ESTIMATION OF KL DIVERGENCE: OPTIMAL MINIMAX RATE

BY

YUHENG BU

THESIS

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

Urbana, Illinois

Adviser:

Professor Venugopal V. Veeravalli

ABSTRACT

The problem of estimating the Kullback-Leibler divergence $D(P\|Q)$ between two unknown distributions P and Q is studied, under the assumption that the alphabet size k of the distributions can scale to infinity. The estimation is based on m independent samples drawn from P and n independent samples drawn from Q. It is first shown that there exists no consistent estimator that guarantees asymptotically small worst-case quadratic risk over the set of all pairs of distributions. A restricted set that contains pairs of distributions, with density ratio bounded by a function $f(k)$, is further considered. An augmented plug-in estimator is proposed, and is shown to be consistent if and only if m has an order greater than $k \vee \log^2(f(k))$, and n has an order greater than $kf(k)$. Moreover, the minimax quadratic risk is characterized to be within a constant factor of $(\frac{k}{m\log k}+\frac{kf(k)}{n\log k})^2+\frac{\log^2 f(k)}{m}+\frac{f(k)}{n}$, if m and n exceed constant factors of $k/\log(k)$ and $kf(k)/\log k$, respectively. The lower bound on the minimax quadratic risk is characterized by employing a generalized Le Cam's method. A minimax optimal estimator is then constructed by employing both the polynomial approximation and plug-in approaches.

To my family

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to Professor Venugopal V. Veeravalli for his support throughout this research process. Without his counsel and guidance this work would not have been possible.

Lastly, I would like to express the thanks I owe to my family and my girlfriend for their patience and support throughout my life.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background

As an important quantity in information theory, the Kullback-Leibler (KL) divergence between two distributions has a wide range of applications in various domains. 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 available data samples. This thesis studies such a problem.

1.2 Problem Statement

Consider the estimation of KL divergence between two probability distributions P and Q defined as

$$D(P\|Q) = \sum_{i=1}^{k} P_i \log \frac{P_i}{Q_i}, \tag{1.1}$$

where P and Q are supported on 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 $\mathcal{M}_k$ to denote the collection of all such pairs of distributions.

Suppose P and Q are unknown, and that 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. The sufficient statistics for estimating $D(P\|Q)$ are the histograms of the samples

$M \triangleq (M_1, \dots, M_k)$ and $N \triangleq (N_1, \dots, N_k)$, where

$$M_j = \sum_{i=1}^{m} \mathbb{1}_{\{X_i=j\}} \quad \text{and} \quad N_j = \sum_{i=1}^{n} \mathbb{1}_{\{Y_i=j\}} \tag{1.2}$$

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:

$$R(\hat{D}, k, m, n) \triangleq \sup_{(P,Q)\in\mathcal{M}_k} \mathbb{E}[(\hat{D}(M, N) - D(P\|Q))^2]. \tag{1.3}$$

We further define the minimax quadratic risk as:

$$R^*(k, m, n) \triangleq \inf_{\hat{D}} R(\hat{D}, k, m, n). \tag{1.4}$$

In this thesis, we are interested in the large-alphabet regime with $k \to \infty$. Furthermore, the numbers m and n of samples are functions of k, which are allowed to scale with k to infinity.

Definition 1. *A sequence of estimators $\hat{D}$, indexed by k, is said to be consistent under sample complexity $m(k)$ and $n(k)$ if*

$$\lim_{k\to\infty} R(\hat{D}, k, m, n) = 0. \tag{1.5}$$

We are also interested in the following set:

$$\mathcal{M}_{k,f(k)} = \left\{(P, Q) : |P| = |Q| = k, \frac{P_i}{Q_i} \le f(k),\ \forall\ 1 \le i \le k\right\}, \tag{1.6}$$

which contains distributions (P, Q) with density ratio bounded by $f(k)$. We define the worst-case quadratic risk over $\mathcal{M}_{k,f(k)}$ as

$$R(\hat{D}, k, m, n, f(k)) \triangleq \sup_{(P,Q)\in\mathcal{M}_{k,f(k)}} \mathbb{E}[(\hat{D}(M, N) - D(P\|Q))^2], \tag{1.7}$$

and define the corresponding minimax quadratic risk as

$$R^*(k,m,n,f(k)) \triangleq \inf_{\hat{D}} R(\hat{D},k,m,n,f(k)). \tag{1.8}$$

1.3 Notations

We adopt the following notation to express asymptotic scaling of quantities with n: $f(n) \lesssim g(n)$ represents that there exists a constant c s.t. $f(n) \leq cg(n)$; $f(n) \gtrsim g(n)$ represents that there exists a constant c s.t. $f(n) \geq cg(n)$; $f(n) \asymp g(n)$ when $f(n) \gtrsim g(n)$ and $f(n) \lesssim g(n)$ hold simultaneously; $f(n) \gg g(n)$ represents that for all $c > 0$, there exists $n_0 > 0$ s.t. for all $n > n_0$, $|f(n)| \geq c|g(n)|$; and $f(n) \ll g(n)$ represents that for all $c > 0$, there exists $n_0 > 0$ s.t. for all $n > n_0$, $|f(n)| \leq cg(n)$.

1.4 Related Work

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 on the densities, the estimator proposed in [6] is based on a k-nearest neighbor approach, 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 for variance. All of these approaches exploit the smoothness of continuous densities or density ratios, which guarantees that samples falling into a certain neighborhood 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, an example is provided in [5] to show that the estimation of KL divergence can be difficult even for continuous distributions if the density has sharp dips.

A more general problem of estimating the f-divergence for d-dimensional distributions is studied in [8]. A weighted ensemble divergence estimator is proposed, which is based on a weighted ensemble of plug-in estimators to trade bias for variance. Such an estimator is shown to converge faster than the simple plug-in estimator, perform well for high dimensions and be easily implemented.

Estimation of KL divergence when the distributions P and Q are discrete has been studied in [9–11] for the regime with *fixed* alphabet size 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 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 is, regarding 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 particular, such distributions dominate the risk in (1.3) and make the construction of consistent estimators challenging.

1.5 Contribution of Thesis

We summarize our main contribution in the following three theorems, whose detailed proofs are given respectively in Chapters 2, 3 and 4.

Our first result, based on Le Cam's two-point method [16], is that there is no consistent estimator of KL divergence over the distribution set $\mathcal{M}_k$.

Theorem 1. *For any $m, n \in \mathbb{N}$, and $k \geq 2$, $R^*(k, m, n)$ is infinite. Therefore, there exists no consistent estimator of KL divergence over the set $\mathcal{M}_k$.*

The intuition behind this result is that the set $\mathcal{M}_k$ contains distributions Q that have arbitrarily small components that contribute significantly to KL divergence but require arbitrarily large number of samples to estimate accurately. However, in practical applications, it is reasonable to assume that the ratio of P to Q is bounded. Thus, we further focus on the set $\mathcal{M}_{k,f(k)}$ given in (1.6) that contains distribution pairs (P, Q) with their density ratio bounded by $f(k)$.

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

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

$$m \gg k \vee \log^2(f(k)) \quad \textit{and} \quad n \gg kf(k). \tag{1.9}$$

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

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. This naturally inspires the question 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 show that the augmented plug-in estimator is not minimax optimal, and that an estimator that employs both the polynomial approximation and plug-in approaches is minimax optimal, and the following theorem characterizes the minimax risk.

Theorem 3. *If $f(k) \geq \log^2 k$, $\log m \lesssim \log k$, $\log^2 n \lesssim k^{1-\epsilon}$, $m \gtrsim \frac{k}{\log k}$ and $n \gtrsim \frac{kf(k)}{\log k}$, where ϵ is any positive constant, then the minimax risk satisfies*

$$R^*(k,m,n,f(k)) \asymp \left(\frac{k}{m\log k} + \frac{kf(k)}{n\log k}\right)^2 + \frac{\log^2 f(k)}{m} + \frac{f(k)}{n}. \tag{1.10}$$

The key idea in the construction of the minimax optimal estimator is the application of a 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 a joint polynomial approximation is difficult 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 a generalized Le Cam's method involving two composite hypotheses, as in the case of entropy estimation [13]. But the challenge here that requires special technical treatment is the construction of prior distributions for (P, Q) that satisfy the bounded density ratio constraint.

We note that the first term $\left(\frac{k}{m\log k}+\frac{kf(k)}{n\log k}\right)^2$ in (1.10) captures the squared bias, and the remaining terms correspond to the variance. If we compare the upper bound on the risk in (3.3) for the augmented plug-in estimator with the minimax risk in (1.10), 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 squared error, the sufficient and necessary conditions on the sample complexity are given by

$$m \gg (\log^2 f(k) \vee \frac{k}{\log k}), \text{ and } n \gg \frac{kf(k)}{\log k}. \tag{1.11}$$

The comparison of (1.11) with (1.9) shows that the augmented plug-in estimator is strictly sub-optimal.

CHAPTER 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 $\mathcal{M}_k$.

We will provide a rigorous proof in the following section. 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 squared difference between the divergence values.

2.1 Lower Bound for the Worst Case Risk over Set $\mathcal{M}_k$

Proof. For any fixed (k, m, n), applying Le Cam's two-point method, we have

$$\begin{aligned} R^*(k,m,n) \geq \frac{1}{16}(D(P^{(1)}\|Q^{(1)}) - D(P^{(2)}\|Q^{(2)}))^2 \qquad (2.1)\\ \exp\left(-mD(P^{(1)}\|P^{(2)}) - nD(Q^{(1)}\|Q^{(2)})\right). \end{aligned}$$

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 keep $(D(P^{(1)}\|Q^{(1)}) - D(P^{(2)}\|Q^{(2)}))^2$ large. We construct the following two pairs of distributions:

$$P^{(1)} = P^{(2)} = \left(\frac{1}{2(k-1)}, \ldots, \frac{1}{2(k-1)}, \frac{1}{2}\right), \qquad (2.2)$$

$$Q^{(1)} = \left(\frac{1-\epsilon_1}{k-1}, \ldots, \frac{1-\epsilon_1}{k-1}, \epsilon_1\right), \qquad (2.3)$$

$$Q^{(2)} = \left(\frac{1-\epsilon_2}{k-1}, \ldots, \frac{1-\epsilon_2}{k-1}, \epsilon_2\right), \qquad (2.4)$$

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

$$D(P^{(1)} \| P^{(2)}) = 0, \tag{2.5}$$

$$D(Q^{(1)} \| Q^{(2)}) = (1-\epsilon_1) \log \frac{1-\epsilon_1}{1-\epsilon_2} + \epsilon_1 \log \frac{\epsilon_1}{\epsilon_2}. \tag{2.6}$$

Furthermore,

$$\begin{aligned} D(Q^{(1)} \| Q^{(2)}) &= (1-\epsilon_1) \log \left(1 + \frac{\epsilon_2 - \epsilon_1}{1-\epsilon_2}\right) + \epsilon_1 \log \frac{\epsilon_1}{\epsilon_1 + \frac{1}{4n}} \\ &= (1-\epsilon_1) \log \left(1 + \frac{1}{4n(1-\epsilon_2)}\right) + \epsilon_1 \log \frac{\epsilon_1}{\epsilon_1 + \frac{1}{4n}} \\ &< \frac{1-\epsilon_1}{4n(1-\epsilon_2)}. \end{aligned} \tag{2.7}$$

Since $\epsilon_1 > 0$ 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 $(P^{(2)}, Q^{(2)})$,

$$\begin{aligned} \big(D(P^{(1)} \| Q^{(1)}) - D(P^{(2)} \| Q^{(2)})\big)^2 &= \left(\frac{1}{2} \log \big(\frac{1-\epsilon_2}{1-\epsilon_1}\big) + \frac{1}{2} \log \frac{\epsilon_2}{\epsilon_1}\right)^2 \\ &= \left(\frac{1}{2} \log \big(\frac{1-\epsilon_2}{1-\epsilon_1}\big) + \frac{1}{2} \log \big(1 + \frac{1}{4n\epsilon_1}\big)\right)^2. \end{aligned} \tag{2.8}$$

Note that $|\frac{1}{2} \log \big(\frac{1-\epsilon_2}{1-\epsilon_1}\big)|$ is upper bounded by $\log 2$. The only constraint for (2.8) is $0 < \epsilon_1 < 1/4$. Hence, we can choose ϵ_1 to be arbitrarily small, such that $\frac{1}{4n\epsilon_1}$ is arbitrarily large for any fixed k, m and n. Consequently,

$$\big(D(P^{(1)} \| Q^{(1)}) - D(P^{(2)} \| Q^{(2)})\big)^2 = \left(\frac{1}{2} \log \big(\frac{1-\epsilon_2}{1-\epsilon_1}\big) + \frac{1}{2} \log \big(1 + \frac{1}{4n\epsilon_1}\big)\right)^2 \to \infty, \tag{2.9}$$

as $\epsilon_1 \to 0$. Therefore, the minimax quadratic risk lower bound is infinity for any fixed k, m and n, which implies that there does not exist any consistent estimator over the set $\mathcal{M}_k$. □

2.2 Example

We next give an example of binary distributions, i.e., $k = 2$, to illustrate how the 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 $Q_2 = (\frac{1}{2s}, 1 - \frac{1}{2s})$, where $s > 0$. For any $n \in \mathbb{N}$, we choose s sufficiently large such that $D(Q_1 \| Q_2) < \frac{1}{n}$. Thus, the error probability for distinguishing Q_1 and Q_2 with n samples is greater than a constant. However, $D(P_1 \| Q_1) \asymp s$ and $D(P_2 \| Q_2) \asymp \log s$. Hence, the minimax risk, which is lower bounded by the difference of the above divergences, can be made arbitrarily large by letting $s \to \infty$. This 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 the 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 the KL divergence accurately over the set $\mathcal{M}_k$ under the minimax setting.

CHAPTER 3

AUGMENTED PLUG-IN ESTIMATOR

Since there exists no consistent estimator of KL divergence over the set $\mathcal{M}_k$, we study estimators over the set $\mathcal{M}_{k,f(k)}$.

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

The "plug-in" approach is a natural way 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

$$\hat{D}_{\mathrm{plug-in}}(M,N) = D(\hat{P}\|\hat{Q}), \tag{3.1}$$

where $\hat{P} = (\hat{P}_1, \dots, \hat{P}_k)$ and $\hat{Q} = (\hat{Q}_1, \dots, \hat{Q}_k)$ denote the empirical distributions with $\hat{P}_i = \frac{M_i}{m}$ and $\hat{Q}_i = \frac{N_i}{n}$, respectively, for $i = 1, \cdots, k$.

As frequently observed in functional estimation problems, the plug-in estimator is simple but may cause a lot of problems. Unlike the entropy estimation problem, where the plug-in estimator $\hat{H}_{\mathrm{plug-in}}$ is asymptotically efficient in the "fixed P, large n" regime, the direct plug-in estimator $\hat{D}_{\mathrm{plug-in}}$ in (3.1) 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 $\hat{D}_{\mathrm{plug-in}}$.

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 $\hat{Q}'_i$:

$$\hat{D}_{\mathrm{A-plug-in}}(M,N) = \sum_{i=1}^{k} \frac{M_i}{m} \log \frac{M_i/m}{(N_i+1)/n}. \tag{3.2}$$

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 the worst case risk of augmented plug-in estimator, respectively.

3.2 Upper Bound for the Worst Case Risk of Augmented Plug-in Estimator

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

$$R(\hat{D}_{\mathrm{A-plug-in}}, k, m, n, f(k)) \lesssim \left(\frac{kf(k)}{n} + \frac{k}{m}\right)^2 + \frac{\log^2(k)}{m} + \frac{\log^2 f(k)}{m} + \frac{f(k)}{n}. \tag{3.3}$$

Therefore, if $m \gg (k \vee \log^2 f(k))$ and $n \gg kf(k)$,

$$R(\hat{D}_{\mathrm{A-plug-in}}, k, m, n, f(k)) \to 0, ask \to \infty.$$

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

It can be seen that in the risk bound (3.3), the first term captures the squared bias, and the remaining terms correspond to the variance.

3.3 Lower Bound for the Worst Case Risk of Augmented Plug-in Estimator

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

$$R(\hat{D}_{\mathrm{A-plug-in}}, k, m, n, f(k)) \gtrsim 1. \tag{3.4}$$

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

$$\sup_{(P,Q)\in\mathcal{M}_{k,f(k)}} \mathbb{E}[\hat{D}_{\mathrm{A-plug-in}}(m,n) - D(P\|Q)] \geq (\frac{k}{m} \wedge 1) - \frac{kf(k)}{n} \tag{3.5a}$$

$$\mathbb{E}[\hat{D}_{\mathrm{A-plug-in}}(m,n) - D(P\|Q)] \leq \log\left(1+\frac{k}{m}\right) - \frac{k-1}{k}\exp(-\frac{1.05n}{kf(k)}). \tag{3.5b}$$

1) If $m \lesssim k$ and $n \gg kf(k)$, the lower bound in (3.5a) 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 $m \gg k$ and $n \lesssim kf(k)$, the upper bound in (3.5b) is upper bounded by a negative constant, for large k. This implies that the risk is lower bounded by a positive constant.

3) If $m \lesssim k$ and $n \lesssim kf(k)$, the lower bound (3.5a) and the upper bound (3.5b) provide no useful information. Hence, we design another approach for this case as follows.

The bias of the augmented plug-in estimator can be decomposed into:

1. bias due to estimating $\sum_{i=1}^k P_i \log P_i$;
2. bias due to estimating $\sum_{i=1}^k P_i \log Q_i$.

It can be shown that the first bias term is always positive for any distribution P. The second bias term is always negative for any distribution Q. Hence, the two bias terms may cancel out partially or even fully. Thus, to show that the risk is bounded away from zero, the idea is to first determine which bias dominates, and then to accordingly construct a pair of distributions such that the dominant bias is either lower bounded by a positive constant or upper bounded by a negative constant.

If $\frac{k}{m} \geq (1+\epsilon)\frac{\alpha kf(k)}{n}$, where $\epsilon > 0$ and $0 < \alpha < 1$ are constants, and which implies that the number of samples drawn from P is smaller than the number of samples drawn from Q, the first bias term dominates. If P is uniform and $Q = \left(\frac{1}{\alpha kf(k)}, \cdots, \frac{1}{\alpha kf(k)}, 1 - \frac{k-1}{\alpha kf(k)}\right)$, then it can be shown that the bias (and hence the risk) is lower bounded by a positive constant.

If $\frac{k}{m} < (1+\epsilon)\frac{\alpha k f(k)}{n}$, which implies that the number of samples drawn from P is larger than the number of samples drawn from Q, the second bias term dominates.

- If $n \le kf(k)$, set P to be uniform and $Q = \left(\frac{1}{kf(k)},\ \cdots,\ \frac{1}{kf(k)},\ 1-\frac{k-1}{kf(k)}\right)$.
- If $n > kf(k)$, set $P = \left(\frac{f(k)}{n}, \cdots, \frac{f(k)}{n}, 1-\frac{(k-1)f(k)}{n}\right)$, and $Q = \left(\frac{1}{n}, \ldots, \frac{1}{n}, 1-\frac{k-1}{n}\right)$.

It can be shown that the bias is upper bounded by a negative constant. Hence, the risk is lower bounded by a positive constant.

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

$$P^{(1)} = \left(\frac{1}{3(k-1)}, \cdots, \frac{1}{3(k-1)}, \frac{2}{3}\right), \tag{3.6}$$

$$P^{(2)} = \left(\frac{1-\epsilon}{3(k-1)}, \cdots, \frac{1-\epsilon}{3(k-1)}, \frac{2+\epsilon}{3}\right), \tag{3.7}$$

$$Q^{(1)} = Q^{(2)} = \left(\frac{1}{3(k-1)f(k)}, \cdots, \frac{1}{3(k-1)f(k)}, 1-\frac{1}{3f(k)}\right). \tag{3.8}$$

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

CHAPTER 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 chapter, we describe ideas and central arguments underlying this theorem, with detailed proofs relegated to the appendix.

4.1 Poisson Sampling

The sufficient statistics for estimating $D(P\|Q)$ are the histograms of the samples $M = (M_1, \dots, M_k)$ and $N = (N_1, \dots, N_k)$, and M and N are multinomial distributed. However, the histograms are not independent across different bins, which is hard to analyze. In this subsection, we introduce the *Poisson sampling* technique to handle the dependency of the multinomial distribution across different bins, as in [13] 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 replace the deterministic sample sizes m and n with 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 then independent across different bins, which significantly simplifies the analysis.

Analogous to the minimax risk (1.8), we define its counterpart under the Poisson sampling model as

$$\widetilde{R}^*(k,m,n,f(k)) \triangleq \inf_{\hat{D}} \sup_{(P,Q)\in\mathcal{M}_{k,f(k)}} \mathbb{E}[(\hat{D}(M,N) - D(P\|Q))^2], \tag{4.1}$$

where the expectation is taken over $M_i \sim \text{Poi}(nP_i)$ and $N_i \sim \text{Poi}(nQ_i)$ for

$i = 1, \dots, k$. Since the Poisson 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 $c > \frac{1}{4}$ such that*

$$\begin{aligned}\widetilde{R}^*(k, 2m, 2n, f(k)) - e^{-cm}\log^2 f(k) - e^{-cn}\log^2 f(k) \qquad\qquad (4.2)\\ \leq R^*(k, m, n, f(k)) \leq 4\widetilde{R}^*(k, m/2, n/2, f(k)).\end{aligned}$$

Proof. See Appendix C. □

Thus, in order to show Theorem 3, it suffices to bound the Poisson risk $\widetilde{R}^*(k, m, n, f(k))$. In Section 4.2, a lower bound on the minimax risk with deterministic sample size is derived, and in Section 4.3, an upper bound on the minimax risk with Poisson sampling is derived, which further yields an upper bound on the minimax risk with deterministic sample size. It can be shown that the upper and lower bounds 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 $\log^2 n \lesssim k$, $m \gtrsim \frac{k}{\log k}$, $n \gtrsim \frac{kf(k)}{\log k}$,*

$$R^*(k, m, n, f(k)) \gtrsim \left(\frac{k}{m\log k} + \frac{kf(k)}{n\log k}\right)^2 + \frac{\log^2 f(k)}{m} + \frac{f(k)}{n}. \qquad (4.3)$$

Outline of Proof. We describe the main idea in the development of the lower bound, with the detailed proof provided in Appendix D.

To prove Proposition 3, it suffices to show that the minimax risk is lower bounded separately by each individual terms in (4.3) in the order sense. The proof for the last two terms requires the Le Cam's two-point method, and the proof for the first term requires a more general method, as we outline in the following.

Le Cam's two-point method: The last two terms in the lower bound correspond to the variance of the estimator.

