Minimax Estimation of the L_1 Distance

Jiantao Jiao
Stanford University
jiantao@stanford.edu

Yanjun Han
Stanford University
yjhan@stanford.edu

Tsachy Weissman
Stanford University
tsachy@stanford.edu

***Abstract*—We consider the problem of estimating the L_1 distance between two discrete probability measures P and Q from empirical data in a nonasymptotic and large alphabet setting. We construct minimax rate-optimal estimators for $L_1(P,Q)$ when Q is either known or unknown, and show that the performance of the optimal estimators with n samples is essentially that of the Maximum Likelihood Estimators (MLE) with $n \ln n$ samples. Hence, we demonstrate that the *effective sample size enlargement* phenomenon, discovered and discussed in Jiao *et al.* (2015), holds for this problem as well. However, the construction of optimal estimators for $L_1(P,Q)$ requires new techniques and insights outside the scope of the *Approximation* methodology of functional estimation in Jiao *et al.* (2015).**

I. Introduction

A. Problem formulation

Statistical functionals are usually used to quantify the fundamental limits of data processing tasks such as data compression (e.g. Shannon entropy [2]), data transmission (e.g. mutual information [2]), estimation and testing (e.g. Kullback–Leibler divergence), etc. They measure the difficulties of the corresponding data processing tasks and shed light on how much improvement one may expect beyond the current state-of-the-art approaches. In this sense, it is of great value to obtain accurate estimates of these functionals in various problems.

In this paper, we consider estimating the L_1 distance between two discrete distributions $P = (p_1, p_2, \ldots, p_S), Q = (q_1, q_2, \ldots, q_S)$, which is defined as:

$$L_1(P,Q) \triangleq \sum_{i=1}^{S} |p_i - q_i|. \tag{1}$$

Throughout we use the squared error loss, i.e., the risk function for an estimator $\hat{L}$ is defined as

$$R(P,Q;\hat{L}) \triangleq \mathbb{E}|\hat{L}(X^n, Y^n) - L_1(P,Q)|^2, \tag{2}$$

where $(X_i, Y_i) \overset{\text{i.i.d.}}{\sim} P \times Q$. The maximum risk of an estimator $\hat{L}$, and the minimax risk in estimating $L_1(P,Q)$ are defined as

$$R_{\text{maximum}}(\mathcal{P}, \mathcal{Q}; \hat{L}) \triangleq \sup_{P\in\mathcal{P}, Q\in\mathcal{Q}} R(P,Q;\hat{L}), \tag{3}$$

$$R_{\text{minimax}}(\mathcal{P}, \mathcal{Q}) \triangleq \inf_{\hat{L}} \sup_{P\in\mathcal{P}, Q\in\mathcal{Q}} R(P,Q;\hat{L}), \tag{4}$$

respectively, where $\mathcal{P}, \mathcal{Q}$ are given collections of probability measures P and Q, respectively, and the infimum is taken over all possible estimators $\hat{L}$.

The L_1 distance is closely related to the Bayes error, i.e., the fundamental limit, in classification problems. Specifically, for a two-class classification problem, if the prior probabilities for each class are equal, then the minimum probability of error achieved using the optimal classifier is given by

$$L^* = \frac{1}{2} - \frac{1}{4} L_1(P_{X|Y=1}, P_{X|Y=0}), \tag{5}$$

where $Y \in \{0,1\}$ indicates the class, and $P_{X|Y}$ are the class-conditional distributions. Hence, the problem of estimating L^* in this classification problem is reduced to estimating the L_1 distance between the two class-conditional distributions $P_{X|Y=1}, P_{X|Y=0}$ from the empirical data. In the statistical learning theory literature, most work on Bayes classification error estimation deals with the case that $P_{X|Y=1}$ and $P_{X|Y=0}$ are continuous distributions, and it turns out that even in an asymptotic setting it is very difficult to estimate this quantity in this general continuous case. Indeed, we know from [3, Section 8.5] the negative result that for every sample size n, any estimate of the Bayes error $\hat{L}_n$, and any $\epsilon > 0$, there exist some class-conditional distributions such that $\mathbb{E}|\hat{L}_n - L^*| \geq \frac{1}{4} - \epsilon$.

