# OPTIMAL ENTROPY ESTIMATION ON LARGE ALPHABET: FUNDAMENTAL LIMITS AND FAST ALGORITHMS

BY

PENGKUN YANG

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2016

Urbana, Illinois

Adviser:

Assistant Professor Yihong Wu

# ABSTRACT

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 of Valiant and Valiant (2011) that the minimal sample size for consistent entropy estimation scales. 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.

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

# ACKNOWLEDGMENTS

I would like to thank my adviser, Professor Yihong Wu, who shapes my way of thinking, and even living. Every piece of progress in this thesis comes from hours of meeting every week with him. I learn from him not only immense knowledge but also passion and courage to solve the seemingly insurmountable problems. I would also like to thank my friends and the wonderful graduate students in the Coordinated Science Laboratory for their support. For the main part of this work, I would like to thank Yury Polyanskiy for the motivation and inspiring discussions in the early stage of the project, Tony Cai and Mark Low for many conversations on their result, R. Srikant for the discussion on optimization problems, and Jiantao Jiao for sharing his observation that the variance bound for the empirical entropy can be improved from $\frac{\log^2 n}{n}$ to $\frac{\log^2(k \wedge n)}{n}$.

# TABLE OF CONTENTS

LIST OF ABBREVIATIONS . . . . . . . . . . . . . . . . . . . . . . . . vi

CHAPTER 1 INTRODUCTION . . . . . . . . . . . . . . . . . . . . . . 1
1.1 Property estimation on large alphabet . . . . . . . . . . . . . 1
1.2 Entropy estimation . . . . . . . . . . . . . . . . . . . . . . . . . 2
1.3 Previous results on entropy estimation . . . . . . . . . . . . . 5

CHAPTER 2 BEST POLYNOMIAL APPROXIMATION . . . . . . . 9
2.1 Estimator design via polynomial approximation . . . . . . . . 10
2.2 Moment matching and best polynomial approximation . . . . 10
2.3 Best polynomial approximation of the logarithm function . . . 12

CHAPTER 3 POISSON SAMPLING . . . . . . . . . . . . . . . . . . 17

CHAPTER 4 MINIMAX LOWER BOUND . . . . . . . . . . . . . . 20
4.1 Le Cam's two-point method . . . . . . . . . . . . . . . . . . . . 20
4.2 Le Cam's method involving composite hypotheses . . . . . . . 21
4.3 Total variation distance between Poisson mixtures . . . . . . . 32

CHAPTER 5 OPTIMAL ESTIMATOR VIA BEST POLYNOMIAL APPROXIMATION . . . . . . . . . . . . . . . . . . . . . . . . 35
5.1 Plug-in estimator and Bernstein polynomial approximation . . 35
5.2 Unbiased estimator for the best polynomial . . . . . . . . . . . 39

CHAPTER 6 NUMERICAL EXPERIMENTS . . . . . . . . . . . . . 49

REFERENCES . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 53

# LIST OF ABBREVIATIONS

| | |
|---|---|
| BUB | best upper bound |
| CDF | cumulative distribution function |
| CLT | central limit theorem |
| i.i.d. | independently and identically distributed |
| KL | Kullback-Leibler |
| LP | linear programming |
| MLE | maximum likelihood estimator |
| MSE | mean-squared error |
| RMSE | root mean squared error |

| | |
|---|---|
| $\phi$ | the entropy function $x \mapsto x \log \frac{1}{x}$ |
| $a \vee b$ | maximum between $a$ and $b$ |
| $a \wedge b$ | minimum between $a$ and $b$ |
| $a_n \gtrsim b_n$ | $a_n \geq cb_n$ for some absolute positive constant $c$ |
| $a_n \asymp b_n$ | $a_n \gtrsim b_n$ and $b_n \gtrsim a_n$ |
| $\mathrm{Bern}(p)$ | the Bernoulli distribution with mean $p$ |
| $D(P\|Q)$ | 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 $\pi$ |

| | |
|---|---|
| $\mathbb{E}\left[\mathrm{Poi}\left(U\right)\right]$ | the Poisson mixture with respect to the distribution of a positive random variable $U$ |
| $h$ | histogram of $N$, also known as profile or fingerprint |
| $H(P)$ | Shannon entropy of a distribution $P$ |
| $\hat{H}_{\text{plug-in}}$ | empirical entropy |
| $k$ | alphabet size |
| $[k]$ | 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 |
| $\mathcal{M}_k$ | the set of probability distributions on $[k]$ |
| $N$ | histogram of original samples |
| $n$ | sample size |
| $\hat{P}$ | 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 $L$ |
| $\mathrm{Poi}(\lambda)$ | the Poisson distribution with mean $\lambda$ whose probability mass function is $\mathrm{poi}(\lambda, j) \triangleq \frac{\lambda^j e^{-\lambda}}{j!}, j \in \mathbb{Z}_+$ |
| $R^*(k, n)$ | minimax quadratic risk of Shannon entropy estimation on $[k]$ using $n$ samples |
| $R_{\text{plug-in}}(k, n)$ | worst-case MSE of empirical entropy on $[k]$ using $n$ samples |
| $\mathsf{TV}(P, Q)$ | the total variation between probability measures $P$ and $Q$ |
| $X$ | original samples: i.i.d. realizations of a distribution $P$ |

# CHAPTER 1

# INTRODUCTION

## 1.1 Property estimation on large alphabet

Learning complicated objects is difficult and sometimes requires intolerably many resources: In a data center network, learning the entire network traffic consisting of billions of flows is impossible in substance. However, in many cases, we are not most interested in the object per se but certain properties thereof, which is more tractable: In a complicated network, the most important performance measures are the throughput and the latency, which are mostly impacted by only a tiny number of large flows.

For various purposes, properties are the key evaluation criterion: In card games in a casino, fairness is partly contributed to by the uniformity of the card sequence, and the number of card shuffles needed is referred to as the mixing time; in the study of the human genome, the amount of unknown variations is connected to the total number distinct genes (both known and unknown), i.e., its support size; in the design of large scale networks, connectivity is often characterized by the graph expansion property; in the storage of big files, compressibility is measured by the randomness of the data, which is often quantified by the entropy. Understanding those properties is the key to precise evaluation.

Property estimation is one major subject studied by statisticians for hundreds of years. In classical applications the objects are often simple. Learning their properties is naturally accomplished by first estimating the objects themselves very well and then extracting the desired properties. In modern tasks this intuitive approach often fails due to the complication of objects: the estimation of the entire object is often highly inaccurate. The complication in the problem of estimating properties of a distribution, such as entropy and support size, is mainly reflected by the large alphabet. For example, in the study of the human genome, sample collection is difficult and expensive, where samples are insufficient to capture the whole alphabet: the genes. Those new tasks on large alphabet urge new and fast algorithms, and also demand new theory to quantify what is the best we can do.

The main focus of this thesis is the estimation of the Shannon entropy. Analogous techniques have subsequently been used in [1] to obtain sharp minimax risk for estimating the power sum, and used in [2] to obtain the sharp sample complexity for estimating the support size (number of distinct elements).

## 1.2 Entropy estimation

Entropy estimation has found numerous applications across various fields, such as neuroscience [3], physics [4], telecommunication [5], biomedical research [6], 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 [7, 8, 9, 10].

Let $P$ be a distribution over an alphabet of cardinality $k$. Let $X_1, \ldots, X_n$ be independently and identically distributed (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

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

records the number of occurrences of $j \in [k]$ in the sample. Then $N \sim \text{Multinomial}(n, P)$. The problem of focus is to estimate the Shannon entropy of the distribution $P$:

$$H(P) = \sum_{i=1}^{k} p_i \log \frac{1}{p_i}.$$

From a statistical standpoint, the problem of entropy estimation falls under the category of *functional estimation*, where we are not interested in directly estimating the high-dimensional parameter (the distribution $P$) per se, but rather a function thereof (the entropy $H(P)$). Estimating a scalar functional has been intensively studied in nonparametric statistics, e.g., estimate a scalar function of a regression function such as linear functional [11, 12], quadratic functional [13], $L_q$ norm [14], 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. This leads to the commonly used plug-in estimator, i.e., the empirical entropy,

$$\hat{H}_{\text{plug-in}} = H(\hat{P}), \tag{1.1}$$

where $\hat{P} = (\hat{p}_1, \ldots, \hat{p}_k)$ denotes the empirical distribution with $\hat{p}_i = \frac{N_i}{n}$. As frequently observed in functional estimation problems, the plug-in estimator can suffer from severe bias (see [15, 16] and the references therein). Indeed, although $\hat{H}_{\text{plug-in}}$ is asymptotically efficient and minimax (cf., e.g., [17, Sections 8.7 and 8.9]), in the "fixed-$k$-large-$n$" regime, it can be highly suboptimal in high dimensions, where, due to the 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 [18];
- network traffic analysis: many customers or website users are only seen a small number of times [19];
- analyzing neural spike trains: natural stimuli generate neural responses of high timing precision resulting in a massive space of meaningful responses [20, 21, 22].

Statistical inference on large alphabets with insufficient samples has a rich history in information theory, statistics and computer science, with early contributions dating back to Fisher [23], Good and Turing [24], Efron and Thisted [18] and recent renewed interest in compression, prediction, classification and estimation aspects for large-alphabet sources [25, 26, 27, 28, 29]. However, none of the current results allow a general understanding of the fundamental limits of estimating information quantities of distributions on large alphabets. The particularly interesting case is when the sample size scales *sublinearly* with the alphabet size.

To investigate the decision-theoretic fundamental limit, we consider the minimax quadratic risk of entropy estimation:

$$R^*(k,n) \triangleq \inf_{\hat{H}} \sup_{P \in \mathcal{M}_k} \mathbb{E}_P[(\hat{H}(N) - H(P))^2], \tag{1.2}$$

where $\mathcal{M}_k$ denotes the set of probability distributions on $[k]$. The goal is a) to provide a constant-factor approximation of the minimax risk $R^*(k,n)$, b) to devise a linear-time estimator that provably attains $R^*(k,n)$ within universal constant factors. Our main result is the characterization of the minimax risk within universal constant factors:

**Theorem 1.** *If $n \gtrsim \frac{k}{\log k}$,[1] then*

$$R^*(k,n) \asymp \left(\frac{k}{n \log k}\right)^2 + \frac{\log^2 k}{n}. \tag{1.3}$$

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

To interpret the minimax rate in Equation (1.3), 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 $R^*(k,n) \asymp (\frac{k}{n \log k})^2$ if and only if $n \lesssim \frac{k^2}{\log^4 k}$, where the bias dominates. As a consequence, the minimax rate in Theorem 1 implies that to estimate the entropy within $\epsilon$ bits with probability, say 0.9, the minimal sample size is given by

$$n \asymp \frac{\log^2 k}{\epsilon^2} \vee \frac{k}{\epsilon \log k}. \tag{1.4}$$

Next we evaluate the performance of plug-in estimator in terms of its worst-case mean-square error

$$R_{\text{plug-in}}(k,n) \triangleq \sup_{P \in \mathcal{M}_k} \mathbb{E}_P[(\hat{H}_{\text{plug-in}}(N) - H(P))^2]. \tag{1.5}$$

Analogous to Theorem 1 which applies to the optimal estimator, the risk of

[1] For any sequences $\{a_n\}$ and $\{b_n\}$ of positive numbers, we write $a_n \gtrsim b_n$ or $b_n \lesssim a_n$ when $a_n \geq c b_n$ for some absolute constant $c$. Finally, we write $a_n \asymp b_n$ when both $a_n \gtrsim b_n$ and $a_n \lesssim b_n$ hold.

the plug-in estimator admits a similar characterization (see Section 5.1 for details):

**Proposition 1.** *If $n \gtrsim k$, then*

$$R_{\text{plug-in}}(k,n) \asymp \left(\frac{k}{n}\right)^2 + \frac{\log^2 k}{n}. \tag{1.6}$$

*If $n \lesssim k$, then $\hat{H}_{\textit{plug-in}}$ is inconsistent, i.e., $R_{\text{plug-in}}(k,n) \gtrsim 1$.*

Note that the first and second terms in the risk in Equation (1.6) again correspond to the squared bias and variance respectively. While it is known that the bias can be as large as $\frac{k}{n}$ [30], the variance of the plug-in estimator is at most a constant factor of $\frac{\log^2 n}{n}$, regardless of the alphabet size (see, e.g., [31, Remark (iv), p. 168]). This variance bound can in fact be improved to $\frac{\log^2(k \wedge n)}{n}$ by a more careful application of Steele's inequality [32], and hence the mean-square error (MSE) is upper bounded by $\left(\frac{k}{n}\right)^2 + \frac{\log^2(k\wedge n)}{n} \asymp \left(\frac{k}{n}\right)^2 + \frac{\log^2 k}{n}$, which turns out to be the sharp characterization.

Comparing Equation (1.3) and Equation (1.6), 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.

## 1.3 Previous results on entropy estimation

Below we give a concise overview of the previous results on entropy estimation. There also exists a vast amount of literature on estimating (differential) entropy on continuous alphabets which is outside the present focus (see the survey [33] and the references therein).

**Fixed alphabet** For fixed distribution $P$ and $n \to \infty$, Antos and Kontoyiannis [31] showed that the plug-in estimator is always consistent and the asymptotic variance of the plug-in estimator is obtained in [34]. However, the convergence rate of the bias can be arbitrarily slow on a possibly infinite alphabet. The asymptotic expansion of the bias is obtained in, e.g., [35, 36]:

$$\mathbb{E}[\hat{H}_{\text{plug-in}}(N)] = H(P) - \frac{S(P)-1}{2n} + \frac{1}{12n^2}\left(1 - \sum_{i=1}^{k} \frac{1}{p_i}\right) + O(n^{-3}), \quad (1.7)$$

where $S(P) = \sum_i \mathbf{1}_{\{p_i > 0\}}$ denote the support size. This inspired various types of bias reduction to the plug-in estimator, such as the Miller-Madow estimator [35]:

$$\hat{H}_{\text{MM}} = \hat{H}_{\text{plug-in}} + \frac{\hat{S}-1}{2n}, \quad (1.8)$$

where $\hat{S}$ is the number of observed distinct symbols.

