Estimation of KL Divergence Between Large-Alphabet Distributions

Yuheng Bu* Shaofeng Zou† Yingbin Liang† Venugopal V. Veeravalli*
* University of Illinois at Urbana-Champaign † Syracuse University
Email: bu3@illinois.edu, szou02@syr.edu, yliang06@syr.edu, vvv@illinois.edu

***Abstract*—To be considered for the 2016 IEEE Jack Keil Wolf ISIT Student Paper Award. The problem of estimating the KL divergence between two unknown distributions is studied. The alphabet size k of the distributions can scale to infinity. The estimation is based on m and n independent samples respectively drawn from the two distributions. It is first shown that there does not exist any consistent estimator to guarantee asymptotic small worst-case quadratic risk over the set of all pairs of distributions. A restricted set that contains pairs of distributions with bounded ratio $f(k)$ is further considered. An augmented plug-in estimator is proposed, and is shown to be consistent if and only if $m=\omega(k\vee \log^2(f(k))$ and $n=\omega(kf(k))$. Furthermore, if $f(k)\geq \log^2 k$ and $\log^2(f(k))=o(k)$, it is shown that any consistent estimator must satisfy the necessary conditions: $m=\omega(\frac{k}{\log k}\vee \log^2(f(k))$ and $n=\omega(\frac{kf(k)}{\log k})$.**

I. Introduction

Consider estimation of Kullback-Leibler (KL) divergence between the probability distributions P and Q defined as

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

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

Suppose P and Q are unknown. Instead, m independent and identically distributed (i.i.d.) samples $X_1,\ \ldots\ ,X_m$ drawn from P and n i.i.d. samples $Y_1,\ \ldots\ ,Y_n$ drawn from Q are available for estimation. The sufficient statistics for estimating $D(P\|Q)$ are the histograms of the samples $M\triangleq(M_1,\ldots,M_k)$ and $N\triangleq(N_1,\ldots,N_k)$, where

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

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:

$$R(\hat{D},k,m,n)\triangleq\sup_{P,Q\in\mathcal{M}_k}\mathbb{E}[(\hat{D}(M,N)-D(P\|Q))^2]. \quad (1)$$

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

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

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

We further define the minimax quadratic risk as:

$$R^*(k,m,n)\triangleq\inf_{\hat{D}}R(\hat{D},k,m,n). \quad (2)$$

We are also interested in the following set

$$\mathcal{M}_{k,f(k)}=\left\{(P,Q):|P|=|Q|=k,\frac{P_i}{Q_i}\leq f(k),\ \forall\ 1\leq i\leq k\right\}, \quad (3)$$

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

$$R(\hat{D},k,m,n,f(k))\triangleq\sup_{(P,Q)\in\mathcal{M}_{k,f(k)}}\mathbb{E}[(\hat{D}(M,N)-D(P\|Q))^2], \quad (4)$$

and define the corresponding minimax quadratic risk as

$$R^*(k,m,n,f(k))\triangleq\inf_{\hat{D}}R(\hat{D},k,m,n,f(k)). \quad (5)$$

A. Comparison to Related Problems

Several estimators of KL divergence when P and Q are *continuous* have been proposed and shown to be consistent. The estimator proposed in [1] is based on data-dependent partition for density estimation, the estimator proposed in [2] is based on a k-nearest neighbor approach for density estimation, and the estimator developed in [3] utilizes a kernel-based approach for estimating the density ratio. A more general problem of estimating the f-divergence was studied in [4], where an estimator based on a weighted ensemble of plug-in estimators was proposed to trade bias with variance. All these approaches exploit the smoothness of continuous densities or

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

We adopt the following notations to express asymptotic scaling of quantities with n: $f(n)=\mathcal{O}(g(n))$ represents that there exist $k,n_0>0$ s.t. for all $n>n_0$, $|f(n)|\leq k|g(n)|$; $f(n)=\Omega(g(n))$ represents that there exist $c,n_0>0$ s.t. for all $n>n_0$, $f(n)\geq cg(n)$; $f(n)=\Theta(g(n))$ represents that there exist $c_1,c_2,n_0>0$ s.t. for all $n>n_0$, $c_1g(n)\leq f(n)\leq c_2g(n)$; $f(n)=\omega(g(n))$ represents that for all $c>0$, there exists $n_0>0$ s.t. for all $n>n_0$, $|f(n)|\geq c|g(n)|$; and $f(n)=o(g(n))$ represents that for all $c>0$, there exists $n_0>0$ s.t. for all $n>n_0$, $|f(n)|\leq cg(n)$.

