Minimax rates of entropy estimation on large alphabets via best polynomial approximation

Yihong Wu and Pengkun Yang*

Abstract

Consider the problem of estimating the Shannon entropy of a distribution on $k$ elements from $n$ independent samples. We show that the minimax mean-square error is within universal multiplicative constant factors of
\[
\left( \frac{k}{n \log k} \right)^2 + \frac{\log^2 k}{n}
\]
as long as $n$ grows no faster than a polynomial of $k$. This implies the recent result of Valiant-Valiant [VV11] that the minimal sample size for consistent entropy estimation scales according to $\Theta\left( \frac{k}{\log k} \right)$. The apparatus of best polynomial approximation plays a key role in both the minimax lower bound and the construction of optimal estimators.

1 Introduction

Let $P$ be a distribution over an alphabet of cardinality $k$. Let $X_1, \ldots, X_n$ be i.i.d. samples drawn from $P$. Without loss of generality, we shall assume that the alphabet is $[k] \triangleq \{1, \ldots, k\}$. To perform statistical inference on the unknown distribution $P$ or any functional thereof, a sufficient statistic is $N \triangleq (N_1, \ldots, N_k)$, where
\[
N_j = \sum_{i=1}^n 1\{X_i=j\}
\]
records the number of occurrence 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 input distribution $P$:
\[
H(P) = \sum_{i=1}^k p_i \log \frac{1}{p_i}.
\]

Entropy estimation has many applications in various fields, such as neuroscience [RBWvS99], physics [VBB+12], telecommunication [PW96], biomedical research [PGM+01], etc. To investigate the decision-theoretic fundamental limit, we consider the minimax quadratic risk of entropy estimation:
\[
R^*(k, n) \triangleq \inf_{\hat{H}} \sup_{P \in \mathcal{M}_k} \mathbb{E}[ (\hat{H}(N) - H(P))^2 ]
\]
(1)

where $\mathcal{M}_k$ denotes the set of probability distributions on $[k]$. The goal of the paper is to provide non-asymptotic characterization of the minimax risk $R^*(k, n)$ within constant factors.

*The authors are with the Department of ECE, University of Illinois at Urbana-Champaign, Urbana, IL, \{yihongwu,pyang14\}@illinois.edu.
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 per se (the distribution \( P \)), 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 a linear functional [Sto80, DL91], quadratic functional [CL05], \( L_q \) norm [LNS99], etc. To estimate a function, perhaps the most natural idea is the “plug-in” approach, namely, first estimate the parameter then substitute into the function. This leads to the commonly used plug-in estimator, i.e., the empirical entropy,

\[
\hat{H}_{\text{plug-in}} = H(\hat{P}),
\]

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 suffers from severe bias. Indeed, although \( \hat{H}_{\text{plug-in}} \) is asymptotically efficient in the “fixed-\( P \)-large-\( n \)” regime, it can be highly suboptimal in high dimensions.

Our main result is the characterization of the minimax quadratic risk with universal constant factors as long as the sample size grows no faster than a polynomial of the alphabet size:

**Theorem 1.** If \( n \gtrsim \frac{k}{\log k} \) and \( \log n \lesssim \log k, \)

\[
R^*(k, n) \asymp \left( \frac{k}{n \log k} \right)^2 + \frac{\log^2 k}{n},
\]

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

To interpret the minimax rate (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 \( R^*(k, n) \asymp \left( \frac{k}{n \log k} \right)^2 \) 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 with an additive error of \( \epsilon \), the minimal sample size is given by \( n \geq \frac{\log^2 k}{\epsilon^2} \lor \frac{k}{\epsilon \log k} \).

Using Theorem 1, next we evaluate the performance of plug-in estimator in terms of its worst-case mean-square error

\[
R_{\text{plug-in}}(k, n) \equiv \sup_{P \in M_k} \mathbb{E}[(\hat{H}_{\text{plug-in}}(N) - H(P))^2].
\]

Analogous to Theorem 1 which applies to the optimal estimator, the risk of the plug-in estimator admits a similar (perhaps well-known) characterization (see Appendix A for a short proof based on Pinsker’s inequality): If \( n \gtrsim k \) and \( \log n \lesssim \log k \), then

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

If \( n \lesssim k \), then \( \hat{H}_{\text{plug-in}} \) is inconsistent, i.e., \( R_{\text{plug-in}}(k, n) \gtrsim 1 \). Again the first and second term corresponds to the bias and variance respectively. It is known that the variance of the plug-in estimator is always upper bounded by a constant factor of \( \frac{\log^2 n}{n} \), regardless of the alphabet size [AK01, Remark (iv), p. 168], while the bias can be at large as \( \frac{k}{n} \).

\(^1\) 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 \( a_n \geq cb_n \) 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.
Comparing (3) and (5), we reach the following verdict on the plug-in estimator: In the regime where the sample size $n$ grows at most polynomially in $k$, empirical entropy is rate-optimal, i.e., within a constant factor of the minimax risk, if and only if we are in the “data-rich” regime $n = \Theta\left(\frac{k^2}{\log k}\right)$. In the “data-starved” regime of $n = o\left(\frac{k^2}{\log k}\right)$, empirical entropy is strictly rate-suboptimal.

1.1 Previous results

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 scope of the present paper (see the survey [WKV09] and the references therein).

**Fixed alphabet** For fixed distribution $P$ and $n \to \infty$, Antos and Kontoyiannis showed that [AK01] the plug-in estimator is always consistent and the asymptotic variance of the plug-in estimator is obtained in [Bas59]. 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., [Mil55, Har75], which inspired various types of biased correction to the plug-in estimator.

**Large alphabet** It is well-known that to estimate the distribution $P$ itself, say, under the total variation loss, we need at least $\Theta(k)$ samples. However, to estimate the entropy $H(P)$ which is a scalar function, it is unclear from first principles whether $n = \Theta(k)$ is necessary. 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 $R^*(k, n) \to 0$ as $k \to \infty$ [Pan04]. Valiant proved that no consistent estimator exists, i.e., $R^*(k, n_k) \geq 1$ if $n \lesssim \frac{k}{\exp\left(\sqrt{\log k}\right)}$ [Val08]. The exact scaling of for the minimal sample size of consistent estimation is shown to be $\frac{k}{\log k}$ in the breakthrough results of Valiant and Valiant [VV10, VV11]. Theorem 1 generalizes this result by characterizing the full minimax rate.

We briefly discuss the difference between our lower bound strategy and that used in [VV10]. Since the entropy is a permutation-invariant functional of the distribution, a sufficient statistics for entropy estimation is a further summary of the sufficient statistics $N$:

$$h_i = \sum_{j=1}^{k} 1_{\{X_i = j\}}, \quad i \in [n],$$

known as histogram order statistics [Pan03], profile [OSZ04], or fingerprint [VV10], which counts the number of symbols that appear exactly $i$ times in the sample. A canonical approach to obtain minimax functional estimation lower bound is Le Cam’s two-point argument [LC86, Chapter 2], which involves finding two distributions with very different entropy and induce almost the same distribution for the sufficient statistics, in this case, the count statistics $N^*_i$ of fingerprint $h^*_i$, both of which have non-produce distributions. A frequently used technique to reduce dependence is Poisson sampling (see Section 2), where we relax the sample size to a Poisson random variable with mean $n$. This has almost no impact on the statistical behavior due to the exponential concentration of the Poisson distribution near its mean. Under Poisson sample, the sufficient statistics $N_1, \ldots, N_k$ are independent Poissons with mean $np_i$. However, even with Poisson sampling, the entries of the fingerprint are still highly dependent. To contend with the difficulty of computing statistical distance between high-dimensional distribution with dependent entries, the major tool [VV10] is new CLT for approximating the fingerprint distribution by quantized Gaussian distribution, which are parameterized by the mean and covariance matrices and hence more tractable.
In contrast, in this paper we shall not deal with the fingerprint directly, but rather use the original sufficient statistics $N^k$ due to the independence. Our lower bound relies on choosing two random distributions with almost iid entries which effectively reduces the problem to one dimension, thus circumventing the hurdle of dealing with high-dimensional non-product distributions.

1.2 Best polynomial approximation

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 [LNS99, CL11] for in Gaussian mean models. Nemirovski (credited in [INK87]) pioneered the use of polynomial approximation in functional estimation. It is shown that unbiased estimators for the truncated Taylor series of the smooth functionals is asymptotically efficient. This strategy is generalized to non-smooth functionals in [LNS99] using best polynomial approximation and in [CL11] for estimating the $\ell_1$-norm in Gaussian mean model.