**Large alphabet** 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., [37]). 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 [38], 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 $k \to \infty$ [39]. Valiant proved that no consistent estimator exists, i.e., $R^*(k, n_k) \gtrsim 1$ if $n \lesssim \frac{k}{\exp(\sqrt{\log k})}$ [40]. 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 [41, 42]. However, the optimal sample size as a function of alphabet size $k$ and estimation error $\epsilon$ has not been completely resolved. Indeed, an estimator based on linear programming is shown to achieve an additive error of $\epsilon$ using $\frac{k}{\epsilon^2 \log k}$ samples [29, Theorem 1], while $\frac{k}{\epsilon \log k}$ samples are shown to be necessary [41, Corollary 10]. This gap is partially amended in [43] by a different estimator, which requires $\frac{k}{\epsilon \log k}$ samples but only valid when $\epsilon > k^{-0.03}$. Theorem 1 generalizes their result by characterizing the full minimax rate and the sharp sample complexity is given by Equation (1.4).

We briefly discuss the difference between the lower bound strategy of [41] 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 $N$:

$$h_i = \sum_{j=1}^{k} \mathbf{1}_{\{N_j = i\}}, \quad i \in [n], \tag{1.9}$$

also known as *histogram order statistics* [30], *profile* [25], or *fingerprint* [41], 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 [44, Chapter 2], i.e., finding two distributions which have very different entropy but induce almost the same distribution for the sufficient statistics, in this case, the histogram $N_1^k$ or the fingerprints $h_1^n$, both of which have non-product distributions. A frequently used technique to reduce dependence is *Poisson sampling* (see Chapter 3), 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 $N_1, \dots, N_k$ 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 [41] 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 [40] obtained using Poisson approximation.

In contrast, we shall not deal with the fingerprint directly, but rather use the original sufficient statistics $N_1^k$ 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 the focus of this thesis is estimating the entropy under the additive error criterion, approximating the entropy multiplicatively has been considered in [45]. 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 $\gamma$ using $n \lesssim k^{(1+\eta)/\gamma^2} \log k$ number of samples ([45, Theorem 2]).

# CHAPTER 2

# BEST POLYNOMIAL APPROXIMATION

The theory of approximation has a long history. It represents one logic of mathematical analysis that pursues the simplification of more complex objects. Taylor's expansion is one approximation of abstract differentiable functions by polynomials. It also has profound and extensive impact in scientific and engineering fields. Truncated Fourier series is one approximation of periodic functions by trigonometrics which is extended to Fourier transform laying a foundation of signal processing. Closely related discrete cosine transform (DCT) after quantization is used in JPEG files we view every day. With little perceptible loss of quality, it saves a lot of storage space. More generally, the theory of approximation deals with the projection of a complex space to a simpler subspace, often a finite-dimensional subspace. A fundamental theorem is that in linear normed space, the best approximation by finite linearly independent elements does exist. Further, in Hilbert space, the best approximation has a nice geometric interpretation characterized by orthogonal principle.

The proof of both the upper and the lower bound in Theorem 1 relies on the apparatus of *best polynomial approximation*. Our inspiration comes from previous work on functional estimation in Gaussian mean models [14, 46]. Nemirovski (credited in [47]) 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 [14] using best polynomial approximation and in [46] for estimating the $\ell_1$-norm in Gaussian mean model.

## 2.1 Estimator design via polynomial approximation

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., [30, Proposition 8]) that to estimate a functional $f(P)$ using $n$ samples, an unbiased estimator exists if and only if $f(P)$ is a polynomial in $P$ of degree at most $n$. Similarly, under Poisson sample model, $f(P)$ admits an unbiased estimator if and only if $f$ is real analytic. Consequently, there exists no unbiased entropy estimator with or without Poissonized sampling. Therefore, a natural idea is to approximate the entropy 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 entropy estimation is not new. In [4], 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 [30], which forms the basis for proving the existence of consistent estimators with sublinear sample complexity in [39].

Shortly before we posted our result to arXiv, we learned that Jiao et al. [1] independently used the idea of best polynomial approximation 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 [48, 49, 2, 50, 51]. In particular, [51] sharpened Theorem 1 by giving a constant-factor characterization of the minimax risk in the regime of $n \lesssim \frac{k}{\log k}$ using similar techniques.

## 2.2 Moment matching and best polynomial approximation

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 [14, 46], 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. For entropy estimation, this approach yields the optimal minimax lower bound, although the argument is considerably more involved due to the extra constraint imposed by probability vectors.

In the remainder of this section we discuss the relationship between moment matching and best polynomial approximation and, in particular, provide a short proof that they are dual of each other. Denote by $\mathcal{P}_L$ the set of polynomials of degree $L$ and let $g$ be a continuous function on the interval $[a,b]$. Abbreviate by $\hat{\mathcal{E}}^*$ the best uniform approximation error

$$\hat{\mathcal{E}}^* \triangleq E_L(g,[a,b]) \triangleq \inf_{p\in\mathcal{P}_L} \sup_{x\in[a,b]} |g(x)-p(x)| \,.$$

Let $\mathcal{S}_L = \{(X,X') \in [a,b]^2 : \mathbb{E}[X^j] = \mathbb{E}[X'^j], j=1,\ldots,L\}$. For any polynomial $p \in \mathcal{P}_L$, we have

$$\begin{aligned}\mathcal{E}^* &\triangleq \sup_{(X,X')\in\mathcal{S}_L} \mathbb{E}[g(X)] - \mathbb{E}[g(X')] \\ &= \sup_{(X,X')\in\mathcal{S}_L} \mathbb{E}[g(X)-p(X)] - \mathbb{E}[g(X')-p(X')],\end{aligned}$$

and therefore by triangle inequality

$$\begin{aligned}\mathcal{E}^* &= \inf_{p\in\mathcal{P}_L} \sup_{(X,X')\in\mathcal{S}_L} \mathbb{E}[g(X)-p(X)] - \mathbb{E}[g(X')-p(X')] \\ &\le 2 \inf_{p\in\mathcal{P}_L} \sup_{x\in[a,b]} |g(x)-p(x)| = 2E_L(g,[a,b]).\end{aligned}$$

For the achievability part, Chebyshev alternating theorem [52, Theorem 1.6] states that there exists a (unique) polynomial $p^* \in \mathcal{P}_L$ and at least $L+2$ points $a \le x_1 < \cdots < x_{L+2} \le b$ and $\alpha \in \{0,1\}$ such that $g(x_i) - p^*(x_i) = (-1)^{i+\alpha}\hat{\mathcal{E}}^*$. Fix any $l = 0,1,\ldots,L$, define a Lagrange interpolation polynomial

$$f_l(x) \triangleq \sum_{j=1}^{L+2} x_j^l \frac{\prod_{v\neq j}(x-x_v)}{\prod_{v\neq j}(x_j-x_v)}$$

satisfying that $f_l(x_j) = x_j^l$ for $j = 1,\ldots,L+2$. Since $f_l$ has degree $L+1$, it must be that $f_l(x) = x^l$. Note that the coefficient of $x^{L+1}$ of polynomial $f_l$ is 0, i.e., $\sum_i x_i^l b_i = 0$ where $b_i \triangleq (\prod_{v \neq i}(x_i - x_v))^{-1}$. Define $w_i = \frac{2b_i}{\sum_j |b_j|}$, then $\sum_i |w_i| = 2$. When $l = 0$ then $\sum_i b_i = 0$ so $\sum_i w_i = 0$. Note that $w_i$ change signs alternatively. Construct discrete random variables $X, X'$ with distributions $\mathbb{P}[X = x_i] = |w_i|$ for $i$ odd and $\mathbb{P}[X' = x_i] = |w_i|$ for $i$ even. Then $(X, X') \in S_L$. The property of those $L+2$ points that $g(x_i) - p^*(x_i) = (-1)^{i+\alpha}\hat{\mathcal{E}}^*$ yields that $|\mathbb{E}[g(X) - p^*(X)] - \mathbb{E}[g(X') - p^*(X')]| = 2\hat{\mathcal{E}}^*$.

**Remark 1.** Alternatively, the achievability part can be argued from an optimization perspective (zero duality gap, see [53, Exercise 8.8.7, p. 236]), or using the Riesz representation of linear operators as in [54], which has been used in [14] and [46].

## 2.3 Best polynomial approximation of the logarithm function

As a concrete example of best polynomial approximation, we consider the approximation of logarithmic function. In particular we provide a proof that, for some universal positive constants $c, c', L_0$ such that for any $L \geq L_0$,

$$E_{\lfloor cL \rfloor}(\log, [L^{-2}, 1]) \geq c', \tag{2.1}$$

which will be useful in the proof of our minimax lower bound.

For definiteness let $E_m(f) \triangleq E_m(f, [-1, 1])$. 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

$$E_{cL}(f_L) \gtrsim 1 \tag{2.2}$$

lies in the fact that the approximand $f_L$ changes with the degree $L$. In fact, the following asymptotic result has been shown in [55, 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 Equation (2.2) 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 Equation (2.2), 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., [52, 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 $\tau_1$ for $f_L$:

**Lemma 1** (Direct bound)**.**

$$\tau_1(f_L, \Delta_m) \le \log\left(\frac{2L^2}{m^2}\right), \ \forall m \le 0.1L. \tag{2.3}$$

**Lemma 2** (Converse bound)**.**

$$\tau_1(f_L, \Delta_L) \ge 1, \forall L \ge 10. \tag{2.4}$$

From [52, 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 1 gives us

$$\begin{aligned}\frac{1}{L}\sum_{m=1}^{cL} E_m(f_L) \le \frac{100}{L}\sum_{m=1}^{cL}\log\left(\frac{2L^2}{m^2}\right) &= 100c\log 2 + \frac{200}{L}\log\frac{L^{cL}}{(cL)!} \\ &< \frac{1}{400} - \frac{100}{L}\log(2\pi cL), \end{aligned}\tag{2.5}$$

where the last inequality follows from Stirling's approximation $n! > \sqrt{2\pi n}(\frac{n}{e})^n$. We apply the converse result for approximation in [52, Theorem 3.14] that

$$\tau_1(f_L, \Delta_L) \le \frac{100}{L}\sum_{m=0}^{L} E_m(f_L), \tag{2.6}$$

where $E_0(f_L) = \log L$. Assembling Equation (2.4)–Equation (2.6), we obtain that, for all $c \le 10^{-7}$ and $L > 10 \vee \left(100 \times 400 \log \frac{1}{2\pi c}\right)$,

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

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 $L > 4 \times 10^4 \log \frac{1}{2\pi c}$,

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

**Remark 2.** From the direct bound Lemma 1 we know that $E_{cL}(\log, [1/L^2, 1]) \lesssim 1$. Therefore the bound Equation (2.1) is in fact tight: $E_{cL}(\log, [1/L^2, 1]) \asymp 1$.

*Proof of Lemmas 1 and 2.* First we show Equation (2.3). 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 $\xi_1$ 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

$$x_m = \frac{m^2 - m^4 + \sqrt{-m^2 + 3m^4}}{m^2 + m^4}. \tag{2.7}$$

Note that $\Delta_m$ 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 $m \ge 2$ since $\xi_1(0) > -1$ then $x_m < 0$ and $-x_m > 0$. Therefore,

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

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,

$$\begin{aligned}\tau_1(f_L, \Delta_m) &= \sup_{x\in[-1,x_m)} \{f_L(x) - f_L(\xi_2(x))\} \vee \sup_{x\in[x_m,1]} \{f_L(\xi_1(x)) - f_L(x)\} \\ &= \sup_{x\in[-1,x_m)} \{\log(1+\beta_L(x))\} \vee \sup_{x\in[x_m,1]} \{-\log(1-\beta_L(x))\}, \quad (2.8)\end{aligned}$$

where $\beta_L(x) \triangleq \frac{\Delta_m(x)}{x+\frac{L^2+1}{L^2-1}}$. If $m = 1$ then $x_1 > 0$ and $-x_1 < 0$ by Equation (2.7) and

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

Since $f_L(\xi_2(x)\wedge 1) \geq f_L(\xi_2(x))$, by the same argument, Equation (2.8) remains a valid upper bound of $\tau_1(f_L, \Delta_1)$. Next we will show separately that the two terms in Equation (2.8) both satisfy the desired upper bound.

For the first term in Equation (2.8), 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 $\sqrt{1-x^2}+\frac{1}{L} \leq L(x+1)+\frac{2}{L}$ 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,

$$\sup_{x\in[-1,x_m)} \{\log(1+\beta_L(x))\} \leq \log\left(\frac{2L^2}{m^2}\right), \ \forall m \leq L. \quad (2.9)$$

For the second term in Equation (2.8), it follows from the derivative of $\beta_L(x)$ that it is decreasing when $x > \frac{1-L^2}{1+L^2}$. From Equation (2.7) we have $x_m > \frac{1-m^2}{1+m^2}$ and hence $x_m > \frac{1-L^2}{1+L^2}$ when $m \leq L$. So the supremum is achieved exactly at the left end of $[x_m, 1]$, that is:

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

From Equation (2.7) we know that $x_m \geq -1$ and $x_m < -1 + \frac{3.8}{m^2}$. Therefore $\frac{1-x_m}{2} \leq 1$ and $\frac{x_m+1}{2} < \frac{1.9}{m^2}$. For $m \leq 0.1L$, we have

$$\sup_{x \in [x_m, 1]} \{-\log(1 - \beta_L(x))\} \leq \log\left(1 + \frac{1.9L^2}{m^2}\right) \leq \log\left(\frac{2L^2}{m^2}\right). \tag{2.10}$$

Plugging Equation (2.9) and Equation (2.10) into Equation (2.8), we complete the proof of Lemma 1.

Next we prove Equation (2.4). Recall that $x_L - \Delta_L(x_L) = -1$. By definition,

$$\begin{aligned}\tau_1(f_L, \Delta_L) \geq f_L(x_L - \Delta_L(x_L)) - f_L(x_L) &= \log\left(\frac{1+x_L}{2}L^2 + \frac{1-x_L}{2}\right)\\ &\geq \log\left(\frac{2L^2 + \sqrt{-L^2+3L^4}}{2(L^2+1)} + \frac{2L^4 - \sqrt{-L^2+3L^4}}{2(L^2+L^4)}\right) \geq 1\end{aligned}$$

when $L \geq 10$, where we used the close-form expression of $x_L$ in Equation (2.7).

□

# CHAPTER 3

# POISSON SAMPLING

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. In view of the marginal distribution of histogram, this is the commonly used Poisson approximation for binomial distribution: the histogram under fixed samples size $N_i \sim \mathrm{Binomial}(n, p_i)$ is approximated by $N_i \sim \mathrm{Poi}(np_i)$.