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

$$P^{(1)} = \left(\frac{1}{3(k-1)}, \dots, \frac{1}{3(k-1)}, \frac{2}{3}\right), \tag{4.4}$$

$$P^{(2)} = \left(\frac{1-\epsilon}{3(k-1)}, \dots, \frac{1-\epsilon}{3(k-1)}, \frac{2+\epsilon}{3}\right), \tag{4.5}$$

$$Q^{(1)} = Q^{(2)} = \left(\frac{1}{3(k-1)f(k)}, \dots, \frac{1}{3(k-1)f(k)}, 1 - \frac{1}{3f(k)}\right), \tag{4.6}$$

where $\epsilon = \frac{1}{\sqrt{m}}$.

The bound $R^*(k,m,n,f(k)) \gtrsim \frac{f(k)}{n}$ can be shown by choosing

$$P^{(1)} = P^{(2)} = \left(\frac{1}{3(k-1)},\ 0,\ \dots,\ \frac{1}{3(k-1)},\ 0,\ \frac{5}{6}\right), \tag{4.7}$$

$$Q^{(1)} = \left(\frac{1}{2(k-1)f(k)}, \dots, \frac{1}{2(k-1)f(k)}, 1 - \frac{1}{2f(k)}\right), \tag{4.8}$$

$$Q^{(2)} = \left(\frac{1-\epsilon}{2(k-1)f(k)}, \frac{1+\epsilon}{2(k-1)f(k)}, \dots, \frac{1-\epsilon}{2(k-1)f(k)}, \frac{1+\epsilon}{2(k-1)f(k)}, 1 - \frac{1}{2f(k)}\right), \tag{4.9}$$

where $\epsilon = \sqrt{\frac{f(k)}{n}}$.

Generalized Le Cam's method: In order to show that $R^*(k,m,n,f(k)) \gtrsim \left(\frac{k}{m\log k} + \frac{kf(k)}{n\log k}\right)^2$, it suffices to show that $R^*(k,m,n,f(k)) \gtrsim \left(\frac{k}{m\log k}\right)^2$ and $R^*(k,m,n,f(k)) \gtrsim \left(\frac{kf(k)}{n\log k}\right)^2$. These two lower bounds can be shown by applying a generalized Le Cam's method, which involves the following two composite hypotheses [16]:

$$H_0 : D(P\|Q) \le t \quad \text{versus} \quad H_1 : D(P\|Q) \ge t + d. \tag{4.10}$$

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 Bayesian 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 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 unknown. In this way, we only need to design priors on one distribution, which can be shown to be sufficient for the proof of the lower bound.

In order to show that $R^*(k, m, n, f(k)) \gtrsim \left(\frac{k}{m \log k}\right)^2$, 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 steps similar to those in [13], we can obtain the desired result.

In order to show that $R^*(k, m, n, f(k)) \gtrsim \left(\frac{kf(k)}{n \log k}\right)^2$, we set

$$P = \Big(\frac{f(k)}{n \log k}, \dots, \frac{f(k)}{n \log k}, 1 - \frac{(k-1)f(k)}{n \log k}\Big), \tag{4.11}$$

and assume P is known. 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 the 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 conditions $f(k) \geq \log^2 k$ and $\log^2 n \lesssim k$ in Proposition 3.

4.3 Minimax Upper Bound via Optimal Estimator

Comparing the lower bound in Proposition 3 with the upper bound in Proposition 1 that characterizes an upper bound on the risk for the augmented plug-in estimator, it is clear that there is a difference of a $\log k$ 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 design an improved estimator. Entropy estimation [13, 15] suggests incorporating a polynomial approximation into the estimator in order to reduce the bias with price of the variance. In this subsection, we construct an estimator using this approach, and characterize an upper bound on the minimax risk in Proposition 4.

The KL divergence $D(P\|Q)$ can be written as

$$D(P\|Q) = \sum_{i=1}^{k} P_i \log P_i - \sum_{i=1}^{k} P_i \log Q_i. \tag{4.12}$$

The first term equals the entropy of P, and the minimax optimal entropy estimator (denoted by $\hat{D}_1$) in [13] can be applied to estimate it. The major challenge in estimating $D(P\|Q)$ arises due to the second term. We overcome the challenge by using a polynomial approximation to reduce the bias when Q_i is small. Under Poisson sampling model, unbiased estimators can be constructed for any polynomials of P_i and Q_i. Thus, if we approximate $P_i \log Q_i$ by polynomials, and then construct unbiased estimator for the polynomials, the bias of estimating $P_i \log Q_i$ is reduced to the error in the approximation of $P_i \log Q_i$ using polynomials.

A natural idea is to construct polynomial approximation for $|P_i \log Q_i|$ in two dimensions, exploiting the fact that $|P_i \log Q_i|$ is bounded by $f(k) \log k$. However, it is challenging to find the explicit form of the best polynomial approximation in this case [17].

On the other hand, a one-dimensional polynomial approximation of $\log Q_i$ also appears challenging to develop. First of all, the function $\log x$ on interval $(0, 1]$ is not bounded due to the 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 implicitly 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.

Another approach is to rewrite 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 using polynomial approximation and then estimated accurately (see [18, Section 7.5.4] and [13]), it is difficult to find a good estimator for $\frac{P_i}{Q_i}$.

Motivated by the unsuccessful approaches, we design our estimator as follows. We rewrite $P_i \log Q_i$ as $P_i \frac{1}{Q_i} Q_i \log Q_i$. When Q_i is small, we construct a polynomial approximation $\mu_L(Q_i)$ for $Q_i \log Q_i$, which does not contain a zero-degree term. Then, $\frac{\mu_L(Q_i)}{Q_i}$ is also a polynomial, which can be used to approximate $\frac{1}{Q_i} Q_i \log Q_i$. Thus, an unbiased estimator for $\frac{\mu_L(Q_i)}{Q_i}$ is constructed. Note that the error in the approximation of $\log Q_i$ using $\frac{\mu_L(Q_i)}{Q_i}$ is not bounded, which implies that the bias of using unbiased estimator of $\frac{\mu_L(Q_i)}{Q_i}$ to estimate $\log Q_i$ is not bounded. However, we can show that the bias of estimating $P_i \log Q_i$ is bounded, which is due to the density ratio constraint $f(k)$. The fact that when Q_i is small, P_i is also small, helps to reduce the bias. 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 $m_1' \sim$Poi(m), and $m_2' \sim$Poi(m), and then draw m_1' and m_2' i.i.d. samples from distribution P, where we use $M = (M_1, \ldots, M_k)$ and $M' = (M_1', \ldots, M_k')$ to denote the histograms of m_1' samples and m_2' samples, respectively. We then use these samples to estimate $\sum_{i=1}^k P_i \log P_i$ following the entropy estimator proposed in [13]. Next, we draw $n_1' \sim$Poi(n) and $n_2' \sim$Poi(n) independently. We then draw n_1' and n_2' i.i.d. samples from distribution Q, where 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 note that $N_i \sim$Poi(nQ_i), and $N_i' \sim$Poi(nQ_i).

We then focus on the estimation of $\sum_{i=1}^k P_i \log Q_i$. If $Q_i \in [0, \frac{c_1 \log k}{n}]$, we construct a polynomial approximation for the function $P_i \log Q_i$ and further estimate the polynomial function. And if $Q_i \in [\frac{c_1 \log k}{n}, 1]$, we use the bias corrected augmented plug-in estimator. We use N' to determine whether to use a polynomial estimator or plug-in estimator, and we use N to estimate $\sum_{i=1}^k P_i \log Q_i$. Intuitively, if N_i' is large, then Q_i is more likely to be large, and vice versa. Based on the generation scheme, N and N' are independent. Such independence significantly simplifies the analysis.

We let $L = \lfloor c_0 \log k \rfloor$, where c_0 is a constant to be determined later, and denote the degree-L best polynomial approximation of the function $x \log x$ over the interval $[0, 1]$ as $\sum_{j=0}^L a_j x^j$. We further scale the interval $[0, 1]$ to $[0, \frac{c_1 \log k}{n}]$. Then we have the best polynomial approximation of the function $x \log x$ over the interval $[0, \frac{c_1 \log k}{n}]$ as follows:

$$\gamma_L(x) = \sum_{j=0}^{L} \frac{a_j n^{j-1}}{(c_1 \log k)^{j-1}} x^j - \left(\log \frac{n}{c_1 \log k} \right) x. \tag{4.13}$$

Following the result in [18, Section 7.5.4], the error in approximating $x \log x$

by $\gamma_L(x)$ over the interval $[0, \frac{c_1 \log k}{n}]$ can be upper bounded as follows:

$$\sup_{x \in [0, \frac{c_1 \log k}{n}]} |\gamma_L(x) - x \log x| \lesssim \frac{1}{n \log k}. \tag{4.14}$$

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:

$$\frac{a_0 c_1 \log k}{n} \lesssim \frac{1}{n \log k}. \tag{4.15}$$

Now, subtracting the zero-degree term from $\gamma_L(x)$ in (4.13) yields the following polynomial:

$$\begin{aligned} \mu_L(x) &\triangleq \gamma_L(x) - \frac{a_0 c_1 \log k}{n} \\ &= \sum_{j=1}^{L} \frac{a_j n^{j-1}}{(c_1 \log k)^{j-1}} x^j - \left(\log \frac{n}{c_1 \log k} \right) x. \end{aligned} \tag{4.16}$$

The error in approximating $x \log x$ by $\mu_L(x)$ over the interval $[0, \frac{c_1 \log k}{n}]$ can also be upper bounded by $\frac{1}{n \log k}$, because

$$\begin{aligned} \sup_{x \in [0, \frac{c_1 \log k}{n}]} |\mu_L(x) - x \log x| &= \sup_{x \in [0, \frac{c_1 \log k}{n}]} \left| \gamma_L(x) - x \log x - \frac{a_0 c_1 \log k}{n} \right| \\ &\leq \sup_{x \in [0, \frac{c_1 \log k}{n}]} |\gamma_L(x) - x \log x| + \left| \frac{a_0 c_1 \log k}{n} \right| \\ &\lesssim \frac{1}{n \log k}. \end{aligned} \tag{4.17}$$

The bound in (4.17) implies that although $\mu_L(x)$ is not the best polynomial approximation of $x \log x$, the error in the approximation by $\mu_L(x)$ has the same order as that by $\gamma_L(x)$. Compared to $\gamma_L(x)$, there is no zero-degree term in $\mu_L(x)$, and hence $\frac{\mu_L(x)}{x}$ is a valid polynomial approximation of $\log x$. Although the approximation error of $\log x$ using $\frac{\mu_L(x)}{x}$ is unbounded, the error in the approximation of $P_i \log Q_i$ using $P_i \frac{\mu_L(Q_i)}{Q_i}$ can be bounded. More importantly, by the way in which we constructed $\mu_L(x)$, $P_i \frac{\mu_L(Q_i)}{Q_i}$ is a polynomial function of P_i and Q_i, for which an unbiased estimator can be constructed. More specifically, the error in using $P_i \frac{\mu_L(Q_i)}{Q_i}$ to approximate $P_i \log Q_i$ can be bounded as follows:

$$\left|P_i\frac{\mu_L(Q_i)}{Q_i} - P_i\log Q_i\right| = \frac{P_i}{Q_i}|\mu_L(Q_i) - Q_i\log Q_i| \lesssim \frac{f(k)}{n\log k}, \tag{4.18}$$

if $Q_i \in [0, \frac{c_1\log k}{n}]$. We further define the factorial moment of x by $(x)_j \triangleq \frac{x!}{(x-j)!}$. If X ∼Poi(λ), $\mathbb{E}[(X)_j] = \lambda^j$. Based on this fact, we construct an unbiased estimator for $\frac{\mu_L(Q_i)}{Q_i}$ as follows:

$$g_L(N_i) = \sum_{j=1}^{L}\frac{a_j}{(c_1\log k)^{j-1}}(N_i)_{j-1} - \left(\log\frac{n}{c_1\log k}\right). \tag{4.19}$$

We then construct our estimator for $\sum_{i=1}^k P_i\log Q_i$ as follows:

$$\hat{D}_2 = \sum_{i=1}^{k}\left(\frac{M_i}{m}g_L(N_i)\mathbb{1}_{\{N_i'\le c_2\log k\}} + \frac{M_i}{m}\Big(\log\frac{N_i+1}{n} - \frac{1}{2(N_i+1)}\Big)\mathbb{1}_{\{N_i'>c_2\log k\}}\right). \tag{4.20}$$

For the term $\sum_{i=1}^k P_i\log P_i$ in $D(P\|Q)$, we use the minimax optimal entropy estimator proposed in [13]. We note that $\gamma_L(x)$ is the best polynomial approximation of the function $x\log x$. And an unbiased estimator of $\gamma_L(x)$ is as follows:

$$g_L'(M_i) = \frac{1}{m}\sum_{j=1}^{L'}\frac{a_j}{(c_1'\log k)^{j-1}}(M_i)_j - \left(\log\frac{m}{c_1'\log k}\right)M_i. \tag{4.21}$$

Based on $g_L'(M_i)$, the estimator $\hat{D}_1$ for $\sum_{i=1}^k P_i\log P_i$ is constructed as follows:

$$\hat{D}_1 = \sum_{i=1}^{k}\left(g_L'(M_i)\mathbb{1}_{\{M_i'\le c_2'\log k\}} + (\frac{M_i}{m}\log\frac{M_i}{m} - \frac{1}{2m})\mathbb{1}_{\{M_i'>c_2'\log k\}}\right). \tag{4.22}$$

Combining the estimator $\hat{D}_1$ in (4.22) for $\sum_{i=1}^k P_i\log P_i$ [13] and the estimator $\hat{D}_2$ in (4.20) for $\sum_{i=1}^k P_i\log Q_i$, we obtain the estimator $\widetilde{D}_{\text{opt}}$ for KL divergence $D(P\|Q)$ as

$$\widetilde{D}_{\text{opt}} = \hat{D}_1 - \hat{D}_2. \tag{4.23}$$

Due to the density ratio constraint, we can show that $0 \leq D(P\|Q) \leq \log f(k)$. We therefore construct an estimator $\hat{D}_{\text{opt}}$ as follows:

$$\hat{D}_{\text{opt}} = \widetilde{D}_{\text{opt}} \vee 0 \wedge \log f(k). \tag{4.24}$$

The following proposition characterizes an upper bound on the worse case risk of $\hat{D}_{\text{opt}}$.

Proposition 4. *If* $\log^2 n \lesssim k^{1-\epsilon}$, *where* ϵ *is any positive constant, and* $\log m \leq C \log k$ *for some constant* C, *then there exist* c_0, c_1 *and* c_2 *depending on* C *only, such that*

$$\sup_{(P,Q)\in\mathcal{M}_{k,f(k)}} \mathbb{E}\left[\left(\hat{D}_{\text{opt}}(M,N) - D(P\|Q)\right)^2\right] \lesssim \left(\frac{k}{m\log k} + \frac{kf(k)}{n\log k}\right)^2 + \frac{\log^2 f(k)}{m} + \frac{f(k)}{n}. \tag{4.25}$$

Proof. See Appendix E. □

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.

CHAPTER 5

NUMERICAL EXPERIMENTS

In this chapter, we provide numerical results to demonstrate the performance of our estimators, and compare our augmented plug-in estimator and minimax optimal estimator with a number of other KL divergence estimators.

To implement the minimax optimal estimator, we first compute the coefficients of the best polynomial approximation by applying the Remez algorithm [19]. In our experiments, we replace the N_i' and M_i' in (4.20) and (4.22) with N_i and M_i, which means we use all the samples for both selecting estimators (polynomial or plug-in) and estimation. We choose the constants c_0, c_1 and c_2 following ideas in [15]. More specifically, we set $c_0 = 1.2$, $c_2 \in [0.05, 0.2]$ and $c_1 = 2c_2$.

We compare the performance of the following five estimators: 1) our augmented plug-in estimator (BZLV A-plugin) in (3.2); 2) our minimax optimal estimator (BZLV opt) in (4.23); 3) Han, Jiao and Weissman's modified plug-in estimator (HJW M-plugin) in [20] ; 4) Han, Jiao and Weissman's minimax optimal estimator (HJW opt) [20]; 5) Zhang and Grabchak's estimator (ZG) in [9] which is constructed for fix-alphabet size setting.

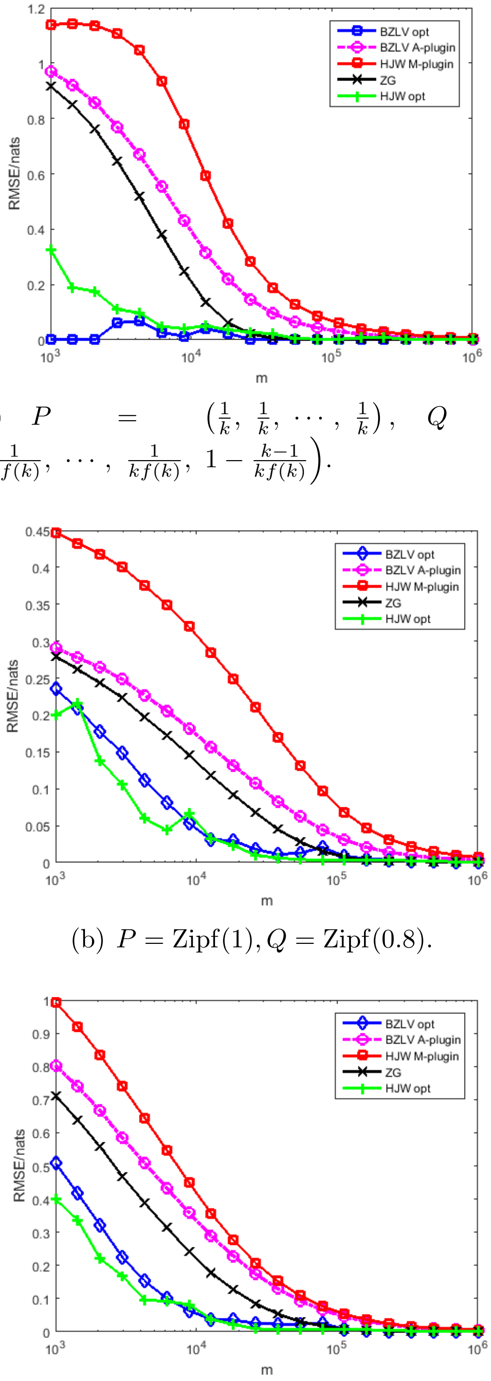
We first compare the performance of the five estimators under the traditional setting in which we set $k = 10^4$ and let m and n change. We choose two types of distributions (P, Q). The first type is given by

$$P = \left(\frac{1}{k}, \frac{1}{k}, \cdots, \frac{1}{k}\right), \quad Q = \left(\frac{1}{kf(k)}, \cdots, \frac{1}{kf(k)}, 1 - \frac{k-1}{kf(k)}\right), \quad (5.1)$$

where $f(k) = 5$. For this pair of (P, Q), the density ratio is $f(k)$ for all but one of the bins, which is in a sense a worst-case for the KL divergence estimation problem. We let m range from 10^3 to 10^6 and set $n = 3f(k)m$. The second type is given by $(P, Q) = (\text{Zipf}(1), \text{Zipf}(0.8))$ and $(P, Q) = (\text{Zipf}(1), \text{Zipf}(0.6))$. The Zipf distribution is a discrete distribution that is commonly used in linguistics, insurance, and the modeling of rare events. If $P = \mathrm{Zipf}(\alpha)$, then $P_i = \frac{i^{-\alpha}}{\sum_{j=1}^{k} j^{-\alpha}}$, for $1 \leq i \leq k$. We let m range from 10^3 to 10^6 and set $n = 0.5f(k)m$, where $f(k)$ is computed for these two pairs of Zipf distributions, respectively.

In Fig. 5.1, we plot the root mean square errors (RMSE) of the five estimators as a function of the sample size m for these three pairs of distributions. It is clear from the figure that our minimax optimal estimator (BZLV opt) and the HJW minimax optimal estimator (HJW opt) outperform the other three approaches. Such a performance improvement is significant especially when the sample size is small. Furthermore, our augmented plug-in estimator (BZLV A-plugin) has a much better performance than the HJW modified plug-in estimator (HJW M-plugin), because the bias of estimating $\sum_{i=1}^{k} P_i \log P_i$ and the bias of estimating $\sum_{i=1}^{k} P_i \log Q_i$ may cancel each other out by the design of our augmented plug-in estimator. Furthermore, the RMSEs of all five estimators converge to zero when the number of samples is sufficiently large.

We next compare the performance of the five estimators under the large-alphabet setting, in which we let k range from 10^3 to 10^6, and set $m = \frac{2k}{\log k}$ and $n = \frac{kf(k)}{\log k}$. We use the same three pairs of distributions as in the previous setting. In Fig. 5.2, we plot the RMSEs of the five estimators as a function of k. It is clear from the figure that our minimax optimal estimator (BZLV opt) and the HJW minimax optimal estimator (HJW opt) have very small estimation errors, which is consistent with our theoretical results of the minimax risk bound. However, the RMSEs of the other three approaches increase with k, which implies that $m = \frac{2k}{\log k}$, $n = \frac{kf(k)}{\log k}$ are insufficient for those estimators.

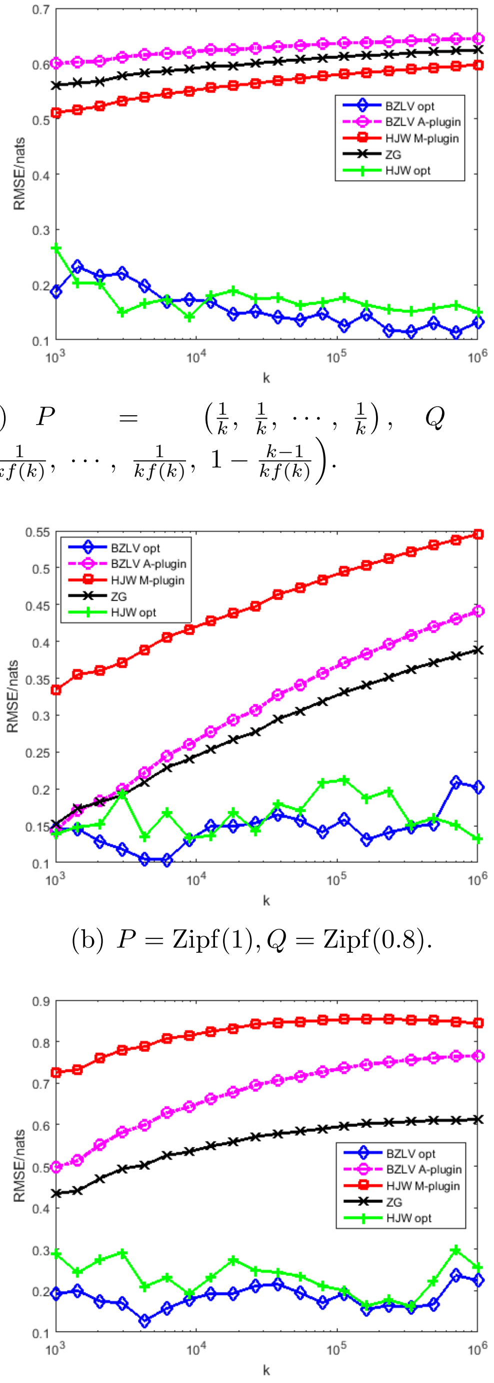


(a) $P = \left(\frac{1}{k}, \frac{1}{k}, \cdots, \frac{1}{k}\right)$, $Q = \left(\frac{1}{kf(k)}, \cdots, \frac{1}{kf(k)}, 1 - \frac{k-1}{kf(k)}\right)$.