This negative result shows that one needs to look at special classes of the class-conditional distributions in order to obtain consistent estimates. In the discrete setting, the seminal work of Valiant and Valiant [4] deserves special mention. They constructed an estimator for $L_1(P,Q)$ and showed that when $S/\ln S \lesssim n \lesssim S$, it achieves L_1 error $\sqrt{S/(n\ln n)}$, and it is the best possible rate for the constant L_1 error regime. It is quite simple to argue, as we do in this paper, that the simplest estimator for $L_1(P,Q)$, namely plugging in the empirical distribution P_n, Q_n and obtaining $L_1(P_n, Q_n)$ achieves L_1 error rate $\sqrt{S/n}$ for $n \gtrsim S$. In this sense, the optimal estimator seems to enlarge the sample size n to $n \ln n$ in the error rate expression.

B. Approximation: the general recipe

We emphasize that the observed *effective sample size enlargement* here is another manifestation of the recently discovered phenomenon in functional estimation of high dimensional objects. There has been a recent wave of study on functional estimation of high dimensional parameters [1], [5]–[7], and it was shown in Jiao *et al.* [1] that for a wide class of functional estimation problems (including Shannon entropy $H(P) =$

This work was supported in part by the NSF Center for Science of Information under grant agreement CCF-0939370. J. Jiao's work is partially supported by a 3Com Stanford Graduate Fellowship.

$\sum_{i=1}^{S} -p_i \ln p_i$, $F_\alpha \triangleq \sum_{i=1}^{S} p_i^\alpha$, and mutual information), there exists a general methodology, termed *Approximation*, that can be applied to design minimax rate-optimal estimators whose performance with n samples is essentially that of the MLE (maximum likelihood estimator, or the plug-in approach) with $n \ln n$ samples.

The general methodology of *Approximation* in [1] is as follows. Consider estimating $G(\theta)$ of a parameter $\theta \in \Theta \subset \mathbb{R}^p$ for an experiment $\{P_\theta : \theta \in \Theta\}$, with a consistent estimator $\hat{\theta}_n$ for θ, where n is the number of observations. Suppose the functional $G(\theta)$ is analytic[1] everywhere except at $\theta \in \Theta_0$. A natural estimator for $G(\theta)$ is $G(\hat{\theta}_n)$, and we know from classical asymptotics [8, Lemma 8.14] that given the benign LAN (Local Asymptotic Normality) condition [8], $G(\hat{\theta}_n)$ is asymptotically efficient for $G(\theta)$ for $\theta \notin \Theta_0$ if $\hat{\theta}_n$ is asymptotically efficient for θ. In the estimation of functionals of discrete distributions, Θ is the S-dimensional probability simplex, and a natural candidate for $\hat{\theta}_n$ is the empirical distribution, which is unbiased for any $\theta \in \Theta$.

We propose to conduct the following two-step procedure in estimating $G(\theta)$.

1) **Classify the Regime**: Compute $\hat{\theta}_n$, and declare that we are in the "non-smooth" regime if $\hat{\theta}_n$ is "close" enough to Θ_0. Otherwise declare we are in the "smooth" regime;
2) **Estimate**:
   - If $\hat{\theta}_n$ falls in the "smooth" regime, use an estimator "similar" to $G(\hat{\theta}_n)$ to estimate $G(\theta)$;
   - If $\hat{\theta}_n$ falls in the "non-smooth" regime, replace the functional $G(\theta)$ in the "non-smooth" regime by an approximation $G_{\text{appr}}(\theta)$ (another functional) which can be estimated without bias, then apply an unbiased estimator for the functional $G_{\text{appr}}(\theta)$.

Approaches of this nature appeared before [1] in Lepski, Nemirovski, and Spokoiny [9], Cai and Low [10], Vinck *et al.* [11], Valiant and Valiant [4], developed independently for entropy estimation by Wu and Yang [6], and later utilized by Acharya *et al.* [7]. However, we emphasize that in all the examples above except for the L_1 distance estimator in Valiant and Valiant [4], the functionals considered all take the form $G(\sum_{i=1}^{p} f(\theta_i))$ or $G(\int f(p(x))dx)$, where $p(x)$ is a *univariate* density or function, and each $\theta_i \in \mathbb{R}$. In particular, the functions $f(\cdot)$ considered are everywhere analytic except at zero, e.g., $x^\alpha, |x|^\alpha$ for $\alpha > 0$ and $x \ln x$. Most of these features are violated in the L_1 distance estimation problem. If we write $L_1(P,Q) = \sum_{i=1}^{S} f(p_i, q_i)$ with $f(x,y) = |x-y| \in C([0,1]^2)$, then we have:

1) a *bivariate* function $f(x,y)$ in the sum;
2) a function $f(x,y)$ which is analytic except on a *segment* $x = y \in [0,1]$.

