon \| e^{-n/k}) \geq d(\epsilon \| \epsilon^{1-\delta}) \geq d(k^{-1} \| k^{-1+\delta}) \asymp k^{-1+\delta},$$

where the middle inequality follows from the fact that $\epsilon \mapsto d(\epsilon \| \epsilon^{1-\delta})$ is increasing near zero. Therefore $\mathbb{P}[|\hat{S}_{\text{seen}} - S(P)| \leq \epsilon k] \leq \exp(-\Omega(k^{\delta}))$. $\square$

Proof of Proposition 5.2. First we consider the bias. By (5.13) the bias of $\hat{S}$ is

$$|\mathbb{E}[\hat{S}-S]| \le \sum_{i:p_i>0} |e^{-np_i} P_L(p_i)| \le S \max_{x\in[\frac{1}{k},1]} |e^{-nx} P_L(x)|, \tag{5.36}$$

where P_L is the Chebyshev polynomial in (5.14). Recall that $L = \lfloor c_0 \log k \rfloor, l = \frac{1}{k}, r = \frac{c_1 \log k}{n}$. Then

$$\max_{x\in[l,r]} |P_L(x)| = \frac{1}{|T_L(-\frac{r+l}{r-l})|}, \tag{5.37}$$

$$\max_{x\in(r,1]} |e^{-nx} P_L(x)| = \frac{\max_{x\in(r,1]} e^{-nx} |T_L(\frac{2x-r-l}{r-l})|}{|T_L(-\frac{r+l}{r-l})|}. \tag{5.38}$$

We need Lemma 5.2 to upper bound (5.38).

Lemma 5.2. *If $\alpha \triangleq L/\beta = \Omega(1)$, then*

$$\max_{x\ge 1} e^{-\beta x} T_L(x) = \frac{1}{2}\left(\frac{\alpha+\sqrt{\alpha^2+1}}{e^{\sqrt{1+1/\alpha^2}}}(1+o_L(1))\right)^L, \quad L\to\infty.$$

Applying Lemma 5.2 to (5.38) with $L = \lfloor c_0 \log k \rfloor$, $\beta = \frac{nr(1-\delta)}{2} = \frac{c_1 \log k(1-\delta)}{2}$, we obtain that,

$$\max_{x\ge r} \left| e^{-nx} T_L\left(\frac{2x-r-l}{r-l}\right)\right| \le \frac{1}{2}\left(\frac{2\rho+\sqrt{(2\rho)^2+1}}{e^{\sqrt{1+1/(2\rho)^2}+1/(2\rho)}}(1+o_k(1)+o_\delta(1))\right)^L, \tag{5.39}$$

where $\rho \triangleq c_0/c_1$. Combining (5.37) to (5.39),

$$\max_{x\in[l,1]} |e^{-nx} P_L(x)| \le \frac{1+o_k(1)+o_\delta(1)}{|T_L(-\frac{1+\delta}{1-\delta})|},$$

as long as we pick the constant ρ such that $\frac{2\rho+\sqrt{(2\rho)^2+1}}{e^{\sqrt{1+1/(2\rho)^2}+1/(2\rho)}} < 1 \Leftrightarrow \operatorname{arcsinh}(2\rho) < \frac{1+\sqrt{1+4\rho^2}}{2\rho}$, or equivalently, $\rho < \rho^* \approx 1.1$. Then, by (5.36), the bias of $\hat{S}$ is at most

$$\begin{aligned}|\mathbb{E}[\hat{S}-S]| &\le S\frac{1+o_k(1)+o_\delta(1)}{|T_L(-\frac{1+\delta}{1-\delta})|} \le 2S(1+o_k(1)+o_\delta(1))\left(1-\frac{2\sqrt{\delta}}{1+\sqrt{\delta}}\right)^L \\ &= 2S(1+o_k(1))\exp\left(-(1+o_\delta(1))\sqrt{4c_0\rho\frac{n\log k}{k}}\right).\end{aligned} \tag{5.40}$$

Now we turn to the variance of $\hat{S}$:

$$\begin{aligned}\mathsf{var}[\hat{S}] = \sum_{i:p_i>0} \mathsf{var}\left[u_{N_i}\mathbf{1}_{\{N_i\le L\}}\right] &\le \sum_{i:p_i>0} \mathbb{E}\left[u_{N_i}^2\mathbf{1}_{\{N_i\le L\}}\right] \\ &\le \|u\|_\infty^2 \sum_{i:p_i>0} \mathbb{P}[N_i \le L], \end{aligned} \tag{5.41}$$

where $\Phi_j \triangleq \sum_i \mathbf{1}_{\{N_i=j\}}$ is the fingerprint of samples. The following lemma shows that $|u_j|$ is at most exponential in the degree of the polynomial.

Lemma 5.3. *Let a_j be defined as (5.14) and u_j be defined as (5.17). Then,*

$$\|u\|_\infty \le \frac{e\sqrt{L}}{2}\exp\left(\tau\left(\frac{L}{nr}\right)L\right), \tag{5.42}$$

where $\tau(x) \triangleq \operatorname{arcsinh}(2x) - \frac{\sqrt{1+4x^2}-1}{2x}$.

From (5.41) and (5.42) the variance of $\hat{S}$ is at most

$$\mathsf{var}[\hat{S}] \le S\frac{e^2L}{4}k^{2c_0\tau(\rho)}. \tag{5.43}$$

Then, from (5.40) and (5.43), we obtain that

$$\begin{aligned}\sup_{P\in\mathcal{D}_k}\mathbb{E}(\hat{S}-S)^2 \le 4k^2(1+o_k(1))\exp\left(-2(1+o_\delta(1))\sqrt{\frac{2\rho}{\tau(\rho)}\frac{n\log k}{k}}\right) \\ +\frac{e^2c_0\log k}{4}k^{1+2c_0\tau(\rho)}.\end{aligned}$$

Note that the first term is $4k^{2-o_\delta(1)}$. Therefore as long as we pick constant c_0 such that $2c_0\tau(\rho)<1$ the second term is lower order than the first term, and thus

$$\sup_{P\in\mathcal{D}_k}\mathbb{E}(\hat{S}-S)^2 \le 4k^2(1+o_k(1))\exp\left(-2(1+o_\delta(1))\sqrt{\frac{2\rho}{\tau(\rho)}\frac{n\log k}{k}}\right).$$

The conclusion follows from the fact that $\sup_{\rho<\rho^*} 2\rho/\tau(\rho) \approx 2.494$, which corresponds to choosing $c_0 \approx 0.558$ and $c_1 = 0.5$. □

Proof of Proposition 5.3. Let $\delta = l/r$, which is less than some absolute constant C/c_0 when $\epsilon > n^{-C}$. The upper bound of mean squared error is essentially the same as the proof of Proposition 5.2. The bias of $\tilde{S}$ is at most

$S \max_{x\in[p_{\min},1]} e^{-nx}|P_L(x)|$ given in (5.36). For $p_i \in [l,r]$, the bias is upper bounded by the uniform approximation error

$$\max_{x\in[l,r]} |P_L(x)| \le \frac{1}{|T_L(-\frac{1+\delta}{1-\delta})|} \le 2\left(1-\frac{2\sqrt{\delta}}{1+\sqrt{\delta}}\right)^L \le 2\epsilon.$$

For $p_i > r$, following (5.38)-(5.39), we have $e^{-np_i}|P_L(p_i)| = o(\epsilon)$ as long as $c_0/c_1 < \rho^* \approx 1.1$. For $p_i \in [p_{\min}, l]$, since the shifted Chebyshev polynomial P_L is monotone on $(-\infty, l)$, we have

$$\begin{aligned}|P_L(x)| &\le \frac{|T_L(\frac{2p_{\min}-r-l}{r-l})|}{|T_L(\frac{-r-l}{r-l})|} = \frac{|T_L(1+\frac{2\alpha\delta}{1-\delta})|}{|T_L(1+\frac{2\delta}{1-\delta})|} \\ &= \exp\left(-(1-o_\delta(1))2(1-\sqrt{\alpha})L\sqrt{\delta}\right) \le \epsilon^{1-\sqrt{\alpha}},\end{aligned}$$

where $\alpha = \frac{l-p_{\min}}{l} \in (0,1)$ denotes the relative deviation of l from $p_{\min}$, and we used the following equation of the Chebyshev polynomial evaluated at $1+x$ for $x > 0$:

$$\begin{aligned}T_L(1+x) &= \frac{1}{2}\left(\left(1+x-\sqrt{x^2+2x}\right)^L + \left(1+x+\sqrt{x^2+2x}\right)^L\right) \\ &= \frac{1}{2}\exp\left((1+o_x(1))L\sqrt{2x}\right).\end{aligned}$$

To conclude, the bias of $\tilde{S}$ is at most

$$\max_{x\in[p_{\min},1]} e^{-nx}|P_L(x)| \le S\epsilon^{1-\sqrt{(1-p_{\min}/l)\vee 0}}.$$

By similar analysis to (5.41) and (5.42) the variance is at most $O(Sn^c)$ for some constant $c < 1$. □

5.2.9 Proof of lower bounds

Proof of Proposition 5.4. For $0 < \nu < 1$, define the set of *approximate* probability vectors by

$$\mathcal{D}_k(\nu) \triangleq \left\{P = (p_1, p_2, \dots) : \left|\sum_i p_i - 1\right| \le \nu, p_i \in \{0\} \cup \left[\frac{1+\nu}{k}, 1\right]\right\},$$

which reduces to the original probability distribution space $\mathcal{D}_k$ if $\nu = 0$. Generalizing the sample complexity $n^*(k, \epsilon)$ in Definition 5.1 to the Poisson sampling model over $\mathcal{D}_k(\nu)$, we define

$$\tilde{n}^*(k, \epsilon, \nu) \triangleq \min\{n \geq 0 \colon \exists \hat{S}, \text{ s.t. } \mathbb{P}[|\hat{S} - S(P)| \geq \epsilon k] \leq 0.2, \forall P \in \mathcal{D}_k(\nu)\}, \tag{5.44}$$

where $\hat{S}$ is an integer-valued estimator measurable with respect to $N = (N_1, N_2, \dots) \overset{\text{ind}}{\sim} \text{Poi}(np_i)$. The sample complexity of the fixed-sample-size and Poissonized model is related by Lemma 5.4.

Lemma 5.4. *For any $\nu \in (0, 1)$ and any $\epsilon \in (0, \frac{1}{2})$,*

$$n^*(k, \epsilon) \geq (1 - \nu)\tilde{n}^*(k, \epsilon, \nu) \left(1 - O\left(\frac{1}{\sqrt{(1 - \nu)\tilde{n}^*(k, \epsilon, \nu)}}\right)\right). \tag{5.45}$$

To establish a lower bound of $\tilde{n}^*(k, \epsilon, \nu)$, we apply generalized Le Cam's method involving two composite hypothesis. Given two random variables $U, U' \in [0, k]$ with unit mean we can construct two random vectors by $\mathsf{P} = \frac{1}{k}(U_1, \dots, U_k)$ and $\mathsf{P}' = \frac{1}{k}(U'_1, \dots, U'_k)$ with i.i.d. entries. Then $\mathbb{E}[S(\mathsf{P})] - \mathbb{E}[S(\mathsf{P}')] = k(\mathbb{P}[U > 0] - \mathbb{P}[U' > 0])$. Furthermore, both $S(\mathsf{P})$ and $S(\mathsf{P}')$ are binomially distributed, which are tightly concentrated at the respective means. We can reduce the problem to the separation on mean values, as shown in Lemma 5.5.

Lemma 5.5. *Let $U, U' \in \{0\} \cup [1 + \nu, \lambda]$ be random variables such that $\mathbb{E}[U] = \mathbb{E}[U'] = 1$, $\mathbb{E}[U^j] = \mathbb{E}[U'^j]$ for $j \in [L]$, and $|\mathbb{P}[U > 0] - \mathbb{P}[U' > 0]| = d$, where $\nu \in (0, 1), L \in \mathbb{N}, d \in (0, 1)$, and $\lambda > 1 + \nu$. Then, for any $\alpha < 1/2$,*

$$\frac{2\lambda}{k\nu^2} + \frac{2}{k\alpha^2 d^2} + k\left(\frac{en\lambda}{2kL}\right)^L \leq 0.6 \Rightarrow \tilde{n}^*\left(k, \frac{(1 - 2\alpha)d}{2}, \nu\right) \geq n. \tag{5.46}$$

Applying Lemma 5.1 in Section 5.2.7, we obtain two random variables $U, U' \in \{0\} \cup [1 + \nu, \lambda]$ such that $\mathbb{E}[U] = \mathbb{E}[U'] = 1$, $\mathbb{E}[U^j] = \mathbb{E}[U'^j], j =$

$1, \ldots, L$ and

$$\begin{aligned}\mathbb{P}[U > 0] - \mathbb{P}[U' > 0] &= 2E_{L-1}\left(\frac{1}{x}, [1+\nu, \lambda]\right) \\ &= \frac{\left(1+\sqrt{\frac{1+\nu}{\lambda}}\right)^2}{1+\nu}\left(1 - \frac{2\sqrt{\frac{1+\nu}{\lambda}}}{1+\sqrt{\frac{1+\nu}{\lambda}}}\right)^L \triangleq d,\end{aligned}$$

where the value of $E_{L-1}(\frac{1}{x}, [1+\nu, \lambda])$ follows from [34, 2.11.1]. To apply Lemma 5.5 and obtain a lower bound of $\tilde{n}^*(k, \epsilon, \nu)$, we need to pick the parameters depending on the given k and ϵ to fulfill:

$$\frac{(1-2\alpha)d}{2} \geq \epsilon, \tag{5.47}$$

$$\frac{2\lambda}{k\nu^2} + \frac{2}{k\alpha^2 d^2} + k\left(\frac{en\lambda}{2kL}\right)^L \leq 0.6. \tag{5.48}$$

Let

$$\begin{aligned}L = \lfloor c_0 \log k \rfloor, \quad \lambda = \left(\frac{\gamma \log k}{\log(1/2\epsilon)}\right)^2, \quad n = C\frac{k}{\log k}\log^2\frac{1}{2\epsilon}, \\ \alpha = \frac{1}{k^{1/3}}, \quad \nu = \sqrt{\sqrt{\lambda/k}(1-2\epsilon)},\end{aligned}$$

for some $c_0, \gamma, C \asymp 1$ to be specified, and by assumption $L, \lambda \to \infty$, $\frac{\alpha}{1-2\epsilon} = o_k(1)$, $\frac{\nu}{1-2\epsilon} = o_\tau(1) + o_k(1)$, $1/\lambda = o_\delta(1)$. Since $d \geq \frac{1}{1+\nu}(1 - 2\sqrt{\frac{1+\nu}{\lambda}})^L$, a sufficient condition for (5.47) is that

$$\left(1 - 2\sqrt{\frac{1+\nu}{\lambda}}\right)^L \geq 2\epsilon\frac{1+\nu}{1-2\alpha} \Leftrightarrow \frac{\gamma}{c_0} > 2 + o_\tau(1) + o_\delta(1) + o_k(1). \tag{5.49}$$

Now we consider (5.48). By the choice of ν and α, we have

$$\nu \gg \sqrt{\lambda/k}, \quad \alpha \gg 1/\sqrt{k}d,$$

since $1 - 2\epsilon \gg \frac{\sqrt{\log k}}{k^{1/4}}$, $d \geq \frac{2\epsilon}{1-2\alpha}$ and $\epsilon = k^{-o(1)}$. Then the first two terms in (5.48) vanish. The last term in (5.48) vanishes as long as the constant

$C < \frac{2c_0}{e\gamma^2} e^{-1/c_0}$. By the fact that

$$\sup\left\{\frac{2c_0}{e\gamma^2} e^{-1/c_0} : 0 < 2c_0 < \gamma\right\} = \frac{1}{2e^2},$$

the optimal C satisfying (5.49) is $\frac{1+o_\delta(1)+o_\tau(1)+o_k(1)}{2e^2}$. Therefore, combining (5.47) – (5.48) and applying (5.46), we obtain a lower bound of $\tilde{n}^*$ that

$$\tilde{n}^*(k,\epsilon,\nu) \geq \frac{1+o_\delta(1)+o_\tau(1)+o_k(1)}{2e^2} \frac{k}{\log k} \log^2 \frac{1}{2\epsilon}.$$

Since $1-2\epsilon \gg \frac{\sqrt{\log k}}{k^{1/4}}$, we have $\tilde{n}^*(k,\epsilon,\nu) \gg \sqrt{k}$. Applying Lemma 5.4, we conclude the desired lower bound of $n^*(k,\epsilon)$. □

5.2.10 Proof of lemmas

Proof of Lemma 5.2. By assumption, $\alpha = \frac{L}{\beta}$ is strictly bounded away from zero. Let $f(x) \triangleq e^{-\beta x} T_L(x) = e^{-\beta x} \cosh(L \operatorname{arccosh}(x))$ when $x \geq 1$. By taking the derivative of f, we obtain that f is decreasing if and only if

$$\frac{\tanh(L \operatorname{arccosh}(x))}{\sqrt{x^2-1}} = \frac{\tanh(Ly)}{\sinh(y)} < \frac{1}{\alpha},$$

where $x = \cosh(y)$. Let $g(y) = \frac{\tanh(Ly)}{\sinh(y)}$. Note that g is strictly decreasing on $\mathbb{R}_+$ with $g(0) = L$ and $g(\infty) = 0$. Therefore f attains its maximum at x^* which is the unique solution of $\frac{\tanh(L \operatorname{arccosh}(x))}{\sqrt{x^2-1}} = \frac{1}{\alpha}$. It is straightforward to verify that the solution satisfies $x^* = \sqrt{1+\alpha^2}(1-o_L(1))$ when α is strictly bounded away from zero. Therefore the maximum value of f is

$$e^{-\beta x^*} T_L(x^*) = e^{-L\sqrt{1+1/\alpha^2}(1-o_L(1))} \frac{1}{2}(z^L + z^{-L}),$$

where we used (2.19) and $z = x^* + \sqrt{x^{*2}-1} = (\sqrt{1+\alpha^2}+\alpha)(1-o_L(1))$ is strictly bounded away from 1. This proves the lemma. □

Proof of Lemma 5.3. Recall that the polynomial coefficients a_j can be expressed in terms of the derivatives of P_L by (5.15). It is well known that the maximum of the derivatives of a polynomial on a compact interval can be upper bounded in terms of the maximum of the polynomial itself; one of

such results is Markov brothers' inequality (see, e.g., [139]):

$$\max_{-1\le x\le 1} |P^{(k)}(x)| \le 2^k k! \frac{n}{n+k}\binom{n+k}{2k} \max_{-1\le x\le 1} |P(x)|,$$

where P is any polynomial of degree at most n. Applying the above inequality to the degree-L polynomial $P(x) = \frac{T_L(\frac{r+l}{r-l}x)}{T_L(\frac{r+l}{r-l})}$ that is at most one on $[-1,1]$, we obtain from (5.15) that

$$|a_j| \le \left(\frac{4}{r+l}\right)^j \frac{L}{L+j}\binom{L+j}{2j}. \tag{5.50}$$

We use the following bound on binomial coefficients [140, Lemma 4.7.1]:

$$\frac{\sqrt{\pi}}{2} \le \frac{\binom{n}{k}}{(2\pi n\lambda(1-\lambda))^{-1/2}\exp(nh(\lambda))} \le 1, \tag{5.51}$$

where $\lambda = \frac{k}{n} \in (0,1)$ and $h(\lambda) \triangleq -\lambda\log\lambda - (1-\lambda)\log(1-\lambda)$ denotes the binary entropy function. Therefore, from (5.50) and (5.51), for $j = 1,\ldots,L-1$,

$$\begin{aligned} |a_j| &\le \left(\frac{4}{r+l}\right)^j \frac{L}{L+j} \frac{\exp((L+j)h(\frac{2j}{L+j}))}{\sqrt{2\pi\cdot 2j\frac{L-j}{L+j}}} \\ &\le \frac{1}{2}\left(\frac{4}{r}\right)^j \exp\left((L+j)h\left(\frac{2j}{L+j}\right)\right), \end{aligned} \tag{5.52}$$

where we used the fact that $\max_{j\in[L-1]} \frac{L}{\sqrt{4\pi j(L-j)(L+j)}} = \frac{L}{\sqrt{4\pi(L^2-1)}} \le \frac{1}{2}$ for $L \ge 2$. From (5.50), the upper bound (5.52) also holds for $j = L$. Using (5.52) and Stirling's approximation that $n! < e\sqrt{n}(\frac{n}{e})^n$, we upper bound $|u_j| = \frac{|a_j|j!}{n^j}$ by, with $\rho \triangleq \frac{L}{nr}$ and $\beta \triangleq j/L$,

$$\begin{aligned} |u_j| &\le \left(\frac{4\rho}{L}\right)^j \frac{e\sqrt{j}}{2}\left(\frac{j}{e}\right)^j \exp\left((L+j)h\left(\frac{2j}{L+j}\right)\right) \\ &= \frac{e\sqrt{j}}{2} e^{L(\beta\log\frac{4\rho\beta}{e} + (1+\beta)h(\frac{2\beta}{1+\beta}))} \le \frac{e\sqrt{L}}{2}\exp(L\tau(\rho)), \end{aligned} \tag{5.53}$$

where

$$\tau(\rho) \triangleq \sup_{\beta\in[0,1]} (\beta \log \frac{4\rho\beta}{e} + (1+\beta)h(\frac{2\beta}{1+\beta})) = \operatorname{arcsinh}(2\rho) - \frac{\sqrt{1+4\rho^2}-1}{2\rho}. \tag{5.54}$$

The conclusion follows. □

Remark 5.5. The upper bound (5.53) on the coefficients $\|u\|_\infty$ using Markov brothers' inequality is in fact tight when $l \ll r$. Note that $|T_L(1-\frac{2x}{r})| \le 1$ for $x \in [0,r] \supseteq [l,r]$. By [34, 2.9.11], the magnitude of each coefficient of $T_L(1-\frac{2x}{r})$ is at most that of the corresponding coefficient in $T_L(\frac{x-r-l}{r-l})$. Note that

$$T_L\left(1-\frac{2x}{r}\right) = \sum_{j=0}^{L} \frac{L}{L+j}\left(\frac{-4}{r}\right)^j \binom{L+j}{2j} x^j.$$

Applying [34, 2.9.11], we can lower bound the magnitude of coefficients of P_L in (5.14) by

$$|a_j| \ge \frac{1}{|T_L(-\frac{r+l}{r-l})|} \frac{L}{L+j}\left(\frac{4}{r}\right)^j \binom{L+j}{2j}. \tag{5.55}$$

From (5.40), we have $|T_L(-\frac{r+l}{r-l})| = \exp(\Theta(L\sqrt{\delta}))$, where $\delta = l/r = o(1)$. Analogous to (5.52) and (5.53), applying Stirling's approximation yields a matching lower bound of $\|u\|_\infty$.

Proof of Lemma 5.4. Fix an arbitrary $P = (p_1, p_2, \ldots) \in \mathcal{D}_k(\nu)$. Let $N = (N_1, N_2, \ldots) \overset{\text{ind}}{\sim} \text{Poi}(np_i)$ and let $n' = \sum N_i \sim \text{Poi}(n\sum p_i) \ge_{\text{s.t.}} \text{Poi}(n(1-\nu))$. Let $\hat{S}_n$ be the optimal estimator of support size for fixed sample size n, such that whenever $n \ge n^*(k,\epsilon)$ we have $\mathbb{P}[|\hat{S}_n - S(P)| \ge \epsilon k] \le 0.1$ for any $P \in \mathcal{D}_k$. We construct an estimator for the Poisson sampling model by $\tilde{S}(N) = \hat{S}_{n'}(N)$. We observe that conditioned on $n' = m$, $N \sim \text{multinomial}(m, \frac{P}{\sum_i p_i})$. Note that $\frac{P}{\sum_i p_i} \in \mathcal{D}_k$ by the definition of $\mathcal{D}_k(\nu)$. Therefore

$$\begin{aligned}
\mathbb{P}\left[\left|\tilde{S}(N) - S(P)\right| \ge \epsilon k\right] &= \sum_{m=0}^{\infty} \mathbb{P}\left[\left|\hat{S}_m(N) - S\left(\frac{P}{\sum_i p_i}\right)\right| \ge \epsilon k\right] \mathbb{P}\left[n' = m\right] \\
&\le 0.1\,\mathbb{P}[n' \ge n^*] + \mathbb{P}[n' < n^*] = 0.1 + 0.9\,\mathbb{P}[n' < n^*] \\
&\le 0.1 + 0.9\,\mathbb{P}[\text{Poi}(n(1-\nu)) < n^*].
\end{aligned}$$

If $n = \frac{1+\beta}{1-\nu} n^*$ for $\beta > 0$, then Chernoff bound (see, e.g., [56, Theorem 5.4])

yields that

$$\mathbb{P}[\mathrm{Poi}(n(1-\nu)) < n^*] \le \exp(-n^*(\beta - \log(1+\beta))).$$

By picking $\beta = \frac{C}{\sqrt{n^*}}$ for some absolute constant C, we obtain $\tilde{n}^* \le \frac{n^* + C\sqrt{n^*}}{1-\nu}$ and hence the lemma. □

Proof of Lemma 5.5. Define two random vectors

$$\mathsf{P} = \left(\frac{U_1}{k}, \ldots, \frac{U_k}{k}\right), \quad \mathsf{P}' = \left(\frac{U_1'}{k}, \ldots, \frac{U_k'}{k}\right),$$

where U_i and U_i' are i.i.d. copies of U and U', respectively. Conditioned on P and P' respectively, the corresponding histogram $N = (N_1, \ldots, N_k) \overset{\text{ind}}{\sim} \mathrm{Poi}(nU_i/k)$ and $N' = (N_1', \ldots, N_k') \overset{\text{ind}}{\sim} \mathrm{Poi}(nU_i'/k)$. Define the following high-probability events: for $\alpha < 1/2$,

$$E \triangleq \left\{ \left| \frac{\sum_i U_i}{k} - 1 \right| \le \nu, |S(\mathsf{P}) - \mathbb{E}\left[S(\mathsf{P})\right]| \le \alpha k d \right\},$$
$$E' \triangleq \left\{ \left| \frac{\sum_i U_i'}{k} - 1 \right| \le \nu, |S(\mathsf{P}') - \mathbb{E}\left[S(\mathsf{P}')\right]| \le \alpha k d \right\}.$$

Now we define two priors on the set $\mathcal{D}_k(\nu)$ by the following conditional distributions:

$$\pi = P_{\mathsf{P}|E}, \quad \pi' = P_{\mathsf{P}'|E'}.$$

First we consider the separation of the support sizes under π and π'. Note that $\mathbb{E}[S(\mathsf{P})] = k\mathbb{P}[U > 0]$ and $\mathbb{E}[S(\mathsf{P}')] = k\mathbb{P}[U' > 0]$, so $|\mathbb{E}[S(\mathsf{P})] - \mathbb{E}[S(\mathsf{P}')]| \ge kd$. By the definition of the events E, E' and the triangle inequality, we obtain that under π and π', both $\mathsf{P}, \mathsf{P}' \in \mathcal{D}_k(\nu)$ and

$$|S(\mathsf{P}) - S(\mathsf{P}')| \ge (1 - 2\alpha)kd. \tag{5.56}$$

Now we consider the total variation distance of the distributions of the histogram under the priors π and π'. By the triangle inequality and the fact that total variation of product distribution can be upper bounded by the

summation of individual one,

$$\begin{aligned}
&\mathsf{TV}(P_{N|E}, P_{N'|E'}) \\
&\leq \mathsf{TV}(P_{N|E}, P_N) + \mathsf{TV}(P_N, P_{N'}) + \mathsf{TV}(P_{N'}, P_{N'|E'}) \\
&= \mathbb{P}[E^c] + \mathsf{TV}\left((\mathbb{E}[\mathrm{Poi}(nU/k)])^{\otimes k}, (\mathbb{E}[\mathrm{Poi}(nU'/k)])^{\otimes k}\right) + \mathbb{P}[E'^c] \\
&\leq \mathbb{P}[E^c] + \mathbb{P}[E'^c] + k\mathsf{TV}(\mathbb{E}[\mathrm{Poi}(nU/k)], \mathbb{E}[\mathrm{Poi}(nU'/k)]). \qquad (5.57)
\end{aligned}$$

By the Chebyshev inequality and the union bound, both

$$\begin{aligned}
\mathbb{P}[E^c], \mathbb{P}[E'^c] &\leq \mathbb{P}\left[\left|\sum_i \frac{U_i}{k} - 1\right| > \nu\right] + \mathbb{P}\left[|S(\mathsf{P}) - \mathbb{E}\left[S(\mathsf{P})\right]| > \alpha k d\right] \\
&\leq \frac{\sum_i \mathsf{var}[U_i]}{(k\nu)^2} + \frac{\sum_i \mathsf{var}[\mathbf{1}_{\{U_i>0\}}]}{(\alpha k d)^2} \leq \frac{\lambda}{k\nu^2} + \frac{1}{k\alpha^2 d^2}, \qquad (5.58)
\end{aligned}$$

where we upper bounded the variance of U by $\mathsf{var}[U] \leq \mathbb{E}[U^2] \leq \mathbb{E}[\lambda U] = \lambda$.

Applying the total variation bound for Poisson mixtures in Theorem 3.5 yields that

$$\mathsf{TV}(\mathbb{E}[\mathrm{Poi}(nU/k)], \mathbb{E}[\mathrm{Poi}(nU'/k)]) \leq \left(\frac{en\lambda}{2kL}\right)^L. \qquad (5.59)$$

Plugging (5.58) and (5.59) into (5.57), we obtain that

$$\mathsf{TV}(P_{N|E}, P_{N'|E'}) \leq \frac{2\lambda}{k\nu^2} + \frac{2}{k\alpha^2 d^2} + k\left(\frac{en\lambda}{2kL}\right)^L. \qquad (5.60)$$

Applying Le Cam's lemma [96], the conclusion follows from (5.56) and (5.60). □

5.3 Distinct elements problem

The **Distinct Elements** problem can be viewed as a special case of the **Support Size** problem discussed in Section 5.2. Samples drawn from a k-ball urn with replacement can be viewed as i.i.d. samples from a distribution supported on the set $\{\frac{1}{k}, \frac{2}{k}, \ldots, \frac{k}{k}\}$. From this perspective, any support size estimator, as well as its performance guarantee, is applicable to the **Distinct**

Elements problem. Theorem 5.2 yields a sample complexity upper bound

$$O\left(\frac{k}{\log k}\log^2\frac{k}{\Delta}\right). \tag{5.61}$$

We briefly describe and compare the strategy to construct estimators in the last and the current sections. Both are based on the idea of *polynomial approximation*, a powerful tool to circumvent the nonexistence of unbiased estimators [17]. The key is to approximate the function to be estimated by a polynomial, whose degree is chosen to balance the approximation error (bias) and the estimation error (variance). The worst-case performance guarantee for the **Support Size** problem in the last section is governed by the uniform approximation error over an interval where the probabilities may reside. In contrast, in the **Distinct Elements** problem, samples are generated from a distribution supported on a *discrete* set of values. Uniform approximation over a discrete subset leads to smaller approximation error and, in turn, improved sample complexity. It turns out that $O(\frac{k}{\log k}\log\frac{k}{\Delta})$ samples are sufficient to achieve an additive error of Δ that satisfies $k^{0.5+O(1)} \leq \Delta \leq O(k)$, which strictly improves the sample complexity (5.61) for the **Support Size** problem, thanks to the discrete structure of the **Distinct Elements** problem.

5.3.1 A summary of the sample complexity

The main results of this chapter provide bounds and constant-factor approximations of the sample complexity in various regimes summarized in Table 5.1, as well as computationally efficient algorithms. Below we highlight a few important conclusions drawn from Table 5.1:

From linear to sublinear: From the result for $k^{0.5+\delta} \leq \Delta \leq ck$ in Table 5.1, we conclude that the sample complexity is sublinear in k if and only if $\Delta = k^{1-o(1)}$, which also holds for sampling without replacement. To estimate within a constant fraction of balls $\Delta = ck$ for any small constant c, the sample complexity is $\Theta(\frac{k}{\log k})$, which coincides with the general support size estimation problem. However, in other regimes we can achieve better performance by exploiting the discrete nature of the **Distinct Elements** problem.

From linear to superlinear: The transition from linear to superlinear sample complexity occurs near $\Delta = \sqrt{k}$. Although the exact sample complexity near $\Delta = \sqrt{k}$ is not completely resolved in the current chapter, the lower bound and upper bound in Table 5.1 differ by a factor of at most $\log\log k$. In particular, the estimator via interpolation can achieve $\Delta = \sqrt{k}$ with $n = O(k \log\log k)$ samples, and achieving a precision of $\Delta \le k^{0.5-o(1)}$ requires strictly superlinear sample size.

Table 5.1: Summary of the sample complexity $n^*(k, \Delta)$, where δ is any sufficiently small constant, c is an absolute positive constant less than 0.5 (same over the table), and the notations $a \wedge b$ and $a \vee b$ stand for $\min\{a, b\}$ and $\max\{a, b\}$, respectively. The estimators are linear with coefficients obtained from either interpolation or ℓ_2-approximation.

Δ	Lower bound	Upper bound
≤ 1	$\Theta(k \log k)$	
$\left[1, \sqrt{k}(\log k)^{-\delta}\right]$	$\Theta\left(k \log \frac{k}{\Delta^2}\right)$	
$\left[\sqrt{k}(\log k)^{-\delta}, k^{0.5+\delta}\right]$	$\Omega\left(k\left(1 \vee \log \frac{k}{\Delta^2}\right)\right)$	$O\left(k \log \frac{\log k}{1 \vee \log \frac{\Delta^2}{k}}\right)$
$[k^{0.5+\delta}, ck]$	$\Theta\left(\frac{k}{\log k} \log \frac{k}{\Delta}\right)$	
$[ck, (0.5-\delta)k]$	$k \exp(-\sqrt{O(\log k \log\log k)})$[106]^6	$O\left(\frac{k}{\log k}\right)$

To establish the sample complexity, our lower bounds are obtained under zero-one loss and our upper bounds are under the (stronger) quadratic loss. Hence we also obtain the following characterization of the minimax mean squared error (MSE) of the **Distinct Elements** problem:

$$\min_{\hat{C}} \max_{k\text{-ball urn}} \mathbb{E}\left(\frac{\hat{C} - C}{k}\right)^2 = \exp\left\{-\Theta\left(\left(1 \vee \frac{n \log k}{k}\right) \wedge \left(\log k \vee \frac{n}{k}\right)\right)\right\}$$
$$= \begin{cases} \Theta(1), & n \le \frac{k}{\log k}, \\ \exp(-\Theta(\frac{n \log k}{k})), & \frac{k}{\log k} \le n \le k, \\ \exp(-\Theta(\log k)), & k \le n \le k \log k, \\ \exp(-\Theta(\frac{n}{k})), & n \ge k \log k, \end{cases}$$

where $\hat{C}$ denotes an estimator using n samples with replacements and C is the number of distinct colors in a k-ball urn.

5.3.2 Linear estimators via discrete polynomial approximation

In this section we develop a unified framework to construct linear estimators and analyze its performance. Note that linear estimators (i.e. linear combinations of fingerprints) have been previously used for estimating distribution functionals [68, 94, 95, 72]. As commonly done in the literature, we assume the *Poisson sampling model*, where the sample size is a random variable $\mathrm{Poi}(n)$ instead of being exactly n. Under this model, the histograms of the samples, which count the number of balls in each color, are independent which simplifies the analysis. Any estimator under the Poisson sampling model can be easily modified for fixed sample size, and vice versa, thanks to the concentration of the Poisson random variable near its mean. Consequently, the sample complexities of these two models are close to each other.

Performance guarantees for general linear estimators. Recall that C denotes the number of distinct colors in a urn containing k colored balls. Let k_i denote the number of balls of the i^{th} color in the urn. Then $\sum_i k_i = k$ and $C = \sum_i \mathbf{1}_{\{k_i>0\}}$. Let $X_1, X_2, \ldots$ be independently drawn with replacement from the urn. Equivalently, the X_i's are i.i.d. according to a distribution $P = (p_i)_{i\geq 1}$, where $p_i = k_i/k$ is the fraction of balls of the i^{th} color. The observed data are $X_1, \ldots, X_N$, where the sample size N is independent from $(X_i)_{i\geq 1}$ and is distributed as $\mathrm{Poi}(n)$. Under the Poisson model (or any of the sampling models described in Section 5.1.2), the *histograms* $\{N_i\}$ are sufficient statistics for inferring any aspect of the urn configuration; here N_i is the number of balls of the i^{th} color observed in the sample, which is independently distributed as $\mathrm{Poi}(np_i)$. Furthermore, the *fingerprints* $\{\Phi_j\}_{j\geq 1}$, which are the histogram of the histograms, are also sufficient for estimating any permutation-invariant distributional property [66, 130], in particular, the number of colors. Specifically, the jth fingerprint Φ_j denotes the number of colors that appear exactly j times. Note that $U \triangleq \Phi_0$, the number of unseen colors, is not observed.

The naïve estimator, "what you see is what you get," is simply the number of observed distinct colors, which can be expressed in terms of fingerprints

as

$$\hat{C}_{\text{seen}} = \sum_{j \geq 1} \Phi_j.$$

This is typically an underestimator because $C = \hat{C}_{\text{seen}} + U$. In turn, our estimator is

$$\tilde{C} = \hat{C}_{\text{seen}} + \hat{U}, \tag{5.62}$$

which adds a linear correction term

$$\hat{U} = \sum_{j \geq 1} u_j \Phi_j, \tag{5.63}$$

where the coefficients u_j's are to be specified. Since the fingerprints $\Phi_0, \Phi_1, \ldots$ are dependent (for example, they sum up to C), (5.63) serves as a linear predictor of $U = \Phi_0$ in terms of the observed fingerprints. Equivalently, in terms of histograms, the estimator has the following decomposable form:

$$\tilde{C} = \sum_{i=1}^{\infty} g(N_i), \tag{5.64}$$

where $g : \mathbb{Z}_+ \to \mathbb{R}$ satisfies $g(0) = 0$ and $g(j) = 1 + u_j$ for $j \geq 1$. In fact, any estimator that is linear in the fingerprints can be expressed of the decomposable form (5.64).

The main idea to choose the coefficients u_j is to achieve a good trade-off between the variance and the bias. In fact, it is instructive to point out that linear estimators can easily achieve exactly zero bias, which, however, comes at the price of high variance. To see this, note that the bias of the estimator (5.64) is $\mathbb{E}[\tilde{C}] - C = \sum_{i \geq 1} (\mathbb{E}[g(N_i)] - 1)$, where

$$|\mathbb{E}[g(N_i) - 1]| = e^{-np_i} \left| -1 + \sum_{j=1}^{\infty} k_i^j \frac{u_j (n/k)^j}{j!} \right| \leq e^{-n/k} \max_{a \in [k]} |\phi(a) - 1|, \tag{5.65}$$

and $\phi(a) \triangleq \sum_{j \geq 1} a^j \frac{u_j (n/k)^j}{j!}$ is a (formal) power series with $\phi(0) = 0$. The right-hand side of (5.65) can be made zero by choosing ϕ to be, e.g., the Lagrange interpolating polynomial that satisfies $\phi(0) = -1$ and $\phi(i) = 0$ for $i \in [k]$, namely, $\phi(a) = \frac{(-1)^{k+1}}{k!} \prod_{i=1}^{k} (a - i)$; however, this strategy results in a high-degree polynomial ϕ with large coefficients, which, in turn, leads to a large variance of the estimator.

To reduce the variance of our estimator, we only use the first L fingerprints in (5.63) by setting $u_j = 0$ for all $j > L$, where L is chosen to be proportional to $\log k$. This restricts the polynomial degree in (5.65) to at most L and, while possibly incurring bias, reduces the variance. A further reason for only using the first few fingerprints is that higher-order fingerprints are *almost uncorrelated* with the number of unseens Φ_0. For instance, if red balls are observed for $n/2$ times, the only information this reveals is that approximately half of the urn are red. In fact, the correlation between Φ_0 and Φ_j decays exponentially. Therefore for $L = \Theta(\log k)$, $\{\Phi_j\}_{j>L}$ offer little predictive power about Φ_0. Moreover, if a color is observed at most L times, say, $N_i \le L$, this implies that, with high probability, $k_i \le M$, where $M = O(kL/n)$, thanks to the concentration of Poisson random variables. Therefore, effectively we only need to consider those colors that appear in the urn for at most M times, i.e., $k_i \in [M]$, for which the bias is at most

$$\begin{aligned} |\mathbb{E}[g(N_i) - 1]| &\le e^{-n/k} \max_{a \in [M]} |\phi(a) - 1| = e^{-n/k} \max_{x \in [M]/M} |p(x) - 1| \\ &= e^{-n/k} \left\| Bw - \mathbf{1} \right\|_\infty , \end{aligned} \tag{5.66}$$

where $p(x) \triangleq \phi(Mx) = \sum_{j=1}^L w_j x^j$, $w = (w_1, \ldots, w_L)^\top$, and

$$w_j \triangleq \frac{u_j (Mn/k)^j}{j!}, \quad B \triangleq \begin{pmatrix} 1/M & (1/M)^2 & \cdots & (1/M)^L \\ 2/M & (2/M)^2 & \cdots & (2/M)^L \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}. \tag{5.67}$$

Here B is a (partial) Vandermonde matrix. Lastly, since $\hat{C}_{\text{seen}} \le C \le k$, we define the final estimator to be $\tilde{C}$ projected to the interval $[\hat{C}_{\text{seen}}, k]$. We have the following error bound.

Proposition 5.5. *Assume the Poisson sampling model. Let*

$$L = \alpha \log k, \quad M = \frac{\beta k \log k}{n}, \tag{5.68}$$

for any $\beta > \alpha$ such that L and M are integers. Let $w \in \mathbb{R}^L$. Let $\tilde{C}$ be defined in (5.62) with $u_j = w_j j! (\frac{k}{nM})^j$ for $j \in [L]$ and $u_j = 0$ otherwise. Define

$\hat{C} \triangleq (\tilde{C} \vee \hat{C}_{\text{seen}}) \wedge k$. *Then*

$$\mathbb{E}(\hat{C} - C)^2 \leq k^2 e^{-2n/k} \|Bw - \mathbf{1}\|_\infty^2 + k e^{-n/k} + k \max_{m \in [M]} \mathbb{E}_{N \sim \text{Poi}(nm/k)}[u_N^2] + k^{-(\beta - \alpha \log \frac{e\beta}{\alpha} - 3)}. \tag{5.69}$$

Proof. Since $\hat{C}_{\text{seen}} \leq C \leq k$, $\hat{C}$ is always an improvement of $\tilde{C}$. Define the event $E \triangleq \cap_{i=1}^k \{N_i \leq L \Rightarrow k p_i \leq M\}$, which means that whenever $N_i \leq L$ we have $p_i \leq M/k$. Since $\beta > \alpha$, applying the Chernoff bound and the union bound yields $\mathbb{P}[E^c] \leq k^{1 - \beta + \alpha \log \frac{e\beta}{\alpha}}$, and thus

$$\mathbb{E}(\hat{C} - C)^2 \leq \mathbb{E}((\hat{C} - C)\mathbf{1}_E)^2 + k^2 \mathbb{P}[E^c] \leq \mathbb{E}((\tilde{C} - C)\mathbf{1}_E)^2 + k^{3 - \beta + \alpha \log \frac{e\beta}{\alpha}}. \tag{5.70}$$

The decomposable form of $\tilde{C}$ in (5.64) leads to

$$(\tilde{C} - C)\mathbf{1}_E = \sum_{i : k_i \in [M]} (g(N_i) - 1)\mathbf{1}_{\{N_i \leq L\}} \triangleq \mathcal{E}.$$

In view of the bias analysis in (5.66), we have

$$|\mathbb{E}[\mathcal{E}]| \leq \sum_{i : k_i \in [M]} e^{-n k_i / k} \|Bw - \mathbf{1}\|_\infty \leq k e^{-n/k} \|Bw - \mathbf{1}\|_\infty . \tag{5.71}$$

Recall that $g(0) = 0$ and $g(j) = u_j + 1$ for $j \in [L]$. Since N_i is independently distributed as $\text{Poi}(n k_i / k)$, we have

$$\begin{aligned} \mathsf{var}[\mathcal{E}] &= \sum_{i : k_i \in [M]} \mathsf{var}\left[(g(N_i) - 1)\mathbf{1}_{\{N_i \leq L\}}\right] \leq \sum_{i : k_i \in [M]} \mathbb{E}\left[(g(N_i) - 1)^2 \mathbf{1}_{\{N_i \leq L\}}\right] \\ &= \sum_{i : k_i \in [M]} \left(e^{-n k_i / k} + \mathbb{E}[u_{N_i}^2]\right) \leq k e^{-n/k} + k \max_{m \in [M]} \mathbb{E}_{N \sim \text{Poi}(nm/k)}[u_N^2]. \end{aligned} \tag{5.72}$$

Combining the upper bound on the bias in (5.71) and the variance in (5.72) yields an upper bound on $\mathbb{E}[\mathcal{E}^2]$. Then the MSE in (5.69) follows from (5.70). □

Proposition 5.5 suggests that the coefficients of the linear estimator can be chosen by solving the following linear programming (LP)

$$\min_{w \in \mathbb{R}^L} \|Bw - \mathbf{1}\|_\infty \tag{5.73}$$

and showing that the solution does not have large entries. Instead of the ℓ_∞-approximation problem (5.73), whose optimal value is difficult to analyze, we solve the ℓ_2-approximation problem as a relaxation:

$$\min_{w \in \mathbb{R}^L} \|Bw - \mathbf{1}\|_2, \tag{5.74}$$

which is an upper bound of (5.73), and is in fact within an $O(\log k)$ factor since $M = O(k \log k/n)$ and $n = \Omega(k/\log k)$. In the remainder of this section, we consider two separate cases:

- $M > L$ ($n \lesssim k$): In this case, the linear system in (5.74) is overdetermined and the minimum is non-zero. Surprisingly, as shown later in this subsection, the exact optimal value can be found in closed form using discrete orthogonal polynomials. The coefficients of the solution can be bounded using the minimum singular value of the matrix B, which is analyzed in Section 5.3.3.