On the constructive side, the main idea is to trade bias with variance. In many statistical models, unbiased estimator exists for monomials and hence for any polynomials, but not for entropy involving logarithms (see [Pan03, Proposition 8, p. 1236]). Therefore a natural idea is to approximate the functional by polynomials which enjoy unbiased estimation, and reduce the bias to at most the uniform approximation error. The choice of the degree aims to achieve a good bias-variance balance. In fact the use of polynomial approximation in entropy estimation is not new. In [VBB+12], 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 investigated using approximation by Bernstein polynomials to reduce the bias of the plug-in estimators [Pan03], which forms the basis for proving the existence of consistent estimators with sublinear sample complexity in [Pan04].

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 observed in [LNS99, CL11], it turns out that the lower bound involves choosing two priors with matching moments up to a certain degree, which ensures that the impossibility to test. The lower bound is then given by the maximal separation in the expected functional value subject to the moment matching condition. This problem is the dual of best polynomial approximation in the optimization sense (see Appendix B for a self-contained account). For entropy estimation, this strategy yields the optimal minimax lower bound, although the argument is considerably more involved due to the extra constraint on the mean of the prior.

Notations Throughout the paper all logarithms are with respect to the natural base and the entropy is measured in nats. $\text{Poi}(\lambda)$ denotes the Poisson distribution with mean $\lambda$ whose probability mass function is $\text{poi}(\lambda, j) \triangleq \frac{\lambda^j e^{-\lambda}}{j!}, j \in \mathbb{Z}_+$. Given a distribution $P$, its $n$-fold product is denoted by $P^{\otimes n}$. For a parametrized family of distributions $\{P_\theta\}$ and a prior $\pi$, the mixture is denoted by $\mathbb{E}_\pi [P_\theta] = \int P_\theta \pi(d\theta)$. In particular, $\mathbb{E} [\text{Poi} (U)]$ denotes the Poisson mixture with respect to the distribution of a positive random variable $U$. The total variation and Kullback-Leibler divergence between probability measures $P$ and $Q$ is $\text{TV}(P, Q) = \frac{1}{2} \int |dP - dQ|$ and $D(P \parallel Q) = \int dP \log \frac{dP}{dQ}$.

2 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 \( n' \sim \text{Poi}(n) \), then draw \( n' \) i.i.d. samples from the distribution \( P \). The main benefit is that now the sufficient statistics \( N_i \sim \text{Poi}(np_i) \) are independent, which significantly simplifies the analysis. In view of the exponential tail of Poisson distributions, the sample size is concentrated near its mean \( n \) with high probability, which guarantees that the statistical performance as well as the minimax risk under Poisson sampling are provably close to that with fixed sample size.

Analogous to the minimax risk (1), we define its counterpart under the Poisson sampling model:

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

(7)

where \( N_i \sim \text{Poi}(np_i) \) for \( i = 1, \ldots, k \). Note that by definition we have

\[
\tilde{R}^*(k, n) = \mathbb{E}_{n' \sim \text{Poi}(n)}[R^*(k, n')] = \sum_{m \geq 0} R^*(k, m) \text{poi}(n, m).
\]

(8)

Since \( 0 \leq R^*(k, m) \leq \log^2 k \), in view of the fact that \( m \mapsto R^*(k, m) \) is decreasing and applying Markov’s inequality and the Chernoff bound (see, e.g., [MU05, Theorem 5.4], we have

\[
\tilde{R}^*(k, 2n) - \exp(-n/4) \log^2 k \leq R^*(k, n) \leq 2\tilde{R}^*(k, n/2),
\]

(9)

which allows us to focus on the risk of the Poisson model.

3 Minimax lower bound

In this section 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 (3) separately. This follows from combining Propositions 1 and 2 below.

**Proposition 1.** For all \( k, n \in \mathbb{N} \),

\[
R^*(k, n) \gtrsim \frac{\log^2 k}{n}.
\]

(10)

**Proposition 2.** If \( n \geq \frac{ck}{\log k} \) for some \( c > 0 \), then

\[
R^*(k, n) \geq c' \left( \frac{k}{n \log k} \right)^2
\]

(11)

where \( c' \) only depends on \( c \).

Proposition 1, proved in Section 5.1, 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 value \( |H(P) - H(Q)|^2 \).

The remainder of this section is devoted to illustrating the broad strokes for proving Proposition 2. The proofs as well as the intermediate results are elaborated in Section 5. 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 (11) requires more powerful technique. To this end, we use a generalized version of Le Cam’s method involving two composite hypotheses (also known as fuzzy hypothesis testing in [Tsy09]):

\[
H_0 : H(P) \leq t \quad \text{versus} \quad H_1 : H(P) \geq t + d,
\]

which is more general than the two-point argument using only simple hypothesis testing. Similarly, if we can establish that no test can distinguish (12) 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 prior.

For (12) we need to choose a pair of priors, which, in this case, are distributions on the probability simplex \( \mathcal{M}_k \), is to ensure the entropy values are separated.

### 3.1 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 iid construction. Let \( U \) be a \( \mathbb{R}_+ \)-valued random variable with unit mean. Denote the random vector

\[
P = \frac{1}{k}(U_1, \ldots, U_k),
\]

consisting of iid copies of \( U \). Note that \( P \) itself is not a probability distribution; however, the key observation is that, since \( \mathbb{E}[U] = 1 \), the law of large numbers implies \( P \) is approximately a probability distribution. Use some soft arguments we can show that the distribution of \( P \) can effectively serve as a prior. To gain more intuitions, 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 where the support is uniformly chosen.

Next we consider the main ingredients in Le Cam’s method:

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

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

which also concentrates near its mean \( \mathbb{E}[H(P)] = \mathbb{E}[\phi(U)] + \mathbb{E}[U] \log k \). Therefore, given another random variable \( U' \) with unit mean, we can obtain \( P' \) similarly using iid copies of \( U' \).

Then with high probability, \( H(P) \) and \( H(P') \) are separated by the difference in the respective means

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

which we want to maximize.

2. **Indistinguishably**: Note that given \( P \), the sufficient statistics satisfy \( N_i \overset{\text{iid}}{\sim} \text{Poi}(np_i) \). Therefore, if \( P \) is drawn from the distribution of \( P \), then \( N = (N_1, \ldots, N_k) \) are iid distributed according the Poisson mixture \( \mathbb{E}[\text{Poi}(\frac{n}{k} U)] \). Similarly, if \( P \) is drawn from the prior of \( P' \), then \( N \) is distributed according to \( (\mathbb{E}[\text{Poi}(\frac{n}{k} U')])^{\otimes k} \). To establish the impossibility of testing, we need the total variation distance between the two \( k \)-product distributions to strictly bounded away from one, for which a sufficient conditions is

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

for some small \( c \).
To conclude, we see that the iid construction fully exploits the independence blessed by the Poisson sampling, thereby reduce the problem to one dimension. This allows us to sidestep the difficulty encountered in [VV10] when dealing with fingerprint which are high-dimensional random vector with dependent distribution.

What remains is the following scalar problem: choose $U, U'$ to maximize $|E[\phi(U)] - E[\phi(U')]|$ subject to the constraint (13). A commonly used proxy for bounding the total variation distance is moment matching, i.e., $E[U^j] = E[U'^j]$ for all $j = 1, \ldots, L$. Together with some $L_{\infty}$-norm constraints, a sufficient large $L$ ensures the total variation bound (13). Combining the above steps, our lower bound is proportional to the value of the following convex optimization problem (in fact, infinite-dimensional linear programming):

$$F_L(\lambda) \triangleq \sup E[\phi(U)] - E[\phi(U')]$$

s.t. $E[U] = E[U'] = 1$

$$E[U^j] = E[U'^j], \quad j = 1, \ldots, L,$$

$U, U' \in [0, \lambda]$ (14)

for some $L \in N$ and $\lambda > 1$ depending on $n$ and $k$.

Finally, we connect the optimization problem (14) to the machinery of best polynomial approximation: Denote by $\mathcal{P}_L$ the set of polynomials of degree $L$ and

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

which is the best uniform approximation error of a function $f$ over a finite interval $I$ by polynomials of degree $L$. We prove that

$$F_L(\lambda) \geq 2E_L(\log([1/\lambda, 1])).$$

Due to the singularity of the logarithm function near zero, the approximation error can be made bounded away from zero if $\lambda$ grows quadratically with the degree $L$ (see Appendix C). This leads to the lower bound of $n \gtrsim \frac{k}{\log k}$ for consistent estimation. For $n \gg \frac{k}{\log k}$, the lower bound for the quadratic risk follows from relaxing the unit-mean constraint in (14) to $E[U] = E[U'] \leq 1$. We refer to Section 5 for the details.