(b) $P = \text{Zipf}(1), Q = \text{Zipf}(0.8)$.

(c) $P = \text{Zipf}(1), Q = \text{Zipf}(0.6)$.

Figure 5.1: Comparison of five estimators under traditional setting with $k = 10^4$, m ranging from 10^3 to 10^6 and $n \asymp f(k)m$.



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

(b) $P = \mathrm{Zipf}(1), Q = \mathrm{Zipf}(0.8)$.

(c) $P = \mathrm{Zipf}(1), Q = \mathrm{Zipf}(0.6)$.

Figure 5.2: Comparison of five estimators under large-alphabet setting with k ranging from 10^3 to 10^6, $m = \frac{2k}{\log k}$ and $n = \frac{kf(k)}{\log k}$.

CHAPTER 6

CONCLUSIONS

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

APPENDIX A

PROOF OF PROPOSITION 1

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

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

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

A.1 Bounding the Bias

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

$$\begin{aligned}
&\left|\mathbb{E}\big(\hat{D}_{\mathrm{A-plug-in}}(M,N) - D(P\|Q)\big)\right| \\
=&\left|\mathbb{E}\left(\sum_{i=1}^{k}\left[\frac{M_i}{m}\log\frac{M_i/m}{(N_i+1)/n} - P_i\log\frac{P_i}{Q_i}\right]\right)\right| \\
=&\left|\mathbb{E}\left(\sum_{i=1}^{k}\left[\frac{M_i}{m}\log\frac{M_i}{m} - P_i\log P_i\right]\right) + \mathbb{E}\left(\sum_{i=1}^{k}\left[P_i\log Q_i - \frac{M_i}{m}\log\frac{N_i+1}{n}\right]\right)\right| \\
=&\left|\mathbb{E}\left(\sum_{i=1}^{k}\left[\frac{M_i}{m}\log\frac{M_i}{m} - P_i\log P_i\right]\right) + \mathbb{E}\left(\sum_{i=1}^{k}P_i\log\frac{nQ_i}{N_i+1}\right)\right| \\
\le&\left|\mathbb{E}\left(\sum_{i=1}^{k}\left[\frac{M_i}{m}\log\frac{M_i}{m} - P_i\log P_i\right]\right)\right| + \left|\mathbb{E}\left(\sum_{i=1}^{k}P_i\log\frac{nQ_i}{N_i+1}\right)\right|. \qquad \text{(A.1)}
\end{aligned}$$

The first term in (A.1) is the bias of the plug-in estimator for entropy estimation, which can be bounded as in [21]:

$$\left|\mathbb{E}\left(\sum_{i=1}^{k}\left[\frac{M_i}{m}\log\frac{M_i}{m}-P_i\log P_i\right]\right)\right|\leq\log\left(1+\frac{k-1}{m}\right)<\frac{k}{m}. \tag{A.2}$$

Next, we bound the second term in (A.1) as follows:

$$\begin{aligned}\mathbb{E}\left(\sum_{i=1}^{k}P_i\log\frac{nQ_i}{N_i+1}\right)&=-\sum_{i=1}^{k}P_i\mathbb{E}\left(\log\big(1+\frac{N_i+1-nQ_i}{nQ_i}\big)\right)\\&\overset{(a)}{\geq}-\sum_{i=1}^{k}P_i\mathbb{E}\left(\frac{N_i+1-nQ_i}{nQ_i}\right)\\&=-\sum_{i=1}^{k}P_i\frac{1}{nQ_i}\\&\geq-\frac{kf(k)}{n},\end{aligned} \tag{A.3}$$

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

$$\mathbb{E}\left(\sum_{i=1}^{k}P_i\log\frac{nQ_i}{N_i+1}\right)=\sum_{i=1}^{k}P_i\mathbb{E}\left(\log\frac{nQ_i}{N_i+1}\right)\leq\sum_{i=1}^{k}P_i\log\mathbb{E}\left[\frac{nQ_i}{N_i+1}\right]. \tag{A.4}$$

Let $B(n,p)$ denote the binomial distribution where n is the total number of experiments, and p is the probability that each experiment yields a desired outcome. Note that since $N_i\sim B(n,Q_i)$, then the expectation in (A.4) can be computed as follows:

$$
\begin{aligned}
\mathbb{E}\left[\frac{1}{N_i+1}\right] &= \sum_{j=0}^{n} \frac{1}{j+1}\binom{n}{j} Q_i^j (1-Q_i)^{n-j} \\
&= \sum_{j=0}^{n} \frac{1}{j+1}\frac{n!}{(n-j)!j!} Q_i^j (1-Q_i)^{n-j} \\
&= \frac{1}{n+1}\sum_{j=0}^{n} \frac{(n+1)!}{(n-j)!(j+1)!} Q_i^j (1-Q_i)^{n-j} \\
&= \frac{1}{(n+1)Q_i}\sum_{j=0}^{n} \binom{n+1}{j+1} Q_i^{j+1} (1-Q_i)^{n-j} \\
&= \frac{1}{(n+1)Q_i}(1-(1-Q_i)^{n+1}) < \frac{1}{nQ_i}. \qquad \text{(A.5)}
\end{aligned}
$$

Thus, we obtain

$$
\mathbb{E}\left(\sum_{i=1}^{k} P_i \log \frac{nQ_i}{N_i+1}\right) \leq \sum_{i=1}^{k} P_i \log \mathbb{E}\left[\frac{nQ_i}{N_i+1}\right] < \sum_{i=1}^{k} P_i \log \frac{nQ_i}{nQ_i} = 0. \quad \text{(A.6)}
$$

Combining (A.3) and (A.6), we obtain the following upper bound for the second term in the bias:

$$
\left|\mathbb{E}\left(\sum_{i=1}^{k} P_i \log \frac{nQ_i}{N_i+1}\right)\right| \leq \frac{kf(k)}{n}. \qquad \text{(A.7)}
$$

Hence,

$$
\left|\mathbb{E}\Big(\hat{D}_{\mathrm{A-plug-in}}(M,N) - D(P\|Q)\Big)\right| < \frac{k}{m} + \frac{kf(k)}{n}. \qquad \text{(A.8)}
$$

A.2 Bounding the Variance

Applying the Efron-Stein inequality [22, Theorem 3.1], we have:

$$
\begin{aligned}
\mathrm{Var}[\hat{D}_{\mathrm{A-plug-in}}(M,N)] \leq & \frac{m}{2}\mathbb{E}\left[(\hat{D}_{\mathrm{A-plug-in}}(M,N) - \hat{D}_{\mathrm{A-plug-in}}(M',N))^2\right] \\
& + \frac{n}{2}\mathbb{E}\left[(\hat{D}_{\mathrm{A-plug-in}}(M,N) - \hat{D}_{\mathrm{A-plug-in}}(M,N'))^2\right],
\end{aligned}
\qquad \text{(A.9)}
$$

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

$$\phi(x,y)\triangleq x\log x - x\log y. \tag{A.10}$$

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

$$\hat{D}_{\mathrm{A-plug-in}}(M,N)=\sum_{i=1}^{k}\phi(\frac{M_i}{m},\frac{N_i+1}{n}). \tag{A.11}$$

Let $\tilde{M}_{X_m}$ be the number of samples in bin X_m. We can bound the first term in (A.9) as follows:

$$\begin{aligned}
&\mathbb{E}\left[(\hat{D}_{\mathrm{A-plug-in}}(M,N)-\hat{D}_{\mathrm{A-plug-in}}(M',N))^2\right]\\
=&\mathbb{E}\Bigg[\mathbb{E}\Bigg[\Bigg(\phi(\frac{\tilde{M}_{X_m}+1}{m},\frac{N_{X_m}+1}{n})+\phi(\frac{\tilde{M}_{X'_m}}{m},\frac{N_{X'_m}+1}{n})\\
&\qquad-\phi(\frac{\tilde{M}_{X_m}}{m},\frac{N_{X_m}+1}{n})-\phi(\frac{\tilde{M}_{X'_m}+1}{m},\frac{N_{X'_m}+1}{n})\Bigg)^2\Bigg|X_m,X'_m\Bigg]\Bigg]\\
\overset{(a)}{\le}&4\mathbb{E}\left[\mathbb{E}\left[\left(\phi(\frac{\tilde{M}_{X_m}+1}{m},\frac{N_{X_m}+1}{n})-\phi(\frac{\tilde{M}_{X_m}}{m},\frac{N_{X_m}+1}{n})\right)^2\Bigg|X_m\right]\right]\\
=&4\sum_{j=1}^{k}\mathbb{E}\left[\left(\phi(\frac{\tilde{M}_j+1}{m},\frac{N_j+1}{n})-\phi(\frac{\tilde{M}_j}{m},\frac{N_j+1}{n})\right)^2\right]P_j\\
=&\frac{4}{m^2}\sum_{j=1}^{k}\mathbb{E}\left[\left((\tilde{M}_j+1)\log\frac{\tilde{M}_j+1}{m}-\tilde{M}_j\log\frac{\tilde{M}_j}{m}-\log\frac{N_j+1}{n}\right)^2\right]P_j\\
=&\frac{4}{m^2}\sum_{j=1}^{k}\mathbb{E}\left[\left(\log\frac{\tilde{M}_j+1}{m}+\tilde{M}_j\log(1+\frac{1}{\tilde{M}_j})-\log\frac{N_j+1}{n}\right)^2\right]P_j\\
\overset{(b)}{\le}&\frac{8}{m^2}+\frac{8}{m^2}\sum_{j=1}^{k}\mathbb{E}\left[\left(\log\frac{\tilde{M}_j+1}{m}-\log\frac{N_j+1}{n}\right)^2\right]P_j,
\end{aligned} \tag{A.12}$$

where (a) is due to the fact that X_m is independent and identically distributed as X'_m, and (b) is due to the fact that $0 \le x \log(1+\frac{1}{x}) \le 1$ for all $x > 0$. We rewrite the second term in (A.12) as follows:

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

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

$$\begin{aligned}
&\text{If } \ \tilde{M}_j \le \frac{mP_j}{2}, \ \text{then } \frac{1}{m} \le \frac{\tilde{M}_j+1}{m} \le \frac{P_j}{2}+\frac{1}{m}; \\
&\text{If } \ \tilde{M}_j > \frac{mP_j}{2}, \ \text{then } \frac{P_j}{2}+\frac{1}{m} < \frac{\tilde{M}_j+1}{m} \le 1; \\
&\text{If } \ N_j > \frac{nQ_j}{2}, \ \text{then } \frac{Q_j}{2}+\frac{1}{n} < \frac{N_j+1}{n} \le 1+\frac{1}{n}; \\
&\text{If } \ N_j \le \frac{nQ_j}{2}, \ \text{then } \frac{1}{n} \le \frac{N_j+1}{n} \le \frac{Q_j}{2}+\frac{1}{n}.
\end{aligned}$$

With the above bounds, and assuming that $m > 2$, $n > 2$, we next analyze the following four cases:

1. If $\tilde{M}_j \le \frac{mP_j}{2}$ and $N_j > \frac{nQ_j}{2}$, then we have $\frac{n}{m(n+1)} \le \frac{\tilde{M}_j+1}{m}\frac{n}{N_j+1} \le \frac{\frac{P_j}{2}+\frac{1}{m}}{\frac{Q_j}{2}+\frac{1}{n}}$,

and

$$
\begin{aligned}
&\mathbb{E}\left[\left(\log\frac{\tilde{M}_j+1}{m}\frac{n}{N_j+1}\mathbb{1}_{\{\tilde{M}_j\le\frac{m}{2}P_j\}}\mathbb{1}_{\{N_j>\frac{nQ_j}{2}\}}\right)^2\right]\\
\le&\left[\log^2(\frac{n}{m(n+1)})+\left(\log(\frac{P_j}{2}+\frac{1}{m})-\log(\frac{Q_j}{2}+\frac{1}{n})\right)^2\right]\mathbb{P}\left(\tilde{M}_j\le\frac{m}{2}P_j\right)\\
\overset{(c)}{\le}&\left[\log^2(\frac{1}{2m})+\left(\log(\frac{P_j}{2}+\frac{1}{m})-\log(\frac{Q_j}{2}+\frac{1}{n})\right)^2\right]\exp\big(-\frac{(m-2)P_j}{8}\big)\\
\le&2\left[\log^2(\frac{1}{2m})+\left(\log^2(\frac{P_j}{2}+\frac{1}{m})+\log^2(\frac{Q_j}{2}+\frac{1}{n})\right)\right]\exp\big(-\frac{(m-2)P_j}{8}\big)\\
\le&2\left[\log^2(2m)+\left(\log^2(\frac{P_j}{2})+\log^2(\frac{Q_j}{2})\right)\right]\exp\big(-\frac{(m-2)P_j}{8}\big),
\end{aligned}
\tag{A.13}
$$

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

2. If $\tilde{M}_j>\frac{mP_j}{2}$ and $N_j>\frac{nQ_j}{2}$, then we have $\frac{n}{n+1}(\frac{P_j}{2}+\frac{1}{m})\le\frac{\tilde{M}_j+1}{m}\frac{n}{N_j+1}\le\frac{1}{\frac{Q_j}{2}+\frac{1}{n}}$, and

$$
\begin{aligned}
&\mathbb{E}\left[\left(\log\frac{\tilde{M}_j+1}{m}\frac{n}{N_j+1}\mathbb{1}_{\{\tilde{M}_j>\frac{m}{2}P_j\}}\mathbb{1}_{\{N_j>\frac{nQ_j}{2}\}}\right)^2\right]\\
\le&\left[\log^2(\frac{n}{n+1}(\frac{P_j}{2}+\frac{1}{m}))+\log^2(\frac{Q_j}{2}+\frac{1}{n})\right]\\
\le&2\left[\log^2(\frac{P_j}{4})+\log^2(\frac{Q_j}{2})\right].
\end{aligned}
\tag{A.14}
$$

3. If $\tilde{M}_j\le\frac{mP_j}{2}$ and $N_j\le\frac{nQ_j}{2}$, then we have $\frac{1}{m(\frac{Q_j}{2}+\frac{1}{n})}\le\frac{\tilde{M}_j+1}{m}\frac{n}{N_j+1}\le\frac{nP_j}{2}+\frac{n}{m}$, and

$$
\begin{aligned}
&\mathbb{E}\left[\left(\log\frac{\tilde{M}_j+1}{m}\frac{n}{N_j+1}\mathbb{1}_{\{\tilde{M}_j\le\frac{m}{2}P_j\}}\mathbb{1}_{\{N_j\le\frac{nQ_j}{2}\}}\right)^2\right]\\
\le&\left[\log^2(m(\frac{Q_j}{2}+\frac{1}{n}))+\log^2(n(\frac{P_j}{2}+\frac{1}{m}))\right]\exp\big(-\frac{(m-2)P_j}{8}-\frac{nQ_j}{8}\big)\\
\le&2\left[\log^2(\frac{Q_j}{2})+\log^2 m+\log^2(\frac{P_j}{2})+\log^2 n\right]\exp\big(-\frac{(m-2)P_j}{8}-\frac{nQ_j}{8}\big).
\end{aligned}
\tag{A.15}
$$

4. If $\tilde{M}_j > \frac{mP_j}{2}$ and $N_j \leq \frac{nQ_j}{2}$, then we have $\frac{\frac{P_j}{2}+\frac{1}{m}}{\frac{Q_j}{2}+\frac{1}{n}} \leq \frac{\tilde{M}_j+1}{m}\frac{n}{N_j+1} \leq n$, and

$$
\begin{aligned}
&\mathbb{E}\left[\left(\log\frac{\tilde{M}_j+1}{m}\frac{n}{N_j+1}\mathbb{1}_{\{\tilde{M}_j>\frac{m}{2}P_j\}}\mathbb{1}_{\{N_j\leq\frac{nQ_j}{2}\}}\right)^2\right]\\
\leq&\left[\left(\log(\frac{P_j}{2}+\frac{1}{m})-\log(\frac{Q_j}{2}+\frac{1}{n})\right)^2+\log^2 n\right]\exp\big(-\frac{nQ_j}{8}\big)\\
\leq&2\left[\log^2 n+\left(\log^2(\frac{P_j}{2})+\log^2(\frac{Q_j}{2})\right)\right]\exp\big(-\frac{nQ_j}{8}\big).
\end{aligned}
\tag{A.16}
$$

Combining the four cases together, we have

$$
\begin{aligned}
&\mathbb{E}\left[(\hat{D}_{\mathrm{A-plug-in}}(M,N)-\hat{D}_{\mathrm{A-plug-in}}(M',N))^2\right]\\
\leq&\frac{8}{m^2}+\frac{8}{m^2}\sum_{j=1}^{k}\mathbb{E}\left[\left(\log\frac{\tilde{M}_j+1}{m}-\log\frac{N_j+1}{n}\right)^2\right]P_j\\
\leq&\frac{16}{m^2}\sum_{j=1}^{k}P_j\left[\log^2(\frac{Q_j}{2})+\log^2 m+\log^2(\frac{P_j}{2})+\log^2 n\right]\exp\big(-\frac{(m-2)P_j}{8}-\frac{nQ_j}{8}\big)\\
&+\frac{16}{m^2}\sum_{j=1}^{k}P_j\left[\log^2(2m)+\left(\log^2(\frac{P_j}{2})+\log^2(\frac{Q_j}{2})\right)\right]\exp\big(-\frac{(m-2)P_j}{8}\big)\\
&+\frac{16}{m^2}\sum_{j=1}^{k}P_j\left[\log^2(n)+\left(\log^2(\frac{P_j}{2})+\log^2(\frac{Q_j}{2})\right)\right]\exp\big(-\frac{nQ_j}{8}\big)\\
&+\frac{16}{m^2}\sum_{j=1}^{k}P_j\left[\log^2(\frac{P_j}{4})+\log^2(\frac{Q_j}{2})\right]+\frac{8}{m^2}\\
\leq&\frac{16}{m^2}\sum_{j=1}^{k}P_j\Big[4\log^2(\frac{4}{P_j})+4\log^2(\frac{2}{Q_j})+2\log^2(2m)\exp\big(-\frac{(m-2)P_j}{8}\big)\\
&+2\log^2 n\exp\big(-\frac{nQ_j}{8}\big)\Big]+\frac{8}{m^2}.
\end{aligned}
\tag{A.17}
$$

Now, we analyze the asymptotic behavior of the above four terms in (A.17):

1. It can be shown that $\sum_{j=1}^{k}-P_j\log P_j \leq \log k$ and $\sum_{j=1}^{k}P_j\log^2 P_j \leq \log^2 k$. Hence, we obtain

$$
\sum_{j=1}^{k}P_j\log^2(\frac{4}{P_j})=\sum_{j=1}^{k}P_j(\log^2(P_j)+\log^2 4-2\log P_j\log 4)\leq(\log k+\log 4)^2.
\tag{A.18}
$$

2. Given the bounded ratio constraint $\frac{1}{Q_j} \leq \frac{f(k)}{P_j}$, we have

$$\begin{aligned}\sum_{j=1}^{k} P_i \log^2(\frac{2}{Q_j}) &\leq \sum_{j=1}^{k} P_j \log^2 \frac{2f(k)}{P_j} \\ &= \sum_{j=1}^{k} P_i(\log^2 2f(k) + \log^2 P_j - 2\log 2f(k) \log P_j) \\ &\leq (\log k + \log 2f(k))^2.\end{aligned} \tag{A.19}$$

3. Since $\sup_{x>0} x \exp(-nx/8) = \frac{8}{ne}$, we have

$$\sum_{i=1}^{k} P_i \log^2(2m) \exp\big(-\frac{(m-2)P_j}{8}\big) \leq \sum_{i=1}^{k} \frac{8\log^2(2m)}{(m-2)e} \leq \frac{8k\log^2(2m)}{(m-2)e}. \tag{A.20}$$

4. Since $Q_j \geq \frac{P_j}{f(k)}$, and $\sup_{x>0} x \exp(-nx/8) = \frac{8}{ne}$, we have

$$\sum_{i=1}^{k} P_i \log^2 n \exp\big(-\frac{nQ_j}{8}\big) \leq \sum_{i=1}^{k} \log^2 n P_i \exp\big(-\frac{nP_j}{8f(k)}\big) \leq \frac{8kf(k)\log^2 n}{ne}. \tag{A.21}$$

Thus,

$$\begin{aligned}&\mathbb{E}\left[(\hat{D}_{\mathrm{A-plug-in}}(M,N) - \hat{D}_{\mathrm{A-plug-in}}(M',N))^2\right] \\ \lesssim & \frac{(\log f(k) + \log k)^2}{m^2} + \frac{k\log^2 m}{m^3} + \frac{kf(k)\log^2 n}{m^2 n} \\ \lesssim & \frac{(\log f(k) + \log k)^2}{m^2}\big(1 + \frac{k\log^2 m}{m\log^2 k} + \frac{kf(k)\log^2 n}{\log^2(kf(k))n}\big),\end{aligned} \tag{A.22}$$

where the second term applies $kf(k) \geq k$. Note that the assumption $m \gtrsim k$ and $n \gtrsim kf(k)$ implies that $\frac{k\log^2 m}{m\log^2 k} \lesssim 1$ and $\frac{kf(k)\log^2 n}{\log^2(kf(k))n} \lesssim 1$, because $\frac{x}{\log x}$ is an increasing function. We then obtain

$$\mathbb{E}\left[(\hat{D}_{\mathrm{A-plug-in}}(M,N) - \hat{D}_{\mathrm{A-plug-in}}(M',N))^2\right] \lesssim \frac{\log^2 f(k) + \log^2 k}{m^2}. \tag{A.23}$$