- $M \leq L$ ($n \gtrsim k$): In this case, the linear system is underdetermined and the minimum in (5.74) is zero. To bound the variance, it turns out that the coefficients bound obtained from the minimum singular value is not precise enough in this regime. Instead, we express the coefficients in terms of Lagrange interpolating polynomials and use Stirling numbers to obtain sharp variance bounds. This analysis in carried out in Section 5.3.4.

We finish this subsection with two remarks.

Remark 5.6 (Discrete versus continuous approximation)**.** The optimal estimator for the **Support Size** problem in [72] has the same linear form as (5.62); however, since the probabilities can take any values in an interval, the coefficients are found to be the solution of the continuous polynomial approximation problem

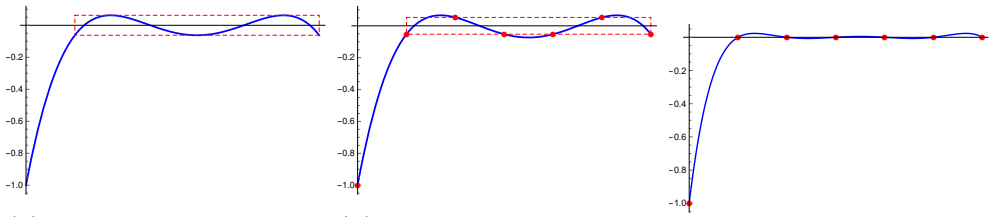
$$\inf_{p} \max_{x \in [\frac{1}{M}, 1]} |p(x) - 1| = \exp\Big(-\Theta\Big(\frac{L}{\sqrt{M}}\Big)\Big), \tag{5.75}$$

where the infimum is taken over all degree-L polynomials such that $p(0) = 0$, achieved by the (appropriately shifted and scaled) Chebyshev polynomial [34]. In contrast, we will show that the discrete version of (5.75), which is

equivalent to the LP (5.73), satisfies

$$\inf_{p} \max_{x \in \{\frac{1}{M}, \frac{2}{M}, \ldots, 1\}} |p(x) - 1| = \mathsf{poly}(M) \exp\Big(- \Theta\Big(\frac{L^2}{M}\Big)\Big), \tag{5.76}$$

provided $L < M$. The difference between (5.75) and (5.76) explains why the sample complexity (5.61) for the **Support Size** problem has an extra log factor compared to that of the **Distinct Elements** problem in Table 5.1. When the sample size n is large enough, interpolation is used in lieu of approximation. See Figure 5.5 for an illustration.



(a) Continuous approximation

(b) Discrete approximation

(c) Interpolation

Figure 5.5: Continuous and discrete polynomial approximations for $M = 6$ and degree $L = 4$, where (a) and (b) plot the optimal solution to (5.75) and (5.76) respectively. The interpolating polynomial in (c) requires a higher degree $L = 6$.

Remark 5.7 (Time complexity)**.** The time complexity of the estimator (5.62) consists of: (a) Computing histograms N_i and fingerprints Φ_j of n samples: $O(n)$; (b) Computing the coefficients w by solving the least square problem in (5.66): $O(L^2(M + L))$; (c) Evaluating the linear combination (5.62): $O(n \wedge k)$. As shown in Table 5.1, for an accurate estimation the sample complexity is $n = \Omega(\frac{k}{\log k})$, which implies $L = O(\log k)$ and $M = O(\log^2 k)$. Therefore, the overall time complexity is $O(n + \log^4 k) = O(n)$.

Exact solution to the ℓ_2-approximation. Next we give an explicit solution to the ℓ_2-approximation problem (5.74). In general, the optimal solution is given by $w^* = (B^\top B)^{-1} B^\top \mathbf{1}$ and the minimum value is the Euclidean distance between the all-one vector $\mathbf{1}$ and the column span of B, which, in the case of $M > L$, is non-zero (since B has linearly independent columns). Taking advantage of the Vandermonde structure of the matrix B in (5.67), we note that (5.74) can be interpreted as finding the orthogonal projection of

the constant function onto the linear space of polynomials of degree between 1 and L defined on the discrete set $[M]/M$. Using the orthogonal polynomials with respect to the counting measure, known as *discrete Chebyshev (or Gram) polynomials* (see [53, Section 2.8] or [141, Section 2.4.2]), we show that, surprisingly, the optimal value of the ℓ_2-approximation can be found in closed form.

Lemma 5.6. *For all $L \geq 1$ and $M \geq L+1$,*

$$\min_{w \in \mathbb{R}^L} \|Bw - \mathbf{1}\|_2 = \left[\frac{\binom{M+L+1}{L+1}}{\binom{M}{L+1}} - 1 \right]^{-1/2} = \left[\exp\left(\Theta\left(\frac{L^2}{M} \right) \right) - 1 \right]^{-1/2}. \tag{5.77}$$

Proof. Define the following inner product between functions f and g

$$\langle f, g \rangle \triangleq \sum_{i=1}^{M} f\left(\frac{i}{M}\right) g\left(\frac{i}{M}\right) \tag{5.78}$$

and the induced norm $\|f\| \triangleq \sqrt{\langle f, f \rangle}$. The least square problem (5.77) can be equivalently formulated as

$$\min_{w \in \mathbb{R}^L} \|-1 + w_1 x + w_2 x^2 + \cdots + w_L x^L\|. \tag{5.79}$$

This can be analyzed using the orthogonal polynomials under the inner product (5.78), which we describe next.

Recall the discrete Chebyshev polynomials (2.26). By appropriately shifting and scaling the set of polynomials t_m, we define an orthonormal basis for the set of polynomials of degree at most $L \leq M-1$ under the inner product (5.78) by

$$\phi_m(x) = \frac{t_m(Mx-1)}{\sqrt{c(M,m)}}, \quad m = 0, \ldots, L. \tag{5.80}$$

Since $\{\phi_m\}_{m=0}^{L}$ constitute a basis for polynomials of degree at most L, the least square problem (5.79) can be equivalently formulated as

$$\min_{a: \sum_{i=1}^{L} a_i \phi_i(0) = -1} \left\| \sum_{i=0}^{L} a_i \phi_i \right\| = \min_{a: \langle a, \phi(0) \rangle = -1} \|a\|_2,$$

where $\phi(0) \triangleq (\phi_0(0), \ldots, \phi_L(0))$, $a = (a_0, \ldots, a_L)$, and $\langle \cdot, \cdot \rangle$ denotes the

vector inner product. Thus, the optimal value is clearly $\frac{1}{\|\phi(0)\|_2}$, achieved by $a^* = -\frac{\phi(0)}{\|\phi(0)\|_2^2}$.

From (2.27) we have $p_m(0) = p_m(1) = \cdots = p_m(m-1) = 0$. By the formula of t_m in (2.26), we obtain

$$t_m(-1) = \frac{1}{m!}(-1)^m p_m(-1) = (-1)^m \prod_{j=1}^{m}(M+j).$$

In view of the definition of ϕ_m in (5.80), we have

$$\phi_m(0) = \frac{t_m(-1)}{\sqrt{c(M,m)}} = \frac{(-1)^m \prod_{j=1}^{m}(M+j)}{\sqrt{\frac{M\prod_{j=1}^{m}(M^2-j^2)}{2m+1}}} = (-1)^m \sqrt{\frac{2m+1}{M}\prod_{j=1}^{m}\frac{M+j}{M-j}}.$$

Therefore

$$\|\phi(0)\|_2^2 = \sum_{m=0}^{L}\frac{2m+1}{M}\prod_{j=1}^{m}\frac{M+j}{M-j} = \frac{\binom{M+L+1}{L+1}}{\binom{M}{L+1}} - 1,$$

where the last equality follows from induction since

$$\frac{\binom{M+L+1}{L+1}}{\binom{M}{L+1}} - \frac{\binom{M+L}{L}}{\binom{M}{L}} = \frac{2L+1}{M}\prod_{j=1}^{L}\frac{M+j}{M-j}.$$

This proves the first equality in (5.77).

The second equality in (5.77) is a direct consequence of Stirling's approximation. If $M = L+1$, then

$$\frac{\binom{M+L+1}{L+1}}{\binom{M}{L+1}} = \binom{2(L+1)}{L+1} = \exp(\Theta(L)). \tag{5.81}$$

If $M \geq L+2$, denoting $x = \frac{L+1}{M}$ and applying $n! = \sqrt{2\pi n}(\frac{n}{e})^n(1+\Theta(\frac{1}{n}))$ when $n \geq 1$, we have

$$\frac{\binom{M+L+1}{L+1}}{\binom{M}{L+1}} = \exp\left(\Theta(Mx^2) + \frac{1}{2}\log(1-x^2) + \log\frac{1+\Theta(\frac{1}{M(1-x^2)})}{1+\Theta(\frac{1}{M})}\right), \tag{5.82}$$

where the last step follows from $(1+x)\log(1+x)+(1-x)\log(1-x) = \Theta(x^2)$ when $0 \leq x \leq 1$. In the exponent of (5.82), the term $\Theta(Mx^2)$ dominates

when $M \geq L + 2$. Applying (5.81) and (5.82) to the exact solution (5.77) yields the desired approximation. □

5.3.3 Minimum singular values of real rectangle Vandermonde matrices

In Proposition 5.5 the variance of our estimator is bounded by the magnitude of coefficients u, which is related to the polynomial coefficients w by (5.67). A classical result from approximation theory is that if a polynomial is bounded over a compact interval, its coefficients are at most exponential in the degree [34, Theorem 2.9.11]: for any degree-L polynomial $p(x) = \sum_{i=0}^{L} w_i x^i$,

$$\max_{0 \leq i \leq L} |w_i| \leq \max_{x \in [0,1]} |p(x)| \exp(O(L)), \tag{5.83}$$

which is tight when p is the Chebyshev polynomial. This fact has been applied in statistical contexts to control the variance of estimators obtained from best polynomial approximation [39, 55, 72, 69]. In contrast, for the **Distinct Elements** problem, the polynomial is only known to be bounded over the discretized interval. Nevertheless, we show that the bound (5.83) continues to hold as long as the discretization level exceeds the degree:

$$\max_{0 \leq i \leq L} |w_i| \leq \max_{x \in \{\frac{1}{M}, \frac{2}{M}, \ldots, 1\}} |p(x)| \exp(O(L)), \tag{5.84}$$

provided that $M \geq L + 1$ (see Remark 5.8 after Lemma 5.7). Clearly, (5.84) implies (5.83) by sending $M \to \infty$. If $M \leq L$, a coefficient bound like (5.84) is impossible, because one can add to p an arbitrary degree-L interpolating polynomial that evaluates to zero at all M points.

To bound the coefficients, note that the optimal solution of ℓ_2-approximation is $w^* = (B^\top B)^{-1} B^\top \mathbf{1}$, and consequently

$$\|w^*\|_2 \leq \frac{\|\mathbf{1}\|_2}{\sigma_{\min}(B)}, \tag{5.85}$$

where $\sigma_{\min}(B)$ denotes the smallest singular value of B. Let

$$\bar{B} \triangleq [\mathbf{1}, B] = \begin{pmatrix} 1 & 1/M & (1/M)^2 & \cdots & (1/M)^L \\ 1 & 2/M & (2/M)^2 & \cdots & (2/M)^L \\ 1 & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & 1 \end{pmatrix},$$

which is an $M \times (L+1)$ Vandermonde matrix and satisfies $\sigma_{\min}(\bar{B}) \leq \sigma_{\min}(B)$ since $\bar{B}$ has one extra column. The Gram matrix of $\bar{B}$ is an instance of *moment matrices*. A moment matrix associated with a probability measure μ is a Hankel matrix M given by $M_{i,j} = m_{i+j-2}$, where $m_\ell = \int x^\ell \mathrm{d}\mu$ denotes the ℓth moment of μ. Then $\frac{1}{M}\bar{B}^\top \bar{B}$ is the moment matrix associated with the uniform distribution over the discrete set $\{\frac{1}{M}, \frac{2}{M}, \ldots, 1\}$, which converges to the uniform distribution over the interval $(0, 1)$. The moment matrix of the uniform distribution is the famous *Hilbert matrix* H, with

$$H_{ij} = \frac{1}{i+j-1},$$

which is a well-studied example of ill-conditioned matrices in the numerical analysis literature. In particular, it is known that the condition number of the $L \times L$ Hilbert matrix is $O(\frac{(1+\sqrt{2})^{4L}}{\sqrt{L}})$ [142] and the operator norm is $\Theta(1)$, and thus the minimum singular value is exponentially small in the degree. Therefore we expect the discrete moment matrix $\frac{1}{M}\bar{B}^\top \bar{B}$ to behave similarly to the Hilbert matrix when M is large enough. Interestingly, we show that this is indeed the case as soon as M exceeds L (otherwise the minimum singular value is zero).

Lemma 5.7. *For all $M \geq L + 1$,*

$$\sigma_{\min}\left(\frac{\bar{B}}{\sqrt{M}}\right) \geq \frac{1}{L^2 2^{7L}(2L+1)} \left(\frac{M+L}{eM}\right)^{L+0.5}. \tag{5.86}$$

Remark 5.8. The inequality (5.84) follows from Lemma 5.7 since the coefficient vector $w = (w_0, \ldots, w_L)$ satisfies $\|w\|_\infty \leq \|w\|_2 \leq \frac{1}{\sigma_{\min}(\bar{B})}\|\bar{B}w\|_2 \leq \frac{\sqrt{M}}{\sigma_{\min}(\bar{B})}\|\bar{B}w\|_\infty$.

Remark 5.9. The extreme singular values of square Vandermonde matrices have been extensively studied (c.f. [143, 144] and the references therein). For

rectangular Vandermonde matrices, the focus was mainly with nodes on the unit circle in the complex domain [145, 146, 147] with applications in signal processing. In contrast, Lemma 5.7 is on rectangular Vandermonde matrices with real nodes. The result on integers nodes in [148] turns out to be too crude for the purpose of this chapter.

Proof. Note that $\bar{B}^\top \bar{B}$ is the Gramian of monomials $\mathbf{x} = (1, x, x^2, \ldots, x^L)^\top$ under the inner product defined in (5.78). When $M \geq L+1$, the orthonormal basis $\phi = (\phi_0, \ldots, \phi_L)^\top$ under the inner product (5.78) are given in (5.80). Let $\phi = \mathbf{L}\mathbf{x}$ where $\mathbf{L} \in \mathbb{R}^{(L+1)\times(L+1)}$ is a lower triangular matrix and $\mathbf{L}$ consists of the coefficients of ϕ. Taking the Gramian of ϕ yields that $I = \mathbf{L}(\bar{B}^\top \bar{B})\mathbf{L}^\top$, i.e., $\mathbf{L}^{-1}$ can be obtained from the Cholesky decomposition: $\bar{B}^\top \bar{B} = (\mathbf{L}^{-1})(\mathbf{L}^{-1})^\top$. Then[7]

$$\sigma^2_{\min}(\bar{B}) = \frac{1}{\|\mathbf{L}\|^2_{op}} \geq \frac{1}{\|\mathbf{L}\|^2_F}, \tag{5.87}$$

where $\|\cdot\|_{op}$ denotes the ℓ_2 operator norm, which is the largest singular value of L, and $\|\cdot\|_F$ denotes the Frobenius norm. By definition, $\|\mathbf{L}\|^2_F$ is the sum of all squared coefficients of $\phi_0, \ldots, \phi_L$. A useful method to bound the sum-of-squares of the coefficients of a polynomial is by its maximal modulus over the unit circle on the complex plane. Specifically, for any polynomial $p(z) = \sum_{i=0}^n a_i z^i$, we have

$$\sum_{i=0}^n |a_i|^2 = \frac{1}{2\pi} \oint_{|z|=1} |p(z)|^2 \mathrm{d}z \leq \sup_{|z|=1} |p(z)|^2. \tag{5.88}$$

Therefore

$$\begin{aligned} \sigma_{\min}(\bar{B}) &\geq \frac{1}{\|\mathbf{L}\|_F} \geq \frac{1}{\sqrt{\sum_{m=0}^L \sup_{|z|=1} |\phi_m(z)|^2}} \\ &\geq \frac{1}{\sqrt{L+1}} \frac{1}{\sup_{0\leq m\leq L, |z|=1} |\phi_m(z)|}. \end{aligned} \tag{5.89}$$

For a given M, the orthonormal basis $\phi_m(x)$ in (5.80) is proportional to the discrete Chebyshev polynomials $t_m(Mx - 1)$. The classical asymptotic

[7]The lower bound (5.87), which was also obtained in [149, (1.13)] using Cauchy-Schwarz inequality, is tight up to polynomial terms in view of the fact that $\|\mathbf{L}\|_F \leq (L+1)\|\mathbf{L}\|_{op}$.

result for the discrete Chebyshev polynomials shows that [53, (2.8.6)]

$$\lim_{M\to\infty} M^{-m} t_m(Mx) = P_m(2x-1),$$

where P_m is the Legendre polynomial of degree m. This gives the intuition that $t_m(x) \approx M^m$ for real-valued $x \in [0, M]$. We have the following non-asymptotic upper bound for t_m over the complex plane.

Lemma 5.8. *For all* $0 \le m \le M-1$,

$$|t_m(z)| \le m^2 2^{6m} \sup_{0\le\xi\le m} \left(|z+\xi| \vee M\right)^m. \tag{5.90}$$

Applying (5.90) on the definition of ϕ_m in (5.80), for any $|z| = 1$ and any $M \ge L+1$, we have

$$|\phi_m(z)| = \frac{|t_m(Mz-1)|}{\sqrt{c(M,m)}} \le \frac{m^2 2^{7m} M^m}{\sqrt{\frac{M(M^2-1^2)(M^2-2^2)\cdots(M^2-m^2)}{2m+1}}}.$$

The right-hand side is increasing with m. Therefore,

$$\begin{aligned}\sup_{0\le m\le L, |z|=1} |\phi_m(z)| &\le \frac{L^2 2^{7L} M^L}{\sqrt{\frac{M(M^2-1^2)(M^2-2^2)\cdots(M^2-L^2)}{2L+1}}} \\ &= \frac{1}{\sqrt{M}} L^2 2^{7L} \sqrt{2L+1} \sqrt{\frac{M^{2L+1}}{\binom{M+L}{2L+1}(2L+1)!}}.\end{aligned}$$

Combining (5.89), we obtain

$$\begin{aligned}\sigma_{\min}\left(\frac{\bar{B}}{\sqrt{M}}\right) &\ge \frac{1}{L^2 2^{7L}\sqrt{(L+1)(2L+1)}} \sqrt{\frac{\binom{M+L}{2L+1}(2L+1)!}{M^{2L+1}}} \\ &\ge \frac{1}{L^2 2^{7L}(2L+1)} \left(\frac{M+L}{eM}\right)^{L+0.5},\end{aligned}$$

where in the last inequality we used $\binom{n}{k} \ge (\frac{n}{k})^k$ and $n! \ge (\frac{n}{e})^n$. □

Using the optimal solution w^* to the ℓ_2-approximation problem (5.74) as the coefficient of the linear estimator $\hat{C}$, the following performance guarantee is obtained by applying Lemma 5.6 and Lemma 5.7 to bound the bias and variance, respectively.

Theorem 5.3. *Assume the Poisson sampling model. Then,*

$$\mathbb{E}(\hat{C}-C)^2 \le k^2 \exp\left(-\Theta\left(1 \vee \frac{n\log k}{k} \wedge \log k\right)\right). \tag{5.91}$$

Proof. If $n \le \frac{k}{\log k}$, then the upper bound in (5.91) is $\Theta(k^2)$, which is trivial thanks to the thresholds that $\hat{C} = (\tilde{C} \vee \hat{C}_{\text{seen}}) \wedge k$. It is hereinafter assumed that $n \ge \frac{k}{\log k}$, or equivalently $M \le \frac{\beta}{\alpha^2}L^2$; here M, L are defined in (5.68) and the constants α, β are to be determined later. Then, from Lemma 5.6,

$$\|Bw^* - \mathbf{1}\|_\infty \le \|Bw^* - \mathbf{1}\|_2 \le \exp\left(-\Theta\left(\frac{L^2}{M}\right)\right). \tag{5.92}$$

In view of (5.85) and Lemma 5.7, we have

$$\|w^*\|_\infty \le \|w^*\|_2 \le \frac{\|\mathbf{1}\|_2}{\sigma_{\min}(B)} \le \exp(O(L)).$$

Recall the connection between u_j and w_j in (5.67). For $1 \le j \le L < \beta \log k$, we have $u_j = w_j \frac{j!}{(\beta \log k)^j} \le \frac{w_j}{\beta \log k}$. Therefore,

$$\|u^*\|_\infty \le \frac{\|w^*\|_\infty}{\beta \log k} \le \frac{\exp(O(L))}{\beta \log k}. \tag{5.93}$$

Applying (5.92) and (5.93) to Proposition 5.5, we obtain

$$\begin{aligned}\mathbb{E}(\hat{C}-C)^2 \le k^2 \exp\left(-\frac{2n}{k} - \Theta\left(\frac{n\log k}{k}\right)\right) + ke^{-n/k} \\ + k\frac{\exp(O(\log k))}{(\beta \log k)^2} + k^{-(\beta - \alpha \log \frac{e\beta}{\alpha} - 3)}.\end{aligned}$$

Then the desired (5.91) holds as long as β is sufficiently large and α is sufficiently small. □

5.3.4 Lagrange interpolating polynomials and Stirling numbers

When we sample at least a constant faction of the urn, i.e., $n = \Omega(k)$, we can afford to choose α and β in (5.68) so that $L = M$ and B is an invertible matrix. We choose the coefficient $w = B^{-1}\mathbf{1}$ which is equivalent to applying

Lagrange interpolating polynomial and achieves exact zero bias. To control the variance, we can follow the approach in Section 5.3.3 by using the bound on minimum singular value of the matrix B, which implies that the coefficients are $\exp(O(L))$ and yields a coarse upper bound $O(k\frac{\log k}{1\vee\log\frac{\Delta^2}{k}})$ on the sample complexity. As previously announced in Table 5.1, this bound can be improved to $O(k\log\frac{\log k}{1\vee\log\frac{\Delta^2}{k}})$ by a more careful analysis of the Lagrange interpolating polynomial coefficients expressed in terms of the Stirling numbers, which we introduce next.

The Stirling numbers of the first kind are defined as the coefficients of the falling factorial $(x)_n$ where

$$(x)_n = x(x-1)\dots(x-n+1) = \sum_{j=1}^{n} s(n,j)x^j.$$

Compared to the coefficients w expressed by the Lagrange interpolating polynomial:

$$\sum_{j=1}^{M} w_j x^j - 1 = -\frac{(1-xM)(2-xM)\dots(M-xM)}{M!},$$

we obtain a formula for the coefficients w in terms of the Stirling numbers:

$$w_j = \frac{(-1)^{M+1}M^j}{M!}s(M+1,j+1), \quad 1 \le j \le M.$$

Consequently, the coefficients of our estimator u_j are given by

$$u_j = (-1)^{M+1}\frac{j!}{M!}\left(\frac{k}{n}\right)^j s(M+1,j+1). \tag{5.94}$$

The precise asymptotics the Stirling number is rather complicated. In particular, the asymptotic formula of $s(n,m)$ as $n\to\infty$ for fixed m is given by [150] and the uniform asymptotics over all m is obtained in [151] and [152]. The following lemma is a coarse non-asymptotic version, which suffices for the purpose of constant-factor approximations of the sample complexity.

Lemma 5.9.

$$|s(n+1,m+1)| = n!\left(\Theta\left(\frac{1}{m}\left(1\vee\log\frac{n}{m}\right)\right)\right)^m. \tag{5.95}$$

We construct $\hat{C}$ as in Proposition 5.5 using the coefficients u_j in (5.94) to

achieve zero bias. The variance upper bound by the coefficients u is a direct consequence of the upper bound of Stirling numbers in Lemma 5.9. Then we obtain the following mean squared error (MSE).

Theorem 5.4 (Interpolation)**.** *Assume the Poisson sampling model. If $n > \eta k$ for some sufficiently large constant η, then*

$$\mathbb{E}(\hat{C} - C)^2 \le k e^{-\Theta(\frac{n}{k})} + k^{-0.5-3.5\frac{k}{n}\log\frac{k}{en}} + \epsilon(k, n),$$

where

$$\epsilon(k,n) \triangleq \begin{cases} k \exp\left(\frac{k^2 \log k}{n^2} e^{-\Theta(\frac{n}{k})}\right), & n \lesssim k \log\log k, \\ k\left(\Theta\left(\frac{k}{n}\right)\log\frac{k^2\log k}{n^2}\right)^{2n/k}, & k\log\log k \lesssim n \lesssim k\sqrt{\log k}, \\ 0, & n \gtrsim k\sqrt{\log k}. \end{cases}$$

Proof. In Proposition 5.5, fix $\beta = 3.5$ and $\alpha = \frac{\beta k}{n}$ so that $L = M$. Our goal is to show an upper bound of

$$\max_{\lambda \in \frac{n}{k}[M]} \mathbb{E}_{N\sim\mathrm{Poi}(\lambda)}[u_N^2] = \max_{\lambda \in \frac{n}{k}[M]} \sum_{j=1}^{M} u_j^2 e^{-\lambda}\frac{\lambda^j}{j!}. \tag{5.96}$$

Here the coefficients u_j are obtained from (5.94) and, in view of (5.95), satisfy:

$$|u_j| \le \left(\frac{\eta k}{n}\left(1 \vee \log\frac{M}{j}\right)\right)^j, \quad 1 \le j \le M, \tag{5.97}$$

for some universal constant η. We consider three cases separately:

Case I: $n \ge \sqrt{\beta} k\sqrt{\log k}$**.** In this case we have $\frac{n}{k} \ge M$. The maximum of each summand in (5.96) as a function of $\lambda \in \mathbb{R}$ occurs at $\lambda = j$. Since $j \le \frac{n}{k}$, the maximum over $\lambda \in \frac{n}{k}[M]$ is attained at $\lambda = \frac{n}{k}$. Then,

$$\max_{\lambda \in \frac{n}{k}[M]} \mathbb{E}_{N\sim\mathrm{Poi}(\lambda)}[u_N^2] = \mathbb{E}_{N\sim\mathrm{Poi}(\frac{n}{k})}[u_N^2]. \tag{5.98}$$

In view of (5.97) and $j \geq 1$, we have $|u_j| \leq (\Theta(k/n) \log M)^j$. Then,

$$\begin{aligned}\mathbb{E}_{N\sim\mathrm{Poi}(\frac{n}{k})}[u_N^2] &\leq \mathbb{E}_{N\sim\mathrm{Poi}(\frac{n}{k})}\left(\Theta\left(\frac{k\log M}{n}\right)^2\right)^N \\ &= \exp\left(\frac{n}{k}\left(\Theta\left(\frac{k\log M}{n}\right)^2 - 1\right)\right) = e^{-\Theta(n/k)},\end{aligned}$$

as long as $n \gtrsim k \log\log k$ and thus $\frac{k\log M}{n} \lesssim 1$. Therefore,

$$\max_{\lambda\in\frac{n}{k}[M]} \mathbb{E}_{N\sim\mathrm{Poi}(\lambda)}[u_N^2] \leq e^{-\Theta(n/k)}, \quad n \gtrsim k\sqrt{\log k}. \tag{5.99}$$

Case II: $\eta k \log\log k \leq n \leq \sqrt{\beta} k\sqrt{\log k}$**.** We apply the following upper bound:

$$\begin{aligned}&\max_{\lambda\in\frac{n}{k}[M]} \mathbb{E}_{N\sim\mathrm{Poi}(\lambda)}[u_N^2] \\ &= \max_{\lambda\in\frac{n}{k}[M]} \mathbb{E}_{N\sim\mathrm{Poi}(\lambda)}[u_N^2 \mathbf{1}_{\{N\geq n/k\}}] + \max_{\lambda\in\frac{n}{k}[M]} \mathbb{E}_{N\sim\mathrm{Poi}(\lambda)}[u_N^2 \mathbf{1}_{\{N<n/k\}}] \\ &\leq \max_{\frac{n}{k}\leq j\leq M} |u_j|^2 + e^{-\Theta(n/k)},\end{aligned} \tag{5.100}$$

where the upper bound of the second addend is analogous to (5.98) and (5.99). Since $\frac{\eta k}{n} \leq 1$, the right-hand side of (5.97) is decreasing with j when $j \geq M/e$. It suffices to consider $j \leq M/e$, when the maximum as a function of $j \in \mathbb{R}$ occurs at $j^* \leq Me^{-\frac{n}{\eta k}}$. Since $Me^{-\frac{n}{\eta k}} \leq \frac{n}{k}$ when $n \geq \eta k \log\log k$, the maximum over $\frac{n}{k} \leq j \leq M$ is attained at $j = \frac{n}{k}$. Applying (5.97) with $j = \frac{n}{k}$ to (5.100) yields

$$\max_{\lambda\in\frac{n}{k}[M]} \mathbb{E}_{N\sim\mathrm{Poi}(\lambda)}[u_N^2] \leq \left(\Theta\left(\frac{k}{n}\right)\log\frac{k^2\log k}{n^2}\right)^{2n/k} + e^{-\Theta(n/k)}. \tag{5.101}$$

Case III: $\eta k \leq n \leq \eta k \log\log k$**.** We apply the upper bound of expectation by the maximum:

$$\max_{\lambda\in\frac{n}{k}[M]} \mathbb{E}_{N\sim\mathrm{Poi}(\lambda)}[u_N^2] \leq \max_{j\in[M]} u_j^2.$$

Since $\frac{\eta k}{n} \leq 1$, the right-hand side of (5.97) is decreasing with j when $j \geq M/e$, so it suffices to consider $j \leq M/e$. Denoting $x = \log\frac{M}{j}$ and $\tau = \Theta(\frac{k}{n})$, in view of (5.97), we have $|u_j| \leq \exp(Me^{-x}\log(\tau x))$, which attains maximum

at x^* satisfying $\frac{e^{1/x^*}}{x^*} = \tau$. Then,

$$|u_j| \le \exp(Me^{-x^*}\log(\tau x^*)) = \exp(Me^{-x^*}/x^*) < \exp(M\tau e^{-1/\tau}),$$

where the last inequality is because of $\tau > \frac{1}{x^*}$. Therefore,

$$\max_{\lambda \in \frac{n}{k}[M]} \mathbb{E}_{N\sim\mathrm{Poi}(\lambda)}[u_N^2] \le \exp\left(\frac{k^2\log k}{n^2} e^{-\Theta(\frac{n}{k})}\right), \quad k \lesssim n \lesssim k\log\log k. \tag{5.102}$$

Applying the upper bounds in (5.99), (5.101) and (5.102) to Proposition 5.5 concludes the proof. □

Remark 5.10. It is impossible to bridge the gap near $\Delta = \sqrt{k}$ in Table 5.1 using the technology of interpolating polynomials that aims at zero bias, since its worst-case variance is at least $k^{1+\Omega(1)}$ when $n = O(k)$. To see this, note that the variance term given by (5.72) is

$$\sum_{p_i} \mathbb{E}_{N\sim\mathrm{Poi}(np_i)}[u_N^2] = \sum_{p_i}\sum_{j=1}^{L} u_j^2 e^{-np_i}\frac{(np_i)^j}{j!}. \tag{5.103}$$

Consider the distribution uniform$[n/j_0]$ with $j_0 = Le^{-2n/k} = \Omega(\log k)$, which corresponds to an urn where each of the n/j_0 colors appears equal number of times. By the formula of coefficient u_j in (5.94) and the characterization from Lemma 5.9, the $j = j_0$ term in the summation of (5.103) is of order $\frac{n}{j_0}(\frac{k}{n}\log\frac{M}{j_0})^{2j_0} = \frac{n}{j_0}2^{2j_0}$, which is already $k^{1+\Omega(1)}$.

5.3.5 Optimality of the sample complexity

In this subsection we develop lower bounds of the sample complexity which certify the optimality of estimators constructed in Section 5.3.2. We first give a brief overview of the lower bound in [107, Theorem 1], which gives the optimal sample complexity under the multiplicative error criterion. The lower bound argument boils down to considering two hypotheses: in the null hypothesis, the urn consists of only one color; in the alternative, the urn contains $2\Delta+1$ distinct colors, where $k-2\Delta$ balls share the same color as in the null hypothesis, and all other balls have distinct colors. These two scenarios are distinguished if and only if a second color appears in the samples, which typically requires $\Omega(k/\Delta)$ samples. This lower bound is optimal for

estimating within a multiplicative factor of $\sqrt{\Delta}$, which, however, is too loose for additive error Δ.

In contrast, instead of testing whether the urn is monochromatic, our first lower bound is given by testing whether the urn is maximally colorful, that is, containing k distinct colors. The alternative contains $k - 2\Delta$ colors, and the numbers of balls of two different colors differ by at most one. In other words, the null hypothesis is the uniform distribution on $[k]$ and the alternative is close to uniform distribution with smaller support size. The sample complexity, which is shown in Theorem 5.5, gives the lower bound in Table 5.1 for $\Delta \le \sqrt{k}$.

Theorem 5.5. *If* $1 \le \Delta \le \frac{k}{2}$, *then*

$$n^*(k, \Delta) \ge \Omega\left(\frac{k - 2\Delta}{\sqrt{k}}\right). \tag{5.104}$$

If $1 \le \Delta < \frac{k}{4}$, *then*

$$n^*(k, \Delta) \ge \Omega\left(k \operatorname{arccosh}\left(1 + \frac{k}{4\Delta^2}\right)\right) \asymp \begin{cases} k \log(1 + \frac{k}{\Delta^2}), & \Delta \le \sqrt{k}, \\ \frac{k^{3/2}}{\Delta}, & \Delta \ge \sqrt{k}. \end{cases} \tag{5.105}$$

Proof. Consider the following two hypotheses: The null hypothesis H_0 is an urn consisting of k distinct colors; the alternative H_1 consists of $k - 2\Delta$ distinct colors, and each color appears either $b_1 \triangleq \lfloor \frac{k}{k-2\Delta} \rfloor$ or $b_2 \triangleq \lceil \frac{k}{k-2\Delta} \rceil$ times. In terms of distributions, H_0 is the uniform distribution $Q = (\frac{1}{k}, \ldots, \frac{1}{k})$; H_1 is the closest perturbation from the uniform distribution: randomly pick disjoint sets of indices $I, J \subseteq [k]$ with cardinality $|I| = c_1$ and $|J| = c_2$, where c_1 and c_2 satisfy

$$\begin{aligned} \text{(number of colors)} \quad & c_1 + c_2 = k - 2\Delta, \\ \text{(number of balls)} \quad & c_1 b_1 + c_2 b_2 = k. \end{aligned}$$

Conditional on $\theta \triangleq (I, J)$, the distribution $P_\theta = (p_{\theta,1}, \ldots, p_{\theta,k})$ is given by

$$p_\theta = \begin{cases} b_1/k, & i \in I, \\ b_2/k, & i \in J. \end{cases}$$

Put the uniform prior on the alternative. Denote the marginal distributions of the n samples $X = (X_1, \ldots, X_n)$ under H_0 and H_1 by Q_X and P_X, respectively. Since the distinct colors in H_0 and H_1 are separated by 2Δ, to show that the sample complexity $n^*(k, \Delta) \geq n$, it suffices to show that no test can distinguish H_0 and H_1 reliably using n samples. A further sufficient condition is a bounded χ^2 divergence [32]

$$\chi^2(P_X \| Q_X) \triangleq \int \frac{P_X^2}{Q_X} - 1 \leq O(1).$$

The remainder of this proof is devoted to upper bounds of the χ^2 divergence.

Since $P_{X|\theta} = P_\theta^{\otimes n}$ and $Q_X = Q^{\otimes n}$, we have

$$\begin{aligned}\chi^2(P_X \| Q_X) + 1 = \int \frac{P_X^2}{Q_X} &= \int \frac{(\mathbb{E}_\theta P_{X|\theta})(\mathbb{E}_{\theta'} P_{X|\theta'})}{Q_X} \\ &= \mathbb{E}_{\theta,\theta'} \int \frac{P_{X|\theta} P_{X|\theta'}}{Q_X} = \mathbb{E}_{\theta,\theta'} \left(\int \frac{P_\theta P_{\theta'}}{Q} \right)^n,\end{aligned}$$

where θ' is an independent copy of θ. By the definition of P_θ and Q,

$$\int \frac{P_\theta P_{\theta'}}{Q} = \frac{b_1^2}{k}|I \cap I'| + \frac{b_2^2}{k}|J \cap J'| + \frac{b_1 b_2}{k}(|I \cap J'| + |J \cap I'|) = 1 + \sum_{i=1}^4 A_i, \quad (5.106)$$

where $A_1 \triangleq \frac{b_1^2}{k}(|I \cap I'| - \frac{c_1^2}{k})$, $A_2 \triangleq \frac{b_2^2}{k}(|J \cap J'| - \frac{c_2^2}{k})$, $A_3 = \frac{b_1 b_2}{k}(|I \cap J'| - \frac{c_1 c_2}{k})$, and $A_4 = \frac{b_1 b_2}{k}(|J \cap I'| - \frac{c_1 c_2}{k})$ are centered random variables. Applying $1 + x \leq e^x$ and Cauchy-Schwarz inequality, we obtain

$$\chi^2(P_X \| Q_X) + 1 \leq \mathbb{E}[e^{n \sum_{i=1}^4 A_i}] \leq \prod_{i=1}^4 (\mathbb{E}[e^{4nA_i}])^{\frac{1}{4}}. \quad (5.107)$$

Consider the first term $\mathbb{E}[e^{4nA_1}]$. Note that $|I \cap I'| \sim \text{hypergeometric}(k, c_1, c_1)$,[8] which is the distribution of the sum of c_1 samples drawn without replacement from a population of size k which consists of c_1 ones and $k - c_1$ zeros. By the convex stochastic dominance of the binomial over the hypergeometric

[8] hypergeometric(N, K, n) denotes the hypergeometric distribution with probability mass function $\binom{K}{k}\binom{N-K}{n-k}/\binom{N}{n}$, for $0 \vee (n + K - N) \leq k \leq n \wedge K$

distribution [153, Theorem 4], for $Y \sim \text{binomial}(c_1, \frac{c_1}{k})$, we have

$$\begin{aligned}(\mathbb{E}[e^{4nA_1}])^{\frac{1}{4}} &\le \left(\mathbb{E}\left[\exp\left(\frac{4nb_1^2}{k}(Y - c_1^2/k)\right)\right]\right)^{\frac{1}{4}} \\ &\le \exp\left(\frac{c_1^2}{4k}\left(\exp\left(\frac{4nb_1^2}{k}\right) - 1 - \frac{4nb_1^2}{k}\right)\right) \\ &\le \exp\left(\frac{c_1^2}{4k}\left(\exp\left(\frac{4nb_2^2}{k}\right) - 1 - \frac{4nb_2^2}{k}\right)\right), \end{aligned} \tag{5.108}$$

where the last inequality follows from the fact that $x \mapsto e^x - 1 - x$ is increasing when $x > 0$. Other terms in (5.107) are bounded analogously and we have

$$\begin{aligned}&\chi^2(P_X \| Q_X) + 1 \\ &\le \exp\left(\frac{c_1^2 + c_2^2 + 2c_1c_2}{4k}\left(\exp\left(\frac{4nb_2^2}{k}\right) - 1 - \frac{4nb_2^2}{k}\right)\right) \\ &= \exp\left(\frac{(k-2\Delta)^2}{4k}\left(\exp\left(\frac{4n}{k}\left\lceil\frac{k}{k-2\Delta}\right\rceil^2\right) - 1 - \frac{4n}{k}\left\lceil\frac{k}{k-2\Delta}\right\rceil^2\right)\right). \end{aligned} \tag{5.109}$$

If $k - 2\Delta \ge \sqrt{k}$, the upper bound (5.109) implies that $n^*(k, \Delta) \ge \Omega(\frac{k-2\Delta}{\sqrt{k}})$ since the χ^2-divergence is finite with $O(\frac{k-2\Delta}{\sqrt{k}})$ samples, using the inequality that $e^x - 1 - x \le \frac{x^2}{2}$ for $x \ge 0$; if $k - 2\Delta \le \sqrt{k}$, the lower bound is trivial since $\frac{k-2\Delta}{\sqrt{k}} \le 1$.

Now we prove the refined estimate (5.105) for $1 \le \Delta < k/4$, in which case $|I| = c_1 = k - 4\Delta, |J| = c_2 = 2\Delta$ and $b_1 = 1, b_2 = 2$. When c_1 is close to k, hypergeometric(k, c_1, c_1) is no longer well approximated by binomial$(c_1, \frac{c_1}{k})$, and the upper bound in (5.108) yields a loose lower bound for the sample complexity. To fix this, note that in this case the set $K \triangleq (I \cup J)^c$ has small cardinality $|K| = 2\Delta$. The equality in (5.106) can be equivalently represented in terms of J, J' and K, K' by

$$\int \frac{P_\theta P_{\theta'}}{Q} = 1 + \frac{|J \cap J'| + |K \cap K'| - |J \cap K'| - |K \cap J'|}{k}.$$

By upper bounds analogous to (5.107) – (5.109), $\chi^2(P_X \| Q_X) + 1 \le \prod_{i=1}^4 (\mathbb{E}[e^{4nB_i}])^{\frac{1}{4}}$, where $B_1 \triangleq \frac{1}{k}(|J \cap J'| - \frac{(2\Delta)^2}{k})$, $B_2 \triangleq \frac{1}{k}(|K \cap K'| - \frac{(2\Delta)^2}{k})$, $B_3 \triangleq -\frac{1}{k}(|J \cap K'| - \frac{(2\Delta)^2}{k})$, and $B_4 \triangleq -\frac{1}{k}(|K \cap J'| - \frac{(2\Delta)^2}{k})$. Note that $|J \cap J'|, |K \cap K'|, |J \cap K'|, |K \cap J'|$ are all distributed as hypergeometric$(k, 2\Delta, 2\Delta)$,

which is dominated by binomial$(2\Delta, \frac{2\Delta}{k})$. For $Y \sim$ binomial$(2\Delta, \frac{2\Delta}{k})$, we have

$$\begin{aligned}(\mathbb{E}[e^{4nB_i}])^{\frac{1}{4}} &\le \left(\mathbb{E}\left[\exp\left(t\left(Y - \frac{(2\Delta)^2}{k}\right)\right)\right]\right)^{1/4} \\ &\le \exp\left(\frac{(2\Delta)^2}{4k}\left(e^t - 1 - t\right)\right),\end{aligned}$$

with $t = \frac{4n}{k}$ for $i = 1, 2$ and $t = -\frac{4n}{k}$ for $i = 3, 4$. Therefore,

$$\begin{aligned}\chi^2(P_X \| Q_X) + 1 &\le \exp\left(\frac{\Delta^2}{k}\left(2e^{4n/k} + 2e^{-4n/k} - 4\right)\right) \\ &= \exp\left(\frac{4\Delta^2}{k}(\cosh(4n/k) - 1)\right). \qquad (5.110)\end{aligned}$$

The upper bound (5.110) yields the sample complexity $n^*(k, \Delta) \ge \Omega(k \operatorname{arccosh}(1 + \frac{k}{4\Delta^2}))$. □

Now we establish another lower bound for the sample complexity of the **Distinct Elements** problem for sampling without replacement. Since we can simulate sampling with replacement from samples obtained without replacement (see (5.111) for details), it is also a valid lower bound for $n^*(k, \Delta)$ defined in Definition 5.2. On the other hand, as observed in [106, Lemma 3.3] (see also [154, Lemma 5.14]), any estimator $\hat{C}$ for the **Distinct Elements** problem with sampling without replacement leads to an estimator for the **Support Size** problem with slightly worse performance: Suppose we have n i.i.d. samples drawn from a distribution P whose minimum non-zero probability is at least $1/\ell$. Let $\hat{C}_{\text{seen}}$ denote the number of distinct elements in these samples. Equivalently, these samples can be viewed as being generated in two steps: first, we draw k i.i.d. samples from P, whose realizations form an instance of a k-ball urn with $\hat{C}_{\text{seen}}$ distinct colors; next, we draw n samples from this urn without replacement ($n \le k$), which clearly are distributed according to $P^{\otimes n}$. Suppose $\hat{C}_{\text{seen}}$ is close to the actual support size of P. Then applying any algorithm for the **Distinct Elements** problem to these n i.i.d. samples constitutes a good support size estimator. Lemma 5.10 formalizes this intuition.

