Estimation of KL Divergence: Optimal Minimax Rate

Yuheng Bu*  Shaofeng Zou* Yingbin Liang† Venugopal V. Veeravalli*

* University of Illinois at Urbana-Champaign  † Syracuse University

Email: bu3@illinois.edu, szou3@illinois.edu, yliang06@syr.edu, vvv@illinois.edu

Abstract

The problem of estimating the Kullback-Leibler divergence between two unknown distributions is studied. The alphabet size $k$ of the distributions can scale to infinity. The estimation is based on $m$ and $n$ independent samples respectively drawn from the two distributions. It is first shown that there does not exist any consistent estimator to guarantee asymptotically small worst-case quadratic risk over the set of all pairs of distributions. A restricted set that contains pairs of distributions with bounded density ratio $f(k)$ is further considered. An augmented plug-in estimator is proposed, and is shown to be consistent if and only if $m = \omega(k \sqrt{\log k})$ and $n = \omega(k f(k))$. Moreover, the minimax quadratic risk is characterized to be within a constant factor of $(\frac{k^2}{m \log k} + \frac{k f(k)}{n \log k})^2 + \frac{\log^2 f(k)}{n \log k}$ if $f(k) \geq \log^2 k$,

log $m \leq C \log k$ and $\log^2 n \leq k$. A minimax optimal estimator is constructed by jointly employing polynomial approximation and plug-in approach. The lower bound on the minimax quadratic risk is characterized by employing the generalized Le Cam’s method.

1 Introduction

As an important concept in information theory, Kullback-Leibler (KL) divergence has a wide range of applications in various domains. Since KL divergence measures the difference between two distributions, it is instinctively useful in many classification and pattern recognition problems. For example, KL divergence can be used as a similarity measure in nonparametric outlier detection [1], multimedia classification [2], text classification [3], and the two-sample problem [4]. In these contexts, it is often desired to estimate KL divergence efficiently based on data samples available. This paper studies such a problem.

More formally, consider estimation of KL divergence between the probability distributions $P$ and $Q$ defined as

$$D(P\|Q) = \sum_{i=1}^{k} P_i \log \frac{P_i}{Q_i},$$

(1)

where $P$ and $Q$ are over a common alphabet set $[k] \triangleq \{1, \ldots, k\}$, and $P$ is absolutely continuous with respect to $Q$, i.e., if $Q_i = 0$, $P_i = 0$, for $1 \leq i \leq k$. We use $M_k$ to denote the collection of all such pairs of distributions.

Suppose $P$ and $Q$ are unknown. Instead, $m$ independent and identically distributed (i.i.d.) samples $X_1, \ldots, X_m$ drawn from $P$ and $n$ i.i.d. samples $Y_1, \ldots, Y_n$ drawn from $Q$ are available for estimation.

This work will be presented in part at the International Symposium on Information Theory (ISIT), July 2016.

The first two student authors have contributed equally to this work.

The work of Y. Bu and V. V. Veeravalli was supported by the Air Force Office of Scientific Research under Grant FA9550-10-1-0458, and by the National Science Foundation under the Grant NSF 11-11342, through the University of Illinois at Urbana-Champaign. The work of S. Zou and Y. Liang was supported by the Air Force Office of Scientific Research under Grant FA9550-16-1-0077 and by an National Science Foundation CAREER Award under Grant CCF-10-26565.

We adopt the following notations to express asymptotic scaling of quantities with $n$: $f(n) \leq g(n)$ represents that there exists a constant $c$ s.t. $f(n) \leq cg(n)$; $f(n) \geq g(n)$ represents that there exists a constant $c$ s.t. $f(n) \geq cg(n)$; $f(n) \asymp g(n)$ when $f(n) \geq \omega(g(n))$ and $f(n) \leq \omega(g(n))$ hold simultaneously; $f(n) = \omega(g(n))$ represents that for all $c > 0$, there exists $n_0 > 0$ s.t. for all $n > n_0$, $|f(n)| \geq |c|g(n)$; $f(n) = \Theta(g(n))$ represents that for all $c > 0$, there exists $n_0 > 0$ s.t. for all $n > n_0$, $|f(n)| \leq c|g(n)|$; and $f(n) = o(g(n))$ represents that for all $c > 0$, there exists $n_0 > 0$ s.t. for all $n > n_0$, $|f(n)| \leq cg(n)$.
The sufficient statistics for estimating $D(P\|Q)$ are the histograms of the samples $M \triangleq (M_1, \ldots, M_k)$ and $N \triangleq (N_1, \ldots, N_k)$, where
\begin{equation}
M_j = \sum_{i=1}^{m} 1_{\{X_i = j\}} \quad \text{and} \quad N_j = \sum_{i=1}^{n} 1_{\{Y_i = j\}}
\end{equation}
record the numbers of occurrences of $j \in [k]$ in samples drawn from $P$ and $Q$, respectively. Then $M \sim \text{Multinomial}(m, P)$ and $N \sim \text{Multinomial}(n, Q)$. An estimator $\hat{D}$ of $D(P\|Q)$ is then a function of the histograms $M$ and $N$, denoted by $\hat{D}(M, N)$.

We adopt the following worst-case quadratic risk to measure the performance of estimators of the KL divergence:
\begin{equation}
R(\hat{D}, k, m, n) \triangleq \sup_{(P,Q) \in \mathcal{M}_k} \mathbb{E}[(\hat{D}(M,N) - D(P\|Q))^2].
\end{equation}

In this paper, we are interested in the large-alphabet regime with $k \to \infty$. And, the number $m$ and $n$ of samples are functions of $k$, which can scale with $k$ to infinity.

**Definition 1.** A sequence of estimators $\hat{D}$ is said to be consistent under sample complexity $m(k)$ and $n(k)$ if
\begin{equation}
\lim_{k \to \infty} R(\hat{D}, k, m, n) = 0.
\end{equation}

We further define the minimax quadratic risk as:
\begin{equation}
R^*(k, m, n) \triangleq \inf_{\hat{D}} R(\hat{D}, k, m, n).
\end{equation}

We are also interested in the following set:
\begin{equation}
\mathcal{M}_k,f(k) = \left\{ (P,Q) : |P| = |Q| = k, \frac{P_i}{Q_i} \leq f(k), \forall 1 \leq i \leq k \right\},
\end{equation}
which contains distributions $(P,Q)$ with bounded density ratio $f(k)$. We define the worst-case quadratic risk over $\mathcal{M}_k,f(k)$ as
\begin{equation}
R(\hat{D}, k, m, n, f(k)) \triangleq \sup_{(P,Q) \in \mathcal{M}_k,f(k)} \mathbb{E}[(\hat{D}(M,N) - D(P\|Q))^2],
\end{equation}
and define the corresponding minimax quadratic risk as
\begin{equation}
R^*(k, m, n, f(k)) \triangleq \inf_{\hat{D}} R(\hat{D}, k, m, n, f(k)).
\end{equation}

1.1 Comparison to Related Problems

Several estimators of KL divergence when $P$ and $Q$ are continuous have been proposed and shown to be consistent. The estimator proposed in [5] is based on data-dependent partition for density estimation, the estimator proposed in [6] is based on a k-nearest neighbor approach for density estimation, and the estimator developed in [7] utilizes a kernel-based approach for estimating the density ratio. A more general problem of estimating the $f$-divergence was studied in [8], where an estimator based on a weighted ensemble of plug-in estimators was proposed to trade bias with variance. All these approaches exploit the smoothness of continuous densities or density ratios, which guarantees that samples falling into a certain neighborhood area can be used to estimate the local density or density ratio accurately. However, such a smoothness property does not hold for discrete distributions, whose probabilities over adjacent point masses can vary significantly. In fact, [5] provides an example to show that estimation of KL divergence can be difficult even for continuous distributions if the density has sharp dips.

Estimation of KL divergence when the distributions $P$ and $Q$ are discrete has been studied in [9][10] for the regime with fixed alphabet cardinality $k$ and large sample sizes $m$ and $n$. Such a regime is very different from
the large-alphabet regime in which we are interested, with \( k \) scaling to infinity. Clearly, as \( k \) increases, the scaling of the sample sizes \( m \) and \( n \) must be fast enough with respect to \( k \) in order to guarantee consistent estimation.

In the large-alphabet regime, KL divergence estimation is closely related to the entropy estimation with a large alphabet recently studied in [12–15]. Compared to entropy estimation, KL divergence estimation has one more dimension of uncertainty, that about the distribution \( Q \). Some distributions \( Q \) can contain very small point masses that contribute significantly to the value of divergence, but are difficult to estimate because samples of these point masses occur rarely. In fact, such distributions dominate the risk in (3), and make the construction of consistent estimators challenging.

### 1.2 Summary of Main Results

We summarize the main results in the following three theorems. We further discuss the technical implication and describe the proof arguments of these theorems in detail respectively in Sections 2, 3 and 4.

We first show in the following theorem, using Le Cam’s two-point method [16], that there is no consistent estimator of KL divergence over the distribution set \( M_k \).

**Theorem 1.** For any \( m, n \in \mathbb{N} \), and \( k \geq 2 \), \( R^*(k, m, n) \) is infinite. Therefore, there does not exist any consistent estimator of KL divergence over the set \( M_k \).

The above result is mainly due to the fact that the set \( M_k \) contains distributions \( Q \), which have arbitrarily small components that contribute significantly to KL divergence but require arbitrarily large number of samples to estimate accurately. However, practical applications are typically captured by better behaved distributions, for example, the ratio of \( P \) to \( Q \) is bounded. Thus, we further focus on a set \( M_{k,f(k)} \) given in (6) that contains distribution pairs \((P, Q)\) with their density ratio bounded by \( f(k) \).

We first construct an augmented plug-in estimator and characterize the sufficient conditions on the sample complexity such that such an estimator is consistent in the following theorem.

**Theorem 2.** The augmented plug-in estimator of KL divergence is consistent over the set \( M_{k,f(k)} \) if and only if

\[
m = \omega(k \lor \log^2 f(k)) \quad \text{and} \quad n = \omega(kf(k)).
\]

Our proof of the sufficient conditions is based on evaluating the bias and variance separately. Our proof of the necessary condition \( m = \omega(\log^2 f(k)) \) is based on Le Cam’s two-point method with a judiciously chosen pair of distributions. And our proof of the necessary conditions \( m = \omega(k) \) and \( n = \omega(kf(k)) \) is based on analyzing the bias of the estimator and constructing different pairs of “worst case” distributions for the cases when either the bias caused by insufficient samples from \( P \) or the bias caused by insufficient samples from \( Q \) dominates, respectively.

The above result suggests that the required samples \( m \) and \( n \) should be larger than the alphabet size \( k \) for the plug-in estimator to be consistent (due to the fact that \( f(k) \geq 1 \)). This naturally inspires further exploration of whether the plug-in estimator achieves the minimax risk, and if not, what estimator is minimax optimal and what is the corresponding minimax risk.

We thus further show that an estimator that jointly employs the polynomial approximation and plug-in approach is minimax optimal, and the following theorem characterizes the minimax risk.

**Theorem 3.** If \( f(k) \geq \log^2 k \), \( \log m \leq C \log k \), \( \log^2 n \lesssim k \), \( m \geq \frac{k}{\log k} \) and \( n \geq \frac{kf(k)}{\log k} \), then the minimax risk satisfies

\[
R^*(k, m, n, f(k)) \simeq \left( \frac{k}{m \log k} + \frac{kf(k)}{n \log k} \right)^2 + \frac{\log^2 f(k)}{m} + \frac{f(k)}{n}.
\]
The idea to construct the minimax optimal estimator is to apply polynomial approximation to reduce the bias in the regime where the bias of the plug-in estimator is large. Compared to entropy estimation [13,15], the challenge here is that the KL divergence is a function of two variables, for which the polynomial approximation is very hard to derive. We solve this problem by employing separate polynomial approximations for functions involving \( P \) and \( Q \) as well as judiciously using the density ratio constraint to bound the estimation error. The proof of the lower bound on the minimax risk is based on the generalized Le Cam’s method involving two composite hypotheses as for entropy estimation [13]. But the challenge here that requires special technical treatment is to construct prior distributions for \((P, Q)\) composite hypotheses as for entropy estimation [13]. However, in practice, it is often the case that a smaller set of distributions are of interest. Thus, our second main result Theorem 2 focuses on the set \( \mathcal{M}_{k,f(k)} \) and studies an augmented plug-in estimator.

We note that the first term \( \left( \frac{k}{m \log k} + \frac{k f(k)}{m \log k} \right)^2 \) in (10) capture the squared bias, and the remaining terms correspond to the variance. If we compare the upper bound on the risk in (14) for the augmented plug-in estimator with the minimax risk in (10), there is a \( \log k \) factor rate improvement in the bias.

Theorem 3 directly implies that in order to estimate the KL divergence over the set \( \mathcal{M}_{k,f(k)} \) with vanishing mean square error, the sufficient and necessary conditions on the sample complexity is given by

\[
 m = \omega((\log^2 f(k) + k f(k)), \quad n = \omega(\frac{k f(k)}{\log \log k}).
\]

(11)

The comparison of (11) with (10) suggests that the augmented plug-in estimator is strictly sub-optimal.

We note that after our initial conference submission was accepted by ISIT 2016 and during our preparation of this full version of the work, an independent study of the same problem of KL divergence estimation was posted on arXiv [17]. While [17] establishes the results similar to those in this paper, our technical conditions for the minimax rate to hold in Theorem 3 appear to be more relaxed than those in [17].

2 No Consistent Estimator over \( \mathcal{M}_k \)

Theorem 1 states that the minimax risk over the set \( \mathcal{M}_k \) is unbounded for arbitrary alphabet size \( k \) and \( m \) and \( n \) samples, which suggests that there is no consistent estimator for the minimax risk over \( \mathcal{M}_k \).

The proof of Theorem 1 is provided in Appendix A. The idea follows from Le Cam’s two-point method [16]: If two pairs of distributions \((P_1, Q_1)\) and \((P_2, Q_2)\) are sufficiently close such that it is impossible to reliably distinguish between them using \( m \) samples from \( P \) and \( n \) samples from \( Q \) with error probability less than some constant, then any estimator suffers a quadratic risk proportional to the difference between the divergence values \( D(P_1||Q_1) - D(P_2||Q_2) \).

We next give an example for binary distributions, i.e., \( k = 2 \), to illustrate how distributions in the proof can be constructed. We let \( P_1 = P_2 = (\frac{1}{2}, \frac{1}{2}) \), \( Q_1 = (e^{-s}, 1 - e^{-s}) \) and \( Q_2 = (\frac{1}{2}, 1 - \frac{1}{2}) \), where \( s > 0 \). For any \( n \in \mathbb{N} \), choose \( s \) sufficiently large such that \( D(Q_1||Q_2) < \frac{1}{s} \). Thus, the error probability of distinguishing \( Q_1 \) and \( Q_2 \) with \( n \) samples is greater than a constant. However, \( D(P_1||Q_1) \sim s \) and \( D(P_2||Q_2) \sim \log s \). Hence, the minimax risk, which is lower bounded by the difference of the above divergences, can be made arbitrarily large by letting \( s \to \infty \). Such an example demonstrates that two pairs of distributions \((P_1, Q_1)\) and \((P_2, Q_2)\) can be very close so that the data samples are almost indistinguishable, but KL divergences \( D(P_1||Q_1) \) and \( D(P_2||Q_2) \) can still be far away. In such a case, it is not possible to estimate KL divergence accurately based on the samples.

3 Augmented Plug-in Estimator over \( \mathcal{M}_{k,f(k)} \)

As we have shown in Section 2 there does not exist any consistent estimator of KL divergence over the set \( \mathcal{M}_k \). However, in practice, it is often the case that a smaller set of distributions are of interest. Thus, our second main result Theorem 2 focuses on the set \( \mathcal{M}_{k,f(k)} \), and studies an augmented plug-in estimator.
In fact, the “plug-in” approach is a natural idea to estimate the KL divergence, namely, first estimate the distributions and then substitute these estimates into the divergence function. This leads to the following plug-in estimator, i.e., the empirical divergence

\[
\hat{D}_{\text{plug-in}}(M, N) = D(\hat{P}\|\hat{Q}),
\]  

(12)

where \(\hat{P} = (\hat{P}_1, \ldots, \hat{P}_k)\) and \(\hat{Q} = (\hat{Q}_1, \ldots, \hat{Q}_k)\) denote the empirical distributions with \(\hat{P}_i = \frac{M_i}{m}\) and \(\hat{Q}_i = \frac{N_i}{n}\), respectively.

Unlike the entropy estimation problem, where the plug-in estimator \(\hat{H}_{\text{plug-in}}\) is asymptotically efficient in the “fixed \(P\) large \(n\)” regime, the direct plug-in estimator \(\hat{D}_{\text{plug-in}}\) in (12) of KL divergence has an infinite bias. This is because of the non-zero probability of \(N_j = 0\) and \(M_j \neq 0\) for some \(j \in [k]\), which leads to infinite \(\hat{D}_{\text{plug-in}}\).

We can get around the above issue associated with the direct plug-in estimator, if we add one more sample to each mass point of \(Q\), and take \(\hat{Q}'_i = \frac{N_i + 1}{n}\) as an estimate of \(Q_i\) so that \(\hat{Q}'_i\) is non-zero for all \(i\). We therefore propose the following “augmented plug-in” estimator based on \(\hat{Q}'_i\)

\[
\hat{D}_{A-\text{plug-in}}(M, N) = \sum_{i=1}^{k} \frac{M_i}{m} \log \frac{M_i/m}{(N_i + 1)/n}.
\]

(13)

**Remark 1.** For technical convenience, \(\hat{Q}'_i\) is not normalized after adding samples. It can be shown that normalization does not provide order-level smaller risk for the plug-in estimator.

**Remark 2.** The add-constant estimator \([18]\) of \(Q\), which adds a fraction sample to each mass point of \(Q\), can also be used as an estimator of divergence. Though intuitively such an estimator should not provide order-level improvement in the risk, the analysis of the risk appears to be difficult.

Theorem 2 characterizes sufficient and necessary conditions on the sample complexity to guarantee consistency of the augmented plug-in estimator over \(\mathcal{M}_{k,f(k)}\). The proof of Theorem 2 involves the proofs of the following two propositions, which provide upper and lower bounds on \(R(\hat{D}_{A-\text{plug-in}}, k, m, n, f(k))\), respectively.

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

\[
R(\hat{D}_{A-\text{plug-in}}, k, m, n, f(k)) \lesssim \left( \frac{kf(k)}{n} + \frac{k}{m} \right)^2 + \frac{\log^2(k)}{m} + \frac{\log^2 f(k)}{m} + \frac{f(k)}{n}.
\]

Therefore, if \(m = \omega(k \vee \log^2 f(k))\) and \(n = \omega(kf(k))\), \(R(\hat{D}_{A-\text{plug-in}}, k, m, n, f(k)) \to 0\) as \(k\) goes to infinity.

**Outline of Proof.** The proof consists of separately bounding the bias and variance of the augmented plug-in estimator. The details are provided in Appendix B.

It can be seen that in the risk bound (12), the first term captures the squared bias, and the remaining terms correspond to the variance.

**Proposition 2.** If \(m \lesssim (k \vee \log^2 f(k))\), or \(n \lesssim (kf(k))\), then for sufficiently large \(k\)

\[
R(\hat{D}_{A-\text{plug-in}}, k, m, n, f(k)) \geq c'
\]

where \(c'\) is a positive constant.

**Outline of Proof.** We provide the central idea of the proof here with the details provided in Appendix C. It can be shown that the bias of the augmented plug-in estimator is lower and upper bounded as follows:

\[
\sup_{(P,Q) \in \mathcal{M}_{k,f(k)}} \mathbb{E}[\hat{D}_{A-\text{plug-in}}(m,n) - D(P\|Q)] \geq \left( \frac{k}{m} \wedge 1 \right) - \frac{kf(k)}{n}
\]

(16a)

\[
\mathbb{E}[\hat{D}_{A-\text{plug-in}}(m,n) - D(P\|Q)] \leq \log \left( 1 + \frac{k}{m} \right) - \frac{k-1}{k} \exp(-\frac{2n}{kf(k)}).
\]

(16b)
1) If \( m \leq k \) and \( n = \omega(k f(k)) \), the lower bound in (16a) is lower bounded by a positive constant, for large \( k \). Hence, the bias as well as the risk is lower bounded by a positive constant.

2) If \( m = \omega(k) \) and \( n \leq k f(k) \), the upper bound in (16b) is upper bounded by a negative constant. This implies that the risk is lower bounded by a positive constant.

3) If \( m \leq k \) and \( n \leq k f(k) \), the lower bound (16a) and the upper bound (16b) do not provide useful information. Hence, we design another approach for this case as follows.

We now focus on the third case above. We choose \( P \) to be the uniform distribution. The bias of the augmented plug-in estimator can be decomposed into: 1) bias due to estimating \( \sum_{i=1}^{k} P_i \log P_i \); and 2) bias due to estimating \( \sum_{i=1}^{k} P_i \log Q_i \). It can be shown that the first bias is always positive, because the uniform distribution achieves the largest entropy for a given alphabet size \( k \). The second bias is always negative for any distribution \( Q \). Hence, the two bias terms may cancel out partially or even fully. Thus, to show the risk is bounded away from zero, the idea is to first determine which bias dominates, and then to construct a pair of distributions accordingly such that the dominant bias is either lower bounded by a positive constant or upper bounded by a negative constant.

If \( \frac{k}{m} \geq (1 + \epsilon) \frac{\omega f(k)}{n} \), where \( \epsilon > 0 \) and \( 0 < \alpha < 1 \) are constants, and which implies that the number of samples drawn from \( P \) is relatively smaller than the number of samples drawn from \( Q \), the first bias dominates. We construct \((P, Q)\): \( P \) is uniform and \( Q = \left( \frac{1}{\alpha k f(k)}, \ldots, \frac{1}{\alpha k f(k)}, 1 - \frac{k-1}{\alpha k f(k)} \right) \).

It can be shown that for the above \((P, Q)\), the bias (and hence the risk) is lower bounded by a positive constant \( \log(1 + \epsilon) \).