The second term in (A.9) can be bounded similarly as follows:

$$\begin{aligned}
&\mathbb{E}\left[(\hat{D}_{\mathrm{A-plug-in}}(M,N)-\hat{D}_{\mathrm{A-plug-in}}(M,N'))^2\right]\\
=&\mathbb{E}\Bigg[\mathbb{E}\Bigg[\Big(\phi(\frac{M_{Y_m}}{m},\frac{\tilde{N}_{Y_m}+2}{n})+\phi(\frac{M_{Y'_m}}{m},\frac{\tilde{N}_{Y'_m}+1}{n})\\
&\qquad-\phi(\frac{M_{Y_m}}{m},\frac{\tilde{N}_{Y_m}+1}{n})-\phi(\frac{M_{Y'_m}}{m},\frac{\tilde{N}_{Y'_m}+2}{n})\Big)^2\Bigg|Y_m,Y'_m\Bigg]\Bigg]\\
\leq&4\mathbb{E}\left[\mathbb{E}\left[\left(\phi(\frac{M_{Y_m}}{m},\frac{\tilde{N}_{Y_m}+2}{n})-\phi(\frac{M_{Y_m}}{m},\frac{\tilde{N}_{Y_m}+1}{n})\right)^2\Bigg|Y_m\right]\right]\\
=&4\sum_{j=1}^{k}\mathbb{E}\left[\left(\phi(\frac{M_j}{m},\frac{\tilde{N}_j+2}{n})-\phi(\frac{M_j}{m},\frac{\tilde{N}_j+1}{n})\right)^2\right]Q_j\\
=&4\sum_{j=1}^{k}\mathbb{E}\left[\left(\frac{M_j}{m}\log\left(1+\frac{1}{\tilde{N}_j+1}\right)\right)^2\right]Q_j\\
=&\frac{4}{m^2}\sum_{j=1}^{k}\mathbb{E}\left[M_j^2\right]\mathbb{E}\left[\log^2\left(1+\frac{1}{\tilde{N}_j+1}\right)\right]Q_j. \qquad (A.24)
\end{aligned}$$

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

$$\mathbb{E}\left[M_j^2\right]=\mathbb{E}[M_j]^2+\mathrm{Var}(M_j)=m^2P_j^2+mP_j(1-P_j). \qquad (A.25)$$

We can also derive

$$\begin{aligned}
\mathbb{E}\left[\log^2\left(1+\frac{1}{\tilde{N}_j+1}\right)\right]&\leq\mathbb{E}\left[\Big(\frac{1}{\tilde{N}_j+1}\Big)^2\right]\\
&\leq\mathbb{E}\left[\frac{2}{(\tilde{N}_j+1)(\tilde{N}_j+2)}\right]\\
&\leq\frac{2}{(n-1)^2Q_j^2}, \qquad (A.26)
\end{aligned}$$

where the last inequality follows similarly from (A.5). Thus,

$$\begin{aligned}
&\mathbb{E}\left[(\hat{D}_{\mathrm{A-plug-in}}(M,N)-\hat{D}_{\mathrm{A-plug-in}}(M,N'))^2\right]\\
=&\frac{4}{m^2}\sum_{j=1}^{k}\mathbb{E}\left[M_j^2\right]\mathbb{E}\left[\log^2\left(1+\frac{1}{\tilde{N}_j+1}\right)\right]Q_j\\
\leq&4\sum_{j=1}^{k}\left(P_j^2+\frac{P_j(1-P_j)}{m}\right)\frac{2}{(n-1)^2Q_j^2}Q_j\\
\lesssim&\sum_{j=1}^{k}\frac{P_j}{Q_j}\left(P_j+\frac{1}{m}\right)\frac{2}{n^2}\\
\lesssim&\frac{f(k)}{n^2}+\frac{kf(k)}{n^2m}.
\end{aligned} \tag{A.27}$$

Combing (A.23) and (A.27), we obtain the following upper bound on the variance:

$$\begin{aligned}
&\mathrm{Var}[\hat{D}_{\mathrm{A-plug-in}}(M,N)]\\
\leq&\frac{m}{2}\mathbb{E}\left[(\hat{D}_{\mathrm{A-plug-in}}(M,N)-\hat{D}_{\mathrm{A-plug-in}}(M',N))^2\right]\\
&+\frac{n}{2}\mathbb{E}\left[(\hat{D}_{\mathrm{A-plug-in}}(M,N)-\hat{D}_{\mathrm{A-plug-in}}(M,N'))^2\right]\\
\lesssim&\frac{\log^2 k}{m}+\frac{\log^2 f(k)}{m}+\frac{f(k)}{n}+\frac{kf(k)}{nm}.
\end{aligned} \tag{A.28}$$

Note that the term $\frac{kf(k)}{nm}$ in the variance can be further upper bounded as follows:

$$\frac{kf(k)}{nm}\leq\frac{kf(k)}{n}\frac{k}{m}\leq\left(\frac{kf(k)}{n}+\frac{k}{m}\right)^2. \tag{A.29}$$

Combining (A.8), (A.28) and (A.29), we obtain the following upper bound on the worse case quadratic risk for augmented plug-in estimator:

$$R(\hat{D}_{\mathrm{A-plug-in}},k,m,n,f(k))\lesssim\left(\frac{kf(k)}{n}+\frac{k}{m}\right)^2+\frac{\log^2 k}{m}+\frac{\log^2 f(k)}{m}+\frac{f(k)}{n}. \tag{A.30}$$

APPENDIX B

PROOF OF PROPOSITION 2

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

B.1 $m \gg k$ and $n \gg kf(k)$ Are Necessary

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

$$\mathbb{E}\left[\left(\hat{D}_{\text{A}-\text{plug}-\text{in}}(M,N) - D(P\|Q)\right)^2\right] \geq \left(\mathbb{E}\left[\hat{D}_{\text{A}-\text{plug}-\text{in}}(M,N) - D(P\|Q)\right]\right)^2. \tag{B.1}$$

Following steps in (A.1), we have:

$$\begin{aligned}\mathbb{E}[\hat{D}_{\text{A}-\text{plug}-\text{in}}(M,N) - D(P\|Q)] =& \mathbb{E}\left(\sum_{i=1}^{k}\left(\frac{M_i}{m}\log\frac{M_i}{m} - P_i\log P_i\right)\right) \\ &+ \mathbb{E}\left(\sum_{i=1}^{k} P_i \log\frac{nQ_i}{N_i+1}\right). \end{aligned} \tag{B.2}$$

The first term in (B.2) is the bias of the plug-in entropy estimator. As shown in [13] and [21], the worst case quadratic risk of the first term can be bounded as follows:

$$\mathbb{E}\left(\sum_{i=1}^{k}\left(\frac{M_i}{m}\log\frac{M_i}{m}-P_i\log P_i\right)\right)\geq(\frac{k}{m}\wedge 1),\quad \text{if } P \text{ is uniform distribution,} \tag{B.3a}$$

$$\mathbb{E}\left(\sum_{i=1}^{k}\left(\frac{M_i}{m}\log\frac{M_i}{m}-P_i\log P_i\right)\right)\leq\log\left(1+\frac{k-1}{m}\right),\quad \text{for any } P. \tag{B.3b}$$

As shown in (A.3), we have the following bound on the second term in (B.2):

$$-\frac{kf(k)}{n}\leq\mathbb{E}\left(\sum_{i=1}^{k}P_i\log\frac{nQ_i}{N_i+1}\right),\quad \text{for any } (P,Q). \tag{B.4}$$

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

$$P=\left(\frac{1}{k},\ \frac{1}{k},\ \cdots,\ \frac{1}{k}\right),\quad Q=\left(\frac{1}{kf(k)},\ \cdots,\ \frac{1}{kf(k)},\ 1-\frac{k-1}{kf(k)}\right). \tag{B.5}$$

It can be verified that P and Q satisfy the density ratio constraint. For this a (P,Q) pair, we have

$$\begin{aligned}\mathbb{E}\left(\sum_{i=1}^{k}P_i\log\frac{nQ_i}{N_i+1}\right)&\leq\sum_{i=1}^{k}P_i\log\mathbb{E}\left[\frac{nQ_i}{N_i+1}\right]\\&=\sum_{i=1}^{k}P_i\log\left(\frac{nQ_i}{(n+1)Q_i}(1-(1-Q_i)^{n+1})\right)\\&\leq\sum_{i=1}^{k}P_i\log(1-(1-Q_i)^{n+1})\\&\leq\frac{k-1}{k}\log(1-(1-\frac{1}{kf(k)})^{n+1})\\&\leq-\frac{k-1}{k}(1-\frac{1}{kf(k)})^{n+1}\\&=-\frac{k-1}{k}(1-\frac{1}{kf(k)})^{kf(k)(n+1)\frac{1}{kf(k)}}.\end{aligned} \tag{B.6}$$

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, $1/(kf(k))$ is close to 0 , and thus we have

$$e^{-1} > (1 - \frac{1}{kf(k)})^{kf(k)} > e^{-\beta_0}, \tag{B.7}$$

where $\beta_0 > 1$ is a constant. Thus,

$$\mathbb{E}\left(\sum_{i=1}^{k} P_i \log \frac{nQ_i}{N_i+1}\right) \leq -\frac{k-1}{k} \exp(-\frac{\beta_0 n}{kf(k)}), \quad \text{for } (P,Q) \text{ in (B.5)}. \tag{B.8}$$

Combining (B.3a) and (A.3), we have

$$(\frac{k}{m} \wedge 1) - \frac{kf(k)}{n} \leq \sup_{(P,Q)\in\mathcal{M}_{k,f(k)}} \mathbb{E}[\hat{D}_{\text{A}-\text{plug}-\text{in}}(M,N) - D(P\|Q)], \tag{B.9}$$

and combining (B.3b) and (B.8), we obtain

$$\mathbb{E}[\hat{D}_{\text{A}-\text{plug}-\text{in}}(M,N) - D(P\|Q)] \leq \log\left(1+\frac{k}{m}\right) - \frac{k-1}{k}\exp(-\frac{\beta_0 n}{kf(k)}), \tag{B.10}$$

for (P,Q) in (B.5).

1) If $m \lesssim k$ and $n \gg kf(k)$, let $m \leq C_1 k$, where C_1 is a positive constant. Then (B.9) suggests

$$\sup_{(P,Q)\in\mathcal{M}_{k,f(k)}} \mathbb{E}[\hat{D}_{\text{A}-\text{plug}-\text{in}}(M,N) - D(P\|Q)] \geq (\frac{k}{m}\wedge 1) - \frac{kf(k)}{n} \to (\frac{1}{C_1}\wedge 1). \tag{B.11}$$

The bias is lower bounded by a positive constant, and hence, for sufficiently large k, the augmented plug-in estimator is not consistent.

2) If $m \gg k$ and $n \lesssim kf(k)$, let $n \leq C_2 kf(k)$, where C_2 is a positive constant. Then (B.10) suggests

$$\begin{aligned}\mathbb{E}[\hat{D}_{\text{A}-\text{plug}-\text{in}}(M,N) - D(P\|Q)] &\leq \log\left(1+\frac{k}{m}\right) - \frac{k-1}{k}\exp(-\frac{\beta_0 n}{kf(k)}) \\ &\to -\frac{k-1}{k}\exp(-\frac{\beta_0 n}{kf(k)}) \\ &\leq -\frac{k-1}{k}e^{-\beta_0 C_2}. \end{aligned} \tag{B.12}$$

The bias is upper bounded by a negative constant, and hence, for sufficiently large k, the augmented plug-in estimator is not consistent.

3) If $m \lesssim k$ and $n \lesssim kf(k)$, we cannot get a useful lower bound on the squared bias from (B.9) and (B.10) using the chosen pair (P, Q). Hence, we need to choose other pairs (P, Q).

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

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

$$P = \left(\frac{1}{k}, \frac{1}{k}, \cdots, \frac{1}{k}\right), \quad Q = \left(\frac{1}{\alpha k f(k)}, \ \cdots, \ \frac{1}{\alpha k f(k)}, \ 1 - \frac{k-1}{\alpha k f(k)}\right). \tag{B.13}$$

Let $\alpha > \frac{1}{f(k)}$, and then $1 - \frac{k-1}{\alpha k f(k)} > \frac{1}{k}$. 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, the bias of entropy estimation can be written as

$$\mathbb{E}\left(\sum_{i=1}^k \left(\frac{M_i}{m} \log \frac{M_i}{m} - P_i \log P_i\right)\right) = \log k + \mathbb{E}\left(\sum_{i=1}^k \frac{M_i}{m} \log \frac{M_i}{m}\right). \tag{B.14}$$

It can be shown that

$$\sum_{i=1}^k \frac{M_i}{m} \log \frac{M_i}{m} \geq -\log m. \tag{B.15}$$

Combining with (B.3a) and (B.3b), we have

$$(\frac{k}{m} \wedge 1) \vee \log \frac{k}{m} \le \mathbb{E}\left(\sum_{i=1}^{k}\left(\frac{M_i}{m}\log\frac{M_i}{m} - P_i \log P_i\right)\right) \le \log\left(1+\frac{k}{m}\right). \tag{B.16}$$

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

$$\begin{aligned}
&\mathbb{E}\left(\sum_{i=1}^{k}\left(P_i \log Q_i - \frac{M_i}{m}\log\frac{N_i+1}{n}\right)\right)\\
=&-\sum_{i=1}^{k} P_i \mathbb{E}\left[\log\frac{N_i+1}{nQ_i}\right]\\
\ge&-\sum_{i=1}^{k} P_i \log \mathbb{E}\left[\frac{N_i+1}{nQ_i}\right]\\
=&-\sum_{i=1}^{k} P_i \log\left(1+\frac{1}{nQ_i}\right)\\
\ge&-\frac{k-1}{k}\log\left(1+\frac{\alpha k f(k)}{n}\right) - \frac{1}{k}\log\left(1+\frac{k}{n}\right)\\
\ge&-\log\left(1+\frac{\alpha k f(k)}{n}\right) - \frac{\log 2k}{k}.
\end{aligned} \tag{B.17}$$

Combining with (B.16), we obtain the following lower bound:

$$\begin{aligned}
&\mathbb{E}[\hat{D}_{\mathrm{A-plug-in}}(M,N) - D(P\|Q)]\\
\ge&(\frac{k}{m}\wedge 1)\vee\log\frac{k}{m} - \log\left(1+\frac{\alpha k f(k)}{n}\right) - \frac{\log 2k}{k}\\
\ge&(\frac{k}{m}\wedge 1)\vee\log\frac{k}{m} - \log\left(1+\frac{k}{m(1+\epsilon)}\right) - \frac{\log 2k}{k}.
\end{aligned} \tag{B.18}$$

Note that $m \lesssim k$. Let $m \le C_1 k$, where C_1 is a positive constant. Without loss of generality, we can assume that $C_1 > 1$, since the case $C_1 \le 1$ is included in the following discussion.

Denote $x = \frac{k}{m}$, $x \in [\frac{1}{C_1}, \infty]$. If $x \in [\frac{1}{C_1}, 1]$, then $(x \wedge 1) \vee \log x = x$, and it can be shown that

$$x - \log\left(1+\frac{x}{1+\epsilon}\right) \ge x - \frac{x}{1+\epsilon} \ge \frac{\epsilon}{C_1(1+\epsilon)}. \tag{B.19}$$

If $x \in (1, e]$, then $(x \wedge 1) \vee \log x = 1$, and it can be shown that

$$1 - \log\left(1 + \frac{x}{1+\epsilon}\right) \geq 1 - \log\left(1 + \frac{e}{1+\epsilon}\right). \tag{B.20}$$

If $x \in (e, \infty)$, then $(x \wedge 1) \vee \log x = \log x$, and it can be shown that

$$\log x - \log\left(1 + \frac{x}{1+\epsilon}\right) = \log\left(\frac{1}{\frac{1}{x} + \frac{1}{1+\epsilon}}\right) \geq 1 - \log\left(1 + \frac{e}{1+\epsilon}\right). \tag{B.21}$$

Combining (B.19), (B.20) and (B.21), we obtain

$$(x \wedge 1) \vee \log x - \log\left(1 + \frac{x}{1+\epsilon}\right) \geq \min\left(\frac{\epsilon}{C_1(1+\epsilon)}, 1 - \log\left(1 + \frac{e}{1+\epsilon}\right)\right). \tag{B.22}$$

If $\epsilon > \frac{1}{e-1}$, the right-hand side in the above inequality is positive. And for sufficiently large k, $\frac{\log(2k)}{k}$ is arbitrarily small. 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.

Case II: If $\frac{k}{m} < (1+\epsilon)\frac{\alpha k f(k)}{n}$, which implies that the number of samples drawn from P is larger than the number of samples drawn from Q, then the second bias term dominates.

Since $n \lesssim kf(k)$, assume that $n \leq C_2 k f(k)$, where C_2 is a positive constant. Without loss of generality, we assume that $C_2 > 1$.

For $n \leq kf(k)$, we set:

$$P = \left(\frac{1}{k}, \frac{1}{k}, \cdots, \frac{1}{k}\right), \quad Q = \left(\frac{1}{kf(k)}, \ \cdots, \ \frac{1}{kf(k)}, \ 1 - \frac{k-1}{kf(k)}\right). \tag{B.23}$$

Following the steps in (B.6) and (B.7), we have

$$\begin{aligned}
&\mathbb{E}[\hat{D}_{\mathrm{A-plug-in}}(M, N) - D(P\|Q)] && \text{(B.24)}\\
\leq &\log\left(1 + \frac{k}{m}\right) + \frac{k-1}{k}\log(1 - \exp(-\frac{\beta_0 n}{kf(k)}))\\
= &\frac{k-1}{k}\left(\log\left(1 + \frac{k}{m}\right) + \log(1 - \exp(-\frac{\beta_0 n}{kf(k)}))\right) + \frac{1}{k}\log\left(1 + \frac{k}{m}\right)\\
\leq &\frac{k-1}{k}\log\left(\left(1 + (1+\epsilon)\frac{\alpha k f(k)}{n}\right)\left(1 - \exp(-\frac{\beta_0 n}{kf(k)})\right)\right) + \frac{\log(2k)}{k}. && \text{(B.25)}
\end{aligned}$$

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

$$h(t) \triangleq (1+\frac{\beta}{t})(1-\exp(-\beta_0 t)), \quad t \in (0,1). \tag{B.26}$$

For sufficiently large k, we choose $\beta_0 = 1.05$. Then for any $\beta < 0.3$, we have

$$h(t) = (1+\frac{\beta}{t})(1-\exp(-1.05t)) < 0.9, \quad \forall t \in (0,1). \tag{B.27}$$

Thus, if $f(k)$ is large, e.g., $f(k) > 6$, then we can find α that satisfies the condition $\alpha > \frac{1}{f(k)}$ and $(1+\epsilon)\alpha < 0.3$, with $\epsilon > \frac{1}{e-1}$. Then, for sufficiently large k,

$$\begin{aligned}
&\mathbb{E}[\hat{D}_{\mathrm{A-plug-in}}(M,N) - D(P\|Q)] \\
\leq &\frac{k-1}{k}\log\left(\big(1+(1+\epsilon)\frac{\alpha k f(k)}{n}\big)\big(1-\exp(-\frac{\beta_0 n}{kf(k)})\big)\right) + \frac{\log(2k)}{k} \\
\rightarrow &\log\left(\big(1+(1+\epsilon)\frac{\alpha k f(k)}{n}\big)\big(1-\exp(-\frac{1.05n}{kf(k)})\big)\right) \\
\leq &\log 0.9 < 0.
\end{aligned} \tag{B.28}$$

For $kf(k) < n \leq C_2 k f(k)$, we set:

$$P = \left(\frac{f(k)}{n}, \dots, \frac{f(k)}{n}, 1 - \frac{(k-1)f(k)}{n}\right), \quad Q = \left(\frac{1}{n}, \dots, \frac{1}{n}, 1-\frac{k-1}{n}\right). \tag{B.29}$$

It can be shown that (P,Q) satisfy the density ratio constraint. Following the steps for deriving (B.6), we have

$$\begin{aligned}
&\mathbb{E}[\hat{D}_{\mathrm{A-plug-in}}(M,N) - D(P\|Q)] \\
\leq &\log\left(1+\frac{k}{m}\right) + \frac{(k-1)f(k)}{n}\log(1-(1-\frac{1}{n})^{n+1}) \\
\overset{(a)}{\leq} &\frac{k}{m} - 0.3\frac{(k-1)f(k)}{n} \\
\leq &(1+\epsilon)\frac{\alpha k f(k)}{n} - 0.3\frac{kf(k)}{n} + \frac{0.3f(k)}{n} \\
\overset{(b)}{\leq} &((1+\epsilon)\alpha - 0.3)\frac{1}{C_2} + \frac{0.3}{k} \\
\overset{(c)}{<} &0,
\end{aligned} \tag{B.30}$$

where (a) is due to the fact that $\log(1-(1-\frac{1}{n})^{n+1}) \leq -0.3$ for $n \geq 5$; (b) holds if $(1+\epsilon)\alpha - 0.3 < 0$; and (c) holds for sufficiently large k.

This implies that for large k,

$$\mathbb{E}[\hat{D}_{\mathrm{A-plug-in}}(M,N) - D(P\|Q)] < c < 0, \tag{B.31}$$

where c is a negative constant. Hence, the worst case quadratic risk is lower bounded by a positive constant, and the augmented plug-in estimator is not consistent in this case.

B.2 $m \gg \log^2 f(k)$ Is Necessary

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

$$P^{(1)} = \left(\frac{1}{3(k-1)}, \ldots, \frac{1}{3(k-1)}, \frac{2}{3}\right), \tag{B.32}$$

$$P^{(2)} = \left(\frac{1-\epsilon}{3(k-1)}, \ldots, \frac{1-\epsilon}{3(k-1)}, 1-\frac{1-\epsilon}{3}\right), \tag{B.33}$$

$$Q^{(1)} = Q^{(2)} = \left(\frac{1}{3(k-1)f(k)}, \ldots, \frac{1}{3(k-1)f(k)}, 1-\frac{1}{3f(k)}\right). \tag{B.34}$$

The above distributions satisfy:

$$D(P^{(1)}\|Q^{(1)}) = \frac{1}{3}\log f(k) + \frac{2}{3}\log\frac{\frac{2}{3}}{1-\frac{1}{3f(k)}}, \tag{B.35}$$

$$D(P^{(2)}\|Q^{(2)}) = \frac{1-\epsilon}{3}\log(1-\epsilon)f(k) + \left(1-\frac{1-\epsilon}{3}\right)\log\frac{1-\frac{1-\epsilon}{3}}{1-\frac{1}{3f(k)}}, \tag{B.36}$$

$$D(P^{(1)}\|P^{(2)}) = \frac{1}{3}\log\frac{1}{1-\epsilon} + \frac{2}{3}\log\frac{\frac{2}{3}}{1-\frac{1-\epsilon}{3}}. \tag{B.37}$$

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

$$\begin{aligned} D(P^{(1)}\|P^{(2)}) &= \frac{1}{3}\log\left(1+\frac{\epsilon}{1-\epsilon}\right) + \frac{2}{3}\log\left(1-\frac{\epsilon}{2+\epsilon}\right) \\ &\le \frac{\epsilon}{3(1-\epsilon)} - \frac{2}{3}\frac{\epsilon}{2+\epsilon} \\ &= \frac{\epsilon^2}{(1-\epsilon)(2+\epsilon)} \le \frac{1}{m}. \end{aligned} \tag{B.38}$$

Furthermore,

$$\begin{aligned} & D(P^{(1)}\|Q^{(1)}) - D(P^{(2)}\|Q^{(2)}) \\ =& \frac{1}{3}\log f(k) + \frac{2}{3}\log\frac{\frac{2}{3}}{1-\frac{1}{3f(k)}} - \frac{1-\epsilon}{3}\log(1-\epsilon)f(k) - \left(1-\frac{1-\epsilon}{3}\right)\log\frac{1-\frac{1-\epsilon}{3}}{1-\frac{1}{3f(k)}} \\ =& \frac{1}{3}\log\frac{1}{1-\epsilon} + \frac{\epsilon}{3}\log(1-\epsilon)f(k) + \frac{2}{3}\log\frac{2}{2+\epsilon} - \frac{\epsilon}{3}\log\frac{2+\epsilon}{3-\frac{1}{f(k)}} \\ =& \frac{1}{3}\log\frac{1}{1-\epsilon}\frac{4}{(2+\epsilon)^2} - \frac{\epsilon}{3}\log\frac{2+\epsilon}{(1-\epsilon)(3f(k)-1)}, \end{aligned} \tag{B.39}$$

which implies that

$$(D(P^{(1)}\|Q^{(1)}) - D(P^{(2)}\|Q^{(2)}))^2 \gtrsim \epsilon^2 \log^2\frac{2}{(3f(k)-1)} \asymp \frac{\log^2 f(k)}{m}, \tag{B.40}$$

as $m \to \infty$. Now applying Le Cam's two-point method, we obtain

$$\begin{aligned} R^*(k,m,n,f(k)) \ge \frac{1}{16}&(D(P^{(1)}\|Q^{(1)}) - D(P^{(2)}\|Q^{(2)}))^2 \\ &\exp\left(-mD(P^{(1)}\|P^{(2)}) - nD(Q^{(1)}\|Q^{(2)})\right). \end{aligned} \tag{B.41}$$

Clearly, if $m \lesssim \log^2 f(k)$, 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.

