Nonparametric estimation of the likelihood ratio and divergence functionals

XuanLong Nguyen[1], Martin J. Wainwright[1,2] and Michael I. Jordan[1,2]
[1] Department of Electrical Engineering and Computer Science
[2] Department of Statistics
University of California, Berkeley
{xuanlong,wainwrig,jordan}@eecs.berkeley.edu

***Abstract*— We develop and analyze a nonparametric method for estimating the class of f-divergence functionals, and the density ratio of two probability distributions. Our method is based on a non-asymptotic variational characterization of the f-divergence, which allows us to cast the problem of estimating divergences in terms of risk minimization. We thus obtain an M-estimator for divergences, based on a convex and differentiable optimization problem that can be solved efficiently We analyze the consistency and convergence rates for this M-estimator given conditions only on the ratio of densities.**

I. Introduction

Given samples from two (multivariate) probability distributions $\mathbb{P}$ and $\mathbb{Q}$, it is frequently of interest to estimate the values of functionals measuring the divergence between the unknown $\mathbb{P}$ and $\mathbb{Q}$. Of particular interest is the Kullback-Leibler (KL) divergence, but the approach of this paper applies to the more general class of Ali-Silvey or f-divergences [1], [7]. An f-divergence, to be defined formally in the sequel, is of the form $D_\phi(\mathbb{P}, \mathbb{Q}) = \int \phi(d\mathbb{Q}/d\mathbb{P}) d\mathbb{P}$, where ϕ is a convex function of the likelihood ratio.

These divergences play fundamental roles in statistics and information theory. In particular, divergences are often used as measures of discrimination in binary hypothesis testing and classification applications. Examples include signal selection [12] and decentralized detection [15], where f-divergences are used to solve experimental design problems. The Shannon mutual information (a particular type of KL divergence), in addition to its role in coding theorems, is often used as a measure of independence to be extremized in dimensionality reduction and feature selection. In all of these cases, if divergences are to be used as objective functionals, one has to be able to estimate them efficiently from data.

There are two ways in which divergences are typically characterized. The classical characterization is an asymptotic one; for example, the KL divergence emerges as the asymptotic rate of the probability of error in Neyman-Pearson binary hypothesis testing (a result known as Stein's lemma). But it is also possible to provide non-asymptotic characterizations of the divergences; in particular, Fano's lemma shows that KL divergence provides a lower bound on the error probability for decoding error [6]. This paper presents a method for divergence estimation, motivated by a non-asymptotic characterization of f-divergence, in the spirit of Fano's lemma, first explicated in our earlier work [15]. This characterization states that that there is an one-to-one correspondence between the family of f-divergences and the family of "surrogate loss functions", such that the (optimum) Bayes risk is equal to the negative of the divergence. In other words, any negative f-divergence can serve as a lower bound of a risk minimization problem. This *variational characterization* of divergence, stated formally in Lemma 1, allows us estimate a divergence $D_\phi(\mathbb{P}, \mathbb{Q})$ by solving a binary decision problem. Not surprisingly, we show how the problem of estimating f-divergence is intrinsically linked to that of estimating the likelihood ratio $g_0 = d\mathbb{P}/d\mathbb{Q}$. Overall, we obtain an M-estimator, whose optimal value estimates the divergence and optimizing argument estimates the likelihood ratio.

Our estimator is nonparametric, in that it imposes no strong assumptions on the form of the densities for $\mathbb{P}$ and $\mathbb{Q}$. We establish consistency of this estimator by exploiting analysis techniques for M-estimators in the setting of nonparametric density estimation and regression [19], [21]. At a high level, the key to the proof is suitable control on the modulus of continuity of the suprema of two empirical processes, one for each of $\mathbb{P}$ and $\mathbb{Q}$, with respect to a metric defined over density ratios. This metric turns out to be a surrogate lower bound of a Bregman divergence defined on a pair of density ratios. In this way, we not only establish consistency of

our estimator, but also obtain convergence rates. As one concrete example, when the likehood ratio g_0 lies in a function class $\mathcal{G}$ of smoothness α with $\alpha > d/2$, where d is the number of data dimensions, our estimator of the likelihood ratio achieves the optimal minimax rate $n^{-\alpha/(2\alpha+d)}$ according to the Hellinger metric, while the divergence estimator achieves the rate $n^{-2\alpha/(2\alpha+d)}$.