Lemma 5.10. *Suppose an estimator $\hat{C}$ takes n samples from a k-ball urn ($n \le k$) without replacement and provides an estimation error of less than Δ with probability at least $1 - \delta$. Applying $\hat{C}$ with n i.i.d. samples from any*

distribution P with minimum non-zero mass $1/\ell$ and support size $S(P)$, we have

$$|\hat{C} - S(P)| \le 2\Delta$$

with probability at least $1 - \delta - \binom{\ell}{\Delta}\left(1 - \frac{\Delta}{\ell}\right)^k$.

Proof. Suppose that we take k i.i.d. samples from $P = (p_1, p_2, \dots)$, which form a k-ball urn consisting of C distinct colors. By the union bound,

$$\mathbb{P}[|C - S(P)| \ge \Delta] \le \sum_{\substack{I:|I|=\Delta,\\ p_i \ge \frac{1}{\ell}, i \in I}} \left(1 - \sum_{i \in I} p_i\right)^k \le \binom{\ell}{\Delta}\left(1 - \frac{\Delta}{\ell}\right)^k.$$

Next we take n samples without replacement from this urn and apply the given estimator $\hat{C}$. By assumption, conditioned on any realization of the k-ball urn, $|\hat{C} - C| \le \Delta$ with probability at least $1 - \delta$. Then $|\hat{C} - S(P)| \le 2\Delta$ with probability at least $1 - \delta - \binom{\ell}{\Delta}\left(1 - \frac{\Delta}{\ell}\right)^k$. Marginally, these n samples are identically distributed as n i.i.d. samples from P. □

Combining with the sample complexity of the **Support Size** problem in (5.61), Lemma 5.10 leads to the following lower bound for the **Distinct Elements** problem.

Theorem 5.6. *Fix a sufficiently small constant c. For any $1 \le \Delta \le ck$,*

$$n^*(k, \Delta) \ge \Omega\left(\frac{k}{\log k} \log \frac{k}{\Delta}\right).$$

The same lower bound holds for sampling without replacement.

Proof. By the lower bound of the support size estimation problem obtained in [72, Theorem 2], if $n \le \frac{\alpha\ell}{\log \ell} \log^2 \frac{\ell}{2\Delta}$ and $2\Delta \le c_0\ell$ for some fixed constants $c_0 < \frac{1}{2}$ and α, then for any $\hat{C}$, there exists a distribution P with minimum non-zero mass $1/\ell$ such that $|\hat{C} - S(P)| \le 2\Delta$ with probability at most 0.8. Applying Lemma 5.10 yields that, using n samples without replacement, no estimator can provide an estimation error of Δ with probability 0.9 for an arbitrary k-ball urn, provided $\binom{\ell}{\Delta}\left(1 - \frac{\Delta}{\ell}\right)^k \le 0.1$. Consequently, as long as $2\Delta \le c_0\ell$ and $\binom{\ell}{\Delta}\left(1 - \frac{\Delta}{\ell}\right)^k \le 0.1$, we have

$$n^*(k, \Delta) \ge \frac{\alpha\ell}{\log \ell} \log^2 \frac{\ell}{2\Delta}.$$

The desired lower bound follows from choosing $\ell \asymp \frac{k}{\log(k/\Delta)}$. $\square$

5.3.6 Proof of results in Table 5.1

Below we explain how the sample complexity bounds summarized in Table 5.1 are obtained from various results in Section 5.3.2 and Section 5.3.5:

- The upper bounds are obtained from the worst-case MSE in Section 5.3.2 and the Markov inequality. In particular, the case of $\Delta \leq \sqrt{k}(\log k)^{-\delta}$ follows from the second and the third upper bounds of Theorem 5.4; the case of $\sqrt{k} \leq \Delta \leq k^{0.5+\delta}$ follows from the first upper bound of Theorem 5.4; the case of $k^{1-\delta} \leq \Delta \leq ck$ follows from Theorem 5.3. By monotonicity, we have the $O(k \log\log k)$ upper bound when $\sqrt{k}(\log k)^{-\delta} \leq \Delta \leq \sqrt{k}$, the $O(\frac{k}{\log k})$ upper bound when $\Delta \geq ck$, and the $O(k)$ upper bound when $k^{0.5+\delta} \leq \Delta \leq k^{1-\delta}$.

- The lower bound for $\Delta \leq \sqrt{k}$ follows from Theorem 5.5; the lower bound for $k^{0.5+\delta} \leq \Delta \leq ck$ follows from Theorem 5.6. These further imply the $\Omega(k)$ lower bound for $\sqrt{k} \leq \Delta \leq k^{0.5+\delta}$ by monotonicity.

5.3.7 Connections between various sampling models

As mentioned in Section 5.1.2, four popular sampling models have been introduced in the statistics literature: the multinomial model, the hypergeometric model, the Bernoulli model, and the Poisson model. The connections between those models are explained in detail in this section, as well as relations between the respective sample complexities.

The connections between different models are illustrated in Figure 5.6. Under the Poisson model, the sample size is a Poisson random variable; conditioned on the sample size, the samples are i.i.d. which is identical to the multinomial model. The same relation holds as the Bernoulli model to the hypergeometric model. Given samples $(Y_1, \ldots, Y_n)$ uniformly drawn from a k-ball urn without replacement (hypergeometric model), we can simulate $(X_1, \ldots, X_n)$ drawn with replacement (multinomial model) as follows: for

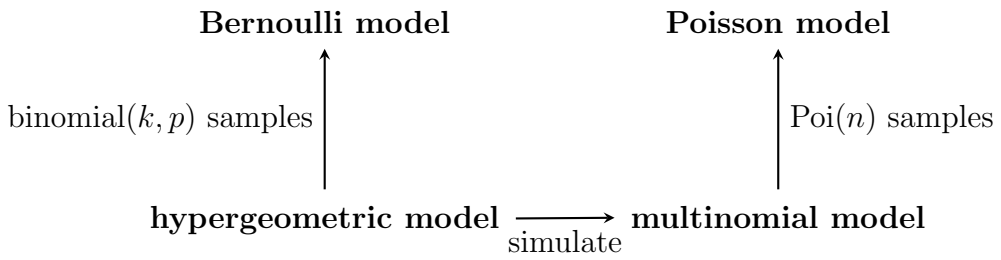


Figure 5.6: Relations between the four sampling models. In particular, hypergeometric (resp. multinomial) model reduces to the Bernoulli (resp. Poisson) model when the sample size is binomial (resp. Poisson) distributed.

each $i = 1, \ldots, n$, let

$$X_i = \begin{cases} Y_i, & \text{with probability } 1 - \frac{i-1}{k}, \\ Y_m, & \text{with probability } \frac{i-1}{k}, \quad m \sim \text{Uniform}([i-1]). \end{cases} \tag{5.111}$$

In view of the connections in Figure 5.6, any estimator constructed for one specific model can be adapted to another. The adaptation from multinomial to hypergeometric model is provided by the simulation in (5.111), and the other direction is given by Lemma 5.10 (without modifying the estimator). The following result provides a recipe for going between fixed and randomized sample size.

Lemma 5.11. *Let N be an $\mathbb{N}$-valued random variable.*

1. *Given any $\hat{C}$ that uses n samples and succeeds with probability at least $1 - \delta$, there exists $\hat{C}'$ using N samples that succeeds with probability at least $1 - \delta - \mathbb{P}[N < n]$.*

2. *Given any $\tilde{C}$ using N samples that succeeds with probability at least $1 - \delta$, there exists $\tilde{C}'$ that uses n samples and succeeds with probability at least $1 - \delta - \mathbb{P}[N > n]$.*

Proof. 1. Denote the samples by $X_1, \ldots, X_N$. Following [106, Lemma 5.3(a)], define $\hat{C}'$ as

$$\hat{C}' = \begin{cases} \hat{C}(X_1, \ldots, X_n), & N \geq n, \\ 0, & N < n. \end{cases}$$

Then $\hat{C}'$ succeeds as long as $N \geq n$ and $\hat{C}$ succeeds, which has probability at least $1 - \delta - \mathbb{P}[N < n]$.

2. Denote the samples by $X_1, \ldots, X_n$. Draw a random variable m from the distribution of N and define $\tilde{C}'$ as

$$\tilde{C}' = \begin{cases} \tilde{C}(X_1, \ldots, X_m), & m \leq n, \\ 0, & m > n. \end{cases}$$

The given estimator $\tilde{C}$ fails with probability $\sum_{j \geq 0} \mathbb{P}[\tilde{C} \text{ fails} | N = j] \mathbb{P}[N = j] \leq \delta$. Consequently, $\sum_{j=0}^{n} \mathbb{P}[\tilde{C} \text{ fails} | N = j] \mathbb{P}[N = j] \leq \delta$. The estimator $\tilde{C}'$ fails with probability at most

$$\sum_{j=0}^{n} \mathbb{P}[\tilde{C} \text{ fails} | m = j] \mathbb{P}[m = j] + \mathbb{P}[m > n] \leq \delta + \mathbb{P}[m > n],$$

which completes the proof.

□

The adaptations of estimators between different sampling models imply the relations of the fundamental limits on the corresponding sample complexities. Extending Definition 5.2, let $n_M^*(k, \Delta, \delta)$, $n_H^*(k, \Delta, \delta)$, $n_B^*(k, \Delta, \delta)$, and $n_P^*(k, \Delta, \delta)$ be the minimum expected sample size under the multinomial, hypergeometric, Bernoulli, and Poisson sampling model, respectively, such that there exists an estimator $\hat{C}$ satisfying $\mathbb{P}[|\hat{C} - C| \geq \Delta] \leq \delta$. Combining Chernoff bounds (see, e.g., [56, Theorem 4.4, 4.5, and 5.4]), we obtain Corollary 5.1, in which the connection between multinomial and Poisson models gives a rigorous justification of the assumption on the Poisson sampling model in Section 5.3.2.

Corollary 5.1. *The following relations hold:*

- n_H^* *versus* n_M^**:*
  - *(a)* $n_H^*(k, \Delta, \delta) \leq n_M^*(k, \Delta, \delta)$*;*
  - *(b)* $n_H^*(k, \Delta, \delta) \leq n \Rightarrow n_M^*(k', 2\Delta, \delta + \binom{k'}{\Delta}(1 - \frac{\Delta}{k'})^k) \leq n$*, for any* $k' \in \mathbb{N}$*. In particular, if* δ *is a constant, then we can choose* $k' = \Theta(k / \log \frac{k}{\Delta})$*.*
- n_P^* *versus* n_M^**:*

(c) $n_P^*(k,\Delta,\delta) \le n \Rightarrow n_M^*(k,\Delta,\delta+(e/4)^n) \le 2n$;

(d) $n_M^*(k,\Delta,\delta) \le n \Rightarrow n_P^*(k,\Delta,\delta+(2/e)^n) \le 2n$.

- n_B^* *versus* n_H^*:

(e) $n_B^*(k,\Delta,\delta) \le n \Rightarrow n_H^*(k,\Delta,\delta+(e/4)^n) \le 2n$;

(f) $n_H^*(k,\Delta,\delta) \le n \Rightarrow n_B^*(k,\Delta,\delta+(2/e)^n) \le 2n$.

5.3.8 Proof of auxiliary lemmas

Proof of Lemma 5.8. For any $z \in \mathbb{C}$, we can represent the forward difference in (2.26) as an integral:

$$\begin{aligned}\Delta^m f(z) &= f(z+m) - \binom{m}{1} f(z+m-1) + \cdots + (-1)^m f(z) \\ &= \int_{[0,1]^m} f^{(m)}(z+x_1+\cdots+x_m)\mathrm{d}x_1\cdots\mathrm{d}x_m.\end{aligned}$$

Therefore,

$$|t_m(z)| = \left|\frac{1}{m!}\Delta^m p_m(z)\right| \le \frac{1}{m!}\sup_{0\le\xi\le m}|p_m^{(m)}(z+\xi)|. \tag{5.112}$$

Recall the definition of p_m in (2.27). Let $p_m(z) = \sum_{l=0}^{2m} a_\ell z^\ell$. Let $z(z-1)\cdots(z-m+1) = \sum_{i=0}^m b_i z^i$ and $(z-M)(z-M-1)\cdots(z-M-m+1) = \sum_{i=0}^m c_i z^i$. Expanding the product and collecting the coefficients yields a simple upper bound:

$$|b_i| \le 2^m(m-1)^{m-i}, \qquad |c_i| \le 2^m(M+m-1)^{m-i} \le 2^m(2M)^{m-i} \le 2^{2m}M^{m-i}.$$

Since $\sum_{\ell=0}^{2m} a_\ell z^\ell = (\sum_{i=0}^m b_i z^i)(\sum_{j=0}^m c_j z^j)$, for $\ell \ge m$,

$$\begin{aligned}|a_\ell| &= \left|\sum_{i=\ell-m}^m b_i c_{\ell-i}\right| \le \sum_{i=\ell-m}^m 2^{3m}(m-1)^{m-i}M^{m-\ell+i} \\ &= 2^{3m}M^{2m-\ell}\sum_{i=\ell-m}^m \left(\frac{m-1}{M}\right)^{m-i} \le m2^{3m}M^{2m-\ell}.\end{aligned}$$

Taking m-th derivative of p_m, we obtain

$$\begin{aligned}|p_m^{(m)}(z)| &= \left|\sum_{j=0}^{m} a_{j+m}\frac{(j+m)!}{j!}z^j\right| \\ &\le \sum_{j=0}^{m}|a_{j+m}M^j|\binom{m+j}{m}m!\left|\frac{z}{M}\right|^j \le m2^{3m}M^m m!(2e)^m\sum_{j=0}^{m}\left|\frac{z}{M}\right|^j \\ &\le m^2 2^{6m}M^m m!\left(\frac{|z|}{M}\vee 1\right)^m = m^2 2^{6m}m!\left(|z|\vee M\right)^m.\end{aligned}$$

Then the desired (5.90) follows from (5.112). □

Proof of Lemma 5.9. The following uniform asymptotic expansions of the Stirling numbers of the first kind was obtained in [155, Theorem 2]:

$$|s(n+1,m+1)| = \begin{cases}\frac{n!}{m!}(\log n+\gamma)^m(1+o(1)), & 1\le m\le\sqrt{\log n},\\ \frac{\Gamma(n+1+R)}{\Gamma(R)R^{m+1}\sqrt{2\pi H}}(1+o(1)), & \sqrt{\log n}\le m\le n-n^{1/3},\\ \binom{n+1}{m+1}(\frac{m+1}{2})^{n-m}(1+o(1)), & n-n^{1/3}\le m\le n,\end{cases}$$

where γ is Euler's constant, R is the unique positive solution to $h'(x)=0$ with $h(x)\triangleq\log\frac{\Gamma(x+n+1)}{\Gamma(x+1)x^m}$, $H=R^2h''(R)$, and all $o(1)$ terms are uniform in m. In the following we consider each range separately and prove the non-asymptotic approximation in (5.95).

Case I. For $1\le m\le\sqrt{\log n}$, Stirling's approximation gives

$$\frac{n!}{m!}(\log n+\gamma)^m = n!\left(\Theta\left(\frac{\log n}{m}\right)\right)^m.$$

Case II. For $n-n^{1/3}\le m\le n$,

$$\begin{aligned}&\binom{n+1}{m+1}\left(\frac{m+1}{2}\right)^{n-m}\\ &= \frac{n!}{m!}\left(\Theta\left(\frac{m}{n-m}\right)\right)^{n-m}\\ &= n!\exp\left(m\left(\frac{n-m}{m}\log\left(\Theta\left(\frac{m}{n-m}\right)\right)-\log\Theta(m)\right)\right)\\ &= n!\left(\Theta\left(\frac{1}{m}\right)\right)^m.\end{aligned}$$

Case III. For $\sqrt{\log n}\le m\le n-n^{1/3}$, note that $h(x)=\sum_{i=1}^{n}\log(x+i)-$

$m \log x$, and thus $H = R^2 h''(R) = m - \sum_{i=1}^{n} \frac{R^2}{(R+i)^2} \leq m$. By [151, Lemma 4.1], $H = \omega(1)$ in this range. Hence,

$$|s(n+1, m+1)| = \frac{\Gamma(n+1+R)}{\Gamma(R) R^{m+1}} (\Theta(1))^m = \frac{n!}{R^m} \frac{\Gamma(n+1+R)}{n! \Gamma(R+1)} (\Theta(1))^m, \tag{5.113}$$

where R is the solution to $x(\frac{1}{x+1} + \cdots + \frac{1}{x+n}) = m$. Bounding the sum by integrals, we have

$$R \log \left(1 + \frac{n}{R+1}\right) \leq m \leq R \log \left(1 + \frac{n}{R}\right).$$

If $\sqrt{\log n} \leq m \leq \frac{n}{e}$, then $R \asymp \frac{m}{\log(n/m)}$, and hence

$$1 \leq \frac{\Gamma(n+1+R)}{n! \Gamma(R+1)} \leq \left(O\left(\frac{n+R}{R}\right)\right)^R = \exp(O(m)).$$

In view of (5.113), we have $|s(n+1, m+1)| = \frac{n!}{(\Theta(R))^m}$, which is exactly (5.95) when $m \leq n/e$. If $n/e \leq m \leq n - n^{1/3}$, then $R \asymp \frac{n^2}{n-m}$, and

$$\begin{aligned}
\frac{1}{R^m} \frac{\Gamma(n+1+R)}{n! \Gamma(R+1)} &= R^{-m} \left(\Theta\left(\frac{n+R}{n}\right)\right)^n \\
&= \exp\left(-m \log \Theta\left(\frac{n^2}{n-m}\right) + n \log \Theta\left(\frac{n}{n-m}\right)\right) \\
&= \exp\left(-m \log \Theta(n) + (n-m) \log \Theta\left(\frac{n}{n-m}\right)\right) \\
&= \exp\left(-m \log \Theta(n)\right).
\end{aligned}$$

Combining (5.113) yields that $|s(n+1, m+1)| = n!(\Theta(\frac{1}{n}))^m$, which coincides with (5.95) since $n \asymp m$ is this range. □

Part II

Learning Gaussian Mixtures

CHAPTER 6

A FRAMEWORK FOR LEARNING MIXTURE MODELS

Learning mixture models has a long history in statistics and computer science with early contributions dating back to Pearson [11] and recent renewed interest in latent variable models. In a k-component mixture model from a family of distributions $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, each observation is distributed as

$$X \sim \sum_{i=1}^{k} w_i P_{\theta_i}.$$

Here w_i is the mixing weight such that $w_i \geq 0$ and $\sum_i w_i = 1$, $\theta_i \in \Theta$ is the parameter of the i^{th} component. Equivalently, we can write the distribution of an observation X as

$$X \sim P_U = \int P_\theta \mathrm{d}\nu(\theta), \tag{6.1}$$

where $\nu = \sum_{i=1}^{k} w_i \delta_{\theta_i}$ denotes the *mixing distribution* and $U \sim \nu$ is referred to as the latent variable.

Generally speaking, there are three common formulations of learning mixture models:

- **Parameter estimation**: estimate the parameter θ_i's and the weights w_i's up to a global permutation.
- **Density estimation**: estimate the probability density function of the mixture model under certain loss such as L_2 or Hellinger distance. This task is further divided into the cases of *proper* and *improper* learning, depending on whether the estimate is required to be a mixture of distributions in $\mathcal{P}$ or not; in the latter case, there is more flexibility in designing the estimator but less interpretability.
- **Clustering**: estimate the latent variable of each sample (i.e. U_i, if the

ith sample is distributed as P_{U_i}) with a small misclassification rate.

It is clear that to ensure the possibility of clustering it is necessary to impose certain separation conditions between the clusters; however, as far as estimation is concerned, both parametric and non-parametric, no separation condition should be needed and one can obtain accurate estimates of the parameters even when clustering is impossible. Furthermore, one should be able to learn from the data the order of the mixture model, that is, the number of components. However, in the present literature, most of the estimation procedures with finite sample guarantees are either clustering-based, or rely on separation conditions in the analysis (e.g. [156, 157, 158]). Bridging this conceptual divide is one of the main motivations of the present chapter.

6.1 Estimating the mixing distribution

Following the framework proposed in [159, 160], in this chapter we consider the estimation of the mixing distribution, rather than estimating the parameters of each component. The main benefits of this formulation include the following:

- Assumption-free: to recover individual components it is necessary to impose certain assumptions to ensure identifiability, such as lower bounds on the mixing weights and separations between components, none of which is needed for estimating the mixing distribution. Furthermore, under the usual assumption such as separation conditions, statistical guarantees on estimating the mixing distribution can be naturally translated to those for estimating the individual parameters.

- Inference on the number of components: this formulation allows us to deal with misspecified models and estimate the order of the mixture model.

In this framework, a meaningful and flexible loss function for estimating the mixing distribution is the 1-*Wasserstein distance* defined by

$$W_1(\nu, \nu') \triangleq \inf\{\mathbb{E}[\|X - Y\|] : X \sim \nu, Y \sim \nu'\}, \tag{6.2}$$

where the infimum is taken over all couplings, i.e., joint distributions of X and Y which are marginally distributed as ν and ν' respectively. In one dimension, the W_1 distance coincides with the L_1-distance between the cumulative distribution functions (CDFs) [161]. This is a natural criterion, which is not too stringent to yield trivial result (e.g. the Kolmogorov-Smirnov (KS) distance[1]) and, at the same time, strong enough to provide meaningful guarantees on the means and weights. In fact, the commonly used criterion $\min_\Pi \sum_i \|\theta_i - \hat{\theta}_{\Pi(i)}\|$ over all permutations Π is precisely (k times) the Wasserstein distance between two equally weighted distributions [161].

Furthermore, we can obtain statistical guarantees on the support sets and weights of the estimated mixing distribution under the usual assumptions in literature [162, 29, 31] that include separation between the means and lower bound on the weights. See Section 6.2 for a detailed discussion. We highlight the following result, phrased in terms of the parameter estimation error up to a permutation.

Lemma 6.1. *Let*

$$\nu = \sum_{i=1}^{k} w_i \delta_{\theta_i}, \quad \hat{\nu} = \sum_{i=1}^{k} \hat{w}_i \delta_{\hat{\theta}_i}.$$

Suppose that

$$\begin{aligned} \epsilon &= W_1(\nu, \hat{\nu}), \\ \epsilon_1 &= \min\{\|\theta_i - \theta_j\|, \|\hat{\theta}_i - \hat{\theta}_j\| : 1 \le i < j \le k\}, \\ \epsilon_2 &= \min\{w_i, \hat{w}_i : i \in [k]\}. \end{aligned}$$

If $\epsilon < \epsilon_1\epsilon_2/4$, then, there exists a permutation Π such that

$$\|\theta_i - \hat{\theta}_{\Pi(i)}\| \le \epsilon/\epsilon_2, \quad |w_i - \hat{w}_{\Pi(i)}| \le 2\epsilon/\epsilon_1, \quad \forall\ i.$$

6.2 Wasserstein distance

A central quantity in the theory of optimal transportation, the Wasserstein distance is the minimum cost of mapping one distribution to another. In this part, we will be mainly concerned with the 1-Wasserstein distance defined in

[1] Consider two mixing distributions δ_0 and δ_ϵ with arbitrarily small ϵ, whose KS distance is always one.

(6.2), which can be equivalently expressed, through the Kantorovich duality [161], as

$$W_1(\nu, \nu') = \sup\{\mathbb{E}_\nu[\varphi] - \mathbb{E}_{\nu'}[\varphi] : \varphi \text{ is 1-Lipschitz}\}. \tag{6.3}$$

The optimal coupling in (6.2) has many equivalent characterization [163] but is often difficult to compute analytically in general. Nevertheless, the situation is especially simple for distributions on the real line, where the quantile coupling is known to be optimal and hence

$$W_1(\nu, \nu') = \int |F_\nu(t) - F_{\nu'}(t)| \mathrm{d}t, \tag{6.4}$$

where F_ν and $F_{\nu'}$ denote the CDFs of ν and ν', respectively. Both (6.3) and (6.4) provide convenient characterizations to bound the Wasserstein distance in Chapter 7.

As previously mentioned in Section 6.1, two discrete distributions close in the Wasserstein distance have similar support sets and weights. This is made precise in Lemma 6.2 and 6.3 next. In Lemma 6.2 the distance between two support sets is in terms of the Hausdorff distance defined as

$$d_H(S, S') = \max\left\{\sup_{x\in S}\inf_{x'\in S'}\|x - x'\|, \sup_{x'\in S'}\inf_{x\in S}\|x - x'\|\right\}. \tag{6.5}$$

Lemma 6.2. *Suppose ν and ν' are discrete distributions supported on S and S', respectively. Let $\epsilon = \min\{\nu(x) : x \in S\} \wedge \min\{\nu'(x) : x \in S'\}$. Then,*

$$d_H(S, S') \le W_1(\nu, \nu')/\epsilon.$$

Proof. For any coupling P_{XY} such that $X \sim \nu$ be $Y \sim \nu'$,

$$\begin{aligned}\mathbb{E}|X - Y| &= \sum_x \mathbb{P}[X = x]\mathbb{E}[|X - Y||X = x] \\ &\ge \sum_x \epsilon \cdot \inf_{x'\in S'}\|x - x'\| \ge \epsilon \cdot \sup_{x\in S}\inf_{x'\in S'}\|x - x'\|.\end{aligned}$$

Interchanging X and Y completes the proof. □

Lemma 6.3. *For any $\delta > 0$,*

$$\nu(x) - \nu'([x \pm \delta]) \le W_1(\nu, \nu')/\delta,$$
$$\nu'(x) - \nu([x \pm \delta]) \le W_1(\nu, \nu')/\delta.$$

Proof. Using the optimal coupling P^*_{XY} such that $X \sim \nu$ be $Y \sim \nu'$, applying Markov inequality yields that

$$\mathbb{P}[|X - Y| > \delta] \le \mathbb{E}|X - Y|/\delta = W_1(\nu, \nu')/\delta.$$

By Strassen's theorem (see [161, Corollary 1.28]), for any Borel set B, we have $\nu(B) \le \nu'(B^\delta) + W_1(\nu, \nu')/\delta$ and $\nu'(B) \le \nu(B^\delta) + W_1(\nu, \nu')/\delta$, where $B^\delta \triangleq \{x : \inf_{y \in B} |x - y| \le \delta\}$ denotes the δ-fattening of B. The conclusion follows by considering a singleton $B = \{x\}$. □

Lemmas 6.2 and 6.3 together yield a bound on the parameter estimation error (up to a permutation) in terms of the Wasserstein distance, which was previously given in Lemma 6.1:

Proof. Denote the support sets of ν and ν' by $S = \{\theta_1, \ldots, \theta_k\}$ and $S' = \{\hat{\theta}_1, \ldots, \hat{\theta}_k\}$, respectively. Applying Lemma 6.2 yields that $d_H(S, S') \le \epsilon/\epsilon_2$, which is less than $\epsilon_1/4$ by the assumption $\epsilon < \epsilon_1\epsilon_2/4$. Since $\|\theta_i - \theta_j\| \ge \epsilon$ for every $i \ne j$, then there exists a permutation Π such that

$$\|\theta_i - \hat{\theta}_{\Pi(i)}\| \le \epsilon/\epsilon_2, \quad \forall\ i.$$

Applying Lemma 6.3 twice with $\delta = \epsilon/2$, $x = \theta_i$ and $x = \hat{\theta}_{\Pi(i)}$, respectively, we obtain that

$$w_i - \hat{w}_{\Pi(i)} \le 2\epsilon/\epsilon_1, \quad \hat{w}_{\Pi(i)} - w_i \le 2\epsilon/\epsilon_1.$$ □

CHAPTER 7

MOMENT COMPARISON THEOREMS

Moment comparison is a classical topic in the probability theory. Classical moments comparison theorems aim to show convergence of distributions by comparing a *growing* number of moments. For example, Chebyshev's theorem states if $\mathbf{m}_r(\pi) = \mathbf{m}_r(N(0,1))$, then (see [164, Theorem 2])

$$\sup_{x\in\mathbb{R}} |F_\pi(x) - \Phi(x)| \leq \sqrt{\frac{\pi}{2r}},$$

where F_π and Φ denote the CDFs of π and $N(0,1)$, respectively. The notation will be used throughout this chapter. For two compactly supported distributions, the above estimate can be sharpened to $O(\frac{\log r}{r})$ [165]. In contrast, in the context of estimating finite mixtures we are dealing with finitely supported mixing distributions, which can be identified by a *fixed* number of moments. However, with finitely many samples, it is impossible to exactly determine the moments, and measuring the error in the KS distance is too much to ask (see Section 6.1). It turns out that W_1-distance is a suitable metric for this purpose, and the closeness of moments does imply the closeness of distribution in the W_1 distance, which is the integrated difference (L_1-distance) between two CDFs as opposed the uniform error (L_∞-distance).

7.1 Wasserstein distance between discrete distributions

A discrete distribution with k atoms has $2k-1$ free parameters. Therefore it is reasonable to expect that it can be uniquely determined by its first $2k-1$ moments. Indeed, we have the following simple identifiability results for discrete distributions.

Lemma 7.1. *Let ν and ν' be distributions on the real line.*

1. *If ν and ν' are both k-atomic, then $\nu = \nu'$ if and only if $\mathbf{m}_{2k-1}(\nu) = \mathbf{m}_{2k-1}(\nu')$.*

2. *If ν is k-atomic, then $\nu = \nu'$ if and only if $\mathbf{m}_{2k}(\nu) = \mathbf{m}_{2k}(\nu')$.*

Proof. We only need to prove the "if" part. We prove this lemma using the apparatus of interpolating polynomials.

1. Denote the union of the support sets of ν and ν' by S. Here S is of size at most $2k$. For any $t \in \mathbb{R}$, there exists a polynomial P of degree at most $2k-1$ to interpolate $x \mapsto \mathbf{1}_{\{x \le t\}}$ on S. Since $m_i(\nu) = m_i(\nu')$ for $i = 1, ..., 2k-1$, we have

$$F_\nu(t) = \mathbb{E}_\nu[\mathbf{1}_{\{X \le t\}}] = \mathbb{E}_\nu[P(X)] = \mathbb{E}_{\nu'}[P(X)] = \mathbb{E}_{\nu'}[\mathbf{1}_{\{X \le t\}}] = F_{\nu'}(t).$$

2. Denote the support set of ν by $S' = \{x_1, \ldots, x_k\}$. Let $Q(x) = \prod_i (x - x_i)^2$, a non-negative polynomial of degree $2k$. Since $m_i(\nu) = m_i(\nu')$ for $i = 1, ..., 2k$, we have

$$\mathbb{E}_{\nu'}[Q(X)] = \mathbb{E}_\nu[Q(X)] = 0.$$

   Therefore, ν' is also supported on S' and thus is k-atomic. The conclusion follows from the first case of Lemma 7.1. □

In the context of statistical estimation, we only have access to samples and noisy estimates of moments. To solve the inverse problems from moments to distributions, our theory relies on the following stable version of the identifiability in Lemma 7.1, which show that closeness of moments implies closeness of distributions in Wasserstein distance. In the sequel we refer to Propositions 7.1 and 7.2 as moment comparison theorems.

Proposition 7.1. *Let ν and ν' be k-atomic distributions supported on $[-1, 1]$. If $|m_i(\nu) - m_i(\nu')| \le \delta$ for $i = 1, \ldots, 2k-1$, then*

$$W_1(\nu, \nu') \le O\left(k\delta^{\frac{1}{2k-1}}\right).$$

Proposition 7.2. *Let ν be a k-atomic distribution supported on $[-1, 1]$. If $|m_i(\nu) - m_i(\nu')| \le \delta$ for $i = 1, \ldots, 2k$, then*

$$W_1(\nu, \nu') \le O\left(k\delta^{\frac{1}{2k}}\right).$$

Remark 7.1. The exponents in Proposition 7.1 and 7.2 are optimal. To see this, we first note that the number of moments needed for identifiability in Lemma 7.1 cannot be reduced:

1. Given any $2k$ distinct points, there exist two k-atomic distributions with disjoint support sets but identical first $2k-2$ moments (see Lemma 8.24).
2. Given any continuous distribution, its k-point Gauss quadrature is k-atomic and have identical first $2k-1$ moments (see Section 2.3).

By the first observation, there exists two k-atomic distributions ν and ν' such that

$$\begin{aligned} m_i(\nu) &= m_i(\nu'), \quad i = 1, \ldots, 2k-2, \\ |m_{2k-1}(\nu) - m_{2k-1}(\nu')| &= c_k, \quad W_1(\nu, \nu') = d_k, \end{aligned}$$

where c_k and d_k are strictly positive constants that depend on k. Let $\tilde{\nu}$ and $\tilde{\nu}'$ denote the distributions of ϵX and $\epsilon X'$ such that $X \sim \nu$ and $X' \sim \nu'$, respectively. Then, we have

$$\max_{i \in [2k-1]} |m_i(\tilde{\nu}) - m_i(\tilde{\nu})| = \epsilon^{2k-1} c_k, \quad W_1(\tilde{\nu}, \tilde{\nu}') = \epsilon d_k.$$

This concludes the tightness of the exponent in Proposition 7.1. Similarly, the exponent in Proposition 7.2 is also tight using the second observation.

When the atoms of the discrete distributions are separated, we have the following adaptive version of the moment comparison theorems (cf. Propositions 7.1 and 7.2).

Proposition 7.3. *Suppose both ν and ν' are supported on a set of ℓ atoms in $[-1, 1]$, and each atom is at least γ away from all but at most ℓ' other atoms. Let $\delta = \max_{i \in [\ell-1]} |m_i(\nu) - m_i(\nu')|$. Then,*

$$W_1(\nu, \nu') \le \ell \left(\frac{\ell 4^{\ell-1} \delta}{\gamma^{\ell-\ell'-1}} \right)^{\frac{1}{\ell'}}.$$

Proposition 7.4. *Suppose ν is supported on k atoms in $[-1, 1]$ and any $t \in \mathbb{R}$ is at least γ away from all but k' atoms. Let $\delta = \max_{i \in [2k]} |m_i(\nu) - m_i(\nu')|$. Then,*

$$W_1(\nu, \nu') \le 8k \left(\frac{k 4^{2k} \delta}{\gamma^{2(k-k')}} \right)^{\frac{1}{2k'}}.$$

7.1.1 Proofs

First we prove Proposition 7.5, which is slightly stronger than Proposition 7.1. We provide three proofs: the first two are based on the primal (coupling) formulation of W_1 distance (6.4), and the third proof uses the dual formulation (6.3). Specifically,

- The first proof uses polynomials to interpolate step functions, whose expected values are the CDFs. The closeness of moments imply the closeness of distribution functions and thus, by (6.4), a small Wasserstein distance. Similar idea applies to the proof of Proposition 7.2 later.
- The second proof finds a polynomial that preserves the sign of the difference between two CDFs, and then relate the Wasserstein distance to the integral of that polynomial. Similar idea is used in [30, Lemma 20] which uses a polynomial that preserves the sign of the difference between two density functions.
- The third proof uses polynomials to approximate 1-Lipschitz functions, whose expected values are related to the Wasserstein distance via the dual formulation (6.3).

Proposition 7.5. *Let ν and ν' be discrete distributions supported on ℓ atoms in $[-1, 1]$. If*

$$|m_i(\nu) - m_i(\nu')| \le \delta, \quad i = 1, \dots, \ell - 1, \tag{7.1}$$

then

$$W_1(\nu, \nu') \le O\left(\ell \delta^{\frac{1}{\ell-1}}\right).$$

First proof of Proposition 7.5. Suppose ν and ν' are supported on

$$S = \{t_1, \dots, t_\ell\}, \quad t_1 < t_2 < \cdots < t_\ell. \tag{7.2}$$

Then, using the integral representation (6.4), the W_1 distance reduces to

$$W_1(\nu, \nu') = \sum_{r=1}^{\ell-1} |F_\nu(t_r) - F_{\nu'}(t_r)| \cdot |t_{r+1} - t_r|. \tag{7.3}$$

For each r, let $f_r(x) = \mathbf{1}_{\{x \le t_r\}}$, and P_r be the unique polynomial of degree $\ell - 1$ to interpolate f_r on S. In this way we have $f_r = P_r$ almost surely under

both ν and ν', and thus

$$|F_\nu(t_r) - F_{\nu'}(t_r)| = |\mathbb{E}_\nu f_r - \mathbb{E}_{\nu'} f_r| = |\mathbb{E}_\nu P_r - \mathbb{E}_{\nu'} P_r|. \tag{7.4}$$

P_r can expressed using Newton formula (2.7) as

$$P_r(x) = 1 + \sum_{i=r+1}^{\ell} f_r[t_1, \ldots, t_i] g_{i-1}(x), \tag{7.5}$$

where $g_r(x) = \prod_{j=1}^{r}(x - t_j)$ and we used $f_r[t_1, \ldots, t_i] = 0$ for $i = 1, \ldots, r$. In (7.5), the absolute values of divided differences are obtained in Lemma 7.2:

$$|f_r[t_1, \ldots, t_i]| \le \frac{\binom{i-2}{r-1}}{(t_{r+1} - t_r)^{i-1}}. \tag{7.6}$$

In the summation of (7.5), let $g_{i-1}(x) = \sum_{j=0}^{i-1} a_j x^j$. Since $|t_j| \le 1$ for every j, we have $\sum_{j=0}^{i-1} |a_j| \le 2^{i-1}$ (see Lemma 7.3). Applying (7.1) yields that

$$|\mathbb{E}_\nu[g_{i-1}] - \mathbb{E}_{\nu'}[g_{i-1}]| \le \sum_{j=1}^{i-1} |a_j|\delta \le 2^{i-1}\delta. \tag{7.7}$$

Then we obtain from (7.4) and (7.5) that

$$|F_\nu(t_r) - F_{\nu'}(t_r)| \le \sum_{i=r+1}^{\ell} \frac{\binom{i-2}{r-1} 2^{i-1}\delta}{(t_{r+1} - t_r)^{i-1}} \le \frac{\ell 4^{\ell-1}\delta}{(t_{r+1} - t_r)^{\ell-1}}. \tag{7.8}$$

Also, $|F_\nu(t_r) - F_{\nu'}(t_r)| \le 1$ trivially. Therefore,

$$W_1(\nu, \nu') \le \sum_{r=1}^{\ell-1} \left(\frac{\ell 4^{\ell-1}\delta}{(t_{r+1} - t_r)^{\ell-1}} \wedge 1 \right) \cdot |t_{r+1} - t_r| \le 4e(\ell - 1)\delta^{\frac{1}{\ell-1}}, \tag{7.9}$$

where we used $\max\{\frac{\alpha}{x^{\ell-2}} \wedge x : x > 0\} = \alpha^{\frac{1}{\ell-1}}$ and $x^{\frac{1}{x-1}} \le e$ for $x \ge 1$. □

Second proof of Proposition 7.5. Suppose on the contrary that

$$W_1(\nu, \nu') \ge C\ell\delta^{\frac{1}{\ell-1}}, \tag{7.10}$$

for some absolute constant C. We will show that $\max_{i \in [\ell-1]} |m_i(\nu) - m_i(\nu')| \ge \delta$. We continue to use S in (7.2) to denote the support of ν and ν'. Let

$\Delta F(t) = F_\nu(t) - F_{\nu'}(t)$ denote the difference between two CDFs. Using (7.3), there exists $r \in [\ell - 1]$ such that

$$|\Delta F(t_r)| \cdot |t_{r+1} - t_r| \geq C\delta^{\frac{1}{\ell-1}}. \tag{7.11}$$

We first construct a polynomial L that preserves the sign of ΔF. To this end, let $S' = \{s_1, \ldots, s_m\} \subseteq S$ such that $t_1 = s_1 < s_2 < \cdots < s_m = t_\ell$ be the set of points where ΔF changes sign, i.e., $\Delta F(x)\Delta F(y) \leq 0$ for every $x \in (s_i, s_{i+1})$, $y \in (s_{i+1}, s_{i+2})$, for every i. Let $L(x) \in \pm\prod_{i=2}^{m-1}(x - s_i)$ be a polynomial of degree at most $\ell - 2$ that also changes sign on S' such that

$$\Delta F(x)L(x) \geq 0, \quad t_1 \leq x \leq t_\ell.$$

Consider the integral of the above positive function. Applying integral by parts, and using $\Delta F(t_\ell) = \Delta F(t_1) = 0$ yields that

$$\int_{t_1}^{t_\ell} \Delta F(x)L(x)\mathrm{d}x = -\int_{t_1}^{t_\ell} P(x)\mathrm{d}\Delta F(x) = \mathbb{E}_{\nu'}[P(X)] - \mathbb{E}_\nu[P(X)], \tag{7.12}$$

where $P(x)$ is a polynomial of degree at most $\ell - 1$ such that $P'(x) = L(x)$. If we write $L(x) = \sum_{j=0}^{\ell-2} a_j x^j$, then $P(x) = \sum_{j=0}^{\ell-2} \frac{a_j}{j+1} x^{j+1}$. Since $|s_j| \leq 1$ for every j, we have $\sum_{j=0}^{\ell-2} |a_j| \leq 2^{\ell-2}$ (see Lemma 7.3), and thus $\sum_{j=0}^{\ell-2} \frac{|a_j|}{j+1} \leq 2^{\ell-2}$. Hence,

$$|\mathbb{E}_{\nu'}[P(X)] - \mathbb{E}_\nu[P(X)]| \leq 2^{\ell-2} \max_{i\in[\ell-1]} |m_i(\nu) - m_i(\nu')|. \tag{7.13}$$

Since $\Delta F(x)L(x)$ is always non-negative, applying (7.11) to (7.12) yields that

$$|\mathbb{E}_{\nu'}[P(X)] - \mathbb{E}_\nu[P(X)]| \geq \int_{t_r}^{t_{r+1}} |\Delta F(x)L(x)|\mathrm{d}x \geq \frac{C\delta^{\frac{1}{\ell-1}}}{|t_{r+1} - t_r|} \int_{t_r}^{t_{r+1}} |L(x)|\mathrm{d}x. \tag{7.14}$$

Recall that $|L(x)| = \prod_{i=2}^{m-1} |x - s_i|$. Then for $x \in (t_r, t_{r+1})$, we have $|x - s_i| \geq x - t_r$ if $s_i \leq t_r$, and $|x - s_i| \geq t_{r+1} - x$ if $s_i \geq t_{r+1}$. Hence,

$$|L(x)| \geq (t_{r+1} - x)^a (x - t_r)^b,$$

for some $a, b \in \mathbb{N}$ such that $a, b \geq 1$ and $a + b \leq \ell - 2$. The integral of the

right-hand side of the above inequality can be expressed as (see [52, 6.2.1])

$$\int_{t_r}^{t_{r+1}} (t_{r+1}-x)^a (x-t_r)^b \mathrm{d}x = \frac{(t_{r+1}-t_r)^{a+b+1}}{(a+1)\binom{a+b+1}{b}}.$$

Since $|t_{r+1}-t_r| \geq |\Delta F(t_r)| \cdot |t_{r+1}-t_r| \geq C\delta^{\frac{1}{\ell-1}}$ and $\binom{a+b+1}{b} \leq 2^{a+b+1}$, and $a+b+1 \leq \ell-1$, we obtain from (7.14) that

$$|\mathbb{E}_{\nu'}[P(X)] - \mathbb{E}_{\nu}[P(X)]| \geq \delta \frac{(C/2)^{\ell-1}}{\ell}. \tag{7.15}$$

We obtain from (7.13) and (7.15) that

$$\max_{i\in[\ell-1]} |m_i(\nu) - m_i(\nu')| \geq \delta \frac{(C/4)^{\ell-1}}{\ell}. \qquad \square$$

Third proof of Proposition 7.5. We continue to use S in (7.2) to denote the support of ν and ν'. For any 1-Lipschitz function f, $\mathbb{E}_\nu f$ and $\mathbb{E}_{\nu'} f$ only pertain to function values $f(t_1), \ldots, f(t_\ell)$, which can be interpolated by a polynomial of degree $\ell-1$. However, the coefficients of the interpolating polynomial can be arbitrarily large.[1] To fix this issue, we slightly modify the function f on S to $\tilde{f}$, and then interpolate $\tilde{f}$ with bounded coefficients. In this way we have

$$|\mathbb{E}_\nu f - \mathbb{E}_{\nu'} f| \leq 2 \max_{x\in\{t_1,\ldots,t_\ell\}} |\tilde{f}(x) - f(x)| + |\mathbb{E}_\nu P - \mathbb{E}_{\nu'} P|.$$

To this end, we define the values of $\tilde{f}$ recursively by

$$\tilde{f}(t_1) = f(t_1), \quad \tilde{f}(t_i) = \tilde{f}(t_{i-1}) + (f(t_i) - f(t_{i-1}))\mathbf{1}_{\{t_i - t_{i-1} > \tau\}}, \tag{7.16}$$

where $\tau \leq 2$ is a parameter we will optimize later. From the above definition $|\tilde{f}(x) - f(x)| \leq \tau\ell$ for $x \in S$. The interpolating polynomial P can be expressed using Newton formula (2.7) as

$$P(x) = \sum_{i=1}^{\ell} \tilde{f}[t_1, \ldots, t_i] g_{i-1}(x),$$

where $g_r(x) = \prod_{j=1}^{r}(x-t_j)$ such that $|\mathbb{E}_\nu[g_r] - \mathbb{E}_{\nu'}[g_r]| \leq 2^r\delta$ by (7.7) for

[1] For example, the polynomial to interpolate $f(-\epsilon) = f(\epsilon) = \epsilon, f(\epsilon) = 0$ is $P(x) = x^2/\epsilon$.