If \( \frac{k}{m} < (1 + \epsilon) \frac{\omega f(k)}{n} \), which implies that the number of samples drawn from \( P \) is relatively larger than the number of samples drawn from \( Q \), the second bias dominates. We construct the following distributions \((P, Q)\): \( P \) is uniform and \( Q = \left( \frac{1}{k f(k)}, \ldots, \frac{1}{k f(k)}, 1 - \frac{k-1}{k f(k)} \right) \). It can be shown that for the above \((P, Q)\), the bias is upper bounded by a negative constant. Hence, the risk is lower bounded by a positive constant.

4) If \( m \leq \log^2 f(k) \), we construct two pairs of distributions as follows:

\[
P_1 = \left( \frac{1}{3(k-1)}, \ldots, \frac{1}{3(k-1)}, \frac{2}{3} \right); \\
P_2 = \left( \frac{1-\epsilon}{3(k-1)}, \ldots, \frac{1-\epsilon}{3(k-1)}, \frac{2+\epsilon}{3} \right), \\
Q_1 = Q_2 = \left( \frac{1}{3(k-1)f(k)}, \ldots, \frac{1}{3(k-1)f(k)}, 1 - \frac{1}{3 f(k)} \right). 
\]

By Le Cam’s two-point method [16], it can be shown that if \( m \leq \log^2 f(k) \), no estimator can be consistent, which implies that the augmented plug-in estimator is not consistent.

### 4 Minimax Quadratic Risk over \( \mathcal{M}_{k,f(k)} \)

Our third main result Theorem 3 characterizes the minimax quadratic risk (within a constant factor) of estimating KL divergence over \( \mathcal{M}_{k,f(k)} \). In this section, we describe ideas and central arguments to show this theorem with detailed proofs relegated to Appendix sections.

#### 4.1 Poisson Sampling

In this subsection, we introduce the Poisson sampling technique to handle the dependency of the multinomial distribution, as in [15] for entropy estimation. Such a technique is used in our proofs to develop the lower and upper bounds on the minimax risk in Sections 4.2 and 4.3.
In Poisson sampling, we relax the deterministic sample sizes \(m\) and \(n\) to Poisson random variables \(m' \sim \text{Poi}(m)\) with mean \(m\) and \(n' \sim \text{Poi}(n)\) with mean \(n\), respectively. Under this model, we draw \(m'\) and \(n'\) i.i.d. samples from \(P\) and \(Q\), respectively. The sufficient statistics \(M_i \sim \text{Poi}(nP_i)\) and \(N_i \sim \text{Poi}(nQ_i)\) are independent, which significantly simplifies the analysis.

Analogous to the minimax risk \(\mathcal{R}\), we define its counterpart under the Poisson sampling model as

\[
\tilde{\mathcal{R}}^*(k, m, n, f(k)) \triangleq \inf_{\mathcal{D}} \sup_{(P,Q)\in\mathcal{M}_{k,f(k)}} \mathbb{E}[(D(M, N) - D(P||Q))^2]
\]

where the expectation is taken over \(M_i \sim \text{Poi}(nP_i)\) and \(N_i \sim \text{Poi}(nQ_i)\) for \(i = 1, \ldots, k\). Since the Poissonized sample sizes are concentrated near their means \(m\) and \(n\) with high probability, the minimax risk under Poisson sampling is close to that with fixed sample sizes as stated in the following lemma.

**Lemma 1.** There exists a constant \(c > \frac{1}{2}\) such that

\[
\tilde{\mathcal{R}}^*(k, 2m, 2n, f(k)) - e^{-cm} \log^2 f(k) - e^{-cn} \log^2 f(k) \leq \mathcal{R}^*(k, m, n, f(k)) \leq 4\tilde{\mathcal{R}}^*(k, m/2, n/2, f(k)).
\]

**Proof.** See Appendix [D]

Thus, in order to show Theorem [3], it suffices to bound the Poisson risk \(\tilde{\mathcal{R}}^*(k, m, n, f(k))\) due to Lemma [1]. The next two sections respectively develop lower and upper bounds on the Poisson risk, which match each other (up to a constant factor).

### 4.2 Minimax Lower Bound

In this subsection, we develop the following lower bound on the minimax risk for the estimation of KL divergence over the set \(\mathcal{M}_{k,f(k)}\).

**Proposition 3.** If \(f(k) \geq \log^2 k\) and \(\log^2 n \lesssim k\), \(m \gtrsim \frac{k}{\log k}\), \(n \gtrsim \frac{kf(k)}{\log k}\),

\[
\tilde{\mathcal{R}}^*(k, m, n, f(k)) \gtrsim \left( \frac{k}{m \log k} + \frac{kf(k)}{n \log k} \right)^2 + \frac{\log^2 f(k)}{m} + \frac{f(k)}{n}.
\]

**Outline of Proof.** We next describe the main idea to develop the lower bound with the detailed proof provided in Appendix [E].

To show Proposition [3] it suffices to show that the minimax risk is lower bounded by each of the individual term in [22] separately. It turns out that the proof of the last two terms requires the Le Cam’s two-point method and the proof of the first term requires more general method as we outlined below.

**Le Cam’s two-point method:** The last two terms in the lower bound correspond to the variance of the estimator. We apply the Le Cam’s two-point method by properly choosing two pairs of distributions.

The bound \(\tilde{\mathcal{R}}^*(k, m, n, f(k)) \gtrsim \frac{\log^2 f(k)}{m}\) can be shown by setting

\[
P_1 = \left( \frac{1}{3(k-1)}, \ldots, \frac{1}{3(k-1)}, \frac{2}{3} \right),
\]

\[
P_2 = \left( \frac{1-\epsilon}{3(k-1)}, \ldots, \frac{1-\epsilon}{3(k-1)}, \frac{1}{3} - \frac{1-\epsilon}{3} \right),
\]

\[
Q_1 = Q_2 = \left( \frac{1}{3(k-1)f(k)}, \ldots, \frac{1}{3(k-1)f(k)}, 1 - \frac{1}{3f(k)} \right),
\]

where \(\epsilon = \frac{1}{\sqrt{m}}\).
The bound $\tilde{R}^*(k, m, n, f(k)) \geq \frac{f(k)}{n}$ can be shown by choosing
\[
P_1 = P_2 = \left( \frac{1}{3(k-1)}, 0, \ldots, \frac{1}{3(k-1)}, 0, \frac{5}{6} \right),
\]
\[
Q_1 = \left( \frac{1}{2(k-1)f(k)}, \ldots, \frac{1}{2(k-1)f(k)}, 1 - \frac{1}{2f(k)} \right),
\]
\[
Q_2 = \left( \frac{1 + \epsilon}{2(k-1)f(k)}, \frac{1 - \epsilon}{2(k-1)f(k)}, \ldots, \frac{1 + \epsilon}{2(k-1)f(k)}, \frac{1 - \epsilon}{2(k-1)f(k)}, 1 - \frac{1}{2f(k)} \right),
\]
where $\epsilon = \sqrt{\frac{f(k)}{n}}$.

**Generalized Le Cam’s method:** In order to show $\tilde{R}^*(k, m, n, f(k)) \geq \left( \frac{k}{m \log k} + \frac{ kf(k)}{n \log k} \right)^2$, it suffices to show that $\tilde{R}^*(k, m, n, f(k)) \geq \left( \frac{k}{m \log k} \right)^2$ and $\tilde{R}^*(k, m, n, f(k)) \geq \left( \frac{ kf(k)}{n \log k} \right)^2$. These two lower bounds can be shown by applying the generalized Le Cam’s method, which involves two composite hypothesis \cite{10}:
\[
H_0 : D(P||Q) \leq t \quad \text{versus} \quad H_1 : D(P||Q) \geq t + d.
\]

It is clear that Le Cam’s two-point approach is a special case of this generalized method. If no test can distinguish $H_0$ and $H_1$ reliably, then we obtain a lower bound on the quadratic risk with order $d^2$. Furthermore, the optimal probability of error for composite hypothesis testing is equivalent to the Bayes risk under the least favorable priors. Our goal here is to construct two prior distributions on $(P, Q)$ (respectively for two hypothesis), such that the two corresponding divergence values are separated (by $d$), but the error probability of distinguishing between the two hypotheses is large. However, it is very difficult to design joint prior distributions on $(P, Q)$ that satisfy the above desired property. In order to simplify this procedure, we set one of the distributions $P$ and $Q$ to be known. Then the minimax risk when both $P$ and $Q$ are unknown is lower bounded by the minimax risk with only either $P$ or $Q$ being known. In this way, we only need to design priors on one distribution, which turns out to be good enough for the proof of the lower bound.

In order to show $\tilde{R}^*(k, m, n, f(k)) \geq \left( \frac{k}{m \log k} \right)^2$, we set $Q$ to be the uniform distribution and assume it is known. Therefore, the estimation of $D(P||Q)$ reduces to the estimation of $\sum_{i=1}^{k} P_i \log P_i$, which is the entropy of $P$. Following similar steps in \cite{13}, we can obtain the desired result.

In order to show $\tilde{R}^*(k, m, n, f(k)) \geq \left( \frac{ kf(k)}{n \log k} \right)^2$, we let $P$ be known and set
\[
P = \left( \frac{f(k)}{n \log k}, \ldots, \frac{f(k)}{n \log k}, 1 - \frac{(k-1)f(k)}{n \log k} \right).
\]
Therefore, the estimation of $D(P||Q)$ reduces to the estimation of $\sum_{i=1}^{k} P_i \log Q_i$. We then properly design priors on $Q$ and apply generalized Le Cam’s method to obtain the desired result.

We note that the proof of Proposition 3 may be strengthened by designing jointly distributed priors on $(P, Q)$, instead of treating them separately. This may help to relax or remove the condition $f(k) \geq \log^2 k$ in Proposition 3.

### 4.3 Minimax Upper Bound via Optimal Estimator

Comparing the lower bound in Proposition 3 with the upper bound in Proposition 4 that characterizes an upper bound on the risk for the augmented plug-in estimator, it is clear that there is a difference of $\log k$ factor in the bias terms, which implies that the augmented plug-in estimator is not minimax optimal. A promising approach to fill in this gap is to incorporate polynomial approximation into the estimator in order to reduce the bias with the price of the variance, similarly to entropy estimation \cite{13,15}. In this subsection, we construct an estimator by jointly employing polynomial approximation and the plug-in approach, and characterize an upper bound on the minimax risk in the following proposition.
Proposition 4. If $\log m \leq C \log k$ and $\log^2 n \lesssim k$, then the worst case quadratic risk for $\hat{D}$ is upper bounded as follows:

$$
\sup_{(P,Q)\in \mathcal{M}_k,f(k)} \mathbb{E} \left[ \left( \hat{D}(M,N) - D(P||Q) \right)^2 \right] \lesssim \left( \frac{k}{m \log k} + \frac{kf(k)}{n \log k} \right)^2 + \frac{\log^2 f(k)}{m} + \frac{f(k)}{n}.
$$

(31)

It is clear that the upper bound in Proposition 4 matches the lower bound in Proposition 3 (up to a constant factor), and thus the constructed estimator is minimax optimal, and the minimax risk in Theorem 3 is established.

Outline of Proof. We outline the central idea here to construct the optimal minimax estimator with the detailed proof provided in Appendix E. The KL divergence $D(P||Q)$ can be written as

$$
D(P||Q) = \sum_{i=1}^{k} P_i \log P_i - \sum_{i=1}^{k} P_i \log Q_i.
$$

(32)

The first term equals the entropy of $P$, and the minimax optimal entropy estimator (denoted by $\hat{D}_1$) in [13] can be applied. The major challenge arises due to the second term. The idea is to use polynomial approximation to reduce the bias when $Q_i$ is small. With bounded density ratio, the function $|P_i \log Q_i|$ is bounded by $f(k) \log k$. A natural idea is to construct polynomial approximation for $|P_i \log Q_i|$ in two dimensions. However, it is very challenging to find the explicit form of the best polynomial approximation [19] in this case. On the other hand, one-dimensional polynomial approximation of $\log Q_i$ also appears challenging. First of all, the function $\log x$ on interval $[0, 1]$ is not bounded with a singularity point at $x = 0$. Hence, the approximation of $\log x$ when $x$ is near the point $x = 0$ is inaccurate. Secondly, such an approach ignores the fact that $\frac{P_i}{Q_i} \leq f(k)$ which implies that when $Q_i$ is small, the value of $P_i$ should also be small. Such a fact should play a very important role to control the bias caused by inaccurate estimation of $\log Q_i$.

Another approach is to express the function $P_i \log Q_i$ as $\frac{P_i}{Q_i} Q_i \log Q_i$, and then estimate $\frac{P_i}{Q_i}$ and $Q_i \log Q_i$ separately. Although the function $Q_i \log Q_i$ can be approximated accurately (see [20] Section 7.5.4), it is difficult to find a good estimator for $\frac{P_i}{Q_i}$. However, such an approach motivates us to think whether we can approximate $\frac{1}{Q_i} Q_i \log Q_i$ jointly using one-dimensional polynomial function of $Q_i$ and then derive the approximation error for the function $P_i \frac{1}{Q_i} Q_i \log Q_i$, rather than deriving the approximation error of $\frac{1}{Q_i} Q_i \log Q_i$ on its own. Due to the density ratio constraint $f(k)$, when $Q_i$ is very small, $P_i$ is upper bounded by $f(k) Q_i$, which is also very small. By multiplying $P_i$ to $\frac{1}{Q_i} Q_i \log Q_i$, we are able to control the approximation error of $P_i \frac{1}{Q_i} Q_i \log Q_i$. In the following, we will introduce how we construct our estimator in detail.

By Lemma 1 we apply Poisson sampling to simplify the analysis. We first draw $m' \sim \text{Poi}(2m)$, and then draw $m'$ i.i.d. samples from distribution $P$. We use the $m'$ samples to estimate $\sum_{i=1}^{k} P_i \log P_i$ using the entropy estimator proposed in [13]. Next, we draw $n'_1 \sim \text{Poi}(n)$ and $n'_2 \sim \text{Poi}(n)$ independently. We then draw $n'_1$ and $n'_2$ i.i.d. samples from distribution $Q$, which we use $N = (N_1, \ldots, N_k)$ and $N' = (N'_1, \ldots, N'_k)$ to denote the histograms of $n'_1$ samples and $n'_2$ samples, respectively. We use $N'$ to determine whether to use polynomial estimator or plug-in estimator, and use $N$ to estimate the function. Based on the generation scheme, $N$ and $N'$ are independent, where $N_i \sim \text{Poi}(n Q_i)$, and $N'_i \sim \text{Poi}(n Q_i)$.

We then focus on the estimation of $\sum_{i=1}^{k} P_i \log Q_i$. We construct a polynomial approximation for the function $P_i \log Q_i$ if $Q_i \in [0, c_1 \log k/n]$, and we use the bias corrected augmented plug-in estimator if $Q_i \in [c_1 \log k/n, 1]$. We let $L = \lfloor c_0 \log k \rfloor$, where $c_0$ is a constant to be determined later. We denote the degree-$L$ best polynomial approximation of the function $x \log x$ on interval $[0, 1]$ as $\sum_{j=0}^{L} a_j x^j$. We further scale the interval $[0, 1]$ to $[0, c_1 \log k/n]$. Then we have the best polynomial approximation of the function $x \log x$ on
interval \([0, \frac{c_1 \log k}{n}]\) as follows:

\[
\gamma_L(x) = \sum_{j=0}^{L} \frac{a_j n^{j-1}}{(c_1 \log k)^{j-1}} x^j + \left( \log \frac{n}{c_1 \log k} \right) x. \tag{33}
\]

Following the result in [20] Section 7.5.4, the approximation error of \(\gamma_L(x)\) on interval \([0, \frac{c_1 \log k}{n}]\) can be upper bounded as follows:

\[
\sup_{x \in [0, \frac{c_1 \log k}{n}]} |\gamma_L(x) - x \log x| \lesssim \frac{1}{n \log k}. \tag{34}
\]

Therefore, we have \(|\gamma_L(0) - 0 \log 0| \lesssim \frac{1}{n \log k}\), which implies that the zero-degree term in \(\gamma_L(x)\) satisfies:

\[
\frac{a_1 c_1 \log k}{n} \lesssim \frac{1}{n \log k}. \tag{35}
\]

Now, subtracting the zero-degree term from \(\gamma_L(x)\) yields the following polynomial \(\mu_L(x)\)

\[
\mu_L(x) \triangleq \gamma_L(x) - \frac{a_1 c_1 \log k}{n} = \sum_{j=1}^{L} \frac{a_j n^{j-1}}{(c_1 \log k)^{j-1}} x^j + \left( \log \frac{n}{c_1 \log k} \right) x. \tag{36}
\]

The approximation error of \(\mu_L(x)\) for function \(x \log x\) on interval \([0, \frac{c_1 \log k}{n}]\) can also be upper bounded by \(\frac{1}{n \log k}\), because

\[
\sup_{x \in [0, \frac{c_1 \log k}{n}]} |\mu_L(x) - x \log x|
\]

\[
= \sup_{x \in [0, \frac{c_1 \log k}{n}]} |\gamma_L(x) - x \log x| - \frac{a_1 c_1 \log k}{n}
\]

\[
\leq \sup_{x \in [0, \frac{c_1 \log k}{n}]} |\gamma_L(x) - x \log x| + \left| \frac{a_1 c_1 \log k}{n} \right|
\]

\[
\lesssim \frac{1}{n \log k}. \tag{37}
\]

The bound in [37] implies that although \(\mu_L(x)\) is not the best polynomial approximation of \(x \log x\), the approximation error of \(\mu_L(x)\) has the same order as \(\gamma_L(x)\). Compared to \(\gamma_L(x)\), there is no zero-degree term in \(\mu_L(x)\), and hence \(\mu_L(x)\) is still a valid polynomial approximation of \(\log x\). Although the approximation error of \(\log x\) using \(\frac{\mu_L(x)}{x}\) is unbounded, the approximation error of \(P_i \log Q_i\) using \(P_i \frac{\mu_L(Q_i)}{Q_i}\) can be bounded.

More importantly, by how we construct \(\mu_L(x)\), \(P_i \frac{\mu_L(Q_i)}{Q_i}\) is a polynomial function of \(P_i\) and \(Q_i\), for which an unbiased estimator can be constructed.

More specifically, the approximation error of using \(P_i \frac{\mu_L(Q_i)}{Q_i}\) to approximate \(P_i \log Q_i\) can be bounded as follows:

\[
|P_i \frac{\mu_L(Q_i)}{Q_i} - P_i \log Q_i| = \frac{P_i}{Q_i} |\mu_L(Q_i) - Q_i \log Q_i| \lesssim \frac{f(k)}{n \log k}. \tag{38}
\]

if \(Q_i \in [0, \frac{c_1 \log k}{n}]\). We further define the factorial moment of \(x\) by \(\langle x \rangle_m \triangleq \frac{x^m}{(m!)^{\frac{1}{m}}}\). If \(X \sim \text{Poi}(\lambda), \mathbb{E}[X]^m = \lambda^m\). Based on such a fact, we construct an unbiased estimator for \(\frac{\mu_L(Q_i)}{Q_i}\) as follows:

\[
g_L(N_i) = \sum_{j=1}^{L} \frac{a_j}{(c_1 \log k)^{j-1}} (N_i)^{j-1} + \left( \log \frac{n}{c_1 \log k} \right). \tag{39}
\]
Therefore, we construct our estimator for $\sum_{i=1}^{k} P_i \log Q_i$ as follows:

$$
\hat{D}_2 = \sum_{i=1}^{k} \left( \frac{M_i}{m} g_L(N_i) \mathbb{1}_{\{N'_i \leq c_2 \log k\}} + \frac{M_i}{m} \left( \log \frac{N_i + 1}{n} - \frac{1}{2(N_i + 1)} \right) \mathbb{1}_{\{N'_i > c_2 \log k\}} \right).
$$

Combining the estimator $\hat{D}_1$ for $\sum_{i=1}^{k} P_i \log P_i$ and the estimator $\hat{D}_2$ in (40) for $\sum_{i=1}^{k} P_i \log Q_i$, we obtain the estimator $\bar{D}_{\text{opt}}$ for $D(P||Q)$ for KL divergence as follows

$$
\bar{D}_{\text{opt}} = \hat{D}_1 - \hat{D}_2.
$$

Due to the density ratio constraint, we can show that $0 \leq D(P||Q) \leq \log f(k)$. Hence, we construct the estimator $\hat{D}$ as follows, which is minimax optimal.

$$
\hat{D}_{\text{opt}} = \bar{D}_{\text{opt}} \lor 0 \land \log f(k).
$$

The corresponding risk for the above estimator provides an upper bound on the minimax risk. The bound is developed in Appendix E.

5 Conclusion

In this paper, we have studied the estimation of KL divergence between large-alphabet distributions. We have shown that there exists no consistent estimator for KL divergence under the worst-case quadratic risk over the set of all pairs of distributions. We then have studied a more practical set of pairs of distributions with bounded density ratio. We have proposed an augmented plug-in estimator, and characterized tight sufficient and necessary conditions on sample complexity for such an estimator to be consistent. We have further designed an minimax optimal estimator by jointly employing polynomial approximation and plug-in approach, and established the optimal minimax rate. We anticipate that the designed KL divergence estimator can be applied to studying various problems including classification, anomaly detection, community clustering, and nonparametric hypothesis testing.
A Proof of Theorem 1

Theorem 1 follows from Le Cam’s two-point method [16]: If two pairs of distributions \((P_1, Q_1)\) and \((P_2, Q_2)\) are sufficiently close such that it is impossible to reliably distinguish between them using \(m\) samples from \(P\) and \(n\) samples from \(Q\) (with error probability less than some constant), then any estimator suffers a risk proportional to the square of the difference between the divergence values \(|D(P_1 \| Q_1) - D(P_2 \| Q_2)|^2\).