As discussed in Jiao *et al.* [1], approximation of multivariate functions is much more involved than that of univariate functions, and the fact that the "non-smooth" regime is around a line segment here makes the application of the *Approximation* approach quite difficult: what shape should we use to specify the "non-smooth" regime? We provide a comprehensive answer to this problem in this paper, thereby substantially generalizing the applicability of the *Approximation* methodology and demonstrate the intricacy of functional estimation problems in high dimensions.

The rest of the paper is organized as follows. In Section II and III, we present a thorough performance analysis of the MLE and explicitly construct the minimax rate-optimal estimators, where Section II covers the known Q case and Section III generalizes to the unknown Q case. Discussions in Section IV highlight the significance and novelty of our approaches by reviewing several other approaches which are shown to be suboptimal. We adopt the following notation for positive sequences $\{a_n\}, \{b_n\}$: $a_n \lesssim b_n$ means $\sup_n a_n/b_n < \infty$, $a_n \gtrsim b_n$ means $b_n \lesssim a_n$, $a_n \asymp b_n$ means $a_n \lesssim b_n$ and $a_n \gtrsim b_n$. We omit the proofs due to space limitations, and refer the readers to the full version [12] for details.

II. Divergence Estimation with Known Q

First we consider the case where $Q = (q_1, \cdots, q_S)$ is known to us while P is an unknown distribution with support $\{1, 2, \cdots, S\}$. In other words, $\mathcal{P} = \mathcal{M}_S$ and $\mathcal{Q} = \{Q\}$. We analyze the performance of the MLE in this case, and construct the approximation-based minimax rate-optimal estimator.

A. Performance of the MLE

The MLE serves as a natural estimator for the L_1 distance which can be expressed as $L_1(P_n, Q) = \sum_{i=1}^{S} |\hat{p}_i - q_i|$, where $P_n = (\hat{p}_1, \hat{p}_2, \cdots, \hat{p}_S)$ is the empirical distribution with $n\hat{p}_i \sim \mathsf{B}(n, p_i)$. We analyze the bias and the variance of $L_1(P_n, Q)$ separately. For the variance, the bounded difference inequality [13] gives $\mathsf{Var}(L_1(P_n, Q)) \le 1/n$, independent of Q. For the bias, since $|\hat{p}_i - q_i|$ is almost unbiased in estimating $|p_i - q_i|$ when p_i is far away from q_i, the case $P = Q$ is almost the worst case. The following lemma provides sharp estimates on $\mathbb{E}|\hat{q} - q|$, where $n\hat{q} \sim \mathsf{B}(n, q)$.

Lemma 1 *[14] For $n\hat{q} \sim \mathsf{B}(n, q)$, we have*

$$\mathbb{E}|\hat{q} - q| \le \sqrt{\frac{q(1-q)}{n}}. \tag{6}$$

Furthermore, if $\frac{1}{n} \le q \le 1 - \frac{1}{n}$, there is also a lower bound

$$\mathbb{E}|\hat{q} - q| \ge \sqrt{\frac{q(1-q)}{2n}}. \tag{7}$$

Based on this lemma, we obtain both the upper and lower bounds for the mean squared error of $L_1(P_n, Q)$.

Theorem 1 *The maximum likelihood estimator satisfies*

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P |L_1(P_n, Q) - L_1(P, Q)|^2 \le \frac{(F_{1/2}(Q))^2 + 1}{n} \tag{8}$$

[1] A function f is analytic at a point x_0 if and only if its Taylor series about x_0 converges to f in some neighborhood of x_0.

where $F_{1/2}(Q) = \sum_{i=1}^{S} \sqrt{q_i}$. Moreover, if $q_i \in [1/n, 1-1/n]$ for all $i = 1, 2, \cdots, S$, there is also a lower bound

$$\sup_{P\in\mathcal{M}_S} \mathbb{E}_P |L_1(P_n, Q) - L_1(P,Q)|^2 \geq \frac{(F_{1/2}(Q))^2}{2n}. \qquad (9)$$

Evidently, the mean squared error of the MLE is of the order $(F_{1/2}(Q))^2/n$, which is closely related to the order-$(1/2)$ power sum of the known distribution Q. The following corollary is straightforward since $F_{1/2}(Q) \leq \sqrt{S}$.

Corollary 1 *If $n \gtrsim S$, we have*

$$\sup_{P,Q\in\mathcal{M}_S} \mathbb{E}_P |L_1(P_n, Q) - L_1(P,Q)|^2 \asymp \frac{S}{n}. \qquad (10)$$

Hence, it is necessary and sufficient for the MLE to have $n \gg S$ samples to be consistent, and we note that the necessity can also be derived using the result that the empirical distribution requires $n \gg S$ samples to have a vanishing L_1 risk in estimating the true distribution [15].