$r \le \ell - 1$. Since f is 1-Lipschitz, we have $|\tilde{f}[t_i, t_{i+1}]| \le 1$ for every i. Higher-order divided differences are recursively evaluated by (2.8). We now prove

$$\tilde{f}[t_i, \ldots, t_{i+j}] \le (2/\tau)^{j-1}, \ \forall\ i, j, \tag{7.17}$$

by induction on j. Assume (7.17) holds for every i and some fixed j. The recursion (2.8) gives

$$\tilde{f}[t_i, \ldots, t_{i+j+1}] = \frac{\tilde{f}[t_{i+1}, \ldots, t_{i+j+1}] - \tilde{f}[t_i, \ldots, t_{i+j}]}{t_{i+j+1} - t_i}.$$

If $t_{i+j+1} - t_i < \tau$, then $\tilde{f}[t_i, \ldots, t_{i+j+1}] = 0$ by (7.16); otherwise, $\tilde{f}[t_i, \ldots, t_{i+j+1}] \le (\frac{2}{\tau})^j$ by triangle inequality. Using (7.17), we obtain that

$$|\mathbb{E}_\nu f - \mathbb{E}_{\nu'} f| \le 2\tau\ell + \sum_{i=2}^{\ell} \left(\frac{2}{\tau}\right)^{i-2} 2^{i-1}\delta \le 2\ell \left(\tau + \frac{4^{\ell-2}}{\tau^{\ell-2}}\delta\right).$$

The conclusion follows by letting $\tau = 4\delta^{\frac{1}{\ell-1}}$. □

The proof of Proposition 7.2 uses a similar idea as the first proof of Proposition 7.5 to approximate step functions for all values of ν and ν'; however, this is clearly impossible for non-discrete ν'. For this reason, we turn from interpolation to majorization. A classical method to bound a distribution function by moments is to construct two polynomials that majorizes and minorizes a step function, respectively. Then the expectations of these two polynomials provide a sandwich bound for the distribution function. This idea is used, for example, in the proof of Chebyshev-Markov-Stieltjes inequality (cf. [45, Theorem 2.5.4]).

Proof of Proposition 7.2. Suppose ν is supported on $x_1 < x_2 < \ldots < x_k$. Fix $t \in \mathbb{R}$ and let $f_t(x) = \mathbf{1}_{\{x \le t\}}$. Suppose $x_m < t < x_{m+1}$. Similar to Example 2.2, we construct polynomial majorant and minorant using Hermite interpolation. To this end, let P_t and Q_t be the unique degree-$2k$ polynomials to interpolate f_t with values in Table 7.1. As a consequence of Rolle's theorem, $P_t \ge f_t \ge Q_t$ (cf. [45, p. 65], and an illustration in Figure 7.1). Using Lagrange formula of Hermite interpolation [41, pp. 52–53], P_t and Q_t

Table 7.1: Interpolation values of f_t.

	x_1	...	x_m	t	x_{m+1}	...	x_k
P	1	...	1	1	0	...	0
P'	0	...	0	any	0	...	0
Q	1	...	1	0	0	...	0
Q'	0	...	0	any	0	...	0

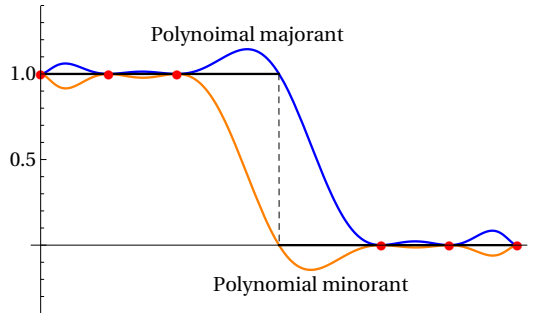


Figure 7.1: Polynomial majorant P_t and minorant Q_t that coincide with the step function on 6 red points. The polynomials are of degree 12, obtained by Hermite interpolation in Section 2.2.

differ by

$$P_t(x) - Q_t(x) = R_t(x) \triangleq \prod_i \left(\frac{x - x_i}{t - x_i}\right)^2.$$

The sandwich bound for f_t yields a sandwich bound for the CDFs:

$$\begin{aligned}
&\mathbb{E}_{\nu'}[Q_t] \le F_{\nu'}(t) \le \mathbb{E}_{\nu'}[P_t] = \mathbb{E}_{\nu'}[Q_t] + \mathbb{E}_{\nu'}[R_t], \\
&\mathbb{E}_{\nu}[Q_t] \le F_{\nu}(t) \le \mathbb{E}_{\nu}[P_t] = \mathbb{E}_{\nu}[Q_t].
\end{aligned}$$

Then the CDFs differ by

$$\begin{aligned}
|F_\nu(t) - F_{\nu'}(t)| \le (f(t) + g(t)) \wedge 1 \le f(t) \wedge 1 + g(t) \wedge 1, \qquad (7.18) \\
f(t) \triangleq |\mathbb{E}_{\nu'}[Q_t] - \mathbb{E}_{\nu}[Q_t]|, \quad g(t) \triangleq \mathbb{E}_{\nu'}[R_t].
\end{aligned}$$

The conclusion will be obtained from the integral of CDF difference using (6.4). Since R_t is almost surely zero under ν, we also have $g(t) = |\mathbb{E}_{\nu'}[R_t] - \mathbb{E}_{\nu}[R_t]|$. Similar to (7.7), we obtain that

$$g(t) = |\mathbb{E}_{\nu'}[R_t] - \mathbb{E}_{\nu}[R_t]| \le \frac{2^{2k}\delta}{\prod_{i=1}^k (t - x_i)^2}.$$

Hence,

$$\int (g(t) \wedge 1) \mathrm{d}t \leq \int \left(\frac{2^{2k}\delta}{\prod_{i=1}^{k}(t-x_i)^2} \wedge 1 \right) \mathrm{d}t \leq 16k\delta^{\frac{1}{2k}}, \tag{7.19}$$

where the last inequality is proved in Lemma 8.23.

Next we analyze $f(t)$. The polynomial Q_t (and also P_t) can be expressed using Newton formula (2.7) as

$$Q_t(x) = 1 + \sum_{i=2m+1}^{2k+1} f_t[t_1, \ldots, t_i] g_{i-1}(x), \tag{7.20}$$

where $t_1, \ldots, t_{2k+1}$ denotes the expanded sequence

$$x_1, x_1, \ldots, x_m, x_m, t, x_{m+1}, x_{m+1}, \ldots, x_k, x_k$$

obtained by (2.15), $g_r(x) = \prod_{j=1}^{r}(x - t_j)$, and we used $f_t[t_1, \ldots, t_i] = 0$ for $i = 1, \ldots, 2m$. In (7.20), the absolute values of divided differences are obtained in Lemma 7.2:

$$f_t[t_1, \ldots, t_i] \leq \frac{\binom{i-2}{2m-1}}{(t - x_m)^{i-1}}.$$

Using (7.20), and applying the upper bound for $|\mathbb{E}_\nu[g_{i-1}] - \mathbb{E}_{\nu'}[g_{i-1}]|$ in (7.7), we obtain that

$$f(t) = |\mathbb{E}_{\nu'}[Q_t] - \mathbb{E}_\nu[Q_t]| \leq \sum_{i=2m+1}^{2k+1} \frac{\binom{i-2}{2m-1} 2^{i-1}\delta}{(t - x_m)^{i-1}} \leq \frac{k 4^{2k}\delta}{(t - x_m)^{2k}},$$
$$\forall\ x_m < t < x_{m+1},\ m \geq 1.$$

If $t < x_1$, then $Q_t = 0$ and thus $f(t) = 0$. Then, analogous to (7.19), we obtain that

$$\int (f(t) \wedge 1) \mathrm{d}t \leq 16k\delta^{\frac{1}{2k}}. \tag{7.21}$$

Using (7.19) and (7.21), the conclusion follows by applying (7.18) to the integral representation of Wasserstein distance (6.4). □

Proof of Proposition 7.3. The proof is analogous to the first proof of Proposition 7.5, apart from a more careful analysis of polynomial coefficients. When

each atom is at least γ away from all but at most ℓ' other atoms, the left-hand side of (7.8) is upper bounded by

$$|F_\nu(t_r) - F_{\nu'}(t_r)| \le \frac{\ell 4^{\ell-1}\delta}{(t_{r+1}-t_r)^{\ell'}\gamma^{\ell-\ell'-1}}.$$

The remaining proof is similar. □

Proof of Proposition 7.4. Similar to the proof of Proposition 7.3, this proof is analogous to Proposition 7.2 apart from a more careful analysis of polynomial coefficients. When every $t \in \mathbb{R}$ is at least γ away from all but k' atoms, the left-hand sides of (7.19) and (7.21) are upper bounded by

$$\int (g(t) \wedge 1)\mathrm{d}t \le 4k \left(\frac{2^{2k}\delta}{\gamma^{2(k-k')}}\right)^{1/(2k')},$$
$$\int (f(t) \wedge 1)\mathrm{d}t \le 4k \left(\frac{k4^{2k}\delta}{\gamma^{2(k-k')}}\right)^{1/(2k')}.$$

The remaining proof is similar. □

7.1.2 Auxiliary lemmas

Lemma 7.2. *Let $t_1 \le t_2 \le \ldots$ be an ordered sequence (not necessarily distinct) and $t_r < t < t_{r+1}$. Let $f(x) = \mathbf{1}_{\{x \le t\}}$. Then*

$$f[t_i, \ldots, t_j] = (-1)^{i-r} \sum_{L \in \mathcal{L}(i,j)} \prod_{(x,y)\in L} \frac{1}{t_x - t_y}, \quad i \le r < r+1 \le j, \tag{7.22}$$

where $\mathcal{L}(i,j)$ is the set of lattice paths from $(r, r+1)$ to (i,j) using steps $(0,1)$ and $(-1,0)$.[2] *Furthermore,*

$$|f[t_1, \ldots, t_i]| \le \frac{\binom{i-2}{r-1}}{(t_{r+1}-t_r)^{i-1}}, \quad i \ge r+1. \tag{7.23}$$

Proof. Denote by $a_{i,j} = f[t_i, \ldots, t_j]$ when $i \le j$. It is obvious that $a_{i,i} = 1$

[2]Formally, for $a, b \in \mathbb{N}^2$, a lattice path from a to b using a set of steps S is a sequence $a = x_1, x_2, \ldots, x_n = b$ with all increments $x_{j+1} - x_j \in S$. In the matrix representation shown in the proof, this corresponds to a path from $a_{r,r+1}$ to $a_{i,j}$ going up and right. This path consists of entries (i,j) such that $i \le r < r+1 \le j$, and thus in (7.22) we always have $t_x \le t_r < t_{r+1} \le t_y$.

for $i \le r$; $a_{i,i} = 0$ for $i \ge r+1$; $a_{i,j} = 0$ for both $i < j \le r$ and $j > i \ge r+1$. For $i \le r < r+1 \le j$, the values can be obtained recursively by

$$a_{i,j} = \frac{a_{i,j-1} - a_{i+1,j}}{t_i - t_j}. \tag{7.24}$$

The above recursion can be represented in Neville's diagram as in Section 2.2. In this proof, it is equivalently represented in a upper triangular matrix as follows:

$$\begin{bmatrix} 1 & 0 & \cdots & 0 & a_{1,r+1} & \cdots & \\ & 1 & \ddots & \vdots & \vdots & & \\ & & 1 & 0 & a_{r-1,r+1} & \cdots & \\ & & & 1 & a_{r,r+1} & \cdots & \\ & & & & 0 & \cdots & 0 \\ & & \mathbf{0} & & & \ddots & \vdots \\ & & & & & & 0 \end{bmatrix}.$$

In the matrix, every $a_{i,j}$ is calculated using the two values left to it and below it. The values on any path from $a_{r,r+1}$ to $a_{i,j}$ going up and right will contribute to the formula of $a_{i,j}$ in (7.22). The paths consist of two types: first go to $a_{i,j-1}$ and then go right; first go to $a_{i+1,j}$ and then go up. Formally, $\{L, (i,j) : L \in \mathcal{L}_{i,j-1}\} \cup \{L, (i,j) : L \in \mathcal{L}_{i+1,j}\} = \mathcal{L}_{i,j}$. This will be used in the proof of (7.22) by induction present next. The base cases (r^{th} row and $(r+1)^{\text{th}}$ column) can be directly computed:

$$a_{r,j} = \prod_{v=r+1}^{j} \frac{1}{t_r - t_v}, \quad a_{i,r+1} = (-1)^{i-r} \prod_{v=i}^{r} \frac{1}{t_v - t_{r+1}}.$$

Suppose (7.22) holds for both $a_{i,j-1}$ and $a_{i+1,j}$. Then $a_{i,j}$ can be evaluated by

$$\begin{aligned} a_{i,j} &= \frac{(-1)^{i-r}}{t_i - t_j} \left(\sum_{L \in \mathcal{L}(i,j-1)} \prod_{(x,y) \in L} \frac{1}{t_x - t_y} + \sum_{L \in \mathcal{L}(i+1,j)} \prod_{(x,y) \in L} \frac{1}{t_x - t_y} \right) \\ &= (-1)^{i-r} \left(\sum_{L \in \mathcal{L}(i,j)} \prod_{(x,y) \in L} \frac{1}{t_x - t_y} \right). \end{aligned}$$

For the upper bound in (7.23), we note that $|\mathcal{L}(i,j)| \le \binom{(r-1)+(i-(r+1))}{r-1}$ in

(7.22), and each summand is at most $\frac{1}{(t_{r+1}-t_r)^{i-1}}$ in magnitude. □

Lemma 7.3. *Let*

$$P(x) = \prod_{i=1}^{\ell}(x - x_i) = \sum_{j=0}^{\ell} a_j x^j.$$

If $|x_i| \le \beta$ *for every* i*, then*

$$|a_j| \le \binom{\ell}{j}\beta^{\ell-j}.$$

Proof. P can be explicitly expanded and we obtain that

$$a_{\ell-j} = (-1)^j \sum_{\{i_1,i_2,\ldots,i_j\}\subseteq[\ell]} x_{i_1} \cdot x_{i_2} \cdot \ldots \cdot x_{i_j}.$$

The summation consists of $\binom{\ell}{j}$ terms, and each term is at most β^j in magnitude. □

7.2 Higher-order moments, and density functions

Lemma 7.4. *If* U *and* U' *each takes at most* k *values in* $[-1, 1]$*, and* $|\mathbb{E}[U^j] - \mathbb{E}[U'^j]| \le \epsilon$ *for* $j = 1, \ldots, 2k-1$*, then, for any* $\ell \ge 2k$*,*

$$|\mathbb{E}[U^\ell] - \mathbb{E}[U'^\ell]| \le 3^\ell\epsilon.$$

Proof. Let $f(x) = x^\ell$ and denote the atoms of U and U' by $x_1 < \cdots < x_{k'}$ for $k' \le 2k$. The function f can be interpolated on $x_1, \ldots, x_{k'}$ using a polynomial P of degree at most $2k-1$, which, in the Newton form (2.7), is

$$P(x) = \sum_{i=1}^{k'} f[x_1, \ldots, x_i]g_{i-1}(x) = \sum_{i=1}^{k'} \frac{f^{(i-1)}(\xi_i)}{(i-1)!}g_{i-1}(x),$$

for some $\xi_i \in [x_1, x_i]$, where $g_r(x) = \prod_{j=1}^{r}(x - x_j)$ and we used the intermediate value theorem for the divided differences (see [41, (2.1.4.3)]). Note that for $\xi_i \in [-1, 1]$, $|f^{(i-1)}(\xi_i)| \le \frac{\ell!}{(\ell-1+i)!}$. Similar to (7.7), we obtain that

$$|\mathbb{E}[U^\ell] - \mathbb{E}[U'^\ell]| = |\mathbb{E}[P(U)] - \mathbb{E}[P(U')]| \le \sum_{i=1}^{k'} \binom{\ell}{i-1} 2^{i-1}\epsilon \le 3^\ell\epsilon. \quad \square$$

In the context of learning Gaussian mixture models, we can obtain the distance between two density functions by comparing their moments.

Lemma 7.5 (Bound χ^2-divergence using moments difference)**.** *Suppose all moments of ν and ν' exist, and ν' is centered with variance σ^2. Then,*

$$\chi^2(\nu * N(0,1) \| \nu' * N(0,1)) \leq e^{\frac{\sigma^2}{2}} \sum_{j \geq 1} \frac{(\Delta m_j)^2}{j!},$$

where $\Delta m_j = m_j(\nu) - m_j(\nu')$ denotes the j^{th} moment difference.

Proof. The densities of two mixture distributions $\nu * N(0,1)$ and $\nu' * N(0,1)$ are

$$f(x) = \int \phi(x-u) \mathrm{d}\nu(u) = \phi(x) \sum_{j \geq 1} H_j(x) \frac{m_j(\nu)}{j!},$$
$$g(x) = \int \phi(x-u) \mathrm{d}\nu'(u) = \phi(x) \sum_{j \geq 1} H_j(x) \frac{m_j(\nu')}{j!},$$

respectively, where ϕ denotes the density of $N(0,1)$, and we used $\phi(x-u) = \phi(x) \sum_{j \geq 0} H_j(x) \frac{u^j}{j!}$ (see the exponential generating function of Hermite polynomials [52, 22.9.17]). Since $x \mapsto e^x$ is convex, applying Jensen's inequality yields that

$$g(x) = \phi(x) \mathbb{E}[\exp(U'x - U'^2/2)] \geq \phi(x) \exp(-\sigma^2/2).$$

Consequently,

$$\chi^2(\nu * N(0,1) \| \nu' * N(0,1)) = \int \frac{(f(x) - g(x))^2}{g(x)} \mathrm{d}x$$
$$\leq e^{\frac{\sigma^2}{2}} \mathbb{E}\left[\left(\sum_{j \geq 1} H_j(Z) \frac{\Delta m_j}{j!} \right)^2 \right] = e^{\frac{\sigma^2}{2}} \sum_{j \geq 1} \frac{(\Delta m_j)^2}{j!},$$

where $Z \sim N(0,1)$ and the last step follows from the orthogonality of Hermite polynomials (2.20). □

CHAPTER 8

LEARNING GAUSSIAN MIXTURES

Consider a k-component Gaussian location mixture model, where each observation is distributed as

$$X \sim \sum_{i=1}^{k} w_i N(\mu_i, \sigma^2). \tag{8.1}$$

Here w_i is the mixing weight such that $w_i \geq 0$ and $\sum_i w_i = 1$, μ_i is the mean (center) of the i^{th} component, and σ is the common standard deviation. Equivalently, we can write the distribution of an observation X as a convolution

$$X \sim \nu * N(0, \sigma^2), \tag{8.2}$$

where $\nu = \sum_{i=1}^{k} w_i \delta_{\mu_i}$ denotes the *mixing distribution*. Thus, we can write $X = U + \sigma Z$, where $U \sim \nu$ is referred to as the latent variable, and Z is standard normal and independent of U. We adopt the framework in Chapter 6 and the goal is to estimate the mixing distribution ν. Equivalently, estimating the mixing distribution can be viewed as a deconvolution problem, where the goal is to recover the distribution ν using observations drawn from the convolution (8.2). Throughout this chapter we consider estimating the mixing distribution ν with respect to the Wasserstein distance (6.2).

8.1 Related work and main results

Existing methodologies for mixture models are largely divided into likelihood-based and moment-based methods; see Section 8.1.3 for a detailed review. Among likelihood-based methods, the *Maximum Likelihood Estimate* (MLE) is not efficiently computable due to the non-convexity of the likelihood function. The most popular heuristic procedure to approximate the MLE is the

Expectation-Maximization (EM) algorithm [166]; however, absent separation conditions, no theoretical guarantee is known in general. Moment-based methods include the classical *method of moments* (MM) [11] and many extensions [167, 168]; however, the usual method of moments suffers from many issues as elaborated next.

8.1.1 Failure of the usual method of moments

The method of moments, commonly attributed to Pearson [11], produces an estimator by equating the population moments to the sample moments. While conceptually simple, this method suffers from the following problems, especially in the context of mixture models:

- *Solubility*: the method of moments entails solving a multivariate polynomial system, in which one frequently encounters non-existence or non-uniqueness of statistically meaningful solutions.
- *Computation*: solving moment equations can be computationally intensive. For instance, for k-component Gaussian mixture models, the system of moment equations consist of $2k-1$ polynomial equations with $2k-1$ variables.
- *Accuracy*: existing statistical literature on the method of moments [20, 167] either shows mere consistency under weak assumptions, or proves asymptotic normality assuming very strong regularity conditions (so that delta method works), which generally do not hold in mixture models since the convergence rates can be slower than parametric. Some results on nonparametric rates are known (cf. [20, Theorem 5.52] and [169, Theorem 14.4]) but the conditions are extremely hard to verify.

To explain the failure of the vanilla method of moments in Gaussian mixture models, we analyze the following simple two-component example.

Example 8.1. Consider a Gaussian mixture model with two unit variance components: $X \sim w_1 N(\mu_1, 1) + w_2 N(\mu_2, 1)$. Since there are three parameters μ_1, μ_2 and $w_1 = 1 - w_2$, we use the first three moments and solve the following

system of equations:

$$\begin{aligned}
\mathbb{E}_n[X] &= \mathbb{E}[X] = w_1\mu_1 + w_2\mu_2, \\
\mathbb{E}_n[X^2] &= \mathbb{E}[X^2] = w_1\mu_1^2 + w_2\mu_2^2 + 1, \\
\mathbb{E}_n[X^3] &= \mathbb{E}[X^3] = w_1\mu_1^3 + w_2\mu_2^3 + 3(w_1\mu_1 + w_2\mu_2),
\end{aligned} \tag{8.3}$$

where $\mathbb{E}_n[X^i] \triangleq \frac{1}{n}\sum_{j=1}^n X_j^i$ denotes the i^{th} moment of the empirical distribution from n i.i.d. samples. The right-hand sides of (8.3) are related to the moments of the mixing distribution by a linear transformation, which allow us to equivalently rewrite the moment equations (8.3) as:

$$\begin{aligned}
\mathbb{E}_n[X] &= \mathbb{E}[U] = w_1\mu_1 + w_2\mu_2, \\
\mathbb{E}_n[X^2 - 1] &= \mathbb{E}[U^2] = w_1\mu_1^2 + w_2\mu_2^2, \\
\mathbb{E}_n[X^3 - 3X] &= \mathbb{E}[U^3] = w_1\mu_1^3 + w_2\mu_2^3,
\end{aligned} \tag{8.4}$$

where $U \sim w_1\delta_{\mu_1} + w_1\delta_{\mu_2}$. It turns out that with finitely many samples, there is always a non-zero chance that (8.4) has no solution; even with infinite samples, it is possible that the solution does not exist with constant probability. To see this, note that, from the first two equations of (8.4), the solution does not exist whenever

$$\mathbb{E}_n[X^2] - 1 < \mathbb{E}_n^2[X], \tag{8.5}$$

that is, the Cauchy-Schwarz inequality fails. Consider the case $\mu_1 = \mu_2 = 0$, i.e., $X \sim N(0, 1)$. Then (8.5) is equivalent to

$$n(\mathbb{E}_n[X^2] - \mathbb{E}_n^2[X]) \leq n,$$

where the left-hand side follows the χ^2-distribution with $n - 1$ degrees of freedom. Thus, (8.5) occurs with probability approaching $\frac{1}{2}$ as n diverges, according to the central limit theorem.

In view of the above example, we note that the main issue with the usual method of moments is the following: although individually each moment estimate is accurate ($\sqrt{n}$-consistent), jointly they do not correspond to the moments of any distribution. Moment vectors satisfy many geometric constraints, e.g., the Cauchy-Schwarz and Hölder inequalities, and lie in a convex

set known as the *moment space*. Thus for any model parameters, with finitely many samples the method of moments fails with non-zero probability whenever the noisy estimates escape the moment space; even with infinitely many samples, it also provably happens with constant probability when the order of the mixture model is strictly less than k, or equivalently, the population moments lie on the boundary of the moment space (see Lemma 8.33 for a justification).

8.1.2 Main results

We propose the *denoised method of moments* (DMM), which consists of three main steps: (1) compute noisy estimates of moments, e.g., the unbiased estimates; (2) jointly denoise the moment estimates by project them onto the moment space; (3) execute the usual method of moments. It turns out that the extra step of projection resolves the three issues of the vanilla version of the method of moments identified in Section 8.1.1 simultaneously:

- *Solubility*: a unique statistically meaningful solution is guaranteed to exist by the classical theory of moments;
- *Computation*: the solution can be found through an efficient algorithm (Gauss quadrature) instead of invoking generic solvers of polynomial systems;
- *Accuracy*: the solution provably achieves the optimal rate of convergence, and automatically adaptive to the clustering structure of the population.

We emphasize that the denoising (projection) step is explicitly carried out via a convex optimization in Section 8.2.1, and implicitly used in analyzing Lindsay's algorithm [49] in Section 8.2.2, when the variance parameter is known and unknown, respectively.

Next we present the theoretical results. Throughout this chapter, we assume that the number of components satisfy

$$k = O\left(\frac{\log n}{\log \log n}\right). \tag{8.6}$$

Denote the underlying model as a convolution of $\nu = \sum_i w_i \delta_{\mu_i}$ and $N(0, \sigma^2)$. Our main result is Theorem 8.1.

Theorem 8.1 (Optimal rates)**.** *Suppose that $|\mu_i| \le M$ for $M \ge 1$ and σ is bounded by a constant, and both k and M are given.*

- *If σ is known, then there exists an estimator $\hat{\nu}$ computable in $O(kn)$ time such that, with probability at least $1 - \delta$,*

$$W_1(\nu, \hat{\nu}) \le O\left(Mk^{1.5}\left(\frac{n}{\log(1/\delta)}\right)^{-\frac{1}{4k-2}}\right); \tag{8.7}$$

- *If σ is unknown, then there exists an estimator $(\hat{\nu}, \hat{\sigma})$ computable in $O(kn)$ time such that, with probability at least $1 - \delta$,*

$$W_1(\nu, \hat{\nu}) \le O\left(Mk^2\left(\frac{n}{\log(1/\delta)}\right)^{-\frac{1}{4k}}\right), \tag{8.8}$$

and

$$|\sigma^2 - \hat{\sigma}^2| \le O\left(M^2 k\left(\frac{n}{\log(1/\delta)}\right)^{-\frac{1}{2k}}\right). \tag{8.9}$$

The above convergence rates are minimax optimal for constant k as shown in Section 8.3 (the optimality of (8.7) has been previously shown in [160]). Note that these results are proved under the worst-case scenario where the centers can be arbitrarily close, e.g., components completely overlap. It is reasonable to expect a faster convergence rate when the components are better separated, and, in fact, a parametric rate in the best-case scenario where the components are fully separated and weights are bounded away from zero. To capture the clustering structure of the mixture model, we introduce the following definition.

Definition 8.1. The Gaussian mixture (8.1) has k_0 (γ, ω)-separated clusters if there exists a partition $S_1, \ldots, S_{k_0}$ of $[k]$ such that

- $|\mu_i - \mu_{i'}| \ge \gamma$ for any $i \in S_\ell$ and $i' \in S_{\ell'}$ such that $\ell \ne \ell'$;
- $\sum_{i \in S_\ell} w_i \ge \omega$ for each ℓ.

In the absence of the minimal weight condition (i.e. $\omega = 0$), we say the Gaussian mixture has k_0 γ-separated clusters.

The next result shows that the DMM estimators attain the following adaptive rates.

Theorem 8.2 (Adaptive rate)**.** *Under the conditions of Theorem 8.1, suppose there are k_0 (γ, ω)-separated clusters such that $\gamma\omega \geq C\epsilon$ for some absolute constant $C > 2$, where ϵ denotes the right-hand side of (8.7) and (8.8) when σ is known and unknown, respectively.*

- *If σ is known, then, with probability at least $1 - \delta$,*[1]

$$W_1(\nu, \hat{\nu}) \leq O_k\left(M\gamma^{-\frac{2k_0-2}{2(k-k_0)+1}}\left(\frac{n}{\log(k/\delta)}\right)^{-\frac{1}{4(k-k_0)+2}}\right). \tag{8.10}$$

- *If σ is unknown, then, with probability at least $1 - \delta$,*[2]

$$\sqrt{|\sigma^2 - \hat{\sigma}^2|},\ W_1(\nu, \hat{\nu}) \leq O_k\left(M\gamma^{-\frac{k_0-1}{k-k_0+1}}\left(\frac{n}{\log(k/\delta)}\right)^{-\frac{1}{4(k-k_0+1)}}\right). \tag{8.11}$$

The result (8.10) is also minimax rate-optimal when k, k_0 and γ are constants, in view of the lower bounds in [160]. We also provide a simple proof in Remark 8.1 by extending the lower bound argument in Section 8.3. For the case of unknown σ, we do not have a matching lower bound for (8.11). In the fully separated case ($k_0 = k$), (8.11) reduces to $n^{-\frac{1}{4}}$ while a parametric rate is achievable.

Next we discuss the implication on density estimation (*proper* learning), where the goal is to estimate the density function of the Gaussian mixture by another k-Gaussian mixture density. Given that the estimated mixing distribution $\hat{\nu}$ from Theorem 8.1, a natural density estimate is the convolution $\hat{f} = \hat{\nu} * N(0, \sigma^2)$. Theorem 8.3 shows that the density estimate $\hat{f}$ is $O(\frac{1}{n})$-close to the true density f in χ^2-divergence, which bounds other common distance measures such as the Kullback-Leibler divergence, total variation, and Hellinger distance.

[1]Here $O_k(\cdot)$ denotes a constant factor that depends on k only.

[2]Note that the estimation rate for the mean part ν is the square root of the rate for estimating the variance parameter σ^2. Intuitively, this phenomenon is due to the infinite divisibility of the Gaussian distribution: note that for the location mixture model $\nu * N(0, \sigma^2)$ with $\nu \sim N(0, \epsilon^2)$ and $\sigma^2 = 1$ has the same distribution as that of $\nu \sim \delta_0$ and $\sigma^2 = 1 + \epsilon^2$.

Theorem 8.3 (Density estimation)**.** *Under the conditions of Theorem 8.1, denote the density of the underlying model by $f = \nu * N(0, \sigma^2)$. If σ is given, then there exists an estimate $\hat{f}$ such that*

$$\chi^2(\hat{f} \| f) + \chi^2(f \| \hat{f}) \leq O_k(\log(1/\delta)/n),$$

with probability $1 - \delta$.

So far we have been focusing on well-specified models. To conclude this subsection, we discuss misspecified models, where the data need not be generated from a k-Gaussian mixture. In this case, the DMM procedure still reports a meaningful estimate that is close to the best k-Gaussian mixture fit of the unknown distribution. This is made precise by the next result of oracle inequality style.

Theorem 8.4 (Misspecified model)**.** *Assume that $X_1, \ldots, X_n$ is independently drawn from a density f which is 1-subgaussian. Suppose there exists a k-component Gaussian location mixture g with variance σ^2 such that $\mathsf{TV}(f, g) \leq \epsilon$. Then, there exists an estimate $\hat{f}$ such that*

$$\mathsf{TV}(\hat{f}, f) \leq O_k\left(\epsilon\sqrt{\log(1/\epsilon)} + \sqrt{\log(1/\delta)/n}\right),$$

with probability $1 - \delta$.

8.1.3 Related work

There exist a vast literature on mixture models, in particular Gaussian mixtures, and the method of moments. For a comprehensive review see [170, 171]. In the following, we highlight a few existing results that are related to the present chapter.

Likelihood-based methods. Maximum likelihood estimation (MLE) is one of the most useful method for parameter estimation. Under strong separation assumptions, MLE is consistent and asymptotically normal [172]; however, those assumptions are difficult to verify, and it is computationally hard to obtain the global maximizer due to the non-convexity of the likelihood function in the location parameters.

Expectation-Maximization (EM) [166] is an iterative algorithm that aims to approximate the MLE. It has been widely applied in Gaussian mixture models [172, 173] and more recently in high-dimensional settings [156]. In general, this method is only guaranteed to converge to a local maximizer of the likelihood function rather than the global MLE. In practice we need to employ heuristic choices of the initialization [174] and stopping criteria [175], as well as possibly data augmentation techniques [176, 177]. Furthermore, its slow convergence rate is widely observed in practice [172, 174]. Additionally, the EM algorithm accesses the entire dataset in each iteration, which is particularly expensive for large sample size and high dimensions.

Lastly, we mention the nonparametric maximum likelihood estimation (NPMLE) in mixture models proposed by [178], where the maximization is taken over all mixing distributions which need not be k-atomic. This is an infinite-dimensional convex optimization problem, which has been studied in [179, 180, 170] and more recently in [181] on its computation based on discretization. One of the drawbacks of NPMLE is its lack of interpretability since the solution is a discrete distribution with at most n atoms cf. [181, Theorem 2]. Furthermore, few statistical guarantees in terms of convergence rate are available.

Moment-based methods. The simplest moment-based method is the method of moments (MM) introduced by Pearson [11]. The failure of the vanilla MM described in Section 8.1.1 has motivated various modifications including, notably, the *generalized method of moments* (GMM) introduced by Hansen [167]. GMM is a widely used methodology for analyzing economic and financial data (cf. [12] for a thorough review). Instead of exactly solving the MM equations, GMM aims to minimize the sum of squared differences between the sample moments and the fitted moments. While it enjoys various nice asymptotic properties [167], GMM involves a non-convex optimization problem which is computationally challenging to solve. In practice, heuristics such as gradient descent are used which converge slowly and lack theoretical guarantees.

For Gaussian mixture models (and more generally finite mixture models), our results can be viewed as a solver for GMM which is provably exact and computationally efficient, which significantly improves over the existing heuristic solvers in terms of both speed and accuracy; this is another algorith-

mic contribution of the present chapter. We also note that minimizing the sum of squares in GMM is not crucial and minimizing any distance yields the same theoretical guarantee. We discuss the connections to GMM in details in Section 8.2.1.

There are a number of recent work in the theoretical computer science literature on provable results for moment-based estimators in Gaussian location-scale mixture models, see, e.g., [30, 29, 182, 31, 183]. For instance, [30] considers the exhaustive search over the discretized parameter space such that the population moments is close to the empirical moments. This method achieves the estimation accuracy $n^{-\Theta(1/k)}$, which is optimal up to constant factors in the exponent, but is computationally expensive in practice. By carefully analyzing Pearson's method of moments equations [11], [31] showed the optimal rate $\Theta(n^{-1/12})$ for two-component location-scale mixtures; however, this approach is difficult to generalize to more components. Finally, for moment-based methods in multiple dimensions, such as spectral and tensor decomposition, we defer the discussion to Section 8.4.2.

Other methods. In the case of known variance, the minimum distance estimator is studied by [184, 159, 160]. Specifically, the estimator is a k-atomic distribution $\hat{\nu}$ such that $\hat{\nu} * N(0, \sigma^2)$ is the closest to the empirical distribution of the samples. The minimax optimal rate $O(n^{-\frac{1}{4k-2}})$ for estimating the mixing distribution under the Wasserstein distance is shown in [160] (which corrects the previous result in [159]), by bounding the W_1 distance between the mixing distributions in terms of the KS distance of the Gaussian mixtures [160, Lemma 4.5]. However, the minimum distance estimator is in general computationally expensive and suffers from the same non-convexity issue of the MLE. In contrast, denoised method of moments is efficiently computable and adaptively achieves the optimal rate of accuracy as given in Theorem 8.2.

Finally, we discuss density estimation, which has been studied for the MLE in [185, 186]. If the estimator is allowed to be any density (improper learning), it is known that as long as the mixing distribution has a bounded support, the rate of convergence is close to parametric regardless of the number of components; specifically, the optimal squared L_2-risk is $\Theta(\frac{\sqrt{\log n}}{n})$ [187], achieved by the kernel density estimator designed for analytic densities [188]. Of course, the optimal proper density estimate (which is required to be a k-

Gaussian mixture) enjoys the same rate of convergence; however, finding the k-Gaussian mixture that best approximates a given density is computationally challenging again due to the non-convexity. From this perspective, another contribution of Theorems 8.3-8.4 is that by approximating moments the best approximation can be found within logarithmic factors.

8.2 Estimators and statistical guarantees

8.2.1 Known variance

The denoised method of moments for estimating Gaussian location mixture models (8.2) with known variance parameter σ^2 consists of three main steps:

1. estimate $\mathbf{m}_{2k-1}(\nu)$ by $\tilde{m} = (\tilde{m}_1, \ldots, \tilde{m}_{2k-1})$ (using Hermite polynomials);

2. denoise $\tilde{m}$ by its projection $\hat{m}$ onto the moment space (semidefinite programming);

3. find a k-atomic distribution $\hat{\nu}$ such that $\mathbf{m}_{2k-1}(\hat{\nu}) = \hat{m}$ (Gauss quadrature).

The complete algorithm is summarized in Algorithm 8.1.

Algorithm 8.1 Denoised method of moments (DMM) with known variance.

Input: n independent samples $X_1, \ldots, X_n$, order k, variance σ^2, interval $I = [a, b]$.

Output: estimated mixing distribution.

1: **for** $r = 1$ **to** $2k - 1$ **do**

2: $\quad \hat{\gamma}_r = \frac{1}{n} \sum_i X_i^r$

3: $\quad \tilde{m}_r = r! \sum_{i=0}^{\lfloor r/2 \rfloor} \frac{(-1/2)^i}{i!(r-2i)!} \hat{\gamma}_{r-2i} \sigma^{2i}$

4: **end for**

5: Let $\hat{m}$ be the optimal solution of the following:

$$\min\{\|\tilde{m} - \hat{m}\| : \hat{m} \text{ satisfies (2.16)}\}, \tag{8.12}$$

where $\tilde{m} = (\tilde{m}_1, \ldots, \tilde{m}_{2k-1})$.

6: Report the outcome of Gauss quadrature (Algorithm 2.2) with input $\hat{m}$.

We estimate the moments of the mixing distribution in lines 1 to 4. The unique unbiased estimators for the polynomials of the mean parameter in a Gaussian location model are Hermite polynomials (2.21) such that $\mathbb{E}H_r(X) = \mu^r$ when $X \sim N(\mu, 1)$. Thus, if we define

$$\gamma_r(x, \sigma) = \sigma^r H_r(x/\sigma) = r! \sum_{j=0}^{\lfloor r/2 \rfloor} \frac{(-1/2)^j}{j!(r-2j)!} \sigma^{2j} x^{r-2j}, \tag{8.13}$$

then $\mathbb{E}\gamma_r(X, \sigma) = \mu^r$ when $X \sim N(\mu, \sigma^2)$. Hence, by linearity, $\tilde{m}_r$ is an unbiased estimate of $m_r(\nu)$. The variance of $\tilde{m}_r$ is analyzed in Lemma 8.1.

Lemma 8.1. *If* $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \nu * N(0, \sigma^2)$ *and* ν *is supported on* $[-M, M]$, *then*

$$\mathsf{var}[\tilde{m}_r] \leq \frac{1}{n}(O(M + \sigma\sqrt{r}))^{2r}.$$

As observed in Section 8.1.1, the major reason for the failure of the usual method of moments is that the unbiased estimate $\tilde{m}$ needs not constitute a legitimate moment sequence, despite the consistency of each individual $\tilde{m}_i$. To resolve this issue, we project $\tilde{m}$ to the moment space using (8.12). As explained in Section 2.3, (2.16) consists of positive semidefinite constraints, and thus the optimal solution of (8.12) can be obtained by semidefinite programming (SDP).[3] In fact, it suffices to solve a *feasibility* program and find any valid moment vector $\hat{m}$ that is within the desired $\frac{1}{\sqrt{n}}$ statistical accuracy.

Now that $\hat{m}$ is indeed a valid moment sequence, we use the Gauss quadrature introduced in Section 2.3 (see Algorithm 2.2) to find the unique k-atomic distribution $\hat{\nu}$ such that $\mathbf{m}_{2k-1}(\hat{\nu}) = \hat{m}$. Using Algorithm 8.1, $\tilde{m}$ is computed in $O(kn)$ time, the semidefinite programming is solvable in $O(k^{6.5})$ time using the interior-point method (see [190]), and the Gauss quadrature can be evaluated in $O(k^3)$ time [50]. In view of the global assumption (8.6), Algorithm 8.1 can be executed in $O(kn)$ time.

We now prove the statistical guarantee (8.7) for the DMM estimator previously announced in Theorem 8.1:

Proof. By scaling it suffices consider $M = 1$. We use Algorithm 8.1 with Euclidean norm in (8.12). Using the variance of $\tilde{m}$ in Lemma 8.1 and Chebyshev

[3]The formulation (8.12) with Euclidean norm can already be implemented in popular modeling languages for convex optimization problem such as CVXPY [189]. A standard form of SDP is given in Section 8.6.7.

inequality yield that, for each $r = 1, \ldots, 2k-1$, with probability $1 - \frac{1}{8k}$,

$$|\tilde{m}_r - m_r(\nu)| \le \sqrt{k/n}(c\sqrt{r})^r, \tag{8.14}$$

for some absolute constant c. By the union bound, with probability $3/4$, (8.14) holds simultaneously for every $r = 1, \ldots, 2k-1$, and thus

$$\|\tilde{m} - \mathbf{m}_{2k-1}(\nu)\|_2 \le \epsilon, \quad \epsilon \triangleq \frac{(\sqrt{ck})^{2k+1}}{\sqrt{n}}.$$

Since $\mathbf{m}_{2k-1}(\nu)$ satisfies (2.16) and thus is one feasible solution for (8.12), we have $\|\tilde{m} - \hat{m}\|_2 \le \epsilon$. Note that $\hat{m} = \mathbf{m}_{2k-1}(\hat{\nu})$. Hence, by triangle inequality, we obtain the following statistical accuracy:

$$\|\mathbf{m}_{2k-1}(\hat{\nu}) - \mathbf{m}_{2k-1}(\nu)\|_2 \le \epsilon. \tag{8.15}$$

Applying Proposition 7.1 yields that, with probability $3/4$,

$$W_1(\hat{\nu}, \nu) \le O\left(k^{1.5} n^{-\frac{1}{4k-2}}\right).$$

The confidence $1 - \delta$ in (8.7) can be obtained by the usual "median trick": divide the samples into $T = \log \frac{2k}{\delta}$ batches, apply Algorithm 8.1 to each batch of n/T samples, and take $\tilde{m}_r$ to be the median of these estimates. Then Hoeffding's inequality and the union bound imply that, with probability $1-\delta$,

$$|\tilde{m}_r - m_r(\nu)| \le \sqrt{\frac{\log(2k/\delta)}{n}}(c\sqrt{r})^r, \quad \forall\ r = 1, \ldots, 2k-1, \tag{8.16}$$

and the conclusion follows. □

To conclude this subsection, we discuss the connection to the generalized method of moments (GMM). Instead of solving the moment equations, GMM aims to minimize the difference between estimated and fitted moments:

$$Q(\theta) = (\hat{m} - m(\theta))^\top W (\hat{m} - m(\theta)), \tag{8.17}$$

where $\hat{m}$ is the estimated moment, θ is the model parameter, and W is a positive semidefinite weighting matrix. The minimizer of $Q(\theta)$ serves as the GMM estimate for the unknown model parameter θ_0. In general the

objective function Q is nonconvex in θ, notably under the Gaussian mixture model with θ corresponding to the unknown means and weights, which is hard to optimize. Note that (8.12) with the Euclidean norm is *equivalent* to GMM with the identity weighting matrix. Therefore Algorithm 8.1 is an exact solver for GMM in the Gaussian location mixture model.