Analogous to the minimax risk Equation (1.2), we define its counterpart under the Poisson sampling model:

$$\tilde{R}^*(k,n) \triangleq \inf_{\hat{H}} \sup_{P \in \mathcal{M}_k} \mathbb{E}(\hat{H}(N) - H(P))^2, \tag{3.1}$$

where $N_i \overset{\text{ind}}{\sim} \mathrm{Poi}(np_i)$ 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 have the following inequalities which allow us to focus on the risk of the Poisson model:

**Proposition 2.** *For any $\alpha > 0$,*

$$R^*(k,n) \geq \tilde{R}^*(k, (1+\alpha)n) - \exp\left(-n(\alpha - \log(1+\alpha))\right) \log^2 k$$

*Proof.* Fix an arbitrary distribution $P$. Let $N = (N_1, N_2, \ldots) \overset{\text{ind}}{\sim} \mathrm{Poi}((1+\alpha)np_i)$ and let $n' = \sum N_i \sim \mathrm{Poi}((1+\alpha)n)$. Let $\hat{H}_n(\cdot)$ be the optimal estimator of Shannon entropy for fixed sample size $n$, i.e.,

$$\mathbb{E}(\hat{H}_n(N) - H(P))^2 \leq R^*(k,n), \quad \forall\ P \in \mathcal{M}_k.$$

We construct an estimator for the Poisson sampling model by $\tilde{H}(N) = \hat{H}_{n'}(N)$. We observe that conditioned on $n' = m$, $N \sim \text{Multinomial}(m, P)$. Therefore,

$$\begin{aligned}\mathbb{E}(\tilde{H}(N) - H(P))^2 &= \sum_{m=0}^{\infty} \mathbb{E}\left[\left(\hat{H}_{n'}(N) - H\left(P\right)\right)^2 \middle| n' = m\right] \mathbb{P}[n' = m] \\ &\leq \sum_{m=0}^{\infty} R^*(k,m)\mathbb{P}\left[n' = m\right].\end{aligned}$$

Note that for fixed $k$, the minimax risk $n \mapsto R^*(k,n)$ is decreasing and $0 \leq R^*(k,n) \leq \log^2 k$. Then,

$$\begin{aligned}\tilde{R}^*(k,(1+\alpha)n) &\leq \sum_{m \geq n} R^*(k,m)\mathbb{P}[n' = m] + \log^2 k\mathbb{P}[\text{Poi}((1+\alpha)n) < n] \\ &\leq R^*(k,n) + \exp(-n(\alpha - \log(1+\alpha)))\log^2 k,\end{aligned}$$

where in the last inequality we used the Chernoff bound (see, e.g., [56, Theorem 5.4]). The conclusion follows. □

**Proposition 3.** *For any* $0 < \beta < 1$,

$$R^*(k,n) \leq \frac{\tilde{R}^*(k,(1-\beta)n)}{1 - \exp(-n\beta^2/2)}$$

*Proof.* This inequality is slightly more involved. First, by the minimax theorem (cf. e.g. [57, Theorem 46.5]),

$$R^*(k,n) = \sup_{\pi} \inf_{\hat{H}_n} \mathbb{E}[(\hat{H}_n - H(P))^2], \tag{3.2}$$

where $\pi$ ranges over all probability distributions (priors) on the simplex $\mathcal{M}_k$ and the expectation is over $P \sim \pi$ and $X_1, \ldots \overset{\text{i.i.d.}}{\sim} P$ conditioned on $P$.

To this end, it is convenient to express the estimator as a function of the original samples instead of the sufficient statistic (histogram). Consequently, under the Poisson sampling model we have a sequence of estimators $\{\hat{H}_m\}$.

The Bayesian risk is a lower bound of the minimax risk, so, for any $\beta < 1$,

$$\tilde{R}^*(k, (1-\beta)n) \geq \sup_{\pi} \inf_{\{\hat{H}_m\}} \mathbb{E}[(\hat{H}_{n'} - H(P))^2], \tag{3.3}$$

where $n' \sim \mathrm{Poi}((1-\beta))$. For any sequence of estimators $\{\hat{H}_m\}$,

$$\begin{aligned} \mathbb{E}[(\hat{H}_{n'} - H(P))^2] &= \sum_{m \geq 0} \mathbb{E}[(\hat{H}_m - H(P))^2]\mathbb{P}[n' = m] \\ &\geq \sum_{m \geq 0}^{n} \mathbb{E}[(\hat{H}_m - H(P))^2]\mathbb{P}[n' = m]. \end{aligned}$$

Taking infimum on both sides, we obtain that

$$\begin{aligned} \inf_{\{\hat{H}_m\}} \mathbb{E}[(\hat{H}_{n'} - H(P))^2] &\geq \inf_{\{\hat{H}_m\}} \sum_{m \geq 0}^{n} \mathbb{E}[(\hat{H}_m - H(P))^2]\mathbb{P}[n' = m] \\ &\geq \sum_{m \geq 0}^{n} \inf_{\hat{H}_m} \mathbb{E}[(\hat{H}_m - H(P))^2]\mathbb{P}[n' = m]. \end{aligned}$$

Note that the Bayesian risk $\inf_{\hat{H}_m} \mathbb{E}[(\hat{H}_m - H(P))^2]$ is monotonic decreasing in the sample size $m$. Therefore,

$$\begin{aligned} \inf_{\{\hat{H}_m\}} \mathbb{E}[(\hat{H}_{n'} - H(P))^2] &\geq \inf_{\hat{H}_n} \mathbb{E}[(\hat{H}_n - H(P))^2]\mathbb{P}[n' \leq n] \\ &\geq \inf_{\hat{H}_n} \mathbb{E}[(\hat{H}_n - H(P))^2](1 - \exp(n(\beta + \log(1-\beta)))) \\ &\geq \inf_{\hat{H}_n} \mathbb{E}[(\hat{H}_n - H(P))^2](1 - \exp(-n\beta^2/2)), \end{aligned} \tag{3.4}$$

where we used Chernoff bound (see, e.g., [56, Theorem 5.4]) and the fact that $\log(1-x) \leq -x - x^2/2$. Taking supremum over $\pi$ on both sides of Equation (3.4), the conclusion follows from Equation (3.3) and minimax theorem Equation (3.2). □

# CHAPTER 4

# MINIMAX LOWER BOUND

In this chapter we give converse results for entropy estimation and prove the lower bound part of Theorem 1. It suffices to show that the minimax risk is lower bounded by the two terms in Equation (1.3) separately, i.e.,

$$R^*(k,n) \gtrsim \frac{\log^2 k}{n},$$

and

$$R^*(k,n) \gtrsim \left(\frac{k}{n\log k}\right)^2.$$

## 4.1 Le Cam's two-point method

Our first lower bound 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 $|H(P) - H(Q)|^2$.

**Proposition 4.** *For all $k, n \in \mathbb{N}$,*

$$R^*(k,n) \gtrsim \frac{\log^2 k}{n}. \tag{4.1}$$

*Proof.* For any pair of distributions $P$ and $Q$, Le Cam's two-point method (see, e.g., [58, Section 2.4.2]) yields

$$R^*(k,n) \geq \frac{1}{4}(H(P) - H(Q))^2 \exp(-nD(P\|Q)). \tag{4.2}$$

Therefore it boils down to solving the optimization problem:

$$\sup\{H(P) - H(Q) : D(P\|Q) \leq 1/n\}. \tag{4.3}$$

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

$$\begin{aligned} P &= \left(\frac{1}{3(k-1)}, \ldots, \frac{1}{3(k-1)}, \frac{2}{3}\right), \\ Q &= \left(\frac{1+\epsilon}{3(k-1)}, \ldots, \frac{1+\epsilon}{3(k-1)}, \frac{2-\epsilon}{3}\right). \end{aligned} \tag{4.4}$$

Direct computation yields that

$$D(P\|Q) = \frac{2}{3}\log\frac{2}{2-\epsilon} + \frac{1}{3}\log\frac{1}{\epsilon+1} \leq \epsilon^2$$

and

$$\begin{aligned} H(Q) - H(P) &= \frac{1}{3}\left(\epsilon\log(k-1) + \log 4 + (2-\epsilon)\log\frac{1}{2-\epsilon} + (1+\epsilon)\log\frac{1}{\epsilon+1}\right) \\ &\geq \frac{1}{3}\log(2(k-1))\epsilon - \epsilon^2. \end{aligned}$$

Choosing $\epsilon = \frac{1}{\sqrt{n}}$ and applying Equation (4.2), we obtain the desired Equation (4.1). □

**Remark 3.** In view of the Pinsker inequality $D(P\|Q) \geq 2\mathsf{TV}^2(P,Q)$ [59, 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 $\mathsf{TV}(P,Q) \leq \frac{1}{4}$ [59, Lemma 2.7], we conclude that the best lower bound given by the two-point method, i.e., the supremum in Equation (4.3), is on the order of $\frac{\log k}{\sqrt{n}}$. Therefore the choice of the pair Equation (4.4) is optimal.

## 4.2 Le Cam's method involving composite hypotheses

This section is devoted to outlining the broad strokes for proving the lower bound by the first term of Equation (1.3). Since it can be shown that the best lower bound provided by the two-point method is $\frac{\log^2 k}{n}$ (see Remark 3), proving Equation (4.11) 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 [58]):

$$H_0 : H(P) \le t \quad \text{versus} \quad H_1 : H(P) \ge t + d, \tag{4.5}$$

which is more general than the two-point argument using only simple hypothesis testing. Similarly, if we can establish that no test can distinguish Equation (4.5) 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 Equation (4.5) 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.

### 4.2.1 Construction of the priors

The main idea for constructing the priors is as follows: First of all, 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 concision, 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 $\mathsf{P}$ 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 $\mathsf{P}$ is *approximately* a probability vector. Using a conditioning argument we can show that the distribution of $\mathsf{P}$ can effectively serve as a prior. To gain more insight, note that, for example, a deterministic $U = 1$ generates a uniform distribution over $[k]$, while a binary $U \sim \frac{1}{2}(\delta_0 + \delta_2)$ generates a uniform distribution over roughly half the alphabet with the support set uniformly chosen at random. From this viewpoint, the cumulative distribution function (CDF) of the random variable $\frac{U}{k}$ plays the role of the *histogram of the distribution* $\mathsf{P}$, which is the central object in the Valiant-Valiant lower bound construction (see [41, Definition 3]).

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

$$H(\mathsf{P}) = \sum_{i=1}^{k} \phi\left(\frac{U_i}{k}\right) = \frac{1}{k}\sum_{i=1}^{k}\phi(U_i) + \frac{\log k}{k}\sum_{i=1}^{k} U_i, \tag{4.6}$$

   which concentrates near its mean $\mathbb{E}\left[H(\mathsf{P})\right] = \mathbb{E}\left[\phi(U)\right] + \mathbb{E}\left[U\right]\log k$ by law of large numbers. Therefore, given another random variable $U'$ with unit mean, we can obtain $\mathsf{P}'$ 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,

$$\mathbb{E}\left[H(\mathsf{P})\right] - \mathbb{E}\left[H(\mathsf{P}')\right] = \mathbb{E}\left[\phi(U)\right] - \mathbb{E}\left[\phi(U')\right], \tag{4.7}$$

   which we aim to maximize.

2. *Indistinguishability*: Note that given $P$, the sufficient statistics satisfy $N_i \overset{\text{ind}}{\sim} \text{Poi}(np_i)$. Therefore, if $P$ is drawn from the distribution of $\mathsf{P}$, then $N = (N_1, \ldots, N_k)$ are i.i.d. distributed according the *Poisson mixture* $\mathbb{E}[\text{Poi}(nU/k)]$. Similarly, if $P$ is drawn from the prior of $\mathsf{P}'$, then $N$ is distributed according to $(\mathbb{E}[\text{Poi}(nU'/k)])^{\otimes k}$. 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

$$\mathsf{TV}(\mathbb{E}[\text{Poi}(nU/k)], \mathbb{E}[\text{Poi}(nU'/k)]) \leq c/k \tag{4.8}$$

   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 [41] 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 $|\mathbb{E}\left[\phi(U)\right] - \mathbb{E}\left[\phi(U')\right]|$ subject to the constraint in Equation (4.8). A commonly used proxy for bounding the total variation distance is *moment matching*, i.e., $\mathbb{E}\left[U^j\right] = \mathbb{E}\left[U'^j\right]$ for all $j = 1, \ldots, L$. Together with $L_\infty$-norm constraints, a sufficiently large degree $L$ ensures the total variation bound in Equation (4.8). 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):

$$\begin{aligned} \mathcal{F}_L(\lambda) \triangleq \sup\ & \mathbb{E}\left[\phi(U)\right] - \mathbb{E}\left[\phi(U')\right] \\ \text{s.t. } & \mathbb{E}[U] = \mathbb{E}[U'] = 1 \\ & \mathbb{E}[U^j] = \mathbb{E}[U'^j], \quad j = 1, \ldots, L, \\ & U, U' \in [0, \lambda] \end{aligned} \tag{4.9}$$

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

Finally, we connect the optimization problem in Equation (4.9) to the machinery of *best polynomial approximation*: We prove that

$$\mathcal{F}_L(\lambda) \geq 2E_L(\log, [1/\lambda, 1]). \tag{4.10}$$

Due to the singularity of the logarithm at zero, the approximation error can be made bounded away from zero if $\lambda$ grows *quadratically* with the degree $L$ (see Section 2.3). 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 Equation (4.9) to $\mathbb{E}[U] = \mathbb{E}[U'] \leq 1$ and a simple scaling argument. We refer to the proofs in Section 4.2.2 for details.

Applying the steps described above, we have the following proposition:

**Proposition 5.** *For all $k, n \in \mathbb{N}$,*

$$R^*(k, n) \gtrsim \left(\frac{k}{n \log k}\right)^2 \vee 1. \tag{4.11}$$

### 4.2.2 Proof of Proposition 5

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

$$\mathcal{M}_k(\nu) \triangleq \left\{ P \in \mathbb{R}_+^k : \left| \sum_{i=1}^k p_i - 1 \right| \leq \nu \right\}, \tag{4.12}$$

which reduces to the probability simplex $\mathcal{M}_k$ if $\nu = 0$.

Generalizing the minimax quadratic risk in Equation (3.1) for Poisson sampling, we define

$$\tilde{R}^*(k,n,\nu) \triangleq \inf_{\hat{H}'} \sup_{P \in \mathcal{M}_k(\nu)} \mathbb{E}(\hat{H}'(N) - H(P))^2, \tag{4.13}$$

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 Equation (1.2) for multinomial sampling by the following lemma, which is an extension of Proposition 2.