For any fixed \((k, m, n)\), applying Le Cam’s two-point method, we have

\[
R^*(k, m, n) \geq \frac{1}{16} (D(P_1 \| Q_1) - D(P_2 \| Q_2))^2 \exp (-m D(P_1 \| P_2) - n D(Q_1 \| Q_2)).
\]  

(43)

The idea here is to keep \(P_1\) close to \(P_2\), and \(Q_1\) close to \(Q_2\), so that \(D(P_1 \| P_2) \leq \frac{1}{m}\), \(D(Q_1 \| Q_2) \leq \frac{1}{n}\), but \((D(P_1 \| Q_1) - D(P_2 \| Q_2))^2\) is large. We construct the following two pairs of distributions:

\[
P_1 = P_2 = \left( \frac{1}{2}, \frac{1}{2(k-1)} \right), \quad \frac{1}{2(k-1)} \right), \quad (44)
\]

\[
Q_1 = \left( \frac{1}{k-1}, \frac{1}{k-1} \right), \quad (45)
\]

\[
Q_2 = \left( \frac{1}{k-1}, \frac{1}{k-1} \right), \quad (46)
\]

where \(0 < \epsilon_1 < 1/4\), and \(\epsilon_2 = \epsilon_1 + \frac{1}{4n} < \frac{1}{2}\). By such a construction, we obtain,

\[
D(P_1 \| P_2) = 0,
\]

(47)

\[
D(Q_1 \| Q_2) = (1 - \epsilon_1) \log \frac{1 - \epsilon_1}{1 - \epsilon_2} + \epsilon_1 \log \frac{\epsilon_1}{\epsilon_2}.
\]

(48)

The constraint \(D(Q_1 \| Q_2) \leq \frac{1}{n}\) is satisfied:

\[
D(Q_1 \| Q_2) = (1 - \epsilon_1) \log \left( 1 + \frac{\epsilon_2 - \epsilon_1}{1 - \epsilon_2} \right) + \epsilon_1 \log \frac{\epsilon_1}{\epsilon_2} + \frac{\epsilon_1}{4n}
\]

\[
= (1 - \epsilon_1) \log \left( 1 + \frac{1}{4n(1 - \epsilon_2)} \right) + \frac{\epsilon_1}{1 + \frac{1}{4n\epsilon_1}}
\]

\[
< \frac{1 - \epsilon_1}{4n(1 - \epsilon_2)}. \quad (49)
\]

Because \(\epsilon_1 > 0\) and \(\epsilon_2 < 1/2\), we obtain \(D(Q_1 \| Q_2) \leq \frac{1}{2n}\).

By the construction of \((P_1, Q_1)\) and \((P_2, Q_2)\),

\[
(D(P_1 \| Q_1) - D(P_2 \| Q_2))^2 = \left( \frac{1}{2} \log \left( \frac{1 - \epsilon_2}{1 - \epsilon_1} \right) + \frac{1}{2} \log \frac{\epsilon_2}{\epsilon_1} \right)^2
\]

\[
= \left( \frac{1}{2} \log \left( \frac{1 - \epsilon_2}{1 - \epsilon_1} \right) + \frac{1}{2} \log \left( 1 + \frac{1}{4n\epsilon_1} \right) \right)^2. \quad (50)
\]

Note that \(\frac{1}{2} \log \left( \frac{1 - \epsilon_2}{1 - \epsilon_1} \right)\) is upper bounded by \(\log 2\). The only constraints here is \(0 < \epsilon_1 < 1/4\). Hence, we can choose \(\epsilon_1\) to be arbitrarily small, such that \(\frac{1}{4n\epsilon_1}\) is arbitrarily large for any fixed \(k\), \(m\) and \(n\). Therefore,

\[
(D(P_1 \| Q_1) - D(P_2 \| Q_2))^2 = \left( \frac{1}{2} \log \left( \frac{1 - \epsilon_2}{1 - \epsilon_1} \right) + \frac{1}{2} \log \left( 1 + \frac{1}{4n\epsilon_1} \right) \right)^2 = \infty. \quad (51)
\]

Hence, the minimax quadratic risk lower bound is infinity for any fixed \(k\), \(m\) and \(n\). Thus, there does not exist any consistent estimator over the set \(\mathcal{M}_k\).
B Proof of Proposition 1

The quadratic risk can be decomposed into the variance and the square of the bias as follows:

\[
E[(\hat{D}_{A\text{-plug-in}}(M, N) - D(P\|Q))^2] = \left(E[\hat{D}_{A\text{-plug-in}}(M, N) - D(P\|Q)]\right)^2 + \text{Var}[\hat{D}_{A\text{-plug-in}}(M, N)].
\]

We next bound the bias and the variance in the following two subsections, respectively.

B.1 Bounding the Bias

The bias of the augmented plug-in estimator can be written as,

\[
|E(\hat{D}_{A\text{-plug-in}}(M, N) - D(P\|Q))| \leq \left|E\left(\sum_{i=1}^{k} \frac{M_i}{m} \log \frac{M_i/m}{(N_i+1)/n} - P_i \log \frac{P_i}{Q_i}\right)\right| \leq \left|E\left(\sum_{i=1}^{k} P_i \log \frac{M_i/m}{N_i+1} - P_i \log \frac{nQ_i}{N_i+1}\right)\right|.
\]

The first term in (52) is the bias of the direct plug-in estimator for entropy estimation, which is bounded as in [21]:

\[
|\text{Bias}(\hat{H}_{\text{plug-in}})| = \left|E\left(\sum_{i=1}^{k} \frac{M_i}{m} \log \frac{M_i/m}{m} - P_i \log P_i\right)\right| \leq \log \left(1 + \frac{k-1}{m}\right) < \frac{k}{m}. \quad (53)
\]

Next, we bound the second term in (52) as follows:

\[
E\left(\sum_{i=1}^{k} P_i \log \frac{nQ_i}{N_i+1}\right) = -\sum_{i=1}^{k} P_i E\left(\log \left(1 + \frac{N_i+1-nQ_i}{nQ_i}\right)\right)
\geq -\sum_{i=1}^{k} P_i E\left(\frac{N_i+1-nQ_i}{nQ_i}\right)
= -\sum_{i=1}^{k} P_i \frac{1}{nQ_i}
\geq -\frac{kf(k)}{n}, \quad (54)
\]

where (a) is due to the fact that \(\log(1 + x) \leq x\). Furthermore, by Jensen’s inequality, we have,

\[
E\left(\sum_{i=1}^{k} P_i \log \frac{nQ_i}{N_i+1}\right) = \sum_{i=1}^{k} P_i E\left(\log \frac{nQ_i}{N_i+1}\right) \leq \sum_{i=1}^{k} P_i \log E\left[\frac{nQ_i}{N_i+1}\right]. \quad (55)
\]

Let \(B(n, p)\) denote the Binomial distribution where \(n\) is the total number of experiment, and \(p\) is the probability that each experiment yields a desired result. Note that \(N_i \sim B(n, Q_i)\), the expectation in (55)
Using this notation, the augmented plug-in estimator can be written as

\[
\mathbb{E}\left[\frac{1}{N_i + 1}\right] = \sum_{j=0}^{n} \frac{1}{j+1} \binom{n}{j} Q_i^j (1 - Q_i)^{n-j}
\]

\[
= \sum_{j=0}^{n} \frac{1}{j+1} \frac{n!}{(n-j)!j!} Q_i^j (1 - Q_i)^{n-j}
\]

\[
= \frac{1}{n+1} \sum_{j=0}^{n} \frac{(n+1)!}{(n-j)!j!} Q_i^j (1 - Q_i)^{n-j}
\]

\[
= \frac{1}{(n+1)Q_i} \sum_{j=0}^{n} \frac{(n+1)}{j+1} Q_i^{j+1} (1 - Q_i)^{n-j}
\]

\[
= \frac{1}{(n+1)Q_i} (1 - (1 - Q_i)^{n+1}) < \frac{1}{nQ_i}.
\]

Thus, we obtain

\[
\mathbb{E} \left( \sum_{i=1}^{k} P_i \log \frac{nQ_i}{N_i + 1} \right) \leq \sum_{i=1}^{k} P_i \log \mathbb{E} \left[ \frac{nQ_i}{N_i + 1} \right] < \sum_{i=1}^{k} P_i \log \frac{nQ_i}{nQ_i} = 0.
\]

Combining (54) and (57), we obtain the upper bound for the second term in the bias,

\[
\left| \mathbb{E} \left( \sum_{i=1}^{k} P_i \log \frac{nQ_i}{N_i + 1} \right) \right| \leq \frac{kf(k)}{n}.
\]

Hence,

\[
\left| \mathbb{E} \left( \hat{D}_{A\text{-plug-in}}(M, N) - D(P||Q) \right) \right| < \frac{k}{m} + \frac{kf(k)}{n}.
\]

### B.2 Bounding the Variance

In this subsection, we derive an upper bound on the variance. Applying Efron-Stein inequality [22, Theorem 3.1], we have:

\[
\text{Var}[\hat{D}_{A\text{-plug-in}}(M, N)] \\
\leq \frac{m}{2} \mathbb{E} \left[ (\hat{D}_{A\text{-plug-in}}(M, N) - \hat{D}_{A\text{-plug-in}}(M', N))^2 \right] + \frac{n}{2} \mathbb{E} \left[ (\hat{D}_{A\text{-plug-in}}(M, N) - \hat{D}_{A\text{-plug-in}}(M, N'))^2 \right],
\]

where \(M'\) and \(N'\) are the histograms of \((X_1, \ldots, X_{m-1}, X'_m)\) and \((Y_1, \ldots, Y_{n-1}, Y'_n)\), respectively. Here, \(X'_m\) is an independent copy of \(X_m\) and \(Y'_n\) is an independent copy of \(Y_n\).

Let \(\bar{M} = (\bar{M}_1, \ldots, \bar{M}_k)\) be the histogram of \((X_1, \ldots, X_{m-1})\), \(\bar{N} = (\bar{N}_1, \ldots, \bar{N}_k)\) be the histogram of \((Y_1, \ldots, Y_{n-1})\). Then \(\bar{M} \sim \text{Multinomial}(m-1, P)\) is independent from \(X_m\) and \(X'_m\), and \(\bar{N} \sim \text{Multinomial}(n-1, Q)\) is independent from \(Y_n\) and \(Y'_n\). Denote the function \(\phi\) as

\[
\phi(x, y) \triangleq x \log x - x \log y.
\]

Using this notation, the augmented plug-in estimator can be written as

\[
\hat{D}_{A\text{-plug-in}}(M, N) = \sum_{i=1}^{k} \phi\left( \frac{M_i}{m}, \frac{N_i + 1}{n} \right).
\]
Let $\tilde{M}_{X_m}$ be the number of samples in bin $X_m$. We can bound the first term in (60) as follows:

$$
\begin{align*}
\mathbb{E}\left[ (\hat{D}_{A-\text{plug-in}}(M, N) - \hat{D}_{A-\text{plug-in}}(M', N))^2 \right] \\
= \mathbb{E}\left[ \left( \phi\left( \frac{\tilde{M}_{X_m} + 1}{m}, \frac{N_{X_m} + 1}{n} \right) + \phi\left( \frac{\tilde{M}_{X_m'}}{m}, \frac{N_{X_m'} + 1}{n} \right) \right)^2 \bigg| X_m, X_m' \right] \\
& \leq 4 \mathbb{E}\left[ \left( \phi\left( \frac{\tilde{M}_{X_m} + 1}{m}, \frac{N_{X_m} + 1}{n} \right) - \phi\left( \frac{\tilde{M}_{X_m'}}{m}, \frac{N_{X_m'} + 1}{n} \right) \right)^2 \bigg| X_m \right] \\
& = 4 \sum_{j=1}^{k} \mathbb{E}\left[ \left( \phi\left( \frac{\tilde{M}_j + 1}{m}, \frac{N_j + 1}{n} \right) - \phi\left( \frac{\tilde{M}_j}{m}, \frac{N_j + 1}{n} \right) \right)^2 \bigg| P_j \right] \\
& = 4 \sum_{j=1}^{k} \mathbb{E}\left[ \left( (\tilde{M}_j + 1) \log \frac{\tilde{M}_j + 1}{m} - \tilde{M}_j \log \frac{\tilde{M}_j}{m} - \log \frac{N_j + 1}{n} \right)^2 \bigg| P_j \right] \\
& \leq 8 \sum_{j=1}^{k} \left[ \log \left( \frac{\tilde{M}_j + 1}{m} \right) - \log \frac{n}{m} \right]^2 P_j. 
\end{align*}
$$

(63)

where $(a)$ is due to the fact that $X_m$ is independent from $X_m'$, and $(b)$ is due to the fact that $0 \leq x \log(1 + \frac{1}{x}) \leq 1$ for all $x > 0$. We rewrite the second term in (63) as follows,

$$
\begin{align*}
\mathbb{E}\left[ \left( \log \frac{\tilde{M}_j + 1}{m} \frac{n}{N_j + 1} \right)^2 \right] \\
= \mathbb{E}\left[ \left( \log \frac{\tilde{M}_j + 1}{m} \frac{n}{N_j + 1} \right) \mathbb{I}_{\{\tilde{M}_j \leq \frac{mP_j}{2} \}} \mathbb{I}_{\{N_j > \frac{nQ_j}{2} \}} \right]^2 + \mathbb{E}\left[ \left( \log \frac{\tilde{M}_j + 1}{m} \frac{n}{N_j + 1} \right) \mathbb{I}_{\{\tilde{M}_j > \frac{mP_j}{2} \}} \mathbb{I}_{\{N_j \leq \frac{nQ_j}{2} \}} \right]^2 \\
+ \mathbb{E}\left[ \left( \log \frac{\tilde{M}_j + 1}{m} \frac{n}{N_j + 1} \right) \mathbb{I}_{\{\tilde{M}_j \leq \frac{mP_j}{2} \}} \mathbb{I}_{\{N_j \leq \frac{nQ_j}{2} \}} \right]^2 + \mathbb{E}\left[ \left( \log \frac{\tilde{M}_j + 1}{m} \frac{n}{N_j + 1} \right) \mathbb{I}_{\{\tilde{M}_j > \frac{mP_j}{2} \}} \mathbb{I}_{\{N_j \leq \frac{nQ_j}{2} \}} \right]^2.
\end{align*}
$$

To analyze the above equation, we first have the following properties that are useful:

- If $\tilde{M}_j \leq \frac{mP_j}{2}$, then $\frac{1}{m} \leq \frac{M_j + 1}{m} \leq \frac{P_j}{m} + \frac{1}{m}$;
- If $\tilde{M}_j > \frac{mP_j}{2}$, then $\frac{P_j}{m} + \frac{1}{m} < \frac{M_j + 1}{m} \leq 1$;
- If $N_j > \frac{nQ_j}{2}$, then $\frac{1}{n} \leq \frac{N_j + 1}{n} \leq 1 + \frac{1}{n}$;
- If $N_j \leq \frac{nQ_j}{2}$, then $\frac{1}{n} \leq \frac{N_j + 1}{n} \leq \frac{Q_j}{n} + \frac{1}{n}$.

With the above bounds, and assume that $m > 2, n > 2$, we next analyze the following four cases,
1. If \( \tilde{M}_j \leq \frac{mP_j}{2} \) and \( N_j > \frac{nQ_j}{2} \), then we have
\[
\frac{n}{m(n+1)} \leq \frac{\tilde{M}_j + 1}{m} \frac{n}{N_j + 1} \leq \frac{P_j + \frac{1}{m}}{Q_j + \frac{1}{n}},
\]
and
\[
\mathbb{E} \left[ \left( \log \frac{\tilde{M}_j + 1}{m} \frac{n}{N_j + 1} \mathbb{1}_{\{\tilde{M}_j \leq \frac{mP_j}{2}\}} \mathbb{1}_{\{N_j > \frac{nQ_j}{2}\}} \right)^2 \right] 
\leq 2 \log^2 \left( \frac{n}{m(n+1)} \right) + \left( \log \left( \frac{P_j}{2} + \frac{1}{m} \right) - \log \left( \frac{Q_j}{2} + \frac{1}{n} \right) \right)^2 \mathbb{P} \left( \tilde{M}_j \leq \frac{mP_j}{2} \right)
\leq 2 \left( \log^2 \left( \frac{1}{2m} \right) + \left( \log \left( \frac{P_j}{2} + \frac{1}{m} \right) - \log \left( \frac{Q_j}{2} + \frac{1}{n} \right) \right)^2 \right) \exp \left( - \frac{(m-2)P_j}{8} \right)
\leq 4 \left[ \log^2 \left( \frac{1}{2m} \right) + \left( \log^2 \left( \frac{P_j}{2} + \frac{1}{m} \right) + \log^2 \left( \frac{Q_j}{2} + \frac{1}{n} \right) \right) \right] \exp \left( - \frac{(m-2)P_j}{8} \right).
\] (64)

where (c) follows from the Chernoff bound of binomial tail.

2. If \( \tilde{M}_j > \frac{mP_j}{2} \) and \( N_j > \frac{nQ_j}{2} \), then we have
\[
\frac{n}{n+1} \left( \frac{P_j}{2} + \frac{1}{m} \right) \leq \frac{\tilde{M}_j + 1}{m} \frac{n}{N_j + 1} \leq \frac{1}{nQ_j},
\]
and
\[
\mathbb{E} \left[ \left( \log \frac{\tilde{M}_j + 1}{m} \frac{n}{N_j + 1} \mathbb{1}_{\{\tilde{M}_j > \frac{mP_j}{2}\}} \mathbb{1}_{\{N_j > \frac{nQ_j}{2}\}} \right)^2 \right] 
\leq 2 \log^2 \left( \frac{n}{n+1} \left( \frac{P_j}{2} + \frac{1}{m} \right) \right) + \log^2 \left( \frac{Q_j}{2} + \frac{1}{n} \right)
\leq 4 \left[ \log^2 \left( \frac{P_j}{4} \right) + \log^2 \left( \frac{Q_j}{2} \right) \right].
\] (65)

3. If \( \tilde{M}_j \leq \frac{mP_j}{2} \) and \( N_j \leq \frac{nQ_j}{2} \), then we have
\[
\frac{1}{mQ_j + \frac{1}{n}} \leq \frac{\tilde{M}_j + 1}{m} \frac{n}{N_j + 1} \leq \frac{nP_j}{2} + \frac{m}{m},
\]
and
\[
\mathbb{E} \left[ \left( \log \frac{\tilde{M}_j + 1}{m} \frac{n}{N_j + 1} \mathbb{1}_{\{\tilde{M}_j \leq \frac{mP_j}{2}\}} \mathbb{1}_{\{N_j \leq \frac{nQ_j}{2}\}} \right)^2 \right] 
\leq 2 \log^2 \left( m \left( \frac{Q_j}{2} + \frac{1}{n} \right) \right) + \log^2 \left( n \left( \frac{P_j}{2} + \frac{1}{m} \right) \right) \exp \left( - \frac{(m-2)P_j}{8} \right) \exp \left( - \frac{nQ_j}{8} \right)
\leq 4 \left[ \log^2 \left( \frac{Q_j}{2} \right) + \log^2 m \right] + \left( \log^2 \left( \frac{P_j}{2} \right) + \log^2 n \right) \exp \left( - \frac{(m-2)P_j}{8} \right) \exp \left( - \frac{nQ_j}{8} \right). \] (66)

4. If \( \tilde{M}_j > \frac{mP_j}{2} \) and \( N_j \leq \frac{nQ_j}{2} \), then we have
\[
\frac{P_j + \frac{1}{m}}{Q_j + \frac{1}{n}} \leq \frac{\tilde{M}_j + 1}{m} \frac{n}{N_j + 1} \leq n,
\]
and
\[
\mathbb{E} \left[ \left( \log \frac{\tilde{M}_j + 1}{m} \frac{n}{N_j + 1} \mathbb{1}_{\{\tilde{M}_j > \frac{mP_j}{2}\}} \mathbb{1}_{\{N_j \leq \frac{nQ_j}{2}\}} \right)^2 \right] 
\leq 4 \left[ \log^2 n + \left( \log^2 \left( \frac{P_j}{2} \right) + \log^2 \left( \frac{Q_j}{2} \right) \right) \right] \exp \left( - \frac{nQ_j}{8} \right). \] (67)
Combining the four cases together, we have
\[
\mathbb{E} \left[ (\hat{D}_{\text{A-\text{plug-in}}}(M, N) - \hat{D}_{\text{A-\text{plug-in}}}(M', N))^2 \right] \\
\leq \frac{8}{m^2} + \frac{8}{m^2} \sum_{j=1}^{k} \mathbb{E} \left[ \left( \log \frac{M_j + 1}{m} - \log \frac{N_j + 1}{n} \right)^2 \right] P_j \\
\leq \frac{32}{m^2} \sum_{j=1}^{k} P_j \left[ \left( \log^2 \left( \frac{Q_j}{2} \right) + \log^2 m \right) + \left( \log^2 \left( \frac{P_j}{2} \right) + \log^2 n \right) \right] \exp \left( - \frac{(m - 2)P_j}{8} \right) \exp \left( - \frac{nQ_j}{8} \right) \\
+ \frac{32}{m^2} \sum_{j=1}^{k} P_j \left[ \log^2 (2m) + \left( \log^2 \left( \frac{P_j}{2} \right) + \log^2 \left( \frac{Q_j}{2} \right) \right) \right] \exp \left( - \frac{(m - 2)P_j}{8} \right) \\
+ \frac{32}{m^2} \sum_{j=1}^{k} P_j \left[ \log^2 (n) + \left( \log^2 \left( \frac{P_j}{2} \right) + \log^2 \left( \frac{Q_j}{2} \right) \right) \right] \exp \left( - \frac{nQ_j}{8} \right) \\
+ \frac{32}{m^2} \sum_{j=1}^{k} P_j \left[ \left( \log^2 \left( \frac{P_j}{4} \right) + \log^2 \left( \frac{Q_j}{2} \right) \right) \right] + \frac{8}{m^2} \\
\leq \frac{32}{m^2} \sum_{j=1}^{k} P_j \left[ 4 \log^2 \left( \frac{4}{P_j} \right) + 4 \log^2 \left( \frac{2}{Q_j} \right) + 2 \log^2 (2m) \exp \left( - \frac{(m - 2)P_j}{8} \right) + 2 \log^2 n \exp \left( - \frac{nQ_j}{8} \right) \right] + \frac{8}{m^2}.
\]