4 Optimal estimator via best polynomial approximation

As observed in various previous results as well as suggested by the minimax lower bound in Section 3, 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 $T : [k]^n \rightarrow \mathbb{R}_+$ is a polynomial of the underlying distribution $P$ and, consequently, no unbiased estimator for the entropy exists (see, e.g., [Pan03, 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 details. In view of the relationship (9) 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 $X = (X_1, \ldots, X_{n'})$ from
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 $B_i \sim \text{Bern}(\frac{1}{2})$. We split samples $X$ by the output of $B$ into two sets and count the samples in each set separately. That is, we acquire $N = (N_1, \ldots, N_k)$ and $N' = (N'_1, \ldots, N'_k)$ by

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

Then $N$ and $N'$ are independent, where $N_i, N'_i \sim \text{Poi}(np_i)$. Let $c_0, c_1, c_2 > 0$ be constants to be specified. Let $L = \lceil c_0 \log k \rceil$. Denote the best polynomial of degree $L$ to uniformly approximate $x \log \frac{1}{x}$ on $[0, 1]$ is $p_L(x) = \sum_{m=0}^{L} a_m x^m$. 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

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

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

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

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

$$\hat{H} \triangleq \sum_{i=1}^{k} \left( g_L(N_i) 1\{N'_i \leq c_2 \log k\} + \left( \frac{N_i}{n} \right) + \frac{1}{2n} \right) 1\{N'_i > c_2 \log k\}, \quad (17)$$

where we apply the estimator from polynomial approximation if $N'_i \leq c_2 \log k$ or the bias-corrected plug-in estimator otherwise. Since $0 \leq H(P) \leq \log k$ for any distribution $P$ with alphabet size $k$, we define our final estimator by:

$$\tilde{H} = \hat{H} \lor 0 \land \log k,$$

The next result gives an upper bound on the above estimator under the Poisson sampling model, which, in view of the right inequality in (9), implies the upper bound on the minimax risk $R^*(n,k)$ in Theorem 1.

**Theorem 2.** Assume that $\log n \leq C \log k$ for some constant $C > 0$. Then there exists $c_0, c_1, c_2$ depending on $C$ only, such that

$$\sup_{P \in M_k} \mathbb{E}(H(P) - \tilde{H}(N))^2 \leq \left( \frac{k}{n \log k} \right)^2 + \frac{\log^2 k}{n},$$

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

8
Remark 1. The benefit of sample splitting is that we can first condition on the realization of $N'$ and treat the indicator as deterministic. This is a frequently-used idea to simplify analysis (see, e.g., the aggregation estimator in [CMW13]) at the price of losing half of the sample thereby inflating the risk by a constant factor. It remains to be shown if we can use the same sample for both selection and estimation. Note that our estimator is a linear estimator in the fingerprint of the second half of the sample.

Proof of Theorem 2. 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 a constant factor of $\log k$. Define two events by $E_1 \triangleq \bigcap_{i=1}^k \{N'_i \leq c_2 \log k \Rightarrow p_i \leq \frac{c_1 \log k}{n}\}$ and $E_2 \triangleq \bigcap_{i=1}^k \{N'_i > c_2 \log k \Rightarrow p_i > \frac{c_3 \log k}{n}\}$, where $c_1 > c_2 > c_3$. Applying the union bound and Chernoff bound for Poissons ([MU05, Theorem 5.4]) gives

\[
\mathbb{P}[E_1^c] = \mathbb{P}\left[\bigcup_{i=1}^k \{N'_i \leq c_2 \log k, p_i > \frac{c_1 \log k}{n}\}\right]
\leq k \mathbb{P}[\text{Poi}(c_1 \log k) \leq c_2 \log k]
\leq \frac{1}{k^{c_1 - c_2 \log \frac{c_1}{c_2} - 1}},
\]

(18)

and, entirely analogously,

\[
\mathbb{P}[E_2^c] \leq \frac{1}{k^{c_3 + c_2 \log \frac{c_3}{c_2} - 1}}.
\]

(19)

Define a event $E \triangleq E_1 \cap E_2$. Again union bound gives us $\mathbb{P}[E^c] \leq \mathbb{P}[E_1^c] + \mathbb{P}[E_2^c]$.

We know that $\bar{H} = \bar{H}^+ \land \log k$ and $H(P) \in [0, \log k]$, therefore $|H(P) - \bar{H}| \leq |H(P) - H|$ and $|H(P) - \bar{H}| \leq \log k$. So MSE can be decomposed and upper bounded by

\[
\begin{align*}
\mathbb{E}(H(P) - \hat{H})^2 &= \mathbb{E}[(H(P) - \hat{H})^2 1_E] + \mathbb{E}[(H(P) - \hat{H})^2 1_{E^c}] \\
&\leq \mathbb{E}[(H(P) - \hat{H})^2 1_E] + (\log k)^2 (\mathbb{P}[E_1^c] + \mathbb{P}[E_2^c]).
\end{align*}
\]

(20)

Define

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

where the (random) index sets defined by

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

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

\[
(H(P) - \hat{H})1_E = \mathcal{E}_1 1_E + \mathcal{E}_2 1_E.
\]

(21)

Combining (20)–(21) we obtain

\[
\mathbb{E}(H(P) - \hat{H})^2 \leq 2\mathbb{E}[\mathcal{E}_1^2] + \mathbb{E}[\mathcal{E}_2^2] + (\log k)^2 (\mathbb{P}[E_1^c] + \mathbb{P}[E_2^c]).
\]

(22)

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., [Tim63, Section 7.5.4]) the optimal uniform approximation error of ϕ by degree-L polynomials on [0, 1] satisfies $L^2 E_L(ϕ, [0, 1]) \to c > 0$ as $L \to \infty$. Therefore $E_L(ϕ, [0, 1]) \lesssim L^{-2}$. By a change of variables, it is easy to show that

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

By definition, $I_1 \subseteq \{i : p_i \leq \frac{c_1 \log k}{n}\}$. 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

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

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

$$\text{var} [\mathcal{E}_1|I_1] = \sum_{i \in I_1} \text{var} [ϕ(p_i) - g_L(N_i)] \leq \sum_{i : p_i \leq \frac{c_1 \log k}{n}} \text{var} [g_L(N_i)]$$

$$= \sum_{i : p_i \leq \frac{c_1 \log k}{n}} \text{var} \left[ \sum_{m \neq 1} \frac{a_m}{(c_1 \log k)^{m-1}} (N_i)_m + \left( a_1 + \log \frac{n}{c_1 \log k} \right) \frac{N_i}{n} \right]$$

$$\leq \frac{1}{n^2} \sum_{i : p_i \leq \frac{c_1 \log k}{n}} \left( \sum_{m \neq 1} \frac{|a_m|}{(c_1 \log k)^{m-1}} \sqrt{\text{var}(N_i)_m} + \left| a_1 + \log \frac{n}{c_1 \log k} \right| \sqrt{\text{var}(N_i)} \right)^2.$$ 

Since $0 \leq ϕ(x) \leq e^{-1}$ on $[0, 1]$ and $\sup_{0 \leq x \leq 1} |P_L(x) - ϕ(x)| = E_L(ϕ, [0, 1]) \leq e^{-1}$, we have $\sup_{0 \leq x \leq 1} |P_L(x)| \leq 2e^{-1}$. From the proof of [CL11, Lemma 2, p. 1035] we know that the polynomial coefficients can be upper bounded by $|a_m| \leq 2e^{-1} \cdot 2L$. Since $\log n \leq C \log k$, we have $|a_1 + \log \frac{n}{c_1 \log k}| \lesssim 2^{2L}$. We also need the following lemma to upper bound the variance of $(N_i)_m$:

**Lemma 1.** Suppose $X \sim \text{Poi}(λ)$ and $(x)_m = \frac{x^m}{(x-m)_m!}$. Then $\text{var}(X)_m$ is increasing in $λ$ and

$$\text{var}(X)_m \leq (λm)^m \left( \frac{(2e)^2 \lambda^m}{\sqrt{m}} \right) \vee 1.$$ 

Recall that $L = c_0 \log k$. Let $c_0 \leq c_1$. The monotonicity in Lemma 1 yields $\text{var}(N_i)_m \leq \text{var}(N)_m$ where $N \sim \text{Poi}(c_1 \log k)$ whenever $p_i \leq \frac{c_1 \log k}{n}$. The upper bound in Lemma 1 shows that the conditional variance can be further upper bounded by the following

$$\text{var} [\mathcal{E}_1|I_1] \lesssim \frac{k}{n^2} \left( \sum_{m=0}^L \frac{2^{2L} \lambda^m m! \log(2e) \log(c_1 \log k)}{(c_1 \log k)^{m-1}} \right)^2$$

$$\leq \frac{k}{n^2} \left( \sum_{m=0}^L \frac{k(c_0 \log 8 + \sqrt{c_0 \log(2e)} \log(c_1 \log k))}{c_1 \log k} \right)^2$$

$$\lesssim \frac{(\log k)^4}{n^2} k^{1+2(c_0 \log 8 + \sqrt{c_0 \log(2e)})}. 