density ratios, which guarantees that samples falling into a certain neighborhood area can be used to estimate the local density or density ratio accurately. However, such a smoothness property does not hold for discrete distributions, whose probabilities over adjacent point masses can vary significantly. In fact, [1] provides an example to show that estimation of KL divergence can be difficult even for continuous distributions if the density has sharp dips.

Estimation of KL divergence when the distributions P and Q are discrete has been studied in [5] for the regime with *fixed* alphabet cardinality k and large sample sizes m and n. Such a regime is very different from the large-alphabet regime in which we are interested, with k scaling to infinity. Clearly, as k increases, the scaling of the sample sizes m and n must be fast enough with respect to k in order to guarantee consistent estimation.

In the large-alphabet regime, KL divergence estimation is closely related to the entropy estimation with a large alphabet recently studied in [6]–[8]. Compared to entropy estimation, KL divergence estimation has one more dimension of uncertainty, that about the distribution Q. Some distributions Q can contain very small point masses that contribute significantly to the value of divergence, but are difficult to estimate because samples of these point masses occur rarely. In fact, such distributions dominate the risk in (1), and make the construction of consistent estimators challenging.

B. Our Contributions

Our contributions contain the following three results.

We first show, using Le Cam's two-point method [9], that there is no consistent estimator of KL divergence over the distribution set $\mathcal{M}_k$. As described above, this is due to the fact that the set $\mathcal{M}_k$ contains distributions Q, which have arbitrarily small components that contribute significantly to KL divergence but require arbitrarily large number of samples to estimate accurately.

Thus, we further focus on the set $\mathcal{M}_{k,f(k)}$ given in (3) that contains distributions (P,Q) with their ratio bounded by $f(k)$. We construct an augmented plug-in estimator and show that such an estimator is consistent over $\mathcal{M}_{k,f(k)}$ if and only if $m = \omega(k \vee \log^2(f(k))$ and $n = \omega(kf(k))$. Our proof of the sufficient conditions is based on evaluating the bias and variance separately. Our proof of the necessary condition $m = \omega(\log^2(f(k))$ is based on Le Cam's two-point method with a judiciously chosen pair of distributions. And our proof of the necessary conditions $m = \omega(k)$ and $n = \omega(kf(k))$ is based on analyzing the bias of the estimator and constructing different pairs of "worst case" distributions for the cases when either the bias caused by insufficient samples from P or the bias caused by insufficient samples from Q dominates, respectively.

We further show that if $f(k) \geq \log^2 k$ and $\log^2(f(k)) = o(k)$, any consistent estimator of KL divergence over $\mathcal{M}_{k,f(k)}$ must satisfy $m = \omega(\frac{k}{\log k} \vee \log^2(f(k))$ and $n = \omega(\frac{kf(k)}{\log k})$. Our proof is based on an extension of Le Cam's two-point method to composite hypotheses. Comparing to entropy estimation problem [7], the challenge here that requires special technical treatment is to construct prior distributions for (P,Q) that satisfy the bounded ratio constraint.

II. Main Results

In this section, we first show that there does not exist any consistent estimator of KL divergence over the set $\mathcal{M}_k$. We then focus on the set $\mathcal{M}_{k,f(k)}$, and study the consistency of an augmented plug-in estimator, and characterize necessary conditions on the sample complexity for any consistent estimator. Due to space limitations, we provide only outlines of our proofs, with detailed proofs available in [10].

A. No Consistent Estimator over $\mathcal{M}_k$

In the following theorem, we show that the minimax risk over the set $\mathcal{M}_k$ is unbounded for arbitrary alphabet size k and m and n samples.