In theory, the optimal weighting matrix W^* that minimizes the asymptotic variance is the inverse of $\lim_{n\to\infty} \mathsf{cov}[\sqrt{n}(\hat{m} - m(\theta_0))]$, which depends the unknown model parameters θ_0. Thus, a popular approach is a two-step estimator [12]:

1. a suboptimal weighting matrix, e.g., identify matrix, is used in the GMM to obtain a consistent estimate of θ_0 and hence a consistent estimate $\hat{W}$ for W^*;

2. θ_0 is re-estimated using the weighting matrix $\hat{W}$.

The above two-step approach can be similarly implemented in the denoised method of moments.

8.2.2 Unknown variance

When the variance parameter σ^2 is unknown, unbiased estimator for the moments of the mixing distribution no longer exists (see Lemma 8.25). It is not difficult to consistently estimate the variance,[4] then plug into the DMM estimator in Section 8.2.1 to obtain a consistent estimate of the mixing distribution ν; however, the convergence rate is far from optimal. In fact, to achieve the optimal rate in Theorem 8.1, it is crucial to simultaneously estimate both the means and the variance parameters. To this end, again we take a moment-based approach. The following result provides a guarantee for any joint estimate of both the mixing distribution and the variance parameter in terms of the moments accuracy.

Proposition 8.1. *Let*

$$\pi = \nu * N(0, \sigma^2), \quad \hat{\pi} = \hat{\nu} * N(0, \hat{\sigma}^2),$$

[4]For instance, the simple estimator $\hat{\sigma} = \frac{\max_i X_i}{\sqrt{2 \log n}}$ satisfies $|\sigma - \hat{\sigma}| = O_P(\log n)^{-\frac{1}{2}}$.

where $\nu, \hat{\nu}$ are k-atomic distributions supported on $[-M, M]$, and $\sigma, \hat{\sigma}$ are bounded. If $|m_r(\pi) - m_r(\hat{\pi})| \le \epsilon$ for $r = 1, \dots, 2k$, then

$$|\sigma^2 - \hat{\sigma}^2| \le O(M^2 \epsilon^{\frac{1}{k}}), \quad W_1(\nu, \hat{\nu}) \le O(Mk^{1.5} \epsilon^{\frac{1}{2k}}).$$

To apply Proposition 8.1, we can solve the method of moments equations, namely, find a k-atomic distribution $\hat{\nu}$ and $\hat{\sigma}^2$ such that

$$\mathbb{E}_n[X^r] = \mathbb{E}_{\hat{\pi}}[X^r], \qquad r = 1, \dots, 2k, \tag{8.18}$$

where $\hat{\pi} = \hat{\mu} * N(0, \hat{\sigma}^2)$ is the fitted Gaussian mixture. Here both the number of equations and the number of variables are equal to $2k$. Suppose (8.18) has a solution $(\hat{\mu}, \hat{\sigma})$. Then applying Proposition 8.1 with $\delta = O_k(\frac{1}{\sqrt{n}})$ achieves the rate $O_k(n^{-1/(4k)})$ in Theorem 8.1, which is minimax optimal (see Section 8.3). In stark contrast to the known σ case, where we have shown in Section 8.1.1 that the vanilla method of moments equation can have no solution unless we denoise by projection to the moment space, here with one extra scale parameter σ, one can show that (8.18) has a solution with probability one![5] Furthermore, an efficient method of finding *a* solution to (8.18) is due to Lindsay [49] and summarized in Algorithm 8.2. Indeed, the sample moments are computable in $O(kn)$ time, and the smallest non-negative root of the polynomial of degree $k(k+1)$ can be found in $O(k^2)$ time using Newton's method (see [191]). So overall Lindsay's estimator can be evaluated in $O(kn)$ time.

In [49] the consistency of this estimator was proved under extra assumptions. In fact we will that it unconditionally achieves the minimax optimal rate (8.8) and (8.9) previously announced in Theorem 8.1. In this section we show that Lindsay's algorithm produces an estimator $\hat{\sigma}$ so that the corresponding the moment estimates lie in the moment space with probability one. In this sense, although no explicit projection is involved, the noisy estimates are *implicitly* denoised.

We first describe the intuition of the choice of $\hat{\sigma}$ in Lindsay's algorithm,

[5] It is possible that the equation (8.18) has no solution, for instance, when $k = 2, n = 7$ and the empirical distribution is $\pi_7 = \frac{1}{7}\delta_{-\sqrt{7}} + \frac{1}{7}\delta_{\sqrt{7}} + \frac{5}{7}\delta_0$. The first four empirical moments are $\mathbf{m}_4(\pi_7) = (0, 2, 0, 14)$, which cannot be realized by any two-component Gaussian mixture (8.1). Indeed, suppose $\hat{\pi} = w_1 N(\mu_1, \sigma^2) + (1-w_1) N(\mu_2, \sigma^2)$ is a solution to (8.18). Eliminating variables leads to the contradiction that $2\mu_1^4 + 2 = 0$. Assuringly, as we will show later in Lemma 8.3, such cases occur with probability zero.

Algorithm 8.2 Lindsay's estimator for normal mixtures with an unknown common variance.

Input: n samples $X_1, \ldots, X_n$.
Output: estimated mixing distribution $\hat{\nu}$, and estimated variance $\hat{\sigma}^2$.

1: **for** $r = 1$ **to** $2k$ **do**
2: $\quad \hat{\gamma}_r = \frac{1}{n}\sum_i X_i^r$
3: $\quad \hat{m}_r(\sigma) = r! \sum_{i=0}^{\lfloor r/2 \rfloor} \frac{(-1/2)^i}{i!(r-2i)!} \hat{\gamma}_{r-2i} \sigma^{2i}$
4: **end for**
5: Let $\hat{d}_k(\sigma)$ be the determinant of the matrix $\{\hat{m}_{i+j}(\sigma)\}_{i,j=0}^{k}$.
6: Let $\hat{\sigma}$ be the smallest positive root of $\hat{d}_k(\sigma) = 0$.
7: **for** $r = 1$ **to** $2k$ **do**
8: $\quad \hat{m}_r = \hat{m}_r(\hat{\sigma})$
9: **end for**
10: Let $\hat{\nu}$ be the outcome of the Gauss quadrature (Algorithm 2.2) with input $\hat{m}_1, \ldots, \hat{m}_{2k-1}$
11: Report $\hat{\nu}$ and $\hat{\sigma}^2$.

i.e., line 6 of Algorithm 8.2. Let $X \sim \nu * N(0, \sigma^2)$. For any $\sigma' \le \sigma$, we have

$$\mathbb{E}[\gamma_j(X, \sigma')] = m_j(\nu * N(0, \sigma^2 - \sigma'^2)).$$

Let $d_k(\sigma')$ denote the determinant of the moment matrix $\{\mathbb{E}[\gamma_{i+j}(X, \sigma')]\}_{i,j=0}^{k}$, which is an even polynomial in σ' of degree $k(k+1)$. According to Theorem 2.12, $d_k(\sigma') > 0$ when $0 \le \sigma' < \sigma$ and becomes zero at $\sigma' = \sigma$, and thus σ is characterized by the smallest positive zero of d_k. In lines 5 – 6, d_k is estimated by $\hat{d}_k$ using the empirical moments, and σ is estimated by the smallest positive zero of $\hat{d}_k$. We first note that $\hat{d}_k$ indeed has a positive zero as shown in Lemma 8.2.

Lemma 8.2. *Assume $n > k$ and the mixture distribution has a density. Then, almost surely, $\hat{d}_k$ has a positive root within $(0, s]$, where $s^2 \triangleq \frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}_n[X])^2$ denotes the sample variance.*

The next result shows that, with the above choice of $\hat{\sigma}$, the moment estimates $\hat{m}_j = \mathbb{E}_n[\gamma_j(X, \hat{\sigma})]$ for $j = 1, \ldots, 2k$ given in line 8 are implicitly denoised and lie in the moment space with probability one. Thus (8.18) has a solution, and the estimated mixing distribution $\hat{\nu}$ can be found by the Gauss quadrature. This result was previously shown in [49] but under extra conditions.

Lemma 8.3. *Assume $n \geq 2k-1$ and the mixture distribution has a density. Then, almost surely, there exists a k-atomic distribution $\hat{\nu}$ such that $m_j(\hat{\nu}) = \hat{m}_j$ for $j \leq 2k$, where $\hat{m}_j$ is from Algorithm 8.2.*

With the above analysis, we now prove the statistical guarantee (8.8) and (8.9) for Lindsay's algorithm announced in Theorem 8.1.

Proof. It suffices to consider $M = 1$. Let $\hat{\pi} = \hat{\nu} * N(0, \hat{\sigma}^2)$ and $\pi = \nu * N(0, \sigma^2)$ denote the estimated mixture distribution and the ground truth, respectively. Let $\hat{m}_r = \mathbb{E}_n[X^r]$ and $m_r = m_r(\pi)$. The variance of $\hat{m}_r$ is upper bounded by

$$\mathsf{var}[\hat{m}_r] = \frac{1}{n}\mathsf{var}[X_1^r] \leq \frac{1}{n}\mathbb{E}[X^{2r}] \leq \frac{(\sqrt{cr})^{2r}}{n},$$

for some absolute constant c. Using Chebyshev inequality, for each $r = 1, \ldots, 2k$, with probability $1 - \frac{1}{8k}$, we have,

$$|\hat{m}_r - m_r| \leq (\sqrt{cr})^r \sqrt{k/n}. \tag{8.19}$$

By the union bound, with probability $3/4$, the above holds simultaneously for every $r = 1, \ldots, 2k$. It follows from Lemmas 8.2 and 8.3 that (8.18) holds with probability one. Therefore,

$$|m_r(\hat{\pi}) - m_r(\pi)| \leq (\sqrt{cr})^r \sqrt{k/n}, \quad r = 1, \ldots, 2k,$$

for some absolute constant c. In the following, the error of variance estimate is denoted by $\tau^2 = |\sigma^2 - \hat{\sigma}^2|$.

- If $\sigma \leq \hat{\sigma}$, let $\nu' = \hat{\nu} * N(0, \tau^2)$. Using $\mathbb{E}_\pi[\gamma_r(X, \sigma)] = m_r(\nu)$ and $\mathbb{E}_{\hat{\pi}}[\gamma_r(X, \sigma)] = m_r(\nu')$, where γ_r is the Hermite polynomial (8.13), we obtain that (see Lemma 8.21)

  $$|m_r(\nu') - m_r(\nu)| \leq (\sqrt{c'k})^{2k} \sqrt{k/n}, \quad r = 1, \ldots, 2k, \tag{8.20}$$

  for an absolute constant c'. Applying Proposition 8.1 yields that

  $$|\sigma^2 - \hat{\sigma}^2| \leq O(kn^{-\frac{1}{2k}}), \quad W_1(\nu, \hat{\nu}) \leq O(k^2 n^{-\frac{1}{4k}}).$$

- If $\sigma \geq \hat{\sigma}$, let $\nu' = \nu * N(0, \tau^2)$. Similar to (8.20), we have

$$|m_r(\hat{\nu}) - m_r(\nu')| \leq (\sqrt{c'k})^{2k}\sqrt{k/n} \triangleq \epsilon, \quad r = 1, \ldots, 2k.$$

To apply Proposition 8.1, we also need to ensure that $\hat{\nu}$ has a bounded support, which is not obvious. To circumvent this issue, we apply a truncation argument thanks to the following tail probability bound for $\hat{\nu}$ (see Lemma 8.16):

$$\mathbb{P}[|\hat{U}| \geq \sqrt{c_0 k}] \leq \epsilon(\sqrt{c_1 k}/t)^{2k}, \quad \hat{U} \sim \hat{\nu}, \tag{8.21}$$

for absolute constants c and c'. To this end, consider $\tilde{U} = \hat{U}\mathbf{1}_{\{|\hat{U}| \leq \sqrt{c_0 k}\}}$ $\sim \tilde{\nu}$. Note that $\tilde{U}$ is k-atomic supported on $[-\sqrt{c_0 k}, \sqrt{c_0 k}]$, we have $W_1(\nu, \hat{\nu}) \leq \epsilon e^{O(k)}$ and $|m_r(\tilde{\nu}) - m_r(\hat{\nu})| \leq k\epsilon(c_1 k)^k$ for $r = 1, \ldots, 2k$. Using the triangle inequality yields that

$$|m_r(\tilde{\nu}) - m_r(\nu')| \leq \epsilon + k\epsilon(c_1 k)^k.$$

Now we apply Proposition 8.1 with $\tilde{\nu}$ and $\nu * N(0, \tau^2)$ where both $\tilde{\nu}$ and ν are k-atomic supported on $[-\sqrt{c_0 k}, \sqrt{c_0 k}]$. In the case $\tilde{\nu}$ is discrete, the dependence on k in Proposition 8.1 can be improved (by improving (8.63) in the proof) and we obtain that

$$|\sigma^2 - \hat{\sigma}^2| \leq O(kn^{-\frac{1}{2k}}), \quad W_1(\nu, \tilde{\nu}) \leq O(k^2 n^{-\frac{1}{4k}}).$$

Using $k \leq O(\frac{\log n}{\log \log n})$, we also obtain $W_1(\nu, \hat{\nu}) \leq O(k^2 n^{-\frac{1}{2k}})$ by the triangle inequality.

To obtain a confidence $1 - \delta$ in (8.8) and (8.9), we can replace the empirical moments $\hat{m}_r$ by the median of $T = \log \frac{1}{\delta}$ independent estimates similar to (8.16). □

8.2.3 Adaptive results

In Sections 8.2.1 and 8.2.2, we proved the statistical guarantees of our estimators under the worst-case scenario where the means can be arbitrarily close. Under separation conditions on the means (see Definition 8.1), our estima-

tors automatically achieve a strictly better accuracy than the one claimed in Theorem 8.1. The goal in this subsection is to show those adaptive results. The key is the adaptive version of the moment comparison theorems Propositions 7.3 and 7.4.

The adaptive result (8.10) in the known variance parameter case is obtained using Proposition 7.3 in place of Proposition 7.1. To deal with unknown variance parameter case, using Proposition 7.4, we first show the following adaptive version of Proposition 8.1.

Proposition 8.2. *Under the conditions of Proposition 8.1, if both Gaussian mixtures both have k_0 γ-separated clusters in the sense of Definition 8.1, then,*

$$\sqrt{|\sigma^2 - \hat{\sigma}^2|},\ W_1(\nu, \hat{\nu}) \leq O_k\left(\left(\frac{\epsilon}{\gamma^{2(k_0-1)}}\right)^{\frac{1}{2(k-k_0+1)}}\right).$$

Using these propositions, we now prove the adaptive rate of the denoised method of moments previously announced in Theorem 8.2.

Proof of Theorem 8.2. By scaling it suffices to consider $M = 1$. Recall that the Gaussian mixture is assumed to have k_0 (γ, ω)-separated clusters in the sense of Definition 8.1, that is, there exists a partition $S_1, \ldots, S_{k_0}$ of $[k]$ such that $|\mu_i - \mu_{i'}| \geq \gamma$ for any $i \in S_\ell$ and $i' \in S_{\ell'}$ such that $\ell \neq \ell'$, and $\sum_{i \in S_\ell} w_i \geq \omega$ for each ℓ.

Let $\hat{\nu}$ be the estimated mixing distribution which satisfies $W_1(\nu, \hat{\nu}) \leq \epsilon$ by Theorem 8.1. Since $\gamma\omega \geq C\epsilon$ by assumption, for each S_ℓ, there exists $i \in S_\ell$ such that μ_i is within distance $c\gamma$, where $c = 1/C$, to some atom of $\hat{\nu}$. Therefore, the estimated mixing distribution $\hat{\nu}$ has k_0 $(1-2c)\gamma$-separated clusters. Denote the union of the support sets of ν and $\hat{\nu}$ by $\mathcal{S}$.

- When σ is known, each atom in $\mathcal{S}$ is $\Omega(\gamma)$ away from at least $2(k_0 - 1)$ other atoms. Then (8.10) follows from Proposition 7.3 with $\ell = 2k$ and $\ell' = (2k-1) - 2(k_0 - 1)$.

- When σ is unknown, (8.11) follows from a similar proof of (8.8) and (8.9) with Proposition 8.1 replaced by Proposition 8.2. □

The rate in (8.10) as well as its optimality is previously obtained in [160], but their minimum-distance estimator is computationally expensive. Finally, we note that if one only assumes the separation condition but not the lower

bound on the weights, we can obtain an intermediate result that is stronger than (8.7) but weaker than (8.10).

Theorem 8.5. *Under the conditions of Theorem 8.1, suppose σ is known and the Gaussian mixture has k_0 γ-separated clusters. Then, with probability at least $1-\delta$,*

$$W_1(\nu, \hat{\nu}) \le O_k\left(M\gamma^{-\frac{k_0-1}{2k-k_0}}\left(\frac{n}{\log(k/\delta)}\right)^{-\frac{1}{4k-2k_0}}\right). \tag{8.22}$$

8.2.4 Unbounded means

In the previous subsections, we assume that the means lie in a bounded interval. In the unbounded case, it is in fact impossible to estimate the mixing distribution under the Wasserstein distance.[6] Nevertheless, provided that the weights are bounded away from zero, it is possible to estimate the support set of the mixing distribution with respect to the Hausdorff distance (cf. (6.5)). This is the goal of this subsection.

In the unbounded case, blindly applying the previous moment-based methods does not work, because the estimated moments suffer from large variance due to the wide range of values of the means (cf. Lemma 8.1). To resolve this issue, we shall apply the *divide and conquer* technique to reduce the range in each subprogram. Specifically, we will divide the real line into small intervals, estimate means in each interval separately, and report the union. The complete algorithm is given in Algorithm 8.3.

The first step is to apply a clustering method that partitions the samples into a small number of groups. There are many clustering algorithms in practice such as the popular Lloyd's k-means clustering [192]. In lines 1 – 4, we present a conservative yet simple clustering with the following guarantees (see Lemma 8.18):

- each interval is of length at most $O(kL)$;
- a sample $X_i = U_i + \sigma Z_i$ is always in the same interval as the latent variable U_i.

[6] Let $\pi_\epsilon = \frac{1+\epsilon}{2}\delta_0 + \frac{1-\epsilon}{2}\delta_M$. Then $W_1(\pi_0, \pi_\epsilon) = M\epsilon$, but $D(\pi_0\|\pi_\epsilon) \le O(\epsilon^2)$. Choosing $\epsilon = o(1/\sqrt{n})$ and $M \gg 1/\epsilon$ leads to arbitrarily large estimation error.

Algorithm 8.3 Estimate means of a Gaussian mixture model in the unbounded case.

Input: n samples $X_1, \dots, X_n$, variance parameter σ^2 (optional), cluster parameter L, and weights threshold τ, test sample size n'.

Output: a set of estimated means $\hat{S}$

1: Merge overlapping intervals $[X_i \pm L]$ for $i \le n'$ into disjoint ones $I_1, \dots, I_s$.
2: **for** $j = 1$ **to** s **do**
3: Let c_j, ℓ_j be such that $I_j = [c_j \pm \ell_j]$.
4: Let $C_j = \{X_i - c_j : X_i \in I_j, i > n'\}$.
5: **if** σ^2 is specified **then**
6: Let $(\hat{w}, \hat{\mu})$ be the outcome of Algorithm 8.1 with input C_j, σ^2, and $[-\ell_j, \ell_j]$.
7: **else**
8: Let $(\hat{w}, \hat{\mu})$ be the outcome of Algorithm 8.2 with input C_j.
9: **end if**
10: Let $\hat{S}_j = \{\hat{x}_i + c_j : \hat{w}_i \ge \tau\}$.
11: **end for**
12: Report $\hat{S} = \cup_j \hat{S}_j$.

In the present clustering method, each cluster C_j only contains samples that are not used in line 1 so that the intervals are independent of each C_j. This is a commonly used *sample splitting* technique in statistics to simplify the analysis. Note that only a small number of samples are needed to determine the intervals (see Theorem 8.6). In the second step, we estimate means in each I_j using samples C_j and report the union of all means.

The statistical guarantee of Algorithm 8.3 is analyzed in Theorem 8.6. Note that Theorem 8.6 holds in the worst case, and can be improved in many situations: The number of samples in each C_j increases proportionally to the total weights. The adaptive rate in Theorem 8.2 is applicable when separation is present within one interval. We can postulate fewer components in one interval based on information from other intervals.

Theorem 8.6. *Assume in the Gaussian mixture (8.1) $w_i \ge \epsilon$, σ is bounded. Let $S = \operatorname{supp}(\nu)$ be the set of means of the Gaussian mixture, and $\hat{S}$ be the output of Algorithm 8.3 with $L = \Theta(\sqrt{\log n})$ and $\tau = \epsilon/(2k)$. If $n \ge 2n' \ge \Omega(\frac{\log(k/\delta)}{\epsilon})$, then, with probability $1 - \delta - n^{-\Omega(1)}$, we have*

$$d_H(\hat{S}, S) \le \begin{cases} O\left(Lk^{3.5}(\frac{\epsilon n}{\log(1/\delta)})^{-\frac{1}{4k-2}}/\epsilon\right), & \sigma \text{ is known}, \\ O\left(Lk^{4}(\frac{\epsilon n}{\log(1/\delta)})^{-\frac{1}{4k}}/\epsilon\right), & \sigma \text{ is unknown}, \end{cases}$$

Table 8.1: Parameters in a random Gaussian mixture model.

Weights	0.123	0.552	0.010	0.080	0.235
Centers	-0.236	-0.168	-0.987	0.299	0.150

where d_H denotes the Hausdorff distance (see (6.5)).

8.2.5 Numerical experiments

The algorithms of the current chapter are implemented in Python.[7] In Algorithm 8.1, the explicit denoising via semidefinite programming uses CVXPY [189] and CVXOPT [193], and the Gauss quadrature is calculated based on [50]. In this section, we compare the performance of our algorithms with the EM algorithm, also implemented in Python, and the GMM algorithm using the popular package gmm [194] implemented in R. We omit the comparison with the vanilla method of moments which constantly fails to output a meaningful solution (see Section 8.1.1). In all figures presented in this section, we omit the running time of gmm, which is on the order of hours as compared to seconds using our algorithms; the slowness of of gmm is mainly due to the heuristic solver of the non-convex optimization (8.17).

We first clarify the parameters used in the experiments. EM and the iterative solver for (8.17) in gmm both require an initialization and a stop criterion. We use the best of five random initializations: The means are drawn independently from a uniform distribution, and the weights are from a Dirichlet distribution. Then we pick the estimate that maximizes the likelihood and the minimal moment discrepancy (8.17) in EM and GMM, respectively. The EM algorithm terminates when log-likelihood increases less than 10^{-3} or 5,000 iterations are reached; we use the default stop criterion in gmm [194].

Known variance. We generated a random instance of Gaussian mixture model with five components and a unit variance. The means are drawn uniformly from $[-1, 1]$; the weights are drawn from the Dirichelet distribution with parameters $(1, 1, 1, 1, 1)$, i.e., uniform over the probability simplex. It has the parameters in Table 8.1. We repeat the experiments 20 times and plot

[7] The implementations are available at `https://github.com/Albuso0/mixture`.

and the average and the standard deviation of the errors in the Wasserstein distance. We also plot the running time at each sample size. The results are shown in Figure 8.1. These three algorithms have comparable accuracies,

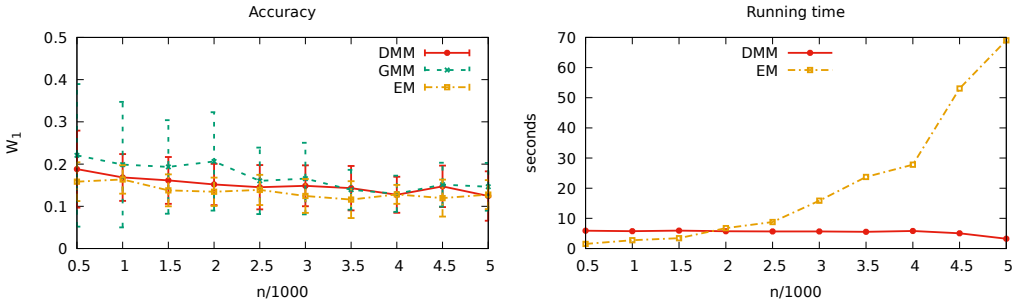


Figure 8.1: Comparison of different methods under a randomly generated five-component Gaussian mixture model.

but EM is significantly slower than DMM: it is 15 times slower with 5,000 samples and is increasing with the number of samples. This is because EM accesses all samples in each iteration, instead of first summarizing data into a few moments.

EM becomes slower when samples from different components have more overlaps since the maximizer of the likelihood function lies in a flat area [172, 174]. In this case, a loose stop criterion will terminate the algorithm early before convergence, while a stringent one incurs substantially longer running time. To show this, we do an experiment in which a two-component Gaussian mixture is to be estimated. However, the two components completely overlap, i.e., samples are drawn from $N(0, 1)$. To see the effect of the stop criterion, we additionally run the EM algorithm that terminates when the log-likelihood increases less than 10^{-4} instead of 10^{-3}, shown as EM^+ in the figures. The setup is the same as before and the results are shown in Figure 8.2. Again the accuracies are similar, but EM^+ is much slower than EM without substantial gain in the accuracy. Specifically, at 5,000 samples, EM is still 15 times slower than DMM, but EM^+ is 60 times slower.

Lastly, we demonstrate a faster rate in the well-separated case as shown in Theorem 8.2. In this experiment, the samples are drawn from $\frac{1}{2}N(1, 1) + \frac{1}{2}N(1, -1)$. The results are shown in Figure 8.3. In this case, the estimation error decays faster than the one shown in Figure 8.2. The larger absolute values of the Wasserstein distance is an artifact of the range of the means.

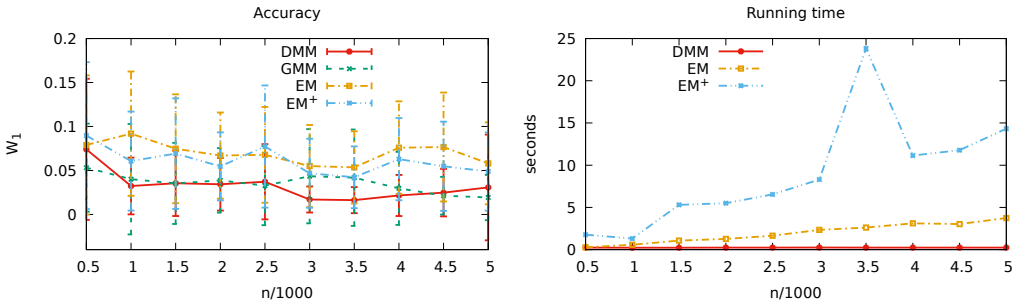


Figure 8.2: Comparison of different methods when components completely overlap.

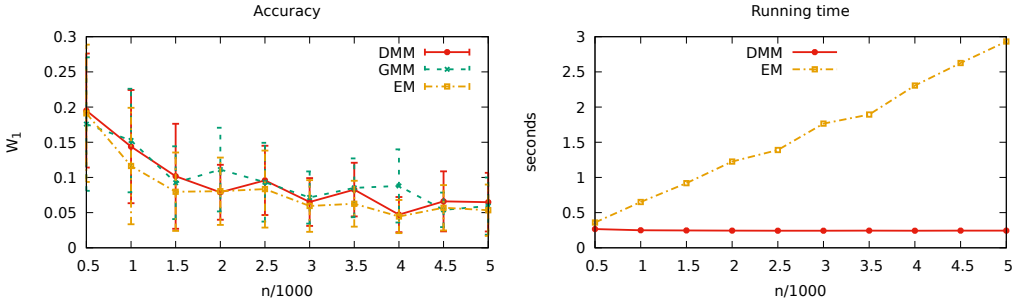


Figure 8.3: Comparison of different methods when components are separated.

Unknown variance. We conduct an experiment under the same five-component Gaussian mixture as before, but now the estimators no longer have access to the true variance parameter. In this case, Lindsay's algorithm (see Algorithm 8.2) involves the empirical moments of degrees up to 10, among which higher-order moments are hard to estimate with limited samples. Indeed, the standard deviation of $\mathbb{E}_n[X^{10}]$ is $\frac{1}{\sqrt{n}}\sqrt{\mathsf{var}[X^{10}]} \approx 473$ under this specified model with $n = 5000$ samples. To resolve this issue, we introduce an extra screening threshold τ to determine whether an empirical moment is too noisy and accept the empirical moment of order j only when its empirical variance satisfies

$$\frac{\mathbb{E}_n[X^{2j}] - (\mathbb{E}_n[X^j])^2}{n} \leq \tau, \tag{8.23}$$

where the left-hand side of (8.23) is an estimate of the variance of $\mathbb{E}_n[X^j]$. The estimated mixture model consists of $\tilde{k}$ components for the largest $\tilde{k}$ such that the first $2\tilde{k}$ empirical moments are all accepted. In the experiment, we

choose $\tau = 0.5$. The results are shown in Figure 8.4. The performance of

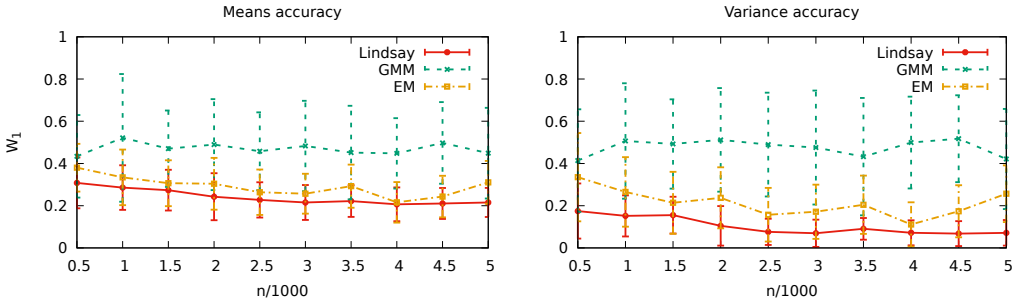


Figure 8.4: Comparison of different methods with unknown variance.

the Lindsay and EM estimators are similar and better than GMM, which is possibly due to the large variance of higher-order empirical moments. The running time comparison are similar to before and thus are omitted. The experiments under the models of Figure 8.2 and Figure 8.3 also yield similar results.

8.3 Lower bounds for estimating Gaussian mixtures

This section introduces minimax lower bounds for estimating Gaussian location mixture models which certify the optimality of our estimators. We will apply Le Cam's two-point method, namely, find two Gaussian mixtures that are statistically close but have different parameters. Then any estimate suffers a loss at least proportional to their parameter difference.

To show a vanishing statistical distance between two mixture models, one commonly used proxy is *moment matching*, i.e., $\nu * N(0,1)$ and $\nu * N(0,1)$ are statistically close if $\mathbf{m}_\ell(\nu) = \mathbf{m}_\ell(\nu')$ for some large ℓ. This is demonstrated in Figure 3.1, and is made precise in Theorem 3.4. The best lower bound follows from two different mixing distributions ν and ν' such that $\mathbf{m}_\ell(\nu) = \mathbf{m}_\ell(\nu')$ with the largest degree ℓ, which is $2k-2$ when both distributions are k-atomic and $2k-1$ when one of them is k-atomic (see Lemma 7.1 and the following Remark 7.1). Next we provide the precise minimax lower bounds for the case of known and unknown variance separately.

Known variance. We shall assume a unit variance. First, we define the space of all k Gaussian location mixtures as

$$\mathcal{P}_k = \{\nu * N(0,1) : \nu \text{ is } k\text{-atomic supported on } [-1,1]\},$$

and we consider the worst-case risk over all mixture models in $\mathcal{P}_k$. From the identifiability of discrete distributions in Lemma 7.1, two different k-atomic distributions can match up to $2k-2$ moments. Therefore, using Theorem 3.4, the best minimax lower bound using Le Cam's method is obtained from the optimal pair of distributions for the following:

$$\begin{aligned} &\max W_1(\nu,\nu'), \\ &\text{s.t. } \mathbf{m}_{2k-2}(\nu) = \mathbf{m}_{2k-2}(\nu'), \\ &\quad \nu,\nu' \text{ are } k\text{-atomic on } [-\epsilon,\epsilon]. \end{aligned} \tag{8.24}$$

The value of the above optimization problem is $\Omega(\epsilon/k)$ (see Lemma 8.20). Using $\epsilon = \sqrt{k}n^{-\frac{1}{4k-2}}$, we obtain the following minimax lower bound.

Proposition 8.3.

$$\inf_{\hat{\nu}} \sup_{P\in\mathcal{P}_k} \mathbb{E}_P W_1(\nu,\hat{\nu}) \geq \Omega\left(\frac{1}{\sqrt{k}} n^{-\frac{1}{4k-2}}\right),$$

where $\hat{\nu}$ is an estimator measurable with respect to $X_1,\ldots,X_n \overset{i.i.d.}{\sim} P = \nu * N(0,1)$.

Remark 8.1. The above lower bound argument can be easily extended to prove the optimality of (8.10) in Theorem 8.2, where the mixture satisfies further separation conditions in the sense of Definition 8.1. In this case, the main difficulty is to estimate parameters in the biggest cluster. When there are k_0 γ-separated clusters, the biggest cluster is of order at most $k' = k - k_0 + 1$. Similar to (8.24), let $\tilde{\nu}$ and $\tilde{\nu}'$ be two k'-atomic distributions on $[-\epsilon,\epsilon]$. Consider the following mixing distributions

$$\nu = \frac{k_0-1}{k_0}\nu_0 + \frac{1}{k_0}\tilde{\nu}, \quad \nu' = \frac{k_0-1}{k_0}\nu_0 + \frac{1}{k_0}\tilde{\nu}',$$

where ν_0 is the uniform distribution over $\{\pm 2\gamma, \pm 3\gamma, \ldots\}$ of cardinality k_0-1. Then both mixture models have k_0 $(\gamma, \frac{1}{k_0})$-separated clusters. Thus the min-

imax lower bound $\Omega(\frac{1}{\sqrt{k'}}n^{-\frac{1}{4k'-2}})$ analogously follows from Le Cam's method.

Unknown variance. In this case the collection of mixture models is defined as

$$\mathcal{P}'_k = \{\nu * N(0,\sigma^2) : \nu \text{ is } k\text{-atomic supported on } [-1,1],\ \sigma \le 1\}.$$

In Theorem 3.4, mixing distributions are not restricted to be k-atomic but can be Gaussian location mixtures themselves, thanks to the infinite divisibility of the Gaussian distributions, e.g., $N(0,\epsilon^2) * N(0,0.5) = N(0,0.5+\epsilon^2)$. Let g_k be the k-point Gauss quadrature of $N(0,\epsilon^2)$. Then g_k has the same first $2k-1$ moments as $N(0,\epsilon^2)$, and $g_k * N(0,0.5)$ is a k Gaussian mixture. Applying (3.10) yields that

$$\chi^2(g_k * N(0,1) \| N(0,1+\epsilon^2)) \le O(\epsilon^{4k}).$$

Using $W_1(g_k,\delta_0) \ge \Omega(\epsilon/\sqrt{k})$ (see Lemma 2.2), and choosing $\epsilon = n^{-\frac{1}{4k}}$, we obtain the following minimax lower bound.

Proposition 8.4. *For $k \ge 2$,*

$$\inf_{\hat\nu}\sup_{P\in\mathcal{P}_k} \mathbb{E}_P W_1(\nu,\hat\nu) \ge \Omega\left(\frac{1}{\sqrt{k}}n^{-\frac{1}{4k}}\right),$$
$$\inf_{\hat\nu}\sup_{P\in\mathcal{P}_k} \mathbb{E}_P|\sigma^2-\hat\sigma^2| \ge \Omega\left(n^{-\frac{1}{2k}}\right),$$

where the infimum is taken over estimators $\hat\nu, \hat\sigma^2$ measurable with respect to $X_1,\dots,X_n \overset{i.i.d.}{\sim} P = \nu * N(0,\sigma^2)$.

8.4 Extensions and discussions

8.4.1 Gaussian location-scale mixtures

In this chapter we focus on the Gaussian location mixture model (8.1), where all components share the same (possibly unknown) variance. One immediate extension is the Gaussian location-scale mixture model with heteroscedastic

components:

$$\sum_{i=1}^{k} w_i N(\mu_i, \sigma_i^2). \tag{8.25}$$

Parameter estimation for this model turns out to be significantly more difficult than the location mixture model, in particular:

- The likelihood function is unbounded. In fact, it is well known that the maximum likelihood estimator is ill-defined [178, p. 905]. For instance, consider $k = 2$, for any sample size n, we have

$$\sup_{p_1,p_2,\theta_1,\theta_2,\sigma} \prod_{i=1}^{n} \left[\frac{p_1}{\sigma_1} \varphi\left(\frac{X_i - \theta_1}{\sigma_1} \right) + \frac{p_2}{\sigma_2} \varphi\left(\frac{X_i - \theta_2}{\sigma_2} \right) \right] = \infty,$$

  achieved by, e.g., $\theta_1 = X_1, p_1 = 1/2$, $\sigma_2 = 1$, and $\sigma_1 \to 0$.

- In this model, the identifiability result based on moments is not completely settled and we do not have a counterpart of Lemma 7.1. Note that the model (8.25) comprises $3k - 1$ free parameters (k means, k variances, and k weights normalized to one), so it is expected to be identified through its first $3k - 1$ moments. However, the intuition of equating the number of parameters and the number of equations is already known to be wrong as pointed out by Pearson [11], who showed that for $k = 2$, five moments are insufficient and six moments are enough. The recent result [195] showed that, if the parameters are in general positions, then $3k - 1$ moments can identify the Gaussian mixture distribution up to finitely many solutions (known as algebraic identifiability). Whether $3k$ moments can uniquely identify the model (known as rational identifiability) in general positions remains open, except for $k = 2$. In the worst case, we need at least $4k - 2$ moments for identifiability since for scale-only Gaussian mixtures all odd moments are zero (see Section 8.4.3 for details).

Besides the issue of identifiability, the optimal estimation rate under the Gaussian location-scale mixture model is resolved only in special cases. The sharp rate is only known in the case of two components to be $\Theta(n^{-1/12})$ for estimating means and $\Theta(n^{-1/6})$ for estimating variances [31], achieved by a robust variation of Pearson's method of moment equations [11]. For k components, the optimal rate is known to be $n^{-\Theta(1/k)}$ [30, 29], achieved

by an exhaustive grid search on the parameter space. In addition, the above results all aim to recover parameters of all components (up to a global permutation), which necessarily requires many assumptions including lower bounds on mixing weights and separation between components; recovering the mixing distribution with respect to, say, Wasserstein distance, remains open.

8.4.2 Multiple dimensions

So far we have focused on Gaussian mixtures in one dimension. The multivariate version of this problem has been studied in the context of clustering, or classification, which typically requires nonoverlapping components [162, 196]. One commonly used approach is dimensionality reduction: projecting data onto some lower-dimensional subspace, clustering samples in that subspace, and mapping back to the original space. Common choices of the subspace include random subspaces and subspaces obtained from the singular value decomposition. The approach using random subspace is analyzed in [162, 197], and requires a pairwise separation polynomial in the dimensions; the subspace from singular value decomposition is analyzed in [196, 198, 199, 200], and requires a pairwise separation that grows polynomially in the number of components. Tensor decomposition for spherical Gaussian mixtures has been studied in [201], which requires the stronger assumption that that means are linear independent and is inapplicable in lower dimensions, say, two or three dimensions.

When components are allowed to overlap significantly, the random projection approach is also adopted by [30, 29, 31], where the estimation problem in high dimensions is reduced to that in one dimension, so that univariate methodologies can be invoked as a primitive. We provide an algorithm (Algorithm 8.4) using similar random projection ideas to estimate the parameters of a Gaussian mixture model in d dimensions for known covariance matrices, using the univariate algorithm in Section 8.2.1 as a subroutine, and obtain the estimation guarantee in Theorem 8.7; the unknown covariance case can be handled analogously using the algorithm in Section 8.2.2 instead. However, the dependency of the performance guarantee on the dimension is highly

suboptimal,[8] which stems from the fact that the method based on random projections estimates each coordinate independently. Moreover, this method needs to match the Gaussian components of the estimated model in each random direction, which necessarily requires lower bounds on the mixing weights and separation between the means.

Algorithm 8.4 Learning a k-component Gaussian mixture in d dimensions.

Input: n samples $X_1, X_2, \ldots, X_n \in \mathbb{R}^d$, common covariance matrix Σ, and separation parameter τ, radius parameter ρ.

Output: estimated mixing distribution $\hat{\pi}$ with weights and means $(\hat{w}_j, \hat{\mu}_j)$ for $j = 1, \ldots, k$.

Let $(b_1, \ldots, b_d)$ be a set of random orthonormal basis in $\mathbb{R}^d$, and $r = b_1$.
Let $\{(w_j, \mu_j)\}$ be the outcome of Algorithm 8.1 using n projected samples $\langle X_1, r\rangle, \ldots, \langle X_n, r\rangle$, variance $r^\top \Sigma r$, and interval $[-\rho, \rho]$.
Reordering the indices such that $\mu_1 < \mu_2 < \cdots < \mu_k$.
Initialize k weights $\hat{w}_j = w_j$ and means $\hat{\mu}_j = (0, \ldots, 0)$.
for $i = 1$ **to** d **do**
    Let $r' = r + \tau b_i$.
    Let $\{\mu'_j\}$ be the estimated means (weights are ignored) from Algorithm 8.1 using n projected samples $\langle X_1, r'\rangle, \ldots, \langle X_n, r'\rangle$, variance $r'^\top \Sigma r'$, and interval $[-\rho - \tau, \rho + \tau]$.
    Reordering the indices such that $\mu'_1 < \mu'_2 < \cdots < \mu'_k$.
    Let $\hat{\mu}_j := \hat{\mu}_j + b_i \frac{\mu'_j - \mu_j}{\tau}$ for $j = 1, \ldots, k$.
end for

Theorem 8.7. *Suppose in a d-dimensional Gaussian mixture $\sum_{j=1}^k w_j N(\mu_j, \Sigma)$,*

$$\|\mu_j\|_2 \le M, \quad \|\mu_i - \mu_j\|_2 \le \epsilon, \quad w_j \ge \epsilon', \quad \forall\ i \ne j.$$

Then Algorithm 8.4 with $n > (\Omega_k(\frac{M}{\tilde{\epsilon}\epsilon'}))^{4k-2} \log \frac{d}{\delta}$ samples, $\tau = \frac{\tilde{\epsilon}}{2M}$, and $\rho = M$, where $\tilde{\epsilon} = \frac{\delta\epsilon}{k^2\sqrt{d}}$, yields $\hat{\pi}$ such that, with probability $1 - 2\delta$,

$$W_1(\pi, \hat{\pi}) < O_k\left(\sqrt{d}\frac{M\epsilon_n}{\tau\epsilon'}\right),$$

where $\pi = \sum_j w_j \delta_{\mu_j}$ and $\epsilon_n = \min\{(\frac{n}{\log(d/\delta)})^{-\frac{1}{4k-2}}, \tilde{\epsilon}^{2-2k}\sqrt{\frac{\log(d/\delta)}{n}}\}$.

[8] Specifically, in d dimensions, estimating each coordinate independently suffers an ℓ_2-loss proportional to $\sqrt{d}$; however, it is possible to achieve $d^{1/4}$. See Lemma 8.32 for an example.