**Lemma 3.** *For any $0 \le \nu \le 1$ and any $\alpha > 0$,*

$$R^*(k,n) \ge \tilde{R}^*\left(k, \frac{1+\alpha}{1-\nu} n, \nu\right) - \exp\left(-n(\alpha - \log(1+\alpha))\right) \log^2 k \\ - \nu(2+\nu)(\log k + 1 + \nu) \log^2 k.$$

To establish a lower bound of $\tilde{R}^*(k,n,\nu)$, we apply generalized Le Cam's method involving two composite hypotheses as in Equation (4.5), which entails choosing 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 the next lemma:

**Lemma 4.** *Let $U$ and $U'$ be random variables such that $U, U' \in [0, \lambda]$ where $\lambda < k/e$. Let $\mathbb{E}[U] = \mathbb{E}[U'] \le 1$ and $|\mathbb{E}[\phi(U)] - \mathbb{E}[\phi(U')]| \ge d$. Then, for any $\beta < 1/2$,*

$$\tilde{R}^*(k,n,\nu) \ge \frac{(1-2\beta)^2 d^2}{4} \Bigg(1 - k\mathsf{TV}(\mathbb{E}[\text{Poi}(nU/k)], \mathbb{E}[\text{Poi}(nU'/k)]) \\ - \frac{2\lambda^2}{k\nu^2} - \frac{2\lambda^2 \log^2 \frac{k}{\lambda}}{k\beta^2 d^2}\Bigg).$$

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 [14, Section 4.3] using Taylor expansion of the KL divergence and orthogonal basis expansion of $\chi^2$-divergence in [46, Proof of Theorem 3]. For Poisson mixtures we directly deal with the total variation as the $\ell_1$-distance between the mixture probability mass functions. The following lemma used the dual problem of moment matching, i.e., best polynomial approximation, and the approximation-theoretical properties of the Poisson distribution functions $x \mapsto \frac{e^{-x}x^j}{j!}$. We refer to Section 4.3 for details.

**Lemma 5.** *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*

$$\mathsf{TV}(\mathbb{E}[\mathrm{Poi}(V)], \mathbb{E}[\mathrm{Poi}(V')]) \leq \frac{(\Lambda/2)^{L+1}}{(L+1)!} \left(2 + 2^{\Lambda/2-L} + 2^{\Lambda/(2\log 2)-L}\right). \quad (4.14)$$

*In particular,*

$$\mathsf{TV}(\mathbb{E}[\mathrm{Poi}(V)], \mathbb{E}[\mathrm{Poi}(V')]) \leq \left(\frac{e\Lambda}{2L}\right)^L.$$

*Also, if $L > \frac{e}{2}\Lambda$ then $\mathsf{TV}(\mathbb{E}[\mathrm{Poi}(V)], \mathbb{E}[\mathrm{Poi}(V')]) \leq \frac{2(\Lambda/2)^{L+1}}{(L+1)!}(1 + o(1))$.*

To apply Lemma 4 and Lemma 5 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 4.2.1 and formulated in Equation (4.9). As shown in Section 2.2, we can obtain $U, U'$ with matching moments from the dual of the best polynomial approximation of $\phi$; 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 5. 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 an 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 $L^{\text{th}}$ order, we can construct $U, U'$ from $X, X'$ with the following distributions:

$$\begin{aligned} P_U(\mathrm{d}u) &= \left(1 - \mathbb{E}\left[\frac{\eta}{X}\right]\right)\delta_0(\mathrm{d}u) + \frac{\alpha}{u}P_{\alpha X/\eta}(\mathrm{d}u), \\ P_{U'}(\mathrm{d}u) &= \left(1 - \mathbb{E}\left[\frac{\eta}{X'}\right]\right)\delta_0(\mathrm{d}u) + \frac{\alpha}{u}P_{\alpha X'/\eta}(\mathrm{d}u), \end{aligned} \quad (4.15)$$

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,

**Lemma 6.** *$\mathbb{E}[\phi(U)] - \mathbb{E}[\phi(U')] = \alpha(\mathbb{E}[\log \frac{1}{X}] - \mathbb{E}[\log \frac{1}{X'}])$ and $\mathbb{E}[U^j] = \mathbb{E}[U'^j]$, $j = 1, \ldots, L+1$. In particular, $\mathbb{E}[U] = \mathbb{E}[U'] = \alpha$.*

*Proof of Lemma 6.* Note that

$$\mathbb{E}[\phi(U)] = \int \left(u \log \frac{1}{u}\right) \frac{\alpha}{u} P_{\alpha X/\eta}(\mathrm{d}u) = \alpha \mathbb{E}\left[\log \frac{\eta}{\alpha X}\right]$$

and, analogously, $\mathbb{E}[\phi(U')] = \alpha \mathbb{E}\left[\log \frac{\eta}{\alpha X'}\right]$. Therefore, $\mathbb{E}[\phi(U)] - \mathbb{E}[\phi(U')] = \alpha(\mathbb{E}\left[\log \frac{1}{X}\right] - \mathbb{E}\left[\log \frac{1}{X'}\right])$. Moreover, for any $j \in [L+1]$,

$$\mathbb{E}\left[U^j\right] = \int u^j \frac{\alpha}{u} P_{\alpha X/\eta}(\mathrm{d}u) = \mathbb{E}\left[(\alpha X/\eta)^{j-1}\alpha\right],$$

which coincides with $\mathbb{E}[U'^j] = \mathbb{E}[(\alpha X'/\eta)^{j-1}\alpha]$, in view of the moment matching condition of $X$ and $X'$ in Equation (4.16). In particular, $\mathbb{E}[U] = \mathbb{E}[U'] = \alpha$ follows immediately. □

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

$$\begin{aligned} \mathcal{E}^* = \max\ & \mathbb{E}\left[\log \frac{1}{X}\right] - \mathbb{E}\left[\log \frac{1}{X'}\right] \\ \text{s.t. } & \mathbb{E}\left[X^j\right] = \mathbb{E}\left[X'^j\right], \quad j = 1, \ldots, L, \\ & X, X' \in [\eta, 1], \end{aligned} \tag{4.16}$$

where $0 < \eta < 1$. Note that Equation (4.16) is an infinite-dimensional linear programming problem with finitely many constraints. Therefore it is natural to turn to its dual. In Section 2.2 we show that the maximum $\mathcal{E}^*$ exists and coincides with twice the best $L_\infty$ approximation error of the log over the interval $[\eta, 1]$ by polynomials of degree $L$:

$$\mathcal{E}^* = 2E_L(\log, [\eta, 1]). \tag{4.17}$$

By definition, this approximation error is decreasing in the degree $L$ when $\eta$ is fixed; on the other hand, since the logarithm function blows up near zero, for fixed degree $L$ the approximation error also diverges as $\eta$ vanishes.

As shown in Equation (2.1), 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 $\eta$ decays according to $L^{-2}$.

Now we are ready to prove our main lower bound in Proposition 5.

*Proof of Proposition 5.* Let $X$ and $X'$ be the maximizer of Equation (4.16). Now we construct $U$ and $U'$ from $X$ and $X'$ according to the recipe Equation (4.15). By Lemma 6, the first $L+1$ moments of $U$ and $U'$ are matched with means equal to $\alpha$ which is less than one; moreover,

$$\mathbb{E}\left[\phi(U)\right] - \mathbb{E}\left[\phi(U')\right] = \alpha \mathcal{E}^*. \tag{4.18}$$

Recall the universal constants $c$ and $c'$ defined in Equation (2.1). Let $L = \lfloor c \log k \rfloor \geq \frac{c \log k}{2}$ and $\eta = \log^{-2} k$ and then we have $\mathcal{E}^* \geq 2c'$. Let $\alpha = \frac{c_1 k}{n \log k}$ and $\lambda = \alpha \eta^{-1} = \frac{c_1 k \log k}{n}$. Using Equation (4.15) and Equation (4.18), 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 $\mathbb{E}\left[\phi(U)\right] - \mathbb{E}\left[\phi(U')\right] = \alpha \mathcal{E}^* \geq 2c'\alpha$. Picking $c_1$ satisfying $c_1 < c/e$ and $\frac{c}{2} \log \frac{c}{ec_1} > 2$, then by Lemma 5 we have $\mathsf{TV}(\mathbb{E}\left[\text{Poi}\left(nU/k\right)\right], \mathbb{E}\left[\text{Poi}\left(nU'/k\right)\right]) \leq 2k^{-2}$. Applying Lemma 4 with $d = 2c'\alpha$ and $\beta = 1/4, \nu = 4\lambda/\sqrt{k}$ we conclude that $\tilde{R}^*(k, n, \frac{4\lambda}{\sqrt{k}}) \gtrsim \alpha^2 \asymp (\frac{k}{n \log k})^2$. Finally applying Lemma 3 yields that $R^*(k,n) \gtrsim (\frac{k}{n \log k})^2$ when $n \geq \frac{k}{\log k}$. For $n \leq \frac{k}{\log k}$ by monotonicity, $R^*(k,n) \geq R^*(k, \frac{k}{\log k}) \gtrsim 1$.

□

**Remark 4** (Structure of the least favorable priors)**.** From the proof of Equation (4.17) in Section 2.2, we conclude that $X, X'$ are in fact discrete random variables with disjoint support each of which has $L+2 \asymp \log k$ atoms. 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 $\mathsf{P}$ and $\mathsf{P}'$ are drawn from the distribution $U$ and $U'$ respectively, which lie in the interval $[0, \frac{\log k}{n}]$. Therefore, although $\mathsf{P}$ and $\mathsf{P}'$ are distributions over $k$ elements, they only have $\log k$ distinct masses and the locations are randomly permuted. Moreover, the entropy of $\mathsf{P}$ and $\mathsf{P}'$ constructed based on $U$ and $U'$ (see Equation (4.23)) are concentrated near the respective mean values, both of which are close to $\log k$ but differ by a constant factor of $\frac{k}{n \log k}$.

### 4.2.3 Proof of lemmas

*Proof of Lemma 3.* This is an extension of the lower bound of $R^*(k,n)$ in Proposition 2 where $\tilde{R}^*(k,n) = \tilde{R}^*(k,n,0)$.

Fix an arbitrary vector $P = (p_1, \ldots, p_k) \in \mathcal{M}_k(\nu)$. Let $N = (N_1, N_2, \ldots) \overset{\text{ind}}{\sim} \text{Poi}(\frac{n(1+\alpha)}{1-v} p_i)$ and let $n' = \sum N_i \sim \text{Poi}(\frac{n(1+\alpha)}{1-v} \sum p_i) \geq_{\text{s.t.}} \text{Poi}((1+\alpha)n)$. Let $\hat{H}_n(\cdot)$ be the optimal estimator of Shannon entropy for fixed sample size $n$, i.e.,

$$\mathbb{E}(\hat{H}_n(N) - H(P))^2 \leq R^*(k,n), \quad \forall\, P \in \mathcal{M}_k.$$

We construct an estimator for the Poisson sampling model by $\tilde{H}(N) = \hat{H}_{n'}(N)$. We observe that conditioned on $n' = m$, $N \sim \text{Multinomial}(m, P')$, where $P' = \frac{P}{\sum p_i}$ is the normalized $P$.

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

$$H(P') = \log\left(\sum p_i\right) + \frac{H(P)}{\sum p_i},$$

which is differed at most by

$$\begin{aligned} |H(P) - H(P')| &\leq \left|\left(\sum p_i - 1\right) H(P')\right| + \left|\left(\sum p_i\right) \log\left(\sum p_i\right)\right| \\ &\leq \nu \log k + (1+\nu)\log(1+\nu) \leq \nu(\log k + 1 + \nu). \end{aligned} \tag{4.19}$$

Since $\mathbb{E}(\hat{H}_n(N) - H(P'))^2 \leq R^*(k,n) \leq \log^2 k$ then

$$|\mathbb{E}(\hat{H}_n(N) - H(P'))| \leq \log k. \tag{4.20}$$

Therefore, by Equation (4.19) and Equation (4.20),

$$\begin{aligned} &\mathbb{E}(\tilde{H}(N) - H(P))^2 \\ &= \sum_{m=0}^{\infty} \mathbb{E}\left[(\hat{H}_{n'}(N) - H(P') + H(P') - H(P))^2 | n' = m\right] \mathbb{P}[n' = m] \\ &\leq \sum_{m=0}^{\infty} R^*(k,m) \mathbb{P}\left[n' = m\right] + \nu^2(\log k + 1 + \nu)^2 + 2\nu(\log k + 1 + \nu)\log k. \end{aligned}$$

Then

$$\tilde{R}^*\left(k, \frac{1+\alpha}{1-\nu}n\right) \le \sum_{m=0}^{\infty} R^*(k,m)\mathbb{P}\left[n' = m\right] + \nu(2+\nu)(\log k + 1 + \nu)^2. \tag{4.21}$$

Note that for fixed $k$, the minimax risk $n \mapsto R^*(k,n)$ is decreasing and $0 \le R^*(k,n) \le \log^2 k$. Then,

$$\begin{aligned}\sum_{m=0}^{\infty} R^*(k,m)\mathbb{P}\left[n' = m\right] &\le \sum_{m \ge n} R^*(k,m)\mathbb{P}[n' = m] + \log^2 k\mathbb{P}[n' < n] \\ &\le R^*(k,n) + \log^2 k\mathbb{P}[\mathrm{Poi}((1+\alpha)n) < n] \\ &\le R^*(k,n) + \exp(-n(\alpha - \log(1+\alpha)))\log^2 k,\end{aligned} \tag{4.22}$$

where in the last inequality we used the Chernoff bound (see, e.g., [56, Theorem 5.4]). Combining Equation (4.21) and Equation (4.22), the conclusion follows. □

*Proof of Lemma 4.* Let $\alpha$ denote the common mean of $U$ and $U'$, which is less than one. Define two random vectors

$$\mathsf{P} = \left(\frac{U_1}{k}, \ldots, \frac{U_k}{k}, 1-\alpha\right), \quad \mathsf{P}' = \left(\frac{U_1'}{k}, \ldots, \frac{U_k'}{k}, 1-\alpha\right), \tag{4.23}$$

where $U_i$ and $U_i'$ are i.i.d. copies of $U$ and $U'$, respectively. Conditioned on $\mathsf{P}$ and $\mathsf{P}'$ 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 concentration events, for $\beta < 1/2$,

$$\begin{aligned}E &\triangleq \left\{\left|\frac{\sum_i U_i}{k} - \alpha\right| \le \nu, |H(\mathsf{P}) - \mathbb{E}\left[H(\mathsf{P})\right]| \le \beta d\right\}, \\ E' &\triangleq \left\{\left|\frac{\sum_i U_i'}{k} - \alpha\right| \le \nu, |H(\mathsf{P}') - \mathbb{E}\left[H(\mathsf{P}')\right]| \le \beta d\right\}.\end{aligned}$$