(68)

Here, we analyze the asymptotic behavior of the following four terms:

1. It can be shown that \( \sum_{j=1}^{k} P_j \log P_j \leq \log k \) and \( \sum_{j=1}^{k} P_j \log^2 P_j \leq \log^2 k \), we have
\[
\sum_{j=1}^{k} P_j \log^2 \left( \frac{4}{P_j} \right) = \sum_{j=1}^{k} P_j (\log^2 (P_j) + \log^2 4 - 2 \log P_j \log 4) \leq (\log k + \log 4)^2.
\]

(69)

2. Given the bounded ratio constraint \( \frac{1}{Q_j} \leq \frac{f(k)}{P_j} \), we have
\[
\sum_{j=1}^{k} P_j \log^2 \left( \frac{2}{Q_j} \right) \leq \sum_{j=1}^{k} P_j \log^2 \frac{2f(k)}{P_j} = \sum_{j=1}^{k} P_j (\log^2 f(k) + \log^2 P_j - 2 \log 2f(k) \log P_j) \leq (\log k + \log 2f(k))^2.
\]

(70)

3. Since \( \sup_{x > 0} x \exp(-nx/8) = \frac{8}{ne} \), we have
\[
\sum_{i=1}^{k} P_i \log^2 (2m) \exp \left( - \frac{(m - 2)P_j}{8} \right) \leq \sum_{i=1}^{k} 8 \log^2 (2m) \leq \frac{8k \log^2 (2m)}{(m - 2)e}.
\]

(71)

4. Since \( Q_j \geq \frac{P_j}{f(k)^3} \), and \( \sup_{x > 0} x \exp(-nx/8) = \frac{8}{ne} \), we have
\[
\sum_{i=1}^{k} P_i \log^2 n \exp \left( - \frac{nQ_j}{8} \right) \leq \sum_{i=1}^{k} \log^2 n P_i \exp \left( - \frac{nP_j}{8f(k)} \right) \leq \frac{8kf(k) \log^2 n}{ne}.
\]

(72)

Thus,
\[
\mathbb{E} \left[ (\hat{D}_{\text{A-\text{plug-in}}}(M, N) - \hat{D}_{\text{A-\text{plug-in}}}(M', N))^2 \right] \lesssim \frac{(\log f(k) + \log k)^2}{m^2} + \frac{k \log^2 m}{m^3} + \frac{ kf(k) \log^2 n}{m^2n}
\]
\[
\lesssim \frac{(\log f(k) + \log k)^2}{m^2} \left( 1 + \frac{k \log^2 m}{m \log^2 k} + \frac{ kf(k) \log^2 n}{\log^2 (kf(k))n} \right).
\]

(73)
Due to the assumption that \( m \geq k \) and \( n \geq k f(k) \), \( \frac{k \log^2 m}{m \log k} \lesssim 1 \) and \( \frac{k f(k) \log^2 n}{m \log (k f(k))} \lesssim 1 \). Thus,

\[
\mathbb{E} \left[ (\hat{D}_{\text{A–plug–in}}(M, N) - \hat{D}_{\text{A–plug–in}}(M', N))^2 \right] \lesssim \frac{\log^2 f(k) + \log^2 k}{m^2}.
\] (74)

The second term in (60) can be bounded similarly as follows:

\[
\begin{align*}
\mathbb{E} \left[ (\hat{D}_{\text{A–plug–in}}(M, N) - \hat{D}_{\text{A–plug–in}}(M, N'))^2 \right] &= \mathbb{E} \left[ \left( \phi \left( \frac{M_Y m}{m}, \frac{\hat{N}_{Y_m} + 2}{n} \right) + \phi \left( \frac{M_{Y_m'}}{m}, \frac{\hat{N}_{Y_m'} + 1}{n} \right) \right) \\
&\quad - \phi \left( \frac{M_Y m}{m}, \frac{\hat{N}_{Y_m} + 1}{n} \right) - \phi \left( \frac{M_{Y_m'}}{m}, \frac{\hat{N}_{Y_m'} + 2}{n} \right) \right)^2 Y_m, Y_m'] } \right] \\
&\leq 4 \mathbb{E} \left[ \left( \phi \left( \frac{M_Y m}{m}, \frac{\hat{N}_{Y_m} + 2}{n} \right) - \phi \left( \frac{M_{Y_m'}}{m}, \frac{\hat{N}_{Y_m'} + 1}{n} \right) \right)^2 | Y_m \right] \\
&= 4 \sum_{j=1}^{k} \mathbb{E} \left[ \left( \frac{M_j m}{m} \log \left( 1 + \frac{1}{N_j + 1} \right) \right)^2 Q_j \\
&= \frac{4}{m^2} \sum_{j=1}^{k} \mathbb{E} [M_j^2] \mathbb{E} \left[ \log^2 \left( 1 + \frac{1}{N_j + 1} \right) \right] Q_j.
\end{align*}
\] (75)

Since \( M_j \) follows the binomial distribution, we compute \( \mathbb{E} [M_j^2] \) as follows:

\[
\mathbb{E} [M_j^2] = \mathbb{E} [M_j]^2 + \text{Var}(M_j) = m^2 P_j^2 + m P_j (1 - P_j).
\] (76)

We can also derive,

\[
\mathbb{E} \left[ \log^2 \left( 1 + \frac{1}{N_j + 1} \right) \right] \leq \mathbb{E} \left[ \frac{2}{(N_j + 1)(N_j + 2)} \right] \leq \frac{2}{(n - 1)^2 Q_j^2},
\] (77)

where the last inequality follows from (50). Thus,

\[
\begin{align*}
\mathbb{E} \left[ (\hat{D}_{\text{A–plug–in}}(M, N) - \hat{D}_{\text{A–plug–in}}(M, N'))^2 \right] &= \frac{4}{m^2} \sum_{j=1}^{k} \mathbb{E} [M_j^2] \mathbb{E} \left[ \log^2 \left( 1 + \frac{1}{N_j + 1} \right) \right] Q_j \\
&\leq 4 \sum_{j=1}^{k} \left( P_j^2 + P_j (1 - P_j) \right) \frac{2}{(n - 1)^2 Q_j^2} Q_j \\
&\leq \sum_{j=1}^{k} \frac{P_j}{Q_j} \left( P_j + \frac{1}{m} \right) \frac{2}{n^2} \\
&\lesssim \frac{f(k)}{n^2} + \frac{k f(k)}{n^2 m}.
\end{align*}
\] (78)

Combining (74) and (78), we obtain the following upper bound for the variance:

\[
\text{Var}[\hat{D}_{\text{A–plug–in}}(M, N)] \\
\lesssim \frac{m}{2} \mathbb{E} \left[ (\hat{D}_{\text{A–plug–in}}(M, N) - \hat{D}_{\text{A–plug–in}}(M', N))^2 \right] + \frac{n}{2} \mathbb{E} \left[ (\hat{D}_{\text{A–plug–in}}(M, N) - \hat{D}_{\text{A–plug–in}}(M, N'))^2 \right] \\
\lesssim \frac{\log^2 k}{m} + \frac{\log^2 f(k)}{m} + \frac{f(k)}{n} + \frac{k f(k)}{n m}.
\] (79)
Note that the term $\frac{ kf(k) }{ nm}$ in the variance can be further upper bounded as follows

$$\frac{ kf(k) }{ nm } \leq \frac{ kf(k) }{ n } \frac{ k }{ m } \leq \left( \frac{ kf(k) }{ n } + \frac{ k }{ m } \right)^2. \quad (80)$$

Combining (59), (79) and (80), we obtain the following upper bound on the worse case quadratic risk for augmented plug-in estimator:

$$R(\hat{D}_{\text{plug-in}}, k, m, n, f(k)) \lesssim \left( \frac{ kf(k) }{ n } + \frac{ k }{ m } \right)^2 + \log \frac{ k }{ m } + \log \frac{ f(k) }{ m } + f(k). \quad (81)$$

C Proof of Proposition 2

In this section, we derive the necessary conditions on the sample complexity to guarantee consistency of the augmented plug-in estimator over $M_{k, f(k)}$. We first show that $m = \omega(k)$ and $n = \omega(k f(k))$ is necessary by lower bounding on the squared bias. We then show that $m = \omega(\log^2 f(k))$ is necessary by Le Cam’s two-point method.

C.1 $m = \omega(k)$ and $n = \omega(k f(k))$ is Necessary

It can be shown that the mean square error is lower bounded by the squared bias, which is as follows:

$$\mathbb{E} \left[ (\hat{D}_{\text{plug-in}}(M, N) - D(P\|Q))^2 \right] \geq \left( \mathbb{E} \left[ \hat{D}_{\text{plug-in}}(M, N) - D(P\|Q) \right] \right)^2. \quad (82)$$

Following steps in (52), we have:

$$\mathbb{E}[\hat{D}_{\text{plug-in}}(M, N) - D(P\|Q)] = \mathbb{E} \left( \sum_{i=1}^{k} \left( \frac{ M_i }{ m } \log \frac{ M_i }{ m } - P_i \log P_i \right) \right) + \mathbb{E} \left( \sum_{i=1}^{k} P_i \log \frac{ n Q_i }{ N_i + 1 } \right). \quad (83)$$

The first term in (83) is the bias of plug-in entropy estimator. As shown in [13] and [21], the worst case risk of the first term can be bounded as follows:

$$\left( \frac{ k }{ m } \wedge 1 \right) \leq \sup_P \mathbb{E} \left( \sum_{i=1}^{k} \left( \frac{ M_i }{ m } \log \frac{ M_i }{ m } - P_i \log P_i \right) \right) \leq \log \left( 1 + \frac{ k - 1 }{ m } \right). \quad (84)$$

As shown in (54) and (57), we have the following bound on the second term in (83):

$$- \frac{ kf(k) }{ n } \leq \mathbb{E} \left( \sum_{i=1}^{k} P_i \log \frac{ n Q_i }{ N_i + 1 } \right) \leq \sum_{i=1}^{k} P_i \log \mathbb{E} \left[ \frac{ n Q_i }{ N_i + 1 } \right]. \quad (85)$$

In order to obtain a tight bound for the bias, we choose the following $(P, Q)$:

$$P = \left( \frac{ 1 }{ k }, \frac{ 1 }{ k }, \cdots, \frac{ 1 }{ k } \right), \quad (86)$$

$$Q = \left( \frac{ 1 }{ kf(k) }, \cdots, \frac{ 1 }{ kf(k) }, 1 - \frac{ k - 1 }{ kf(k) } \right). \quad (87)$$
It can be verified that $P$ and $Q$ satisfy the density ratio constraint. For such a $(P, Q)$ pair, we have

$$
\sum_{i=1}^{k} P_i \log \mathbb{E} \left[ \frac{nQ_i}{N_i + 1} \right] = \sum_{i=1}^{k} P_i \log \left( \frac{nQ_i}{(n+1)Q_i} (1 - (1 - Q_i)^{n+1}) \right)
$$

$$
\leq \sum_{i=1}^{k} P_i \log(1 - (1 - Q_i)^{n+1})
$$

$$
\leq \sum_{i=1}^{k} P_i \log(1 - \frac{1}{k f(k)})^{n+1}
$$

$$
\leq - \frac{k-1}{k} (1 - \frac{1}{k f(k)})^{n+1}
$$

$$
= - \frac{k-1}{k} (1 - \frac{1}{k f(k)})^{k f(k)(n+1) \frac{1}{x f(k)}}. \quad (88)
$$

Since $(1 - x)^{1/x}$ is decreasing on $[0, 1]$, and $\lim_{x \to 0} (1 - x)^{1/x} = \frac{1}{e}$. For sufficiently large $k$, $1/(k f(k))$ is close to $0$, and thus we have,

$$
e^{-1} > (1 - \frac{1}{k f(k)})^{k f(k)} > e^{-\beta_0}, \quad (89)
$$

where $\beta_0 > 1$ is a constant. Thus, we get a tighter bound as follows:

$$
- \frac{k f(k)}{n} \leq \mathbb{E} \left( \sum_{i=1}^{k} \left( P_i \log Q_i - \frac{M_i}{m} \log N_i + 1 \right) \right) \leq - \frac{k-1}{k} \exp(-\frac{\beta_0 n}{k f(k)}). \quad (90)
$$

Combining (88) and (90), we have,

$$
\frac{k}{m} \wedge 1 - k f(k) \leq \sup_{(P, Q) \in [\mathcal{M}_{k,f}(k)]} \mathbb{E}[\hat{D}_{\text{plug-in}}(M, N) - D(P||Q)], \quad (91a)
$$

$$
\mathbb{E}[\hat{D}_{\text{plug-in}}(M, N) - D(P||Q)] \leq \log \left( 1 + \frac{k}{m} \right) - \frac{k-1}{k} \exp(-\frac{\beta_0 n}{k f(k)}). \quad (91b)
$$

1) If $m \ll k$ and $n = \omega(k f(k))$,

$$
\sup_{(P, Q) \in [\mathcal{M}_{k,f}(k)]} \mathbb{E}[\hat{D}_{\text{plug-in}}(M, N) - D(P||Q)] \geq \frac{k}{m} \wedge 1 \frac{k f(k)}{n} \rightarrow c, \quad (92)
$$

where $c$ is some positive constant. The bias is lower bounded by a positive constant. Hence, for sufficiently large $k$, the augmented plug-in estimator is not consistent if $m = \mathcal{O}(k)$ and $n = \omega(k f(k))$.

2) If $m = \omega(k)$ and $n \ll k f(k)$, let $n \leq \beta_1 k f(k)$, with $0 < \beta_1 < 1$,

$$
\mathbb{E}[\hat{D}_{\text{plug-in}}(M, N) - D(P||Q)] \leq \log \left( 1 + \frac{k}{m} \right) - \frac{k-1}{k} \exp(-\frac{\beta_0 n}{k f(k)})
$$

$$
\rightarrow - \frac{k-1}{k} \exp(-\frac{\beta_0 n}{k f(k)})
$$

$$
\leq - \frac{k-1}{k} e^{-\beta_0 \beta_1}. \quad (93)
$$

The bias is upper bounded by a negative constant. Hence, for sufficiently large $k$, the augmented plug-in estimator is not consistent if $m = \omega(k)$ and $n \ll k f(k)$.

3) If $m \ll k$ and $n \ll k f(k)$, We cannot get a useful lower bound on the squared bias from (91a) and (91b) in this case using the chosen pair of $(P, Q)$. Hence, we need to choose other pairs of $(P, Q)$. 

20
Again, we choose $P$ to be the uniform distribution. The bias of the augmented plug-in estimator can be decomposed into: 1) bias due to estimating $\sum_{i=1}^{k} P_i \log P_i$; and 2) bias due to estimating $\sum_{i=1}^{k} P_i \log Q_i$. It can be shown that the first bias is always positive, because the uniform distribution achieves the largest entropy for a given alphabet size $k$. The second bias is always negative for any distribution $Q$. Hence, the two bias terms may cancel out partially or even fully. Thus, to show the risk is bounded away from zero, the idea is to first determine which bias dominates, and then to construct a pair of distributions accordingly such that the dominant bias is either lower bounded by a positive constant or upper bounded by a negative constant.

If $k \frac{m}{n} \geq (1 + \epsilon) \frac{\alpha f(k)}{n}$, where $\epsilon > 0$ and $0 < \alpha < 1$ are constants, and which implies that the number of samples drawn from $P$ is relatively smaller than the number of samples drawn from $Q$, the first bias dominates. We construct:

$$P = \left( \frac{1}{k}, \frac{1}{k}, \cdots, \frac{1}{k} \right), \quad Q = \left( \frac{1}{\alpha k f(k)}, \cdots, \frac{1}{\alpha k f(k)}, 1 - k \frac{1}{\alpha k f(k)} \right). \quad (94)$$

Let $\alpha > \frac{1}{f(k)}$, then $1 - k \frac{1}{\alpha f(k)} > \frac{1}{k}$. It can be verified that the density ratio between $P$ and $Q$ is bounded by $\alpha f(k) \leq f(k)$. Since $P$ is a uniform distribution, which has the maximal entropy, then the bias of entropy estimation can be written as

$$E \left( \sum_{i=1}^{k} \left( \frac{M_i}{m} \log \frac{M_i}{m} - P_i \log P_i \right) \right) = \log k + \log \left( \sum_{i=1}^{k} \frac{M_i}{m} \log \frac{M_i}{m} \right). \quad (95)$$

Since $m \lesssim k$, we assume $m \leq k$. Thus we obtain following inequality:

$$\sum_{i=1}^{k} \frac{M_i}{m} \log \frac{M_i}{m} \geq - \log m. \quad (96)$$

Combining with (84), we have

$$\log \frac{k}{m} \leq E \left( \sum_{i=1}^{k} \left( \frac{M_i}{m} \log \frac{M_i}{m} - P_i \log P_i \right) \right) \leq \log \left( 1 + \frac{k}{m} \right). \quad (97)$$

And for the above choice of $(P, Q)$,

$$E \left( \sum_{i=1}^{k} \left( P_i \log Q_i - \frac{M_i}{m} \log \frac{N_i + 1}{n} \right) \right) = - \sum_{i=1}^{k} P_i E \left[ \log \frac{N_i + 1}{nQ_i} \right]$$

$$\geq - \sum_{i=1}^{k} P_i \log \frac{N_i + 1}{nQ_i}$$

$$= - \sum_{i=1}^{k} \log \left( 1 + \frac{1}{nQ_i} \right)$$

$$\geq - \log \left( 1 + \frac{\alpha f(k)}{n} \right). \quad (98)$$

Combining with (97), we obtain the following lower bound,

$$E[\hat{D}_{A-\text{plug-in}}(M, N) - D(P\|Q)] \geq \log \frac{k}{m} - \log \left( 1 + \frac{\alpha f(k)}{n} \right)$$

$$\geq \log \left( 1 + \epsilon \frac{\alpha f(k)}{n} \right) - \log \left( 1 + \frac{\alpha f(k)}{n} \right)$$

$$\geq \log(1 + \epsilon) > 0. \quad (99)$$
If $\epsilon > c$, then this implies that the worst case quadratic error is also lower bounded by a positive constant, and hence the augmented plug-in estimator is not consistent in this case.

If $\frac{k}{m} < (1 + \epsilon)\frac{akf(k)}{n}$, which implies that the number of samples drawn from $P$ is relatively larger than the number of samples drawn from $Q$, then the second bias dominates. We construct the following distributions:

$$P = \left(\frac{1}{k}, \frac{1}{k}, \cdots, \frac{1}{k}\right), \quad Q = \left(\frac{1}{kf(k)}, \cdots, \frac{1}{kf(k)}, 1 - \frac{k - 1}{kf(k)}\right).$$

(100)

Following from [SS], we have

$$\mathbb{E}[\hat{D}_{A\text{-plug-in}}(M, N) - D(P\|Q)] \leq \log \left(1 + \frac{k}{m}\right) + \frac{k - 1}{k} \log(1 - \exp(-\beta_0 n))$$

$$= \frac{k - 1}{k} \left(\log \left(1 + \frac{k}{m}\right) + \log(1 - \exp(-\beta_0 n))\right) + \frac{1}{k} \log \left(1 + \frac{k}{m}\right)$$

$$\leq \frac{k - 1}{k} \log \left((1 + (1 + \epsilon)\frac{akf(k)}{n})(1 - \exp(-\beta_0 n))\right) + \frac{1}{m}.$$  

(101)

Let $\beta \triangleq (1 + \epsilon)\alpha$, and $t = \frac{n}{kj(k)}$. Since $n \lesssim (kf(k))$, we assume that $n \leq kf(k)$. Then, we define the function

$$h(t) \triangleq (1 + \frac{\beta}{t})(1 - \exp(-\beta_0 t)), \quad t \in (0, 1).$$

(102)

For sufficiently large $k$, we choose $\beta_0 = 1.05$. Then for any $\beta < 0.1$, we have

$$h(t) = (1 + \frac{\beta}{t})(1 - \exp(-1.05t)) < 0.8, \quad \forall t \in (0, 1).$$

(103)

Thus, if $(1 + \epsilon)\alpha < 0.1$, $m \geq 5$, $f(k) > 10$, and $n \leq kf(k)$, for sufficiently large $k$,

$$\mathbb{E}[\hat{D}_{A\text{-plug-in}}(M, N) - D(P\|Q)] \leq \frac{k - 1}{k} \log \left((1 + (1 + \epsilon)\frac{akf(k)}{n})(1 - \exp(-\frac{\beta_0 n}{kf(k)}))\right) + \frac{1}{m}$$

$$\rightarrow \log \left((1 + (1 + \epsilon)\frac{akf(k)}{n})(1 - \exp(-\frac{1.05n}{kf(k)}))\right) + \frac{1}{m}$$

$$\leq \log 0.8 + \frac{1}{m} < 0.$$  

(104)

Hence, for $m \lesssim k$, $n \lesssim kf(k)$ and $\frac{k}{m} < (1 + \epsilon)\frac{akf(k)}{n}$, we can always choose $\alpha$ and $\epsilon$ properly, such that, for large $k$,

$$\mathbb{E}[\hat{D}_{A\text{-plug-in}}(M, N) - D(P\|Q)] \leq \frac{k - 1}{k} \log \left((1 + (1 + \epsilon)\frac{akf(k)}{n})(1 - \exp(-\frac{2n}{kf(k)}))\right) + \frac{1}{m} < c_1,$$

(105)

where $c_1$ is a negative constant. This implies that the worse case quadratic risk is lower bounded by a positive constant, and the augmented plug-in estimator is not consistent in this case.

Therefore, for both cases that $\frac{k}{m} < (1 + \epsilon)\frac{akf(k)}{n}$ and $\frac{k}{m} \geq (1 + \epsilon)\frac{akf(k)}{n}$, we can construct $(P, Q)$ such that if $m \lesssim k$, $n \lesssim kf(k)$, the augmented plug-in estimator is not consistent.