APPENDIX C

PROOF OF LEMMA 1

We prove the inequality (4.2) that connects the minimax risk (1.8) under the deterministic sample size to the risk (4.1) under the Poisson sampling model. We first prove the left-hand side of (4.2). Recall that $0 \leq R^*(k, m, n, f(k)) \leq \log^2 f(k)$ and $R^*(k, m, n, f(k))$ is decreasing with m, n. Therefore,

$$\begin{aligned}
&\widetilde{R}^*(k, 2m, 2n, f(k)) \\
=&\sum_{i\geq 0}\sum_{j\geq 0} R^*(k, i, j, f(k))\mathrm{Poi}(2m, i)\mathrm{Poi}(2n, i) \\
=&\sum_{i\geq m+1}\sum_{j\geq n+1} R^*(k, i, j, f(k))\mathrm{Poi}(2m, i)\mathrm{Poi}(2n, i) \\
&+\sum_{i\geq 0}\sum_{j=0}^{n} R^*(k, i, j, f(k))\mathrm{Poi}(2m, i)\mathrm{Poi}(2n, i) \\
&+\sum_{i=0}^{m}\sum_{j\geq n+1} R^*(k, i, j, f(k))\mathrm{Poi}(2m, i)\mathrm{Poi}(2n, i) \\
\leq& R^*(k, m, n, f(k)) + e^{-(1-\log 2)n}\log^2 f(k) + e^{-(1-\log 2)m}\log^2 f(k), \qquad \text{(C.1)}
\end{aligned}$$

where the last inequality follows from the Chernoff bound $\mathbb{P}[\mathrm{Poi(2n)} \leq n] \leq \exp(-(1-\log 2)n)$. We then prove the right-hand side of (4.2). By the minimax theorem,

$$R^*(k, m, n, f(k)) = \sup_{\pi}\inf_{\hat{D}} \mathbb{E}[(\hat{D}(M, N) - D(P\|Q))^2], \qquad \text{(C.2)}$$

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

Fix a prior π and an arbitrary sequence of estimators $\{\hat{D}_{m,n}\}$ indexed by the sample sizes m and n. It is unclear whether the sequence of Bayesian risks $\alpha_{m,n} = \mathbb{E}[(\hat{D}_{m,n}(M, N) - D(P\|Q))^2]$ with respect to π is decreasing in

m or n. However, we can define $\{\widetilde{\alpha}_{i,j}\}$ as

$$\widetilde{\alpha}_{0,0} = \alpha_{0,0}, \quad \widetilde{\alpha}_{i,j} = \alpha_{i,j} \wedge \alpha_{i-1,j} \wedge \alpha_{i,j-1}. \tag{C.3}$$

Further, define

$$\widetilde{D}_{m,n}(M,N) \triangleq \begin{cases} \hat{D}_{m,n}(M,N), & \text{if } \widetilde{\alpha}_{m,n} = \alpha_{m,n}; \\ \hat{D}_{m-1,n}(M,N), & \text{if } \widetilde{\alpha}_{m,n} = \alpha_{m-1,n}; \\ \hat{D}_{m,n-1}(M,N), & \text{if } \widetilde{\alpha}_{m,n} = \alpha_{m,n-1}. \end{cases} \tag{C.4}$$

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

$$\begin{aligned} & \mathbb{E}\left[(\hat{D}_{m',n'}(M',N') - D(P\|Q))^2\right] \\ =& \sum_{i\geq 0}\sum_{j\geq 0} \mathbb{E}\left[(\hat{D}_{i,j}(M',N') - D(P\|Q))^2\right] \mathrm{Poi}(\frac{m}{2}, i)\mathrm{Poi}(\frac{n}{2}, j) \\ \geq& \sum_{i\geq 0}\sum_{j\geq 0} \mathbb{E}\left[(\widetilde{D}_{i,j}(M,N) - D(P\|Q))^2\right] \mathrm{Poi}(\frac{m}{2}, i)\mathrm{Poi}(\frac{n}{2}, j) \\ \geq& \sum_{i=0}^{m}\sum_{j=0}^{n} \mathbb{E}\left[(\widetilde{D}_{i,j}(M,N) - D(P\|Q))^2\right] \mathrm{Poi}(\frac{m}{2}, i)\mathrm{Poi}(\frac{n}{2}, j) \\ \overset{(a)}{\geq}& \frac{1}{4}\mathbb{E}\left[(\widetilde{D}_{m,n}(M,N) - D(P\|Q))^2\right], \end{aligned} \tag{C.5}$$

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

$$\widetilde{R}^*(k, \frac{m}{2}, \frac{n}{2}, f(k)) \geq \frac{1}{4}R^*(k, m, n, f(k)). \tag{C.6}$$

APPENDIX D

PROOF OF PROPOSITION 3

D.1 Bounds Using Le Cam's Two-Point Method

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

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

$$R^*(k,m,n,f(k)) \gtrsim (D(P^{(1)}\|Q^{(1)}) - D(P^{(2)}\|Q^{(2)}))^2 \gtrsim \frac{\log^2 f(k)}{m}. \tag{D.1}$$

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

We construct two pairs of distributions as follows:

$$P^{(1)} = P^{(2)} = \left(\frac{1}{3(k-1)}, 0, \frac{1}{3(k-1)}, 0, \ldots, \frac{5}{6}\right), \tag{D.2}$$

$$Q^{(1)} = \left(\frac{1}{2(k-1)f(k)}, \ldots, \frac{1}{2(k-1)f(k)}, 1 - \frac{1}{2f(k)}\right), \tag{D.3}$$

$$Q^{(2)} = \left(\frac{1-\epsilon}{2(k-1)f(k)}, \frac{1+\epsilon}{2(k-1)f(k)}, \ldots, \frac{1-\epsilon}{2(k-1)f(k)}, \frac{1+\epsilon}{2(k-1)f(k)}, 1 - \frac{1}{2f(k)}\right). \tag{D.4}$$

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:

$$D(Q^{(1)}\|Q^{(2)}) = \frac{1}{4f(k)} \log \frac{1}{1+\epsilon} + \frac{1}{4f(k)} \log \frac{1}{1-\epsilon}, \tag{D.5}$$

$$D(P^{(1)}\|Q^{(1)}) - D(P^{(2)}\|Q^{(2)}) = \frac{1}{6} \log(1-\epsilon) \leq -\frac{\epsilon}{6}. \tag{D.6}$$

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

$$D(Q^{(1)}\|Q^{(2)}) = \frac{1}{4f(k)}\log(1+\frac{\epsilon^2}{1-\epsilon^2}) \leq \frac{1}{4f(k)}\frac{\epsilon^2}{1-\epsilon^2} < \frac{\epsilon^2}{f(k)} = \frac{1}{n}. \quad \text{(D.7)}$$

We apply Le Cam's two-point method,

$$\begin{aligned} R^*(k,m,n,f(k)) \geq & \frac{1}{16}(D(P^{(1)}\|Q^{(1)}) - D(P^{(2)}\|Q^{(2)}))^2 \\ & \exp\left(-m(D(P^{(1)}\|P^{(2)}) - nD(Q^{(1)}\|Q^{(2)}))\right) \\ \gtrsim & (D(P^{(1)}\|Q^{(1)}) - D(P^{(2)}\|Q^{(2)}))^2 \\ \gtrsim & \epsilon^2 = \frac{f(k)}{n}. \end{aligned} \quad \text{(D.8)}$$

D.2 Bounds Using Generalized Le Cam's Method

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

Let $Q^{(0)}$ denote the uniform distribution. The minimax risk is lower bounded as follows:

$$\begin{aligned} R^*(k,m,n,f(k)) =& \inf_{\hat{D}} \sup_{(P,Q)\in\mathcal{M}_{k,f(k)}} \mathbb{E}[(\hat{D}(M,N) - D(P\|Q))^2] \\ \geq& \inf_{\hat{D}} \sup_{(P,Q^{(0)})\in\mathcal{M}_{k,f(k)}} \mathbb{E}[(\hat{D}(M,Q^{(0)}) - D(P\|Q^{(0)}))^2] \\ \triangleq& R^*(k,m,Q^{(0)},f(k)). \end{aligned} \quad \text{(D.9)}$$

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

$$\begin{aligned} D(P\|Q^{(0)}) =& \sum_{i=1}^{k}\left(P_i\log P_i + P_i\log\frac{1}{Q_i^{(0)}}\right) \\ =& H(P) + \log k. \end{aligned} \quad \text{(D.10)}$$

Hence, $R^*(k,m,Q^{(0)},f(k))$ 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 $Q^{(0)}$ is upper bounded by $f(k)$.

$$R^*(k,m,Q^{(0)},f(k)) = \inf_{\hat{H}} \sup_{P:P_i\le\frac{f(k)}{k}} \mathbb{E}[(\hat{H}(M)-H(P))^2]. \tag{D.11}$$

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

$$\inf_{\hat{H}} \sup_{P} \mathbb{E}[(\hat{H}(M)-H(P))^2] \gtrsim (\frac{k}{m\log k})^2. \tag{D.12}$$

The supremum is achieved for $P_i \le \frac{\log^2 k}{k}$. Comparing this result to (D.11), if $f(k) \ge \log^2 k$, then

$$\frac{\log^2 k}{k} \le \frac{f(k)}{k}. \tag{D.13}$$

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

$$R^*(k,m,n,f(k)) \gtrsim R^*(k,m,Q^{(0)},f(k)) \gtrsim (\frac{k}{m\log k})^2. \tag{D.14}$$

D.2.2 Proof of $R^*(k,m,n,f(k)) \gtrsim (\frac{kf(k)}{n\log k})^2$

Since $n \gtrsim \frac{kf(k)}{\log k}$, we assume that $n \ge \frac{C'kf(k)}{\log k}$. If $C' \ge 1$, we set $P = P^{(0)}$, where

$$P^{(0)} = \left(\frac{f(k)}{n\log k}, \ldots, \frac{f(k)}{n\log k}, 1 - \frac{(k-1)f(k)}{n\log k}\right). \tag{D.15}$$

Then, we have $0 \le 1 - \frac{(k-1)f(k)}{n\log k} \le 1$. Hence, $P^{(0)}$ is a well-defined probability distribution. If $C' < 1$, we set $P^{(0)}$ as follows:

$$P^{(0)} = \left(\frac{C'f(k)}{n\log k}, \ldots, \frac{C'f(k)}{n\log k}, 1 - \frac{C'(k-1)f(k)}{n\log k}\right), \tag{D.16}$$

which is also a well defined probability distribution. In the following, we focus on the case that $C' \ge 1$. And the results can be easily generalized to the case when $C' < 1$.

If $P = P^{(0)}$ given in (D.15) and is known, then estimating the KL divergence between P and Q is equivalent to estimating the following function:

$$D(P^{(0)}\|Q) = \sum_{i=1}^{k-1} \frac{f(k)}{n\log k} \log \frac{\frac{f(k)}{n\log k}}{Q_i} + (1 - \frac{(k-1)f(k)}{n\log k}) \log \frac{1 - \frac{(k-1)f(k)}{n\log k}}{Q_k}, \tag{D.17}$$

which is further equivalent to estimating

$$\sum_{i=1}^{k-1} \frac{f(k)}{n\log k} \log \frac{1}{Q_i} + (1 - \frac{(k-1)f(k)}{n\log k}) \log \frac{1}{Q_k}. \tag{D.18}$$

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

$$\mathcal{N}_{k,f(k)} \triangleq \{(P^{(0)}, Q) \in \mathcal{M}_{k,f(k)} : \frac{1}{n\log k} \leq Q_i \leq \frac{c_4 \log k}{n}, \text{ for } 1 \leq i \leq k-1\}, \tag{D.19}$$

where c_4 is a constant defined later.

The minimax risk can be lower bounded as follows:

$$\begin{aligned} R^*(k,m,n,f(k)) =& \inf_{\hat{D}} \sup_{(P,Q)\in\mathcal{M}_{k,f(k)}} \mathbb{E}[(\hat{D}(M,N) - D(P\|Q))^2] \\ \geq& \inf_{\hat{D}} \sup_{(P^{(0)},Q)\in\mathcal{N}_{k,f(k)}} \mathbb{E}[(\hat{D}(P^{(0)},N) - D(P^{(0)}\|Q))^2] \\ \triangleq& R^*_{\mathcal{N}}(k, P^{(0)}, n, f(k)). \end{aligned} \tag{D.20}$$

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

$$\begin{aligned} \mathcal{N}_{k,f(k)}(\epsilon) \triangleq \{(P^{(0)}, \mathsf{Q}) :& \mathsf{Q} \in \mathbb{R}_+^k, |\sum_{i=1}^{k} \mathsf{Q}_i - 1| \leq \epsilon, \\ & \frac{1}{n\log k} \leq \mathsf{Q}_i \leq \frac{c_4\log k}{n}, \text{ for } 1 \leq i \leq k-1\}. \end{aligned} \tag{D.21}$$

Note that Q is not a distribution. Furthermore, 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 (D.20) under Poisson sampling on the set $\mathcal{N}_{k,f(k)}(\epsilon)$ as follows:

$$\widetilde{R}^*_{\mathcal{N}}(k, P^{(0)}, n, f(k), \epsilon) = \inf_{\hat{D}} \sup_{(P^{(0)},\mathsf{Q})\in\mathcal{N}_{k,f(k)}(\epsilon)} \mathbb{E}[(\hat{D}(P^{(0)}, N) - D(P^{(0)}, \mathsf{Q}))^2], \tag{D.22}$$

where $N_i \sim \mathrm{Poi}(n\mathsf{Q}_i)$, for $1 \leq i \leq k$. The risk (D.22) is connected to the risk (D.20) for multinomial sampling by the following lemma.

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

$$R^*(k, P^{(0)}, \frac{n}{2}, f(k)) \geq \frac{1}{2}\widetilde{R}^*_{\mathcal{N}}(k, P^{(0)}, n, f(k), \epsilon) - \log^2 f(k) \exp{(-\frac{n}{50})} - \log^2(1+\epsilon). \tag{D.23}$$

Proof. See Appendix D.2.3. □

For $(P^{(0)}, \mathsf{Q}) \in \mathcal{N}_{k,f(k)}(\epsilon)$, we then apply the generalized Le Cam's method which involves two composite hypothesis as follows:

$$H_0 : D(P^{(0)}\|\mathsf{Q}) \leq t \quad \text{versus} \quad H_1 : D(P^{(0)}\|\mathsf{Q}) \geq t + \frac{(k-1)f(k)}{n\log k}d. \tag{D.24}$$

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

$$\mathsf{Q} = (V_1, \ldots V_{k-1}, 1-(k-1)\alpha) \text{ and } \mathsf{Q}' = (V_1', \ldots V_{k-1}', 1-(k-1)\alpha) \tag{D.25}$$

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 $(P^{(0)}, \mathsf{Q}), (P^{(0)}, \mathsf{Q}') \in \mathcal{N}_{k,f(k)}(\epsilon)$ satisfy the density ratio constraint. Then the averaged divergences are separated by the distance of

$$|\mathbb{E}[D(P^{(0)}\|\mathsf{Q})] - \mathbb{E}[D(P^{(0)}\|\mathsf{Q}')]| = \frac{(k-1)f(k)}{n\log k}|\mathbb{E}[\log V] - \mathbb{E}[\log V']|. \tag{D.26}$$

Thus, if we construct V and V' such that

$$|\mathbb{E}[\log V] - \mathbb{E}[\log V']| \geq d \tag{D.27}$$

then the constructions in (D.25) satisfy (D.24), serving as the two composite hypothesis which are separated.

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

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

$$\widetilde{R}^*_{\mathcal{N}}(k, P^{(0)}, n, f(k), \epsilon) \geq \frac{(\frac{(k-1)f(k)d}{n\log k})^2}{32}\left(1 - \frac{2(k-1)c_4^2\log^2 k}{n^2\epsilon^2} - \frac{32\log^2(n\log k)}{(k-1)d^2} - k\mathrm{TV}(\mathbb{E}[Poi(nV)], \mathbb{E}[Poi(nV')])\right), \tag{D.28}$$

where $\mathrm{TV}(P,Q) = \frac{1}{2}\sum_{i=1}^k |P_i - Q_i|$ *denotes the total variation between two distributions.*

Proof. See Appendix D.2.4.. □

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 $[\frac{1}{n\log k}, \frac{c_4\log k}{n}]$. If $\mathbb{E}[V^j] = \mathbb{E}[V'^j]$ for $j = 1, \ldots, L$, and $L > \frac{2c_4\log k}{n}$, then*

$$\mathrm{TV}\left(\mathbb{E}[Poi(nV)], \mathbb{E}[Poi(nV')]\right) \leq 2\exp\left(-\left(\frac{L}{2}\log\frac{L}{2ec_4\log k} - 2c_4\log k\right)\right) \wedge 1. \tag{D.29}$$

What remains is to construct V and V' to maximize $d = |\mathbb{E}[\log V'] - \mathbb{E}[\log V]|$, subject to the constraints in Lemma 4. Consider the following optimization problem over random variables X and X'.

$$\begin{aligned} \mathcal{E}^* = &\max \mathbb{E}[\log X] - \mathbb{E}[\log X'] \\ &\text{s.t. } \mathbb{E}[X^j] = \mathbb{E}[X'^j], \quad j = 1, \ldots, L \\ &X, X' \in [\frac{1}{c_4\log^2 k}, 1]. \end{aligned} \tag{D.30}$$

As shown in Appendix E in [13], the maximum $\mathcal{E}^*$ is equal to twice the error in approximating $\log x$ by a polynomial with degree L:

$$\mathcal{E}^* = 2E_L(\log, [\frac{1}{c_4 \log^2 k}, 1]). \tag{D.31}$$

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

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

$$E_{\lfloor cL \rfloor}(\log, [L^{-2}, 1]) > c'. \tag{D.32}$$

Let (X, X') be the maximizer of (D.30). We let $V = \frac{c_4 \log k}{n} X$ and $V' = \frac{c_4 \log k}{n} X'$, such that $V, V' \in [\frac{1}{n \log k}, \frac{c_4 \log k}{n}]$. Then it can be shown that

$$\mathbb{E}[\log V] - \mathbb{E}[\log V'] = \mathcal{E}^*, \tag{D.33}$$

where 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 5, $d = \mathcal{E}^* > 2c'$. We further assume that $\log^2 n \leq c_5 k$, set c_4 and c_5 such that $2c_4^2 + \frac{8c_5}{c'^2} < 1$ and $\frac{c}{2} \log \frac{c}{2ec_4} - 2c_4 > 2$. Then from Lemma 3 and Lemma 2, with $\epsilon = \frac{\sqrt{k} \log k}{n}$, the minimax risk is lower bounded as follows:

$$R^*(k, P^{(0)}, n, f(k)) \gtrsim (\frac{k f(k)}{n \log k})^2. \tag{D.34}$$

D.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

$$\sup_{(P^{(0)},Q) \in \mathcal{N}_{k,f(k)}(0)} \mathbb{E}[(\hat{D}(P^{(0)}, n) - D(P^{(0)} \| Q))^2] \leq \delta + R^*(k, P^{(0)}, n, f(k)). \tag{D.35}$$

For any $(P^{(0)}, \mathsf{Q}) \in \mathcal{N}_{k,f(k)}(\epsilon)$, Q is approximately a distribution. We normalize Q to be a probability distribution, i.e., $\frac{\mathsf{Q}}{\sum_{i=1}^k \mathsf{Q}_i}$, and then we have

$$D(P^{(0)}\|\mathsf{Q}) = \sum_{i=1}^{k} P_{0,i} \log \frac{P_{0,i}}{\mathsf{Q}_i} = -\log \sum_{i=1}^{k} \mathsf{Q}_i + D\Big(P^{(0)}\Big\| \frac{\mathsf{Q}}{\sum_{i=1}^k \mathsf{Q}_i}\Big). \quad \text{(D.36)}$$

Fix distributions $(P^{(0)}, \mathsf{Q}) \in \mathcal{N}_{k,f(k)}(\epsilon)$. Let $N = (N_1, \ldots, N_k)$, and $N_i \sim$Poi$(n\mathsf{Q}_i)$. And define $n' = \sum N_i \sim \text{Poi}(n \sum \mathsf{Q}_i)$. We set an estimator under the Poisson sampling by

$$\widetilde{D}(P^{(0)}, N) = \hat{D}(P^{(0)}, n'). \quad \text{(D.37)}$$

By the triangle inequality, we obtain

$$\begin{aligned}\frac{1}{2}\big(\widetilde{D}(P^{(0)}, N) - D(P^{(0)}\|\mathsf{Q})\big)^2 &\le \left(\widetilde{D}(P^{(0)}, N) - D\Big(P^{(0)}\Big\|\frac{\mathsf{Q}}{\sum_{i=1}^k \mathsf{Q}_i}\Big)\right)^2 \\ &\quad + \left(D\Big(P^{(0)}\Big\|\frac{\mathsf{Q}}{\sum_{i=1}^k \mathsf{Q}_i}\Big) - D(P^{(0)}\|\mathsf{Q})\right)^2 \\ &= \left(\widetilde{D}(P^{(0)}, N) - D\Big(P^{(0)}\Big\|\frac{\mathsf{Q}}{\sum_{i=1}^k \mathsf{Q}_i}\Big)\right)^2 + (\log \sum_{i=1}^{k} \mathsf{Q}_i)^2 \\ &\le \left(\widetilde{D}(P^{(0)}, N) - D\Big(P^{(0)}\Big\|\frac{\mathsf{Q}}{\sum_{i=1}^k \mathsf{Q}_i}\Big)\right)^2 + \log^2(1+\epsilon).\end{aligned} \quad \text{(D.38)}$$

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

$$\begin{aligned}&\mathbb{E}\left[\left(\widetilde{D}(P^{(0)}, N) - D\left(P^{(0)}\Big\|\frac{\mathsf{Q}}{\sum_{i=1}^k \mathsf{Q}_i}\right)\right)^2\right] \\ &= \sum_{j=1}^{\infty} \mathbb{E}\left[\left(\hat{D}(P^{(0)}, j) - D\Big(P^{(0)}\Big\|\frac{\mathsf{Q}}{\sum_{i=1}^k \mathsf{Q}_i}\Big)\right)^2 \middle| n' = j\right] \mathbb{P}(n' = j) \\ &\le \sum_{j=1}^{\infty} R^*(k, P^{(0)}, j, f(k)) \mathbb{P}(n' = j) + \delta.\end{aligned} \quad \text{(D.39)}$$

We note that for fixed k, $R^*(k, P^{(0)}, j, f(k))$ is a monotone decreasing function with respect to n. We also have $R^*(k, P^{(0)}, j, f(k)) \le \log^2 f(k)$, because for any $(P^{(0)}, \mathsf{Q}) \in \mathcal{N}_{k,f(k)}(0)$, $D(P^{(0)}\|\mathsf{Q}) \le \log f(k)$. Furthermore, since

$n' \sim \mathrm{Poi}(n\sum \mathsf{Q}_i)$, and $|\sum \mathsf{Q}_i - 1| \le \epsilon \le 1/3$, we have $P(n' > \frac{n}{2}) \le e^{-\frac{n}{50}}$. Hence, we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\left(\widetilde{D}(P^{(0)}, N) - D\Big(P^{(0)} \Big\| \frac{\mathsf{Q}}{\sum_{i=1}^k \mathsf{Q}_i}\Big)\right)^2\right] \\
\le &\sum_{j=1}^{\infty} R^*(k, P^{(0)}, j, f(k))\mathbb{P}(n' = j) + \delta \\
= &\sum_{j=1}^{n/2} R^*(k, P^{(0)}, j, f(k))\mathbb{P}(n' = j) + \sum_{j=\frac{n}{2}+1}^{\infty} R^*(k, P^{(0)}, j, f(k))\mathbb{P}(n' = j) + \delta \\
\le &R^*(k, P^{(0)}, \frac{n}{2}, f(k)) + (\log^2 f(k))P(n' > \frac{n}{2}) + \delta \\
\le &R^*(k, P^{(0)}, \frac{n}{2}, f(k)) + \log^2 f(k) e^{-\frac{n}{50}} + \delta. \qquad \text{(D.40)}
\end{aligned}
$$

Combining (D.38) and (D.40) completes the proof because δ can be arbitrarily small.