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

Outline of Proof. Theorem 1 follows from Le Cam's two-point method [9]: If two pairs of distributions (P_1, Q_1) and (P_2, Q_2) are sufficiently close such that it is impossible to reliably distinguish between them using m samples from P and n samples from Q with error probability less than some constant, then any estimator suffers a quadratic risk proportional to the difference between the divergence values $|D(P_1\|Q_1) - D(P_2\|Q_2)|^2$. The details of the proof can be found in [10]. □

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

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

As we have shown in Section II-A, there does not exist any consistent estimator of KL divergence over the set $\mathcal{M}_k$. In this subsection, we study the risk of an estimator over the set $\mathcal{M}_{k,f(k)}$, and characterize under what sample complexity such an estimator is consistent.

In order to estimate the KL divergence between a pair of distributions, a natural idea is the "plug-in" approach, 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

$$\hat{D}_{\mathrm{plug-in}}(M,N) = D(\hat{P}\|\hat{Q}), \tag{6}$$

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

Unlike the entropy estimation problem, where the plug-in estimator $\hat{H}_{\mathrm{plug-in}}$ is asymptotically efficient in the "fixed P large n" regime, the direct plug-in estimator $\hat{D}_{\mathrm{plug-in}}$ in (6) of KL divergence has an infinite bias. This is because, with non-zero probability, $N_j = 0$ and $M_j \neq 0$ for some $j \in [k]$, leading to infinite $\hat{D}_{\mathrm{A-plug-in}}$.

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 $\hat{Q}_i'$

$$\hat{D}_{\mathrm{A-plug-in}}(M,N) = \sum_{i=1}^{k} \frac{M_i}{m} \log \frac{M_i/m}{(N_i+1)/n}. \tag{7}$$

Remark 1. *For technical convenience, $\hat{Q}_i'$ is not normalized after adding samples. It can be shown that normalization does not provide order-level smaller risk for the plug-in estimator. Furthermore, the so-called add-constant estimator [11] of Q, which adds a fraction sample to each mass point of Q, can also be used as an estimator of divergence. Although intuitively such an estimator should not provide order-level improvement in the risk, the analysis of the risk appears to be difficult.*

We next characterize sufficient and necessary conditions on the sample complexity to guarantee consistency of the augmented plug-in estimator over $\mathcal{M}_{k,f(k)}$. To this end, we first provide upper and lower bounds, respectively, on $R(\hat{D}_{\mathrm{A-plug-in}}, k, m, n, f(k))$ in the following two propositions.

Proposition 1. *For all $k, m, n \in \mathbb{N}$*

$$R(\hat{D}_{\mathrm{A-plug-in}}, k, m, n, f(k)) = \mathcal{O}\left(\left(\frac{kf(k)}{n} + \log\left(1 + \frac{k-1}{m}\right)\right)^2 + \frac{\log^2 f(k)}{m} + \frac{f(k)}{n}\right).$$

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

Outline of Proof. The proof consists of separately bounding the bias and variance of the augmented plug-in estimator. The details can be found in [10]. □

Proposition 2. *If $m = \mathcal{O}(k \vee \log^2 f(k))$, or $n = \mathcal{O}(kf(k))$, then for sufficiently large k*

$$R(\hat{D}_{\mathrm{A-plug-in}}, k, m, n, f(k)) \geq c' \tag{8}$$

where c' is a positive constant.

Outline of Proof. It can be shown that the bias of the augmented plug-in estimator is upper and lower bounded as follows:

$$\begin{aligned} &\left(\frac{k}{m} \wedge 1\right) - \frac{kf(k)}{n} && \text{(9a)}\\ &\leq \sup_{(P,Q)\in\mathcal{M}_{k,f(k)}} \mathbb{E}[\hat{D}_{\mathrm{A-plug-in}}(m,n) - D(P\|Q)] \\ &\leq \log\left(1 + \frac{k}{m}\right) - \frac{k-1}{k}\exp\left(-\frac{2n}{kf(k)}\right). && \text{(9b)} \end{aligned}$$

1) If $m = \mathcal{O}(k)$ and $n = \omega(kf(k))$, the lower bound in (9a) converges to 1. Hence, the bias as well as the risk is lower bounded by a positive constant.
2) If $m = \omega(k)$ and $n = \mathcal{O}(kf(k))$, the upper bound in (9b) converges to a negative constant. This implies that the risk is lower bounded by a positive constant.
3) If $m = \mathcal{O}(k)$ and $n = \mathcal{O}(kf(k))$, the lower bound (9a) converges to $-\infty$ and the upper bound (9b) converges to $+\infty$, which does not provide useful information. Hence, we design another approach for this case as follows.