Proof. By the distribution of random direction r on the unit sphere (see Lemma 8.31) and the union bound, we obtain that, with probability $1-\delta$,

$$|\langle \mu_i - \mu_j, r\rangle| > 2\tilde{\epsilon}, \quad \forall\, i \neq j.$$

Without loss of generality, assume $\langle \mu_1, r\rangle < \cdots < \langle \mu_k, r\rangle$. Applying Theorem 8.1 yields that, with probability $1 - \frac{\delta}{d+1}$,

$$W_1(\pi_r, \hat{\pi}_r) \leq O_k\left(M\left(\frac{n}{\log(d/\delta)}\right)^{-\frac{1}{4k-2}}\right),$$

where π_r denotes the Gaussian mixture projected on r and $\hat{\pi}_r$ is its estimate. The right-hand side of the above inequality is less than $c\epsilon_r\epsilon'$ for some constant $c < 0.5$ when $n > (\Omega_k(\frac{M}{\tilde{\epsilon}\epsilon'}))^{4k-2}\log\frac{d}{\delta}$. Applying Theorem 8.2 yields that

$$W_1(\pi_r, \hat{\pi}_r) \leq O_k\left(M\tilde{\epsilon}^{2-2k}\sqrt{\frac{\log(d/\delta)}{n}}\right).$$

Hence, we obtained $W_1(\pi_r, \hat{\pi}_r) \leq O_k(M\epsilon_n)$. It follows from Lemma 6.1 that, after reordering indices,

$$|\langle \mu_j, r\rangle - \tilde{\mu}_j| < O_k(M\epsilon_n/\epsilon'), \quad |w_j - \hat{w}_j| < O_k(M\epsilon_n/\tilde{\epsilon}). \tag{8.26}$$

On each direction $r_\ell = r + \tau b_\ell$, the means are separated by $|\langle \mu_i - \mu_j, r_\ell\rangle| > 2\tilde{\epsilon} - 2M\tau > \tilde{\epsilon}$ and the ordering of the means remains the same as on direction r. Therefore the accuracy similar to (8.26) continues to hold for the estimated means $\tilde{\mu}_{\ell,j}$ (μ'_j in lines 7 and 8). Note that $\mu_j = \sum_\ell b_\ell \frac{\langle \mu_j, r_\ell\rangle - \langle \mu_j, r\rangle}{\tau}$ and $\hat{\mu}_j = \sum_\ell b_\ell \frac{\tilde{\mu}_{\ell,j} - \tilde{\mu}_j}{\tau}$. Therefore,

$$\|\hat{\mu}_j - \mu_j\|_2^2 \leq \sum_{\ell=1}^{d}\left(\frac{O_k(M\epsilon_n/\epsilon')}{\tau}\right)^2.$$

Applying the triangle equality yields that

$$W_1(\pi, \hat{\pi}) < \sqrt{d}O_k(M\epsilon_n/\epsilon')/\tau + MO_k\left(M\epsilon_n/\tilde{\epsilon}\right) < O_k\left(\sqrt{d}\frac{M\epsilon_n}{\tau\epsilon'}\right). \qquad \square$$

It is interesting to directly extend the DMM methodology to multiple dimensions, which is challenging both theoretically and algorithmically:

- To apply our method in multiple dimensions, the challenge is to obtain a multidimensional moment comparison theorem analogous to Proposition 7.1 or 7.2, the key step leading to the optimal rate. These results are proved by the primal formulation of the Wasserstein distance and its simple formula (6.4) in one dimension [161]. Alternatively, they can be proved via the dual formula (6.3) which holds in any dimension; however, the proof relies on the Newton's interpolation formula, which is again difficult to generalize or analyze in multiple dimensions.

- To obtain a computationally efficient algorithm, we rely on the semidefinite characterization of the moment space in one dimension to denoise the noisy estimates of moments. In multiple dimensions, however, it remains open how to efficiently describe the moment space [9] as well as how to extend the Gauss quadrature rule to multivariate distributions.

8.4.3 General finite mixtures

Though this chapter focuses on Gaussian location mixture models, the moments comparison theorems in Chapter 7 are independent of properties of Gaussian. As long as moments of the mixing distribution are estimated accurately, similar theory and algorithms can be obtained. Unbiased estimate of moments exists in many useful mixture models, including exponential mixtures [202], Poisson mixtures [203], and more generally the quadratic variance exponential family (QVEF) whose variance is at most a quadratic function of the mean [204, (8.8)].

As a closely related topic of this chapter, we discuss the Gaussian scale mixture model in detail, which has been extensively studied in the statistics literature [205] and is widely used in image and video processing [206, 207]. In a Gaussian scale mixture, a sample is distributed as

$$X \sim \sum_{i=1}^{k} w_i N(0, \sigma_i^2) = \int N(0, \sigma^2) \mathrm{d}\nu(\sigma^2),$$

where $\nu = \sum_{i=1}^{k} w_i \delta_{\sigma_i^2}$ is a k-atomic mixing distribution. Equivalently, a sample can be represented as $X = \sqrt{V} Z$, where $V \sim \nu$ and Z is standard normal independent of V. In this model, samples from different components

significantly overlap, so clustering-based algorithms will fail. Nevertheless, moments of ν can be easily estimated, for instance, using $\mathbb{E}_n[X^{2r}]/\mathbb{E}[Z^{2r}]$ for $m_r(\nu)$ with accuracy $O_r(1/\sqrt{n})$. Applying a similar algorithm to DMM in Section 8.2.1, we obtain an estimate $\hat{\nu}$ such that

$$W_1(\nu, \hat{\nu}) \le O_k(n^{-\frac{1}{4k-2}}),$$

with high probability.

Moreover, using a recipe similar to that in Section 8.3, a minimax lower bound can be established. Analogous to (8.24), let ν and ν' be a pair of k-atomic distributions supported on $[0, \epsilon]$ such that they match the first $2k-2$ moments, and let

$$\pi = \int N(0, \sigma^2) \mathrm{d}\nu(\sigma^2), \quad \pi' = \int N(0, \sigma^2) \mathrm{d}\nu'(\sigma^2),$$

which match their first $4k-3$ moments and are $\sqrt{\epsilon}$-subgaussian. Applying Theorem 3.4 with $\pi * N(0, 0.5)$, $\pi' * N(0, 0.5)$, and $\epsilon = O_k(n^{-\frac{1}{4k-2}})$ yields a minimax lower bound

$$\inf_{\hat{\nu}} \sup_{P \in \mathcal{G}_k} \mathbb{E}_P W_1(\nu, \hat{\nu}) \ge \Omega_k\left(n^{-\frac{1}{4k-2}}\right),$$

where the estimator $\hat{\nu}$ is measurable with respect to $X_1, \ldots, X_n \sim P$, and the space of k Gaussian scale mixtures is defined as

$$\mathcal{G}_k = \left\{ \int N(0, \sigma^2) \mathrm{d}\nu(\sigma^2) : \nu \text{ is } k\text{-atomic supported on } [0, 1] \right\}.$$

8.5 Denoising an empirical distribution

In this section, we consider the related problem of denoising an empirical distribution. Given noisy data $X_i = \theta_i + Z_i$ for $i = 1, \ldots, n$, where $Z_i \sim N(0, 1)$ is an independent Gaussian noise, the goal is to estimate the histogram of $\theta = (\theta_1, \ldots, \theta_n)$, namely, the probability distribution

$$\pi_\theta = \frac{1}{n} \sum_{i=1}^{n} \delta_{\theta_i}.$$

Using the framework of Chapter 6, we will estimate the CDF under the Wasserstein distance. Comparing to Gaussian mixture models, here the mixing distribution is given by the histogram π_θ.

In this problem, the plug-in estimator is the empirical distribution of data. However, the plug-in approach is inconsistent.

Theorem 8.8. *Let $\theta = (\theta_1, \dots, \theta_n) \in \Theta^n$, $X = (X_1, \dots, X_n)$ and $X_i \overset{ind}{\sim} P_{\theta_i}$. Then,*

$$\sup_{\theta \in \Theta^n} \mathbb{E}[W_p^p(\pi_\theta, \pi_X)] = \sup_{\theta \in \Theta} \mathbb{E}|X - \theta|^p.$$

Proof. By the naive coupling that $P_{X|\theta_i} = P_{\theta_i}$,

$$\mathbb{E}[W_p^p(\pi_\theta, \pi_X)] \le \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n |\theta_i - X_i|^p\right] \le \sup_{\theta \in \Theta} \mathbb{E}|X - \theta|^p.$$

Consider $\theta = (\theta_0, \dots, \theta_0)$, hence $\pi_\theta = \delta_{\theta_0}$.

$$\begin{aligned}\sup_{\theta \in \Theta^n} \mathbb{E}[W_p^p(\pi_\theta, \pi_X)] \ge \sup_{\theta_0 \in \Theta} \mathbb{E}[W_p^p(\delta_{\theta_0}, \pi_X)] &= \sup_{\theta_0 \in \Theta} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n |\theta_0 - X_i|^p\right] \\ &= \sup_{\theta \in \Theta} \mathbb{E}|X - \theta|^p. \qquad \square\end{aligned}$$

Corollary 8.1. *Suppose $\Theta \neq \emptyset$. Let $\theta = (\theta_1, \dots, \theta_n) \in \Theta^n$ and $X = (X_1, \dots, X_n) \sim N(\theta, I_n)$. Then,*

$$\sup_{\theta \in \Theta^n} \mathbb{E}[W_p^p(\pi_\theta, \pi_X)] = \mathbb{E}|Z|^p, \quad \forall\ p \ge 1,$$

where $Z \sim N(0, 1)$.

In this section, we will use the moment-based method to denoise the empirical distribution.

8.5.1 Estimation of the empirical moments

The estimation of the empirical moments is the same as estimating the moments of the mixing distribution in Chapter 8. The unbiased estimator of the k^{th} empirical moment $m_k(\pi_\theta) = \frac{1}{n}\sum_{i=1}^n \theta_i^k$ is

$$\tilde{m}_k = \frac{1}{n}\sum_{i=1}^n H_k(X_i). \tag{8.27}$$

The variance of $H_k(X_i)$ is related to the Laguerre polynomials (2.22) by

$$\mathsf{var}[H_k(X_i)] = k!\mathcal{L}_k(-\theta_i^2) - \theta^{2k} = k!\sum_{j=0}^{k-1}\binom{k}{j}\frac{\theta_i^{2j}}{j!}. \tag{8.28}$$

Higher moments of the Hermite polynomials are obtained in Lemma 8.4.

Lemma 8.4. *Let $X \sim N(\theta, 1)$. For all $k, t \geq 1$,*

$$\mathbb{E}|H_k(X) - \theta^k|^t \leq 2^t\mathbb{E}|H_k(X)|^t < \frac{2^t}{3}\left((3\theta)^{kt} + 2\sqrt{2}\left(\frac{9kt}{e}\right)^{kt/2}\right). \tag{8.29}$$

Proof. Let X' be the i.i.d. copy of X. By Jensen's inequality,

$$\begin{aligned}\mathbb{E}|H_k(X) - \theta^k|^t = \mathbb{E}_X|\mathbb{E}_{X'}[H_k(X) - H_k(X')]|^t &\leq \mathbb{E}|H_k(X) - H_k(X')|^t \\ &\leq 2^t\mathbb{E}|H_k(X)|^t,\end{aligned}$$

which is the first inequality of (8.29).

Note one representation of Hermite polynomials that $H_k(x) = \mathbb{E}(x+iW)^k$, where $W \sim N(0,1)$. Then, by Jensen's inequality,

$$\mathbb{E}|H_k(X)|^t = \mathbb{E}_X|\mathbb{E}_W(X+iW)^k|^t \leq \mathbb{E}_X|\mathbb{E}_W|X+iW|^k|^t \leq \mathbb{E}|X+iW|^{kt},\ t \geq 1. \tag{8.30}$$

The right-hand side of (8.30) can be further upper bounded by

$$\mathbb{E}|\theta+W'+iW|^{kt} \leq 3^{kt-1}(\theta^{kt}+2\mathbb{E}|W|^{kt}) = 3^{kt-1}\left(\theta^{kt} + \frac{2}{\sqrt{\pi}}\Gamma\left(\frac{kt+1}{2}\right)2^{kt/2}\right).$$

The conclusion follows by the upper bound of gamma function that $\Gamma(x+1) < \sqrt{2\pi}(\frac{x+1/2}{e})^{x+1/2}$ [208]. □

Using the higher moments in Lemma 8.4, we obtain the following concentration inequalities on $\tilde{m}_k$.

Lemma 8.5. *Suppose $\theta \in [-M, M]^n$.*

$$\mathbb{P}[|\tilde{m}_k - m_k(\pi_\theta)| \geq \epsilon] \leq \frac{k!\mathcal{L}_k(-M^2) - M^{2k}}{n\epsilon^2}. \tag{8.31}$$

If $n\epsilon^2 \geq 144e(18k)^k$,

$$\mathbb{P}[|\tilde{m}_k - m_k(\pi_\theta)| \geq \epsilon] \leq 2\exp\left(-\frac{1}{18}\left(\frac{n\epsilon^2}{72}\right)^{\frac{1}{k+1}}\left(1 \wedge \frac{\log(\frac{kn\epsilon^2}{8(3M)^{2k+2}})}{k+1}\right)\right). \tag{8.32}$$

If $n\epsilon^2 \geq 144e(3M)^{2k}$,

$$\mathbb{P}[|\tilde{m}_k - m_k(\pi_\theta)| \geq \epsilon] \leq 2\exp\left(-\frac{n\epsilon^2}{144e}\frac{1}{(3M)^{2k}}\left(1 \wedge k\log\frac{8e^2(3M)^{2k+2}}{kn\epsilon^2}\right)\right). \tag{8.33}$$

Proof. By Markov inequality,

$$\begin{aligned}\mathbb{P}[|\tilde{m}_k - m_k(\pi_\theta)| \geq \epsilon] &\leq \inf_t \frac{\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n (H_k(X_i) - \theta_i^k)\right|^t}{\epsilon^t} \\ &= \inf_t \frac{\mathbb{E}\left|\sum_{i=1}^n (H_k(X_i) - \theta_i^k)\right|^t}{(n\epsilon)^t}.\end{aligned}$$

The first conclusion (8.31) follows by $t = 2$ (i.e., Chebyshev inequality) and the variance of $\tilde{m}_k$ in (8.28).

Applying Marcinkiewicz-Zygmund inequality that,

$$\begin{aligned}\mathbb{E}\left|\sum_{i=1}^n (H_k(X_i) - \theta_i^k)\right|^t &\leq C(t)n^{t/2-1}\sum_{i=1}^n \mathbb{E}|H_k(X_i) - \theta_i^k|^t \\ &\leq C(t)n^{t/2}\sup_\theta \mathbb{E}|H_k(X_i) - \theta_i^k|^t,\ t \geq 2,\end{aligned}$$

where $C(t) \leq (3\sqrt{2t})^t$ [209], and the moment bound in (8.29), we have

$$\begin{aligned}\mathbb{P}[|\tilde{m}_k - m_k(\pi_\theta)| \geq \epsilon] &\leq \inf_{t\geq 2}\left(\frac{18t}{n\epsilon^2}\right)^{t/2}\frac{2^t}{3}\left((3M)^{kt} + 2\sqrt{2}\left(\frac{9kt}{e}\right)^{kt/2}\right) \\ &\leq 2\inf_{t\geq 2}\left(\frac{72}{n\epsilon^2}\left(\frac{9k}{e}\right)^k t^{k+1}\right)^{t/2} \vee \left(\frac{72}{n\epsilon^2}(3M)^{2k}t\right)^{t/2}.\end{aligned}$$

Then, (8.32) follows by letting $t = \frac{1}{e}(\frac{n\epsilon^2}{72}(\frac{e}{9k})^k)^{\frac{1}{k+1}}$ and applying $(\frac{e}{9k})^{\frac{k}{k+1}} \geq \frac{e}{9k}$, and (8.33) follows by letting $t = \frac{1}{e}\frac{n\epsilon^2}{72}\frac{1}{(3M)^{2k}}$. □

8.5.2 Denoising via Bernstein polynomials

Let $\pi_\theta = \frac{1}{n}\sum_{i=1}^n \delta_{\theta_i}$. For conciseness assume π_θ is supported on $[0,1]$. Suppose the moments of π_θ of degrees up to L are known, we can approximate π_θ by a probability measure $\pi^{(L)}$ supported on equidistant partition of $[0,1]$, namely, $\{i/L : i = 0, 1, \ldots, L\}$, using Bernstein polynomial approximation. Denote by $\theta^{(L)}$ the random variable associating with $\pi^{(L)}$. The probability mass function (pmf) of $\theta^{(L)}$ is given by

$$\begin{aligned} p_k^{(L)} &= \mathbb{P}[\theta^{(L)} = k/L] = \mathbb{E}_{X\sim\pi_\theta}[\mathbb{P}[\text{binomial}(L, X) = k]] \\ &= \binom{L}{k}\sum_{j=k}^{L}\binom{L-k}{j-k}(-1)^{j-k}m_j, \end{aligned} \tag{8.34}$$

for $k = 0, \ldots, L$, where $\mathbb{P}[\text{binomial}(L, X) = k] = \binom{L}{k}X^k(1-X)^{L-k}$ is the Bernstein basis polynomial and $m_j = \frac{1}{n}\sum_i \theta_i^j$ is the jth moment of π_θ. The intuition is that, for any fixed $\alpha \in [0,1]$, by the law of large numbers, as $L \to \infty$,

$$\begin{aligned} \mathbb{P}[\theta^{(L)} \le \alpha] &= \mathbb{E}_{X\sim\pi_\theta}[\mathbb{P}[\text{binomial}(L, X) \le \alpha L]] \\ &= \frac{1}{n}\sum_{i=1}^n \mathbb{P}[\text{binomial}(L, \theta_i) \le \alpha L] \to \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{\theta_i \le \alpha\}}. \end{aligned}$$

We can upper bound the approximation error $W_p(\pi_\theta, \pi^{(L)})$ by the natural coupling from the construction of $P_{\theta^{(L)}}$ that $P_{\theta^{(L)}|X} = \text{binomial}(L, X)/L$.

Lemma 8.6. *For any $p \ge 1$,*

$$W_p(\pi_\theta, \pi^{(L)}) \le \frac{C_p}{\sqrt{L}}, \tag{8.35}$$

where C_p only depends on p, $C_p \le 1/2$ for $1 \le p \le 2$ and $C_p \le 3\sqrt{2p}$ for $p > 2$.

Proof. Let $X \sim \pi_\theta$ and $P_{\theta^{(L)}|X} = \text{binomial}(L, X)/L$. For $p = 2$,

$$W_2^2(\pi_\theta, \pi^{(L)}) \le \mathbb{E}(\theta^{(L)} - X)^2 = \mathbb{E}\left[\mathbb{E}\left[(\theta^{(L)} - X)^2 \middle| X\right]\right] = \mathbb{E}\left[\frac{X(1-X)}{L}\right] \le \frac{1}{4L}. \tag{8.36}$$

For $1 \le p \le 2$, Hölder's inequality and (8.36) imply that

$$W_p(\pi_\theta, \pi^{(L)}) \le W_2(\pi_\theta, \pi^{(L)}) \le \frac{1}{2\sqrt{L}}. \tag{8.37}$$

For $p > 2$, analogous to (8.36),

$$W_p^p(\pi_\theta, \pi^{(L)}) \le \mathbb{E}(\theta^{(L)} - X)^p = \mathbb{E}\left[\mathbb{E}\left[(\theta^{(L)} - X)^p \middle| X\right]\right] \le A_p \frac{L^{p/2}}{L^p} = \frac{A_p}{L^{p/2}}, \tag{8.38}$$

where the second inequality follows from Marcinkiewicz-Zygmund inequality, A_p is a constant that only depends on p and $A_p \le (3\sqrt{2p})^p$ [209]. The conclusion follows from (8.37) and (8.38). □

Using $X_i \overset{\text{ind}}{\sim} N(\theta_i, 1)$ instead of the true moments of π, we can estimate the moments by the $\tilde{m}_j$ as in (8.27), thereby estimating $p_k^{(L)}$ by

$$\tilde{p}_k = \binom{L}{k} \sum_{j=k}^{L} \binom{L-k}{j-k} (-1)^{j-k} \tilde{m}_j. \tag{8.39}$$

Then $\tilde{p} = (\tilde{p}_0, \ldots, \tilde{p}_L)$ is an unbiased estimator for $p^{(L)} = (p_0^{(L)}, \ldots, p_L^{(L)})$, and the risk of $\tilde{p}$ is shown in Lemma 8.7.

Lemma 8.7. *With probability* $1 - e^{-\Omega(L)}$,

$$\|\tilde{p} - p^{(L)}\|_1 \le \sqrt{\frac{O(L)^{L+1}}{n}}. \tag{8.40}$$

Proof. Applying (8.32) with $\epsilon = \sqrt{O(L)^{L+1}/n}$ and the union bound yields that, with probability $1 - e^{-\Omega(L)}$,

$$|\tilde{m}_k - m_k| < \epsilon = \sqrt{\frac{O(L)^{L+1}}{n}}, \quad \forall\ k = 1, \ldots, L.$$

Consequently, by (8.34) and (8.39),

$$\|\tilde{p} - p^{(L)}\|_1 < \sum_{k=0}^{L} \binom{L}{k} \sum_{j=k}^{L} \binom{L-k}{j-k} \epsilon = 3^L \epsilon = \sqrt{\frac{O(L)^{L+1}}{n}}. \qquad \square$$

Note that $\tilde{p} = (\tilde{p}_0, \ldots, \tilde{p}_L)$ defined above may not be a valid pmf. Nevertheless we can project it onto a valid pmf under ℓ_1-distance: find a valid pmf

$\hat{p} = (\hat{p}_0, \ldots, \hat{p}_L)$ that minimizes $\|\tilde{p} - \hat{p}\|_1$. This is accomplished by simply thresholding at zero followed by normalization. By triangle inequality and the optimality of projection,

$$\|\hat{p} - p^{(L)}\|_1 \le \|\tilde{p} - \hat{p}\|_1 + \|\tilde{p} - p^{(L)}\|_1 \le 2\|\tilde{p} - p^{(L)}\|_1. \tag{8.41}$$

Denote by $\hat{\pi}_B$ the probability measure corresponding to $\hat{p}$. By picking $L = (1 + o(1))\frac{\log n}{\log\log n}$, we have the following upper bound on the risk of $\hat{\pi}_B$.

Theorem 8.9. *For any constant $p \ge 1$, with probability $1 - e^{-\Omega(\log n/\log\log n)}$,*

$$W_p(\pi_\theta, \hat{\pi}_B) \le C_p\sqrt{\frac{\log\log n}{\log n}},$$

where C_p only depends on p.

Proof. Since $\theta \in [0, 1]$, we can upper bound of $W_p^p(\hat{\pi}_B, \pi^{(L)})$ by the total variation distance:

$$W_p^p(\hat{\pi}_B, \pi^{(L)}) \le \mathsf{TV}(\hat{\pi}_B, \pi^{(L)}) = \frac{1}{2}\|\hat{p} - p^{(L)}\|_1 \le \|\tilde{p} - p^{(L)}\|_1, \tag{8.42}$$

where the last inequality follows from (8.41). Then, applying triangle inequality and (8.42), the approximation error of $\pi^{(L)}$ in (8.35) and the estimation error of $\tilde{p}$ in (8.40),

$$W_p(\pi_\theta, \hat{\pi}_B) \le W_p(\pi_\theta, \pi^{(L)}) + \|\tilde{p} - p^{(L)}\|_1^{1/p} \le \frac{C_p'}{\sqrt{L}} + \left(\sqrt{\frac{O(L)^{L+1}}{n}}\right)^{1/p},$$

with probability $1 - e^{-\Omega(L)}$. Let $L = \frac{(1+o(1))\log n}{\log\log n}$, we have $(\sqrt{O(L)^{L+1}/n})^{1/p} = o(1/\sqrt{L})$ and thus

$$W_p(\pi_\theta, \hat{\pi}_B) \le \frac{C_p'}{\sqrt{L}}(1 + o(1)) = C_p'(1 + o(1))\sqrt{\frac{\log\log n}{\log n}}.$$

The conclusion follows. □

8.5.3 Optimal denoising under W_1 distance

The estimator in Section 8.5.2 uses equidistant partition of the interval $[0,1]$ which might not be necessary in the optimal denoising. Consider $\theta \in [-M, M]^n$. Recall the dual representation of W_1 distance in (6.3). Let Lip(1) denote the set of functions with best Lipschitz constant one. The idea comes from the observation that if two probability measures match moments up to degree L, then their expectations of $f(X)$ are separated by at most twice the uniform approximation error of f by polynomial of degree no greater than L over the given interval [55]. By Jackson's theorem $E_L(f, [-M, M]) \lesssim M/L$ as long as $f \in \text{Lip}(1)$. Though the exact moments of π_θ are not available, if we can find $\hat{\pi}$ with moments sufficiently close to that of π_θ, then the expectations of $f(X)$ under π_θ and $\hat{\pi}$ are still guaranteed to be close to each other as shown in Lemma 8.8.

Lemma 8.8. *Let μ and ν be two probability measures supported on $[-M, M]$. Denote by $\mathsf{M}_L(\mu) = (\frac{m_1(\mu)}{M}, \ldots, \frac{m_L(\mu)}{M^L})$ the first L normalized moments of μ and similarly for $\mathsf{M}_L(\nu)$. Then*

$$W_1(\mu, \nu) \le \frac{\pi M}{L+1} + 2M(1+\sqrt{2})^L \left\| \mathsf{M}_L(\mu) - \mathsf{M}_L(\nu) \right\|_2 .$$

Proof. Fix any $f \in \text{Lip}(1)$. Let P_L^* be the best polynomial of degree L to uniformly approximate f over $[-M, M]$, and denote its coefficients by $a = (a_1, \ldots, a_L)$.

$$\begin{aligned}
|\mathbb{E}_\mu f - \mathbb{E}_\nu f| &\le |\mathbb{E}_\mu (f - P_L^*)| + |\mathbb{E}_\nu (f - P_L^*)| + |\mathbb{E}_\mu P_L^* - \mathbb{E}_\nu P_L^*| \\
&\le 2 \sup_{-M \le x \le M} |f(x) - P_L^*(x)| + \sum_{i=1}^{L} |a_i| |m_i(\mu) - m_i(\nu)| \\
&= 2E_L(f, [-M, M]) + \sum_{i=1}^{L} |a_i M^i| \left| \frac{m_i(\mu)}{M^i} - \frac{m_i(\nu)}{M^i} \right| \\
&\le \frac{M\pi}{L+1} + \|b\|_2 \|\mathsf{M}_L(\mu) - \mathsf{M}_L(\nu)\|_2,
\end{aligned}$$

where $b = (a_1 M, \ldots, a_L M^L)$ and we applied the upper bound on the uniform approximation error of Lip(1) functions [210, Theorem 4.1.1]

$$E_L(\text{Lip}(1), [-M, M]) = M E_L(\text{Lip}(1), [-1, 1]) \le \frac{M\pi}{2(L+1)}.$$

For any f with $\mathrm{Lip}(f) \le 1$, it has variation no more than $2M$ over $[-M, M]$ then by the optimality of P_L^* its variation is at most $4M$. Then, apart from the constant term, applying (8.61) yields that $\|b\|_2 \le 2M(1+\sqrt{2})^L$ and thus

$$|\mathbb{E}_\mu f - \mathbb{E}_\nu f| \le \frac{M\pi}{L+1} + 2M(1+\sqrt{2})^L \|\mathsf{M}_L(\mu) - \mathsf{M}_L(\nu)\|_2. \tag{8.43}$$

The conclusion follows by applying (8.43) in the dual representation of W_1 distance in (6.3). □

Remark 8.2. It is obtained by [211] that the sharp characterization of the uniform approximation of Lipschitz functions is

$$E_L(\mathrm{Lip}(1), [-1, 1]) = \frac{\pi - o(1)}{2L}.$$

Remark 8.3. If two probability measures match moments up to degree L, then

$$\begin{aligned}
\sup_{\mathsf{M}_L(\mu)=\mathsf{M}_L(\nu)} W_1(\mu, \nu) &= \sup_{\mathsf{M}_L(\mu)=\mathsf{M}_L(\nu)} \sup_{f \in \mathrm{Lip}(1)} |\mathbb{E}_\mu f - \mathbb{E}_\nu f| \\
&= \sup_{f \in \mathrm{Lip}(1)} \sup_{\mathsf{M}_L(\mu)=\mathsf{M}_L(\nu)} |\mathbb{E}_\mu f - \mathbb{E}_\nu f| \\
&= \sup_{f \in \mathrm{Lip}(1)} 2E_L(f, [-M, M]) = 2E_L(\mathrm{Lip}(1), [-M, M]),
\end{aligned}$$

where the third equality follows by the dual problem of best polynomial approximation.

Using the estimator for the empirical moments, let $\tilde{\mathsf{M}} = (\frac{\tilde{m}_1}{M}, \ldots, \frac{\tilde{m}_L}{M^L})$. We project $\tilde{m}$ to the space of moment sequence by (8.12) and obtain a corresponding estimator $\hat{\pi}$. Then, by the optimality of projection and the triangle inequality,

$$\|\mathsf{M}(\pi_\theta) - \mathsf{M}(\hat{\pi})\|_2 \le \|\mathsf{M}(\hat{\pi}) - \tilde{\mathsf{M}}\|_2 + \|\mathsf{M}(\pi_\theta) - \tilde{\mathsf{M}}\|_2 \le 2\|\mathsf{M}(\pi_\theta) - \tilde{\mathsf{M}}\|_2. \tag{8.44}$$

If M is a constant, we can pick $L = (1+o(1))\frac{\log n}{\log\log n}$ and obtain the following upper bound on the risk of $\hat{\pi}$.

Theorem 8.10. *Suppose $\theta \in [-M, M]^n$. If M is a constant, then, with*

probability $1 - e^{-\Omega(\log n/\log\log n)}$,

$$W_1(\pi_\theta, \hat{\pi}) \le \pi M \frac{\log\log n}{\log n}(1 + o(1)). \tag{8.45}$$

If $M = M_n \asymp \sqrt{\log n}$*, then, with high probability,*

$$W_1(\pi_\theta, \hat{\pi}) \lesssim \frac{1}{\sqrt{\log n}}. \tag{8.46}$$

Proof. Applying Lemma 8.8 and (8.44) yields that

$$W_1(\pi_\theta, \hat{\pi}) \le \frac{\pi M}{L+1} + 4M(1+\sqrt{2})^L \|m(\pi_\theta, M) - \tilde{m}\|_2.$$

If M is a constant, by picking $L = \frac{(1+o(1))\log n}{\log\log n}$ such that $(1+\sqrt{2})^L \sqrt{\frac{O(L)^{L+1}}{n}} = o(\frac{1}{L+1})$, we obtain (8.45). If $M = M_n \asymp \sqrt{\log n}$, applying the estimation error of $\tilde{m}$ in (8.33) with $\epsilon = \sqrt{O(M)^{2L+2}/n}$ and the union bound,

$$W_1(\pi_\theta, \hat{\pi}) \le \frac{\pi M}{L+1} + 4M(1+\sqrt{2})^L \sqrt{\frac{O(M)^{2L+2}}{n}}, \tag{8.47}$$

with probability $1 - \exp(-e^{\Omega(L)})$. □

8.5.4 Subsampling

Let $X = (X_1, \ldots, X_n)$ and $X_i \overset{\text{ind}}{\sim} N(\theta_i, 1)$. Let $Y = (Y_1, \ldots, Y_m)$ be m samples from X uniformly at random without replacement. The goal is to estimate π_θ from Y. Though Y_i are dependent, marginally, $Y_i \sim \pi_\theta * N(0,1)$. Hence, an unbiased estimator for the moments of π_θ is

$$\tilde{m}_k = \frac{1}{m}\sum_{i=1}^{m} H_k(Y_i).$$

Project the sequence $\tilde{m} = (\tilde{m}_1, \ldots, \tilde{m}_L)$ to a valid moment vector $\hat{m}$, and finally find the corresponding estimator $\hat{\pi}_{\text{sub}}$.

Theorem 8.11. *Let* $\theta \in [-M, M]^n$ *for any given constant* M*,*

$$\mathbb{E}[W_1(\pi_\theta, \hat{\pi}_{\text{sub}})] \le \pi M \frac{\log\log m}{\log m}(1 + o(1)).$$

Proof. Applying the same argument as (8.47) yields that

$$W_1(\pi_\theta, \hat{\pi}_{\text{sub}}) \le \frac{\pi M}{L+1} + 4M(1+\sqrt{2})^L \|m(\pi) - \tilde{m}\|_2.$$

Then, by Cauchy-Schwartz inequality,

$$\mathbb{E}[W_1(\pi_\theta, \hat{\pi}_{\text{sub}})] \le \frac{\pi M}{L+1} + 4M(1+\sqrt{2})^L \sqrt{\sum_{i=1}^{L} \mathbb{E}(\tilde{m}_k - m_k)^2}. \tag{8.48}$$

The MSE of the moment estimator $\tilde{m}_k$ is

$$\begin{aligned} &\mathbb{E}\left(\frac{1}{m}\sum_{i=1}^{m}(H_k(Y_i) - m_k)\right)^2 \\ &= \frac{1}{m}\mathsf{var}[H_k(Y_1)] + \frac{m-1}{m}\left(\mathbb{E}[H_k(Y_1)H_k(Y_2)] - m_k^2\right). \end{aligned} \tag{8.49}$$

Let $\{I, J\} \subseteq [n]$ be two indices taken uniformly at random. Then

$$\begin{aligned} &\mathbb{E}[H_k(Y_1)H_k(Y_2)] - m_k^2 = \mathbb{E}[H_k(X_I)H_k(X_J)] - m_k^2 = \mathbb{E}[\theta_I^k \theta_J^k] - m_k^2 \\ &= \frac{1}{n(n-1)}\sum_{i\neq j}\theta_i^k\theta_j^k - \left(\frac{1}{n}\sum_{i=1}^{n}\theta_i^k\right)^2 = \frac{1}{n-1}\left(\left(\frac{\sum_i \theta_i^k}{n}\right)^2 - \frac{\sum_i \theta_i^{2k}}{n}\right) \le 0. \end{aligned} \tag{8.50}$$

The variance of $H_k(Y_1)$ is

$$\mathbb{E}(H_k(Y_1) - m_k)^2 \le \mathbb{E}[H_k^2(X_I)] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[H_k^2(X_i)] = \frac{1}{n}\sum_{i=1}^{n} k!\mathcal{L}_k(-\theta_i^2), \tag{8.51}$$

where in the last step we used the second moment of $H_k(X_i)$. Plugging (8.50) and (8.51) into (8.49), we obtain the MSE of $\tilde{m}_k$ that

$$\mathbb{E}(\tilde{m}_k - m_k)^2 \le \frac{k!\mathcal{L}_k(-M^2)}{m}.$$

Then, applying (8.48), we obtain that

$$\mathbb{E}[W_1(\pi_\theta, \hat{\pi}_{\text{sub}})] \le \frac{\pi M}{L+1} + 4M(1+\sqrt{2})^L \sqrt{L\frac{L!\mathcal{L}_L(-M^2)}{m}}.$$

Picking $L = (1+o(1))\frac{\log m}{\log\log m}$, we have $(1+\sqrt{2})^L\sqrt{L\frac{L!\mathcal{L}_L(-M^2)}{m}} = o(\frac{1}{L+1})$, hence the conclusion. □

8.5.5 Minimax rates under W_1 distance

Two composite hypotheses. Recall the dual representation of W_1 distance:

$$W_1(\pi,\hat{\pi}) = \sup_{f:\mathrm{Lip}(f)\le 1} |\mathbb{E}_\pi[f(X)] - \mathbb{E}_{\hat{\pi}}[f(X)]|.$$

For any fixed function f satisfying $\mathrm{Lip}(f)\le 1$, the risk of estimating the additive functional $T(\theta) = \frac{1}{n}\sum_{i=1}^n f(\theta_i)$ also serves a lower bound of estimating π:

$$\begin{aligned}\inf_{\hat{\pi}}\sup_{\theta\in\Theta^n}\mathbb{E}(W_1(\pi,\hat{\pi}))^2 &\ge \inf_{\hat{\pi}}\sup_{\theta\in\Theta^n}\mathbb{E}\left(\mathbb{E}_\pi[f(X)] - \mathbb{E}_{\hat{\pi}}f(X)\right)^2\\ &\ge \inf_{\hat{T}}\sup_{\theta\in\Theta^n}\mathbb{E}\left(T(\theta)-\hat{T}\right)^2.\end{aligned}$$

For example, by taking $f(x) = |x|$, the minimax risk of estimating ℓ_1-norm of Gaussian mean [39] yields that

$$\begin{aligned}\inf_{\hat{\pi}}\sup_{\theta\in[-1,1]^n}\mathbb{E}(W_1(\pi,\hat{\pi}))^2 &\ge \beta_*^2\left(\frac{\log\log n}{\log n}\right)^2(1+o(1)),\\ \inf_{\hat{\pi}}\sup_{\theta\in\mathbb{R}^n}\mathbb{E}(W_1(\pi,\hat{\pi}))^2 &\ge \frac{4\beta_*^2}{9e^2\log n}(1+o(1)),\end{aligned}$$

where $\beta_* \approx 0.28017$ is the Bernstein constant.

Consider $\Theta = [-M,M]$ and $T(\theta) = \frac{1}{n}\sum_{i=1}^n f(\theta_i)$ with f being a function that achieves the approximation error

$$E_L(\mathrm{Lip}(1),[-M,M]) \triangleq \sup_{\mathrm{Lip}(f)\le 1} E_L(f,[-M,M]).$$

The sharp characterization of the above quantity is [211]

$$E_L(\mathrm{Lip}(1),[-M,M]) = M\cdot E_L(\mathrm{Lip}(1),[-1,1]) = M\frac{\pi-o(1)}{2(L+1)}. \tag{8.52}$$

The dual problem of the uniform approximation of f on $[-M,M]$ yields two probability measures supported on $[-M,M]$, denoted by μ and ν, that match

moments of degrees of $1, \ldots, L$, with functional values separated by

$$|\mathbb{E}_\mu f - \mathbb{E}_\nu f| = 2E_L(f, [-M, M]) = 2E_L(\mathrm{Lip}(1), [-M, M]) = \frac{M\pi}{L}(1 - o(1)).$$

Define two priors on $[-M, M]^n$ by $U = (U_1, \ldots, U_n) \sim \mu^{\otimes n}$ and $U' = (U'_1, \ldots, U'_n) \sim \nu^{\otimes n}$. By the separation property of μ and ν, the functional values are separated on average by

$$\mathbb{E}[T(U) - T(U')] = |\mathbb{E}_\mu f - \mathbb{E}_\nu f| = \frac{M\pi}{L}(1 - o(1)). \tag{8.53}$$

The marginal distributions of samples under two priors are n i.i.d. Gaussian mixtures $\mathbb{E}_{X\sim\mu}[N(X, 1)]^{\otimes n}$ and $\mathbb{E}_{X\sim\nu}[N(X, 1)]^{\otimes n}$, respectively. By the moment matching property of μ and ν, the Gaussian mixtures cannot be tested reliably, as shown in Lemma 8.9 [39].

Lemma 8.9. *Suppose μ, ν supported on $[-M, M]$ match moments of degree $1, \ldots, L$. Then*

$$\chi^2(\mathbb{E}_{X\sim\mu}[N(X, 1)] \| \mathbb{E}_{X\sim\nu}[N(X, 1)]) \le e^{M^2/2} \sum_{k>L} \frac{M^{2k}}{k!}. \tag{8.54}$$

Theorem 8.12. *If M is a constant, then*

$$\inf_{\hat\pi} \sup_{\theta\in[-M,M]^n} \mathbb{E}[W_1(\pi, \hat\pi)] \ge \frac{\pi M}{2} \frac{\log\log n}{\log n}(1 + o(1)). \tag{8.55}$$

If $M = M_n = \sqrt{\log n}$, then

$$\inf_{\hat\pi} \sup_{\theta\in[-M_n,M_n]^n} \mathbb{E}[W_1(\pi, \hat\pi)] \gtrsim \frac{1}{\sqrt{\log n}}. \tag{8.56}$$

Proof. Define two high probability concentration events:

$$E = \left\{ \left| \frac{1}{n} \sum_{i=1}^n f(U_i) - \mathbb{E}_\mu f \right| \le \epsilon \right\}, \quad E' = \left\{ \left| \frac{1}{n} \sum_{i=1}^n f(U'_i) - \mathbb{E}_\nu f \right| \le \epsilon \right\},$$

and by the Chebyshev inequality,

$$\mathbb{P}[E^c], \mathbb{P}[E'^c] \le \frac{\mathsf{var}[f(X)]}{n\epsilon^2} \le \frac{M^2}{n\epsilon^2},$$

since $\mathrm{Lip}(f) \le 1$ on $[-M, M]$. Finally, we construct two priors by $\pi = P_{U|E}$ and $\pi' = P_{U'|E'}$, respectively. By the definition of E, E' and the separation of mean values in (8.53), the functional values in two hypotheses are separated by

$$T(U) - T(U') \ge \frac{M\pi}{L}(1 - o(1)) - 2\epsilon. \tag{8.57}$$

By triangle inequality, the total variation distance between two hypotheses is

$$\begin{aligned} \mathsf{TV}(P_{X|E}, P_{X'|E}) &\le \mathbb{P}[E^c] + \mathbb{P}[E'^c] + \mathsf{TV}(P_X, P_{X'}) \\ &\le \mathbb{P}[E^c] + \mathbb{P}[E'^c] + \sqrt{\chi^2(P_X \| P_{X'})}. \end{aligned} \tag{8.58}$$

Applying the upper bound of the χ^2 distance in (8.54) yields that

$$\mathsf{TV}(P_{X|E}, P_{X'|E}) \le \frac{2M^2}{n\epsilon^2} + \sqrt{\exp\left(ne^{M^2/2}\sum_{k>L}\frac{M^{2k}}{k!}\right) - 1}. \tag{8.59}$$

If M is a fixed constant, we can pick $L = \frac{\log n}{\log\log n}(1 + o(1))$ and $\epsilon = n^{-1/4}$ to obtain (8.55); if $M = M_n = \sqrt{\log n}$, we can pick $L \asymp \log n$ and $\epsilon = n^{-1/4}$ to obtain (8.56). □

Fano method. Let $\Theta = [-M, M]^n$. For $\theta = (\theta_1, \dots, \theta_n) \in \Theta$ denote the histogram by $\pi_\theta = \frac{1}{n}\sum \delta_{\theta_i}$ and the law of the observation by $P_\theta = N(\theta, I_n)$. For any $\theta, \theta' \in \Theta$, the Kullback-Leibler divergence between observations is

$$D(P_\theta \| P_{\theta'}) = \frac{\|\theta - \theta'\|_2^2}{2}.$$

The W_p distance for $p \ge 1$ between histograms is (see, e.g., [161, 2.2.2])

$$W_p(\pi_\theta, \pi_{\theta'}) = \frac{\|\tilde\theta - \tilde\theta'\|_p}{n^{1/p}},$$

where $\tilde\theta$ and $\tilde\theta'$ are an ordered sequence of θ and θ', respectively.