Now we define two priors on the set $\mathcal{M}_k(\nu)$ 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 $\pi$ and $\pi'$. 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, $\mathbb{E}\left[H(\mathsf{P}')\right] = \mathbb{E}\left[\phi(U')\right] + \mathbb{E}\left[U'\right]\log k + \phi(1-\alpha)$. Therefore,

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

By the definition of the events $E, E'$ and the triangle inequality, we obtain that under $\pi$ and $\pi'$, both $\mathsf{P}, \mathsf{P}' \in \mathcal{M}_k(\nu)$ and

$$|H(\mathsf{P}) - H(\mathsf{P}')| \geq (1-2\beta)d. \tag{4.24}$$

Now we consider the total variation distance of the distributions of the histogram under the priors $\pi$ and $\pi'$. By the triangle inequality and the fact that total variation of product distribution can be upper bounded by the summation of individual one,

$$\begin{aligned}\mathsf{TV}(P_{N|E}, P_{N'|E'}) &\leq \mathsf{TV}(P_{N|E}, P_N) + \mathsf{TV}(P_N, P_{N'}) + \mathsf{TV}(P_{N'}, P_{N'|E'}) \\ &= \mathbb{P}[E^c] + \mathsf{TV}\left((\mathbb{E}[\mathrm{Poi}(nU/k)])^{\otimes k}, (\mathbb{E}[\mathrm{Poi}(nU'/k)])^{\otimes k}\right) + \mathbb{P}[E'^c] \\ &\leq \mathbb{P}[E^c] + \mathbb{P}[E'^c] + k\mathsf{TV}(\mathbb{E}[\mathrm{Poi}(nU/k)], \mathbb{E}[\mathrm{Poi}(nU'/k)]).\end{aligned} \tag{4.25}$$

By the Chebyshev's inequality and the union bound, both

$$\begin{aligned}\mathbb{P}[E^c], \mathbb{P}[E'^c] &\leq \mathbb{P}\left[\left|\sum_i \frac{U_i}{k} - \alpha\right| > \nu\right] + \mathbb{P}\left[\left|H(\mathsf{P}) - \mathbb{E}\left[H(\mathsf{P})\right]\right| > \beta d\right] \\ &\leq \frac{\sum_i \mathsf{var}[U_i]}{(k\nu)^2} + \frac{\sum_i \mathsf{var}[\phi(U_i/k)]}{(\beta d)^2} \leq \frac{\lambda^2}{k\nu^2} + \frac{k\phi^2(\lambda/k)}{\beta^2 d^2},\end{aligned} \tag{4.26}$$

where the last inequality follows from the fact that $\mathsf{var}[\phi(U/k)] \leq \mathbb{E}(\phi(U/k))^2 \leq \phi^2(\lambda/k)$ when $\lambda/k < e^{-1}$ by assumption.

Plugging Equation (4.26) into Equation (4.25), we obtain that

$$\mathsf{TV}(P_{N|E}, P_{N'|E'}) \leq \frac{2\lambda^2}{k\nu^2} + \frac{2\lambda^2\log^2(k/\lambda)}{k\beta^2 d^2} + k\mathsf{TV}(\mathbb{E}[\mathrm{Poi}(nU/k)], \mathbb{E}[\mathrm{Poi}(nU'/k)]). \tag{4.27}$$

Applying Le Cam's lemma [44], the conclusion follows from Equation (4.24) and Equation (4.27). □

## 4.3 Total variation distance between Poisson mixtures

In this section we prove Lemma 5, which provides a sufficient condition for the indistinguishability between two Poisson mixtures. The proof again relates the problem of moment matching to best polynomial approximation, and then applies Chebyshev polynomial approximation to obtain an achievable approximation error.

*Proof of Lemma 5.* Let

$$f_j(x) \triangleq \frac{e^{-x}x^j}{j!} \tag{4.28}$$

and $\mathcal{S}_L = \{(V, V') \in [0, \Lambda]^2 : \mathbb{E}[V^i] = \mathbb{E}[V'^i], i = 1, \ldots, L\}$. Then

$$\mathsf{TV}(\mathbb{E}[\mathrm{Poi}(V)], \mathbb{E}[\mathrm{Poi}(V')]) = \frac{1}{2}\sum_{j=0}^{\infty} |\mathbb{E}f_j(V) - \mathbb{E}f_j(V')|$$

$$\leq \frac{1}{2}\sum_{j=0}^{\infty} \sup_{(V,V')\in\mathcal{S}_L} |\mathbb{E}f_j(V) - \mathbb{E}f_j(V')| = \sum_{j=0}^{\infty} E_L(f_j, [0, \Lambda]) \tag{4.29}$$

in view of the relation of moment matching and best polynomial approximation in Section 2.2.

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 its $L^{\text{th}}$ derivative. Specifically, we have (cf. e.g., [60, Lecture 20])

$$\begin{aligned} E_L(f, [0, \Lambda]) &\leq \max_{x\in[0,\Lambda]} |f_j(x) - Q_L(f; x)| \\ &\leq \frac{1}{2^L(L+1)!}\left(\frac{\Lambda}{2}\right)^{L+1} \max_{x\in[0,\Lambda]} \left|f^{(L+1)}(x)\right|, \end{aligned} \tag{4.30}$$

where $Q_L(f; x)$ denotes the degree-$L$ interpolating polynomial for $f$ on Chebyshev nodes (roots of Chebyshev polynomial). To apply Equation (4.30) to $f = f_j$ defined in Equation (4.28), note that $f_j^{(L+1)}(x)$ can be conveniently expressed in terms of Laguerre polynomials: Denote the degree-$n$ generalized Laguerre polynomial by $L_n^{(k)}(x)$ and the simple Laguerre polynomial by $L_n(x) = L_n^{(0)}(x)$. The Rodrigues representation is

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

If $j \leq L+1$,

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

Note that $L_j$ is a degree-$j$ polynomial, whose derivative of order higher than $j$ is zero. Applying general Leibniz rule for derivatives yields that

$$\begin{aligned} f_j^{(L+1)}(x) &= \sum_{m=0}^{(L+1-j)\wedge j} \binom{L+1-j}{m} \left( \frac{\mathrm{d}^m L_j(x)}{\mathrm{d}x^m} \right) e^{-x}(-1)^{L+1-j-m} \\ &= (-1)^{L+1-j} e^{-x} \sum_{m=0}^{(L+1-j)\wedge j} \binom{L+1-j}{m} L_{j-m}^{(m)}(x). \end{aligned} \tag{4.31}$$

Applying $|L_n^{(k)}(x)| \leq \binom{n+k}{n} e^{x/2}$ [61, 22.14.13] when $x \geq 0$ and $k \in \mathbb{N}$, we obtain that

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

Therefore $\max_{x\in[0,\Lambda]} |f_j^{(L+1)}(x)| \leq \binom{L+1}{j}$. Observing from Equation (4.31) that $|f_j^{(L+1)}(0)| = \sum_m \binom{L+1-j}{m}\binom{j}{j-m} = \binom{L+1}{j}$, we conclude that

$$\max_{x\in[0,\Lambda]} |f_j^{(L+1)}(x)| = \binom{L+1}{j}, \quad j \leq L+1.$$

Then, applying Equation (4.30),

$$\sum_{j=0}^{L+1} E_L(f_j, [0,\Lambda]) \leq \sum_{j=0}^{L+1} \frac{\binom{L+1}{j}(\Lambda/2)^{L+1}}{2^L(L+1)!} = \frac{2(\Lambda/2)^{L+1}}{(L+1)!}. \tag{4.32}$$

If $j \geq L+2$, the derivatives of $f_j$ is connected to Laguerre polynomial by

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

Again applying $|L_n^{(k)}(x)| \leq \binom{n+k}{n} e^{x/2}$ [61, 22.14.13] when $x \geq 0$ and $k \in \mathbb{N}$, we obtain that

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

where the maximum of right-hand side occurs at $x = (2(j - L - 1)) \wedge \Lambda$. Therefore we obtain an upper bound of $\max_{x\in[0,\Lambda]} |f_j(x)|$ that

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

Then, applying Equation (4.30) and Stirling's approximation that $(\frac{j-L-1}{e})^{j-L-1} < \frac{(j-L-1)!}{\sqrt{2\pi(j-L-1)}}$,

$$\sum_{\substack{j\ge L+2 \\ j<L+1+\Lambda/2}} E_L(f_j,[0,\Lambda]) \le \frac{(\Lambda/2)^{L+1}}{2^L(L+1)!} \sum_{\substack{j\ge L+2 \\ j<L+1+\Lambda/2}} \frac{2^{j-L-1}}{\sqrt{2\pi(j-L-1)}} \le \frac{(\Lambda/2)^{L+1}2^{\Lambda/2}}{2^L(L+1)!}, \tag{4.33}$$

$$\sum_{j\ge L+1+\Lambda/2} E_L(f_j,[0,\Lambda]) \le \frac{(\Lambda/2)^{L+1}e^{-\Lambda/2}}{2^L(L+1)!} \sum_{j\ge L+1+\Lambda/2} \frac{\Lambda^{j-L-1}}{(j-L-1)!} \le \frac{(\Lambda/2)^{L+1}e^{\Lambda/2}}{2^L(L+1)!}. \tag{4.34}$$

Assembling three ranges Equation (4.32) – Equation (4.34) in the total variation bound Equation (4.29), we obtain that

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

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

**Remark 5.** Recall that in the analysis we conclude that $\max_{x\in[0,\Lambda]} |f_j^{(L+1)}(x)| = \binom{L+1}{j}$ when $j \le L+1$, then

$$\frac{(\Lambda/2)^{L+1}}{2^L(L+1)!}\sum_{j=0}^{L+1} \max_{x\in[0,\Lambda]}\left|f_j^{(L+1)}(x)\right| = \frac{2(\Lambda/2)^{L+1}}{(L+1)!}.$$

This is the best possible upper bound if we use Equation (4.30) to upper bound the uniform approximation error $E_L(f_j,[0,\Lambda])$ when $j \le L+1$.

# CHAPTER 5

# OPTIMAL ESTIMATOR VIA BEST POLYNOMIAL APPROXIMATION

In this chapter we prove the achievability of Theorem 1. We first prove the worst-case MSE of plug-in estimator and relate it to the Bernstein polynomial approximation error. Then the estimator based on the best polynomial approximation is proposed and analyzed.

## 5.1 Plug-in estimator and Bernstein polynomial approximation

To estimate a functional the most natural idea is the plug-in approach, i.e., the empirical entropy. It is known that the empirical entropy is always underbiased. Using $n$ i.i.d. samples, the bias is

$$\begin{aligned}\sum_{i=1}^{k}(\phi(p_i) - \mathbb{E}[\phi(N_i/n)]) &= \sum_{i=1}^{k}\left(\phi(p_i) - \sum_{j=0}^{n}\phi(j/n)\binom{n}{j}p_i^j(1-p_i)^{n-j}\right)\\ &= \sum_{i=1}^{k}\left(\phi(p_i) - B_n(p_i)\right), \qquad (5.1)\end{aligned}$$

where $B_n$ is the degree-$n$ Bernstein polynomial to approximate the function $\phi$ given by the following formula:

$$B_n(x) \triangleq \sum_{j=0}^{n}\binom{n}{j}x^j(1-x)^{n-j}f(j/n).$$

Bernstein polynomial approximation error converges to zero uniformly and hence the bias vanishes as $n \to \infty$. However, in the focus of this thesis, the sublinear regime, sample size does not necessarily far exceed the alphabet size. In this case, given a degree, the Bernstein polynomial is often far from the optimal polynomial. Figure 5.1 shows the degree-6 polynomial approximation of the entropy function $\phi$ using Bernstein polynomial versus the best polynomial.

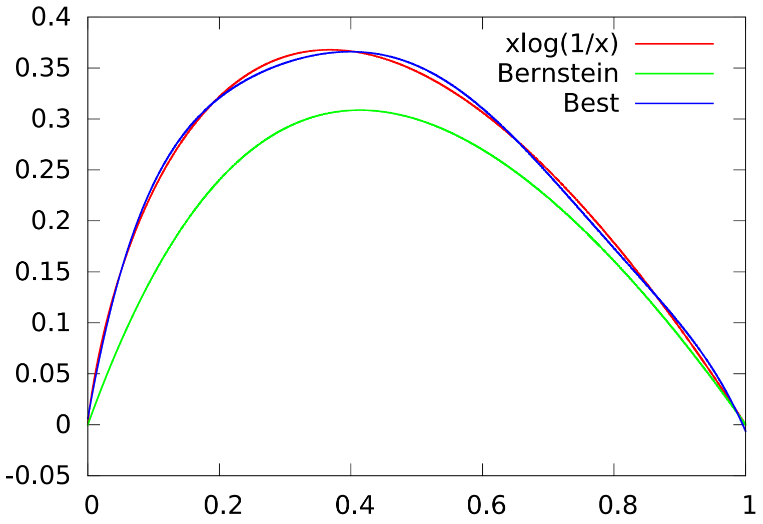

Figure 5.1: Bernstein polynomial and best polynomial to approximate the function $x \mapsto x \log \frac{1}{x}$.

Indeed, as in the statement of Proposition 1 the risk of empirical entropy is: If $n \gtrsim k$, then

$$R_{\text{plug-in}}(k,n) \asymp \left(\frac{k}{n}\right)^2 + \frac{\log^2 k}{n}. \tag{5.2}$$

If $n \lesssim k$, then $\hat{H}_{\text{plug-in}}$ is inconsistent, i.e., $R_{\text{plug-in}}(k,n) \gtrsim 1$.