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

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

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

4) If $m = \mathcal{O}(\log^2 f(k))$, we construct two pairs of distributions as follows:

$$\begin{aligned} P_1 &= \left(\frac{1}{2(k-1)}, \cdots, \frac{1}{2(k-1)}, \frac{2}{3}\right), \\ P_2 &= \left(\frac{1+\epsilon'}{2(k-1)}, \cdots, \frac{1+\epsilon'}{2(k-1)}, \frac{2-\epsilon'}{3}\right), \\ Q_1 = Q_2 &= \left(\frac{1}{3(k-1)f(k)}, \cdots, \frac{1}{3(k-1)f(k)}, 1 - \frac{1}{3f(k)}\right). \end{aligned}$$

By Le Cam's two-point method [9], it can be shown that if $m = \mathcal{O}(\log^2 f(k))$, no estimator can be consistent, which implies that the augmented plug-in estimator is not consistent. □

Combining Propositions 1 and 2, we have the following theorem on the consistency of the augmented plug-in estimator.

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

$$m = \omega(k \vee \log^2(f(k))) \quad \textit{and} \quad n = \omega(kf(k)). \tag{10}$$

C. Minimax Lower Bound over $\mathcal{M}_{k,f(k)}$

In this subsection, we characterize necessary conditions on the sample complexity that all consistent estimators of KL divergence over $\mathcal{M}_{k,f(k)}$ must satisfy. The general idea is to apply generalized Le Cam's two-point method [9] to develop a lower bound on the minimax risk.

1) Poisson sampling: We first utilize the *Poisson sampling* technique to handle the dependency of the multinomial distribution, as is done in [7] for entropy estimation. We relax the deterministic sample sizes m and n to Poisson random variables $m' \sim \mathrm{Poi}(m)$ with mean m and $n' \sim \mathrm{Poi}(n)$ with mean n, respectively. Under this model, we draw m' and n' i.i.d. samples from P and Q, respectively. The sufficient statistics $M_i \sim \mathrm{Poi}(nP_i)$ and $N_i \sim \mathrm{Poi}(nQ_i)$ are independent, which significantly simplifies the analysis.

Analogous to the minimax risk (5), we define its counterpart under the Poisson sampling model as

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

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

Lemma 1. *There exists a constant $c > \frac{1}{4}$ such that*

$$\begin{aligned}&\tilde{R}^*(k,2m,2n,f(k)) - e^{-cm}\log f(k) - e^{-cn}\log f(k)\\ &\quad \le R^*(k,m,n,f(k)) \le 4\tilde{R}^*(k,m/2,n/2,f(k)).\end{aligned} \tag{11}$$

All the results that we prove in this subsection are under the Poisson sampling assumption.

2) Minimax lower bound: We lower bound the minimax risk by the minimax risk with P or Q being known and carefully chosen to tighten the bound. In both cases, we choose the known P or Q to be the uniform distribution. Such a choice yields the following two propositions.

Proposition 3. *For all $k,m,n \in \mathbb{N}$ and $f(k) \ge \log^2 k$,*

$$\tilde{R}^*(k,m,n,f(k)) \ge \tilde{R}^*(k,m,Q,f(k)) = \Theta\left(\frac{k}{m\log k}\right)^2,$$

where

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

is the minimax risk under Poisson sampling with Q being known.

Outline of Proof. Setting $Q = Q_0$ to be uniform distribution on $[k]$, $D(P\|Q_0) = \sum_{i=1}^k P_i \log P_i + \log k = H(P) + \log k$, and the problem reduces to entropy estimation under the minimax risk with (P, Q_0) satisfying the bounded ratio constraint. If $f(k) \ge \log^2 k$, following steps similar to those in [7], it can be shown that $R^*(k,m,Q,f(k))$ is lower bounded by $\left(\frac{k}{m\log k}\right)^2$ at the order level. □

Proposition 4. *If $f(k) \ge \log^2 k$, $\log^2(f(k)) = o(k)$ and $n = \mathcal{O}(\frac{kf(k)}{\log k})$, then for sufficiently large k*

$$\tilde{R}^*(k,m,n,f(k)) \ge \tilde{R}^*(k,P,n,f(k)) \ge c$$

where c is a positive constant, and

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

is the minimax risk under Poisson sampling with P being known.