The goal is to find a c-packing $\mathcal{M}$ of $\{\pi_\theta : \theta \in \Theta\}$ in W_1 distance such that

$$\log|\mathcal{M}| \gtrsim \sup_{\theta,\theta' \in \mathcal{M}} \|\theta - \theta'\|_2^2.$$

If $M = M_n \asymp n$, we can construct an explicit packing: Let the grid be

$G = (g_1, \ldots, g_n)$ be the equipartition of the interval $[-M_n, M_n]$. Consider the θ of form $\theta = G + \alpha\epsilon$ where $\alpha \in \{0,1\}^n$ and $\epsilon \gtrsim 1$. Then

$$\|\theta - \theta'\|_2^2 = \epsilon^2 d_H(\alpha, \alpha'),$$

where d_H denotes the Hamming distance. When ϵ is a small constant that $\epsilon \leq 2M_n/n$, the W_1 distance is simply

$$W_1(P_\theta, P_{\theta'}) = \frac{1}{n}\|\theta - \theta'\|_1 = \frac{\epsilon d_H(\alpha, \alpha')}{n}.$$

By Gilbert-Varshamov bound, the maximal cn-packing of $\{0,1\}^n$ in Hamming distance has size at least

$$|\mathcal{M}| \geq \frac{2^n}{\sum_{j=0}^{cn-1}\binom{n}{j}}.$$

Hence, $\log|\mathcal{M}| \gtrsim n$. By letting ϵ be a small constant and applying Fano method, we conclude that, when $M = M_n \asymp n$,

$$\inf_{\hat{\pi}} \sup_{\theta \in [-M_n, M_n]^n} \mathbb{E}(W_1(\pi, \hat{\pi}))^2 \gtrsim 1.$$

8.5.6 ℓ_2-norm of the coefficients of bounded polynomials

For any polynomial $p(x) = \sum_{i=0}^{L} a_i x^i$, denote the coefficients by $a = (a_0, \ldots, a_L)$, then the sum of squares of its coefficients is given by the following compact formula:

$$\sum_{i=0}^{L} |a_i|^2 = \frac{1}{2\pi}\oint_{|z|=1} |p(z)|^2 \mathrm{d}z.$$

Then, combining the triangle inequality, we have

$$\|a\|_2 \leq \sup_{|z|=1} |p(z)| \leq \|a\|_1. \tag{8.60}$$

Lemma 8.10. *If the polynomial p of degree L satisfies $|p(x)| \leq 1$ on $[-1,1]$, then $|p(z)| \leq (1+\sqrt{2})^L$ on $|z| = 1$.*

Proof. Let $f(y) \triangleq p(\frac{y+y^{-1}}{2})/y^L$ which is analytic and bounded on $|y| \geq 1$. For $y = e^{i\theta}$, $|f(y)| = |p(\cos\theta)| \leq 1$. By the maximum modulus principle,

$|f(y)| \leq 1$ for any $|y| > 1$. Consider $|z| = 1$ and let $\frac{y+y^{-1}}{2} = 2$ for some $|y| \geq 1$. Then $y = z \pm \sqrt{z^2 - 1}$ and by triangle inequality $|y| \leq 1 + \sqrt{2}$. Since $|f(y)| \leq 1$, then $|p(z)| \leq |y|^L \leq (1 + \sqrt{2})^L$. □

Corollary 8.2. *Let $P_L(x) = \sum_{i=0}^{L} a_i x^i$ and suppose $|P_L(x)| \leq M$ on $[-K, K]$. Denote the vector $b = (a_0 K^0, a_1 K^1, \ldots, a_L K^L)$.*

$$\|b\|_2 \leq M(1 + \sqrt{2})^L. \tag{8.61}$$

Remark 8.4. Consider the Chebyshev polynomial $T_L(z) = \frac{1}{2}(y^L + y^{-L})$, where $z = \frac{y+y^{-1}}{2}$. $T_L(x)$ is bounded by one on $[-1, 1]$.

$$|T_L(i)| = \frac{|(\sqrt{2}+1)^L + (-1)^L(\sqrt{2}-1)^L|}{2} \geq \frac{(\sqrt{2}+1)^L - (\sqrt{2}-1)^L}{2}.$$

By (8.60), the upper bound in (8.61) has a tight exponent. This is also observed by the explicit formula for the Chebyshev polynomial:

$$T_L(x) = \frac{L}{2} \sum_{j=0}^{\lfloor L/2 \rfloor} \frac{(-1)^j}{L-j} \binom{L-j}{j} (2x)^{L-2j}.$$

The coefficients at $j = \alpha L$ with $\alpha = \frac{2-\sqrt{2}}{4}$ is

$$\begin{aligned}
&\frac{1}{2(1-\alpha)} \binom{(1-\alpha)L}{\alpha L} 2^{(1-2\alpha)L} \\
&\geq \frac{1}{2(1-\alpha)} \frac{1}{2\sqrt{2\alpha L(1-\frac{\alpha}{1-\alpha})}} \exp\left(L\left((1-\alpha)h\left(\frac{\alpha}{1-\alpha}\right) + (1-2\alpha)\log 2\right)\right) \\
&\asymp \frac{(1+\sqrt{2})^L}{\sqrt{L}},
\end{aligned}$$

where $h(x) \triangleq -x \log x - (1-x)\log(1-x)$ and we used the bound on the binomial coefficient in [140, Lemma 4.7.1].

8.6 Proofs

8.6.1 Proofs of density estimation

Proof of Theorem 8.3. By scaling it suffices to consider $M = 1$. Similar to (8.15) and (8.16), we obtain an estimated mixing distribution $\hat{\nu}$ supported on k atoms in $[-1, 1]$ such that, with probability $1 - \delta$,

$$\|\mathbf{m}_{2k-1}(\hat{\nu}) - \mathbf{m}_{2k-1}(\nu)\|_2 \le \sqrt{c_k \log(1/\delta)/n},$$

for some constant c_k that depends on k. The conclusion follows from Lemmas 7.5 and 7.4. □

Proof of Theorem 8.4. Recall that f is 1-subgaussian and σ is a fixed constant. Similar to (8.16), we obtain an estimate $\tilde{m}_r$ for $\mathbb{E}_f[\gamma_r(X, \sigma)]$ (see the definition of $\gamma_r(\cdot, \sigma)$ in (8.13)) for $r = 1, \ldots, 2k-1$ such that, with probability $1 - \delta$,

$$|\tilde{m}_r - \mathbb{E}_f[\gamma_r(X, \sigma)]| \le \sqrt{c_k \log(1/\delta)/n},$$

for some constant c_k that depends on k. By assumption, $\mathsf{TV}(f, g) \le \epsilon$ where both f and g are 1-subgaussian. Let $g = \nu * N(0, \sigma^2)$. Then, using Lemma 8.11 and the triangle inequality, we have

$$|\tilde{m}_r - m_r(\nu)| \le O_k\left(\epsilon\sqrt{\log(1/\epsilon)} + \sqrt{\log(1/\delta)/n}\right), \quad r = 1, \ldots, 2k-1.$$

Using the projection (8.12), we obtain $\hat{\nu}$ similar to (8.15) such that

$$\|\mathbf{m}_{2k-1}(\hat{\nu}) - \mathbf{m}_{2k-1}(\nu)\|_2 \le O_k\left(\epsilon\sqrt{\log(1/\epsilon)} + \sqrt{\log(1/\delta)/n}\right).$$

Let $\hat{f} = \hat{\nu} * N(0, \sigma^2)$. Using the moment comparison in Lemmas 7.5 and 7.4, and applying the upper bound $\mathsf{TV}(\hat{f}, g) \le \sqrt{\chi^2(\hat{f}\|g)/2}$, we obtain that

$$\mathsf{TV}(\hat{f}, g) \le O_k\left(\epsilon\sqrt{\log(1/\epsilon)} + \sqrt{\log(1/\delta)/n}\right).$$

The conclusion follows from the triangle inequality. □

Lemma 8.11. *Let σ be a constant. If f and g are 1-subgaussian, and*

$\mathsf{TV}(f,g) \le \epsilon$, *then,*

$$|\mathbb{E}_f[\gamma_r(X,\sigma)] - \mathbb{E}_g[\gamma_r(X,\sigma)]| \le O_r(\epsilon\sqrt{\log(1/\epsilon)}).$$

Proof. The total variation distance has the following variational representation:

$$\mathsf{TV}(f,g) = \frac{1}{2}\sup_{\|h\|_\infty \le 1} |\mathbb{E}_f h - \mathbb{E}_g h|. \tag{8.62}$$

Here the function $\gamma_r(\cdot,\sigma)$ is a polynomial and unbounded, so the above representation cannot be directly applied. Instead, we apply a truncation argument, thanks to the subgaussianity of f and g, and obtain that, for both $X \sim f$ and g (see Lemmas 8.27 and 8.29),

$$\mathbb{E}[\gamma_r(X,\sigma)\mathbf{1}_{\{|X|\ge\alpha\}}] \le (O(\sqrt{r}))^r \mathbb{E}|X^r \mathbf{1}_{\{|X|\ge\alpha\}}| \le (O(\alpha\sqrt{r}))^r e^{-\alpha^2/2}.$$

Note that by definition (8.13), $\gamma_r(x,\sigma)$ on $|x| \le \alpha$ is at most $(O(\alpha\sqrt{r}))^r$. Applying (8.62) yields that, for $h(x) = \gamma_r(x,\sigma)\mathbf{1}_{\{|x|\le\alpha\}}$,

$$|\mathbb{E}_f h - \mathbb{E}_g h| \le \epsilon(O(\alpha\sqrt{r}))^r.$$

The conclusion follows by choosing $\alpha = O_r(\sqrt{\log(1/\epsilon)})$ and using the triangle inequality. □

8.6.2 Proofs for Section 8.2.1

Proof of Lemma 8.1. Note that $\tilde{m}_r = \frac{1}{n}\sum_{i=1}^n \gamma_r(X_i,\sigma)$. Then we have

$$\mathsf{var}[\tilde{m}_r] = \frac{1}{n}\mathsf{var}[\gamma_r(X,\sigma)],$$

where $X \sim \nu * N(0,\sigma^2)$. Since the standard deviation of a summation is at most the sum of individual standard deviations, using (8.13), we have

$$\sqrt{\mathsf{var}[\gamma_r(X,\sigma)]} \le r! \sum_{j=0}^{\lfloor r/2\rfloor} \frac{(1/2)^j}{j!(r-2j)!}\sigma^{2j}\sqrt{\mathsf{var}[X^{r-2j}]}.$$

X can be viewed as $U + \sigma Z$ where $U \sim \nu$ and $Z \sim N(0,1)$ independent of U. Since ν is supported on $[-M, M]$, for any $\ell \in \mathbb{N}$, we have

$$\mathsf{var}[X^\ell] \le \mathbb{E}[X^{2\ell}] \le 2^{2\ell-1}(M^{2\ell} + \mathbb{E}|\sigma Z|^{2\ell}) \le ((2M)^\ell + \mathbb{E}|3\sigma Z|^\ell)^2,$$

where in the last step we used the inequality $\mathbb{E}|Z|^{2\ell} \le 2^\ell (\mathbb{E}|Z|^\ell)^2$ (see Lemma 8.12). Therefore,

$$\begin{aligned}\sqrt{\mathsf{var}[\gamma_r(X,\sigma)]} &\le r! \sum_{j=0}^{\lfloor r/2 \rfloor} \frac{(1/2)^j}{j!(r-2j)!} \sigma^{2j}((2M)^{r-2j} + \mathbb{E}|3\sigma Z|^{r-2j}) \\ &= \mathbb{E}(2M + \sigma Z')^r + \mathbb{E}(3\sigma|Z| + \sigma Z')^r,\end{aligned}$$

where $Z' \sim N(0,1)$ independent of Z. The conclusion follows by the moments of the standard normal distribution (see [54]). □

Lemma 8.12. *Let $Z \sim N(0,1)$. For $\ell \in \mathbb{N}$, we have*

$$\sqrt{\frac{\pi}{8}} \le \frac{\mathbb{E}|Z|^{2\ell}}{2^\ell (\mathbb{E}|Z|^\ell)^2} \le \sqrt{\frac{2}{\pi}}.$$

Proof. Direct calculations lead to (see [212, 3.461.2–3]):

$$\frac{\mathbb{E}|Z|^{2\ell}}{2^\ell (\mathbb{E}|Z|^\ell)^2} = \begin{cases} \frac{\binom{2\ell}{\ell}}{\binom{\ell}{\ell/2} 2^\ell}, & \ell \text{ even}, \\ \frac{\pi \ell}{8^\ell} \binom{2\ell}{\ell} \binom{\ell-1}{\frac{\ell-1}{2}}, & \ell \text{ odd}. \end{cases}$$

Using $\frac{2^n}{\sqrt{2n}} \le \binom{n}{n/2} \le 2^n \sqrt{\frac{2}{\pi n}}$ [140, Lemma 4.7.1], we obtain that

$$\begin{aligned}\sqrt{\frac{\pi}{8}} \le \frac{\binom{2\ell}{\ell}}{\binom{\ell}{\ell/2} 2^\ell} \le \sqrt{\frac{2}{\pi}}, \\ \frac{\pi}{4} \sqrt{\frac{\ell}{2(\ell-1)}} \le \frac{\pi \ell}{8^\ell} \binom{2\ell}{\ell} \binom{\ell-1}{\frac{\ell-1}{2}} \le \sqrt{\frac{\ell}{2(\ell-1)}},\end{aligned}$$

which prove this lemma for $\ell \ge 5$. For $\ell \le 4$ the lemma follows from the above equalities. □

8.6.3 Proofs for Section 8.2.2

Proof of Proposition 8.1. By scaling it suffices to consider $M = 1$. Without loss of generality assume $\sigma \geq \hat{\sigma}$ and otherwise we can interchange π and $\hat{\pi}$. Let $\tau^2 = \sigma^2 - \hat{\sigma}^2$ and $\nu' = \nu * N(0, \tau^2)$. Similar to (8.20), we obtain that

$$|m_r(\nu') - m_r(\hat{\nu})| \leq (c\sqrt{k})^{2k}\epsilon, \quad r = 1, \ldots, 2k, \tag{8.63}$$

for some absolute constant c. Using Lemma 8.13 yields that $\tau \leq O(\epsilon^{\frac{1}{2k}})$. It follows from Proposition 7.2 that

$$W_1(\nu', \hat{\nu}) \leq O\left(k^{1.5}\epsilon^{\frac{1}{2k}}\right).$$

The conclusion follows from $W_1(\nu', \nu) \leq O(\tau)$ and the triangle inequality. □

Lemma 8.13. *Suppose $\pi = \nu * N(0, \tau^2)$ and π' is k-atomic supported on $[-1, 1]$. Let $\epsilon = \max_{i \in [2k]} |m_i(\pi) - m_i(\pi')|$. Then,*

$$\tau \leq 2\left(\epsilon/k!\right)^{\frac{1}{2k}}.$$

Proof. Denote the support of π' by $\{x'_1, \ldots, x'_k\}$. Consider the polynomial $P(x) = \prod_{i=1}^{k}(x - x'_i)^2 = \sum_{i=0}^{2k} a_i x^i$ that is almost surely zero under π'. Since every $|x'_i| \leq 1$, similar to (7.7), we obtain that

$$\mathbb{E}_{\pi}[P] = |\mathbb{E}_{\pi}[P] - \mathbb{E}_{\pi'}[P]| \leq 2^{2k}\epsilon.$$

Since $\pi = \nu * N(0, \tau^2)$, we have

$$\mathbb{E}_{\pi}[P] \geq \min_x \mathbb{E}[P(x + \tau Z)] \geq \tau^{2k} \min_{y_1, \ldots, y_k} \mathbb{E}\left[\prod_i (Z + y_i)^2\right] = k!\tau^{2k},$$

where $Z \sim N(0, 1)$, and in the last step we used Lemma 8.14. □

Lemma 8.14. *Let $Z \sim N(0, 1)$. Then,*

$$\min\{\mathbb{E}[p^2(Z)] : \deg(p) \leq k,\ p \text{ is monic}\} = k!$$

achieved by $p = H_k$.

Proof. Since p is monic, it can be written as $p = H_k + \sum_{j=0}^{k-1} \alpha_j H_j$, where

H_j is the Hermite polynomial (2.21). By the orthogonality (2.20), we have $\mathbb{E}[p^2(Z)] = k! + \sum_{j=0}^{k-1} \alpha_j^2 j!$ and the conclusion follows. □

Proof of Lemma 8.2. The proof is similar to [49, Theorem 5B]. Let $\hat{\mathbf{M}}_r(\sigma)$ denote the moment matrix associated with the empirical moments of $\gamma_i(X, \sigma)$ for $i \leq 2r$, and let

$$\hat{\sigma}_r = \inf\{\sigma > 0 : \det(\hat{\mathbf{M}}_r(\sigma)) = 0\}. \tag{8.64}$$

The smallest positive zero of $\hat{d}_k$ is given by $\hat{\sigma}_k$. Direct calculation shows that $\hat{\sigma}_1 = s$. Since the mixture distribution has a density, then almost surely, the empirical distribution has n points of support. By Theorem 2.12, the matrix $\hat{\mathbf{M}}_r(0)$ is positive definite and thus $\hat{\sigma}_r > 0$ for any $r < n$. For any $q < r$, if $\hat{\mathbf{M}}_r(\sigma)$ is positive definite, then $\hat{\mathbf{M}}_q(\sigma)$ as a leading principal submatrix is also positive definite. Since eigenvalues of $\hat{\mathbf{M}}_r(\sigma)$ are continuous functions of σ, we have $\hat{\sigma}_r > \sigma \Rightarrow \hat{\sigma}_q > \sigma$, and thus

$$\hat{\sigma}_q \geq \hat{\sigma}_r, \quad \forall\ q < r. \tag{8.65}$$

In particular, $\hat{\sigma}_k \leq \hat{\sigma}_1$. □

Proof of Lemma 8.3. We continue to use the notation in (8.64). Applying (8.65) and Lemma 8.2 yields that

$$0 < \hat{\sigma} = \hat{\sigma}_k \leq \hat{\sigma}_{k-1} \leq ... \leq \hat{\sigma}_1 = s,$$

and for any $\sigma < \hat{\sigma}_j$, the matrix $\hat{\mathbf{M}}_j(\sigma)$ is positive definite. Since $\det(\hat{\mathbf{M}}_k(\hat{\sigma})) = 0$, then, for some $r \in \{1, \ldots, k\}$, we have $\det(\hat{\mathbf{M}}_j(\hat{\sigma})) = 0$ for $j = r, \ldots, k$, and $\det(\hat{\mathbf{M}}_j(\hat{\sigma})) > 0$ for $j = 0, \ldots, r-1$. By Theorem 2.12, there exists an r-atomic distribution whose j^{th} moment coincides with $\hat{\gamma}_j(\hat{\sigma})$ for $j \leq 2r$. It suffices to show that $r = k$ almost surely.

Since the mixture distribution has a density, in the following we condition on the event that all samples $X_1, \ldots, X_n$ are distinct, which happens almost surely, without loss of generality. We first show that the empirical moments $(\hat{\gamma}_1, \ldots, \hat{\gamma}_n)$, where $\hat{\gamma}_j = \frac{1}{n}\sum_i X_i^j$, have a joint density in $\mathbb{R}^n$. The Jabobian

matrix of this transformation is

$$\frac{1}{n}\begin{bmatrix} 1 & & & \\ & 2 & & \\ & & \ddots & \\ & & & n \end{bmatrix}\begin{bmatrix} 1 & \cdots & 1 \\ X_1 & \cdots & X_n \\ \vdots & \ddots & \vdots \\ X_1^{n-1} & \cdots & X_n^{n-1} \end{bmatrix},$$

which is invertible. Since those n samples $(X_1, \ldots, X_n)$ have a joint density, then the empirical moments $(\hat{\gamma}_1, \ldots, \hat{\gamma}_n)$ also have a joint density.

Suppose, for the sake of contradiction, that $r \leq k-1$. Then $\det(\hat{\mathbf{M}}_{r-1}(\hat{\sigma})) > 0$ and $\det(\hat{\mathbf{M}}_r(\hat{\sigma})) = \det(\hat{\mathbf{M}}_{r+1}(\hat{\sigma})) = 0$. In this case, $\hat{m}_{2r+1}(\hat{\sigma})$ is a deterministic function of $\hat{m}_1(\hat{\sigma}), \ldots, \hat{m}_{2r}(\hat{\sigma})$ (see Lemma 8.26). Since $\hat{\sigma}$ is the smallest positive root of $\hat{d}_r(\sigma) = 0$, it is uniquely determined by $(\hat{\gamma}_1, \ldots, \hat{\gamma}_{2r})$. Therefore, $\hat{m}_{2r+1}(\hat{\sigma})$, and thus $\hat{\gamma}_{2r+1}$, are both deterministic functions of $(\hat{\gamma}_1, \ldots, \hat{\gamma}_{2r})$, which happens with probability zero, since the sequence $(\hat{\gamma}_1, \ldots, \hat{\gamma}_{2r+1})$ has a joint density. Consequently, $r \leq k-1$ with probability zero. □

The proof of (8.21) relies on the following result, which obtains a tail probability bound by comparing moments.

Lemma 8.15. *Let $\epsilon = \max_{i\in[2k]} |m_i(\nu) - m_i(\nu')|$. If either ν or ν' is k-atomic, and ν is supported on $[-1, 1]$, then, for any $t > 1$,*

$$\mathbb{P}[|Y| \geq t] \leq 2^{2k+1}\epsilon/(t-1)^{2k}, \quad Y \sim \nu'.$$

Proof. We only show the upper tail bound $\mathbb{P}[Y \geq t]$. The lower tail bound of Y is equal to the upper tail bound of $-Y$.

- Suppose ν is k-atomic supported on $\{x_1, \ldots, x_k\}$. Consider a polynomial $P(x) = \prod_i (x - x_i)^2$ of degree $2k$ that is almost surely zero under ν. Since every $|x_i| \leq 1$, similar to (7.7), we obtain that

$$\mathbb{E}_{\nu'}[P] = |\mathbb{E}_{\nu}[P] - \mathbb{E}_{\nu'}[P]| \leq 2^{2k}\epsilon.$$

  Using Markov inequality, for any $t > 1$, we have

$$\mathbb{P}[Y \geq t] \leq \mathbb{P}[P(Y) \geq P(t)] \leq \frac{\mathbb{E}[P(Y)]}{P(t)} \leq \frac{2^{2k}\epsilon}{(t-1)^{2k}}.$$

- Suppose ν' is k-atomic supported on $\{x_1, \ldots, x_k\}$. If those values are all within $[-1, 1]$, then we are done. If there are at most $k-1$ values, denoted by $\{x_1, \ldots, x_{k-1}\}$, are within $[-1, 1]$, then we consider a polynomial $P(x) = (x^2 - 1)\prod_i (x - x_i)^2$ of degree $2k$ that is almost surely non-positive under ν. Similar to (7.7), we obtain that

$$\mathbb{E}_{\nu'}[P] \leq \mathbb{E}_{\nu'}[P] - \mathbb{E}_{\nu}[P] \leq 2^{2k}\epsilon.$$

  Since $P \geq 0$ almost surely under ν', the conclusion follows follows analogously using Markov inequality. □

Lemma 8.16. *Let*

$$\pi = \nu * N(0, \tau^2), \quad \hat{\pi} = \hat{\nu},$$

where ν and ν are both k-atomic, ν is supported on $[-1, 1]$, and $\tau \leq 1$. If $|m_i(\pi) - m_i(\hat{\pi})| \leq \epsilon$ for $i \leq 2k$, then, for any $t \geq \sqrt{18k}$,

$$\mathbb{P}[|\hat{U}| \geq t] \leq \frac{2^{2k+1}\epsilon}{(\frac{t}{\sqrt{18k}} - 1)^{2k}}, \quad \hat{U} \sim \hat{\nu}.$$

Proof. Let g be the $(k+1)$-point Gauss quadrature of the standard normal distribution. Furthermore, g is supported on $[-\sqrt{4k+6}, \sqrt{4k+6}]$ for some absolute constant c (see the bound on the zeros of Hermite polynomials in [53, p. 129]). Let $G \sim g$, $U \sim \nu$, and $\hat{U} \sim \hat{\nu}$. Denote the maximum absolute value of $U + \tau G$ by M which is at most $1 + \sqrt{4k+6} \leq \sqrt{18k}$ for $k \geq 1$. Applying Lemma 8.15 with the distributions of $\frac{U+\tau G}{\sqrt{18k}}$ and $\frac{\hat{U}}{\sqrt{18k}}$ yields the conclusion. □

8.6.4 Proofs for Section 8.2.3

Proof of Theorem 8.5. Note that U and $\hat{U}$ are both supported on a set of $2k$ atoms, and the largest cluster of U is of size at most $k - k_0 + 1$. Since different clusters of U are separated by γ, then each atom of either U and $\hat{U}$ is at least $\gamma/2$ away from all but $2k - k_0$ atoms. From the proof of (8.7), we have $|m_r(\hat{U}) - m_r(U)| < (O(\sqrt{k}))^{2k}\sqrt{\frac{\log(k/\delta)}{n}}$. The conclusion follows from Proposition 7.3. □

Proof of Proposition 8.2. The proof is similar to Proposition 8.1, except that

moment comparison theorem Proposition 7.2 is replaced by its adaptive version Proposition 7.4. Recall (8.63):

$$|m_r(\nu') - m_r(\hat{\nu})| \le (c\sqrt{k})^{2k}\epsilon, \quad r = 1, \dots, 2k,$$

where $\nu' = \nu * N(0, \tau^2)$ and $\tau^2 = |\sigma^2 - \hat{\sigma}^2|$. Since $\hat{\nu} * N(0,1)$ has k_0 γ-separated clusters, any $t \in \mathbb{R}$ can be $\gamma/2$ close to at most $k - k_0 + 1$ atoms of $\hat{\nu}$. Applying Proposition 7.4 yields that

$$W_1(\nu', \hat{\nu}) \le 8k \left(\frac{k(4c\sqrt{k})^{2k}\epsilon}{(\gamma/2)^{2(k_0-1)}} \right)^{\frac{1}{2(k-k_0+1)}}.$$

Using Lemma 8.17 yields that $\tau \le O_k(W_1(\nu', \hat{\nu}))$. The conclusion follows from $W_1(\nu', \nu) \le O(\tau)$ and the triangle inequality. □

Lemma 8.17. *Suppose $\pi = \nu * N(0, \tau^2)$ and π' is k-atomic. Then*

$$\tau \le O_k(W_1(\pi, \pi')).$$

Proof. In this proof we write $W_1(X, Y) = W_1(P_X, P_Y)$. Let $Z \sim N(0,1)$, $U \sim \nu$, and $U' \sim \pi'$. For any $x \in \mathbb{R}$, we have

$$W_1(x + \tau Z, U') = \tau W_1(Z, (U' - x)/\tau) \ge c_k \tau,$$

where $c_k = \inf\{W_1(Z, Y) : Y \text{ is } k\text{-atomic}\}$.[9] For any couping between $U + \tau Z$ and U',

$$\mathbb{E}|U + \tau Z - U'| = \mathbb{E}[\mathbb{E}[|U + \tau Z - U'||U]] \ge c_k \tau.$$ □

Proof of Theorem 8.5. By scaling it suffices to consider $M = 1$. Recall that the Gaussian mixture is assumed to have k_0 γ-separated clusters in the sense of Definition 8.1, that is, there exists a partition $S_1, \dots, S_{k_0}$ of $[k]$ such that $|\mu_i - \mu_{i'}| \ge \gamma$ for any $i \in S_\ell$ and $i' \in S_{\ell'}$ such that $\ell \ne \ell'$. Denote the union of the support sets of ν and $\hat{\nu}$ by $\mathcal{S}$. Each atom is $\mathcal{S}$ is at least $\gamma/2$ away from at least $k_0 - 1$ other atoms. Then (8.22) follows from Proposition 7.3 with $\ell = 2k$ and $\ell' = (2k-1) - (k_0 - 1)$. □

[9] We can prove that $c_k \ge \Omega(1/k)$ using the dual formula (6.3).

8.6.5 Proofs for Section 8.2.4

Lemma 8.18. *Assume in the Gaussian mixture (8.1) $w_i \geq \epsilon$, $\sigma = 1$. Suppose $L = \sqrt{c \log n}$ in Algorithm 8.3. Then, with probability at least $1 - ke^{-n'\epsilon} - n^{-(\frac{c}{8}-1)}$, the following holds:*

- *$\ell_j \leq 3kL$ for every j.*
- *Let $X_i = U_i + Z_i$ for $i \in [n]$, where $U_i \sim \nu$ is the latent variable and $Z_i \sim N(0,1)$. Then, $|Z_i| \leq 0.5L$ for every $i \in [n]$; $X_i \in I_j$ if and only if $U_i \in I_j$.*

Proof. By the union bound, with probability $1 - ke^{-n'\epsilon} - n^{-(\frac{c}{8}-1)}$, the following holds:

- $|Z_i| \leq 0.5L$ for every $i \in [n]$.
- For every $j \in [k]$, there exists $i \leq n'$ such that $U_i = \mu_j$.

Recall the disjoint intervals $I_1 \cup \ldots \cup I_s = \cup_{i=1}^{n'}[X_i \pm L]$. Then, we obtain that

$$\bigcup_{j=1}^{k}[\mu_j \pm 0.5L] \;\subseteq\; I_1 \cup \cdots \cup I_s \;\subseteq\; \bigcup_{j=1}^{k}[\mu_j \pm 1.5L].$$

The total length of all intervals is at most $3kL$. Since $|Z_i| \leq 0.5L$, $X_i = U_i + Z_i$ is in the same interval as U_i. □

Proof of Theorem 8.6. Since $n' \geq \Omega(\frac{\log(k/\delta)}{\epsilon})$, applying Lemma 8.18 yields that, with probability at least $1 - \frac{\delta}{3} - n^{-\Omega(1)}$, the following holds:

- $\ell_j \leq O(kL)$ for every j.
- Let $X_i = U_i + \sigma Z_i$ for $i \in [n]$ as in Lemma 8.18. Then, $|Z_i| \leq 0.5L$ for every $i \in [n]$; $X_i \in I_j$ if and only if $U_i \in I_j$.

The intervals $I_1, \ldots, I_s$ are independent of every C_j and are treated as deterministic in the remaining proof. We first evaluate the expected moments of samples in C_j, conditioned on $|Z_i| \leq L' \triangleq 0.5L$. Let $X = U + \sigma Z$ where $U \sim \nu$ and $Z \sim N(0,1)$. Then,

$$\begin{aligned}\mathbb{E}[(X - c_j)^r | X \in I_j, |Z| \leq L'] &= \mathbb{E}[(X - c_j)^r | U \in I_j, |Z| \leq L'] \\ &= \mathbb{E}[(U_j' + \sigma Z)^r | |Z| \leq L'],\end{aligned}$$

where $U_j' = U_j - c_j$, and $U_j \sim P_{U|U\in I_j}$. Since $|U_j'| \le O(kL)$ and $L' = \Theta(\sqrt{\log n})$, the right-hand side differs from the unconditional moment by (see Lemma 8.30)

$$|\mathbb{E}[(U_j'+\sigma Z)^r||Z| \le L'] - \mathbb{E}[(U_j'+\sigma Z)^r]| \le (kL\sigma\sqrt{r})^r n^{-\Omega(1)}, \quad r = 1, \dots, 2k-1,$$

which is less than n^{-1} when $k \le O(\frac{\log n}{\log\log n})$. Therefore, the accuracy of empirical moments in (8.14), (8.19) and thus Theorem 8.1 are all applicable. Since $w_i \ge \epsilon$, with probability at least $1 - \frac{\delta}{3}$, each C_j contains $\Omega(n\epsilon)$ samples, and applying Theorem 8.1 yields that, with probability $1 - \frac{\delta}{3}$,

$$W_1(\hat{\nu}_j, \nu_j) \le \begin{cases} O(Lk^{2.5}(\frac{n\epsilon}{\log(3k/\delta)})^{-\frac{1}{4k-2}}), & \sigma \text{ known}, \\ O(Lk^3(\frac{n\epsilon}{\log(3k/\delta)})^{-\frac{1}{4k}}), & \sigma \text{ unknown}, \end{cases}$$

for every j, where ν_j denotes the distribution of U_j' and $\hat{\nu}_j$ is the estimate in Theorem 8.1. Using the weights threshold $\tau = \epsilon/(2k)$, and applying Lemma 8.19, we obtain that

$$d_H(\mathrm{supp}(\hat{\nu}_j), \mathrm{supp}(\nu_j)) \le \frac{W_1(\hat{\nu}_j, \nu_j)}{\epsilon/(2k)}.$$

The conclusion follows. □

Lemma 8.19. *Let ν be a discrete distribution whose atom has at least ϵ probability. Let S_ν and $S_{\hat{\nu}}$ denote the support sets of ν and $\hat{\nu}$, respectively. For $\hat{S} \subseteq S_{\hat{\nu}}$,*

$$d_H(S_\nu, \hat{S}) \le \frac{W_1(\nu, \hat{\nu})}{(\min_{y\in\hat{S}} \hat{\nu}(y)) \wedge (\epsilon - \hat{\nu}(\hat{S}^c))_+}.$$

Proof. This is a generalization of Lemma 6.2 in the sense that the minimum weight of $\hat{\nu}$ is unknown. For any coupling P_{XY} such that $X \sim \nu$ and $Y \sim \hat{\nu}$, for any $y \in \hat{S}$,

$$\mathbb{E}|X - Y| \ge \hat{\nu}(y)\mathbb{E}[|X - Y||Y = y] \ge \epsilon_1 \min_{x\in S_\nu} |x - y|,$$

where $\epsilon_1 = \min_{y\in\hat{S}} \hat{\nu}(y)$. Note that $\mathbb{P}[Y \notin \hat{S}, X = x] \le \hat{\nu}(\hat{S}^c)$ and $\nu(x) \ge \epsilon$ for any $x \in S_\nu$. Then we have $\mathbb{P}[Y \in \hat{S}, X = x] \ge (\epsilon - \hat{\nu}(\hat{S}^c))_+ \triangleq \epsilon_2$, and thus

$$\mathbb{E}|X - Y| \ge \epsilon_2\mathbb{E}[|X - Y||X = x, Y \in \hat{S}] \ge \epsilon_2 \min_{y\in\hat{S}} |x - y|.$$

Using the definition of d_H in (6.5), the proof is complete. □

8.6.6 Proofs for Section 8.3

Proof of Theorem 3.4. Let $U \sim \nu$ and $U' \sim \nu'$. If ν and ν' are ϵ-subgaussian, then $\mathsf{var}[U'] \le \epsilon^2$, and $\mathbb{E}|U|^p, \mathbb{E}|U'|^p \le 2(\epsilon\sqrt{p/e})^p$ [54]. Applying the χ^2 upper bound from moment difference in Lemma 7.5 yields that

$$\chi^2(\nu * N(0,1) \| \nu' * N(0,1)) \le e^{\epsilon^2/2} \sum_{j \ge \ell+1} \frac{16\epsilon^{2j}}{\sqrt{2\pi j}},$$

where we used Stirling's approximation $n! > \sqrt{2\pi n}(n/e)^n$. If ν and ν' are supported on $[-\epsilon, \epsilon]$, the conclusion is obtained similarly by using $\mathbb{E}|U|^p, \mathbb{E}|U'|^p \le \epsilon^p$. □

Proof of Proposition 8.3. Let ν and ν' be the optimal pair of distributions for (8.24). Applying Theorem 3.4 yields that

$$\chi^2(\nu * N(0,1) \| \nu' * N(0,1)) \le c \left(\frac{e\epsilon^2}{2k-1} \right)^{2k-1},$$

for some absolute constant c. The two mixing distributions satisfy (see Lemma 8.20)

$$W_1(\nu, \nu') \ge \Omega(\epsilon/\sqrt{k}).$$

The conclusion follows by choosing $\epsilon = c'\sqrt{k} n^{-\frac{1}{4k-2}}$ for some absolute constant c' and applying Le Cam's method [96]. □

Lemma 8.20.

$$\sup\{W_1(\nu, \nu') : \mathbf{m}_\ell(\nu) = \mathbf{m}_\ell(\nu'),\ \nu, \nu' \text{ on } [-1,1]\} = \Theta(\beta/(\ell+1)).$$

Furthermore, the supremum is $\frac{\beta(\pi - o(1))}{\ell+1}$ *as* $\ell \to \infty$*, and is achieved by two distributions whose support sizes differ by at most one and sum up to* $\ell + 2$*.*

Proof. It suffices to prove for $\beta = 1$. Using the dual characterization of the W_1 distance in Section 6.2, the supremum is equal to

$$\sup_{f:1-\text{Lipschitz}} \sup \left\{ \mathbb{E}_\nu f - \mathbb{E}_{\nu'} f : \mathbf{m}_\ell(\nu) = \mathbf{m}_\ell(\nu'),\ \nu, \nu' \text{ on } [-\beta, \beta] \right\}.$$

Using the duality between moment matching and best polynomial approximation (see [55, Appendix E]), the optimal value is further equal to

$$2 \sup_{f:1-\text{Lipschitz}} \inf_{P:\text{degree } \le \ell} \sup_{|x|\le 1} |f(x) - P(x)|.$$

The above value is the best uniform approximation error over 1-Lipschitz functions, a well-studied quantity in the approximation theory (see, e.g., [210, section 4.1]), and thus the optimal values in the lemma are obtained. A pair of optimal distributions are supported on the maxima and the mimima of $P^* - f^*$, respectively, where f^* is the optimal 1-Lipschitz function and P^* is the best polynomial approximation for f^*. The numbers of maxima and minima differ by at most one by Chebyshev's alternating theorem (see, e.g., [34, p. 54]). □

Proof of Proposition 8.4. Let $\nu = N(0, \epsilon^2)$ and ν' be its k-point Gauss quadrature. Then $\mathbf{m}_{2k-1}(\nu) = \mathbf{m}_{2k-1}(\nu')$ and ν and ν' are both ϵ-subgaussian (see Lemma 2.1). Applying Theorem 3.4 yields that

$$\chi^2(\nu * N(0,1) \| \nu' * N(0,1)) \le O(\epsilon^{4k}).$$

Note that $\nu * N(0,1) = N(0, 1+\epsilon^2)$ is a valid Gaussian mixture distribution (with single zero mean component). Between the above two mixture models, the variance parameters differ by ϵ^2; the mean parameters satisfy $W_1(g_k, \delta_0) \ge \Omega(\epsilon/\sqrt{k})$ (see Lemma 2.2). The conclusion follows by choosing $\epsilon = cn^{-\frac{1}{4k}}$ for some absolute constant c applying applying Le Cam's method [96]. □

8.6.7 Standard form of the semidefinite programming (8.12)

Given an arbitrary vector $\tilde{m} = (\tilde{m}_1, \ldots, \tilde{m}_r)$, we want to compute its projection onto the moment space $\mathcal{M}_r([a,b])$. By introducing an auxiliary scalar variable t satisfying $t \ge \|x\|_2^2$, (8.12) is equivalent to

$$\begin{aligned} \min\ & t - 2\langle \tilde{m}, x\rangle + \|\tilde{m}\|_2^2, \\ \text{s.t. } & t \ge \|x\|_2^2,\ x \text{ satisfies (2.16)}. \end{aligned}$$

This is a semidefinite programming with decision variable (x, t), since the constraint $t \geq \|x\|_2^2$ is equivalent to $\begin{bmatrix} t & x^\top \\ x & I \end{bmatrix} \succeq 0$ using Schur complement (see, e.g., [213]).

8.6.8 Auxiliary lemmas

Lemma 8.21. *If* $|\mathbb{E}[X^\ell] - \mathbb{E}[X'^\ell]| \leq (C\sqrt{\ell})^\ell \epsilon$ *for* $\ell = 1, \ldots, r$, *then, for* γ_r *in (8.13),*

$$|\mathbb{E}[\gamma_r(X, \sigma)] - \mathbb{E}[\gamma_r(X', \sigma)]| \leq \epsilon \left((2\sigma\sqrt{r/e})^r + (2C\sqrt{r})^r \right).$$

Proof. Note that $|\mathbb{E}[X^\ell] - \mathbb{E}[X'^\ell]| \leq \mathbb{E}|C\sqrt{e}Z'|^r \epsilon$ by Lemma 8.22, where $Z' \sim N(0, 1)$. Then,

$$\begin{aligned} |\mathbb{E}[\gamma_r(X, \sigma)] - \mathbb{E}[\gamma_r(X', \sigma)]| &\leq \sum_{i=0}^{\lfloor r/2 \rfloor} \frac{r! \sigma^{2i}}{i!(r-2i)! 2^i} \mathbb{E}[|C\sqrt{e}Z'|^r] \epsilon \\ &= \epsilon \cdot \mathbb{E}[(\sigma Z + |C\sqrt{e}Z'|)^r], \end{aligned}$$

where $Z \sim N(0, 1)$ independent of Z'. Applying $(a + b)^r \leq 2^{r-1}(|a|^r + |b|^r)$ and Lemma 8.22 completes the proof. □

Lemma 8.22.

$$(p/e)^{p/2} \leq \mathbb{E}|Z|^p \leq \sqrt{2}(p/e)^{p/2}, \quad p \geq 0.$$

Proof. Note that

$$\frac{\mathbb{E}|Z|^p}{(p/e)^{p/2}} = \frac{2^{p/2}\Gamma(\frac{p+1}{2})}{\sqrt{\pi}(p/e)^{p/2}} \triangleq f(p), \quad \forall\ p \geq 0.$$

Since $f(0) = 1$ and $f(\infty) = \sqrt{2}$, it suffices to show that f is increasing in $[0, \infty)$. Equivalently, $\frac{x}{2} \log \frac{2e}{x} + \log \Gamma(\frac{x+1}{2})$ is increasing, which is equivalent to $\psi(\frac{x+1}{2}) \geq \log \frac{x}{2}$ by the derivative, where $\psi(x) \triangleq \frac{\mathrm{d}}{\mathrm{d}x} \log \Gamma(x)$. The last inequality holds for any $x > 0$ (see, e.g., [214, (3)]). □

Lemma 8.23. *Let* $r \geq 2$. *Then,*

$$\int \left(\frac{\delta}{\prod_{i=1}^r |t - x_i|} \wedge 1 \right) \mathrm{d}t \leq 4r\delta^{\frac{1}{r}}.$$

Proof. Without loss of generality, let $x_1 \le x_2 \le \cdots \le x_r$. Note that

$$\int \left(\frac{\delta}{\prod_{i=1}^r |t - x_i|} \wedge 1 \right) \mathrm{d}t = \int_{-\infty}^{x_1} + \int_{x_1}^{\frac{x_1+x_2}{2}} + \int_{\frac{x_1+x_2}{2}}^{x_2} + \cdots + \int_{x_r}^{\infty}.$$

There are $2r$ terms in the summation and each term can be upper bounded by

$$\int_{x_i}^{\infty} \left(\frac{\delta}{|t - x_i|^r} \wedge 1 \right) \mathrm{d}t = \int_0^{\infty} \left(\frac{\delta}{t^r} \wedge 1 \right) \mathrm{d}t = \frac{r}{r-1} \delta^{\frac{1}{r}}.$$

The conclusion follows. □

Lemma 8.24. *Given any $2k$ distinct points $x_1 < x_2 < \cdots < x_{2k}$, there exist two distributions ν and ν' supported on $\{x_1, x_3, \ldots, x_{2k-1}\}$ and $\{x_2, x_4, \ldots, x_{2k}\}$, respectively, such that $\mathbf{m}_{2k-2}(\nu) = \mathbf{m}_{2k-2}(\nu')$.*

Proof. Consider the following linear equation

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{2k-2} & x_2^{2k-2} & \cdots & x_{2k}^{2k-2} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_{2k} \end{pmatrix} = 0.$$

This underdetermined system has a non-zero solution. Let w be a solution with $\|w\|_1 = 2$. Since all weights sum up to zero, then positive weights in w sum up to 1 and negative weights sum up to -1. Let one distribution be supported on x_i with weight w_i for $w_i > 0$, and the other one be supported on the remaining x_i's with the corresponding weights $|w_i|$. Then these two distribution match the first $2k-2$ moments.

It remains to show that the weights in any non-zero solution have alternating signs. Note that all weights are non-zero: if one w_i is zero, then the solution must be all zero since the Vandermonde matrix is of full row rank. To verify the signs of the solution, without loss generality, assume that $w_{2k} = -1$ and then

$$\begin{pmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_{2k-1} \\ \vdots & \ddots & \vdots \\ x_1^{2k-2} & \cdots & x_{2k-1}^{2k-2} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_{2k-1} \end{pmatrix} = \begin{pmatrix} 1 \\ x_{2k} \\ \vdots \\ x_{2k}^{2k-2} \end{pmatrix}.$$

The solution has an explicit formula that $w_i = P_i(x_{2k})$ where P_i is an interpolating polynomial of degree $2k-2$ satisfying $P_i(x_j) = 1$ for $j = i$ and $P_i(x_j) = 0$ for all other $j \le 2k-1$. Specifically, we have $w_i = \frac{\prod_{j\neq i, j\le 2k-1}(x_{2k}-x_j)}{\prod_{j\neq i, j\le 2k-1}(x_i-x_j)}$, which satisfies $w_i > 0$ for odd i and $w_i < 0$ for even i. The proof is complete. □

Lemma 8.25 (Non-existence of an unbiased estimator)**.** *Let $X_1, \dots, X_m$ be independent samples distributed as $pN(s,\sigma^2)+(1-p)N(t,\sigma^2) = \nu * N(0,\sigma^2)$, where $\nu = p\delta_s + (1-p)\delta_t$ and p, s, t, σ are the unknown parameters. For any $r \ge 2$, unbiased estimator for the r^{th} moments of ν, namely, $ps^r + (1-p)t^r$, does not exist.*

Proof. We will derive a few necessary conditions for an unbiased estimator, denoted by $g(x_1, \dots, x_m)$, and then arrive at a contradiction. Expand the function under the Hermite basis

$$g(x_1, \dots, x_m) = \sum_{n_1,\dots,n_m \ge 0} \alpha_{n_1,\dots,n_m} \prod_i H_{n_i}(x_i),$$

and denote by $T_n(\mu, \sigma^2)$ the expected value of the Hermite polynomial $\mathbb{E}H_n(X)$ under Gaussian model $X \sim N(\mu, \sigma^2)$. Without loss of generality we may assume that the function g and the coefficients α are symmetric (permutation invariant). Then, the expected value of the function g under $\sigma^2 = 1$ is

$$\mathbb{E}[g(X_1, \dots, X_m)] = \sum_{n_1,\dots,n_m \ge 0} \alpha_{n_1,\dots,n_m} \prod_i (ps^{n_i} + (1-p)t^{n_i}), \tag{8.66}$$

which can be viewed as a polynomial in p, whereas the target is $ps^r + (1-p)t^r$, a linear function in p. Matching polynomial coefficients yields that

$$\sum_{n_1+\cdots+n_m \ge 0} \alpha_{n_1,\dots,n_m} t^{n_1+\cdots+n_m} = t^r, \tag{8.67}$$

$$\sum_{n_1+\cdots+n_m \ge 0} \alpha_{n_1,\dots,n_m} (s^{n_1} - t^{n_1}) t^{n_2+\cdots+n_m} \cdot m = s^r - t^r, \tag{8.68}$$

$$\sum_{n_1+\cdots+n_m \ge 0} \alpha_{n_1,\dots,n_m} \prod_{i=1}^{j} (s^{n_i} - t^{n_i}) t^{n_{j+1}+\cdots+n_m} = 0, \ \forall\, j = 2, \dots, m, \tag{8.69}$$

where we used the symmetry of the coefficients α. The equality (8.69) with $j = m$ yields that $\alpha_{n_1,\dots,n_m} \neq 0$ only if at least one n_i is zero; then (8.69) with $j = m-1$ yields that $\alpha_{n_1,\dots,n_m} \neq 0$ only if at least two n_i are zero; repeating

this for $j = m, m-1, \ldots, 2$, we obtain that $\alpha_{n_1,\ldots,n_m}$ is non-zero only if at most one n_i is non-zero. Then the equality (8.68) implies that $\alpha_{n_1,\ldots,n_m}$ is non-zero only if exactly one $n_i = r$ and the coefficient is necessarily $\frac{1}{m}$. Therefore, it is necessary that the symmetric function is $g(x_1, \ldots, x_m) = \frac{1}{m}\sum_{i=1}^{m} H_r(x_i)$. However, this function is biased when $\sigma^2 \neq 1$. □

Lemma 8.26. *Given a sequence $\gamma_1, \gamma_2, \ldots$, let $\mathbf{H}_j$ denote the Hankel matrix of order $j+1$ using $1, \gamma_1, \ldots, \gamma_{2j}$. Suppose $\det(\mathbf{H}_{r-1}) \neq 0$, and $\det(\mathbf{H}_r) = \det(\mathbf{H}_{r+1}) = 0$. Then,*

$$\gamma_{2r+1} = (\gamma_{r+1}, \ldots, \gamma_{2r})(\mathbf{H}_{r-1})^{-1}(\gamma_r, \ldots, \gamma_{2r-1})^\top.$$

Proof. The matrices $\mathbf{H}_{r-1}$ and $\mathbf{H}_r$ are both of rank r by their determinants. We first show that the rank of $[\mathbf{H}_r, v]$, which is the first $r+1$ rows of $\mathbf{H}_{r+1}$ and is of dimension $(r+1) \times (r+2)$, is also r, where $v \triangleq (\gamma_{r+1}, \ldots, \gamma_{2r+1})^\top$. Suppose the rank is $r+1$. Then v cannot be in the image of $\mathbf{H}_r$. By symmetry of the Hankel matrix, the transpose of $[\mathbf{H}_r, v]$ is the first $r+1$ columns of $\mathbf{H}_{r+1}$. Those $r+1$ columns are linearly independent when its rank is $r+1$. Since $\det(\mathbf{H}_{r+1}) = 0$, then the last column of $\mathbf{H}_{r+1}$ must be in the image of the first $r+1$ columns, which is a contradiction.