(24)$$
From (23)–(24) we conclude that

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

(25)
as long as

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

(26)

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

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

(27)

where $\hat{p}_i = \frac{N_i}{n}$ is an unbiased estimator of $p_i$ since $N_i \sim \text{Poi}(np_i)$. The first term is thus unbiased conditioned on $I_2$. Note the following elementary bounds on the function $x \log x$: For any $x > 0$,

$$x \log x \geq (x - 1) + \frac{1}{2}(x - 1)^2 - \frac{1}{6}(x - 1)^3,$$

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

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

$$\sum_{i \in I_2} p_i \frac{\hat{p}_i}{p_i} \log \frac{\hat{p}_i}{p_i} \geq \sum_{i \in I_2} (\hat{p}_i - p_i) + \frac{(\hat{p}_i - p_i)^2}{2p_i} - \frac{(\hat{p}_i - p_i)^3}{6p_i^2},$$

$$\sum_{i \in I_2} p_i \frac{\hat{p}_i}{p_i} \log \frac{\hat{p}_i}{p_i} \leq \sum_{i \in I_2} (\hat{p}_i - p_i) + \frac{(\hat{p}_i - p_i)^2}{2p_i} - \frac{(\hat{p}_i - p_i)^3}{6p_i^2} + \frac{(\hat{p}_i - p_i)^4}{p_i^3}.$$

Plugging the inequalities above into (27) and taking expectation on both sides conditioned on $I_2$, we obtain

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

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

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

(28)

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

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

(29)

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

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

\[
(\phi(p_i) - \phi(\hat{p}_i))^2 \leq 2(p_i - \hat{p}_i)^2 \left( \log \frac{1}{p_i} \right)^2 + 2 \left( \hat{p}_i \log \frac{\hat{p}_i}{p_i} \right)^2
\]

\[
\leq 2(p_i - \hat{p}_i)^2 \left( \log \frac{1}{p_i} \right)^2 + 4(\hat{p}_i - p_i)^2 + 4 \left( \frac{\hat{p}_i - p_i}{p_i} \right)^2.
\]

Taking expectation on both sides yields that

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

Plugging the above into (29) and summing over \( i \) such that \( p_i \geq \frac{c_3 \log k}{n} \), we have

\[
\text{var}[\mathcal{E}_2|I_2] \lesssim \frac{(log k)^2}{n} + \frac{k}{n^2}
\]

where we used the fact that \( \sup_{P \in \mathcal{M}_k} \sum_{i=1}^k p_i \log \frac{1}{p_i} \lesssim \log^2 k \). Assembling (28)–(30) yields

\[
\mathbb{E}\mathcal{E}_2^2 \lesssim \left( \frac{1}{\log^4 k} \wedge \left( \frac{k}{n \log k} \right)^2 \right) + \frac{\log^2 k}{n}.
\]  

By assumption, \( \log n \leq C \log k \) for some universal constant \( C \). Choose \( c_1 > c_2 > c_3 > 0 \) and \( c_0 \leq c_1 \) such that \( c_1 - c_2 \log \frac{\log n}{\log k} - 1 > C \), \( c_3 + c_2 \log \frac{\log n}{\log k} - 1 > C \) and the condition (26) holds simultaneously. In particular, we can choose \( c_1 = 4(C + 1) \) and \( c_2 = e^{-1}c_1 \), \( c_3 = e^{-2}c_1 \). Plugging (25), (31) and (18) – (19) into (22), we complete the proof.

**Proof of Lemma 1.** First we compute \( \mathbb{E}(X)_m^2 \):

\[
\mathbb{E}(X)_m^2 = \sum_{x=0}^\infty \frac{e^{-\lambda} x^2}{x!} (x-m)!^2 = \sum_{x=0}^\infty \frac{e^{-\lambda} x^m (x+1)!}{x!} = \lambda^m m! \mathbb{E}(\frac{X+m}{X})
\]

\[
= \lambda^m m! \sum_{k=0}^m \frac{m}{k} \mathbb{E}(\frac{X}{X-k}) = \lambda^m m! \sum_{k=0}^m \frac{m}{k} \mathbb{E}(X)_k = \lambda^m m! \sum_{k=0}^m \frac{m}{k} \lambda^k_!
\]

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

\[
\text{var}(X)_m = \lambda^m m! \sum_{k=0}^m \frac{m}{k} \frac{\lambda^k}{k!} - \lambda^{2m} = \lambda^m m! \sum_{k=0}^{m-1} \frac{m}{k} \frac{\lambda^k}{k!} \leq \lambda^m m! \sum_{k=0}^{m-1} \frac{(\lambda m)^k}{(k!)^2}
\]

The monotonicity of \( \lambda \rightarrow \text{var}(X)_m \) follows from the equality part immediately. The maximal term in the summation is attained at \( k^* = \lfloor \sqrt{\lambda m} \rfloor \). Therefore

\[
\text{var}(X)_m \leq \lambda^m m! m \frac{(\lambda m)^{k^*}}{(k^*!)^2} \leq (\lambda m)^m \frac{(\lambda m)^{k^*}}{(k^*!)^2}
\]

If \( \lambda m < 1 \) then \( k^* = 0 \), \( \frac{(\lambda m)^{k^*}}{(k^*!)^2} = 1 \). Otherwise if \( \lambda m \geq 1 \) then \( \frac{\sqrt{\lambda m}}{2} < k^* \leq \sqrt{\lambda m} \). Applying \( k^* > \frac{\sqrt{\lambda m}}{2} \left( \frac{\lambda m}{k^*} \right)^{k^*} \) yields

\[
\frac{(\lambda m)^{k^*}}{(k^*!)^2} \leq \frac{(\lambda m)^{k^*}}{2\pi \sqrt{\lambda m}^2 (\lambda m/4e)^{k^*}} = \frac{(2e)^2 \sqrt{\lambda m}}{\pi \sqrt{\lambda m}}
\]
Remark 2. Note that the right-hand side of (32) 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 [Sze75, Theorem 8.22.3].

5 Proofs in the lower bound

5.1 Proof of Proposition 1

Proof. For any pair of distributions \( P \) and \( Q \), Le Cam’s two-point method (see, e.g., [Tsy09, Section 2.4.2]) yields

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

Therefore it boils down to solving the optimization problem:

\[
\sup \{ H(P) - H(Q) : D(P\|Q) \leq 1/n \}.
\]

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

\[
P = \left( \frac{1}{3(k-1)}, \ldots, \frac{1}{3(k-1)}, \frac{2}{3} \right), \quad Q = \left( \frac{1 + \epsilon}{3(k-1)}, \ldots, \frac{1 + \epsilon}{3(k-1)}, \frac{2 - \epsilon}{3} \right).
\]

Direct computation yields \( D(P\|Q) = 2^3 \log \frac{2}{\epsilon} + \frac{1}{3} \log \frac{1}{\epsilon} \leq \epsilon^2 \) and \( H(Q) - H(P) = \frac{1}{3} (\epsilon \log(k-1) + \log 4 + (2 - \epsilon) \log \frac{1}{\epsilon} + (1 + \epsilon) \log \frac{1}{\epsilon+1}) \geq \frac{1}{3} \log(2(k-1)) \epsilon - \epsilon^2 \). Choosing \( \epsilon = \frac{1}{\sqrt{n}} \) and applying (33), we obtain the desired (10).

\[ \square \]

Remark 3. In view of the Pinsker inequality \( D(P\|Q) \geq 2TV^2(P, Q) \) [CK82, p. 58] as well as the continuity property of entropy with respect to the total variation distance: \( |H(P) - H(Q)| \leq \text{TV}(P, Q) \log \frac{k}{\text{TV}(P, Q)} \) for \( \text{TV}(P, Q) \leq \frac{1}{k} \) [CK82, Lemma 2.7], we conclude that the best lower bound given by the two-point method, i.e., the supremum in (34), is in the order of \( \frac{\log k}{\sqrt{n}} \). Therefore the choice of the pair (35) is optimal. Interestingly, the Cramér-Rao inequality for unbiased estimators gives the same lower bound as (10), which, of course, does not directly constitute a minimax lower bound.

5.2 Proof of Proposition 2

The following result makes rigorous the i.i.d. construction of priors described in Section 3.1.