Outline of Proof. The proof applies the generalized Le Cam's method [9] that involves the two *composite* hypotheses to lower bound the minimax risk and adapts techniques for entropy estimation in [7]. The new challenge here arises due to the bounded ratio constraint on (P,Q), which requires special technical treatments to construct prior distributions, as well as bounding various divergence related quantities.

Let $P = P_0$ be uniform distribution. Then the minimax risk can be bounded as

$$\begin{aligned}&\tilde{R}^*(k,P,n,f(k))\\ &\ge \inf_{\hat{D}} \sup_{P_0,Q\in\mathcal{M}_{k,f(k)}} \mathbb{E}[(\hat{D}(P_0,N) - D(P_0\|Q))^2].\end{aligned}$$

To use the generalized Le Cam's method [9], consider the following two *composite* hypotheses:

$$H_0 : D(P_0\|Q) \le t \quad \text{versus} \quad H_1 : D(P_0\|Q) \ge t + d. \tag{12}$$

If the optimal test cannot distinguish the two hypotheses in (12) reliably, then the quadratic risk is lower bounded by $\Theta(d^2)$. Furthermore, the optimal probability of error for composite hypotheses testing is given by the Bayesian risk with respect to the least favorable prior.

In the following we construct tractable prior distributions. Let V and V' be two $\mathbb{R}^+$ valued random variables defined on the interval $[\eta,\lambda]$ and have equal mean $\mathbb{E}(V) = \mathbb{E}(V') = \alpha$. We construct two random vectors $\mathsf{Q} = \frac{1}{k}(V_1,\ldots V_{k-1}, k-(k-1)\alpha)$ and $\mathsf{Q}' = \frac{1}{k}(V'_1,\ldots V'_{k-1}, k-(k-1)\alpha)$ consisting of $k-1$ i.i.d. copies of V and V' and a deterministic term $1 - \frac{(k-1)\alpha}{k}$, respectively. Since we choose $P = P_0$ to be uniform distribution, and (P,Q) satisfy the bounded ratio constraint, Q_i must be greater than $\frac{1}{kf(k)}$. This yields a different construction from [7]. Note that $V, V' \in [\eta,\lambda]$. To satisfy the bounded ratio constraint, we assume that $\eta \ge \frac{1}{f(k)}$.

Due to the law of large numbers, the vectors Q and Q' are approximately probability distributions. Furthermore, the elements in Q and Q' are independent, which significantly simplifies the analysis.

Next we outline the main ingredients in Le Cam's method with priors Q and Q'. Note that $D(P_0\|\mathsf{Q})$ converges to its expectation $\mathbb{E}[D(P_0\|\mathsf{Q})]$ as k goes to infinity. Then with high

probability, $D(P_0\|\mathsf{Q})$ and $D(P_0\|\mathsf{Q}')$ are separated by the difference of their means

$$d = \mathbb{E}[D(P_0\|\mathsf{Q})] - \mathbb{E}[D(P_0\|\mathsf{Q}')] = \mathbb{E}[\log V'] - \mathbb{E}[\log V].$$

Since Q is drawn from the prior distribution Q, the sufficient statistics $N = (N_1, \ldots, N_k)$ are i.i.d. distributed according to the Poisson mixture $\mathbb{E}[\mathrm{Poi}(\frac{n}{k}V)]$. To establish the impossibility of hypothesis testing (12), the total variation between the two k-product distributions should satisfy

$$\mathrm{TV}(\mathbb{E}[\mathrm{Poi}(nV/k)], \mathbb{E}[\mathrm{Poi}(nV'/k)]) \le c/k. \qquad (13)$$

In fact, the i.i.d. construction of Q and Q' fully exploits independence imposed by Poisson sampling, and reduces the problem to one dimension. What remains is to choose V and V' to maximize $\mathbb{E}[\log V'] - \mathbb{E}[\log V]$, subject to the constraint (13). A commonly used proxy for bounding the total variation is obtained via *moment matching*, i.e., by solving the following optimization problem with moment matching constraints

$$\begin{aligned}\mathcal{E}_L(\eta, \lambda) \triangleq \max\ & \mathbb{E}[\log V'] - \mathbb{E}[\log V] \\ \text{s.t. } & \mathbb{E}[V^j] = \mathbb{E}[V'^j], \quad j = 1, \ldots, L, \\ & V, V' \in [\eta, \lambda],\end{aligned}$$

for some appropriately chosen $L \in \mathbb{N}$, $\eta \ge \frac{1}{f(k)}$ and λ depending on n and k.