Since the first r columns of $\mathbf{H}_{r+1}$ are linearly independent, and the first $r+1$ columns of $\mathbf{H}_{r+1}$ are of rank r, then the $(r+1)^{\text{th}}$ column of $\mathbf{H}_{r+1}$ is in the image of the first r columns, and thus γ_{2r+1} is a linear combination of $\gamma_{r+1}, \ldots, \gamma_{2r}$. Since $\mathbf{H}_{r-1}$ is of full rank, the coefficients can be uniquely determined by $(\mathbf{H}_{r-1})^{-1}(\gamma_r, \ldots, \gamma_{2r-1})^\top$. □

Lemma 8.27. *If $|x| > 1$, then*

$$|H_r(x)| \leq (\sqrt{cr}|x|)^r,$$

for some absolute constant c.

Proof. For $|x| > 1$,

$$\begin{aligned} |H_r(x)| \leq r! \sum_{j=0}^{\lfloor r/2 \rfloor} \frac{(1/2)^j}{j!(r-2j)!}|x|^r = |x|^r |H_n(\mathbf{i})| = |x|^r |\mathbb{E}(\mathbf{i} + \mathbf{i}Z)^r| \\ = |x|^r |\mathbb{E}(1+Z)^r| \leq (\sqrt{cr}|x|)^r, \end{aligned}$$

for some absolute constant c, where $\mathbf{i} = \sqrt{-1}$ and $Z \sim N(0,1)$. □

Lemma 8.28. *Let $Z \sim N(0,1)$.*

$$\mathbb{P}[Z > M] \leq e^{-\frac{M^2}{2}}.$$

Proof. Applying Chernoff bound yields that

$$\mathbb{P}[Z > M] \leq \exp(-\sup_t(tM - t^2/2)) = \exp(-M^2/2). \qquad \square$$

Lemma 8.29. *For r even, and $M \geq 1$,*

$$\mathbb{E}[Z^r \mathbf{1}_{\{|Z|>M\}}] \leq r(O(\sqrt{r}))^r \left(M^{r-1}e^{-\frac{M^2}{2}}\right).$$

Proof. Applying an integral by parts yields that

$$\int_M^\infty x^r e^{-\frac{x^2}{2}} \mathrm{d}x = M^{r-1}e^{-\frac{M^2}{2}} + (r-1)M^{r-3}e^{-\frac{M^2}{2}} + (r-1)(r-3)M^{r-5}e^{-\frac{M^2}{2}} + \cdots + (r-1)!! \int_M^\infty e^{-\frac{x^2}{2}} \mathrm{d}x.$$

Applying Lemma 8.28 and $(r-1)!! \leq (O(\sqrt{r}))^r$, the conclusion follows. $\square$

Lemma 8.30. *For $M \geq 1$,*

$$0 \leq \mathbb{E}[Z^r] - \mathbb{E}[Z^r||Z| \leq M] \leq r(O(\sqrt{r}))^r \left(M^{r-1}e^{-\frac{M^2}{2}}\right).$$

Proof. For r odd, we have $\mathbb{E}[Z^r] - \mathbb{E}[Z^r||Z| \leq M] = 0$. For r even, the left inequality is immediate since $x \mapsto x^r$ is increasing. For the right inequality,

$$\mathbb{E}[Z^r] - \mathbb{E}[Z^r||Z| \leq M] = \mathbb{E}[Z^r] - \frac{\mathbb{E}[Z^r \mathbf{1}_{\{|Z|\leq M\}}]}{\mathbb{P}[|Z| \leq M]} \leq \frac{\mathbb{E}[Z^r] - \mathbb{E}[Z^r \mathbf{1}_{\{|Z|\leq M\}}]}{\mathbb{P}[|Z| \leq M]},$$

and the conclusion follows from Lemma 8.29. $\square$

Lemma 8.31 (Distribution of random projection)**.** *Let X be uniformly distributed over the unit sphere S^{d-1}. For any $a \in S^{d-1}$ and $r > 0$,*

$$\mathbb{P}[|\langle a, X\rangle| < r] < r\sqrt{d}.$$

Proof. Denote the surface area of the d-dimensional unit sphere by $S_{d-1} =$

$\frac{2\pi^{d/2}}{\Gamma(d/2)}$. By symmetry,

$$\begin{aligned}\mathbb{P}[|\langle a, X\rangle| < r] = \mathbb{P}[|X_1| < r] &= \frac{\int_{-r}^{r}(\sqrt{1-x^2})^{d-2} S_{d-2}\sqrt{1-x^2}\mathrm{d}x}{S_{d-1}} \\ &= \frac{2S_{d-2}}{S_{d-1}}\int_0^r (1-x^2)^{\frac{d-3}{2}}\mathrm{d}x < r\sqrt{d},\end{aligned}$$

where X_1 is the first coordinate of X. □

Lemma 8.32 (Accuracy of the spectral method)**.** *Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \frac{1}{2}N(-\theta, I_d) + \frac{1}{2}N(\theta, I_d)$, where $\theta \in \mathbb{R}^d$. Let λ_S be the largest eigenvalue of $S - I_d$, where $S = \frac{1}{n}\sum_i X_i X_i^\top$ denotes the sample covariance matrix, and $\hat{v}$ the corresponding normalized eigenvector, where we decree that $\theta^\top \hat{v} \geq 0$. Let $\hat{s} = \sqrt{(\lambda_S)_+}$ and $\hat{\theta} = \hat{s}\hat{v}$. If $n > d$, then, with high probability,*

$$\|\theta - \hat{\theta}\|_2 \leq O(d/n)^{1/4}.$$

Proof. The samples can be represented in a matrix form $X = \theta\varepsilon^\top + Z \in \mathbb{R}^{d\times n}$, where $\varepsilon \in \mathbb{R}^n$ is a vector of independent Rademacher random variables, and Z has independent standard normal entries. Using $\varepsilon^\top\varepsilon = n$, we have

$$S - I_d = \theta\theta^\top + B + C,$$

where $B = \frac{1}{n}ZZ^\top - I_d$ and $C = \frac{1}{n}(\theta\varepsilon^\top Z^\top + Z\varepsilon\theta^\top)$ are both symmetric. With high probability, the largest eigenvalue of B is at most $d/n + 2\sqrt{d/n}$ (see [215, Theorem II.13]), which is $O(\sqrt{d/n})$ when $n > d$, and the spectral norm of C is also $O(\sqrt{d/n})$. Then, $|\lambda_S - \|\theta\|_2^2| \leq O(\sqrt{d/n})$ by Weyl's inequality, and thus $|\hat{s} - \|\theta\|_2| \leq O(d/n)^{1/4}$. Since $\hat{v}$ maximizes $\|u^\top(S - I_d)u\|$ among all unit vectors $u \in \mathbb{R}^d$, including the direction of θ, then we obtain that $(\theta^\top\hat{v})^2 \geq \|\theta\|_2^2 - O(\sqrt{d/n})$, and consequently,

$$\|\theta - \|\theta\|_2\hat{v}\|_2^2 \leq O(\sqrt{d/n}).$$

The conclusion follows from the triangle inequality. □

Lemma 8.33. *The boundary of the space of the first $2k-1$ moments of all distributions on $\mathbb{R}$ corresponds to distributions with fewer than k atoms, while the interior corresponds to exactly k atoms.*

Proof. Given $m = (m_1, \ldots, m_{2k-1})$ that corresponds to a distribution of ex-

actly k atoms, by [49, Theorem 2A], the moment matrix $\mathbf{M}_{k-1}$ is positive definite. For any vector m' in a sufficiently small ball around m, the corresponding moment matrix $\mathbf{M}'_{k-1}$ is still positive definite. Consequently, the matrix $\mathbf{M}'_{k-1}$ is of full rank, and thus m' is a legitimate moment vector by [9, Theorem 3.4] (or [46, Theorem 3.1]). If m corresponds to a distribution with exactly $r < k$ atoms, by [49, Theorem 2A], $\mathbf{M}_{r-1}$ is positive definite while $\mathbf{M}_r$ is rank deficient. Then, m is no longer in the moment space if m_{2r} is decreased. □

REFERENCES

[1] E. Keogh and A. Mueen, "Curse of dimensionality," in *Encyclopedia of Machine Learning*. Springer, 2011, pp. 257–258.

[2] T. M. Mitchell, *Machine Learning*. McGraw Hill, 1997.

[3] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[4] R. Szeliski, *Computer Vision: Algorithms and Applications*. Springer Science & Business Media, 2010.

[5] A. V. Oppenheim, *Discrete-time Signal Processing*. Pearson Education India, 1999.

[6] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

[7] J. A. Shohat and J. D. Tamarkin, *The Problem of Moments*. American Mathematical Soc., 1943, no. 1.

[8] S. Karlin and L. S. Shapley, *Geometry of Moment Spaces*. American Mathematical Soc., 1953, no. 12.

[9] J. B. Lasserre, *Moments, Positive Polynomials and Their Applications*. World Scientific, 2009, vol. 1.

[10] K. Schmüdgen, *The Moment Problem*. Springer, 2017.

[11] K. Pearson, "Contributions to the mathematical theory of evolution," *Philosophical Transactions of the Royal Society of London. A*, vol. 185, pp. 71–110, 1894.

[12] A. R. Hall, *Generalized Method of Moments*. Oxford University Press, 2005.

[13] B. P. Rao, *Nonparametric Functional Estimation*. Academic Press, 2014.

[14] C. J. Stone, "Optimal rates of convergence for nonparametric estimators," *The Annals of Statistics*, vol. 8, no. 6, pp. 1348–1360, 1980.

[15] D. L. Donoho and R. C. Liu, "Geometrizing rates of convergence, II," *The Annals of Statistics*, vol. 19, pp. 668–701, 1991.

[16] T. T. Cai and M. G. Low, "Nonquadratic estimators of a quadratic functional," *The Annals of Statistics*, vol. 33, no. 6, pp. 2930–2956, 2005.

[17] O. Lepski, A. Nemirovski, and V. Spokoiny, "On estimation of the L_r norm of a regression function," *Probability Theory and Related Fields*, vol. 113, no. 2, pp. 221–253, 1999.

[18] B. Efron, "Maximum likelihood and decision theory," *The Annals of Statistics*, vol. 10, no. 2, pp. pp. 340–356, 1982.

[19] J. Berkson, "Minimum chi-square, not maximum likelihood! (with discussion)," *The Annals of Statistics*, pp. 457–487, 1980.

[20] A. W. Van der Vaart, *Asymptotic Statistics*. Cambridge, United Kingdom: Cambridge University Press, 2000.

[21] B. Efron and R. Thisted, "Estimating the number of unseen species: How many words did Shakespeare know?" *Biometrika*, vol. 63, no. 3, pp. 435–447, 1976.

[22] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida, "Characterizing user behavior in online social networks," in *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement*, 2009, pp. 49–62.

[23] M. J. Berry, D. K. Warland, and M. Meister, "The structure and precision of retinal spike trains," *Proceedings of the National Academy of Sciences*, vol. 94, no. 10, pp. 5411–5416, 1997.

[24] Z. F. Mainen and T. J. Sejnowski, "Reliability of spike timing in neocortical neurons," *Science*, vol. 268, no. 5216, pp. 1503–1506, 1995.

[25] R. R. de Ruyter van Steveninck, G. D. Lewen, S. P. Strong, R. Koberle, and W. Bialek, "Reproducibility and variability in neural spike trains," *Science*, vol. 275, no. 5307, pp. 1805–1808, 1997.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[27] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: http://tensorflow.org/

[28] X. Meng, J. Bradley, B. Yavuz, E. Sparks, S. Venkataraman, D. Liu, J. Freeman, D. Tsai, M. Amde, S. Owen et al., "MLlib: Machine learning in Apache Spark," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1235–1241, 2016.

[29] A. T. Kalai, A. Moitra, and G. Valiant, "Efficiently learning mixtures of two Gaussians," in *Proceedings of the Forty-Second ACM Symposium on Theory of Computing*. ACM, 2010, pp. 553–562.

[30] A. Moitra and G. Valiant, "Settling the polynomial learnability of mixtures of Gaussians," in *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*. IEEE, 2010, pp. 93–102.

[31] M. Hardt and E. Price, "Tight bounds for learning a mixture of two Gaussians," in *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*. ACM, 2015, pp. 753–760.

[32] A. Tsybakov, *Introduction to Nonparametric Estimation*. New York, NY: Springer Verlag, 2009.

[33] V. V. Prasolov, *Polynomials*. Springer Science & Business Media, 2009, vol. 11.

[34] A. F. Timan, *Theory of Approximation of Functions of a Real Variable*. Pergamon Press, 1963.

[35] D. Jackson, *The Theory of Approximation*. American Mathematical Soc., 1930, vol. 11.

[36] R. A. DeVore and G. G. Lorentz, *Constructive Approximation*. Springer, 1993.

[37] R. T. Rockafellar, *Conjugate Duality and Optimization*. Society for Industrial & Applied Mathmatics, 1974, vol. 16.

[38] D. G. Luenberger, *Optimization by Vector Space Methods*. John Wiley & Sons, 1969.

[39] T. Cai and M. G. Low, "Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional," *The Annals of Statistics*, vol. 39, no. 2, pp. 1012–1041, 2011.

[40] P. J. Davis, *Interpolation and Approximation*. Courier Corporation, 1975.

[41] J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*, 3rd ed. New York, NY: Springer-Verlag, 2002.

[42] C. de Boor, "Divided differences," *Surveys in Approximation Theory*, vol. 1, pp. 46–49, 2005.

[43] K. E. Atkinson, *An Introduction to Numerical Analysis*. John Wiley & Sons, 1989.

[44] L. N. Trefethen, *Approximation Theory and Approximation Practice*. Siam, 2013, vol. 128.

[45] N. I. Akhiezer, *The Classical Moment Problem: and Some Related Questions in Analysis*. Oliver & Boyd, 1965, vol. 5.

[46] R. E. Curto and L. A. Fialkow, "Recursiveness, positivity, and truncated moment problems," *Houston Journal of Mathematics*, vol. 17, no. 4, pp. 603–635, 1991.

[47] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed. Cambridge University Press, 2012.

[48] J. V. Uspensky, *Introduction to Mathematical Probability*. McGraw-Hill, 1937.

[49] B. G. Lindsay, "Moment matrices: Applications in mixtures," *The Annals of Statistics*, pp. 722–740, 1989.

[50] G. H. Golub and J. H. Welsch, "Calculation of Gauss quadrature rules," *Mathematics of Computation*, vol. 23, no. 106, pp. 221–230, 1969.

[51] W. Gautschi, *Orthogonal Polynomials: Computation and Approximation*. Oxford University Press on Demand, 2004.

[52] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*. Courier Corporation, 1964.

[53] G. Szegö, *Orthogonal Polynomials*, 4th ed. Providence, RI: American Mathematical Society, 1975.

[54] V. V. Buldygin and Y. V. Kozachenko, "Sub-Gaussian random variables," *Ukrainian Mathematical Journal*, vol. 32, no. 6, pp. 483–489, 1980.

[55] Y. Wu and P. Yang, "Minimax rates of entropy estimation on large alphabets via best polynomial approximation," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3702–3720, 2016.

[56] M. Mitzenmacher and E. Upfal, *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2005.

[57] H. Strasser, *Mathematical Theory of Statistics: Statistical Experiments and Asymptotic Decision Theory*. Berlin, Germany: Walter de Gruyter, 1985.

[58] R. A. Fisher, A. S. Corbet, and C. 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*, pp. 42–58, 1943.

[59] I. J. Good, "The population frequencies of species and the estimation of population parameters," *Biometrika*, vol. 40, no. 3-4, pp. 237–264, 1953.

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

[61] S. Bhat and R. Sproat, "Knowing the unseen: Estimating vocabulary size over unseen samples," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1*, 2009, pp. 109–117.

[62] B. Kelly, A. Wagner, T. Tularak, and P. Viswanath, "Classification of homogeneous data with large alphabets," *IEEE Transactions on Information Theory*, vol. 59, no. 2, pp. 782–795, 2013.

[63] A. B. Wagner, P. Viswanath, and S. R. Kulkarni, "Probability estimation in the rare-events regime," *IEEE Trans. Inf. Theory*, vol. 57, no. 6, pp. 3207–3229, 2011.

[64] P. Valiant and G. Valiant, "Estimating the unseen: Improved estimators for entropy and other properties," in *Advances in Neural Information Processing Systems*, 2013, pp. 2157–2165.

[65] I. Ibragimov, A. Nemirovskii, and R. Khas' minskii, "Some problems on nonparametric estimation in Gaussian white noise," *Theory of Probability & Its Applications*, vol. 31, no. 3, pp. 391–406, 1987.

[66] L. Paninski, "Estimation of entropy and mutual information," *Neural Computation*, vol. 15, no. 6, pp. 1191–1253, 2003.

[67] M. Vinck, F. P. Battaglia, V. B. Balakirsky, A. H. Vinck, and C. M. Pennartz, "Estimation of the entropy based on its polynomial representation," *Physical Review E*, vol. 85, no. 5, p. 051139, 2012.

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

[69] J. Jiao, K. Venkat, Y. Han, and T. Weissman, "Minimax estimation of functionals of discrete distributions," *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2835–2885, 2015.

[70] J. Jiao, K. Venkat, Y. Han, and T. Weissman, "Maximum likelihood estimation of functionals of discrete distributions," *arXiv:1406.6959v4*, 2014.

[71] J. Acharya, H. Das, A. Orlitsky, and A. T. Suresh, "A unified maximum likelihood approach for estimating symmetric properties of discrete distributions," in *International Conference on Machine Learning*, 2017, pp. 11–21.

[72] Y. Wu and P. Yang, "Chebyshev polynomials, moment matching, and optimal estimation of the unseen," *arXiv:1504.01227*, 2015.

[73] Y. Han, J. Jiao, and T. Weissman, "Does Dirichlet prior smoothing solve the Shannon entropy estimation problem?" *arXiv:1502.00327*, 2015.

[74] Y. Han, J. Jiao, and T. Weissman, "Adaptive estimation of Shannon entropy," *arXiv:1502.00326*, 2015.

[75] A. L. Gibbs and F. E. Su, "On choosing and bounding probability metrics," *International Statistical Review*, vol. 70, no. 3, pp. 419–435, 2002.

[76] Y. Wu and S. Verdú, "The impact of constellation cardinality on Gaussian channel capacity," in *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*. IEEE, 2010, pp. 620–628.

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

[78] F. Attneave, *Applications of Information Theory to Psychology: A Summary of Basic Concepts, Methods, and Results*. Holt, Rinehart & Winston, 1959.

[79] F. Rieke, W. Bialek, D. Warland, and R. d. R. van Steveninck, *Spikes: Exploring the Neural Code*. The MIT Press, 1999.

[80] N. T. Plotkin and A. J. Wyner, "An entropy estimator algorithm and telecommunications applications," in *Maximum Entropy and Bayesian Methods*, ser. Fundamental Theories of Physics. Springer Netherlands, 1996, vol. 62, pp. 351–363.

[81] A. Porta, S. Guzzetti, N. Montano, R. Furlan, M. Pagani, A. Malliani, and S. Cerutti, "Entropy, entropy rate, and pattern classification as tools to typify complexity in short heart period variability series," *IEEE Transactions on Biomedical Engineering*, vol. 48, no. 11, pp. 1282–1291, 2001.

[82] C. Chow and C. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Trans. Inf. Theory*, vol. 14, no. 3, pp. 462–467, 1968.

[83] C. J. Quinn, N. Kiyavash, and T. P. Coleman, "Efficient methods to compute optimal tree approximations of directed information graphs," *IEEE Trans. Signal Process.*, vol. 61, no. 12, pp. 3173–3182, 2013.

[84] J. Jiao, H. H. Permuter, L. Zhao, Y.-H. Kim, and T. Weissman, "Universal estimation of directed information," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6220–6242, 2013.

[85] G. Bresler, "Efficiently learning ising models on arbitrary graphs," in *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, ser. STOC '15. New York, NY, USA: ACM, 2015. [Online]. Available: http://doi.acm.org/10.1145/2746539.2746631 pp. 771–782.

[86] G. A. Miller, "Note on the bias of information estimates," *Information Theory in Psychology: Problems and Methods*, vol. 2, pp. 95–100, 1955.

[87] B. Harris, "The statistical estimation of entropy in the non-parametric case," in *Topics in Information Theory*, I. Csiszár and P. Elias, Eds. Springer Netherlands, 1975, vol. 16, pp. 323–355.

[88] D. Braess and T. Sauer, "Bernstein polynomials and learning theory," *Journal of Approximation Theory*, vol. 128, no. 2, pp. 187–206, 2004.

[89] G. Valiant and P. Valiant, "A CLT and tight lower bounds for estimating entropy," *Electronic Colloquium on Computational Complexity (ECCC)*, 2010.

[90] J. M. Steele, "An Efron-Stein inequality for nonsymmetric statistics," *The Annals of Statistics*, pp. 753–758, 1986.

[91] D. Braess, J. Forster, T. Sauer, and H. U. Simon, "How to achieve minimax expected Kullback-Leibler distance from an unknown finite distribution," in *Algorithmic Learning Theory*. Springer, 2002, pp. 380–394.

[92] R. Dobrushin, "A simplified method of experimentally evaluating the entropy of a stationary sequence," *Theory of Probability & Its Applications*, vol. 3, no. 4, pp. 428–430, 1958.

[93] P. Valiant, "Testing symmetric properties of distributions," in *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, ser. STOC '08, 2008, pp. 383–392.

[94] G. Valiant and P. Valiant, "Estimating the unseen: An $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs," in *Proceedings of the 43rd Annual ACM Symposium on Theory of Computing*, 2011, pp. 685–694.

[95] G. Valiant and P. Valiant, "The power of linear estimators," in *Foundations of Computer Science (FOCS), 2011 IEEE 52nd Annual Symposium on*. IEEE, 2011, pp. 403–412.

[96] L. Le Cam, *Asymptotic Methods in Statistical Decision Theory*. New York, NY: Springer-Verlag, 1986.

[97] T. Batu, S. Dasgupta, R. Kumar, and R. Rubinfeld, "The complexity of approximating the entropy," *SIAM Journal on Computing*, vol. 35, no. 1, pp. 132–150, 2005.

[98] A. Nemirovski, "On tractable approximations of randomly perturbed convext constaints," *Proceedings of the 42nd IEEE Conference on Decision and Control*, pp. 2419–2422, 2003.

[99] P. P. Petrushev and V. A. Popov, *Rational Approximation of Real Functions*. Cambridge University Press, 2011.

[100] P. Yang, "Optimal property estimation on large alphabets: fundamental limits and fast algorithms," M.S. thesis, University of Illinois at Urbana-Champaign, 2016.

[101] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Academic Press, Inc., 1982.

[102] D. R. McNeil, "Estimating an author's vocabulary," *Journal of the American Statistical Association*, vol. 68, no. 341, pp. 92–96, 1973.

[103] R. Thisted and B. Efron, "Did Shakespeare write a newly-discovered poem?" *Biometrika*, vol. 74, no. 3, pp. 445–455, 1987.

[104] S.-P. Huang and B. Weir, "Estimating the total number of alleles using a sample coverage method," *Genetics*, vol. 159, no. 3, pp. 1365–1373, 2001.

[105] K. P. Burnham and W. S. Overton, "Robust estimation of population size when capture probabilities vary among animals," *Ecology*, vol. 60, no. 5, pp. 927–936, 1979.

[106] S. Raskhodnikova, D. Ron, A. Shpilka, and A. Smith, "Strong lower bounds for approximating distribution support size and the distinct elements problem," *SIAM Journal on Computing*, vol. 39, no. 3, pp. 813–842, 2009.

[107] M. Charikar, S. Chaudhuri, R. Motwani, and V. Narasayya, "Towards estimation error guarantees for distinct values," in *Proceedings of the nineteenth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*. ACM, 2000, pp. 268–279.

[108] S.-H. Lo, "From the species problem to a general coverage problem via a new interpretation," *The Annals of Statistics*, vol. 20, no. 2, pp. 1094–1109, 1992.

[109] J. Bunge and M. Fitzpatrick, "Estimating the number of species: A review," *Journal of the American Statistical Association*, vol. 88, no. 421, pp. 364–373, 1993.

[110] W. W. Esty, "Estimation of the size of a coinage: A survey and comparison of methods," *The Numismatic Chronicle (1966-)*, pp. 185–215, 1986.

[111] Z. Bar-Yossef, T. Jayram, R. Kumar, D. Sivakumar, and L. Trevisan, "Counting distinct elements in a data stream," in *Proceedings of the 6th Randomization and Approximation Techniques in Computer Science*. Springer-Verlag, 2002, pp. 1–10.

[112] R. C. Lewontin and T. Prout, "Estimation of the number of different classes in a population," *Biometrics*, vol. 12, no. 2, pp. 211–223, 1956.

[113] J. Darroch and D. Ratcliff, "A note on capture-recapture estimation," *Biometrics*, pp. 149–153, 1980.

[114] B. Harris, "Statistical inference in the classical occupancy problem unbiased estimation of the number of classes," *Journal of the American Statistical Association*, pp. 837–847, 1968.

[115] J. Marchand and F. Schroeck Jr, "On the estimation of the number of equally likely classes in a population," *Communications in Statistics-Theory and Methods*, vol. 11, no. 10, pp. 1139–1146, 1982.

[116] E. Samuel, "Sequential maximum likelihood estimation of the size of a population," *The Annals of Mathematical Statistics*, vol. 39, no. 3, pp. 1057–1068, 1968.

[117] A. Gandolfi and C. Sastri, "Nonparametric estimations about species not observed in a random sample," *Milan Journal of Mathematics*, vol. 72, no. 1, pp. 81–105, 2004. [Online]. Available: http://dx.doi.org/10.1007/s00032-004-0031-8

[118] H. E. Robbins, "Estimating the total probability of the unobserved outcomes of an experiment," *The Annals of Mathematical Statistics*, vol. 39, no. 1, pp. 256–257, 1968.

[119] A. Chao, "Nonparametric estimation of the number of classes in a population," *Scandinavian Journal of Statistics*, pp. 265–270, 1984.

[120] A. Chao and S.-M. Lee, "Estimating the number of classes via sample coverage," *Journal of the American Statistical Association*, vol. 87, no. 417, pp. 210–217, 1992.

[121] I. Good and G. Toulmin, "The number of new species, and the increase in population coverage, when a sample is increased," *Biometrika*, vol. 43, no. 1-2, pp. 45–63, 1956.

[122] C. X. Mao and B. G. Lindsay, "Estimating the number of classes," *The Annals of Statistics*, vol. 35, no. 2, pp. 917–930, 2007.

[123] A. Gandolfi and C. Sastri, "Nonparametric estimations about species not observed in a random sample," *Milan Journal of Mathematics*, vol. 72, no. 1, pp. 81–105, 2004. [Online]. Available: http://dx.doi.org/10.1007/s00032-004-0031-8

[124] L. A. Goodman, "On the estimation of the number of classes in a population," *The Annals of Mathematical Statistics*, pp. 572–579, 1949.

[125] B. M. Hill, "Posterior moments of the number of species in a finite population and the posterior probability of finding a new species," *Journal of the American Statistical Association*, vol. 74, no. 367, pp. 668–673, 1979.

[126] O. Frank, "Estimation of the number of connected components in a graph by using a sampled subgraph," *Scandinavian Journal of Statistics*, pp. 177–188, 1978.

[127] W.-C. Hou, G. Ozsoyoglu, and B. K. Taneja, "Statistical estimators for relational algebra expressions," in *Proceedings of the Seventh ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*. ACM, 1988, pp. 276–287.

[128] J. F. Naughton and S. Seshadri, "On estimating the size of projections," in *International Conference on Database Theory*. Springer, 1990, pp. 499–513.

[129] Z. Bar-Yossef, R. Kumar, and D. Sivakumar, "Sampling algorithms: Lower bounds and applications," in *Proceedings of the Thirty-Third Annual ACM Symposium on Theory of Computing*. ACM, 2001, pp. 266–275.

[130] P. Valiant, "Testing symmetric properties of distributions," *SIAM Journal on Computing*, vol. 40, no. 6, pp. 1927–1968, 2011.

[131] P. Flajolet, É. Fusy, O. Gandouet, and F. Meunier, "Hyperloglog: The analysis of a near-optimal cardinality estimation algorithm," in *In AofA'07: Proceedings of the 2007 International Conference on Analysis of Algorithms*. Citeseer, 2007.

[132] D. M. Kane, J. Nelson, and D. P. Woodruff, "An optimal algorithm for the distinct elements problem," in *Proceedings of the Twenty-Ninth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. ACM, 2010, pp. 41–52.

[133] A. Orlitsky, A. T. Suresh, and Y. Wu, "Optimal prediction of the number of unseen species," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 113, no. 47, pp. 13 283–13 288, 2016.

[134] N. J. Gotelli and R. K. Colwell, "Estimating species richness," *Biological Diversity: Frontiers in Measurement and Assessment*, vol. 12, pp. 39–54, 2011.

[135] G. Valiant, Private communication, Mar. 2017.

[136] "*Oxford English Dictionary*," http://public.oed.com/about/, accessed: 2016-02-16.

[137] G. L. Monitor, "Number of words in the English language," http://www.languagemonitor.com/?attachment_id=8505, accessed: 2016-02-16.

[138] V. K. Dzyadyk and I. A. Shevchuk, *Theory of Uniform Approximation of Functions by Polynomials*. Walter de Gruyter, 2008.

[139] R. Duffin and A. Schaeffer, "A refinement of an inequality of the brothers Markoff," *Transactions of the American Mathematical Society*, vol. 50, no. 3, pp. 517–528, 1941.

[140] R. B. Ash, *Information Theory*. New York, NY: Dover Publications Inc., 1965.

[141] A. F. Nikiforov, V. B. Uvarov, and S. K. Suslov, *Classical Orthogonal Polynomials of a Discrete Variable*. Springer, 1991.

[142] J. Todd, "The condition of the finite segments of the Hilbert matrix," *Contributions to the Solution of Systems of Linear Equations and the Determination of Eigenvalues*, vol. 39, pp. 109–116, 1954.

[143] W. Gautschi, "How (un) stable are Vandermonde systems," *Asymptotic and Computational Analysis*, vol. 124, pp. 193–210, 1990.

[144] B. Beckermann, "The condition number of real Vandermonde, Krylov and positive definite Hankel matrices," *Numerische Mathematik*, vol. 85, no. 4, pp. 553–577, 2000.

[145] A. Córdova, W. Gautschi, and S. Ruscheweyh, "Vandermonde matrices on the circle: Spectral properties and conditioning," *Numerische Mathematik*, vol. 57, no. 1, pp. 577–591, 1990.

[146] P. Ferreira, "Super-resolution, the recovery of missing samples and Vandermonde matrices on the unit circle," in *Proceedings of the Workshop on Sampling Theory and Applications, Loen, Norway*, 1999.

[147] A. Moitra, "Super-resolution, extremal functions and the condition number of Vandermonde matrices," in *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*. ACM, 2015, pp. 821–830.

[148] A. Eisinberg, P. Pugliese, and N. Salerno, "Vandermonde matrices on integer nodes: The rectangular case," *Numerische Mathematik*, vol. 87, no. 4, pp. 663–674, 2001.

[149] Y. Chen and N. Lawrence, "Small eigenvalues of large Hankel matrices," *Journal of Physics A: Mathematical and General*, vol. 32, no. 42, p. 7305, 1999.

[150] C. Jordan, *Calculus of Finite Differences*. Chelsea, 1947.

[151] L. Moser and M. Wyman, "Asymptotic development of the Stirling numbers of the first kind," *Journal of the London Mathematical Society*, vol. 1, no. 2, pp. 133–146, 1958.

[152] N. M. Temme, "Asymptotic estimates of Stirling numbers," *Studies in Applied Mathematics*, vol. 89, no. 3, pp. 233–243, 1993.

[153] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, Mar. 1963.

[154] G. Valiant, "Algorithmic approaches to statistical questions," Ph.D. dissertation, EECS Department, University of California, Berkeley, Sep 2012.

[155] R. Chelluri, L. Richmond, and N. Temme, "Asymptotic estimates for generalized Stirling number," *Analysis-International Mathematical Journal of Analysis and Its Application*, vol. 20, no. 1, pp. 1–14, 2000.

[156] S. Balakrishnan, M. J. Wainwright, and B. Yu, "Statistical guarantees for the em algorithm: From population to sample-based analysis," *The Annals of Statistics*, vol. 45, no. 1, pp. 77–120, 2017.

[157] Y. Lu and H. H. Zhou, "Statistical and computational guarantees of Lloyd's algorithm and its variants," *arXiv preprint arXiv:1612.02099*, 2016.

[158] S. B. Hopkins and J. Li, "Mixture models, robustness, and sum of squares proofs," in *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, 2018, pp. 1021–1034.

[159] J. Chen, "Optimal rate of convergence for finite mixture models," *The Annals of Statistics*, pp. 221–233, 1995.

[160] P. Heinrich and J. Kahn, "Optimal rates for finite mixture estimation," *arXiv:1507.04313*, 2015.

[161] C. Villani, *Topics in Optimal Transportation*. Providence, RI: American Mathematical Society, 2003.

[162] S. Dasgupta, "Learning mixtures of Gaussians," in *40th Annual Symposium on Foundations of Computer Science*. IEEE, 1999, pp. 634–644.

[163] C. Villani, *Optimal Transport: Old and New*. Berlin: Springer Verlag, 2008.

[164] P. Diaconis, "Application of the method of moments in probability and statistics," in *Moments in Mathematics*. Amer. Math. Soc.: Providence, RI, 1987, vol. 37, pp. 125–139.

[165] M. Krawtchouk, "Sur le problème de moments," in *ICM Proceedings*, 1932, available at https://www.mathunion.org/fileadmin/ICM/Proceedings/ICM1932.2/ICM1932.2.ocr.pdf. pp. 127–128.

[166] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (methodological)*, pp. 1–38, 1977.

[167] L. P. Hansen, "Large sample properties of generalized method of moments estimators," *Econometrica: Journal of the Econometric Society*, pp. 1029–1054, 1982.

[168] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky, "Tensor decompositions for learning latent variable models," *Journal of Machine Learning Research*, vol. 15, pp. 2773–2832, 2014.

[169] M. R. Kosorok, *Introduction to Empirical Processes and Semiparametric Inference*. Springer Science & Business Media, 2007.

[170] B. G. Lindsay, "Mixture models: Theory, geometry and applications," in *NSF-CBMS Regional Conference Series in Probability and Statistics*. JSTOR, 1995, pp. i–163.

[171] S. Frühwirth-Schnatter, *Finite mixture and Markov switching models*. Springer Science & Business Media, 2006.

[172] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Review*, vol. 26, no. 2, pp. 195–239, 1984.

[173] L. Xu and M. I. Jordan, "On convergence properties of the EM algorithm for Gaussian mixtures," *Neural Computation*, vol. 8, no. 1, pp. 129–151, 1996.

[174] D. Karlis and E. Xekalaki, "Choosing initial values for the EM algorithm for finite mixtures," *Computational Statistics & Data Analysis*, vol. 41, no. 3, pp. 577–590, 2003.

[175] W. Seidel, K. Mosler, and M. Alker, "A cautionary note on likelihood ratio tests in mixture models," *Annals of the Institute of Statistical Mathematics*, vol. 52, no. 3, pp. 481–487, 2000.

[176] X.-L. Meng and D. Van Dyk, "The EM algorithm–An old folk-song sung to a fast new tune," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 59, no. 3, pp. 511–567, 1997.

[177] R. S. Pilla and B. G. Lindsay, "Alternative EM methods for nonparametric finite mixture models," *Biometrika*, vol. 88, no. 2, pp. 535–550, 2001.

[178] J. Kiefer and J. Wolfowitz, "Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters," *The Annals of Mathematical Statistics*, pp. 887–906, 1956.

[179] N. Laird, "Nonparametric maximum likelihood estimation of a mixing distribution," *Journal of the American Statistical Association*, vol. 73, no. 364, pp. 805–811, 1978.

[180] B. G. Lindsay, "Properties of the maximum likelihood estimator of a mixing distribution," in *Statistical Distributions in Scientific Work*. Springer, 1981, pp. 95–109.

[181] R. Koenker and I. Mizera, "Convex optimization, shape constraints, compound decisions, and empirical Bayes rules," *Journal of the American Statistical Association*, vol. 109, no. 506, pp. 674–685, 2014.

[182] M. Belkin and K. Sinha, "Polynomial learning of distribution families," in *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*. IEEE, 2010, pp. 103–112.

[183] J. Li and L. Schmidt, "Robust and proper learning for mixtures of Gaussians via systems of polynomial inequalities," in *Conference on Learning Theory*, 2017, pp. 1302–1382.

[184] J. Deely and R. Kruse, "Construction of sequences estimating the mixing distribution," *The Annals of Mathematical Statistics*, vol. 39, no. 1, pp. 286–288, 1968.

[185] C. R. Genovese and L. Wasserman, "Rates of convergence for the Gaussian mixture sieve," *Annals of Statistics*, vol. 28, no. 4, pp. 1105–1127, 2000.

[186] S. Ghosal and A. W. van der Vaart, "Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities," *Annals of Statistics*, pp. 1233–1263, 2001.

[187] A. K. Kim, "Minimax bounds for estimation of normal mixtures," *Bernoulli*, vol. 20, no. 4, pp. 1802–1818, 2014.

[188] I. Ibragimov, "Estimation of analytic functions," *Lecture Notes-Monograph Series*, pp. 359–383, 2001.

[189] S. Diamond and S. Boyd, "CVXPY: A Python-embedded modeling language for convex optimization," *Journal of Machine Learning Research*, vol. 17, no. 83, pp. 1–5, 2016.

[190] H. Wolkowicz, R. Saigal, and L. Vandenberghe, *Handbook of Semidefinite Programming: Theory, Algorithms, and Applications*. Springer Science & Business Media, 2012, vol. 27.

[191] K. E. Atkinson, *An Introduction to Numerical Analysis*. John Wiley & Sons, 2008.

[192] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[193] M. Andersen, J. Dahl, and L. Vandenberghe, "CVXOPT: A Python package for convex optimization," 2013, abel.ee.ucla.edu/cvxopt.

[194] P. Chaussé, "Computing generalized method of moments and generalized empirical likelihood with R," *Journal of Statistical Software*, vol. 34, no. 11, pp. 1–35, 2010. [Online]. Available: http://www.jstatsoft.org/v34/i11/

[195] C. Améndola, K. Ranestad, and B. Sturmfels, "Algebraic identifiability of Gaussian mixtures," *International Mathematics Research Notices*, 2016.

[196] S. Vempala and G. Wang, "A spectral algorithm for learning mixture models," *Journal of Computer and System Sciences*, vol. 68, no. 4, pp. 841–860, 2004.

[197] S. Arora and R. Kannan, "Learning mixtures of arbitrary Gaussians," in *Proceedings of the Thirty-Third Annual ACM Symposium on Theory of Computing*. ACM, 2001, pp. 247–257.

[198] R. Kannan, H. Salmasian, and S. Vempala, "The spectral method for general mixture models," in *International Conference on Computational Learning Theory*. Springer, 2005, pp. 444–457.

[199] D. Achlioptas and F. McSherry, "On spectral learning of mixtures of distributions," in *International Conference on Computational Learning Theory*. Springer, 2005, pp. 458–469.

[200] S. C. Brubaker and S. Vempala, "Isotropic PCA and affine-invariant clustering," in *IEEE 49th Annual IEEE Symposium on Foundations of Computer Science, 2008.* IEEE, 2008, pp. 551–560.

[201] D. Hsu and S. M. Kakade, "Learning mixtures of spherical Gaussians: moment methods and spectral decompositions," in *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*. ACM, 2013, pp. 11–20.

[202] N. P. Jewell, "Mixtures of exponential distributions," *The Annals of Statistics*, pp. 479–484, 1982.

[203] D. Karlis and E. Xekalaki, "Mixed Poisson distributions," *International Statistical Review*, vol. 73, no. 1, pp. 35–58, 2005.

[204] C. N. Morris, "Natural exponential families with quadratic variance functions," *The Annals of Statistics*, pp. 65–80, 1982.

[205] D. F. Andrews and C. L. Mallows, "Scale mixtures of normal distributions," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 99–102, 1974.

[206] M. J. Wainwright and E. P. Simoncelli, "Scale mixtures of Gaussians and the statistics of natural images," in *Advances in Neural Information Processing Systems*, 2000, pp. 855–861.

[207] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli, "Image denoising using scale mixtures of Gaussians in the wavelet domain," *IEEE Transactions on Image Processing*, vol. 12, no. 11, pp. 1338–1351, 2003.

[208] N. Batir, "Inequalities for the gamma function," *Archiv der Mathematik*, vol. 91, no. 6, pp. 554–563, 2008.

[209] Y.-F. Ren and H.-Y. Liang, "On the best constant in Marcinkiewicz–Zygmund inequality," *Statistics & Probability Letters*, vol. 53, no. 3, pp. 227–233, 2001.

[210] J. Bustamante, *Algebraic Approximation: A Guide to Past and Current Solutions*. Springer Science & Business Media, 2011.

[211] S. Nikolsky, "On the best approximation of functions satisfying Lipshitz's conditions by polynomials," *Izvestiya Rossiiskoi Akademii Nauk. Seriya Matematicheskaya*, vol. 10, no. 4, pp. 295–322, 1946.

[212] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals Series and Products*, 7th ed. New York, NY: Academic, 2007.

[213] L. Vandenberghe and S. Boyd, "Semidefinite programming," *SIAM Review*, vol. 38, no. 1, pp. 49–95, 1996.

[214] H. G. Diamond and A. Straub, "Bounds for the logarithm of the Euler gamma function and its derivatives," *Journal of Mathematical Analysis and Applications*, vol. 433, no. 2, pp. 1072–1083, 2016.

[215] K. R. Davidson and S. J. Szarek, "Local operator theory, random matrices and Banach spaces," *Handbook of the Geometry of Banach Spaces*, vol. 1, no. 317-366, p. 131, 2001.