Theorem 3. Let \( U \) and \( U' \) be random variables such that \( U, U' \in [0, \lambda] \) and \( \mathbb{E}[U] = \mathbb{E}[U'] \leq 1 \) and \( |\mathbb{E}[\phi(U)] - \mathbb{E}[\phi(U')]| \geq d \), where \( \lambda < c'k \) for some small constant \( c' \). Then

\[
R^*(k, n/2) \geq \frac{d^2}{96} \left( \frac{7}{8} - kTV(\mathbb{E}[\text{Poi}(nU/k)], \mathbb{E}[\text{Poi}(nU'/k)]) - \frac{32\lambda^2 \log^2 \frac{k}{\lambda}}{kd^2} \right) - \log^2 k \exp\left( \frac{-n}{50} \right) - \frac{16\lambda^2 \log^2 \frac{k}{\lambda}}{k} - \left( 1 + \frac{\lambda}{\sqrt{k}} \right) \log^2 \left( 1 + \frac{\lambda}{\sqrt{k}} \right).
\]

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 [LNS99, Section 4.3] using Taylor expansion of the KL divergence and orthogonal basis expansion of \( \chi^2 \)-divergence in [CL11, 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.
Theorem 4. Let \( \lambda = \frac{c_1 \log k}{n}, L = c_0 \log k \). Let \( U \) and \( U' \) be random variables on \([0, \lambda]\). If \( E[U_j] = E[U'_j], j = 1, \ldots, L \), then

\[
TV(E[Poi(nU/k)], E[Poi(nU'/k)]) \lesssim k^{-2},
\]

as long as \( c_0^2 \log c_0^2 e^{c_1} - c_1 > 2 \).

Consider the following optimization problem over random variables \( X \) and \( X' \) (or equivalently, the distributions thereof).

\[
E^* = \max E \left[ \log \frac{1}{X} \right] - E \left[ \log \frac{1}{X'} \right]
\]

\[
s.t. E[X_j] = E[X'_j], j = 1, \ldots, L, \]

\[
X, X' \in [\eta, 1],
\]

where \( 0 < \eta < 1 \). Note that (38) is an infinite-dimensional linear programming problem with finitely many constraints. Therefore it is natural to turn to its dual. In Appendix B we show that the maximum \( E^* \) exists and coincides with twice the best \( L_\infty \) approximation error of the log over the interval \([\eta, 1]\) by polynomials of degree \( L \):

\[
E^* = 2E_L(\log, [\eta, 1]).
\]

Let \( X \) and \( X' \) be the maximizer of (38).

Note that the approximation error on the right-hand side is decreasing in the degree \( L \) and increasing in \( \lambda \), due to the singularity of the logarithm function at zero. As shown in Appendix C, the necessary and sufficient condition for the error to be bounded away from zero is \( \lambda = \Omega(L^2) \).

We need the following theorem to prove our lower bound:

Theorem 5. There exist universal positive constants \( c, c', L_0 \) such that for any \( L \geq L_0 \), then

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

Fix \( \alpha \in (0, 1) \) to be specified later. Now construct \( U \) and \( U' \) from \( X \) and \( X' \) with the following distributions

\[
P_U(du) = \left(1 - E \left[ \frac{\eta}{X} \right] \right) \delta_0(du) + \frac{\alpha}{u} P_{\alpha X/\eta}(du),
\]

\[
P_{U'}(du) = \left(1 - E \left[ \frac{\eta}{X'} \right] \right) \delta_0(du) + \frac{\alpha}{u} P_{\alpha X'/\eta}(du).
\]

Since \( X, X' \in [\eta, 1] \) and thus \( E \left[ \frac{\eta}{X} \right], E \left[ \frac{\eta}{X'} \right] \leq 1 \), these distributions are well-defined and \( U, U' \in [0, \alpha \eta^{-1}] \). The following lemma shows that the values of \( E[\phi(U)] \) and \( E[\phi(U')] \) are separated by \( \alpha E^* \), while the moments of \( U \) and \( U' \) are matched up to the \((L + 1)\) order with mean equals to \( \alpha \).

Lemma 2. \( E[\phi(U)] - E[\phi(U')] = \alpha E^* \) and \( E[U_j] = E[U'_j], j = 1, \ldots, L + 1 \). In particular, \( E[U] = E[U'] = \alpha \).

Proof. Note that

\[
E[\phi(U)] = \int \left( u \log \frac{1}{u} \right) \frac{\alpha}{u} P_{\alpha X/\eta}(du) = \alpha E \left[ \log \frac{\eta}{\alpha X} \right]
\]
and, analogously, $E[\phi(U')] = \alpha E[\log \frac{n}{\alpha}]$. Therefore, $E[\phi(U)] - E[\phi(U')] = \alpha E^*.$  

$$E[U^j] = \int u^j \frac{\alpha}{u} P_{\alpha X/\eta}(du) = E[(X/\eta)^{j-1}]$$

which coincides with $E[U^j] = E[(X'/\eta)^{j-1}]$, in view of the moment matching condition of $X$ and $X'$ in (38). In particular, $E[U] = E[U'] = \alpha$ follows immediately.

**Proof of Proposition 2.** Recall the universal constants $c$ and $c'$ defined in Theorem 5. Let $\eta = \log^{-2} k$, $L = [c \log k]$, $\alpha = \frac{c_1 k}{n \log k}$ and $\lambda = \alpha \eta^{-1} = \frac{c_1 k \log k}{n}$. Using (41), we can construct two random variables $U, U' \in [0, \lambda]$ such that $E[U] = E[U'] = \alpha$, $E[U^j] = E[U'^j]$, for all $j \in [L]$, and $E[\phi(U)] - E[\phi(U')] = \alpha E^*$, in view of Lemma 2. It follows from (39) and Theorem 5 that $E^* \geq c'$ and thus $|E[\phi(U)] - E[\phi(U')]| \geq c' \alpha$. Applying Theorem 4 yields $TV(E[\text{Poi}(nU/k)], E[\text{Poi}(nU'/k)]) \leq k^{-2}$ as long as $\frac{4}{k} \log \frac{c_1}{4 \epsilon c_1} - c_1 > 2$. Finally Theorem 3 with $d = c' \alpha$ yields $R^*(k, n/2) \geq \alpha^2 \times \left( \frac{k}{n \log k} \right)^2$ if $c_1 \leq c$, which implies the desired conclusion.

5.3 **Proof of Theorem 3**

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

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

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

Generalizing the minimax quadratic risk (7) for Poisson sampling, we define

$$\tilde{R}^*(k, n, \epsilon) \triangleq \inf_{\tilde{H}^*} \sup_{P \in \mathcal{M}_k(\epsilon)} E[(\tilde{H}^*(N) - H(P))^2],$$

where $N_i \overset{\text{ind}}{\sim} \text{Poi}(n p_i)$ 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 which is related to entropy of the normalized $P$ by

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

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

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

**Proof.** Fix $\delta > 0$. Let $\tilde{H}(\cdot, n)$ be a near-minimax entropy estimator for fixed sample size $n$, i.e.,

$$\sup_{P \in \mathcal{M}_k} E[(\tilde{H}(N, n) - H(P))^2] \leq \delta + R^*(k, n).$$

Using these estimators we construct a estimator for the Poisson model in (8). Fix an arbitrary $P = (p_1, \ldots, p_k) \in \mathcal{M}_k(\epsilon)$. Let $N = (N_1, \ldots, N_k)$ with $N_i \overset{\text{ind}}{\sim} \text{Poi}(n p_i)$ and let $n' = \sum N_i \sim \text{Poi}(n)$. We construct an estimator for the Poisson sampling model by

$$\tilde{H}(N) = \tilde{H}(N, n').$$
Then triangle inequality and (44) give us
\[
\frac{1}{3}(\tilde{H}(N) - H(P))^2 \\
\leq \left( \tilde{H}(N) - H \left( \frac{P}{\sum_i p_i} \right) \right)^2 + \left( \left( 1 - \sum_i p_i \right) H \left( \frac{P}{\sum_i p_i} \right) \right)^2 + \left( \sum_i p_i \log \frac{1}{\sum_i p_i} \right)^2 \\
\leq \left( \tilde{H}(N) - H \left( \frac{P}{\sum_i p_i} \right) \right)^2 + (\epsilon \log k)^2 + ((1 + \epsilon) \log(1 + \epsilon))^2.
\]

For the first term, we observe that conditioned on \( n' = m, N \sim \text{Multinomial} \left( m, \frac{P}{\sum_i p_i} \right) \). Therefore by (45), we have
\[
E \left( \tilde{H}(N) - H \left( \frac{P}{\sum_i p_i} \right) \right)^2 = \sum_{m=0}^{\infty} E \left[ \left( \tilde{H}(N, m) - H \left( \frac{P}{\sum_i p_i} \right) \right)^2 \middle| n' = m \right] \mathbb{P} [ n' = m ] \\
\leq \sum_{m=0}^{\infty} R^*(k, m) \mathbb{P} [ n' = m ] + \delta.
\]

Now note that for fixed \( k \), the minimax risk \( n \mapsto R^*(k, n) \) is decreasing and \( 0 \leq R^*(k, n) \leq (\log k)^2 \). Since \( n' = \sum_{i=1}^{k} N_i \sim \text{Poi} (n \sum_i p_i) \) and \( \left| \sum_i p_i - 1 \right| \leq \epsilon \leq 1/3 \), we have
\[
E \left( \tilde{H}(N) - H \left( \frac{P}{\sum_i p_i} \right) \right)^2 \leq \sum_{m \geq n/2} R^*(k, m) \mathbb{P} [ n' = m ] + (\log k)^2 \mathbb{P} \left[ n' \leq \frac{n}{2} \right] + \delta \\
\leq R^*(k, n/2) + (\log k)^2 \exp(-n/50) + \delta,
\]
where in the last inequality we used the Chernoff bound. By the arbitrariness of \( \delta \), the lemma follows.