D.2.4 Proof of Lemma 3

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

$$
\begin{aligned}
P = P' = P^{(0)} &= \left(\frac{f(k)}{n \log k}, \ldots, \frac{f(k)}{n \log k}, 1 - \frac{(k-1)f(k)}{n \log k}\right), \qquad \text{(D.41)} \\
\mathsf{Q} &= (V_1, \ldots, V_{k-1}, 1 - (k-1)\alpha), \qquad \text{(D.42)} \\
\mathsf{Q}' &= \big(V_1', \ldots, V_{k-1}', 1 - (k-1)\alpha\big). \qquad \text{(D.43)}
\end{aligned}
$$

We further define the following events:

$$
E \triangleq \left\{\left|\sum_{i=1}^{k-1} V_i - (k-1)\alpha\right| \le \epsilon, |D(P\|\mathsf{Q}) - \mathbb{E}(D(P\|\mathsf{Q}))| \le \frac{d(k-1)f(k)}{4n \log k}\right\}, \qquad \text{(D.44)}
$$

$$
E' \triangleq \left\{\left|\sum_{i=1}^{k-1} V_i' - (k-1)\alpha\right| \le \epsilon, |D(P'\|\mathsf{Q}') - \mathbb{E}(D(P'\|\mathsf{Q}'))| \le \frac{d(k-1)f(k)}{4n \log k}\right\}. \qquad \text{(D.45)}
$$

By union bound and Chebyshev's inequality, we have

$$\begin{aligned} P(E^C) &\leq \frac{(k-1)\mathrm{Var}(V)}{\epsilon^2} + \frac{16(k-1)\mathrm{Var}(\frac{f(k)}{n\log k}\log V_i)}{(\frac{(k-1)f(k)}{n\log k}d)^2} \\ &\leq \frac{c_4^2(k-1)\log^2 k}{\epsilon^2 n^2} + \frac{16\log^2(n\log k)}{(k-1)d^2}. \end{aligned} \tag{D.46}$$

Similarly, we have

$$P(E'^C) \leq \frac{c_4^2(k-1)\log^2 k}{\epsilon^2 n^2} + \frac{16\log^2(n\log k)}{(k-1)d^2}. \tag{D.47}$$

Now, we define two priors on the set $\mathcal{N}_{k,f(k)}(\epsilon)$ by the following conditional distributions:$\pi = P_{V|E} \quad \text{and} \quad \pi' = P_{V'|E'}$.

Hence, given π and π' as prior distributions, recall the assumption $|\mathbb{E}[\log V] - \mathbb{E}[\log V']| \geq d$, and we have

$$|D(P\|\mathsf{Q}) - D(P'\|\mathsf{Q}')| \geq \frac{d(k-1)f(k)}{2n\log k}. \tag{D.48}$$

Now, we consider the total variation of observations under π and π'. The observations are Poisson distributed: $N_i \sim \mathrm{Poi}(n\mathsf{Q}_i)$ and $N_i' \sim \mathrm{Poi}(n\mathsf{Q}_i')$. By the triangle inequality, we have

$$\begin{aligned} \mathrm{TV}(P_{N|E}, P_{N'|E'}) &\leq \mathrm{TV}(P_{N|E}, P_N) + \mathrm{TV}(P_N, P_{N'}) + \mathrm{TV}(P_{N'}, P_{N'|E'}) \\ &= P(E^C) + P(E'^C) + \mathrm{TV}(P_N, P_{N'}) \\ &\leq \frac{2c_4^2(k-1)\log^2 k}{\epsilon^2 n^2} + \frac{32\log^2(n\log k)}{(k-1)d^2} + \mathrm{TV}(P_N, P_{N'}). \end{aligned} \tag{D.49}$$

From the fact that total variation of product distribution can be upper bounded by the summation of individual ones, we obtain

$$\begin{aligned} \mathrm{TV}(P_N, P_{N'}) &\leq \sum_{i=1}^{k-1} \mathrm{TV}(\mathbb{E}(\mathrm{Poi}(nV_i)), \mathbb{E}(\mathrm{Poi}(nV_i'))), \\ &= k\mathrm{TV}\left(\mathbb{E}(\mathrm{Poi}(nV)), \mathbb{E}(\mathrm{Poi}(nV'))\right). \end{aligned} \tag{D.50}$$

Applying the generalized Le Cam's method [16], and combining (D.49) and (D.50) completes the proof.

APPENDIX E

PROOF OF PROPOSITION 4

We first denote

$$D_1 \triangleq \sum_{i=1}^{k} P_i \log P_i, \quad D_2 \triangleq \sum_{i=1}^{k} P_i \log Q_i. \tag{E.1}$$

Hence, $D(P\|Q) = D_1 - D_2$. Recall that our estimator $\hat{D}_{\text{opt}}$ for $D(P\|Q)$ is:

$$\hat{D}_{\text{opt}} = \widetilde{D}_{\text{opt}} \vee 0 \wedge \log f(k), \tag{E.2}$$

where

$$\widetilde{D}_{\text{opt}} = \hat{D}_1 - \hat{D}_2, \tag{E.3}$$

$$\hat{D}_1 = \sum_{i=1}^{k} \left(g'_L(M_i)\mathbb{1}_{\{M'_i \le c'_2 \log k\}} + (\frac{M_i}{m} \log \frac{M_i}{m} - \frac{1}{2m})\mathbb{1}_{\{M'_i > c'_2 \log k\}} \right) \triangleq \sum_{i=1}^{k} \hat{D}_{1,i}, \tag{E.4}$$

$$\begin{aligned} \hat{D}_2 = \sum_{i=1}^{k} &\left(\frac{M_i}{m} g_L(N_i)\mathbb{1}_{\{N'_i \le c_2 \log k\}} + \frac{M_i}{m}\Big(\log \frac{N_i + 1}{n} - \frac{1}{2(N_i + 1)}\Big)\mathbb{1}_{\{N'_i > c_2 \log k\}} \right) \\ \triangleq &\sum_{i=1}^{k} \hat{D}_{2,i}. \end{aligned} \tag{E.5}$$

We define the following sets:

$$\begin{aligned} E_{1,i} &\triangleq \{N'_i \le c_2 \log k, Q_i \le \frac{c_1 \log k}{n}\}, \\ E_{2,i} &\triangleq \{N'_i > c_2 \log k, Q_i > \frac{c_3 \log k}{n}\}, \\ F_{1,i} &\triangleq \{N'_i \le c_2 \log k, Q_i > \frac{c_1 \log k}{n}\}, \\ F_{2,i} &\triangleq \{N'_i > c_2 \log k, Q_i \le \frac{c_3 \log k}{n}\}, \end{aligned} \tag{E.6}$$

and

$$\begin{aligned} E'_{1,i} &\triangleq \{M'_i \le c'_2 \log k, P_i \le \frac{c'_1 \log k}{m}\}, \\ E'_{2,i} &\triangleq \{M'_i > c'_2 \log k, P_i > \frac{c'_3 \log k}{m}\}, \\ F'_{1,i} &\triangleq \{M'_i \le c'_2 \log k, P_i > \frac{c'_1 \log k}{m}\}, \\ F'_{2,i} &\triangleq \{M'_i > c'_2 \log k, P_i \le \frac{c'_3 \log k}{m}\}, \end{aligned} \tag{E.7}$$

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

$$E_1 \triangleq \bigcap_{i=1}^{k} E_{1,i}, \quad E_2 \triangleq \bigcap_{i=1}^{k} E_{2,i} \tag{E.8}$$

$$E'_1 \triangleq \bigcap_{i=1}^{k} E'_{1,i}, \quad E'_2 \triangleq \bigcap_{i=1}^{k} E'_{2,i} \tag{E.9}$$

$$E \triangleq E_1 \cup E_2, \quad E' \triangleq E'_1 \cup E'_2, \tag{E.10}$$

$$\bar{E} \triangleq E \cap E' = \bigcap_{i=1}^{k} \Big((E_{1,i} \cup E_{2,i}) \cap (E'_{1,i} \cup E'_{2,i}) \Big). \tag{E.11}$$

It can be shown that

$$\bar{E}^c = \bigcup_{i=1}^{k} (E_{1,i} \cup E_{2,i})^c \cup (E'_{1,i} \cup E'_{2,i})^c = \bigcup_{i=1}^{k} (F_{1,i} \cup F_{2,i}) \cup (F'_{1,i} \cup F'_{2,i}). \tag{E.12}$$

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

$$\begin{aligned} \mathbb{P}(\bar{E}^c) &= \mathbb{P}\left(\bigcup_{i=1}^{k} (F_{1,i} \cup F_{2,i}) \cup (F'_{1,i} \cup F'_{2,i}) \right) \\ &\le k \left(\mathbb{P}(F_{1,i}) + \mathbb{P}(F_{2,i}) + \mathbb{P}(F'_{1,i}) + \mathbb{P}(F'_{2,i}) \right) \\ &\le \frac{1}{k^{c_1 - c_2 \log \frac{ec_1}{c_2} - 1}} + \frac{1}{k^{c_3 - c_2 \log \frac{ec_3}{c_2} - 1}} + \frac{1}{k^{c'_1 - c'_2 \log \frac{ec'_1}{c'_2} - 1}} + \frac{1}{k^{c'_3 - c'_2 \log \frac{ec'_3}{c'_2} - 1}}. \end{aligned} \tag{E.13}$$

We note that $\hat{D}_{\text{opt}}, D(P\|Q) \in [0, \log f(k)]$, and $\hat{D}_{\text{opt}} = \widetilde{D}_{\text{opt}} \vee 0 \wedge \log f(k)$.

Therefore, we have

$$\begin{aligned}
&\mathbb{E}[(\hat{D}_{\text{opt}} - D(P\|Q))^2] \\
=&\mathbb{E}[(\hat{D}_{\text{opt}} - D(P\|Q))^2 \mathbb{1}_{\{\bar{E}\}} + (\hat{D}_{\text{opt}} - D(P\|Q))^2 \mathbb{1}_{\{\bar{E}^c\}}] \\
\le&\mathbb{E}[(\widetilde{D}_{\text{opt}} - D(P\|Q))^2 \mathbb{1}_{\{\bar{E}\}}] + \log^2 f(k) P(\bar{E}^c) \\
=&\mathbb{E}[(\hat{D}_1 - \hat{D}_2 - D_1 + D_2)^2 \mathbb{1}_{\{\bar{E}\}}] + \log^2 f(k) P(\bar{E}^c). \qquad \text{(E.14)}
\end{aligned}$$

We choose constants $c_1, c_2, c_3, c_1', c_2', c_3'$ such that $c_1 - c_2 \log \frac{ec_1}{c_2} - 1 > C$, $c_1' - c_2' \log \frac{ec_1'}{c_2'} - 1 > C$, $c_3 - c_2 \log \frac{ec_3}{c_2} - 1 > C$, and $c_3' - c_2' \log \frac{ec_3'}{c_2'} - 1 > C$. Then together with $\log m \le C \log k$, we have

$$\log^2 f(k) P(\bar{E}^c) \le \frac{\log^2 f(k)}{m}. \qquad \text{(E.15)}$$

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

$$\begin{aligned}
I_1 &\triangleq \{i : N_i' \le c_2 \log k, Q_i \le \frac{c_1 \log k}{n}\}, \\
I_2 &\triangleq \{i : N_i' > c_2 \log k, Q_i > \frac{c_3 \log k}{n}\}, \\
I_1' &\triangleq \{i : M_i' \le c_2' \log k, P_i \le \frac{c_1' \log k}{m}\}, \\
I_2' &\triangleq \{i : M_i' > c_2' \log k, P_i > \frac{c_3' \log k}{m}\}. \qquad \text{(E.16)}
\end{aligned}$$

Using these index set, define

$$\begin{aligned}
A &\triangleq \sum_{i \in I_1 \cap I_1'} \big(\hat{D}_{1,i} - \hat{D}_{2,i} - P_i \log P_i + P_i \log Q_i\big), \\
B &\triangleq \sum_{i \in I_2 \cap I_1'} \big(\hat{D}_{1,i} - \hat{D}_{2,i} - P_i \log P_i + P_i \log Q_i\big), \\
C &\triangleq \sum_{i \in I_1 \cap I_2'} \big(\hat{D}_{1,i} - \hat{D}_{2,i} - P_i \log P_i + P_i \log Q_i\big), \\
D &\triangleq \sum_{i \in I_2 \cap I_2'} \big(\hat{D}_{1,i} - \hat{D}_{2,i} - P_i \log P_i + P_i \log Q_i\big).
\end{aligned}$$

We can further decompose $\mathbb{E}[(\hat{D}_1 - \hat{D}_2 - D_1 + D_2)^2 \mathbb{1}_{\{\bar{E}\}}]$ as follows:

$$\begin{aligned}
&\mathbb{E}[(\hat{D}_1 - \hat{D}_2 - D_1 + D_2)^2 \mathbb{1}_{\{\bar{E}\}}] \\
=&\mathbb{E}\left[\Big(A + B + C + D\Big)^2 \mathbb{1}_{\{\bar{E}\}}\right] \\
\leq&\mathbb{E}\left[\Big(A + B + C + D\Big)^2\right] \\
=&\mathbb{E}\left[\mathbb{E}^2\Big(A + B + C + D\Big|I_1, I_2, I_1', I_2'\Big) + \mathrm{Var}\Big(A + B + C + D\Big|I_1, I_2, I_1', I_2'\Big)\right],
\end{aligned} \tag{E.17}$$

where the last step follows from the conditional variance formula. For the second term in (E.17),

$$\begin{aligned}
&\mathrm{Var}\Big(A + B + C + D\Big|I_1, I_2, I_1', I_2'\Big) \\
\leq&4\mathrm{Var}\left[\sum_{i \in I_1 \cap I_1'} \Big(\hat{D}_{1,i} - \hat{D}_{2,i}\Big)\middle|I_1, I_1'\right] + 4\mathrm{Var}\left[\sum_{i \in I_2 \cap I_1'} \Big(\hat{D}_{1,i} - \hat{D}_{2,i}\Big)\middle|I_2, I_1'\right] \\
&+ 4\mathrm{Var}\left[\sum_{i \in I_1 \cap I_2'} \Big(\hat{D}_{1,i} - \hat{D}_{2,i}\Big)\middle|I_1, I_2'\right] + 4\mathrm{Var}\left[\sum_{i \in I_2 \cap I_2'} \Big(\hat{D}_{1,i} - \hat{D}_{2,i}\Big)\middle|I_2, I_2'\right].
\end{aligned} \tag{E.18}$$

Furthermore, we define $\mathcal{E}_1$, $\mathcal{E}_2$ and $\mathcal{E}'$ as follows:

$$\mathcal{E}_1 \triangleq \sum_{i \in I_1 \cap (I_1' \cup I_2')} (\hat{D}_{2,i} - P_i \log Q_i), \tag{E.19}$$

$$\mathcal{E}_2 \triangleq \sum_{i \in I_2 \cap (I_1' \cup I_2')} (\hat{D}_{2,i} - P_i \log Q_i), \tag{E.20}$$

$$\mathcal{E}' \triangleq \sum_{i \in (I_1 \cup I_2) \cap (I_1' \cup I_2')} (\hat{D}_{1,i} - P_i \log P_i). \tag{E.21}$$

Then, the first term in (E.17) can be bounded by

$$\begin{aligned}
&\mathbb{E}^2\Big(A+B+C+D\Big|I_1,I_2,I_1',I_2'\Big)\\
=&\mathbb{E}^2\Big(\mathcal{E}'-\mathcal{E}_1-\mathcal{E}_2\Big|I_1,I_2,I_1',I_2'\Big)\\
\le&2\mathbb{E}^2\Big(\mathcal{E}'|I_1,I_2,I_1',I_2'\Big)+2\mathbb{E}^2(\mathcal{E}_1+\mathcal{E}_2|I_1,I_2,I_1',I_2')\\
\le&2\mathbb{E}^2\Big(\mathcal{E}'|I_1,I_2,I_1',I_2'\Big)+4\mathbb{E}^2[\mathcal{E}_1|I_1,I_1',I_2']+4\mathbb{E}^2[\mathcal{E}_2|I_2,I_1',I_2']. \qquad \text{(E.22)}
\end{aligned}$$

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

$$\mathbb{E}^2\Big(\mathcal{E}'|I_1,I_2,I_1',I_2'\Big)\lesssim\frac{k^2}{m^2\log^2 k}. \qquad \text{(E.23)}$$

Thus, in order to bound (E.17), we bound the four terms in (E.18) and the two terms in (E.22) one by one.

E.1 Bounds on the Variance

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

We first show that

$$\text{Var}\left[\sum_{i\in I_1\cap I_1'}(\hat{D}_{1,i}-\hat{D}_{2,i})\middle|I_1,I_1'\right]\le 2\text{Var}\left[\sum_{i\in I_1\cap I_1'}\hat{D}_{1,i}\middle|I_1,I_1'\right]+2\text{Var}\left[\sum_{i\in I_1\cap I_1'}\hat{D}_{2,i}\middle|I_1,I_1'\right]. \qquad \text{(E.24)}$$

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

$$\text{Var}\left[\sum_{i\in I_1\cap I_1'}\hat{D}_{1,i}\middle|I_1,I_1'\right]\lesssim\frac{k^2}{m^2\log^2 k}. \qquad \text{(E.25)}$$

In order to bound $\text{Var}[\sum_{i\in I_1\cap I_1'}\hat{D}_{2,i}|I_1,I_1']$, 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,

$$\begin{aligned}
&\mathrm{Var}\left[\sum_{i\in I_1\cap I_1'} \hat{D}_{2,i}\middle| I_1, I_1'\right] \\
&= \sum_{i\in I_1\cap I_1'} \mathrm{Var}\Big(\frac{M_i}{m} g_L(N_i)\Big) \\
&= \sum_{i\in I_1\cap I_1'} \left[\Big(\mathrm{Var}(\frac{M_i}{m}) + \mathbb{E}(\frac{M_i}{m})^2\Big)\mathrm{Var}\big(g_L(N_i)\big) + \mathrm{Var}(\frac{M_i}{m})\Big(\mathbb{E}\big(g_L(N_i)\big)\Big)^2\right].
\end{aligned} \tag{E.26}$$

We note that $\mathrm{Var}(\frac{M_i}{m}) = \frac{P_i}{m}$, and $\mathbb{E}(\frac{M_i}{m}) = P_i$. We need to upper bound $\mathrm{Var}(g_L(N_i))$ and $\Big(\mathbb{E}\big(g_L(N_i)\big)\Big)^2$, for $i \in I_1 \cap I_1'$. Recall that $g_L(N_i) = \sum_{j=1}^{L} \frac{a_j}{(c_1 \log k)^{j-1}} (N_i)_{j-1} - \log \frac{n}{c_1 \log k}$. The following lemma from [13] is also useful, which provides an upper bound on the variance of $(N_i)_j$.