As shown in [7], we have

$$\mathcal{E}_L(\eta, \lambda) = 2E_L(\log, [\eta/\lambda, 1]) \qquad (14)$$

where $E_L(g, I)$ is the best uniform approximation error of a function g over a finite interval I by polynomials of degree L. Due to the singularity of the logarithm at zero, the approximation error can be made bounded away from zero if η/λ grows quadratically with the degree L^{-1}. Choosing $\eta = \frac{1}{f(k)}$, $\lambda = c_1 \frac{\log^2 k}{f(k)}$, $c_1 \le 1$, $L = \log k$ and together with the condition $\log^2(f(k)) = o(k)$, the minimax risk can be shown to be lower bounded away from zero if $n = \mathcal{O}(\frac{kf(k)}{\log k})$. □

Combining Propositions 3, 4 and the necessary condition $m = \omega(\log^2 f(k))$ from Le Cam's two-point method (case (4) in the proof of Proposition 2), we obtain the following theorem on the necessary conditions for an estimator to be consistent.

Theorem 3. *If $\log^2(f(k)) = o(k)$ and $f(k) \ge \log^2 k$, then any consistent estimator of KL divergence over $\mathcal{M}_{k,f(k)}$ must satisfy*

$$m = \omega(\frac{k}{\log k} \vee \log^2 f(k)) \quad \textit{and} \quad n = \omega(\frac{kf(k)}{\log k}). \qquad (15)$$

Comparing Theorem 3 with Theorem 2 that characterizes the sample complexity for consistent augmented plug-in estimator, there is a gap of the order $\log k$. A promising approach to fill in this gap is to incorporate polynomial approximation into estimator construction to trade bias with variance as in entropy estimation. However, such an approach can be difficult to develop for KL divergence (as a function of two distributions) due to the fact that the best polynomial approximation to multi-variable functions is not well understood yet.

We also note that our proof of Proposition 4 may be strengthened by designing a jointly distributed prior on (P, Q), instead of treating them separately. This may help to relax or remove the conditions $\log^2(f(k)) = o(k)$ and $f(k) \ge \log^2 k$ in Proposition 3 and 4 and Theorem 3.

III. Conclusion

We have shown that there exists no consistent estimator for KL divergence under the worst-case quadratic risk over the set of all pairs of distributions, and therefore focused on the set of pairs of distributions with bounded ratio. We have proposed an augmented plug-in estimator, and characterized tight sufficient and necessary conditions for such an estimator to be consistent. We have also developed necessary conditions on the sample complexity for any consistent estimator, which is within a $\log k$ factor from the that of augmented plug-in estimator. In future work, we hope to find an improved estimator that has sample complexity that approaches our lower bound.

Acknowledgment

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

References

[1] 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.

[2] 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.

[3] 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.

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

[5] Z. Zhang and M. Grabchak, "Nonparametric estimation of Küllback-Leibler divergence," *Neural computation*, vol. 26, no. 11, pp. 2570–2593, 2014.

[6] G. Valiant and P. Valiant, "Estimating the unseen: an n/log (n)-sample estimator for entropy and support size, shown optimal via new clts," in *Proc. of the 43rd annual ACM symposium on Theory of computing*. ACM, 2011, pp. 685–694.

[7] Y. Wu and P. Yang, "Minimax rates of entropy estimation on large alphabets via best polynomial approximation," *arXiv:1407.0381*, 2014.

[8] J. Jiao, K. Venkat, and T. Weissman, "Maximum likelihood estimation of functionals of discrete distributions," *arxiv:1406.6959*.

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

[10] Y. Bu, S. Zou, Y. Liang, and V. V. Veeravalli, "KL divergence estimation over large alphabet," *available at https://szou02.mysite.syr.edu/conference/isit2016.pdf*, 2016.

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