**Proof of Theorem 3.** Let \( \alpha \triangleq \mathbb{E} [ U ] = \mathbb{E} [ U' ] \leq 1 \). Define two random vectors
\[
P = \left( \frac{U_1}{k}, \ldots, \frac{U_k}{k}, 1 - \alpha \right), \quad P' = \left( \frac{U'_1}{k}, \ldots, \frac{U'_k}{k}, 1 - \alpha \right),
\]
where \( U_i, U'_i \) are i.i.d. copies of \( U, U' \) respectively. Put \( \epsilon \triangleq \frac{4\alpha}{\sqrt{k}} \geq 4 \sqrt{\frac{\text{var}[U]\text{var}[U']}{k}} \). Define the following events:
\[
E \triangleq \left\{ \left| \sum_i \frac{U_i}{k} - \alpha \right| \leq \epsilon, |H(P) - \mathbb{E} [ H(P) ]| \leq \frac{d}{4} \right\}, E' \triangleq \left\{ \left| \sum_i \frac{U'_i}{k} - \alpha \right| \leq \epsilon, |H(P') - \mathbb{E} [ H(P') ]| \leq \frac{d}{4} \right\}.
\]

Applying Chebyshev’s inequality and the union bound yields that
\[
P [ E^c ] \leq \mathbb{P} \left[ \left| \sum_i \frac{U_i}{k} - \alpha \right| > \epsilon \right] + \mathbb{P} \left[ |H(P) - \mathbb{E} [ H(P) ]| > \frac{d}{4} \right] \\
\leq \frac{\text{var}[U]}{k\epsilon^2} + 16 \sum_i \text{var}[\phi(U_i/k)] \leq \frac{1}{16} + \frac{16\lambda^2 \log^2 \frac{d}{k} \epsilon^2}{kd^2}.
\]
where the last inequality follows from the fact that \( \text{var} \left[ \phi \left( \frac{U_i}{k} \right) \right] \leq \mathbb{E} \left[ \phi \left( \frac{U_i}{k} \right) \right]^2 \leq (\phi \left( \frac{1}{k} \right))^2 \) when \( \lambda < c'k \) and \( c' < e^{-1} \). By the same reasoning,

\[
\Pr \left[ E^c \right] \leq \frac{1}{16} + \frac{16\lambda^2 \log^2 \frac{k}{x}}{kd^2}.
\] (47)

Now we define two priors on the set \( \mathcal{M}_k(e) \) by the following conditional distributions:

\[
\pi = P_{U|E}, \quad \pi' = P_{U'|E'}.
\]

First we consider the separation of the functional value under \( \pi, \pi' \). It follows from \( H(P) = \frac{1}{k} \sum_i \phi(U_i) + \frac{\log k}{k} \sum_i U_i + \phi(1 - \alpha) \) that \( \mathbb{E} [H(P)] = \mathbb{E} [\phi(U)] + \mathbb{E} [U] \log k + \phi(1 - \alpha) \). Similarly, \( \mathbb{E} [H(P')] = \mathbb{E} [\phi(U')] + \mathbb{E} [U'] \log k + \phi(1 - \alpha) \). By the definition of events \( E, E' \) and triangle inequality, we obtain that under \( \pi, \pi' \)

\[
|H(P) - H(P')| \geq \frac{d}{2}.
\] (48)

Now we consider the total variation of observations under \( \pi, \pi' \). Note that the observations \( N_i \sim \text{Poi}(np_i) \). Triangle inequality yields that

\[
\text{TV} \left( P_{N|E}, P_{N'|E'} \right) \leq \text{TV} \left( P_{N|E}, P_N \right) + \text{TV} \left( P_N, P_{N'} \right) + \text{TV} \left( P_{N'}, P_{N'|E'} \right)
\]

\[
= \Pr \left[ E^c \right] + \text{TV} \left( P_N, P_{N'} \right) + \Pr \left[ E^c \right]
\]

\[
\leq \text{TV} \left( P_N, P_{N'} \right) + \frac{1}{8} + \frac{32\lambda^2 \log^2 \frac{k}{x}}{kd^2},
\] (49)

where in the last inequality we apply (46)–(47). Note that \( P_N, P_{N'} \) are marginal distributions under priors \( P_U, P_{U'} \) respectively. From the fact that total variation of product distribution can be upper bounded by the summation of individual ones we obtain

\[
\text{TV} \left( P_N, P_{N'} \right) \leq \sum_{i=1}^{k} \text{TV} \left( P_{N_i}, P_{N_i} \right) + \text{TV}(\text{Poi}(n(1 - \alpha)), \text{Poi}(n(1 - \alpha)))
\]

\[
= k \text{TV}(\mathbb{E} \left[ \text{Poi} \left( nU/k \right) \right], \mathbb{E} \left[ \text{Poi} \left( nU'/k \right) \right]).
\] (50)

It follows from (48)–(50) and Le Cam’s lemma [LC86] that

\[
\tilde{R}^* (k, n, e) \geq \frac{d^2}{32} \left( \frac{7}{8} - k \text{TV}(\mathbb{E} \left[ \text{Poi} \left( nU/k \right) \right], \mathbb{E} \left[ \text{Poi} \left( nU'/k \right) \right]) - \frac{32\lambda^2 \log^2 \frac{k}{x}}{kd^2} \right)
\] (51)

The conclusion comes from (51) and Lemma 3.

\[
\square
\]

### 5.4 Proof of Theorem 4

For any \( u \leq \lambda, nu/k \leq c_1 \log k \) and \( c_1 \leq \frac{c_0}{2} \), Chernoff bound yields that

\[
\Pr \left[ \text{Poi} \left( \frac{nu}{k} \right) \geq \frac{L}{2} \right] \leq \Pr \left[ \text{Poi} \left( c_1 \log k \right) \geq \frac{c_0 \log k}{2} \right] \leq k^{- \left( c_1 + \frac{c_0}{2} \log \frac{c_0}{c_1} \right)} \leq k^{-2},
\] (52)
Similarly, $\sum_{j=\lfloor L/2 \rfloor}^{\infty} \mathbb{E} [\text{poi} (nU/k, j)] = \int \mathbb{P} \left[ \text{Poi} \left( \frac{nU}{k} \right) \geq \frac{L}{2} \right] \, du \leq k^{-2}$.

For $j \leq \lfloor L/2 \rfloor$, Taylor's expansion and moments matching gives

$$\sum_{j=0}^{\lfloor L/2 \rfloor} \left| \mathbb{E} [\text{poi}(nU/k, j)] - \mathbb{E} [\text{poi}(nU'/k, j)] \right|$$

$$= \sum_{j=0}^{\lfloor L/2 \rfloor} \left| \mathbb{E} \left[ (nU/k)^j \right] \mathbb{E} \left[ \frac{(-nU/k)^m}{m!} \right] - \mathbb{E} \left[ \frac{(nU'/k)^j}{j!} \mathbb{E} \left[ \frac{(-nU'/k)^m}{m!} \right] \right] \right|$$

$$\leq \sum_{j=0}^{\lfloor L/2 \rfloor} \frac{1}{j!} \sum_{m=L-j+1}^{\infty} \frac{1}{m!} 2(c_1 \log k)^{m+j}$$

$$\leq 2e^{c_1 \log k} \sum_{j=0}^{\infty} \frac{(c_1 \log k)^j}{j!} \sum_{m=L/2}^{\infty} \frac{(c_1 \log k)^m e^{-c_1 \log k}}{m!}.$$

Again the Poisson tail bound (52) yields that

$$\sum_{j=0}^{\lfloor L/2 \rfloor} \left| \mathbb{E} [\text{poi}(nU/k, j)] - \mathbb{E} [\text{poi}(nU'/k, j)] \right|$$

$$\leq 2k^{c_1} k^{c_1} k^{-\left( c_1 + \frac{c_0}{2} \log \frac{c_0}{2c_1} \right)} = 2k^{-\left( \frac{c_0}{4} \log \frac{c_0}{2c_1} - c_1 \right)} \leq 2k^{-2}, \quad (53)$$

if we choose $c_0, c_1$ such that $\frac{c_0}{2} \log \frac{c_0}{2c_1} - c_1 > 2$, which, in particular, ensures that (52) holds.