B. Construction of the optimal estimator

Since it is by now widely established that the MLE is usually strictly suboptimal in estimating non-smooth functionals of high-dimensional parameters [1], [5]–[7], now we apply our general recipe to construct the minimax rate-optimal estimator.

First we classify regimes. For $f(x,q) = |x-q|$, we are in the "non-smooth" regime if and only if the empirical probability $\hat{p}$ is close to q, or equivalently, $\hat{p} \in U(q)$ for some "uncertainty set" $U(q)$ containing q. A natural question arises: how do we measure the closeness, or equivalently, how do we determine $U(q)$? Our answer is that, the uncertainty set $U(q_i)$ should be chosen such that the unknown parameter p can be localized within $U(q)$. More precisely, $p \in U(q)$ ensures that $\hat{p} \in \tilde{U}(q)$ with overwhelming probability, where $\tilde{U}(q) \triangleq q + 2(U(q) - q)$ is two times larger, and vice versa by exchanging U and $\tilde{U}$. Mathematically, we require

$$\max\{\sup_{p\in U(q)} \mathbb{P}_p(\hat{p} \notin \tilde{U}(q)), \sup_{p\notin\tilde{U}(q)} \mathbb{P}_p(\hat{p} \in U(q))\} \lesssim n^{-A}$$

for some large universal constant A. In our setting, $n\hat{p} \sim \mathsf{B}(n,p)$, and the Binomial tail bounds yield the choice

$$U(x) = \begin{cases} [0, \frac{c_1 \ln n}{n}], & x \leq \frac{c_1 \ln n}{n} \\ [x - \sqrt{\frac{c_1 x \ln n}{n}}, x + \sqrt{\frac{c_1 x \ln n}{n}}], & x > \frac{c_1 \ln n}{n} \end{cases} \qquad (11)$$

for some universal constant $c_1 > 0$ depending on A only.

Now we construct the estimator. In the "smooth" regime, i.e., $\hat{p} \notin U(q)$, we simply employ the MLE $|\hat{p} - q|$ to estimate $f(p,q)$. In the "non-smooth" regime, we need to approximate $f(p,q)$ by another functional which can be estimated without bias. In our Binomial model $n\hat{p} \sim \mathsf{B}(n,p)$, the only functional of p which can be estimated without bias is a polynomial of p with degree no more than n. Hence, we consider the best polynomial approximation of $f(x,q)$ on $U(q)$, which is defined as

$$P_K(x;q) = \arg\min_{P\in\mathsf{Poly}_K} \max_{z\in U(q)} |f(z,q) - P(z)| \qquad (12)$$

where Poly_K denotes the set of polynomials with degree no more than K. Once we obtain $P_K(x;q)$, we can use an unbiased estimate $\tilde{P}_K(\hat{p};q)$ such that $\mathbb{E}\tilde{P}_K(\hat{p};q) = P_K(p;q)$ for $n\hat{p} \sim \mathsf{B}(n,p)$. As a result, the bias of the estimator $\tilde{P}_K(\hat{p};q)$ in the "non-smooth" regime is exactly the approximation error of $P_K(x;q)$ in approximating $f(x,q) = |x-q|$ on $U(q)$, which can be significantly smaller than the MLE. The following lemma gives the bias and variance bound of $\tilde{P}_K(\hat{p};q)$.

Lemma 2 *For $n\hat{p} \sim \mathsf{B}(n,p)$ with $p \in U(q)$, we have*

$$|\mathbb{E}\tilde{P}_K(\hat{p};q) - |p-q|| \lesssim \frac{1}{K}\sqrt{\frac{q\ln n}{n}} \qquad (13)$$

$$\mathsf{Var}(\tilde{P}_K(\hat{p};q)) \lesssim \frac{B^K(\ln n)^2}{n}(p+q) \qquad (14)$$

for some universal constant $B > 0$.

Hence, to balance the bias and the variance, the approximation order should be set as $K \asymp \ln n$, and then the bias is of the order $\sqrt{q/(n\ln n)}$, a logarithmic improvement compared with Lemma 1. In summary, we have the following construction.

Estimator Construction 1 *Randomly split the samples into two half samples with equal size, and obtain the corresponding empirical probabilities $\hat{p}_{i,1}, \hat{p}_{i,2}$ for each $i = 1, \cdots, S$. We use the first half samples to classify regimes and the second half samples for estimation, i.e., $\hat{L}^{(1)} = \sum_{i=1}^{S}[|\hat{p}_{i,2} - q_i|\mathbb{1}(\hat{p}_{i,1} \notin U(q_i)) + \tilde{P}_K(\hat{p}_{i,2}; q_i)\mathbb{1}(\hat{p}_{i,1} \in U(q_i))]$, where $U(q_i)$ is given by (11), $K = c_2 \ln n$, and $c_1, c_2 > 0$ are properly chosen universal constants.*

The performance of this estimator is as follows.