Lemma 6. *[13, Lemma 6] If $X \sim$Poi(λ) and $(x)_j = \frac{x!}{(x-j)!}$, then the variance of $(X)_j$ is increasing in λ and*

$$\mathit{Var}(X)_j \leq (\lambda j)^j \left(\frac{(2e)^{2\sqrt{\lambda j}}}{\pi\sqrt{\lambda j}} \vee 1\right). \tag{E.27}$$

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

$$\begin{aligned}
\mathrm{Var}\Big(g_L(N_i)\Big) &= \mathrm{Var}\Big(\sum_{j=2}^{L} \frac{a_j}{(c_1 \log k)^{j-1}} (N_i)_{j-1}\Big) \\
&\leq \left(\sum_{j=2}^{L} \frac{a_j}{(c_1 \log k)^{j-1}} \sqrt{\mathrm{Var}\big((N_i)_{j-1}\big)}\right)^2 \\
&\leq \left(\sum_{j=2}^{L} \frac{2e^{-1}2^{3L}}{(c_1 \log k)^{j-1}} \sqrt{\mathrm{Var}\big((N_i)_{j-1}\big)}\right)^2.
\end{aligned} \tag{E.28}$$

By Lemma 6, we obtain

$$\begin{aligned}\mathrm{Var}\Big((N_i)_{j-1}\Big) &\le \big(c_1\log k(j-1)\big)^{j-1}\left(\frac{(2e)^{2\sqrt{c_1\log k(j-1)}}}{\pi\sqrt{c_1\log k(j-1)}}\vee 1\right)\\ &\le (c_1c_0\log^2 k)^{j-1}\left(\frac{(2e)^{2\sqrt{c_1c_0\log^2 k}}}{\pi\sqrt{c_1c_0\log^2 k}}\vee 1\right). \qquad \text{(E.29)}\end{aligned}$$

Substituting (E.29) into (E.28), we obtain

$$\begin{aligned}\mathrm{Var}\Big(g_L(N_i)\Big) &\le L\sum_{j=2}^{L}\Big(\frac{2e^{-1}2^{3L}}{(c_1\log k)^{j-1}}\Big)^2\mathrm{Var}\Big((N_i)_{j-1}\Big) \qquad \text{(E.30)}\\ &\lesssim k^{2(c_0\log 8+\sqrt{c_0c_1}\log 2e)}\log k. \qquad \text{(E.31)}\end{aligned}$$

Furthermore, for $i\in I_1\cap I_1'$, we bound $\big|\mathbb{E}\big(g_L(N_i)\big)\big|$ as follows:

$$\begin{aligned}\Big|\mathbb{E}\big(g_L(N_i)\big)\Big| &= \left|\sum_{j=1}^{L}\frac{a_j}{(c_1\log k)^{j-1}}(nQ_i)^{j-1}-\log\frac{n}{c_1\log k}\right|\\ &\le \sum_{j=1}^{L}\frac{2e^{-1}2^{3L}}{(c_1\log k)^{j-1}}(c_1\log k)^{j-1}+\log\frac{n}{c_1\log k}\\ &\lesssim k^{c_0\log 8}\log k+\log n. \qquad \text{(E.32)}\end{aligned}$$

So far, we have all the ingredients we need to bound $\mathrm{Var}\big(\frac{M_i}{m}g_L(N_i)\big)$. Note that $P_i\le f(k)Q_i$, and $Q_i\le\frac{c_1\log k}{n}$ for $i\in I_1$. First, we derive the following bound:

$$\mathrm{Var}(\frac{M_i}{m})\mathrm{Var}\big(g_L(N_i)\big)\lesssim\frac{f(k)\log^2 k k^{2(c_0\log 8+\sqrt{c_0c_1}\log 2e)}}{mn}\lesssim\frac{kf(k)}{mn\log^2 k}, \qquad \text{(E.33)}$$

if $2(c_0\log 8+\sqrt{c_0c_1}\log 2e)<\frac{1}{2}$.

Secondly, we derive

$$\mathbb{E}(\frac{M_i}{m})^2\mathrm{Var}\big(g_L(N_i)\big)\lesssim\frac{f^2(k)\log^3 k k^{2(c_0\log 8+\sqrt{c_0c_1}\log 2e)}}{n^2}\lesssim\frac{kf^2(k)}{n^2\log^2 k}, \qquad \text{(E.34)}$$

if $2(c_0\log 8+\sqrt{c_0c_1}\log 2e)<\frac{1}{2}$.

Thirdly, we have

$$\begin{aligned}\mathrm{Var}(\frac{M_i}{m})\Big(\mathbb{E}\big(g_L(N_i)\big)\Big)^2 &\lesssim \frac{f(k)\log^3 k k^{2c_0\log 8}}{mn} + \frac{f(k)\log k \log^2 n}{mn}\\ &\lesssim \frac{kf(k)}{mn\log^2 k} + \frac{k^{1-\epsilon}f(k)\log k}{mn}\\ &\lesssim \frac{kf(k)}{mn\log^2 k},\end{aligned} \tag{E.35}$$

if $2c_0\log 8 < \frac{1}{2}$ and $\log^2 n \lesssim k^{1-\epsilon}$.

Combining these three terms together, we obtain

$$\mathrm{Var}\left[\sum_{i\in I_1\cap I_1'} \hat{D}_{2,i}\middle| I_1, I_1'\right] \lesssim \frac{k^2 f(k)}{mn\log^2 k} + \frac{k^2 f^2(k)}{n^2\log^2 k}. \tag{E.36}$$

Due to the fact that $\frac{k^2 f(k)}{mn\log^2 k} \lesssim \frac{k^2 f^2(k)}{n^2\log^2 k} + \frac{k^2}{m^2\log^2 k}$,

$$\mathbb{E}\Bigg[\mathrm{Var}\Big[\sum_{i\in I_1\cap I_1'} \hat{D}_{2,i}\Big| I_1, I_1'\Big]\Bigg] \lesssim \frac{f^2(k)k^2}{n^2\log^2 k} + \frac{k^2}{m^2\log^2 k}. \tag{E.37}$$

E.1.2 Bounds on $\mathrm{Var}\Big[\sum_{i\in I_2\cap I_1'}\big(\hat{D}_{1,i} - \hat{D}_{2,i}\big)\Big| I_2, I_1'\Big]$

Note that for $i \in I_2 \cap I_1'$, $Q_i > \frac{c_3\log k}{n}$ and $P_i \leq \frac{c_1'\log k}{m}$. Following steps similar to those in [13], it can be shown that

$$\mathrm{Var}\left[\sum_{i\in I_2\cap I_1'} \hat{D}_{1,i}\middle| I_2, I_1'\right] \lesssim \frac{k^2}{m^2\log^2 k}. \tag{E.38}$$

We further consider $\mathrm{Var}\left[\sum_{i\in I_2\cap I_1'}\hat{D}_{2,i}\right]$. By the definition of $\hat{D}_{2,i}$, for $i\in I_2\cap I_1'$, we have $\hat{D}_{2,i}=\frac{M_i}{m}\Big(\log\frac{N_i+1}{n}-\frac{1}{2(N_i+1)}\Big)$. Therefore,

$$\begin{aligned}
&\mathrm{Var}\left[\sum_{i\in I_2\cap I_1'}\hat{D}_{2,i}\middle|I_2,I_1'\right]\\
&=\sum_{i\in I_2\cap I_1'}\mathrm{Var}\left[\frac{M_i}{m}\Big(\log\frac{N_i+1}{n}-\frac{1}{2(N_i+1)}\Big)\right]\\
&\le 2\sum_{i\in I_2\cap I_1'}\mathrm{Var}\left[\frac{M_i}{m}\Big(\log\frac{N_i+1}{n}\Big)\right]+2\sum_{i\in I_2\cap I_1'}\mathrm{Var}\left[\frac{M_i}{m}\Big(\frac{1}{2(N_i+1)}\Big)\right].
\end{aligned}\tag{E.39}$$

The first term in (E.39) can be bounded as follows:

$$\begin{aligned}
&\sum_{i\in I_2\cap I_1'}\mathrm{Var}\left[\frac{M_i}{m}\Big(\log\frac{N_i+1}{n}\Big)\right]\\
&\le\sum_{i\in I_2\cap I_1'}\mathbb{E}\left[\left(\frac{M_i}{m}\Big(\log\frac{N_i+1}{n}\Big)-P_i\log Q_i\right)^2\right]\\
&=\sum_{i\in I_2\cap I_1'}\mathbb{E}\left[\left(\frac{M_i}{m}\Big(\log\frac{N_i+1}{n}\Big)-\frac{M_i}{m}\log Q_i+\frac{M_i}{m}\log Q_i-P_i\log Q_i\right)^2\right]\\
&\le\sum_{i\in I_2\cap I_1'}2\mathbb{E}\left[\left(\frac{M_i}{m}\right)^2\left(\log\frac{N_i+1}{n}-\log Q_i\right)^2\right]+\sum_{i\in I_2\cap I_1'}2\mathbb{E}\left[\left(\frac{M_i}{m}-P_i\right)^2\log^2 Q_i\right].
\end{aligned}\tag{E.40}$$

Note that for $i\in I_2\cap I_1'$, $Q_i>\frac{c_3\log k}{n}$ and $P_i\le\frac{c_1'\log k}{m}$. We then have the following bound on the first term in (E.40):

$$\begin{aligned}
&\sum_{i\in I_2\cap I_1'}\mathbb{E}\Big[\left(\frac{M_i}{m}\right)^2\left(\log\frac{N_i+1}{n}-\log Q_i\right)^2\Big]\\
&=\sum_{i\in I_2\cap I_1'}\frac{mP_i^2+P_i}{m}\mathbb{E}\Big[\left(\log\frac{N_i+1}{n}-\log Q_i\right)^2\Big]\\
&=\sum_{i\in I_2\cap I_1'}\frac{P_i(1+mP_i)}{m}\mathbb{E}\Big[\left(\log\frac{N_i+1}{n}-\log Q_i\right)^2\mathbb{1}_{\{N_i\le\frac{nQ_i}{2}\}}\\
&\qquad\qquad+\left(\log\frac{N_i+1}{n}-\log Q_i\right)^2\mathbb{1}_{\{N_i>\frac{nQ_i}{2}\}}\Big]\\
&\overset{(a)}{\lesssim}\frac{k^2}{m^2\log^2 k}+\frac{1}{m},
\end{aligned}\tag{E.41}$$

where (a) holds if $\frac{c_3(1-\log 2)}{2}+1-C>0$, and can be shown as follows. First,

$$\begin{aligned}
&\sum_{i\in I_2\cap I_1'} \frac{P_i(1+mP_i)}{m}\mathbb{E}\left[\left(\log\frac{N_i+1}{n}-\log Q_i\right)^2 \mathbb{1}_{\{N_i\le \frac{nQ_i}{2}\}}\right]\\
\overset{(a)}{\lesssim}&\sum_{i\in I_2\cap I_1'} \frac{P_i\log k}{m}(\log^2 n)P(N_i\le\frac{nQ_i}{2})\\
\overset{(b)}{\lesssim}&\frac{k\log k}{m}P(N_i\le\frac{nQ_i}{2})\\
\overset{(c)}{\lesssim}&\frac{k\log k}{m}k^{-\frac{c_3(1-\log 2)}{2}}\\
\overset{(d)}{\lesssim}&\frac{k^2}{m^2\log^2 k},
\end{aligned}\tag{E.42}$$

where (a) is due to the fact that $P_i\le\frac{c_1'\log k}{m}$, $\frac{N_i+1}{n}\in[\frac{1}{n},1]$, $Q_i\in[\frac{c_3\log k}{n},1]$; (b) is due to the assumption that $\log^2 n\lesssim k^{1-\epsilon}$ and $\sum_{i\in I_2\cap I_1'}P_i\le 1$; (c) is due to the Chernoff bound: $P(N_1\le\frac{nQ_i}{2})\le k^{-\frac{c_3(1-\log 2)}{2}}$; and (d) is due to the assumption that $\log m\le C\log k$ and $\frac{c_3(1-\log 2)}{2}+1-C>0$.

Secondly,

$$\begin{aligned}
&\sum_{i\in I_2\cap I_1'} \frac{mP_i(1+mP_i)}{m^2}\mathbb{E}\left[\left(\log\frac{N_i+1}{n}-\log Q_i\right)^2 \mathbb{1}_{\{N_1> \frac{nQ_i}{2}\}}\right]\\
\lesssim&\sum_{i\in I_2\cap I_1'} \frac{P_i\log k}{m}\mathbb{E}\left[\left(\log\frac{N_i+1}{n}-\log Q_i\right)^2 \mathbb{1}_{\{N_1> \frac{nQ_i}{2}\}}\right]\\
\overset{(a)}{\lesssim}&\sum_{i\in I_2\cap I_1'} \frac{P_i\log k}{m}\mathbb{E}\left[(\frac{N_i+1}{n}-Q_i)^2\frac{1}{\xi^2}\mathbb{1}_{\{N_i> \frac{nQ_i}{2}\}}\right]\\
\overset{(b)}{\le}&\sum_{i\in I_2\cap I_1'} \frac{P_i\log k}{m}\mathbb{E}\left[(\frac{N_i+1}{n}-Q_i)^2\frac{4}{Q_i^2}\mathbb{1}_{\{N_i> \frac{nQ_i}{2}\}}\right]\\
\le&\sum_{i\in I_2\cap I_1'} \frac{P_i\log k}{m}\mathbb{E}\left[(\frac{N_i+1}{n}-Q_i)^2\right]\frac{4}{Q_i^2}\\
\le&\sum_{i\in I_2\cap I_1'} \frac{P_i\log k}{m}\frac{nQ_i+1}{n^2}\frac{4}{Q_i^2}\\
\lesssim&\sum_{i\in I_2\cap I_1'} \frac{P_i\log k}{m}\frac{1}{\log k}\\
=&\frac{1}{m},
\end{aligned}\tag{E.43}$$

where (a) is due to the mean value theorem, with ξ satisfying $\min(\frac{N_i+1}{n}, Q_i) \leq \xi \leq \max(\frac{N_i+1}{n}, Q_i)$; (b) is due to the fact that $\xi \geq \frac{Q_i}{2}$.

We next bound the second term in (E.40).

$$\begin{aligned}\sum_{i\in I_2\cap I_1'} \mathbb{E}\left[\left(\frac{M_i}{m}-P_i\right)^2\log^2 Q_i\right] &= \sum_{i\in I_2\cap I_1'} \frac{P_i\log^2 Q_i}{m}\\ &\leq \sum_{i\in I_2\cap I_1'} \frac{P_i\log^2\frac{P_i}{f(k)}}{m}\\ &\leq \sum_{i\in I_2\cap I_1'} \frac{2P_i(\log^2 P_i+\log^2 f(k))}{m}\\ &\overset{(a)}{\leq}\frac{2\log^2 f(k)}{m}+\frac{2k\frac{c_1'\log k}{m}\log^2(\frac{c_1'\log k}{m})}{m}\\ &\overset{(b)}{\lesssim}\frac{\log^2 f(k)}{m}+\frac{k^2}{m^2\log^2 k},\end{aligned} \tag{E.44}$$

where (a) is due to the facts that $x\log^2 x$ is monotone increasing when x is small and $P_i \leq \frac{c_1'\log k}{m}$, and (b) is due to the assumption that $\log m \lesssim \log k$.

Substituting (E.44) and (E.41) into (E.40), we obtain

$$\sum_{i\in I_2\cap I_1'} \mathrm{Var}\left[\frac{M_i}{m}\Big(\log\frac{N_i+1}{n}\Big)\right] \lesssim \frac{\log^2 f(k)}{m}+\frac{k^2}{m^2\log^2 k}. \tag{E.45}$$

We then consider the second term in (E.39).

$$\begin{aligned}&\sum_{i\in I_2\cap I_1'} \mathrm{Var}\left[\frac{M_i}{m}\Big(\frac{1}{2(N_i+1)}\Big)\right]\\ &=\sum_{i\in I_2\cap I_1'}\left(\mathbb{E}^2[\frac{M_i}{m}]\mathrm{Var}[\frac{1}{2(N_i+1)}]+\mathrm{Var}[\frac{M_i}{m}]\mathbb{E}^2[\frac{1}{2(N_i+1)}]+\mathrm{Var}[\frac{M_i}{m}]\mathrm{Var}[\frac{1}{2(N_i+1)}]\right).\end{aligned} \tag{E.46}$$

In order to bound (E.46), we bound each term as follows. Note that M_i ~Poi(mP_i), and N_i ~Poi(nQ_i). Therefore, $\mathbb{E}^2[\frac{M_i}{m}] = P_i^2$, $\mathrm{Var}[\frac{M_i}{m}] = \frac{P_i}{m}$,

and

$$\begin{aligned}\mathrm{Var}[\frac{1}{2(N_i+1)}]+\mathbb{E}^2[\frac{1}{2(N_i+1)}] &= \mathbb{E}[\frac{1}{4(N_i+1)^2}]\\ &\leq \mathbb{E}[\frac{1}{(N_i+1)(N_i+2)}]\\ &= \sum_{i=0}^{\infty}\frac{1}{(i+1)(i+2)}\frac{e^{-nQ_i}(nQ_i)^i}{i!}\\ &= \sum_{i=0}^{\infty}\frac{1}{(nQ_i)^2}\frac{e^{-nQ_i}(nQ_i)^{i+2}}{(i+2)!}\\ &\leq \frac{1}{(nQ_i)^2}. \end{aligned}\tag{E.47}$$

Therefore, (E.46) can be further upper bounded as follows:

$$\begin{aligned}&\sum_{i\in I_2\cap I_1'}\mathrm{Var}\left[\frac{M_i}{m}\Big(\frac{1}{2(N_i+1)}\Big)\right]\\ \leq&\sum_{i\in I_2\cap I_1'}\left(P_i^2\frac{1}{(nQ_i)^2}+\frac{P_i}{m(nQ_i)^2}\right)\\ \lesssim&\frac{f(k)}{n\log k}+\frac{1}{m\log^2 k}\\ \lesssim&\frac{f(k)}{n}+\frac{1}{m}.\end{aligned}\tag{E.48}$$

Substituting (E.48) and (E.45) into (E.39), we obtain

$$\mathrm{Var}\left[\sum_{i\in I_2\cap I_1'}\hat{D}_{2,i}\middle|I_2,I_1'\right]\lesssim\frac{f(k)}{n}+\frac{\log^2 f(k)}{m}+\frac{k^2}{m^2\log^2 k}.\tag{E.49}$$

Therefore,

$$\mathrm{Var}\left[\sum_{i\in I_2\cap I_1'}\left(\hat{D}_{1,i}-\hat{D}_{2,i}\right)\middle|I_2,I_1'\right]\lesssim\frac{f(k)}{n}+\frac{\log^2 f(k)}{m}+\frac{k^2}{m^2\log^2 k}.\tag{E.50}$$

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

We first note that given $i\in I_1\cap I_2'$, $P_i>\frac{c_3'\log k}{m}$, $Q_i\leq\frac{c_1\log k}{n}$, and $\frac{P_i}{Q_i}\leq f(k)$. Hence, $\frac{c_3'\log k}{m}<P_i\leq\frac{c_1 f(k)\log k}{n}$. Following steps similar to those in [13], it

can be shown that

$$\mathrm{Var}\left[\sum_{i\in I_1\cap I_2'}\hat{D}_{1,i}\middle|I_1,I_2'\right]\le\frac{4}{m}+\frac{12k}{m^2}+\frac{4k}{c_3'm^2\log k}+\sum_{i\in I_1\cap I_2'}\frac{2P_i}{m}\log^2 P_i. \quad \text{(E.51)}$$

Consider the last term $\sum_{i\in I_1\cap I_2'}\frac{2P_i}{m}\log^2 P_i$ in (E.51), under the condition that $\frac{c_3'\log k}{m}<P_i\le\frac{c_1 f(k)\log k}{n}$. Then,

$$\begin{aligned}\sum_{i\in I_1\cap I_2'}\frac{P_i}{m}\log^2 P_i &\le \sum_{i\in I_1\cap I_2'}\frac{c_1 f(k)\log k}{mn}\log^2\frac{c_3'\log k}{m}\\ &\le \frac{c_1 k f(k)\log k}{mn}\log^2\frac{c_3'\log k}{m}\\ &\overset{(a)}{\lesssim}\frac{kf(k)\log k}{mn}\log^2 m\\ &\overset{(b)}{\lesssim}\frac{kf(k)\log^3 k}{mn}\lesssim\frac{k^2 f(k)}{mn\log^2 k}\\ &\overset{(c)}{\lesssim}\frac{f^2(k)k^2}{n^2\log^2 k}+\frac{k^2}{m^2\log^2 k}, \end{aligned}\quad \text{(E.52)}$$

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

$$\mathrm{Var}\left[\sum_{i\in I_1\cap I_2'}\hat{D}_{1,i}\middle|I_1,I_2'\right]\lesssim\frac{\log^2 f(k)}{m}+\frac{f^2(k)k^2}{n^2\log^2 k}+\frac{k^2}{m^2\log^2 k}. \quad \text{(E.53)}$$

Following steps similar to those in Appendix E.1.1, we can show that

$$\mathrm{Var}\left[\sum_{i\in I_1\cap I_2'}\hat{D}_{2,i}\middle|I_1,I_2'\right]\lesssim\frac{f(k)}{n}+\frac{f^2(k)k^2}{n^2\log^2 k}+\frac{k^2}{m^2\log^2 k}. \quad \text{(E.54)}$$

Hence,

$$\mathrm{Var}\left[\sum_{i\in I_1\cap I_2'}\left(\hat{D}_{1,i}-\hat{D}_{2,i}\right)\middle|I_1,I_2'\right]\lesssim\frac{\log^2 f(k)}{m}+\frac{f(k)}{n}+\frac{f^2(k)k^2}{n^2\log^2 k}+\frac{k^2}{m^2\log^2 k}. \quad \text{(E.55)}$$

E.1.4 Bounds on $\mathrm{Var}\Big[\sum_{i\in I_2\cap I_2'}\big(\hat{D}_{1,i}-\hat{D}_{2,i}\big)\Big|I_2,I_2'\Big]$

We note that for $i \in I_2 \cap I_2'$, $P_i > \frac{c_3' \log k}{m}$, $Q_i > \frac{c_3 \log k}{n}$, and

$$\hat{D}_{1,i}-\hat{D}_{2,i}=\frac{M_i}{m}\log\frac{M_i}{m}-\frac{1}{2m}-\frac{M_i}{m}\Big(\log\frac{N_i+1}{n}-\frac{1}{2(N_i+1)}\Big). \tag{E.56}$$

It can be shown that

$$\begin{aligned}&\mathrm{Var}\Bigg[\sum_{i\in I_2\cap I_2'}\big(\hat{D}_{1,i}-\hat{D}_{2,i}\big)\Bigg|I_2,I_2'\Bigg]\\ \le&2\mathrm{Var}\Bigg[\sum_{i\in I_2\cap I_2'}\frac{M_i}{m}\log\frac{M_i}{m}-\frac{M_i}{m}\log\frac{N_i+1}{n}\Bigg|I_2,I_2'\Bigg]\\ &+2\mathrm{Var}\Bigg[\sum_{i\in I_2\cap I_2'}\frac{M_i}{m}\frac{1}{2(N_i+1)}\Bigg|I_2,I_2'\Bigg].\end{aligned} \tag{E.57}$$

Following steps similar to those used in showing (E.48), we bound the second term in (E.57) as follows:

$$\mathrm{Var}\Bigg[\sum_{i\in I_2\cap I_2'}\frac{M_i}{m}\frac{1}{2(N_i+1)}\Bigg|I_2,I_2'\Bigg]\lesssim\frac{f(k)}{n}+\frac{1}{m}. \tag{E.58}$$