*Proof of Proposition 1.* Recall the worst-case quadratic risk of the plug-in estimator $R_{\text{plug-in}}(k,n)$ defined in Equation (1.5). We show that for any $k \geq 2$ and $n \geq 2$,

$$\left(\frac{k}{n} \wedge 1\right)^2 + \frac{\log^2 k}{n} \lesssim R_{\text{plug-in}}(k,n) \lesssim \left(\frac{k}{n}\right)^2 + \frac{\log^2 (k \wedge n)}{n}. \tag{5.3}$$

The second term of the lower bound follows from the minimax lower bound Proposition 4 which applies to all $k$ and $n$. To prove the first term of the lower bound, we take $P$ as uniform distribution. We consider its bias here since squared bias is a lower bound for MSE. We denote the empirical distribution as $\hat{P} = \frac{N}{n}$. Applying Pinsker's inequality and Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}\mathbb{E}(\hat{H}_{\text{plug-in}}(N) - H) &= -\mathbb{E}[D(\hat{P}||P)] \le -2\mathbb{E}[(\mathsf{TV}(\hat{P}, P))^2] \\ &\le -2(\mathbb{E}[\mathsf{TV}(\hat{P}, P)])^2 = -2\left(\frac{k}{2n}\mathbb{E}\left|N_1 - \frac{n}{k}\right|\right)^2,\end{aligned}$$

where $N_1 \sim \text{Binomial}\left(n, \frac{1}{k}\right)$. From [62, Theorem 1], we know that $\mathbb{E}\left|N_1 - \frac{n}{k}\right| = \frac{2n}{k}\left(1 - \frac{1}{k}\right)^n$ when $n < k$ and $\mathbb{E}\left|N_1 - \frac{n}{k}\right| \ge \sqrt{\frac{n}{2k}\left(1 - \frac{1}{k}\right)}$ when $n \ge k$. Therefore,

$$\begin{aligned}-\mathbb{E}(\hat{H}_{\text{plug-in}}(N) - H) &\ge 2\left(1 - \frac{1}{k}\right)^{2n} \gtrsim 1, \quad n < k, \\ -\mathbb{E}(\hat{H}_{\text{plug-in}}(N) - H) &\ge \frac{k}{4n}\left(1 - \frac{1}{k}\right) \gtrsim \frac{k}{n}, \quad n \ge k.\end{aligned}$$

Consequently,

$$\mathbb{E}[(\hat{H}_{\text{plug-in}}(N) - H)^2] \ge [\mathbb{E}(\hat{H}_{\text{plug-in}}(N) - H)]^2 \gtrsim \left(\frac{k}{n} \wedge 1\right)^2.$$

The upper bound of MSE follows from the upper bounds of bias and variance. The squared bias can be upper bounded by $(\frac{k-1}{n})^2$ according to [30, Proposition 1]. For the variance we apply Steele's inequality [63]:

$$\mathsf{var}[\hat{H}_{\text{plug-in}}] \le \frac{n}{2}\mathbb{E}(\hat{H}_{\text{plug-in}}(N) - \hat{H}_{\text{plug-in}}(N'))^2, \tag{5.4}$$

where $N'$ is the histogram of $(X_1, \ldots, X_{n-1}, X_n')$ and $X_n'$ is an independent copy of $X_n$. Let $\tilde{N} = (\tilde{N}_1, \ldots, \tilde{N}_k)$ be the histogram of $X_1^{n-1}$, then $\tilde{N} \sim \text{Multinomial}(n-1, P)$ independently of $X_n, X_n'$. Hence, applying triangle inequality,

$$
\begin{aligned}
&\mathbb{E}(\hat{H}_{\text{plug-in}}(N) - \hat{H}_{\text{plug-in}}(N'))^2 \\
&= \mathbb{E}\left(\phi\left(\frac{\tilde{N}_{X_n}+1}{n}\right) - \phi\left(\frac{\tilde{N}_{X_n}}{n}\right) + \phi\left(\frac{\tilde{N}_{X_n'}}{n}\right) - \phi\left(\frac{\tilde{N}_{X_n'}+1}{n}\right)\right)^2 \\
&\le 4\sum_{j=1}^{k} \mathbb{E}\left(\phi\left(\frac{\tilde{N}_j+1}{n}\right) - \phi\left(\frac{\tilde{N}_j}{n}\right)\right)^2 p_j \\
&= \frac{4}{n^2}\sum_{j=1}^{k} \mathbb{E}\left[\left(\tilde{N}_j \log(1+\tilde{N}_j^{-1}) + \log\frac{\tilde{N}_j+1}{n}\right)^2\right] p_j \\
&\le \frac{8}{n^2} + \frac{8}{n^2}\sum_{j=1}^{k} \mathbb{E}\left[\log^2\frac{\tilde{N}_j+1}{n}\right] p_j, \qquad (5.5)
\end{aligned}
$$

where the last step follows from $0 \le x\log(1+x^{-1}) \le 1$ for all $x > 0$.

Now we rewrite and upper bound the last expectation:

$$
\begin{aligned}
&\mathbb{E}\left[\log^2\frac{\tilde{N}_j+1}{n}\right] \\
&= \mathbb{E}\left[\log^2\frac{n}{\tilde{N}_j+1}\mathbf{1}_{\left\{\tilde{N}_j \le \frac{(n-1)p_j}{2}\right\}}\right] + \mathbb{E}\left[\log^2\frac{n}{\tilde{N}_j+1}\mathbf{1}_{\left\{\tilde{N}_j > \frac{(n-1)p_j}{2}\right\}}\right] \\
&\le (\log^2 n)\,\mathbb{P}\left[\tilde{N}_j \le \frac{(n-1)p}{2}\right] + \log^2\frac{2n}{(n-1)p_j}. \qquad (5.6)
\end{aligned}
$$

Applying Chernoff bound for Binomial tail [56, Theorem 4.5] and plugging into Equation (5.5) then Equation (5.4), we obtain

$$
\begin{aligned}
\mathsf{var}\hat{H}_{\text{plug-in}} &\lesssim \frac{1}{n} + \frac{1}{n}\sum_{j=1}^{k} p_j(\log^2 p_j + \log^2 n \exp(-(n-1)p_j/8)) \\
&\lesssim \frac{\log^2 k}{n} + \frac{\log^2 n}{n}\frac{k}{n} = \frac{\log^2 k}{n}\left(1 + \frac{k\log^2 n}{n\log^2 k}\right),
\end{aligned}
$$

where we have used $\sum_{i=1}^{k} p_i \log^2 p_i \lesssim \log^2 k$ and $\sup_{x>0} x\exp(-(n-1)x/8) = \frac{8}{(n-1)e}$. We know that $\frac{k\log^2 n}{n\log^2 k} \lesssim 1$ when $n \ge k$ and thus $\mathsf{var}\hat{H}_{\text{plug-in}} \lesssim \frac{\log^2 k}{n}$. From [31, Remark (iv), p. 168] we also know that $\mathsf{var}\hat{H}_{\text{plug-in}}(N) \lesssim \frac{\log^2 n}{n}$ for all $n$ and consequently $\mathsf{var}\hat{H}_{\text{plug-in}}(N) \lesssim \frac{\log^2(k\wedge n)}{n}$. □

## 5.2 Unbiased estimator for the best polynomial

As observed in various previous results as well as suggested by the minimax lower bound in Chapter 4, 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, e.g., [30, Proposition 8]). This observation inspired us to approximate $\phi$ 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 in Proposition 3 between the risks with fixed and Poisson sample size, we shall assume the Poisson sampling model to simplify the analysis, where we first draw $n' \sim \text{Poi}(2n)$ and then draw $n'$ i.i.d. samples $X = (X_1, \dots, X_{n'})$ 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 $B_i \overset{\text{i.i.d.}}{\sim} \text{Bern}\left(\frac{1}{2}\right)$. 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 $N = (N_1, \dots, N_k)$ and $N' = (N'_1, \dots, N'_k)$ 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 $N_i, N'_i \overset{\text{i.i.d.}}{\sim} \text{Poi}(np_i)$.

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

$$p_L(x) = \sum_{m=0}^{L} a_m x^m. \tag{5.7}$$

Through a change of variables, we see that the best polynomial of degree $L$ to approximate $x \log \frac{1}{x}$ on $[0, \frac{c_1 \log k}{n}]$ is

$$P_L(x) \triangleq \sum_{m=0}^{L} \frac{a_m n^{m-1}}{(c_1 \log k)^{m-1}} x^m + \left( \log \frac{n}{c_1 \log k} \right) x.$$

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 $L$,

$$g_L(N_i) \triangleq \frac{1}{n} \left( \sum_{m=0}^{L} \frac{a_m}{(c_1 \log k)^{m-1}} (N_i)_m + \left( \log \frac{n}{c_1 \log k} \right) N_i \right), \tag{5.8}$$

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

$$\tilde{H} \triangleq \sum_{i=1}^{k} \left( g_L(N_i) \mathbf{1}_{\left\{N_i' \le c_2 \log k\right\}} + \left( \phi\left(\frac{N_i}{n}\right) + \frac{1}{2n} \right) \mathbf{1}_{\left\{N_i' > c_2 \log k\right\}} \right), \tag{5.9}$$

where we apply the estimator from polynomial approximation if $N_i' \le c_2 \log k$ or the bias-corrected plug-in estimator otherwise (cf. the asymptotic expansion Equation (1.7) 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:

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

Since Equation (5.9) can be expressed in terms of a linear combination of the fingerprints Equation (1.9) of the second sample and the coefficients can be pre-computed using fast best polynomial approximation algorithms (e.g., the Remez algorithm), it is clear that the estimator $\hat{H}$ can be computed in linear time in $n$.

The next result gives an upper bound on the above estimator under the Poisson sampling model, which, in view of the inequality in Proposition 3 and Proposition 1, implies the upper bound on the minimax risk $R^*(n, k)$ in Theorem 1.

**Proposition 6.** *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* $N = (N_1, \ldots, N_k) \overset{ind}{\sim} \mathrm{Poi}(np_i)$.

*Proof of Proposition 6.* Given that $N_i'$ is above (resp. below) the threshold $c_2 \log k$, we can conclude with high confidence that $p_i$ is above (resp. below) a constant factor of $\frac{\log k}{n}$. Define two events by

$$E_1 \triangleq \bigcap_{i=1}^{k} \left\{ N_i' \le c_2 \log k \Rightarrow p_i \le \frac{c_1 \log k}{n} \right\},$$
$$E_2 \triangleq \bigcap_{i=1}^{k} \left\{ N_i' > c_2 \log k \Rightarrow p_i > \frac{c_3 \log k}{n} \right\},$$

where $c_1 > c_2 > c_3$. Applying the union bound and the Chernoff bound for Poissons ([56, Theorem 5.4]) yields that

$$\begin{aligned}\mathbb{P}\left[E_1^c\right] &= \mathbb{P}\left[\bigcup_{i=1}^{k} \left\{ N_i' \le c_2 \log k, p_i > \frac{c_1 \log k}{n} \right\}\right] \\ &\le k \mathbb{P}\left[\mathrm{Poi}(c_1 \log k) \le c_2 \log k\right] \le \frac{1}{k^{c_1 - c_2 \log \frac{ec_1}{c_2} - 1}}.\end{aligned}$$

Define an event $E \triangleq E_1 \cap E_2$ and then by union bound

$$\mathbb{P}\left[E^c\right] \le \mathbb{P}\left[E_1^c\right] + \mathbb{P}\left[E_2^c\right] \le \frac{1}{k^{c_1 - c_2 \log \frac{ec_1}{c_2} - 1}} + \frac{1}{k^{c_3 + c_2 \log \frac{ec_2}{c_3} - 1}}. \tag{5.10}$$

By construction $\hat{H} = (\tilde{H} \vee 0) \wedge \log k$, the fact $H(P) \in [0, \log k]$ yields that $|H(P) - \hat{H}| \le |H(P) - \tilde{H}|$ and $|H(P) - \hat{H}| \le \log k$. So the MSE can be decomposed and upper bounded by

$$\begin{aligned}\mathbb{E}(H(P) - \hat{H})^2 &= \mathbb{E}[(H(P) - \hat{H})^2 \mathbf{1}_E] + \mathbb{E}[(H(P) - \hat{H})^2 \mathbf{1}_{E^c}] \\ &\le \mathbb{E}[(H(P) - \tilde{H})^2 \mathbf{1}_E] + (\log k)^2 \mathbb{P}[E^c]. \end{aligned}\tag{5.11}$$

Define

$$\mathcal{E}_1 \triangleq \sum_{i \in I_1} \phi(p_i) - g_L(N_i), \quad \mathcal{E}_2 \triangleq \sum_{i \in I_2} \left(\phi(p_i) - \phi\left(\frac{N_i}{n}\right) - \frac{1}{2n}\right),$$

where the (random) index sets defined by

$$I_1 \triangleq \left\{i : N'_i \le c_2 \log k, p_i \le \frac{c_1 \log k}{n}\right\}, \quad I_2 \triangleq \left\{i : N'_i > c_2 \log k, p_i > \frac{c_3 \log k}{n}\right\}$$

are independent of $N$ due to the independence of $N$ and $N'$. The implications in the event $E$ yield that

$$(H(P) - \tilde{H})\mathbf{1}_E = \mathcal{E}_1 \mathbf{1}_E + \mathcal{E}_2 \mathbf{1}_E. \tag{5.12}$$

Combining Equations (5.11) and (5.12) and applying triangle inequality we obtain that

$$\mathbb{E}(H(P) - \hat{H})^2 \le 2\mathbb{E}[\mathcal{E}_1^2] + 2\mathbb{E}[\mathcal{E}_2^2] + (\log k)^2 \mathbb{P}[E^c]. \tag{5.13}$$

Next we proceed to consider the error terms $\mathcal{E}_1$ and $\mathcal{E}_2$ separately.