Theorem 2 *For $\ln n \lesssim \ln F_{1/2}(Q)$, we have*

$$\sup_{P\in\mathcal{M}_S} \mathbb{E}_P |\hat{L}^{(1)} - L_1(P,Q)|^2 \lesssim \frac{[F_{1/2}(Q)]^2}{n\ln n}. \qquad (15)$$

Corollary 2 *For $\ln n \lesssim \ln S$, we have*

$$\sup_{P,Q\in\mathcal{M}_S} \mathbb{E}_P |\hat{L}^{(1)} - L_1(P,Q)|^2 \lesssim \frac{S}{n\ln n}. \qquad (16)$$

Hence, together with Theorem 1 and Corollary 1, the estimator $\hat{L}^{(1)}$ has a logarithmic improvement over the MLE.

C. Minimax lower bound

Now we prove a minimax lower bound for estimating $L_1(P,Q)$ via the dual of best polynomial approximation and show that $\hat{L}^{(1)}$ is indeed minimax rate-optimal. The main idea is the so-called fuzzy hypothesis testing [16], i.e., we construct two priors μ_1, μ_2 satisfying the following two key ingredients:

1) *Functional value separation*: the expected functional values under $P \sim \mu_1$ and $P \sim \mu_2$ are quite different, i.e., $|\mathbb{E}_{\mu_1} L_1(P,Q) - \mathbb{E}_{\mu_2} L_1(P,Q)|$ is large.
2) *Indistinguishability*: the marginal distributions F_i of the observations generated by $P \sim \mu_i$ are close to each other, i.e., the total variation distance $\mathsf{TV}(F_1, F_2)$ is small.

Specifically, we choose $Q = (1/S, \cdots, 1/S)$ and $\mu_i = \nu_i^{\otimes S}$ as product priors that generate a valid pmf with high probability [6], where $\mathsf{supp}(\nu_i) \subset U(1/S)$. If $\mathbb{E}_{\nu_1} x^l = \mathbb{E}_{\nu_2} x^l$ for $l = 0, 1, \cdots, K \asymp \ln n$, we can obtain that $\mathsf{TV}(F_1, F_2) \lesssim n^{-2}$. Hence, it suffices to maximize $|\mathbb{E}_{\nu_1}|x - S^{-1}| - \mathbb{E}_{\nu_2}|x - S^{-1}||$ subject to the previous moment constraints. Duality gives

$$\sup_{\substack{\nu : \|\nu\|_{\mathsf{TV}} \le 1 \\ \int x^l \nu(dx) = 0, l = 0, 1, \cdots, K}} \left| \int f(x) \nu(dx) \right| = \inf_{p \in \mathsf{Poly}_K} \|f - p\|_\infty,$$

which is the best polynomial approximation error of degree K and is $\asymp 1/\sqrt{Sn \ln n}$ for each symbol. The final lower bound follows from a detailed analysis of μ_i.

Theorem 3 *For $\ln n \lesssim \ln S$ and $S \lesssim n \ln n$, we have*

$$\inf_{\hat{L}} \sup_{P, Q \in \mathcal{M}_S} |\hat{L} - L_1(P, Q)|^2 \gtrsim \frac{S}{n \ln n} \tag{17}$$

where the infimum is taken over all possible estimators.

Hence, a combination of Corollary 2 and Theorem 3 shows that the estimator $\hat{L}^{(1)}$ is minimax rate-optimal.

III. Divergence Estimation with Unknown Q

Now we consider the general case where both P and Q are unknown to us, i.e., $\mathcal{P} = \mathcal{Q} = \mathcal{M}_S$.

A. Performance of the MLE

In this case, the MLE is expressed as $L_1(P_n, Q_n) = \sum_{i=1}^S |\hat{p}_i - \hat{q}_i|$. Since $|L_1(P_n, Q_n) - L_1(P, Q)| \le L_1(P_n, P) + L_1(Q_n, Q)$ by the triangle inequality, Lemma 1 can again be applied here to give the performance of the MLE.

Theorem 4 *If $n \gtrsim S$, the MLE satisfies*

$$\sup_{P, Q \in \mathcal{M}_S} |L_1(P_n, Q_n) - L_1(P, Q)|^2 \asymp \frac{S}{n}. \tag{18}$$

Hence, the MLE achieves the mean squared error S/n, and requires $n \gg S$ samples to be consistent.