Combining (52)–(53), we obtain

$$\text{TV} (\mathbb{E} [\text{Poi} (nU/k)], \mathbb{E} [\text{Poi} (nU'/k)])$$

$$= \frac{1}{2} \sum_{j=0}^{\infty} \left| \mathbb{E} [\text{poi}(nU/k, j)] - \mathbb{E} [\text{poi}(nU'/k, j)] \right| \leq k^{-2}.$$

### A Non-asymptotic risk bounds for the plug-in estimator

Recall the worst-case quadratic risk of the plug-in estimator $R_{\text{plug-in}}(k, n)$ defined in (4). We show that for any $k \geq 2$ and $n \geq 2$,

$$\left( \frac{k}{n} \wedge 1 \right)^2 + \log^2 k \leq R_{\text{plug-in}}(k, n) \leq \left( \frac{k}{n} \right)^2 + \log^2 n \quad (54)$$

The upper bound of MSE follows from the upper bounds of bias and variance. From [AK01, Remark (iv), p. 168] we know that $\text{var}(\hat{H}_{\text{plug-in}}(N)) \leq \frac{C \log^2 n}{n}$ for some universal constant $C$. The squared bias can be upper bounded by $(\frac{k-1}{n})^2$ according to [Pan03, Proposition 1].
The second term of the lower bound follows from the minimax lower bound Proposition 1 which applies to all \( k \) and \( n \). To prove the first term of 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

\[
\mathbb{E}(|\hat{H}_{\text{plug-in}}(N) - H|) = - \mathbb{E}[D(\hat{P}||P)] \leq -2\mathbb{E}[(TV(\hat{P}, P))^2] \\
\leq -2(\mathbb{E}[TV(\hat{P}, P)])^2 = -2 \left( \frac{k}{2n} \mathbb{E} \left| N_1 - \frac{n}{k} \right| \right)^2,
\]

where \( N_1 \sim \text{Binomial} \left( n, \frac{1}{k} \right) \).

From [BK13, 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 \( \mathbb{E} \left| N_1 - \frac{n}{k} \right| \geq \sqrt{\frac{n}{2k}} \left( 1 - \frac{1}{k} \right) \) when \( n \geq k \). Therefore

\[
- \mathbb{E}(\hat{H}_{\text{plug-in}}(N) - H) \geq 2 \left( 1 - \frac{1}{k} \right)^{2n} \gtrsim 1, \quad n < k, \tag{55}
\]

\[
- \mathbb{E}(\hat{H}_{\text{plug-in}}(N) - H) \geq \frac{k}{4n} \left( 1 - \frac{1}{k} \right) \gtrsim \frac{k}{n}, \quad n \geq k. \tag{56}
\]

From (55)–(56) we conclude that

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

### B Moment matching and best polynomial approximation

In this appendix we discuss the relationship between moment matching and best polynomial approximation and, in particular, provide a short proof of (39). Abbreviate by \( \hat{E}^* \) the best uniform approximation error \( E_L(\log, [\eta, 1]) = \inf_{p \in \mathcal{P}_L} \sup_{x \in [\eta, 1]} |\log \frac{1}{x} - p(x)| \).

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

\[
\mathcal{E}^* = \sup_{(X, X') \in \mathcal{S}_L} \mathbb{E} \left[ \log \frac{1}{X} - \mathbb{E} \left[ \log \frac{1}{X} \right] \right] = \sup_{(X, X') \in \mathcal{S}_L} \mathbb{E} \left[ \log \frac{1}{X} - p(X) \right] - \mathbb{E} \left[ \log \frac{1}{X'} - p(X') \right],
\]

and therefore

\[
\mathcal{E}^* = \inf_{p \in \mathcal{P}_L} \sup_{(X, X') \in \mathcal{S}_L} \mathbb{E} \left[ \log \frac{1}{X} - p(X) \right] - \mathbb{E} \left[ \log \frac{1}{X'} - p(X') \right] \leq 2 \inf_{p \in \mathcal{P}_L} \sup_{x \in [\eta, 1]} \left| \log \frac{1}{x} - p(x) \right| = 2E_L(\log, [\eta, 1]).
\]

For the achievability part, Chebyshev alternating theorem [PP11, Theorem 1.6] states that there exists a (unique) polynomial \( p^* \in \mathcal{P}_L \) and at least \( L + 2 \) points \( x_1 < \cdots < x_{L+2} \) and \( \alpha \in \{0, 1\} \) such that

\[
\log \frac{1}{x_i} - p^*(x_i) = (-1)^{i+\alpha} \hat{E}^*.
\]

Let \( b_i = \left[ \prod_{v \neq i} (x_i - x_v) \right]^{-1} \). For any \( l \in \{0, 1, \ldots, L\} \), define a Lagrange interpolate polynomial \( f(x) \triangleq \sum_{j=1}^{L+2} x_j^l \prod_{v \neq j} (x - x_v) \), which satisfies \( f(x_j) = x_j^l \).
for \( j = 1, \ldots, L + 2 \). Since \( f \) has degree \( L \), it must be that \( f(x) = x^j \), and \( \sum_i x_i^j b_i = 0 \) from the coefficient of \( x^{L+1} \). Define \( w_i = \frac{2b_i}{\sum_j |w_j|} \), then \( \sum_i w_i = 0 \) and \( \sum_i |w_i| = 2 \) with alternating signs.

Construct \( X, X' \) with distributions \( \mathbb{P} [X = x_i] = |w_i| \) for \( i \) odd and \( \mathbb{P} [X' = x_i] = |w_i| \) for \( i \) even. Then \( (X, X') \in S_L \) and \( \mathbb{E} \left[ \log \frac{1}{X} - p^*(X) \right] - \mathbb{E} \left[ \log \frac{1}{X'} - p^*(X') \right] = 2\mathcal{E}^* \).

**Remark 4.** From the above achievability argument, we can see that \( X, X' \) are actually discrete random variables and so are \( U, U' \). Alternatively, the achievability can be argued from an optimization perspective (zero duality gap, see [Luc69, Exercise 8.8.7, p. 236]), or using the standard Riesz representation of linear operators as in [DL93], which is repeated in [LNS99] and [CL11].

### C Best polynomial approximation of the logarithm function

**Proof of Theorem 5.** Recall the best uniform polynomial approximation error \( E_m(f, I) \) defined in (15). Put \( E_m(f) \equiv E_m(f, [-1, 1]) \). 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

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

Let \( \Delta_m(x) = \frac{1}{m} \sqrt{1 - x^2} + \frac{1}{m^2} \) and define the following modulus of continuity for \( f \):

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

Initiated from Jackson’s theorem, there are a lot of characterization of approximation error in term of refined modulus of continuity. We choose \( \tau_1 \) from [PP11, 3.4] for our proof. We first state the following two lemmas for \( \tau_1 \):

**Lemma 4** (Direct bound).

\[
\tau_1(f_L, \Delta_m) \leq \log \left( \frac{2L^2}{m^2} \right), \quad \forall m \leq 0.1L.
\] (57)

**Lemma 5** (Converse bound).

\[
\tau_1(f_L, \Delta_L) \geq 1, \forall L \geq 10.
\] (58)

From [PP11, Theorem 3.13, Lemma 3.1] we know that \( E_m(f_L) \leq 100\tau_1(f_L, \Delta_m) \). Therefore, for all \( c \leq 10^{-7} < 0.1 \), the direct bound in Lemma 4 gives us

\[
\frac{1}{L} \sum_{m=1}^{cL} E_m(f_L) \leq \frac{100}{L} \sum_{m=1}^{cL} \log \left( \frac{2L^2}{m^2} \right) = 100c \log 2 + \frac{200}{L} \log \frac{L^c}{(cL)!} < \frac{1}{400} - \frac{100}{L} \log(2\pi cL),
\] (59)

where in the last inequality we apply Stirling’s approximation \( n! > \sqrt{2\pi n}(n/e)^n \).

[PP11, Theorem 3.14] yields that \( \tau_1(f_L, \Delta_L) \leq \frac{100}{L} \sum_{m=0}^{L} E_m(f_L) \). We reorganize it and apply the fact that \( E_0(f_L) = \log L \). For all \( c \leq 10^{-7} \) and \( L > 10 \sqrt{(100 \times 400 \log \frac{1}{2\pi c})} \), we apply (59) and Lemma 5 and obtain

\[
\frac{1}{L} \sum_{m=cL+1}^{L} E_m(f_L) \geq \frac{1}{100} - \left( \frac{1}{L} E_0(f_L) + \frac{1}{L} \sum_{m=1}^{cL} E_m(f_L) \right) \geq \frac{1}{100} - \left( \frac{1}{400} + \frac{100 \log \frac{1}{2\pi c}}{L} \right) > \frac{1}{200}.
\]
By definition, the approximation error \( E_m(f_L) \) is a decreasing function of the degree \( m \). Therefore for all \( c \leq 10^{-7} \) and \( L > 4 \times 10^4 \log \frac{1}{2m} \),

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

**Remark 5.** From the direct bounds we know that \( E_{cL}(\log[1/L^2,1]) \leq 1 \). Therefore the bound (41) is in fact tight: \( E_{cL}(\log[1/L^2,1]) \times 1 \).