C.2 $m = \omega\left(\log^2 f(k)\right)$ is Necessary

To show that the augmented plug-in estimator is not consistent when $m = \mathcal{O}(\log^2 f(k))$, we use the minimax risk as the lower bound of the worst case risk for augmented plug-in estimator. To this end, we
apply the Le Cam’s two-point method. We first construct two pairs of distributions as follows:

\[ P_1 = \left[ \frac{1}{3(k-1)}, \ldots, \frac{1}{3(k-1)} \right], \quad P_2 = \left[ \frac{1}{1 - \epsilon}, \ldots, \frac{1}{1 - \epsilon}, \frac{1}{3} \right], \]

\[ Q_1 = Q_2 = \left[ \frac{1}{3(k-1)f(k)}, \ldots, \frac{1}{3(k-1)f(k)}, 1 - \frac{1}{3f(k)} \right]. \tag{108} \]

The above distributions satisfy:

\[ D(P_1 || Q_1) = \frac{1}{3} \log f(k) + \frac{2}{3} \log \frac{1}{1 - \frac{1}{3f(k)}}, \]

\[ D(P_2 || Q_2) = \frac{1 - \epsilon}{3} \log(1 - \epsilon)f(k) + \left(1 - \frac{1 - \epsilon}{3} \right) \log \frac{1 - \frac{1 - \epsilon}{3}}{1 - \frac{1}{3f(k)}}, \]

\[ D(P_1 || P_2) = \frac{1}{3} \log \frac{1}{1 - \epsilon} + \frac{2}{3} \log \frac{1}{1 - \frac{1}{3f(k)}}. \]

We set \( \epsilon = \frac{1}{\sqrt{m}} \), and then obtain

\[ D(P_1 || P_2) = \frac{1}{3} \log \left(1 + \frac{\epsilon}{1 - \epsilon}\right) + \frac{2}{3} \log \left(1 - \frac{\epsilon}{2 + \epsilon}\right) \]

\[ \leq \frac{\epsilon}{3(1 - \epsilon)} - \frac{2}{3} \frac{\epsilon}{2 + \epsilon} = \frac{3 \epsilon^2}{(1 - \epsilon)(2 + \epsilon)} \leq \frac{3}{m}. \] \tag{112} \]

Furthermore, we show that

\[ D(P_1 || Q_1) - D(P_2 || Q_2) \]

\[ = \frac{1}{3} \log f(k) + \frac{2}{3} \log \frac{1}{1 - \frac{1}{3f(k)}} - \frac{1 - \epsilon}{3} \log(1 - \epsilon)f(k) - \left(1 - \frac{1 - \epsilon}{3} \right) \log \frac{1 - \frac{1 - \epsilon}{3}}{1 - \frac{1}{3f(k)}} \]

\[ = \frac{1}{3} \log \frac{1}{1 - \epsilon} + \frac{2}{3} \log(1 - \epsilon)f(k) + \frac{2}{3} \log \frac{2 + \epsilon}{3 - \frac{1}{f(k)}} - \frac{\epsilon}{3} \log \frac{1}{1 - \frac{1}{3f(k)}} \]

\[ \leq \frac{1}{3} \log \frac{1}{1 - \epsilon} - \frac{4}{3} \log \frac{2 + \epsilon}{(1 - \epsilon)(3f(k) - 1)}. \tag{113} \]

which implies

\[ (D(P_1 || Q_1) - D(P_2 || Q_2))^2 \geq \epsilon^2 \log^2 \frac{2}{(3f(k) - 1)} \times \frac{\log^2 f(k)}{m}, \]

as \( m \to \infty \). Now applying Le Cam’s two-point method, we obtain

\[ R^*(k, m, n, f(k)) \geq \frac{1}{16} (D(P_1 || Q_1) - D(P_2 || Q_2))^2 \exp \left(-m(D(P_1 || P_2) - nD(Q_1 || Q_2))\right). \tag{115} \]

Clearly, if \( m = \mathcal{O}(\log^2 f(k)) \), the minimax quadratic risk does not converge to 0 as \( k \to \infty \), which further implies that the augmented plug-in estimator is not consistent for this case.

**D  Proof of Lemma 1**

Here we prove the inequality (21) that connects the minimax risk under the i.i.d. sampling model (8) to that under the Poisson sampling model (20). We first prove the left hand side of (21). Recall that
Further define, prove the right hand side of (21). By the minimax theorem, then for \( m \) decreasing in \( m \),

\[
\pi \text{ where } \alpha \text{ unclear whether the sequence of Bayesian risks } E[R_{\pi}]
\]

where the last inequality follows from the Chernoff bound \( \mathbb{P} [\text{Poi}(2n) \leq n] \leq \exp(-1 - \log 2)n \). We then prove the right hand side of (21). By the minimax theorem,

\[
R^*(k, m, n, f(k)) = \sup_{\pi} F_{\pi} \left[ \mathbb{E} \left( D(M, N) - D(P \| Q) \right)^2 \right],
\]

where \( \pi \) ranges over all probability distribution pairs on the simplex \( \mathcal{M}_{k, f(k)} \) and the expectation is over \( (P, Q) \sim \pi \).

Fix a prior \( \pi \) and an arbitrary sequence of estimators \( \{ \hat{D}_{m,n} \} \) indexed by the sample size \( m \) and \( n \). It is unclear whether the sequence of Bayesian risks \( \alpha_{m,n} = E[(\hat{D}_{m,n}(M, N) - D(P \| Q))^2] \) with respect to \( \pi \) is decreasing in \( m \) or \( n \). However, we can define \( \{ \tilde{\alpha}_{i,j} \} \) as

\[
\tilde{\alpha}_{0,0} = \alpha_{0,0}, \quad \tilde{\alpha}_{i,j} = \alpha_{i,j} \land \alpha_{i-1,j} \land \alpha_{i,j-1}.
\]

Further define,

\[
\tilde{D}_{m,n}(M, N) \triangleq \begin{cases} 
\hat{D}_{m,n}(M, N), & \text{if } \tilde{\alpha}_{m,n} = \alpha_{m,n}; \\
\hat{D}_{m-1,n}(M, N), & \text{if } \tilde{\alpha}_{m,n} = \alpha_{m-1,n}; \\
\hat{D}_{m,n-1}(M, N), & \text{if } \tilde{\alpha}_{m,n} = \alpha_{m,n-1}.
\end{cases}
\]

Then for \( m' \sim \text{Poi}(m/2) \) and \( n' \sim \text{Poi}(n/2) \), and \( (P, Q) \sim \pi \), we have

\[
\mathbb{E} \left[ (\tilde{D}_{m',n'}(M', N') - D(P \| Q))^2 \right]
\]

\[
\geq \frac{1}{4} \mathbb{E} \left[ (\hat{D}_{m,n}(M, N) - D(P \| Q))^2 \right],
\]

where (a) is due to the Markov’s inequality: \( \mathbb{P}[\text{Poi}(n/2) \geq n] \leq \frac{1}{2} \). If we take infimum of the left hand side over \( \hat{D}_{m,n} \), then take supremum of both sides over \( \pi \), and use the Baysian risk as a lower bound for the minimax risk, then we can show that

\[
\bar{R}^*(k, m, n, f(k)) \geq \frac{1}{4} R^*(k, m, n, f(k)).
\]
E Proof of Proposition 3

E.1 Bounds Using Le Cam’s Two-Point Method

E.1.1 Proof of $R^*(k, m, n, f(k)) \gtrsim \frac{\log^2 f(k)}{m}$

Following the same steps in Appendix C.2, we can show

$$R^*(k, m, n, f(k)) \gtrsim (D(P_1 || Q_1) - D(P_2 || Q_2))^2 \gtrsim \frac{\log^2 f(k)}{m}. \quad (122)$$

E.1.2 Proof of $R^*(k, m, n, f(k)) \gtrsim \frac{f(k)}{n}$

We construct two pairs of distributions as follows:

$$P_1 = P_2 = \left( \frac{1}{3(k-1)}0, \frac{1}{3(k-1)}0, \ldots, \frac{5}{6} \right), \quad (123)$$

$$Q_1 = \left( \frac{1}{2(k-1)f(k)}, \ldots, \frac{1}{2(k-1)f(k)} \right), \quad (124)$$

$$Q_2 = \left( \frac{1 - \epsilon}{2(k-1)f(k)}, \frac{1 + \epsilon}{2(k-1)f(k)}, \ldots, \frac{1 - \epsilon}{2(k-1)f(k)} \right). \quad (125)$$

It can be verified that if $\epsilon < \frac{1}{3}$, then the density ratio is bounded by $\frac{2f(k)}{3(1-\epsilon)} \leq f(k)$. We set $\epsilon = \sqrt{\frac{f(k)}{n}}$. The above distributions satisfy:

$$D(Q_1 || Q_2) = \frac{1}{4f(k)} \log \frac{1}{1+\epsilon} + \frac{1}{4f(k)} \log \frac{1}{1-\epsilon}, \quad (126)$$

$$D(P_1 || Q_1) - D(P_2 || Q_2) = \frac{1}{6} \log(1-\epsilon) \leq \frac{\epsilon}{6}. \quad (127)$$

Due to $\epsilon = \sqrt{\frac{f(k)}{n}}$, it can be shown that

$$D(Q_1 || Q_2) = \frac{1}{4f(k)} \log \left(1 + \frac{\epsilon^2}{1-\epsilon^2}\right) \leq \frac{1}{4f(k)} \frac{\epsilon^2}{1-\epsilon^2} < \frac{\epsilon^2}{f(k)} = \frac{1}{n}. \quad (128)$$

We apply Le Cam’s two point method,

$$R^*(k, m, n, f(k)) \geq \frac{1}{16} (D(P_1 || Q_1) - D(P_2 || Q_2))^2 \exp \left( -m(D(P_1 || P_2) - nD(Q_1 || Q_2)) \right)$$

$$\gtrsim (D(P_1 || Q_1) - D(P_2 || Q_2))^2 \gtrsim \epsilon^2 = \frac{f(k)}{n}. \quad (129)$$
E.2 Bounds Using Generalized Le Cam’s Method

E.2.1 Proof of $\tilde{R}^*(k, m, n, f(k)) \gtrsim \left(\frac{k}{m \log k}\right)^2$

Let $Q_0$ denote the uniform distribution. The minimax risk with Poisson sampling is lower bounded as follows:

$$\tilde{R}^*(k, m, n, f(k)) = \inf_{\hat{D}} \sup_{(P, Q) \in \mathcal{M}_{k,f(k)}} \mathbb{E}[\hat{D}(M, N) - D(P\|Q)]^2$$

$$\gtrsim \inf_{\hat{D}} \sup_{(P, Q_0) \in \mathcal{M}_{k,f(k)}} \mathbb{E}[\hat{D}(M, Q_0) - D(P\|Q)]^2$$

$$\tilde{\tilde{R}}^*(k, m, Q_0, f(k)).$$  \hspace{1cm} (130)

If $Q = Q_0$ is known, then estimating the KL divergence between $P$ and $Q_0$ is equivalent to estimating the entropy of $P$, because

$$D(P\|Q_0) = \sum_{i=1}^{k} P_i \log P_i + P_i \log \frac{1}{Q_i}$$

$$= H(P) + \log k.$$  \hspace{1cm} (131)

Hence, $\tilde{R}^*(k, m, Q_0, f(k))$ is equivalent to the following minimax risk of estimating the entropy of distribution $P$ with $P_i \leq \frac{f(k)}{k}$ for $1 \leq i \leq k$ such that the ratio between $P$ and $Q$ is upper bounded by $f(k)$.

$$\tilde{R}^*(k, m, Q_0, f(k)) = \inf_{\hat{H}} \sup_{P: P_i \leq \frac{f(k)}{k}} \mathbb{E}[\hat{H}(M) - H(P)]^2.$$  \hspace{1cm} (132)

If $m \gtrsim \frac{k}{\log k}$, as shown in [13], the minimax lower bound for estimating entropy is given by

$$\inf_{\hat{H}} \sup_{P} \mathbb{E}[\hat{H}(M) - H(P)]^2 \gtrsim \left(\frac{k}{m \log k}\right)^2.$$  \hspace{1cm} (133)

The supremum is achieved for $P_i \leq \frac{\log^2 k}{k}$. Comparing this result to (132), if $f(k) \geq \log^2 k$, then

$$\frac{\log^2 k}{k} \leq \frac{f(k)}{k}.$$  \hspace{1cm} (134)

Thus, we can use the minimax lower bound of entropy estimation as the lower bound for divergence estimation on $\mathcal{M}_{k,f(k)}$.

$$\tilde{R}^*(k, m, n, f(k)) \gtrsim \tilde{R}^*(k, m, Q_0, f(k)) \gtrsim \left(\frac{k}{m \log k}\right)^2.$$  \hspace{1cm} (135)

E.2.2 Proof of $\tilde{R}^*(k, m, n, f(k)) \gtrsim \left(\frac{f(k)}{n \log k}\right)^2$

We set $P = P_0$, where

$$P_0 = \begin{bmatrix} f(k) \frac{n \log k}{k} & \cdots & f(k) \frac{n \log k}{k} - \frac{(k-1)f(k)}{n \log k} \end{bmatrix}.$$  \hspace{1cm} (136)

Since $n \gtrsim \frac{k f(k)}{\log k}$, we assume that $n \geq \frac{k f(k)}{\log k}$. Then, we have $0 \leq 1 - \frac{(k-1)f(k)}{n \log k} \leq 1$. Hence, $P_0$ is a well-defined probability distribution.
If \( P = P_0 \) and is known, then estimating the KL divergence between \( P \) and \( Q \) is equivalent to estimating the following function:

\[
D(P_0\|Q) = \sum_{i=1}^{k-1} \frac{f(k)}{n \log k} \log \frac{f(k)/Q_i}{(k-1)f(k)/Q_k} + (1 - \frac{(k-1)f(k)}{n \log k}) \log \frac{1 - (k-1)f(k)/Q_k}{Q_k}.
\]  

(137)

which is further equivalent to estimating

\[
\sum_{i=1}^{k-1} \frac{f(k)}{n \log k} \log \frac{1}{Q_i} + (1 - \frac{(k-1)f(k)}{n \log k}) \log \frac{1}{Q_k}.
\]  

(138)

We further consider the following subset of \( \mathcal{M}_{k,f(k)} \):

\[
\mathcal{N}_{k,f(k)} \triangleq \{(P_0, Q) \in \mathcal{M}_{k,f(k)} : \frac{1}{n \log k} \leq Q_i \leq \frac{c_4 \log k}{n}, \text{ for } 1 \leq i \leq k-1\}.
\]  

(139)

The minimax risk can be lower bounded as follows:

\[
R^*(k, m, n, f(k)) = \inf_{\hat{D}} \sup_{(P, Q) \in \mathcal{M}_{k,f(k)}} \mathbb{E}[\hat{D}(M, N) - D(P\|Q)^2] \geq \inf_{\hat{D}} \sup_{(P_0, Q) \in \mathcal{N}_{k,f(k)}} \mathbb{E}[\hat{D}(P_0, N) - D(P_0\|Q)^2] \triangleq R^*_N(k, P_0, n, f(k)).
\]  

(140)

For \( 0 < \epsilon < 1 \), we introduce the following set of approximate probability vectors:

\[
\mathcal{N}_{k,f(k)}(\epsilon) \triangleq \{(P_0, Q) \in \mathbb{R}^k_+ : |\sum_{i=1}^k Q_i - 1| \leq \epsilon, \frac{1}{n \log k} \leq Q_i \leq \frac{c_4 \log k}{n}, \text{ for } 1 \leq i \leq k-1\}.
\]  

(141)

Note that \( Q \) is not a distribution, but \( P_0 \) and \( Q \) still satisfy the density ratio upper bound if \((P_0, Q) \in \mathcal{N}_{k,f(k)}\), due to our construction. And the set \( \mathcal{N}_{k,f(k)}(\epsilon) \) reduces to \( \mathcal{N}_{k,f(k)} \) if \( \epsilon = 0 \).

We further consider the minimax quadratic risk \((140)\) for Poisson sampling on the set \( \mathcal{N}_{k,f(k)}(\epsilon) \) as follows:

\[
\tilde{R}^*_N(k, P_0, n, f(k), \epsilon) = \inf_{\hat{D}} \sup_{(P_0, Q) \in \mathcal{N}_{k,f(k)}(\epsilon)} \mathbb{E}[\hat{D}(P_0, N) - D(P_0, Q)^2],
\]  

(142)

where \( N_i \sim \text{Poi}(nQ_i) \), for \( 1 \leq i \leq k \). The risk \((142)\) is connected to the risk \((140)\) for multinomial sampling by the following lemma:

**Lemma 2.** For any \( k, n \in \mathbb{N} \) and \( \epsilon < 1/3 \),

\[
R^*(k, P_0, \frac{n}{2}, f(k)) \geq \frac{1}{2} \tilde{R}^*_N(k, P_0, n, f(k), \epsilon) - \log^2 f(k) \exp \left( -\frac{n}{50} \right) - \log^2 (1 + \epsilon).
\]  

(143)

**Proof.** See Appendix E.2.3.

For \((P_0, Q) \in \mathcal{N}_{k,f(k)}(\epsilon)\), we then apply the generalized Le Cam’s method which involves two composite hypothesis as follows:

\[
H_0 : D(P_0\|Q) \leq t \quad \text{versus} \quad H_1 : D(P_0\|Q) \geq t + \frac{(k-1)f(k)}{n \log k} d.
\]  

(144)

In the following we construct tractable prior distributions. Let \( V \) and \( V' \) be two \( \mathbb{R}^+ \) valued random variables defined on the interval \( \left[ \frac{1}{n \log k}, \frac{c_4 \log k}{n} \right] \) and have equal mean \( \mathbb{E}(V) = \mathbb{E}(V') = \alpha \). We construct two random vectors

\[
Q = (V_1, \ldots, V_{k-1}, 1 - (k-1)\alpha) \quad \text{and} \quad Q' = (V'_1, \ldots, V'_{k-1}, 1 - (k-1)\alpha)
\]  

(145)
consisting of $k - 1$ i.i.d. copies of $V$ and $V'$ and a deterministic term $1 - (k - 1)\alpha$, respectively. It can be verified that $(P_0, Q), (P_0, Q') \in \mathcal{N}_{k, f(k)}(\epsilon)$ satisfy the density ratio constraint, and the averaged divergences are separated by the distance of

$$\left| \mathbb{E}[D(P_0 \| Q)] - \mathbb{E}[D(P_0 \| Q')] \right| = \frac{(k - 1)f(k)}{n \log k} \left| \mathbb{E}[\log V] - \mathbb{E}[\log V'] \right|. \quad (146)$$

Thus, if we construct $V$ and $V'$ such that

$$\left| \mathbb{E}[\log V] - \mathbb{E}[\log V'] \right| \geq d \quad (147)$$

then the construction in (145) satisfy (144), serving as the two composite hypothesis which are separated.

By such a construction, we have the following lemma via generalized Le Cam’s method:

**Lemma 3.** Let $V$ and $V'$ be random variables such that $V, V' \in \left[ \frac{1}{n \log k}, \frac{c_4 \log k}{n} \right]$, and $\mathbb{E}[V] = \mathbb{E}[V'] = \alpha$, and $\left| \mathbb{E}[\log V] - \mathbb{E}[\log V'] \right| \geq d$. Then,

$$\bar{R}_\alpha(k, P_0, n, f(k), \epsilon) \geq \frac{(k - 1)f(k)^2}{32} \left[ 1 - \frac{2(k - 1)c_4^2 \log^2 k}{n^2 \epsilon^2} - \frac{32(\log n + \log \log k)^2}{(k - 1)d^2} - kTV(\mathbb{E}[Poi(nV)], \mathbb{E}[Poi(nV')]) \right]. \quad (148)$$

**Proof.** See Appendix E.2.3. \hfill \square

To establish the impossibility of hypothesis testing between $V$ and $V'$, we also have the following lemma which provides an upper bound on the total variation of the two mixture Poisson distributions.

**Lemma 4.** [13] Lemma 3] Let $V$ and $V'$ be random variables on $\left[ \frac{1}{n \log k}, \frac{c_4 \log k}{n} \right]$. If $\mathbb{E}[V^j] = \mathbb{E}[V'^j]$ for $j = 1, \ldots, L$, and $L > \frac{2c_4 \log k}{n}$, then,

$$TV(\mathbb{E}[Poi(nV)], \mathbb{E}[Poi(nV')]) \leq 2 \exp \left( - \left( \frac{L}{2} \log \frac{L}{2c_4 \log k} - 2c_4 \log k \right) \right) \wedge 1. \quad (149)$$

What remains is to construct $V$ and $V'$ to maximize $d = \left| \mathbb{E}[\log V'] - \mathbb{E}[\log V] \right|$, subject to the constraints in Lemma 3. Consider the following optimization problem over random variable $X$ and $X'$:

$$\mathcal{E}^* = \max \mathbb{E}[\log X] - \mathbb{E}[\log X']$$

s.t. $\mathbb{E}[X^j] = \mathbb{E}[X'^j], \quad j = 1, \ldots, L$

$$X, X' \in \left[ \frac{1}{c_4 \log^2 k}, 1 \right]. \quad (150)$$

As shown in Appendix E in [13], the maximum $\mathcal{E}^*$ is equal to twice the approximation error of polynomial degree $L$:

$$\mathcal{E}^* = 2E_L(\log, \left[ \frac{1}{c_4 \log^2 k}, 1 \right]). \quad (151)$$

The following lemma provides a lower bound on the approximation error of polynomial degree $L$ over $[L^{-2}, 1]$ for function $\log x$.

**Lemma 5.** [13] Lemma 4] There exist universal positive constants $c, c', L_0$ such that for any $L > L_0$,

$$E_{[cL]}(\log[L^{-2}, 1]) > c'. \quad (152)$$
Let $X$ and $X'$ be the maximizer of (150). We construct $V = \frac{c_1 \log k}{n} X$ and $V' = \frac{c_2 \log k}{n} X'$, such that $V, V' \in [\frac{1}{n \log k}, \frac{c_4 \log k}{n}]$. And it can be shown that

$$E[\log \frac{1}{V}] - E[\log \frac{1}{V'}] = \mathcal{E}^*, \quad (153)$$

and $V$ and $V'$ match up to $L$-th moment. We choose the value of $d$ to be $\mathcal{E}^*$.