B. Construction of the optimal estimator

Again we apply our general recipe to construct the optimal estimator, but encounter several new difficulties: $f(x, y) = |x - y|$ is not analytic on a segment, and both the uncertainty set and the polynomial approximation need to be generalized to the 2D case. We will overcome these obstacles step by step.

As usual, first we classify "smooth" and "non-smooth" regimes. Since the function $f(x, y) = |x - y| \in C([0, 1]^2)$ is not analytic on the segment $x = y \in [0, 1]$, we are looking for the "uncertainty set" U containing this segment such that any $(p, q) \in U$ can be "localized" in the previous sense. Applying the Binomial tail bounds again yields

$$U = \cup_{x \in [0,1]} U(x) \times U(x) \tag{19}$$

where $U(x)$ is given by (11). As a result, we declare that we are in the "non-smooth" regime if and only if $(\hat{p}, \hat{q}) \in U$.

Now we construct the estimator. In the "smooth" regime $(\hat{p}, \hat{q}) \notin U$, we employ the MLE $|\hat{p} - \hat{q}|$ as before. In the "non-smooth" regime, the previous example seems to suggest that we consider the best polynomial approximation of $f(x, y) = |x - y|$ on U. However, this will not work for two reasons:

1) the entire 2D stripe U is too large for the polynomial approximation error to vanish at the correct rate;
2) best polynomial approximation in the 2D case is not unique, and may not achieve the desired pointwise error.

We will explore these reasons in more detail in Section IV. To solve the first problem, we remark that although U is the set such that its element can be localized within U, a specific element $(x, y) \in U$ can be localized in a smaller subset $U(x, y) = U(\frac{x+y}{2}) \times U(\frac{x+y}{2}) \subset U$, where $U(x)$ is given by (11). Hence, for the observation $(\hat{p}, \hat{q})$, we should consider a polynomial $P_K(x, y; \hat{p}, \hat{q})$ with degree K to approximate $f(x, y)$ on $U(\hat{p}, \hat{q})$, and then use an unbiased estimate $\tilde{P}_K(x, y; \hat{p}, \hat{q})$ of $P_K(x, y; \hat{p}, \hat{q})$ for estimation.

For the second problem, we need to find a suitable polynomial $P_K(x, y; \hat{p}, \hat{q})$ with satisfactory approximation properties. The answer is as follows: if $(\hat{p} + \hat{q})/2 > c_1 \ln n/n$, we use $P_K(x, y; \hat{p}, \hat{q}) = Q_K(x - y; \sqrt{2(\hat{p} + \hat{q}) \ln n/n})$, where

$$Q_K(t; s) = \arg \min_{P \in \mathsf{Poly}_K} \max_{z \in [-s, s]} ||z| - P(z)| \tag{20}$$

is the 1D polynomial approximation of $|t|$ in $[-s, s]$ by the variable substitution $t = x - y, s = \sqrt{2(\hat{p} + \hat{q}) \ln n/n}$, which is validated by the fact $f[U(\hat{p}, \hat{q})] \subset [-s, s]$. If $(\hat{p} + \hat{q})/2 \le c_1 \ln n/n$, we consider the decomposition $|x - y| = (\sqrt{x} + \sqrt{y})|\sqrt{x} - \sqrt{y}|$ and use $P_K(x, y; \hat{p}, \hat{q}) = u_K(x, y) v_K(x, y)$, where u_K, v_K are the degree-K best approximating polynomial of $\sqrt{x} + \sqrt{y}$ and $|\sqrt{x} - \sqrt{y}|$, respectively, on $U(\hat{p}, \hat{q}) = [0, c_1 \ln n/n]^2$. In practice, u_K and v_K can be replaced by the efficiently implementable lowpass filtered Chebyshev expansion [17], which achieves the same error rate as the best polynomial approximation. The performance of $\tilde{P}_K(x, y; \hat{p}, \hat{q})$ is presented in the following lemma.

Lemma 3 *For $n(\hat{p}_1, \hat{q}_1), n(\hat{p}_2, \hat{q}_2) \sim \mathsf{B}(n, p) \times \mathsf{B}(n, q)$ with $(p, q) \in U$ and $(\hat{p}_1, \hat{q}_1)$ independent with $(\hat{p}_2, \hat{q}_2)$, we have*

$$|\mathbb{E}\tilde{P}_K(\hat{p}_2, \hat{q}_2; \hat{p}_1, \hat{q}_1) - |p - q|| \lesssim \frac{1}{K} \sqrt{\frac{\ln n}{n}} (\sqrt{p} + \sqrt{q}) \tag{21}$$

$$\mathsf{Var}(\tilde{P}_K(\hat{p}_2, \hat{q}_2; \hat{p}_1, \hat{q}_1)) \lesssim \frac{B^K (\ln n)^2}{n} (p + q) \tag{22}$$

for some universal constant $B > 0$.