**Proof of Lemma 4.** First we observe the important equivalence that

\[
\{ x \in [-1,1] : x - \Delta_m(x) < -1 \} \equiv \{ x \in [-1,x_m] \},
\]

\[
\{ x \in [-1,1] : x - \Delta_m(x) > -1 \} \equiv \{ x \in (x_m,1] \},
\]

where \( x_m \in [-1,1] \) is the solution to \( x_m - \Delta_m(x_m) = -1 \). In fact, \( x_m \) has the close-form expression

\[
x_m = \frac{m^2 - m^4 + \sqrt{-m^2 + 3m^4}}{m^2 + m^4}.
\]

Since \( f_L \) is a decreasing and convex function, we can decompose the supremum and it turns out

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

\[
\leq \sup_{x \in [-1,x_m]} \{ f_L(x) - f_L(x + \Delta_m(x)) \} \vee \sup_{x \in [-1,x_m]} \{ f_L(-1) - f_L(x) \} \vee \sup_{x \in [x_m,1]} \{ f_L(x - \Delta_m(x)) - f_L(x) \}.
\]

Note that the second term in the last inequality is superfluous since it is dominated by the third term by \( f_L(x_m - \Delta_m(x_m)) - f_L(x_m) = f_L(-1) - f_L(x_m) > f_L(-1) - f_L(x) \) for any \( x \in [-1,x_m] \). Hence

\[
\tau_1(f_L, \Delta_m) \leq \sup_{x \in [-1,x_m]} \{ f_L(x) - f_L(x + \Delta_m(x)) \} \vee \sup_{x \in [x_m,1]} \{ f_L(x - \Delta_m(x)) - f_L(x) \}
\]

\[
= \sup_{x \in [-1,x_m]} \{ \log (1 + \beta_L(x)) \} \vee \sup_{x \in [x_m,1]} \{ - \log (1 - \beta_L(x)) \},
\]

where \( \beta_L(x) \triangleq \frac{\Delta_m(x)}{x + \frac{L^2}{x^2-1}} \). Next we will show separately that the two terms in (62) both satisfy the desired upper bound.

For the first term in (62), we can see that

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

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

Therefore

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

Consequently,

\[
\sup_{x \in [-1,x_m]} \{ \log (1 + \beta_L(x)) \} \leq \log \left( \frac{2L^2}{m^2} \right), \forall m \leq L.
\]
For the second term in (62), it follows from the derivative of $\beta_L(x)$ that it is decreasing when $x > \frac{1-L^2}{1+L^2}$. From (61) we can see that $x_m > \frac{1-m^2}{1+L^2}$, therefore $x_m > \frac{1-L^2}{1+L^2}$ when $m \leq L$. So the supremum is achieved exactly the left end of $[x_m, 1]$, that is:

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

From (61) we know that $x_m \geq -1$ and $x_m < -1 + \frac{3L^2}{m^2}$. Therefore $\frac{1-x_m}{2} \leq 1$ and $\frac{x_m+1}{2} < \frac{1}{m^2}$. For $m \leq 0.1L$, we have

$$\sup_{x \in [x_m, 1]} \{- \log (1 - \beta_L(x))\} \leq \log \left( 1 + \frac{1.9m^2}{L^2} \right) \leq \log \left( \frac{2m^2}{L^2} \right). \quad (64)$$

Plugging (63) and (64) into (62), we complete the proof.

**Proof of Lemma 5.** We still use the notations in the proof of Lemma 4 since it is closely related. Recall that $x_L - \Delta_L(x_L) = -1$. By definition,

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

Note that $x_L$ has close form expression in (61) with $m = L$. We further obtain

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

when $L \geq 10$.

**Acknowledgement**

It is a pleasure to thank Yury Polyiansky for pointing out and inspiring discussions in the early stage of the project, Tony Cai and Mark Low for many conversations on their result [CL11], and R. Srikant for the pointers in [Lue69].

**References**

[AK01] András Antos and Ioannis Kontoyiannis. Convergence properties of functional estimates for discrete distributions. *Random Structures & Algorithms*, 19(3-4):163–193, 2001.

[Bas59] G.P. Basharin. On a statistical estimate for the entropy of a sequence of independent random variables. *Theory of Probability & Its Applications*, 4(3):333–336, 1959.

[BK13] Daniel Berend and Aryeh Kontorovich. A sharp estimate of the binomial mean absolute deviation with applications. *Statistics & Probability Letters*, 83(4):1254–1259, 2013.

[CK82] Imre Csiszár and János Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Academic Press, Inc., 1982.
[CL05] T. T. Cai and M. G. Low. Nonquadratic estimators of a quadratic functional. *The Annals of Statistics*, 33(6):2930–2956, 2005.

[CL11] T.T. Cai and M. G. Low. Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional. *The Annals of Statistics*, 39(2):1012–1041, 2011.

[CMW13] T.T. Cai, Zongming Ma, and Yihong Wu. Sparse PCA: Optimal rates and adaptive estimation. *The Annals of Statistics*, 41(6):3074–3110, 2013.

[DL91] David L Donoho and Richard C Liu. Geometrizing rates of convergence, ii. *The Annals of Statistics*, 19:668–701, 1991.

[DL93] Ronald A. DeVore and George G. Lorentz. *Constructive approximation*. Springer, 1993.

[Har75] B. Harris. The statistical estimation of entropy in the non-parametric case. In I Csiszár and P. Elias, editors, *Topics in Information Theory*, volume 16, pages 323–355. Springer Netherlands, 1975.

[INK87] I.A. Ibragimov, A.S. Nemirovskii, and R.Z. Khas’minskii. Some problems on nonparametric estimation in gaussian white noise. *Theory of Probability & Its Applications*, 31(3):391–406, 1987.

[LC86] L. Le Cam. *Asymptotic methods in statistical decision theory*. Springer-Verlag, New York, NY, 1986.

[LNS99] Oleg Lepski, Arkady Nemirovski, and Vladimir Spokoiny. On estimation of the $L_p$ norm of a regression function. *Probability theory and related fields*, 113(2):221–253, 1999.

[Lue69] David G. Luenberger. *Optimization by vector space methods*. John Wiley & Sons, 1969.

[Mil55] George A. Miller. Note on the bias of information estimates. *Information theory in psychology: Problems and methods*, 2:95–100, 1955.

[MU05] Michael Mitzenmacher and Eli Upfal. *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge University Press, 2005.

[OSZ04] Alon Orlitsky, Narayana P. Santhanam, and Junan Zhang. Universal compression of memoryless sources over unknown alphabets. 50(7):1469–1481, 2004.

[Pan03] Liam Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253, 2003.

[Pan04] Liam Paninski. Estimating entropy on $m$ bins given fewer than $m$ samples. *IEEE Transactions on Information Theory*, 50(9):2200–2203, 2004.

[PGM+01] 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*, 48(11):1282–1291, 2001.
[PP11] Penco Petrov Petrushev and Vasil Atanasov Popov. *Rational approximation of real functions*. Cambridge University Press, 2011.

[PW96] Nina T. Plotkin and Abraham J. Wyner. An entropy estimator algorithm and telecommunications applications. In *Maximum Entropy and Bayesian Methods*, volume 62 of *Fundamental Theories of Physics*, pages 351–363. Springer Netherlands, 1996.

[RBWvS99] Fred Rieke, William Bialek, David Warland, and Rob de Ruyter van Steveninck. Spikes: Exploring the neural code. 1999.

[Sto80] Charles J. Stone. Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, 8(6):1348–1360, 1980.

[Sze75] G. Szegő. *Orthogonal polynomials*. American Mathematical Society, Providence, RI, 4th edition, 1975.

[Tim63] Aleksandr Filippovich Timan. *Theory of approximation of functions of a real variable*. Pergamon Press, 1963.

[Tsy09] A.B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Verlag, New York, NY, 2009.

[Val08] Paul Valiant. Testing symmetric properties of distributions. In *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, STOC ’08, 2008.

[VBB+12] Martin Vinck, Francesco P. Battaglia, Vladimir B. Balakirsky, A.J. Han Vinck, and Cyriel M.A. Pennartz. Estimation of the entropy based on its polynomial representation. *Physical Review E*, 85(5):051139, 2012.

[VV10] Gregory Valiant and Paul Valiant. A CLT and tight lower bounds for estimating entropy. In *Electronic Colloquium on Computational Complexity (ECCC)*, volume 17, page 179, 2010.

[VV11] Gregory Valiant and Paul Valiant. Estimating the unseen: an $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs. In *Proceedings of the 43rd annual ACM symposium on Theory of computing*, pages 685–694, 2011.

[WKV09] Qing Wang, Sanjeev R Kulkarni, and Sergio Verdú. Universal estimation of information measures for analog sources. *Foundations and Trends in Communications and Information Theory*, 5(3):265–353, 2009.