Hence, we set $L = \lfloor c \log k \rfloor$, then from Lemma 3 $d = \mathcal{E}^* > 2\epsilon'$. We further assume that $\log^2 n \leq c_5 k$, set $c_4$ and $c_5$ such that $2c_4^2 + \frac{c_5}{2c_4} < 1$ and $\frac{1}{2c_4} - 2c_4 < 1$. Then from Lemma 3 and Lemma 2 with $\epsilon = \frac{\sqrt{c \log k}}{n}$, the minimax risk is lower bounded as follows:

$$\hat{R}^*(k, P_0, n, f(k)) \geq \left( \frac{k f(k)}{n \log k} \right)^2. \quad (154)$$

E.2.3 Proof of Lemma 2

Fix $\delta > 0$ and $(P_0, Q) \in \mathcal{N}_{k, f(k)}(0)$. Let $\hat{D}(P_0, n)$ be a near optimal minimax estimator for $D(P_0||Q)$ with $n$ samples such that

$$\sup_{(P_0, Q) \in \mathcal{N}_{k, f(k)}(0)} E[(\hat{D}(P_0, n) - D(P_0||Q))^2] \leq \delta + R^*(k, P_0, n, f(k)). \quad (155)$$

For any $(P_0, Q) \in \mathcal{N}_{k, f(k)}(\epsilon)$, $Q$ is approximately a distribution. We normalize $Q$ to be a probability distribution, i.e., $\frac{Q}{\sum_{i=1}^{k} Q_i}$, and then we have,

$$D(P_0||Q) = \sum_{i=1}^{k} P_{0,i} \log \frac{P_{0,i}}{Q_i} = -\log \sum_{i=1}^{k} Q_i + D(P_0||Q) = \frac{Q}{\sum_{i=1}^{k} Q_i}). \quad (156)$$

Fix distributions $(P_0, Q) \in \mathcal{M}_{k, f(k)}(\epsilon)$. Let $N = (N_1, \ldots, N_k)$, and $N_i \sim \text{Poi}(n Q_i)$. And let $n' = \sum N_i \sim \text{Poi}(n \sum Q_i)$. We construct an estimator for the Poisson sampling by

$$\tilde{D}(P_0, N) = \hat{D}(P_0, n'). \quad (157)$$

By the triangle inequality, we obtain

$$\frac{1}{2}(\tilde{D}(P_0, N) - D(P_0||Q))^2 \leq (\tilde{D}(P_0, N) - D(P_0||Q))^2 = (D(P_0||Q) - D(P_0||Q))^2 \leq (D(P_0||Q))^2 + \log 2 + \epsilon. \quad (158)$$

Since $n' = \sum N_i \sim \text{Poi}(n \sum Q_i)$, we can show that

$$E[(\tilde{D}(P_0, N) - D(P_0||Q))^2] \leq \sum_{j=1}^{\infty} E[(\tilde{D}(P_0, j) - D(P_0||Q))^2 | n' = j] \mathbb{P}(n' = j) \leq \sum_{j=1}^{\infty} R^*(k, P_0, j, f(k)) \mathbb{P}(n' = j) + \delta. \quad (159)$$
We note that for fixed $k$, $R^*(k, P_0, j, f(k))$ is a monotonically decreasing function with respect to $n$. We also have $R^*(k, P_0, j, f(k)) \leq \log^2 f(k)$, because for any $(P_0, Q) \in \mathcal{N}_{k,f(k)}(0)$, $D(P_0||Q) \leq \log f(k)$. Furthermore, since $n' \sim \text{Poi}(n \sum Q_i)$, and $|\sum Q_i - 1| \leq \epsilon \leq 1/3$, we have $P(n' > n/2) \leq e^{-\kappa}$. Hence, we obtain

\[
\mathbb{E}[(\tilde{D}(P_0, N) - D(P_0, \frac{Q}{\sum_{i=1}^{\infty} Q_i}))^2] 
\leq \sum_{j=1}^{\infty} R^*(k, P_0, j, f(k))P(n' = i) + \delta 
= \sum_{j=1}^{\infty} R^*(k, P_0, j, f(k))P(n' = i) + \sum_{j=\infty+1}^{\infty} R^*(k, P_0, j, f(k))P(n' = i) + \delta 
\leq (\log^2 f(k))P(n' > n/2) + R^*(k, P_0, n/2, f(k)) + \delta 
\leq \log^2 f(k)e^{-\kappa} + R^*(k, P_0, n/2, f(k)) + \delta. 
\]  

Combining (158) and (160) completes the proof because $\delta$ can be arbitrary.

### E.2.4 Proof of Lemma 3

We construct the following pairs of $(P, Q)$ and $(P', Q')$:

\[ P = P' = P_0 = \left[ \frac{f(k)}{n \log k}, \ldots, \frac{f(k)}{n \log k}, 1 - \frac{(k-1)f(k)}{n \log k} \right], \]  

(161)

\[ Q = (V_1, \ldots, V_{k-1}, 1 - (k-1)\alpha), \]  

(162)

\[ Q' = (V'_1, \ldots, V'_{k-1}, 1 - (k-1)\alpha). \]  

(163)

And we assume that $|\mathbb{E}[\log V] - \mathbb{E}[\log V']| \geq d$. We define the following events:

\[ E \triangleq \left\{ \sum_{i=1}^{k-1} V_i - (k-1)\alpha \leq \epsilon, |D(P||Q) - \mathbb{E}(D(P||Q))| \leq \frac{d(k-1)f(k)}{4n \log k} \right\}, \]  

(164)

\[ E' \triangleq \left\{ \sum_{i=1}^{k-1} V'_i - (k-1)\alpha \leq \epsilon, |D(P'||Q') - \mathbb{E}(D(P'||Q'))| \leq \frac{d(k-1)f(k)}{4n \log k} \right\}. \]  

(165)

By union bound and Chebyshev’s inequality, we have

\[
P(E^C) \leq \frac{(k-1)\text{Var}(V)}{\epsilon^2} + \frac{16(k-1)\text{Var}(\frac{f(k)}{n \log k} \log V_i)}{(\frac{(k-1)f(k)}{n \log k})^2} \leq \frac{c^2(k-1) \log^2 k}{\epsilon^2 n^2} + \frac{16 \log^2 (n \log k)}{(k-1)d^2}. \]  

(166)

Similarly, we have

\[
P(E'^C) \leq \frac{c^2(k-1) \log^2 k}{\epsilon^2 n^2} + \frac{16 \log^2 (n \log k)}{(k-1)d^2}. \]  

(167)

Now, we define two priors on the set $\mathcal{N}_{k,f(k)}(\epsilon)$ by the following conditional distributions:

\[ \pi = P_{V|E} \quad \text{and} \quad \pi' = P_{V'|E'}. \]  

(168)
Hence, under \( \pi \) and \( \pi' \), we have

\[
|D(P||Q) - D(P'||Q')| \geq \frac{d(k - 1)f(k)}{2n \log k},
\]

(169)

Now, we consider the total variation of observations under \( \pi \) and \( \pi' \). The observations are Poisson distributed: \( N_i \sim \text{Poi}(nQ_i) \) and \( N'_i \sim \text{Poi}(nQ'_i) \). By the triangle inequality, we have

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

\[
\leq P(E^C) + P(E^{C'}) + TV(P_N, P_{N'})
\]

\[
\leq \frac{2c_2^2(k - 1)\log^2 k}{\epsilon^2 n^2} + \frac{32\log^2(n \log k)}{(k - 1)d^2} + TV(P_N, P_{N'}). \tag{170}
\]

By the i.i.d. construction of \( V_i \) for \( i = 1, \ldots, k - 1 \), we have

\[
TV(P_N, P_{N'}) = \sum_{i=1}^{k-1} TV(\mathbb{E}(\text{Poi}(nV_i)), \mathbb{E}(\text{Poi}(nV'_i)));
\]

\[
\leq kTV(\mathbb{E}(\text{Poi}(nV)), \mathbb{E}(\text{Poi}(nV'))). \tag{171}
\]

Applying generalized Le Cam’s method \cite{10}, and combining \(170\) and \(171\) complete the proof.

**F Proof of Proposition 4**

We first denote \( D_1 = \sum_{i=1}^{k} P_i \log P_i \) and \( D_2 = \sum_{i=1}^{k} P_i \log Q_i \). Hence, \( D(P||Q) = D_1 - D_2 \). Recall that our estimator \( \hat{D}_{opt} \) for \( D(P||Q) \) is defined as:

\[
\hat{D}_{opt} = \tilde{D}_{opt} \lor 0 \land f(k),
\]

(172)

where

\[
\bar{D}_{opt} = \bar{D}_1 - \bar{D}_2,
\]

(173)

\[
\bar{D}_1 = \sum_{i=1}^{k} \left( g_i(L_i)\mathbb{I}(M_i \leq c_2 \log k) + \left( \frac{M_i}{m} \log \frac{M_i}{m} - \frac{1}{2m} \right) \mathbb{I}(M_i > c_2 \log k) \right),
\]

(174)

\[
\bar{D}_2 = \sum_{i=1}^{k} \left( \frac{M_i}{m} g_i(N_i) \mathbb{I}(N_i \leq c_2 \log k) + \frac{M_i}{m} \left( \log \frac{N_i + 1}{n} - \frac{1}{2(N_i + 1)} \right) \mathbb{I}(N_i > c_2 \log k) \right)
\]

(175)

We define the following sets:

\[
E_{1,i} = \{ N'_i \leq c_2 \log k, Q_i \leq \frac{c_1 \log k}{n} \},
\]

\[
E_{2,i} = \{ N'_i > c_2 \log k, Q_i > \frac{c_1 \log k}{n} \},
\]

\[
F_{1,i} = \{ N'_i \leq c_2 \log k, Q_i > \frac{c_1 \log k}{n} \},
\]

\[
F_{2,i} = \{ N'_i > c_2 \log k, Q_i \leq \frac{c_1 \log k}{n} \}, \tag{176}
\]

31
and

\[ E'_{1,i} = \{ M'_i \leq c'_2 \log k, P_i \leq \frac{c'_1 \log k}{m} \}, \]
\[ E'_{2,i} = \{ M'_i > c'_2 \log k, P_i > \frac{c'_1 \log k}{m} \}, \]
\[ F'_{1,i} = \{ M'_i \leq c'_2 \log k, P_i > \frac{c'_1 \log k}{m} \}, \]
\[ F'_{2,i} = \{ M'_i > c'_2 \log k, P_i \leq \frac{c'_1 \log k}{m} \}, \]

(177)

where \( c_1 > c_2 > c_3 \) and \( c'_1 > c'_2 > c'_3 \). We further define the following sets:

\[ E = \bigcap_{i=1}^{k} (E_{1,i} \cup E_{2,i}), \]
\[ E' = \bigcap_{i=1}^{k} (E'_{1,i} \cup E'_{2,i}), \]
\[ \tilde{E} = \bigcap_{i=1}^{k} \left( (E_{1,i} \cup E_{2,i}) \cap (E'_{1,i} \cup E'_{2,i}) \right) \]
\[ = E \cap E'. \]

(178)

(179)

(180)

It can be shown that

\[ \tilde{E}^c = \bigcup_{i=1}^{k} (E_{1,i} \cup E_{2,i})^c \cup (E'_{1,i} \cup E'_{2,i})^c = \bigcup_{i=1}^{k} (F_{1,i} \cup F_{2,i}) \cup (F'_{1,i} \cup F'_{2,i}). \]

(181)

By union bound and Chernoff bound for Poisson distributions [23, Theorem 5.4], we have

\[ P(\tilde{E}^c) = P\left( \bigcup_{i=1}^{k} (F_{1,i} \cup F_{2,i}) \cup (F'_{1,i} \cup F'_{2,i}) \right) \leq \frac{1}{k^{c'_1-c'_2 \log \frac{c'_1}{c'_2} - 1}} + \frac{1}{k^{c'_3-c'_2 \log \frac{c'_3}{c'_2} - 1}} + \frac{1}{k^{c'_1-c'_2 \log \frac{c'_1}{c'_2} - 1}} + \frac{1}{k^{c'_3-c'_2 \log \frac{c'_3}{c'_2} - 1}}. \]

(182)

We note that \( \hat{D}_{opt}, D(P||Q) \in [0, \log f(k)] \), and \( \hat{D}_{opt} = \tilde{D}_{opt} \vee 0 \land \log f(k) \). Therefore, we have

\[ \mathbb{E}[ (\hat{D}_{opt} - D(P||Q))^2 ] \]
\[ = \mathbb{E}[ (\hat{D}_{opt} - D(P||Q))^2 1_{\tilde{E}} + (\hat{D}_{opt} - D(P||Q))^2 1_{\tilde{E}^c} ] \]
\[ \leq \mathbb{E}[ (\hat{D}_{opt} - D(P||Q))^2 1_{\tilde{E}}] + \log^2 f(k) P(\tilde{E}^c) \]
\[ = \mathbb{E}[ (\hat{D} - \tilde{D} - D_1 - D_2)^2 1_{\tilde{E}}] + \log^2 f(k) P(\tilde{E}^c) \]
\[ = \mathbb{E} \left[ (\hat{D} - \tilde{D} - D_1 - D_2)^2 | \tilde{E} \right] + \log^2 f(k) P(\tilde{E}^c). \]

(183)

We choose constants \( c_1, c_2, c_3, c'_1, c'_2, c'_3 \) such that \( c_1 - c_2 \log \frac{c_1}{c_2} - 1 > C, c'_1 - c'_2 \log \frac{c'_1}{c'_2} - 1 > C, c_3 - c_2 \log \frac{c_3}{c_2} - 1 > C, \) and \( c'_3 - c'_2 \log \frac{c'_3}{c'_2} - 1 > C \). Then together with \( \log m \leq C \log k \), we have

\[ \log^2 f(k) P(\tilde{E}^c) \leq \frac{\log^2 f(k)}{m}. \]

(184)
We can further decompose $\mathbb{E}\left[(\hat{D}_1 - \hat{D}_2 - D_1 + D_2)^2|\bar{E}\right]$ as follows:

\[
\mathbb{E}\left[(\hat{D}_1 - \hat{D}_2 - D_1 + D_2)^2|\bar{E}\right] = \mathbb{E}^2\left[(\hat{D}_1 - \hat{D}_2 - D_1 + D_2)|\bar{E}\right] + \text{Var}[(\hat{D}_1 - \hat{D}_2 - D_1 + D_2)|\bar{E}]
\]

\[
\leq 2\mathbb{E}^2\left[(\hat{D}_1 - D_1)|\bar{E}\right] + 2\mathbb{E}^2\left[(\hat{D}_2 - D_2)|\bar{E}\right] + \text{Var}[(\hat{D}_1 - \hat{D}_2 - D_1 + D_2)|\bar{E}]
\]

(185)

Following similar steps in [13], it can be shown that

\[
\mathbb{E}^2[|\hat{D}_1 - D_1|] \lesssim \frac{k^2}{m^2 \log^2 k}.
\]

(186)

We further define the index sets $I_1$, $I_2$, $I_1'$ and $I_2'$ as follows:

\[I_1 \triangleq \{ i : N_i' \leq c_2 \log k, Q_i \leq \frac{c_1 \log k}{n}\},\]

(187)

\[I_2 \triangleq \{ i : N_i' > c_2 \log k, Q_i > \frac{c_3 \log k}{n}\},\]

(188)

\[I_1' \triangleq \{ i : M_i' \leq c_2' \log k, P_i \leq \frac{c_1' \log k}{m}\},\]

(189)

\[I_2' \triangleq \{ i : M_i' > c_2' \log k, P_i > \frac{c_3' \log k}{m}\}.\]

(190)

We derive the following bound

\[
\text{Var} \left[(\hat{D}_1 - \hat{D}_2 - D_1 + D_2)|\bar{E}\right] = \text{Var} \left[\sum_{i \in I_1 \cap I_1'} (\hat{D}_{1,i} - D_{1,i}) + \sum_{i \in I_2 \cap I_1'} (\hat{D}_{1,i} - \hat{D}_{2,i}) + \sum_{i \in I_1 \cap I_2} (\hat{D}_{1,i} - \hat{D}_{2,i}) + \sum_{i \in I_1 \cap I_2} (\hat{D}_{1,i} - \hat{D}_{2,i})|\bar{E}\right]
\]

\[
\leq 4\text{Var} \left[\sum_{i \in I_1 \cap I_1'} (\hat{D}_{1,i} - \hat{D}_{2,i}|I_1, I_1')\right] + 4\text{Var} \left[\sum_{i \in I_2 \cap I_1'} (\hat{D}_{1,i} - \hat{D}_{2,i}|I_2, I_1')\right] + 4\text{Var} \left[\sum_{i \in I_1 \cap I_2} (\hat{D}_{1,i} - \hat{D}_{2,i}|I_1, I_2')\right] + 4\text{Var} \left[\sum_{i \in I_2 \cap I_2} (\hat{D}_{1,i} - \hat{D}_{2,i}|I_2, I_2')\right].
\]

(191)

We then define $\mathcal{E}_1$ and $\mathcal{E}_2$ as follows:

\[\mathcal{E}_1 = \sum_{i \in I_1 \cap (I_1' \cup I_2')} (\hat{D}_{2,i} - P_i \log Q_i),\]

(192)

\[\mathcal{E}_2 = \sum_{i \in I_2 \cap (I_1' \cup I_2')} (\hat{D}_{2,i} - P_i \log Q_i),\]

(193)

Then, we have

\[
\mathbb{E}^2[|\hat{D}_2 - D_2|] = \mathbb{E}^2[|\mathcal{E}_1 + \mathcal{E}_2|] \leq 2\mathbb{E}^2[\mathcal{E}_1 |I_1 \cap (I_1' \cup I_2')] + 2\mathbb{E}^2[\mathcal{E}_2 |I_2 \cap (I_1' \cup I_2')].
\]

(194)

In order to bound (185), we bound the four terms in (191) and the two terms in (194) one by one.
F.1 Bounds on the Variance

F.1.1 Bounds on \( \text{Var}[\sum_{i \in I \cap I'}(\hat{D}_{1,i} - \hat{D}_{2,i})|I_1, I'_1] \)

We first show that

\[
\text{Var}\left[ \sum_{i \in I \cap I'} (\hat{D}_{1,i} - \hat{D}_{2,i}) | I_1, I'_1 \right] 
\leq 2 \text{Var}\left[ \sum_{i \in I \cap I'} \hat{D}_{1,i} | I_1, I'_1 \right] + 2 \text{Var}\left[ \sum_{i \in I \cap I'} \hat{D}_{2,i} | I_1, I'_1 \right].
\]

(195)

Following similar steps in \[13\], it can be shown that

\[
\text{Var}\left[ \sum_{i \in I \cap I'} \hat{D}_{1,i} | I_1, I'_1 \right] \lesssim \frac{k^2}{m^2 \log^2 k}.
\]

(196)

In order to bound \( \text{Var}[\sum_{i \in I \cap I'} \hat{D}_{2,i} | I_1, I'_1] \), we bound \( \text{Var}(\frac{M_i}{m}g_L(N_i)) \) for each \( i \in I_1 \cap I'_1 \). Due to the independence between \( M_i \) and \( N_i \), \( \frac{M_i}{m} \) is independent of \( g_L(N_i) \). Hence,

\[
\text{Var}(\frac{M_i}{m}g_L(N_i)|i \in I_1 \cap I'_1) = \text{Var}(\frac{M_i}{m}) \text{Var}(g_L(N_i)|i \in I_1 \cap I'_1) + \text{Var}(\frac{M_i}{m})(\mathbb{E}(g_L(N_i))|i \in I_1 \cap I'_1)^2.
\]

(197)

We note that \( \text{Var}(\frac{M_i}{m}) = \frac{P_i}{m} \), and \( \mathbb{E}(\frac{M_i}{m}) = P_i \). We need to upper bound \( \text{Var}(g_L(N_i)|i \in I_1 \cap I'_1) \) and \( (\mathbb{E}(g_L(N_i))|i \in I_1 \cap I'_1)^2 \). Recall that \( g_L(N_i) = \sum_{m=1}^{L} \frac{a_m}{(c_1 \log k)^m} (N_{i,m-1} + \log \frac{n}{c_1 \log k}) \). The following lemma from \[13\] is also useful which provides an upper bound on the variance of \((N_i)_m\).