Hence, we still choose $K \asymp \ln n$ to balance the bias and the variance, and construct the estimator as follows.

Estimator Construction 2 *As before, use sample splitting to obtain $(\hat{p}_{i,1}, \hat{q}_{i,1})$ and $(\hat{p}_{i,2}, \hat{q}_{i,2})$. The estimator is defined as $\hat{L}^{(2)} = \sum_{i=1}^S [|\hat{p}_{i,2} - \hat{q}_{i,2}| \mathbb{1}((\hat{p}_{i,1}, \hat{q}_{i,1}) \notin U) + \tilde{P}_K(\hat{p}_{i,2}, \hat{q}_{i,2}; \hat{p}_{i,1}, \hat{q}_{i,1}) \mathbb{1}((\hat{p}_{i,1}, \hat{q}_{i,1}) \in U)]$, where U is given by (19), $K = c_2 \ln n$, and $c_1, c_2 > 0$ are properly chosen universal constants.*

The next theorem presents the performance of $\hat{L}^{(2)}$.

Theorem 5 *For $\ln n \lesssim \ln S$, we have*

$$\sup_{P, Q \in \mathcal{M}_S} |\hat{L}^{(2)} - L_1(P, Q)|^2 \lesssim \frac{S}{n \ln n}. \tag{23}$$

Since the lower bound for the known Q case also serves as a lower bound for the general case, Theorem 3 and Theorem 5 yield that $\hat{L}^{(2)}$ is minimax rate-optimal. Note that $\hat{L}^{(2)}$ achieves the minimax rate without knowing the support size S a priori. Moreover, the *effective sample size enlargement* effect holds again: the performance of the optimal estimator with n samples is essentially that of the MLE with $n \ln n$ samples.

IV. Comparison with Other Approaches

In this section, we review some other possible approaches in estimating the L_1 distance, and apply approximation theory to argue the strict suboptimality of some approaches.

A. Approximation only around the origin

In the previous papers [1], [4]–[7] in estimating entropy, power sum, mutual information, etc, the approximation methodology is conducted only around the origin. However, we remark that this is insufficient in estimating the L_1 distance. Consider the known Q case with $S \ll n/\ln n$ and $P = Q$ uniform, since we are only using approximation for $\hat{p} \in [0, c_1 \ln n/n]$, we will use the plug-in approach in this case. Then the lower bound in Lemma 1 shows that the mean squared error of this estimator is lower bounded by S/n, which is worse than the optimal rate $S/(n \ln n)$. This is exactly the reason why the estimator of Valiant and Valiant [4] can only achieve the optimal error rate when $n \lesssim S \lesssim n \ln n$, but ours merely requires $\ln n \lesssim \ln S$ to achieve the optimal error rate.

B. One-dimensional approximation in the 2D case

In the construction of $\hat{L}^{(2)}$, we split into two cases when $(\hat{p}, \hat{q}) \in U$, i.e., 1D approximation of $|t|$ via the substitution $t = x - y$ if $(\hat{p} + \hat{q})/2 > c_1 \ln n/n$, and the decomposition of $|x - y|$ into $(\sqrt{x} + \sqrt{y})|\sqrt{x} - \sqrt{y}|$ otherwise. Can we always do 1D approximation of $|t|$ with $t = x - y$ to achieve the desired approximation error, i.e., propose some $P(t) \in \mathsf{Poly}_K$ with $K \asymp \ln n$ and $|P(t) - |t|| \lesssim \sqrt{t/(n \ln n)}$ for any $|t| \le c_1 \ln n/n$? We have a proposition for approximating $|t|$ [18].

Proposition 1 *If $Q_K(t) \in \mathsf{Poly}_K$ is even with $Q_K(0) = 0$, and achieves the best uniform error rate $\max_{t \in [-1,1]} |Q_K(t) - |t|| \lesssim 1/K$, we have*

$$\limsup_{K \to \infty} \frac{1}{K} \sup_{0 < |t| \le 1/K} \frac{|t| - |Q_K(t) - |t||}{t^2} < \infty. \tag{24}$$

Obviously $P(0) = 0$ and achieves the best uniform bound, and by $\tilde{P}(t) = (P(t) + P(-t))/2$ we can get $P(t)$ even. Then Proposition 1 gives $|P(t) - |t|| \ge |t| - Cnt^2$ for $|t| \lesssim 1/n$, contradicting the upper bound when $1/(n \ln n) \ll t \ll 1/n$. Hence, any 1D approximation does not work in this case!