We next bound the first term in (E.57) as follows:

$$\begin{aligned}&\mathrm{Var}\Bigg[\sum_{i\in I_2\cap I_2'}\frac{M_i}{m}\log\frac{M_i}{m}-\frac{M_i}{m}\log\frac{N_i+1}{n}\Bigg|I_2,I_2'\Bigg]\\ &\le\sum_{i\in I_2\cap I_2'}\mathbb{E}\Bigg[\Big(\frac{M_i}{m}\log\frac{M_i}{m}-\frac{M_i}{m}\log\frac{N_i+1}{n}-P_i\log\frac{P_i}{Q_i}\Big)^2\Bigg]\\ &=\sum_{i\in I_2\cap I_2'}\mathbb{E}\Bigg[\Big(\frac{M_i}{m}\log\frac{M_i}{m}-\frac{M_i}{m}\log\frac{N_i+1}{n}-\frac{M_i}{m}\log P_i+\frac{M_i}{m}\log\frac{P_i}{Q_i}\\ &\qquad\qquad+\frac{M_i}{m}\log Q_i-P_i\log\frac{P_i}{Q_i}\Big)^2\Bigg]\\ &\le\sum_{i\in I_2\cap I_2'}\Bigg(3\mathbb{E}\Bigg[\Big(\frac{M_i}{m}(\log\frac{M_i}{m}-\log P_i)\Big)^2\Bigg]+3\mathbb{E}\Bigg[\Big((\frac{M_i}{m}-P_i)\log\frac{P_i}{Q_i}\Big)^2\Bigg]\\ &\quad+3\mathbb{E}\Bigg[\Big(\frac{M_i}{m}(\log\frac{N_i+1}{n}-\log Q_i)\Big)^2\Bigg]\Bigg).\end{aligned} \tag{E.59}$$

We further bound the three terms in (E.59) one by one. Note that $\log x \le x-1$ for any $x>0$. Therefore,

$$\frac{M_i}{m} - P_i \le \frac{M_i}{m}\log\frac{\frac{M_i}{m}}{P_i} \le \frac{M_i}{m} - P_i + \frac{(\frac{M_i}{m}-P_i)^2}{P_i}, \tag{E.60}$$

which implies that

$$\left(\frac{M_i}{m}\log\frac{\frac{M_i}{m}}{P_i}\right)^2 \le 2(\frac{M_i}{m}-P_i)^2 + 2\frac{(\frac{M_i}{m}-P_i)^4}{P_i^2}. \tag{E.61}$$

Taking expectations on both sides, we have

$$\mathbb{E}\left[\left(\frac{M_i}{m}\log\frac{\frac{M_i}{m}}{P_i}\right)^2\right] \le \frac{2P_i}{m} + \frac{6}{m^2} + \frac{2}{m^3 P_i} \le \frac{2P_i}{m} + \frac{6}{m^2} + \frac{2}{m^2 c_3'\log k}, \tag{E.62}$$

where the last inequality is due to the condition that $P_i \ge \frac{c_3'\log k}{m}$. Therefore,

$$\sum_{i\in I_2\cap I_2'} \mathbb{E}\left[\left(\frac{M_i}{m}(\log\frac{M_i}{m} - \log P_i)\right)^2\right] \lesssim \frac{1}{m} + \frac{k}{m^2}. \tag{E.63}$$

For the second term in (E.59), we derive the following bound:

$$\sum_{i\in I_2\cap I_2'} \mathbb{E}\left[\left((\frac{M_i}{m}-P_i)\log\frac{P_i}{Q_i}\right)^2\right] = \sum_{i\in I_2\cap I_2'} \frac{P_i}{m}\log^2\frac{P_i}{Q_i} \lesssim \frac{\log^2 f(k)}{m}, \tag{E.64}$$

where the last inequality is because

$$\begin{aligned}\sum_{i\in I_2\cap I_2'} P_i\log^2\frac{P_i}{Q_i} &= \sum_{i\in I_2\cap I_2'}\left(P_i\log^2\frac{P_i}{Q_i}\mathbb{1}_{\{\frac{1}{f(k)}\le\frac{P_i}{Q_i}\le f(k)\}} + Q_i\frac{P_i}{Q_i}\log^2\frac{P_i}{Q_i}\mathbb{1}_{\{\frac{P_i}{Q_i}\le\frac{1}{f(k)}\}}\right)\\ &\lesssim \log^2 f(k).\end{aligned} \tag{E.65}$$

where the last inequality is because the function $x\log^2 x$ is bounded by a constant on the interval $[0,1]$.

We next bound the third term in (E.59). Note that $\frac{\hat{x}-x}{\hat{x}} \le \log\frac{\hat{x}}{x} \le \frac{\hat{x}-x}{x}$, and therefore

$$
\begin{aligned}
&\sum_{i\in I_2\cap I_2'}\mathbb{E}\left[\left(\frac{M_i}{m}(\log\frac{N_i+1}{n}-\log Q_i)\right)^2\right]\\
=&\sum_{i\in I_2\cap I_2'}\mathbb{E}\left[\left(\frac{M_i}{m}(\log\frac{N_i+1}{nQ_i})\right)^2\mathbb{1}_{\{N_i\le\frac{nQ_i}{2}\}}+\left(\frac{M_i}{m}(\log\frac{N_i+1}{nQ_i})\right)^2\mathbb{1}_{\{N_i>\frac{nQ_i}{2}\}}\right]\\
\overset{(a)}{\le}&\sum_{i\in I_2\cap I_2'}\mathbb{E}\left[\frac{M_i^2}{m^2}\right]\mathbb{E}\left[\left(\frac{N_i+1-nQ_i}{1}\right)^2\mathbb{1}_{\{N_i\le\frac{nQ_i}{2}\}}+\left(\frac{N_i+1-nQ_i}{\frac{nQ_i}{2}}\right)^2\mathbb{1}_{\{N_i>\frac{nQ_i}{2}\}}\right]\\
\overset{(b)}{\le}&\sum_{i\in I_2\cap I_2'}(P_i^2+\frac{P_i}{m})\left(2nQ_iP(N_i\le\frac{nQ_i}{2})+\frac{8}{nQ_i}\right)\\
\le&\sum_{i\in I_2\cap I_2'}(P_i^2+\frac{P_i}{m})\left(2nQ_ie^{-\frac{nQ_i(1-\log 2)}{2}}+\frac{8}{nQ_i}\right)\\
\overset{(c)}{\lesssim}&\sum_{i\in I_2\cap I_2'}(P_i^2+\frac{P_i}{m})\frac{1}{nQ_i}\\
\lesssim&\frac{f(k)}{n}+\frac{kf(k)}{mn}.
\end{aligned}
\tag{E.66}
$$

where (a) is due to the mean value theorem and the fact $N_i+1\ge 1$; (b) uses the Chernoff bound of Poisson distribution; (c) is due to the fact that $x^2e^{-\frac{x(1-\log 2)}{2}}$ is bounded by a constant for $x>0$.

Combining (E.58), (E.63), (E.64) and (E.66), we obtain

$$
\mathrm{Var}\left[\sum_{i\in I_2\cap I_2'}\left(\hat{D}_{1,i}-\hat{D}_{2,i}\right)\middle| I_2,I_2'\right]\lesssim\frac{k}{m^2}+\frac{\log^2 f(k)}{m}+\frac{f(k)}{n}+\frac{kf(k)}{mn}.
\tag{E.67}
$$

E.2 Bounds on the Bias:

Consider the first term in (E.22). Based on the definition of the set I_1, $\mathcal{E}_1$ can be written as follows:

$$
\mathcal{E}_1=\sum_{i\in I_1\cap(I_1'\cup I_2')}\left(\frac{M_i}{m}g_L(N_i)-P_i\log Q_i\right).
\tag{E.68}
$$

Hence, $\left|\mathbb{E}[\mathcal{E}_1|I_1, I_1', I_2']\right| = \left|\sum_{i\in I_1\cap(I_1'\cup I_2')}\mathbb{E}[\frac{M_i}{m}g_L(N_i) - P_i\log Q_i\big|I_1, I_1', I_2']\right|$. For $i \in I_1\cap(I_1'\cup I_2')$, we have $0 \le Q_i \le \frac{c_1\log k}{n}$ and $\left|P_i\frac{\mu_L(Q_i)}{Q_i} - \frac{P_i}{Q_i}Q_i\log Q_i\right| \lesssim \frac{f(k)}{n\log k}$. Therefore,

$$\begin{aligned}\left|\mathbb{E}\Big[\frac{M_i}{m}g_L(N_i) - P_i\log Q_i\Big|I_1, I_1', I_2'\Big]\right| &= \left|P_i\frac{\mu_L(Q_i)}{Q_i} - P_i\log Q_i\right|\\ &\lesssim \frac{f(k)}{n\log k}.\end{aligned} \tag{E.69}$$

Hence, $|\mathbb{E}[\mathcal{E}_1|I_1, I_1', I_2']|$ can be bounded as follows:

$$\begin{aligned}\left|\mathbb{E}(\mathcal{E}_1|I_1, I_1', I_2')\right| &= \left|\sum_{i\in I_1\cap(I_1'\cup I_2')}\mathbb{E}\Big[\frac{M_i}{m}g_L(N_i) - P_i\log Q_i\Big|I_1, I_1', I_2'\Big]\right|\\ &\lesssim \frac{kf(k)}{n\log k}.\end{aligned} \tag{E.70}$$

Therefore,

$$\mathbb{E}\Big[\mathbb{E}^2\big[\mathcal{E}_1|I_1, I_1', I_2'\big]\Big] \lesssim \frac{k^2f^2(k)}{n^2\log^2 k}. \tag{E.71}$$

Now consider the second term in (E.22). Based on how we define I_2, $\mathcal{E}_2$ can be written as follows:

$$\begin{aligned}\mathcal{E}_2 &= \sum_{i\in I_2\cap(I_1'\cup I_2')}\left(\frac{M_i}{m}\Big(\log\frac{N_i+1}{n} - \frac{1}{2(N_i+1)}\Big) - P_i\log Q_i\right)\\ &= \sum_{i\in I_2\cap(I_1'\cup I_2')}\left((\frac{M_i}{m} - P_i)\log Q_i + \frac{M_i}{m}\log\frac{N_i+1}{nQ_i} - \frac{P_i}{2(N_i+1)}\right).\end{aligned} \tag{E.72}$$

Taking expectations on both sides, we obtain

$$\mathbb{E}\big[\mathcal{E}_2|I_2, I_1', I_2'\big] = \sum_{i\in I_2\cap(I_1'\cup I_2')}\mathbb{E}\left[P_i\log\frac{N_i+1}{nQ_i} - \frac{P_i}{2(N_i+1)}\middle|I_1, I_1', I_2'\right]. \tag{E.73}$$

Consider $\sum_{i\in I_2\cap(I_1'\cup I_2')}\mathbb{E}\Big[P_i\log\frac{N_i+1}{nQ_i}\Big|I_1, I_1', I_2'\Big]$. Note that for any $x > 0$,

$$\log x \le (x-1) - \frac{1}{2}(x-1)^2 + \frac{1}{3}(x-1)^3. \tag{E.74}$$

Since N_i ~Poi(nQ_i),

$$\mathbb{E}\left[P_i \log\frac{N_i+1}{nQ_i}\right] \le P_i\mathbb{E}\left[\left((\frac{N_i+1}{nQ_i}-1)-\frac{1}{2}(\frac{N_i+1}{nQ_i}-1)^2+\frac{1}{3}(\frac{N_i+1}{nQ_i}-1)^3\right)\right]$$
$$= P_i(\frac{1}{2nQ_i}+\frac{5}{6(nQ_i)^2}+\frac{1}{3(nQ_i)^3}). \tag{E.75}$$

It can be shown that

$$\mathbb{E}\left[\frac{P_i}{2(N_i+1)}\right] = \frac{P_i}{2nQ_i}(1-e^{-nQ_i}). \tag{E.76}$$

Hence, we obtain

$$\begin{aligned}\mathbb{E}\big[\mathcal{E}_2\big|I_2,I_1',I_2'\big] &\le \sum_{i\in I_2\cap(I_1'\cup I_2')} P_i(\frac{1}{2nQ_i}+\frac{5}{6(nQ_i)^2}+\frac{1}{3(nQ_i)^3})-\frac{P_i}{2nQ_i}(1-e^{-nQ_i})\\ &\overset{(a)}{\lesssim} \sum_{i\in I_2\cap(I_1'\cup I_2')} \frac{P_i}{n^2Q_i^2}\\ &\lesssim \frac{kf(k)}{n\log k},\end{aligned} \tag{E.77}$$

where (a) is due to the fact that xe^{-x} is bounded by a constant for $x\ge 0$.

We further derive a lower bound on $\mathbb{E}\big[\mathcal{E}_2\big|I_2,I_1',I_2'\big]$. For any $x\ge\frac{1}{5}$, it can be shown that

$$\log x \ge (x-1)-\frac{1}{2}(x-1)^2+\frac{1}{3}(x-1)^3-(x-1)^4. \tag{E.78}$$

Define the following event: $A_i=\{\frac{N_i}{nQ_i}>\frac{1}{5}\}$. We then rewrite $\mathbb{E}\big[\mathcal{E}_2\big|I_2,I_1',I_2'\big]$ as follows:

$$\begin{aligned}&\mathbb{E}\big[\mathcal{E}_2\big|I_2,I_1',I_2'\big]\\ &=\sum_{i\in I_2\cap(I_1'\cup I_2')}\mathbb{E}\left[P_i\log\frac{N_i+1}{nQ_i}\mathbb{1}_{\{A_i\}}+P_i\log\frac{N_i+1}{nQ_i}\mathbb{1}_{\{A_i^c\}}-\frac{P_i}{2(N_i+1)}\middle|I_2,I_1',I_2'\right]\\ &\ge\sum_{i\in I_2\cap(I_1'\cup I_2')}\mathbb{E}\left[P_i\log\frac{N_i+1}{nQ_i}\mathbb{1}_{\{A_i\}}-\frac{P_i}{2(N_i+1)}\middle|I_2,I_1',I_2'\right]\\ &\quad-\sum_{i\in I_2\cap(I_1'\cup I_2')}\left|\mathbb{E}\left[P_i\log\frac{N_i+1}{nQ_i}\mathbb{1}_{\{A_i^c\}}\middle|I_2,I_1',I_2'\right]\right|.\end{aligned} \tag{E.79}$$

Using (E.78), we obtain

$$\begin{aligned}&\mathbb{E}\Big[P_i \log\frac{N_i+1}{nQ_i}\mathbb{1}_{\{A_i\}}\Big|I_2,I_1',I_2'\Big]\\ \geq&\mathbb{E}\Big[P_i\Big((\frac{N_i+1}{nQ_i}-1)-\frac{1}{2}(\frac{N_i+1}{nQ_i}-1)^2\\ &\qquad+\frac{1}{3}(\frac{N_i+1}{nQ_i}-1)^3-(\frac{N_i+1}{nQ_i}-1)^4\Big)\mathbb{1}_{\{A_i\}}\Big|I_2,I_1',I_2'\Big].\end{aligned}\tag{E.80}$$

Note that

$$\begin{aligned}&\mathbb{E}\Big[(\frac{N_i+1}{nQ_i}-1)\mathbb{1}_{\{A_i\}}\Big|I_2,I_1',I_2'\Big]\\ =&\mathbb{E}\Big[(\frac{N_i+1}{nQ_i}-1)\Big|I_2,I_1',I_2'\Big]-\mathbb{E}\Big[(\frac{N_i+1}{nQ_i}-1)\mathbb{1}_{\{A_i^c\}}\Big|I_2,I_1',I_2'\Big]\\ \overset{(a)}{\geq}&\mathbb{E}\Big[(\frac{N_i+1}{nQ_i}-1)\Big|I_2,I_1',I_2'\Big]\\ =&\frac{1}{nQ_i},\end{aligned}\tag{E.81}$$

where (a) follows because $(\frac{N_i+1}{nQ_i}-1)\mathbb{1}_{\{A_i^c\}}\leq 0$. Similarly,

$$\mathbb{E}\left[\left(\frac{N_i+1}{nQ_i}-1\right)^3\mathbb{1}_{\{A_i\}}\Big|I_2,I_1',I_2'\right]\geq\mathbb{E}\left[\left(\frac{N_i+1}{nQ_i}-1\right)^3\right]=\frac{4}{(nQ_i)^2}+\frac{1}{(nQ_i)^3}.\tag{E.82}$$

For the term $\mathbb{E}\Big[(\frac{N_i+1}{nQ_i}-1)^2\Big|I_2,I_1',I_2'\Big]$, it can be shown that

$$\begin{aligned}&\mathbb{E}\Big[(\frac{N_i+1}{nQ_i}-1)^2\mathbb{1}_{\{A_i\}}\Big|I_2,I_1',I_2'\Big]\\ \leq&\mathbb{E}\Big[(\frac{N_i+1}{nQ_i}-1)^2\Big|I_2,I_1',I_2'\Big]=\frac{1}{nQ_i}+\frac{1}{(nQ_i)^2}.\end{aligned}\tag{E.83}$$

Similarly, it can be shown that

$$\begin{aligned}&\mathbb{E}\Big[(\frac{N_i+1}{nQ_i}-1)^4\mathbb{1}_{\{A_i\}}\Big|I_2,I_1',I_2'\Big]\leq\mathbb{E}\Big[(\frac{N_i+1}{nQ_i}-1)^4\Big|I_2,I_1',I_2'\Big]\\ =&\frac{1+3nQ_i}{(nQ_i)^3}+\frac{10}{(nQ_i)^3}+\frac{1}{(nQ_i)^4}.\end{aligned}\tag{E.84}$$

Combining these results together, we obtain

$$\mathbb{E}\Big[P_i \log\big(\frac{N_i+1}{nQ_i}\big)\mathbb{1}_{\{A_i\}}\Big| I_2, I_1', I_2'\Big] \geq \frac{P_i}{2nQ_i} - \frac{13P_i}{6(nQ_i)^2} - \frac{32P_i}{3(nQ_i)^3} - \frac{P_i}{(nQ_i)^4}. \tag{E.85}$$

From the previous results, we know that

$$\mathbb{E}\Big[-\frac{P_i}{2(N_i+1)}\Big| I_2, I_1', I_2'\Big] = -\frac{P_i}{2nQ_i}(1-e^{-nQ_i}). \tag{E.86}$$

Combining (E.85) and (E.86), it can be shown that

$$\begin{aligned}
&\sum_{i\in I_2\cap(I_1'\cup I_2')} \mathbb{E}\Big[P_i \log(\frac{N_i+1}{nQ_i})\mathbb{1}_{\{A_i\}} - \frac{P_i}{2(N_i+1)}\Big| I_2, I_1', I_2'\Big] \\
\geq &\sum_{i\in I_2\cap(I_1'\cup I_2')} \left(\frac{P_i}{2nQ_i} - \frac{13P_i}{6(nQ_i)^2} - \frac{32P_i}{3(nQ_i)^3} - \frac{P_i}{(nQ_i)^4} - \frac{P_i}{2nQ_i}(1-e^{-nQ_i})\right) \\
= &\sum_{i\in I_2\cap(I_1'\cup I_2')} \left(-\frac{13P_i}{6(nQ_i)^2} - \frac{32P_i}{3(nQ_i)^3} - \frac{P_i}{(nQ_i)^4} + \frac{P_i}{2nQ_i}e^{-nQ_i}\right).
\end{aligned} \tag{E.87}$$

We further bound the absolute value of the right-hand side of (E.87) as follows:

$$\left|\sum_{i\in I_2\cap(I_1'\cup I_2')} \left(-\frac{13P_i}{6(nQ_i)^2} - \frac{32P_i}{3(nQ_i)^3} - \frac{P_i}{(nQ_i)^4} + \frac{P_i}{2nQ_i}e^{-nQ_i}\right)\right| \lesssim \frac{kf(k)}{n\log k}, \tag{E.88}$$

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

For the $\mathbb{E}\big[P_i \log \frac{N_i+1}{nQ_i} \mathbb{1}_{\{A_i^c\}} \big| I_2, I_1', I_2'\big]$, it can be shown that

$$\begin{aligned}
&\sum_{i \in I_2 \cap (I_1' \cup I_2')} \left| \mathbb{E}\left[P_i \log \frac{N_i+1}{nQ_i} \mathbb{1}_{\{A_i^c\}} \middle| I_2, I_1', I_2' \right] \right| \\
\overset{(a)}{\leq}& \sum_{i \in I_2 \cap (I_1' \cup I_2')} P_i \log(nQ_i) P(A_i^c) \\
\overset{(b)}{\leq}& \sum_{i \in I_2 \cap (I_1' \cup I_2')} \frac{P_i}{(nQ_i)^2} (nQ_i)^2 \log(nQ_i) e^{-(1-\frac{\log(5e)}{5})nQ_i} \\
\overset{(c)}{\lesssim}& \frac{kf(k)}{n \log k},
\end{aligned} \tag{E.89}$$

where (a) is due to the fact that $N_i + 1 \geq 1$, and the fact that $Q_i > \frac{c_3 \log k}{n}$, hence $|\log \frac{N_i+1}{nQ_i}| \leq \log(nQ_i)$ for large k; (b) is due to the Chernoff bound, where $1 - \frac{\log(5e)}{5} > 0$; c is due to the fact that $(nQ_i)^2 \log(nQ_i) e^{-(1-\frac{\log(5e)}{5})nQ_i}$ is bounded by a constant for $nQ_i > 1$, and the fact that $nQ_i > c_3 \log k$. Thus, (E.88) and (E.89) yield

$$\mathbb{E}\big[\mathcal{E}_2 \big| I_2, I_1', I_2'\big] \gtrsim -\frac{kf(k)}{n \log k}. \tag{E.90}$$

Combining (E.77) and (E.90), we obtain,

$$\big|\mathbb{E}\big[\mathcal{E}_2 \big| I_2, I_1', I_2'\big]\big| \lesssim \frac{kf(k)}{n \log k}. \tag{E.91}$$

For the constant c_0, c_1, c_2 and c_3, note that $\log m \leq C \log k$ for some constant C, and we can choose $c_1 = 50(C+1)$, $c_2 = e^{-1}c_1$, $c_3 = e^{-1}c_2$, such that $c_1 - c_2 \log \frac{ec_1}{c_2} - 1 > C$, $c_3 - c_2 \log \frac{ec_3}{c_2} - 1 > C$ and $\frac{c_3(1-\log 2)}{2} + 1 - C > 0$ hold simultaneously. Also, we can choose $c_0 > 0$ sufficiently small, satisfying condition $2c_0 \log 8 < \frac{1}{2}$ and $2(c_0 \log 8 + \sqrt{c_0 c_1} \log 2e) < \frac{1}{2}$. Thus, we show the existence of c_0, c_1 and c_2.

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, pp. 1265–1287, Mar 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 Int. Symp. Information Theory (ISIT)*, 2014, pp. 356–360.

[9] Z. Zhang and M. Grabchak, "Nonparametric estimation of Küllback-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. the 43rd annual ACM Symposium on Theory of Computing*, 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] V. Totik, *Polynomial Approximation on Polytopes.* American Mathematical Society, 2014, vol. 232, no. 1091.

[18] A. F. Timan, *Theory of Approximation of Functions of a Real Variable*, ser. Dover books on Advanced Mathematics. Pergamon Press, 1963.

[19] P. P. Petrushev and V. A. Popov, *Rational Approximation of Real Functions.* Cambridge University Press, 2011, vol. 28.

[20] Y. Han, J. Jiao, and T. Weissman, "Minimax estimation of KL divergence between discrete distributions," *arxiv: 1605.09124*, 2016.

[21] L. 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.