**Lemma 6.** \[13\] Lemma 6] If \( X \sim \text{Poi}(\lambda) \) and \( (x)_m = \frac{x}{(x-m)}! \). Then the variance of \((X)_m\) is increasing in \( \lambda \) and

\[
\text{Var}(X)_m \leq (\lambda m)^{m} \left( \frac{(2e)^{2\sqrt{\lambda m}}}{\pi \sqrt{\lambda m}} \lor 1 \right).
\]

(198)

Furthermore, the polynomial coefficients can be upper bounded as \( |a_m| \leq 2e^{-1/2}3L \) \[24\]. Due to the fact that the variance of sum of random variables is upper bounded by the square of the sum of individual standard deviations, we obtain

\[
\text{Var}(g_L(N_i)|i \in I_1 \cap I'_1) = \text{Var}(\sum_{m=2}^{L} \frac{a_m}{(c_1 \log k)^m} (N_{i,m-1} | i \in I_1 \cap I'_1)
\leq \left( \sum_{m=2}^{L} \frac{a_m}{(c_1 \log k)^m} \right)^2 \text{Var}((N_{i,m-1} | Q_i \leq \frac{c_1 \log k}{n})
\leq \left( \sum_{m=2}^{L} \frac{2e^{-1/2}3L}{(c_1 \log k)^m} \right)^2 \text{Var}((N_{i,m-1} | Q_i \leq \frac{c_1 \log k}{n})
\leq (c_1 \log k)^{m-1} \left( \frac{(2e)^{2\sqrt{c_1 \log k}}}{{\pi \sqrt{c_1 \log k}}} \lor 1 \right).
\]

(199)

By Lemma \[6\] we show that

\[
\text{Var}((N_{i,m-1} | Q_i \leq \frac{c_1 \log k}{n})
\leq \frac{c_1 \log k}{n})
\leq (c_1 \log k)^{m-1} (\frac{(2e)^{2\sqrt{c_1 \log k}}}{{\pi \sqrt{c_1 \log k}}} \lor 1).
\]

(200)
Substituting (200) into (199), we obtain
\[
\Var(g_L(N_i)|i \in I_1 \cap I'_1) \lesssim k^{2(c_0 \log 8 + \sqrt{c_0} \log 2e) \log k}. \tag{201}
\]
Furthermore, we bound \(|\E g_L(N_i)|i \in I_1 \cap I'_1|\) as follows:
\[
|\E(g_L(N_i)|i \in I_1 \cap I'_1)| = \left| \sum_{m=1}^{L} \frac{a_m}{(c_1 \log k)^{m-1}}(nQ_i)^{m-1} + \log n \frac{n}{c_1 \log k} |Q_i| \leq c_1 \log k \right|
\leq \sum_{m=1}^{L} 2e^{-1} 2^m (c_1 \log k)^{m-1} + \log n \frac{n}{c_1 \log k}
\lesssim k^{c_0 \log 8 \log k + \log n}. \tag{202}
\]

So far, we have all the ingredients we need to bound \(\Var(\frac{M_i}{n} g_L(N_i)|i \in I_1 \cap I'_1)\). We note that \(P_i/Q_i \leq f(k)\), i.e., \(P_i \leq f(k)Q_i\), and \(Q_i \leq c_1 \log k/n\) for \(i \in I_1\). First, we derive the following bound:
\[
\Var(\frac{M_i}{n} g_L(N_i)|i \in I_1 \cap I'_1)
\lesssim \frac{f(k) \log^2 k k^{2(c_0 \log 8 + \sqrt{c_0} \log 2e)}}{mn}
\lesssim \frac{f(k)}{n}, \tag{203}
\]
if \(2(c_0 \log 8 + \sqrt{c_0} \log 2e) < 1\).

Secondly, we derive
\[
(\E(\frac{M_i}{n} g_L(N_i)|i \in I_1 \cap I'_1))^2
\lesssim \frac{f^2(k) \log^3 k k^{2(c_0 \log 8 + \sqrt{c_0} \log 2e)}}{n^2}
\lesssim \frac{f^2(k)k^2}{n^2 \log^2 k}. \tag{204}
\]
if \(2(c_0 \log 8 + \sqrt{c_0} \log 2e) < 1\).

Thirdly, we have
\[
\Var(\frac{M_i}{n})(\E(g_L(N_i)|i \in I_1 \cap I'_1))^2
\lesssim \frac{f(k) \log^3 k k^{2c_0 \log 8} + f(k) \log k \log^2 n}{mn}
\lesssim \frac{2f(k)k^2}{mn} + \frac{f(k) \log k}{mn}
\lesssim \frac{2f(k)k^2}{mn}, \tag{205}
\]
if \(2c_0 \log 8 < 1\) and \(\log^2 n \lesssim k\).

Combining these three terms together, we obtain
\[
\Var[\sum_{i \in I_1 \cap I'_1} \hat{D}_{2,i}|I_1 \cap I'_1] \lesssim \frac{f(k)}{n} + \frac{f^2(k)}{n^2 \log^2 k} + \frac{k f(k) \log k}{mn}. \tag{206}
\]
Due to the fact that \(\frac{k f(k) \log k}{mn} \lesssim \frac{k^2 f^2(k)}{mn \log^2 k} \lesssim \frac{k^2 f^2(k)}{n^2 \log^2 k} + \frac{k^2}{m^2 \log^2 k}\), we further show that
\[
\E[\Var[\sum_{i \in I_1 \cap I'_1} \hat{D}_{2,i}|I_1 \cap I'_1]] \lesssim \frac{f(k)}{n} + \frac{f^2(k)}{n^2 \log^2 k} + \frac{k^2}{m^2 \log^2 k}. \tag{207}
\]
F.1.2 Bounds on \( \text{Var} \left[ \sum_{i \in I_2 \cap I'_1} \hat{D}_{1,i} - \hat{D}_{2,i} | I_2, I'_1 \right] \)

We note that for \( i \in I_2 \cap I'_1, Q_i > \frac{c_3 \log k}{n} \) and \( P_i \leq \frac{c'_3 \log k}{m} \). Following similar steps in [13], it can be shown that

\[
\text{Var} \left[ \sum_{i \in I_2 \cap I'_1} \hat{D}_{1,i} | I_2, I'_1 \right] \lesssim \frac{k^2}{m^2 \log^2 k}. \tag{208}
\]

We further consider \( \text{Var} \left[ \sum_{i \in I_2 \cap I'_1} \hat{D}_{2,i} | I_2, I'_1 \right] \). By the definition of \( D_{2,i} \), for \( i \in I_2 \cap I'_1 \), we have \( D_{2,i} = \frac{M_i}{m} \left( \log \frac{N_i + 1}{n} - \frac{1}{2(N_i + 1)} \right) \). Therefore we derive

\[
\text{Var} \left[ \sum_{i \in I_2 \cap I'_1} \hat{D}_{2,i} | I_2, I'_1 \right] = \sum_{i \in I_2 \cap I'_1} \text{Var} \left[ \frac{M_i}{m} \left( \log \frac{N_i + 1}{n} - \frac{1}{2(N_i + 1)} \right) \right] \leq 2 \sum_{i \in I_2 \cap I'_1} \text{Var} \left[ \frac{M_i}{m} \left( \log \frac{N_i + 1}{n} \right) \right] + 2 \sum_{i \in I_2 \cap I'_1} \text{Var} \left[ \frac{M_i}{m} \left( \frac{1}{2(N_i + 1)} \right) \right]. \tag{209}
\]

The first term in (209) can be bounded as follows:

\[
\sum_{i \in I_2 \cap I'_1} \text{Var} \left[ \frac{M_i}{m} \left( \log \frac{N_i + 1}{n} \right) \right] \leq \sum_{i \in I_2 \cap I'_1} \mathbb{E} \left[ \left( \frac{M_i}{m} \left( \log \frac{N_i + 1}{n} \right) - P_i \log Q_i \right)^2 \right] = \sum_{i \in I_2 \cap I'_1} \mathbb{E} \left[ \left( \frac{M_i}{m} \left( \log \frac{N_i + 1}{n} \right) - \frac{M_i}{m} \log Q_i + \frac{M_i}{m} \log Q_i - P_i \log Q_i \right)^2 \right] \leq \sum_{i \in I_2 \cap I'_1} 2 \mathbb{E} \left[ \left( \frac{M_i}{m} \right)^2 \left( \log \frac{N_i + 1}{n} - \log Q_i \right)^2 \right] + \sum_{i \in I_2 \cap I'_1} 2 \mathbb{E} \left[ \left( \frac{M_i}{m} - P_i \right)^2 \log^2 Q_i \right]. \tag{210}
\]

We note that for \( i \in I_2 \cap I'_1, Q_i > \frac{c_3 \log k}{n} \) and \( P_i \leq \frac{c'_3 \log k}{m} \). We then have the following bound on the first term in (210):

\[
\sum_{i \in I_2 \cap I'_1} \mathbb{E} \left[ \left( \frac{M_i}{m} \right)^2 \left( \log \frac{N_i + 1}{n} - \log Q_i \right)^2 \right] = \sum_{i \in I_2 \cap I'_1} \frac{m^2 P_i^2 + mP_i}{m^2} \mathbb{E} \left[ \left( \log \frac{N_i + 1}{n} - \log Q_i \right)^2 \right] = \sum_{i \in I_2 \cap I'_1} \frac{mP_i(1 + mP_i)}{m^2} \mathbb{E} \left[ \left( \log \frac{N_i + 1}{n} - \log Q_i \right) \mathbb{I}_{\{N_i \leq \frac{nQ_i}{\log k}\}} + \left( \log \frac{N_i + 1}{n} - \log Q_i \right)^2 \mathbb{I}_{\{N_i > \frac{nQ_i}{\log k}\}} \right] \leq a \frac{k^2}{m^2 \log^2 k} + \frac{1}{m}. \tag{211}
\]
where (a) is shown as follows, if \( \frac{c_1(1-\log 2)}{2} + 1 - C > 0 \). First, we derive the following bound:

\[
\sum_{i \in I_2 \cap I_1'} \frac{mP_i(1 + mP_i)}{m^2} \mathbb{E} \left[ \left( \log \frac{N_i + 1}{n} - \log Q_i \right)^2 I_{\{N_i \leq \frac{nQ_i}{2}\}} \right]
\]

\( \lesssim \sum_{i \in I_2 \cap I_1'} P_i \log k \frac{\log n}{m} P(N_i \leq \frac{nQ_i}{2}) \)

\( \lesssim \frac{k \log k}{m} \log k \frac{\log n}{\log m} \)

\( \lesssim \frac{m^2 \log^2 k}{k^2} \)

where (a) is due to the fact that \( P_i \leq \frac{c_1 \log k}{m}, \frac{N_i + 1}{n} \in \left[ \frac{1}{n}, 1 \right], Q_i \in \left[ \frac{\log k}{n}, 1 \right] \); (b) is due to the assumption that \( \log^2 n \lesssim k \) and \( \sum_{i \in I_2 \cap I_1'} P_i \leq 1 \); (c) is due to the Chernoff Bound: \( P(N_1 \leq \frac{nQ_1}{2}) \leq k^{-\frac{c_1(1-\log 2)}{2}} \); and (d) is due to the assumption that \( \log m \leq C \log k \) and the assumption that \( \frac{c_1(1-\log 2)}{2} + 1 - C > 0 \).

Secondly, we derive the following bound:

\[
\sum_{i \in I_2 \cap I_1'} \frac{mP_i(1 + mP_i)}{m^2} \mathbb{E} \left[ \left( \log \frac{N_i + 1}{n} - \log Q_i \right)^2 I_{\{N_i > \frac{nQ_i}{2}\}} \right]
\]

\( \lesssim \sum_{i \in I_2 \cap I_1'} \frac{P_i \log k}{m} \mathbb{E} \left[ \left( \frac{N_i + 1}{n} - Q_i \right)^2 \frac{1}{\xi^2} I_{\{N_1 > \frac{nQ_1}{2}\}} \right] \)

\( \lesssim \frac{P_i \log k}{m} \mathbb{E} \left[ \left( \frac{N_i + 1}{n} - Q_i \right)^2 \frac{4}{Q_i^2} I_{\{N_1 > \frac{nQ_1}{2}\}} \right] \)

\( \lesssim \frac{P_i \log k}{m} \frac{nQ_i + 1}{n^2} \frac{4}{Q_i^2} \)

\( \lesssim \frac{P_i \log k}{m} \frac{1}{\log k} \)

\( = \frac{1}{m}, \)

where (a) is due to the mean value theorem, and \( \min(\frac{N_i + 1}{n}, Q_i) \leq \xi \leq \max(\frac{N_i + 1}{n}, Q_i) \); (b) is due to the fact that \( \xi \geq \frac{Q_i}{2} \).
We next bound the second term in (210).

\[ \sum_{i \in I_2 \cap I'_1} E \left[ \left( \frac{M_i}{m} - P_i \right)^2 \log^2 Q_i \right] \]

\[ = \sum_{i \in I_2 \cap I'_1} \frac{P_i \log^2 Q_i}{m} \]

\[ \leq \sum_{i \in I_2 \cap I'_1} \frac{P_i \log^2 \frac{P_i}{f(k)}}{m} \]

\[ \leq \sum_{i \in I_2 \cap I'_1} \frac{2P_i (\log^2 P_i + \log f(k))}{m} \]

\[ \leq \frac{2 \log^2 f(k)}{m} + \frac{2k' \log k \log (\frac{2k'}{m})}{m} \]

\[ \leq \frac{\log^2 f(k)}{m} + \frac{k^2}{m^2 \log^2 k} \]

(214)

where (a) is due to the facts that \( x \log^2 x \) is monotone increasing when \( x \) is small and \( P_i \leq \frac{c' \log k}{m} \); (b) is due to the assumption that \( \log m \lesssim \log k \).

Substituting (214) and (211) into (210), we obtain

\[ \sum_{i \in I_2 \cap I'_1} \text{Var} \left[ \frac{M_i}{m} \left( \log \frac{N_i + 1}{n} \right) \right] \lesssim \frac{\log^2 f(k)}{m} + \frac{k^2}{m^2 \log^2 k}. \]

(215)

We then consider the second term in (209).

\[ \sum_{i \in I_2 \cap I'_1} \text{Var} \left[ \frac{M_i}{m} \left( \frac{1}{2(N_i + 1)} \right) \right] \]

\[ = \sum_{i \in I_2 \cap I'_1} \left( \text{Var} \left[ \frac{M_i}{m} \right] \text{Var} \left[ \frac{1}{2(N_i + 1)} \right] + \text{Var} \left[ \frac{M_i}{m} \right] \text{E}^2 \left[ \frac{1}{2(N_i + 1)} \right] + \text{Var} \left[ \frac{M_i}{m} \right] \text{Var} \left[ \frac{1}{2(N_i + 1)} \right] \right). \]

(216)

In order to bound (210), we bound each term as follows. We note that \( M_i \sim \text{Poi}(mP_i) \), and \( N_i \sim \text{Poi}(nQ_i) \). Therefore, we have \( \text{E}^2 \left[ \frac{M_i}{m} \right] = P_i^2 \), \( \text{Var} \left[ \frac{M_i}{m} \right] = \frac{P_i}{m} \), and

\[ \text{Var} \left[ \frac{1}{2(N_i + 1)} \right] + \text{E}^2 \left[ \frac{1}{2(N_i + 1)} \right] \]

\[ = \text{E} \left[ \frac{1}{4(N_i + 1)^2} \right] \]

\[ \leq \frac{1}{(N_i + 1)(N_i + 2)} \]

\[ = \sum_{i=0}^{\infty} \frac{1}{(i+1)(i+2)} \frac{e^{-nQ_i} (nQ_i)^i}{i!} \]

\[ = \sum_{i=0}^{\infty} \frac{1}{(nQ_i)^2} \frac{e^{-nQ_i} (nQ_i)^{i+2}}{(i+2)!} \]

\[ \leq \frac{1}{(nQ_i)^2}. \]

(217)
Therefore, (216) can be further upper bounded as follows:

\[
\sum_{i \in I_2 \cap I_1'} \text{Var} \left[ \frac{M_i}{m} \left( \frac{1}{2(N_i + 1)} \right) \right] \\
\leq \sum_{i \in I_2 \cap I_1'} \left( P_i^2 \frac{1}{(nQ_i)^2} + \frac{P_i}{m(nQ_i)^2} \right) \\
\lesssim \frac{f(k)}{n} + \frac{1}{m \log^2 k} \\
\lesssim \frac{f(k)}{n} + \frac{1}{m}. 
\]

Substituting (218) and (215) into (209), we obtain

\[
\text{Var} \left[ \sum_{i \in I_2 \cap I_1'} \hat{D}_{2,i} | I_2, I_1' \right] \lesssim \frac{f(k)}{n} + \log f(k) m + k^2 m^2 \log^2 k. 
\]

Therefore, we derive

\[
\text{Var} \left[ \sum_{i \in I_2 \cap I_1'} \hat{D}_{1,i} - \hat{D}_{2,i} | I_2, I_1' \right] \lesssim \frac{f(k)}{n} + \log f(k) m + k^2 m^2 \log^2 k. 
\]

F.1.3 Bounds on \(
\text{Var} \left[ \sum_{i \in I_1 \cap I_2'} \hat{D}_{1,i} - \hat{D}_{2,i} | I_1, I_2' \right] 
\)

We first note that given \(i \in I_1 \cap I_2', P_i > \frac{c_1 \log k}{m}, Q_i \leq \frac{c_1 \log k}{m}, \) and \(\frac{P_i}{Q_i} \leq f(k).\) Hence, we have \(\frac{c_1 \log k}{m} < P_i \leq \frac{c_1 f(k) \log k}{n}.\) Following similar steps in [13], it can be shown that

\[
\text{Var} \left[ \sum_{i \in I_1 \cap I_2'} \hat{D}_{1,i} \right] \leq \frac{4}{m} + \frac{12k}{m^2} + \frac{k}{c_3 m^2 \log k} + \sum_{i \in I_1 \cap I_2'} \frac{2P_i}{m} \log^2 P_i. 
\]

We further consider the last term \(\sum_{i \in I_1 \cap I_2'} \frac{2P_i}{m} \log^2 P_i\) in (221), under the condition that \(\frac{c_1 \log k}{m} < P_i \leq \frac{c_1 f(k) \log k}{n}.\) We show that

\[
\sum_{i \in I_1 \cap I_2'} \frac{P_i \log^2 P_i}{m} \\
\leq \sum_{i \in I_1 \cap I_2'} \frac{c_1 f(k) \log k}{mn} \log^2 \frac{c_1 \log k}{m} \\
\leq \frac{c_1 k f(k) \log k}{mn} \log^2 \frac{c_1 \log k}{m} \\
\lesssim \frac{k f(k) \log k}{mn} m \\
\lesssim \frac{k f(k) \log^3 k}{mn} \\
\lesssim \frac{k^2 f(k)}{mn \log^2 k} \\
\lesssim \frac{k f(k) \log k}{m \log k} + \frac{(k f(k) \log k)}{n \log k}^2. 
\]

39
where (a) is due to the assumption that \( m \gtrsim \frac{k}{\log k} \), (b) is due to the assumption that \( \log m \leq C \log k \), and (c) is due to the fact that \( 2ab \leq a^2 + b^2 \). Therefore, we obtain

\[
\text{Var}\left[ \sum_{i \in I_1 \cap I_2^*} \hat{D}_{1,i} \right] \lesssim (\frac{k}{m \log k})^2 + (\frac{kf(k)}{n \log k})^2. \tag{223}
\]

Following similar steps in Appendix F.1.1, we can show that

\[
\text{Var}\left[ \sum_{i \in I_1 \cap I_2^*} \hat{D}_{2,i} \right] \lesssim \frac{f(k)}{n} + \frac{f^2(k)k^2}{n^2 \log^2 k} + \frac{k^2}{m^2 \log^2 k}. \tag{224}
\]

Hence, we have

\[
\text{Var}\left[ \sum_{i \in I_1 \cap I_2^*} \hat{D}_{1,i} - \hat{D}_{2,i} \mid I_1, I_2^* \right] \lesssim \frac{f(k)}{n} + \frac{f^2(k)k^2}{n^2 \log^2 k} + \frac{k^2}{m^2 \log^2 k}. \tag{225}
\]

**F.1.4 Bounds on \( \text{Var}\left[ \sum_{i \in I_2^* \cap I_2^*} \hat{D}_{1,i} - \hat{D}_{2,i} \mid I_2, I_2^* \right] \)**

We note that for \( i \in I_2 \cap I_2^* \), \( P_i > \frac{c_i \log k}{m}, Q_i > \frac{c_i \log k}{n} \), and

\[
\hat{D}_{1,i} - \hat{D}_{2,i} = \frac{M_i}{m} \log \frac{M_i}{m} - \frac{M_i}{m} \log \frac{N_i + 1}{n} - \frac{1}{2m} \left( \log \frac{N_i + 1}{n} - 2(N_i + 1) \right). \tag{226}
\]

It can be shown that

\[
\text{Var}\left[ \sum_{i \in I_2 \cap I_2^*} \hat{D}_{1,i} - \hat{D}_{2,i} \mid I_2, I_2^* \right] \leq 2 \text{Var}\left[ \sum_{i \in I_2 \cap I_2^*} \frac{M_i}{m} \log \frac{M_i}{m} - \frac{M_i}{m} \log \frac{N_i + 1}{n} \mid I_2, I_2^* \right] + 2 \text{Var}\left[ \sum_{i \in I_2 \cap I_2^*} \frac{1}{2(N_i + 1)} \mid I_2, I_2^* \right]. \tag{227}
\]

Following similar steps as showing (215), we bound the second term in (227) as follows:

\[
\text{Var}\left[ \sum_{i \in I_2 \cap I_2^*} \frac{M_i}{m} \frac{1}{2(N_i + 1)} \mid I_2, I_2^* \right] \lesssim \frac{f(k)}{n} + \frac{1}{m}. \tag{228}
\]

We next bound the first term in (227) as follows:

\[
\text{Var}\left[ \sum_{i \in I_2 \cap I_2^*} \frac{M_i}{m} \log \frac{M_i}{m} - \frac{M_i}{m} \log \frac{N_i + 1}{n} \mid I_2, I_2^* \right] \leq \sum_{i \in I_2 \cap I_2^*} \mathbb{E} \left[ \left( \frac{M_i}{m} \log \frac{M_i}{m} - \frac{M_i}{m} \log \frac{N_i + 1}{n} - P_i \log \frac{P_i}{Q_i} \right)^2 \right]
\]

\[
= \sum_{i \in I_2 \cap I_2^*} \mathbb{E} \left[ \left( \frac{M_i}{m} \log \frac{M_i}{m} - \frac{M_i}{m} \log \frac{N_i + 1}{n} - \frac{M_i}{m} \log P_i + \frac{M_i}{m} \log \frac{P_i}{Q_i} + \frac{M_i}{m} \log Q_i - P_i \log \frac{P_i}{Q_i} \right)^2 \right]
\]

\[
\leq \sum_{i \in I_2 \cap I_2^*} 3 \mathbb{E} \left[ \left( \frac{M_i}{m} \log \frac{M_i}{m} - \log P_i \right)^2 \right] + 3 \mathbb{E} \left[ \left( \frac{M_i}{m} - P_i \log \frac{P_i}{Q_i} \right)^2 \right] + 3 \mathbb{E} \left[ \left( \frac{M_i}{m} \log \frac{N_i + 1}{n} - \log Q_i \right)^2 \right]. \tag{229}
\]
We further bound the three terms in (229) one by one. We note that \( \log x \leq x - 1 \) for any \( x > 0 \). Therefore, we have

\[
\frac{M_i}{m} - P_i \leq \frac{M_i}{m} \log \frac{M_i}{m} P_i \leq \frac{M_i}{m} - P_i + \left( \frac{M_i}{m} - P_i \right)^2,
\]

which implies

\[
\left( \frac{M_i}{m} \log \frac{M_i}{m} P_i \right)^2 \leq 2 \left( \frac{M_i}{m} - P_i \right)^2 + 2 \left( \frac{M_i}{m} - P_i \right)^2.
\]

Taking expectation on both sides, we have

\[
\mathbb{E} \left[ \left( \frac{M_i}{m} \log \frac{M_i}{m} P_i \right)^2 \right] \leq 4 \frac{P_i}{m} + \frac{12}{m^2} + \frac{4}{n^3 P_i} \leq 4 \frac{P_i}{m} + \frac{12}{m^2} + \frac{4}{m^2 c_3 \log k},
\]

where the last inequality is due to the condition that \( P_i \geq \frac{c_1 \log k}{m} \). Therefore, we obtain the following bound:

\[
\sum_{i \in I_2 \cap I_2'} \mathbb{E} \left[ \left( \frac{M_i}{m} \log \frac{M_i}{m} - \log P_i \right)^2 \right] \leq \frac{1}{m} + \frac{k}{m^2}.
\]

Consider the second term in (229), we derive the following bound:

\[
\sum_{i \in I_2 \cap I_2'} \mathbb{E} \left[ \left( \frac{M_i}{m} - P_i \right) \log \frac{P_i}{Q_i} \right]^2 \leq \sum_{i \in I_2 \cap I_2'} \frac{P_i}{m} \log^2 \frac{P_i}{Q_i} \leq \frac{\log^2 f(k)}{m},
\]

where the last inequality is due to the following fact that

\[
\sum_{i \in I_2 \cap I_2'} P_i \log^2 \frac{P_i}{Q_i} = \sum_{i \in I_2 \cap I_2'} \left( P_i \log^2 \frac{P_i}{Q_i} \mathbb{1}_{\{nQi \leq f(k)\}} + Q_i \log^2 \frac{P_i}{Q_i} \mathbb{1}_{\{f(k) < nQi \}} \right) \leq \log^2 f(k).
\]

The last inequality in (230) is due to the fact that the function \( x \log x \) is bounded by a constant on the interval \([0, 1]\).