C. Approximation on the entire 2D stripe

In the unknown Q case we have decomposed the stripe U into subsets where polynomial approximations take place. Is it possible that we use a single polynomial $P(x, y)$ of degree $K \asymp \ln n$ to approximate $|x - y|$ such that $|P(x, y) - |x - y|| \lesssim \sqrt{(x + y)/(n \ln n)}$ for any $(x, y) \in U$? We prove that the answer is negative even for $U' = \cup_{x \in [c_1 \ln n/n, t_n]} U(x) \times U(x) \subset U$ and any $t_n \gg (\ln n)^3/n$.

Proposition 2 *If $(\ln n)^3/n \ll t_n \le 1/2$, $K \asymp \ln n$, we have*

$$\liminf_{n \to \infty} \sqrt{n \ln n} \cdot \inf_{P \in \mathsf{Poly}_K} \sup_{(x,y) \in U'} \frac{|P(x, y) - |x - y||}{\sqrt{x + y}} = +\infty.$$

Proposition 2 shows that the subset $U(c_1 \ln n/n, c_1 \ln n/n)$ of U is the correct set to approximate $|x - y|$ over when our observation $(\hat{p}, \hat{q})$ is in it. For a too large set U' (e.g., $U' = U$), every polynomial fails to achieve the desired approximation error bound $\sqrt{(x + y)/(n \ln n)}$.

D. The estimator in Valiant and Valiant [4]

The estimator for the L_1 distance in Valiant and Valiant [4] also achieves the optimal MSE $S/(n \ln n)$ for $n \lesssim S \lesssim n \ln n$. Our estimator, as an linear estimator in the language of [4], improves over [4] in two aspects: it achieves the optimal error rate in more general cases by approximating over the whole non-smooth segment, and achieves a tighter upper bound $\sqrt{(p + q)/(n \ln n)}$ by a better polynomial approximation (sharper than the bound $\sqrt{S/(n \ln n)}(p + q + 1/S)$ in [4]).

References

[1] J. Jiao, K. Venkat, Y. Han, and T. Weissman, "Minimax estimation of functionals of discrete distributions," *Information Theory, IEEE Transactions on*, vol. 61, no. 5, pp. 2835–2885, 2015.

[2] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.

[3] L. Devroye, L. Györfi, and G. Lugosi, "A probabilistic theory of pattern recognition," 1996.

[4] G. Valiant and P. Valiant, "The power of linear estimators," in *Foundations of Computer Science (FOCS), 2011 IEEE 52nd Annual Symposium on.* IEEE, 2011, pp. 403–412.

[5] P. Valiant and G. Valiant, "Estimating the unseen: improved estimators for entropy and other properties," in *Advances in Neural Information Processing Systems*, 2013, pp. 2157–2165.

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

[7] J. Acharya, A. Orlitsky, A. T. Suresh, and H. Tyagi, "The complexity of estimating Rényi entropy." SODA, 2015.

[8] A. W. Van der Vaart, *Asymptotic statistics*. Cambridge university press, 2000, vol. 3.

[9] O. Lepski, A. Nemirovski, and V. Spokoiny, "On estimation of the L_r norm of a regression function," *Probability theory and related fields*, vol. 113, no. 2, pp. 221–253, 1999.

[10] 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.

[11] M. Vinck, F. P. Battaglia, V. B. Balakirsky, A. H. Vinck, and C. M. Pennartz, "Estimation of the entropy based on its polynomial representation," *Physical Review E*, vol. 85, no. 5, p. 051139, 2012.

[12] J. Jiao, Y. Han, and T. Weissman, "Minimax estimation of divergence functions," *in preparation*.

[13] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

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

[15] Y. Han, J. Jiao, and T. Weissman, "Minimax estimation of discrete distributions under l_1 loss," *Information Theory, IEEE Transactions on*, vol. 61, no. 11, pp. 6343–6354, Nov 2015.

[16] A. Tsybakov, *Introduction to Nonparametric Estimation*. Springer-Verlag, 2008.

[17] H. N. Mhaskar, P. Nevai, and E. Shvarts, "Applications of classical approximation theory to periodic basis function networks and computational harmonic analysis," *Bulletin of Mathematical Sciences*, vol. 3, no. 3, pp. 485–549, 2013.

[18] Y. Han, J. Jiao, and T. Weissman, "Are the usual pointwise bounds in approximation theory optimal?" *in preparation*, 2016.