**Case 1: Polynomial estimator** It is known that (see, e.g., [55, Section 7.5.4]) the optimal uniform approximation error of $\phi$ by degree-$L$ polynomials on $[0, 1]$ satisfies $L^2 E_L(\phi, [0, 1]) \to c > 0$ as $L \to \infty$. Therefore $E_L(\phi, [0, 1]) \lesssim L^{-2}$. By a change of variables, it is easy to show that

$$E_L\left(\phi, \left[0, \frac{c_1 \log k}{n}\right]\right) = \frac{c_1 \log k}{n} E_L(\phi, [0, 1]) \lesssim \frac{1}{n \log k}.$$

By definition, $I_1 \subseteq \{i : p_i \le \frac{c_1 \log k}{n}\}$. Since $g_L(N_i)$ is an unbiased estimator of $P_L(p_i)$, the bias can be bounded by the uniform approximation error almost surely as

$$|\mathbb{E}[\mathcal{E}_1 | I_1]| = \left| \sum_{i \in I_1} p_i \log \frac{1}{p_i} - P_L(p_i) \right| \le k E_L\left(\phi, \left[0, \frac{c_1 \log k}{n}\right]\right) \lesssim \frac{k}{n \log k}. \tag{5.14}$$

Next we consider the conditional variance of $\mathcal{E}_1$. In view of the fact that the standard deviation of sum of random variables is at most the sum of individual standard deviations, we obtain that

$$
\begin{aligned}
&\mathsf{var}\left[\mathcal{E}_1|I_1\right] = \sum_{i\in I_1} \mathsf{var}[g_L(N_i)] \\
&= \sum_{i\in I_1} \mathsf{var}\left[\sum_{m\neq 1} \frac{a_m}{(c_1\log k)^{m-1}} \frac{(N_i)_m}{n} + \left(a_1 + \log\frac{n}{c_1\log k}\right)\frac{N_i}{n}\right] \\
&\le \frac{1}{n^2} \sum_{i:p_i\le\frac{c_1\log k}{n}} \left(\sum_{m\neq 1} \frac{|a_m|\sqrt{\mathsf{var}(N_i)_m}}{(c_1\log k)^{m-1}} + \left|a_1+\log\frac{n}{c_1\log k}\right|\sqrt{\mathsf{var}(N_i)}\right)^2.
\end{aligned}
$$

Since $0 \le \phi(x) \le e^{-1}$ on $[0,1]$ then $\sup_{0\le x\le 1}|p_L(x) - \phi(x)| = E_L(\phi,[0,1]) \le e^{-1}$. Therefore $\sup_{0\le x\le 1}|p_L(x)| \le 2e^{-1}$. From the proof of [46, Lemma 2, p. 1035] we know that the polynomial coefficients can by upper bounded by $|a_m| \le 2e^{-1}2^{3L}$. Since $\log n \le C\log k$, we have $\left|a_1 + \log\frac{n}{c_1\log k}\right| \lesssim 2^{3L}$. Therefore all polynomial coefficients can be upper bounded by a constant factor of $2^{3L}$. We also need the following lemma to upper bound the variance of $(N_i)_m$:

**Lemma 7.** *Suppose* $X \sim \mathrm{Poi}(\lambda)$ *and* $(x)_m = \frac{x!}{(x-m)!}$. *Then* $\mathsf{var}(X)_m$ *is increasing in* $\lambda$ *and*

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

*Proof of Lemma 7.* First we compute $\mathbb{E}(X)_m^2$:

$$
\begin{aligned}
\mathbb{E}(X)_m^2 &= \sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!}\frac{x!^2}{(x-m)!^2} = \sum_{j=0}^{\infty}\frac{e^{-\lambda}\lambda^{j+m}}{j!}\frac{(j+m)!}{j!} = \lambda^m m!\mathbb{E}\binom{X+m}{X} \\
&= \lambda^m m!\mathbb{E}\left[\sum_{k=0}^{m}\binom{m}{k}\binom{X}{X-k}\right] = \lambda^m m!\sum_{k=0}^{m}\binom{m}{k}\frac{\mathbb{E}(X)_k}{k!} \\
&= \lambda^m m!\sum_{k=0}^{m}\binom{m}{k}\frac{\lambda^k}{k!},
\end{aligned}
\tag{5.15}
$$

where we have used $\mathbb{E}(X)_k = \lambda^k$. Therefore the variance of $(X)_m$ is

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

The monotonicity of $\lambda \mapsto \mathsf{var}(X)_m$ follows from the equality part immediately. Since the maximal term in the summation is attained at $k^* = \lfloor \sqrt{\lambda m} \rfloor$, we have

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

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

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

**Remark 6.** Note that the right-hand side of Equation (5.15) coincides with $\lambda^m m! L_m(-\lambda)$, where $L_m$ denotes the Laguerre polynomial of degree $m$. The term $e^{\sqrt{\lambda m}}$ agrees with the sharp asymptotics of the Laguerre polynomial on the negative axis [64, Theorem 8.22.3].

Recall that $L = c_0 \log k$. Let $c_0 \leq c_1$. The monotonicity in Lemma 7 yields that $\mathsf{var}(N_i)_m \leq \mathsf{var}(\tilde{N})_m$, where $\tilde{N} \sim \mathrm{Poi}(c_1 \log k)$ whenever $p_i \leq \frac{c_1 \log k}{n}$. Applying the upper bound in Lemma 7 and in view of the relation that $m \leq c_0 \log k \leq c_1 \log k$, the conditional variance can be further upper bounded by the following:

$$\begin{aligned}
\mathsf{var}\left[\mathcal{E}_1 | I_1\right] \lesssim & \frac{k}{n^2} \left( \sum_{m=0}^{L} \frac{2^{3L}}{(c_1 \log k)^{m-1}} \sqrt{((c_1 \log k)(c_1 \log k))^m (2e)^{2\sqrt{(c_0 \log k)(c_1 \log k)}}} \right)^2 \\
= & \frac{k}{n^2} \left( \sum_{m=0}^{L} k^{(c_0 \log 8 + \sqrt{c_0 c_1} \log(2e))} c_1 \log k \right)^2 \\
\lesssim & \frac{(\log k)^4}{n^2} k^{1+2(c_0 \log 8 + \sqrt{c_0 c_1} \log(2e))}. \qquad (5.16)
\end{aligned}$$

From Equation (5.14)–Equation (5.16) we conclude that

$$\mathbb{E}[\mathcal{E}_1^2] = \mathbb{E}\left[\mathbb{E}[\mathcal{E}_1 | I_1]^2 + \mathsf{var}(\mathcal{E}_1 | I_1)\right] \lesssim \left( \frac{k}{n \log k} \right)^2 \qquad (5.17)$$

as long as

$$c_0 \log 8 + \sqrt{c_0 c_1} \log(2e) < \frac{1}{4}. \qquad (5.18)$$

**Case 2: Bias-corrected plug-in estimator** First note that $\mathcal{E}_2$ can be written as

$$\mathcal{E}_2 = \sum_{i \in I_2} \left( (p_i - \hat{p}_i) \log \frac{1}{p_i} + \hat{p}_i \log \frac{\hat{p}_i}{p_i} - \frac{1}{2n} \right), \tag{5.19}$$

where $\hat{p}_i = \frac{N_i}{n}$ is an unbiased estimator of $p_i$ since $N_i \sim \mathrm{Poi}(np_i)$. The first term is thus unbiased conditioned on $I_2$. Note the following elementary bounds on the function $x \log x$:

**Lemma 8.** *For any $x > 0$,*

$$0 \le x \log x - (x-1) - \frac{1}{2}(x-1)^2 + \frac{1}{6}(x-1)^3 \le \frac{(x-1)^4}{3}.$$

*Proof of Lemma 8.* It follows from Taylor's expansion of $x \mapsto x \log x$ at $x = 1$ that

$$x \log x = (x-1) + \frac{1}{2}(x-1)^2 - \frac{1}{6}(x-1)^3 + \frac{1}{3} \int_1^x \left( \frac{x}{t} - 1 \right)^3 \mathrm{d}t.$$

Hence it suffices to show $0 \le \int_1^x \left( \frac{x}{t} - 1 \right)^3 \mathrm{d}t \le (x-1)^4$ for all $x > 0$. If $x > 1$, the conclusion is obvious since the integrand is always positive and no greater than $(x-1)^3$. If $x < 1$, we rewrite the integral as $\int_x^1 \left(1 - \frac{x}{t}\right)^3 \mathrm{d}t$. Then the conclusion follows from the same reason that the integrand is always positive and at most $(1-x)^3$. □

Applying the above facts to $x = \frac{\hat{p}_i}{p_i}$, we obtain that

$$\begin{aligned}
\sum_{i \in I_2} p_i \frac{\hat{p}_i}{p_i} \log \frac{\hat{p}_i}{p_i} &\ge \sum_{i \in I_2} (\hat{p}_i - p_i) + \frac{(\hat{p}_i - p_i)^2}{2p_i} - \frac{(\hat{p}_i - p_i)^3}{6p_i^2}, \\
\sum_{i \in I_2} p_i \frac{\hat{p}_i}{p_i} \log \frac{\hat{p}_i}{p_i} &\le \sum_{i \in I_2} (\hat{p}_i - p_i) + \frac{(\hat{p}_i - p_i)^2}{2p_i} - \frac{(\hat{p}_i - p_i)^3}{6p_i^2} + \frac{(\hat{p}_i - p_i)^4}{3p_i^3}.
\end{aligned}$$

Plugging the inequalities above into Equation (5.19) and taking expectation on both sides conditioned on $I_2$, using the central moments of Poisson distribution that $\mathbb{E}(X - \mathbb{E}[X])^2 = \lambda, \mathbb{E}(X - \mathbb{E}[X])^3 = \lambda, \mathbb{E}(X - \mathbb{E}[X])^4 = \lambda(1 + 3\lambda)$ when $X \sim \mathrm{Poi}(\lambda)$, we obtain that

$$-\sum_{i \in I_2} \frac{1}{6n^2 p_i} \le \mathbb{E}\left[\mathcal{E}_2 | I_2\right] \le \sum_{i \in I_2} \frac{1 + 3np_i}{3n^3 p_i^2} - \frac{1}{6n^2 p_i}.$$

By definition, $I_2 \subseteq \{i : p_i > \frac{c_3 \log k}{n}\}$ and $|I_2| \leq k$. Hence, almost surely,

$$|\mathbb{E}\left[\mathcal{E}_2|I_2\right]| \lesssim \sum_{i\in I_2} \frac{1}{n^2 p_i} + \sum_{i \in I_2} \frac{1}{n^3 p_i^2} \lesssim \frac{k}{n \log k}. \tag{5.20}$$

It remains to bound the variance of the plug-in estimator. Note that

$$\mathsf{var}\left[\mathcal{E}_2|I_2\right] \leq \sum_{i:p_i > \frac{c_3 \log k}{n}} \mathsf{var}\left[\phi(p_i) - \phi(\hat{p}_i)\right] \leq \sum_{i:p_i > \frac{c_3 \log k}{n}} \mathbb{E}\left(\phi(p_i) - \phi(\hat{p}_i)\right)^2. \tag{5.21}$$

In view of the fact that $\log x \leq x - 1$ and $x \log x \geq x - 1$ for any $x > 0$, we have

$$\hat{p}_i - p_i = p_i \left(\frac{\hat{p}_i}{p_i} - 1\right) \leq p_i \frac{\hat{p}_i}{p_i} \log \frac{\hat{p}_i}{p_i} = \hat{p}_i \log \frac{\hat{p}_i}{p_i} \leq \hat{p}_i \left(\frac{\hat{p}_i}{p_i} - 1\right) = \hat{p}_i - p_i + \frac{(\hat{p}_i - p_i)^2}{p_i}.$$

Recall that $\phi(p_i) - \phi(\hat{p}_i) = (p_i - \hat{p}_i) \log \frac{1}{p_i} + \hat{p}_i \log \frac{\hat{p}_i}{p_i}$. Then, by triangle inequality,

$$\begin{aligned}(\phi(p_i) - \phi(\hat{p}_i))^2 \leq\ & 2(p_i - \hat{p}_i)^2 \log^2 \frac{1}{p_i} + 2\left(\hat{p}_i \log \frac{\hat{p}_i}{p_i}\right)^2 \\ \leq\ & 2(p_i - \hat{p}_i)^2 \log^2 \frac{1}{p_i} + 4(\hat{p}_i - p_i)^2 + \frac{4(\hat{p}_i - p_i)^4}{p_i^2}.\end{aligned}$$

Taking expectation on both sides yields that

$$\mathbb{E}(\phi(p_i) - \phi(\hat{p}_i))^2 \leq \frac{2p_i}{n}\left(\log \frac{1}{p_i}\right)^2 + \frac{4p_i}{n} + \frac{12}{n^2} + \frac{4}{n^3 p_i}.$$

Plugging the above into Equation (5.21) and summing over $i$ such that $p_i \geq \frac{c_3 \log k}{n}$, we have

$$\mathsf{var}[\mathcal{E}_2|I_2] \lesssim \frac{(\log k)^2}{n} + \frac{k}{n^2}, \tag{5.22}$$

where we used the fact that $\sup_{P\in\mathcal{M}_k} \sum_{i=1}^k p_i \log^2 \frac{1}{p_i} \lesssim \log^2 k$. Assembling Equation (5.20)–Equation (5.22) yields that

$$\mathbb{E}\mathcal{E}_2^2 \lesssim \left(\frac{k}{n \log k}\right)^2 + \frac{\log^2 k}{n}. \tag{5.23}$$

By assumption, $\log n \leq C \log k$ for some constant $C$. Choose $c_1 > c_2 >$

$c_3 > 0$ such that $c_1 - c_2 \log \frac{ec_1}{c_2} - 1 > C$ and $c_3 + c_2 \log \frac{ec_2}{c_3} - 1 > C$ hold simultaneously, e.g., $c_1 = 4(C+1), c_2 = e^{-1}c_1$, $c_3 = e^{-2}c_1$, and $c_0 \le c_1$ satisfying the condition Equation (5.18), e.g., $c_0 = \frac{1}{300c_1} \wedge c_1 \wedge 0.01$. Plugging Equation (5.17), Equation (5.23), Equation (5.10) into Equation (5.13), we complete the proof. □

**Remark 7.** The estimator Equation (5.9) uses the polynomial approximation of $x \mapsto x \log \frac{1}{x}$ for those masses below $\frac{\log k}{n}$ and the bias-corrected plug-in estimator otherwise. In view of the fact that the lower bound in Proposition 5 is based on a pair of randomized distributions whose masses are below $\frac{\log k}{n}$ (except for possibly a fixed large mass at the last element), 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. See Remark 4 and the proof of Proposition 5 for details.

**Remark 8.** The estimator in Equation (5.9) 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 obtained from quantizing an analog source and $k$ is naturally determined by the quantization level [22]. 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 [1], we can simply let the constant term $a_0$ of the approximating polynomial Equation (5.7) go to zero and obtain the corresponding unbiased estimator Equation (5.8) through factorial moments, which satisfies $g_L(0) = 0$ by construction.[1] The bias upper bound becomes $\sum_i (P_L(p_i) - \phi(p_i) - P_L(0))$ which is at most twice original upper bound since $P_L(0) \le \|P_L - \phi\|_\infty$. The minimax rate in Proposition 6 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 Section 5.2.1 for a short proof). Therefore, removing the constant term leads to $g_L(N_i)$ which is always underbiased as shown in Figure 5.2. A

[1] Alternatively, we can directly set $g_L(0) = 0$ and use the original $g_L(j)$ in Equation (5.8) 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.