We next bound the third term in (229). We note that \( \frac{x - x}{x} \leq \log \frac{x}{x} \leq \frac{x - x}{x} \), and derive the following bound

\[
\sum_{i \in I_2 \cap I_2'} \mathbb{E} \left[ \left( \frac{M_i}{m} \log \frac{N_i + 1}{n} - \log Q_i \right)^2 \right]
\]

\[
= \sum_{i \in I_2 \cap I_2'} \mathbb{E} \left[ \left( \frac{M_i}{m} \log \frac{N_i + 1}{nQ_i} \right)^2 \mathbb{1}_{\{N_i + 1 \leq \frac{nQi}{2} + 1\}} + \left( \frac{M_i}{m} \log \frac{N_i + 1}{nQ_i} \right)^2 \mathbb{1}_{\{N_i + 1 > \frac{nQi}{2} + 1\}} \right]
\]

\[
\leq \sum_{i \in I_2 \cap I_2'} \left( \frac{M_i}{m} \right)^2 \left( \frac{N_i + 1 - nQ_i}{1} \right)^2 \mathbb{1}_{\{N_i + 1 \leq \frac{nQi}{2} + 1\}} + \left( \frac{M_i}{m} \right)^2 \left( \frac{N_i + 1 - nQ_i}{nQ_i} \right)^2 \mathbb{1}_{\{N_i + 1 > \frac{nQi}{2} + 1\}}
\]

\[
\leq \sum_{i \in I_2 \cap I_2'} \left( P_i^2 + \frac{P_i}{m} \right) \left( 2nQ_i P(N_i \leq \frac{nQi}{2}) + 8 \right)
\]

\[
\leq \sum_{i \in I_2 \cap I_2'} \left( P_i^2 + \frac{P_i}{m} \right) \left( 2nQ_i e^{-nQ_i} + 8 \frac{nQ_i}{nQ_i} \right)
\]

\[
\leq \frac{f(k)}{n} + \frac{kf(k)}{mn},
\]

(236)
where (a) is due to the fact that \( N_i + 1 \geq 1 \); (b) use the fact that \( Q_i \geq \frac{c_k \log k}{n} \); (c) is due to the fact that \( x^2 e^{-x} \) is bounded by a constant for \( x > 0 \).

Hence, combining (225), (226), (227) and (228), we obtain

\[
\text{Var} \left[ \sum_{i \in I_1 \cap (I'_1 \cup I'_2)} \hat{D}_{1,i} - \hat{D}_{2,i} | I_2, I'_2 \right] \leq \frac{k}{m^2} + \frac{\log^2 f(k)}{m} + \frac{f(k)}{n} + \frac{kf(k)}{mn}. \tag{237}
\]

### F.2 Bounds on the Bias:

We consider the first term in [194]. Based on the definition of the set \( I_1 \), \( \mathcal{E}_1 \) can be written as follows:

\[
\mathcal{E}_1 = \sum_{i \in I_1 \cap (I'_1 \cup I'_2)} \left( \frac{M_i}{m} g_L(N_i) - P_i \log Q_i \right). \tag{238}
\]

Hence, \( |\mathbb{E}[\mathcal{E}_1 | I_1 \cap (I'_1 \cup I'_2)]| = |\sum_{i \in I_1 \cap (I'_1 \cup I'_2)} \mathbb{E}[\frac{M_i}{m} g_L(N_i) - P_i \log Q_i]|. \) For \( i \in I_1 \cap (I'_1 \cup I'_2) \), we have \( 0 \leq Q_i \leq \frac{c_k \log k}{n} \) and \( P_i \frac{g_L(Q_i)}{Q_i} - P_i \log Q_i \leq \frac{f(k)}{n \log k} \). Therefore, we derive the following bound:

\[
\mathbb{E}[\frac{M_i}{m} g_L(N_i) - P_i \log Q_i | i \in I_1 \cap (I'_1 \cup I'_2)] = |P_i \frac{g_L(Q_i)}{Q_i} - P_i \log Q_i| \mathbb{1}_{0 \leq Q_i \leq \frac{c_k \log k}{n}} \leq \frac{f(k)}{n \log k}. \tag{239}
\]

Hence, \( |\mathbb{E}[\mathcal{E}_1 | I_1 \cap (I'_1 \cup I'_2)]| \) can be bounded as follows:

\[
|\mathbb{E}(\mathcal{E}_1 | I_1 \cap (I'_1 \cup I'_2))| = \left| \sum_{i \in I_1} \mathbb{E}[\frac{M_i}{m} g_L(N_i) - P_i \log Q_i | i \in I_1 \cap (I'_1 \cup I'_2)] \right| \leq \frac{k \log f(k)}{n \log k}. \tag{240}
\]

Therefore, we have

\[
\mathbb{E}[\mathbb{E}[\mathcal{E}_1 | I_1 \cap (I'_1 \cup I'_2)] \leq \frac{k^2 \log^2 f(k)}{n^2 \log^2 k}. \tag{241}
\]

We then consider the second term in [194]. Based on how we define \( I_2 \), \( \mathcal{E}_2 \) can be written as follows:

\[
\mathcal{E}_2 = \sum_{i \in I_2 \cap (I'_1 \cup I'_2)} \left( \frac{M_i}{m} \left( \log \frac{N_i + 1}{n} - \frac{1}{2(N_i + 1)} \right) - P_i \log Q_i \right) \leq \sum_{i \in I_2 \cap (I'_1 \cup I'_2)} \left( \frac{M_i}{m} - P_i \right) \log Q_i + \frac{M_i}{m} \log \frac{N_i + 1}{n Q_i} - \frac{P_i}{2(N_i + 1)} \tag{242}
\]

Taking expectation on both sides, we obtain

\[
\mathbb{E} \left[ \mathcal{E}_2 | I_2, I'_1, I'_2 \right] = \sum_{i \in I_2 \cap (I'_1 \cup I'_2)} \mathbb{E} \left[ P_i \log \frac{N_i + 1}{n Q_i} - \frac{P_i}{2(N_i + 1)} | I_2, I'_1, I'_2 \right]. \tag{243}
\]

We first consider \( \sum_{i \in I_2 \cap (I'_1 \cup I'_2)} \mathbb{E} \left[ P_i \log \frac{N_i + 1}{n Q_i} | I_2, I'_1, I'_2 \right] \). We note that for any \( x > 0 \),

\[
\log x \leq (x - 1) - \frac{1}{2}(x - 1)^2 + \frac{1}{3}(x - 1)^3. \tag{244}
\]
It can be shown that
\[
\frac{1}{2n_{Q_i}} + \frac{5}{6(n_{Q_i})^2} + \frac{1}{3(n_{Q_i})^3}.
\]
(245)

It can be shown that
\[
\mathbb{E} \left[ \frac{P_i}{2(N_i + 1)} \right] = \frac{P_i}{2n_{Q_i}} (1 - e^{-n_{Q_i}}).
\]
(246)

Hence, we obtain
\[
\mathbb{E} \left[ E_2|I_2, I_1', I_2' \right] \leq \sum_{i \in I_2 \cap (I_1' \cup I_2')} P_i \left( \frac{1}{2n_{Q_i}} + \frac{5}{6(n_{Q_i})^2} + \frac{1}{3(n_{Q_i})^3} \right) - \frac{P_i}{2n_{Q_i}} (1 - e^{-n_{Q_i}}) \\
\leq k f(k) \leq \frac{n^2 Q_i}{n \log k}.
\]
(247)

where \((a)\) is due to the fact that \(x^2 e^{-x}\) is bounded by a constant for \(x \geq 0\).

We further consider the lower bound on \(\mathbb{E} \left[ E_2|I_2, I_1', I_2' \right]\). Similarly, for any \(x \geq \frac{1}{2}\), we can show that
\[
\log x \geq (x - 1) - \frac{1}{2}(x - 1)^2 + \frac{1}{3}(x - 1)^3 - (x - 1)^4.
\]
(248)

We define the following event: \(A_i = \{ \frac{N_i}{n_{Q_i}} > \frac{1}{2} \}\). We then rewrite \(\mathbb{E} \left[ E_2|I_2, I_1', I_2' \right]\) as follows:
\[
\mathbb{E} \left[ E_2|I_2, I_1', I_2' \right] = \sum_{i \in I_2 \cap (I_1' \cup I_2')} \mathbb{E} \left[ P_i \log \frac{N_i + 1}{n_{Q_i}} \mathbb{I}_{\{A_i\}} + P_i \log \frac{N_i + 1}{n_{Q_i}} \mathbb{I}_{\{A_i'\}} - \frac{P_i}{2(N_i + 1)} |I_2, I_1', I_2' \right].
\]
(249)

It can be shown that
\[
\mathbb{E}[P_i \log \frac{N_i + 1}{n_{Q_i}} \mathbb{I}_{\{A_i\}} |I_2, I_1', I_2'] \\
\geq \mathbb{E}[P_i \left( \left( \frac{N_i + 1}{n_{Q_i}} - 1 \right) - \frac{1}{2}\left( \frac{N_i + 1}{n_{Q_i}} - 1 \right)^2 + \frac{1}{3}\left( \frac{N_i + 1}{n_{Q_i}} - 1 \right)^3 - \left( \frac{N_i + 1}{n_{Q_i}} - 1 \right)^4 \right) \mathbb{I}_{\{A_1\}} |I_2, I_1', I_2']
\]
(250)

We note that
\[
\mathbb{E}[P_i \left( \frac{N_i + 1}{n_{Q_i}} - 1 \right) \mathbb{I}_{\{A_1\}} |I_2, I_1', I_2'] \\
= \mathbb{E}[P_i \left( \frac{N_i + 1}{n_{Q_i}} - 1 \right) |I_2, I_1', I_2'] - \mathbb{E}[P_i \left( \frac{N_i + 1}{n_{Q_i}} - 1 \right) \mathbb{I}_{\{A_1'\}} |I_2, I_1', I_2'] \\
\geq \mathbb{E}[P_i \left( \frac{N_i + 1}{n_{Q_i}} - 1 \right) |I_2, I_1', I_2'] \\
= \frac{1}{n_{Q_i}}.
\]
(251)
Similarly, we obtain
\[
E \left[ \left( \frac{N_i + 1}{nQ_i} - 1 \right)^3 \mathbb{I}_{(A_i')} | I_2, I_1', I_2' \right] \geq E \left[ \left( \frac{N_i + 1}{nQ_i} - 1 \right)^3 \right] = \frac{4}{(nQ_i)^2} + \frac{1}{(nQ_i)^4} \tag{252}
\]

For the term \(E[(\frac{N_i + 1}{nQ_i} - 1)^2 | I_2, I_1', I_2']\), it can be shown that
\[
E[(\frac{N_i + 1}{nQ_i} - 1)^2 \mathbb{I}_{(A_i')} | I_2, I_1', I_2'] \leq E[(\frac{N_i + 1}{nQ_i} - 1)^2 | I_2, I_1', I_2'] = \frac{1}{nQ_i} + \frac{1}{(nQ_i)^2}. \tag{253}
\]

Similarly, it can be shown that
\[
E[(\frac{N_i + 1}{nQ_i} - 1)^4 \mathbb{I}_{(A_i')} | I_2, I_1', I_2'] \leq E[(\frac{N_i + 1}{nQ_i} - 1)^4 | I_2, I_1', I_2'] = \frac{1 + 3nQ_i}{(nQ_i)^3} + \frac{10}{(nQ_i)^3} + \frac{1}{(nQ_i)^4}. \tag{254}
\]

Combining these results together, we obtain
\[
E[P_i \log(\frac{N_i + 1}{nQ_i}) \mathbb{I}_{(A_i')} | I_2, I_1', I_2'] \geq \frac{P_i}{2nQ_i} - \frac{13P_i}{6(nQ_i)^2} - \frac{32P_i}{3(nQ_i)^3} - \frac{P_i}{(nQ_i)^4} \tag{255}
\]

For the \(E[P_i \log(\frac{N_i + 1}{nQ_i}) \mathbb{I}_{(A_i')} | I_2, I_1', I_2']\), it can be shown that
\[
\sum_{i \in I_2 \cap (I_1' \cup I_2')} |E[P_i \log(\frac{N_i + 1}{nQ_i}) \mathbb{I}_{(A_i')} | I_2, I_1', I_2']| \leq
\]
\[
\leq \sum_{i \in I_2 \cap (I_1' \cup I_2')} P_i \log(nQ_i) P(A_i^c)
\]
\[
\leq \sum_{i \in I_2 \cap (I_1' \cup I_2')} \frac{P_i}{(nQ_i)^2} \log(nQ_i) e^{-(1 - \frac{\log(nQ_i)}{2}) nQ_i}
\]
\[
\leq \frac{k f(k)}{n \log k} \tag{256}
\]

where (a) is due to the fact that \(N_i + 1 \geq 1\), and the fact that \(Q_i > \frac{c_1 \log k}{n}\), hence \(|\log(\frac{N_i + 1}{nQ_i})| \leq \log(nQ_i)\) for large \(k\); (b) is due to Chernoff bound, where \(1 - \frac{\log(nQ_i)}{2} > 0\); (c) is due to the fact that \((nQ_i)^2 \log(nQ_i) e^{-(1 - \frac{\log(nQ_i)}{2}) nQ_i}\) is bounded by a constant for \(nQ_i > 1\), and the fact \(nQ_i > c_3 \log k\).

From previous results, we know that
\[
E \left[ - \frac{P_i}{2(N_i + 1)} | I_2, I_1', I_2' \right] = - \frac{P_i}{2nQ_i} (1 - e^{-nQ_i}). \tag{257}
\]

Combining (255) and (257), it can be shown that
\[
\sum_{i \in I_2 \cap (I_1' \cup I_2')} E \left[ P_i \log(\frac{N_i + 1}{nQ_i}) \mathbb{I}_{(A_i')} - \frac{P_i}{2(N_i + 1)} | I_2, I_1', I_2' \right]
\]
\[
\geq \sum_{i \in I_2 \cap (I_1' \cup I_2')} \frac{P_i}{2nQ_i} - \frac{13P_i}{6(nQ_i)^2} - \frac{32P_i}{3(nQ_i)^3} - \frac{P_i}{(nQ_i)^4} - \frac{P_i}{2nQ_i} (1 - e^{-nQ_i})
\]
\[
= \sum_{i \in I_2 \cap (I_1' \cup I_2')} - \frac{13P_i}{6(nQ_i)^2} - \frac{32P_i}{3(nQ_i)^3} - \frac{P_i}{(nQ_i)^4} + \frac{P_i}{2nQ_i} e^{-nQ_i}. \tag{258}
\]
We further bound the absolute value of the right hand side of (258) as follows:

\[
\left| \sum_{i \in I_2 \cap (I_1' \cup I_2')} - \frac{13 P_i}{6(nQ_i)^2} - \frac{32 P_i}{3(nQ_i)^3} - \frac{P_i}{(nQ_i)^4} + \frac{P_i}{2nQ_i} e^{-nQ_i} \right| \lesssim \frac{k f(k)}{n \log k},
\]

(259)

where we use the facts that \( \frac{P_i}{nQ_i} \leq f(k), nQ_i > c_3 \log k \), and \( nQ_i e^{-nQ_i} \) is upper bounded by a constant for any value of \( nQ_i \).

Combining (217), (256) and (259), it can be shown that

\[
\left| \mathbb{E} [E_{2|I_2, I_1', I_2'}] \right| \lesssim \frac{k f(k)}{n \log k}.
\]

(260)

References

[1] Y. Bu, S. Zou, Y. Liang, and V. V. Veeravalli, “Universal outlying sequence detection for continuous observations,” in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), March 2016, pp. 4254–4258.

[2] P. J. Moreno, P. P. Ho, and N. Vasconcelos, “A kullback-leibler divergence based kernel for svm classification in multimedia applications,” in Proc. Advances in Neural Information Processing Systems (NIPS), 2003.

[3] I. S. Dhillon, S. Mallela, and R. Kumar, “A divisive information-theoretic feature clustering algorithm for text classification,” Journal of machine learning research, vol. 3, no. Mar, pp. 1265–1287, 2003.

[4] N. H. Anderson, P. Hall, and D. M. Titterington, “Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates,” Journal of Multivariate Analysis, vol. 50, no. 1, pp. 41–54, 1994.

[5] Q. Wang, S. R Kulkarni, and S. Verdú, “Divergence estimation of continuous distributions based on data-dependent partitions,” IEEE Trans. Inform. Theory, vol. 51, no. 9, pp. 3064–3074, 2005.

[6] Q. Wang, S. R Kulkarni, and S. Verdú, “Divergence estimation for multidimensional densities via k-nearest-neighbor distances,” IEEE Trans. Inform. Theory, vol. 55, no. 5, pp. 2392–2405, May 2009.

[7] X. Nguyen, M. J Wainwright, and M. Jordan, “Estimating divergence functionals and the likelihood ratio by convex risk minimization,” IEEE Trans. Inform. Theory, vol. 56, no. 11, pp. 5847–5861, 2010.

[8] K. R. Moon and A. O. Hero, “Ensemble estimation of multivariate f-divergence,” in Proc. IEEE International Symposium on Information Theory (ISIT), 2014, pp. 356–360.

[9] Z. Zhang and M. Grabchak, “Nonparametric estimation of Kullback-Leibler divergence,” Neural computation, vol. 26, no. 11, pp. 2570–2593, 2014.

[10] H. Cai, S. R. Kulkarni, and S. Verdú, “Universal estimation of entropy and divergence via block sorting,” in Proc. IEEE Int. Symp. Information Theory (ISIT), 2002, p. 433.

[11] H. Cai, S. R. Kulkarni, and S. Verdú, “Universal divergence estimation for finite-alphabet sources,” IEEE Trans. Inform. Theory, vol. 52, no. 8, pp. 3456–3475, 2006.

[12] 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 Proc. of the 43rd annual ACM symposium on Theory of computing. ACM, 2011, pp. 685–694.

[13] Y. Wu and P. Yang, “Minimax rates of entropy estimation on large alphabets via best polynomial approximation,” IEEE Trans. Inform. Theory, vol. 62, no. 6, pp. 3702–3720, June 2016.
[14] J. Jiao, K. Venkat, and T. Weissman, “Maximum likelihood estimation of functionals of discrete distributions,” arxiv:1406.6959.

[15] J. Jiao, K. Venkat, Y. Han, and T. Weissman, “Minimax estimation of functionals of discrete distributions,” IEEE Trans. Inform. Theory, vol. 61, no. 5, pp. 2835–2885, 2015.

[16] A. B. Tsybakov, Introduction to nonparametric estimation, Springer Science & Business Media, 2008.

[17] Y. Han, J. Jiao, and T. Weissman, “Minimax estimation of kl divergence between discrete distributions,” arxiv: 1605.09124, 2016.

[18] A. Orlitsky and A. T. Suresh, “Competitive distribution estimation: Why is good-turing good,” in Proc. Advances in Neural Information Processing Systems, pp. 2134–2142. 2015.

[19] V. Totik, Polynomial approximation on polytopes, vol. 232, American Mathematical Society, 2014.

[20] A. F. Timan, Theory of Approximation of Functions of a Real Variable, Dover books on advanced mathematics. Pergamon Press, 1963.

[21] Liam Paninski, “Estimation of entropy and mutual information,” Neural computation, vol. 15, no. 6, pp. 1191–1253, 2003.

[22] S. Boucheron, G. Lugosi, and P. Massart, Concentration inequalities: A nonasymptotic theory of independence, Oxford university press, 2013.

[23] M. Mitzenmacher and E. Upfal, Probability and computing: Randomized algorithms and probabilistic analysis, Cambridge University Press, 2005.

[24] T. 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.