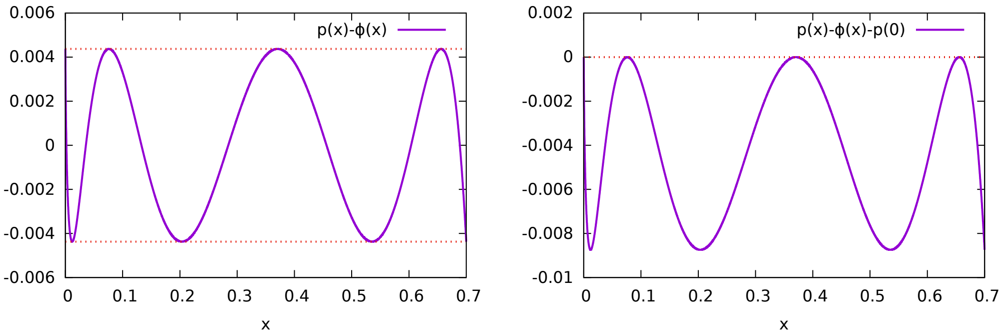

Figure 5.2: Bias of the degree-6 polynomial estimator with and without the constant term.

better choice for adaptive estimation is to find the best polynomial satisfying $p_L(0) = 0$ that uniformly approximates $\phi$.

### 5.2.1 Approximation error at the end points

We prove the claim in Remark 8. By Chebyshev alternating theorem [52, 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 $g''(x) = \frac{xP_L''(x)+1}{x}$. 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 $a_0 = g(0) = E_L(\phi)$.

# CHAPTER 6

# NUMERICAL EXPERIMENTS

In this chapter we compare the performance of our estimator described in Chapter 5 to other estimators using synthetic data.[1] Note that the coefficients of best polynomial to approximate $\phi$ 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., [52, 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 Equation (5.9) 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 $\alpha$ 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 $p_{i+k/2} \propto (1 - \frac{2}{k})^{i-1}$, $1 \leq i \leq \frac{k}{2}$. 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 Equation (5.9) 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 pre-

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

liminary estimate $\tilde{H}$ in Equation (5.9), our final output is $\tilde{H} \vee 0$.[2] Since the plug-in 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 Equation (1.8). We also compare the performance with the linear programming estimator in [29], the best upper bound (BUB) estimator [30], and the estimator based on similar polynomial approximation techniques[3] proposed by [1] using their implementations with default parameters. Our estimator is implemented in C++ which is much faster than those from [29, 1, 30] implemented in MATLAB so the running time comparison is ignored. We notice that the linear programming in [29] is much slower than the polynomial estimator in [1], 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 6.1 where the sample size ranges from one percent to 300 folds of the alphabet size. In Figure 6.2 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 [29, 1, 30], outperforms the classical Miller-Madow estimator significantly; furthermore, our estimator performs better than those in [1, 30] in most cases tested and comparably with that in [29].

When the samples are abundant all estimators achieve very small error; however, it has been empirically observed in [1] that the performance of linear programming starts to deteriorate when the sample size is very large, which is also observed in our experiments, see Figure 6.3. By Equation (5.9), for large sample size our estimator tends to the Miller-Madow estimator when every symbol is observed many times.

[2] We can, as in Proposition 6, output $(\tilde{H} \vee 0) \wedge \log k$, which yields a better performance. We elect not to do so for a stricter comparison.

[3] The estimator in [1] uses a smooth cutoff function in lieu of the indicator function in Equation (5.9); this seems to improve neither the theoretical error bound nor the empirical performance.

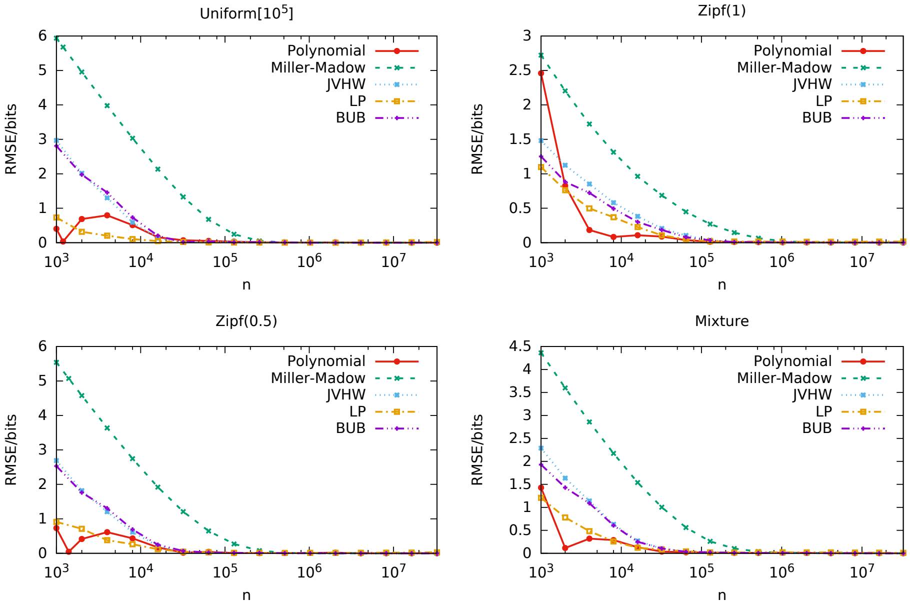

Figure 6.1: Performance comparison with sample size $n$ ranging from $10^3$ to $3 \times 10^7$.

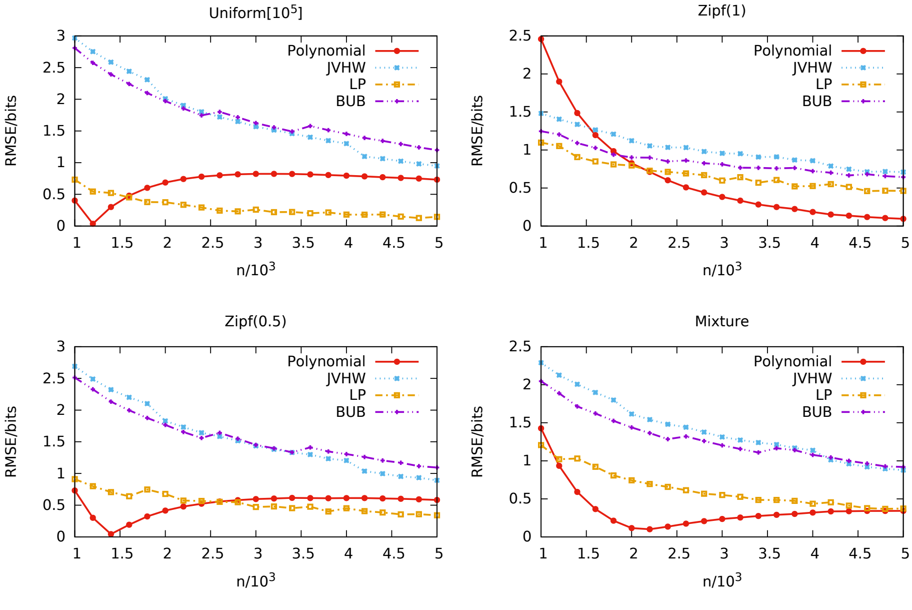

Figure 6.2: Performance comparison when sample size $n$ ranges from 1000 to 5000.

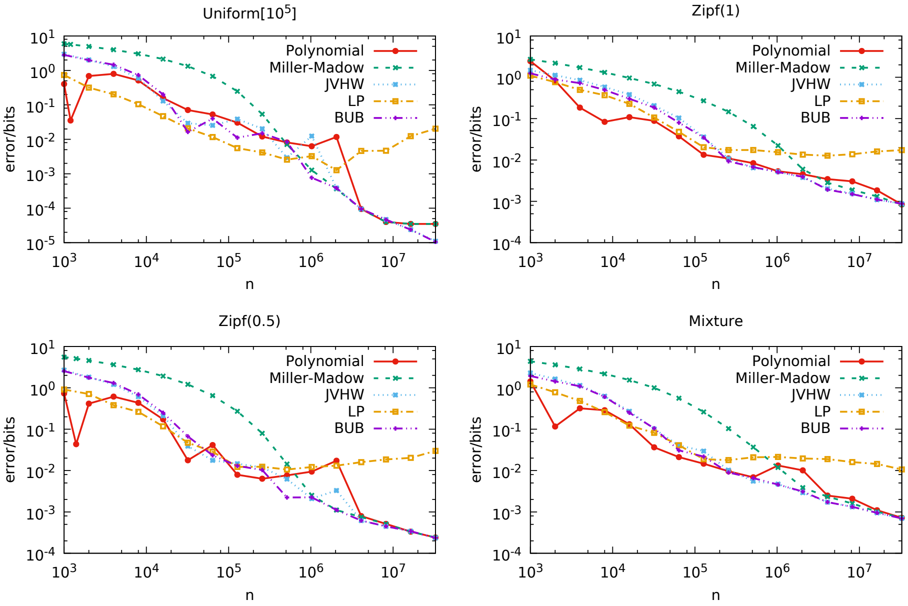

Figure 6.3: Performance comparison with sample size $n$ ranging from $10^3$ to $3 \times 10^7$ with logarithmic y-axis.

# REFERENCES

[1] 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.

[2] Y. Wu and P. Yang, “Chebyshev polynomials, moment matching, and optimal estimation of the unseen,” *arXiv:1504.01227*, 2015.

[3] F. Rieke, W. Bialek, D. Warland, and R. d. R. van Steveninck, *Spikes: Exploring the Neural Code.* The MIT Press, 1999.

[4] 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.

[5] 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.

[6] 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.

[7] C. Chow and C. Liu, “Approximating discrete probability distributions with dependence trees,” *IEEE Trans. Inf. Theory*, vol. 14, no. 3, pp. 462–467, 1968.

[8] 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.

[9] 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.

[10] 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.

[11] C. J. Stone, "Optimal rates of convergence for nonparametric estimators," *The Annals of Statistics*, vol. 8, no. 6, pp. 1348–1360, 1980.

[12] D. L. Donoho and R. C. Liu, "Geometrizing rates of convergence, II," *The Annals of Statistics*, vol. 19, pp. 668–701, 1991.

[13] 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.

[14] 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.

[15] B. Efron, "Maximum likelihood and decision theory," *The Annals of Statistics*, vol. 10, no. 2, pp. pp. 340–356, 1982.

[16] J. Berkson, "Minimum chi-square, not maximum likelihood! (with discussion)," *The Annals of Statistics*, pp. 457–487, 1980.

[17] A. W. Van der Vaart, *Asymptotic Statistics.* Cambridge, United Kingdom: Cambridge university press, 2000.

[18] 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.

[19] 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.

[20] 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.

[21] Z. F. Mainen and T. J. Sejnowski, "Reliability of spike timing in neocortical neurons," *Science*, vol. 268, no. 5216, pp. 1503–1506, 1995.

[22] 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.

[23] 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.

[24] I. J. Good, "The population frequencies of species and the estimation of population parameters," *Biometrika*, vol. 40, no. 3-4, pp. 237–264, 1953.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] L. Paninski, "Estimation of entropy and mutual information," *Neural Computation*, vol. 15, no. 6, pp. 1191–1253, 2003.

[31] A. Antos and I. Kontoyiannis, "Convergence properties of functional estimates for discrete distributions," *Random Structures & Algorithms*, vol. 19, no. 3-4, pp. 163–193, 2001.

[32] J. Jiao, Private communication, Oct. 2014.

[33] Q. Wang, S. R. Kulkarni, and S. Verdú, "Universal estimation of information measures for analog sources," *Foundations and Trends in Communications and Information Theory*, vol. 5, no. 3, pp. 265–353, 2009.

[34] G. Basharin, "On a statistical estimate for the entropy of a sequence of independent random variables," *Theory of Probability & Its Applications*, vol. 4, no. 3, pp. 333–336, 1959.

[35] G. A. Miller, "Note on the bias of information estimates," *Information Theory in Psychology: Problems and Methods*, vol. 2, pp. 95–100, 1955.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] G. Valiant and P. Valiant, "A CLT and tight lower bounds for estimating entropy," in *Electronic Colloquium on Computational Complexity (ECCC)*, vol. 17, 2010, p. 179.

[42] 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.

[43] 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.

[44] L. Le Cam, *Asymptotic Methods in Statistical Decision Theory.* New York, NY: Springer-Verlag, 1986.

[45] 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.

[46] 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.

[47] 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.

[48] J. Jiao, K. Venkat, Y. Han, and T. Weissman, "Maximum likelihood estimation of functionals of discrete distributions," *arXiv:1406.6959v4*, 2014.

[49] J. Acharya, A. Orlitsky, A. T. Suresh, and H. Tyagi, "The complexity of estimating Rényi entropy," in *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms.* SIAM, 2015, pp. 1855–1869.

[50] Y. Han, J. Jiao, and T. Weissman, "Does Dirichlet prior smoothing solve the Shannon entropy estimation problem?" *arXiv:1502.00327*, 2015.

[51] Y. Han, J. Jiao, and T. Weissman, "Adaptive estimation of Shannon entropy," *arXiv:1502.00326*, 2015.

[52] P. P. Petrushev and V. A. Popov, *Rational Approximation of Real Functions.* Cambridge University Press, 2011.

[53] D. G. Luenberger, *Optimization by Vector Space Methods.* John Wiley & Sons, 1969.

[54] R. A. DeVore and G. G. Lorentz, *Constructive Approximation.* Springer, 1993.

[55] A. F. Timan, *Theory of Approximation of Functions of a Real Variable.* Pergamon Press, 1963.

[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] A. Tsybakov, *Introduction to Nonparametric Estimation.* New York, NY: Springer Verlag, 2009.

[59] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems.* Academic Press, Inc., 1982.

[60] G. Stewart, *Afternotes on Numerical Analysis.* Society for Industrial and Applied Mathematics, 1996.

[61] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables.* Courier Corporation, 1964, no. 55.

[62] D. Berend and A. Kontorovich, “A sharp estimate of the binomial mean absolute deviation with applications,” *Statistics & Probability Letters*, vol. 83, no. 4, pp. 1254–1259, 2013.

[63] J. M. Steele, “An Efron-Stein inequality for nonsymmetric statistics,” *The Annals of Statistics*, pp. 753–758, 1986.

[64] G. Szegö, *Orthogonal Polynomials*, 4th ed. Providence, RI: American Mathematical Society, 1975.