In abstract terms, an f-divergence can be viewed as an integral functional of a pair of densities. While there relatively little work focusing on integral functionals for pairs of densities (such as the f-divergences of interest here), there is an extensive literature on the estimation of an integral functionals of the form $\int \phi(p)p$, where p the density of an unknown probability distribution. Work on this topic dates back to the 1970s [10], [14]; see also [3], [4], [13] and the references therein. There are also a number of papers that focus specifically on the entropy functional (see, e.g., [8], [11], [9]).

In a separate line of work, Wang et al. [22] proposed a histogram-based KL estimator. Their method is based on the estimation of the likelihood ratio by building partitions of equivalent (empirical) $\mathbb{Q}$-measure. The estimator is easy to implement, and was empirically shown to outperform direct plug-in methods, but no theoretical convergence rate analysis was given. A concern with histogram-based methods are their possibly inefficiency, in both statistical and computational terms, when applied to higher dimensional data. Our preliminary experimental results [16] suggest that our estimator exhibits comparable or superior convergence rates in a number of examples.

The remainder of this paper is organized as follows. In Section II, we describe a general variational characterization of f-divergence, and derive an M-estimator for the KL divergence and the likelihood ratio. Section III is devoted to the analysis of consistency and convergence rates. In Section IV we briefly discuss how our analysis extends to general f-divergences. Additional results and complete proofs of all theorems can be be found in the technical report [16].

II. M-ESTIMATOR FORMULATION

We begin by describing how the estimation of KL divergence and the density ratio can be formulated as an M-estimator.

A. Variational characterization of f-divergence

Let $X_1, \ldots, X_n$ be n i.i.d. random variables drawn from an unknown distribution $\mathbb{P}$; similarly, let $Y_1, \ldots, Y_n$ be n random variables drawn from an unknown distribution $\mathbb{Q}$. We assume that both are absolutely continuous with respect to Lebesgue measure μ, with positive densities p_0 and q_0, respectively, on some compact domain $\mathcal{X} \subset \mathbb{R}^d$. The KL divergence between $\mathbb{P}$ and $\mathbb{Q}$ is defined as:

$$D_K(\mathbb{P}, \mathbb{Q}) = \int p_0 \log(p_0/q_0)\, d\mu.$$

The KL divergence is a special case of a broader class of divergences known as Ali-Silvey distances, or f-divergences [7], [1]:

$$D_\phi(\mathbb{P}, \mathbb{Q}) = \int p_0 \phi(q_0/p_0)\, d\mu,$$

where $\phi : \mathbb{R} \to \bar{\mathbb{R}}$ is a convex function. Different choices of ϕ result in many divergences that play important roles in information theory and statistics, including the variational distance, Hellinger distance, KL divergence and so on (see, e.g., [18]).

Since ϕ is a convex function, by Legendre-Fenchel convex duality [17] we can write:

$$\phi(u) = \sup_{v \in \mathbb{R}} (uv - \phi^*(v)),$$

where ϕ^* is the convex conjugate of ϕ. As a result,

$$\begin{aligned} D_\phi(\mathbb{P}, \mathbb{Q}) &= \int p_0 \sup_f (f q_0/p_0 - \phi^*(f))\, d\mu \\ &= \sup_f \left(\int f\, d\mathbb{Q} - \int \phi^*(f)\, d\mathbb{P} \right), \end{aligned}$$

where the supremum is taken over all measurable functions $f : \mathcal{X} \to \mathbb{R}$, and $\int f\, d\mathbb{P}$ denotes the expectation of f under distribution $\mathbb{P}$. It is easy to see that equality in the supremum is attained for functions f such that $q_0/p_0 \in \partial\phi^*(f)$, where q_0, p_0 and f are evaluated at any $x \in \mathcal{X}$. By convex duality, this is true if $f \in \partial\phi(q_0/p_0)$ for any $x \in \mathcal{X}$. Thus, we have proved the following lemma:

Lemma 1. *Letting $\mathcal{F}$ be any function class in $\mathcal{X} \to \mathbb{R}$, there holds:*

$$D_\phi(\mathbb{P}, \mathbb{Q}) \geq \sup_{f \in \mathcal{F}} \int f\, d\mathbb{Q} - \phi^*(f)\, d\mathbb{P}. \tag{1}$$

Furthermore, equality holds if $\mathcal{F} \cap \partial\phi(q_0/p_0) \neq \emptyset$.

B. An M-estimator of density ratio and KL divergence

Returning to the KL divergence, ϕ has the form $\phi(u) = -\log(u)$ for $u > 0$ and $+\infty$ for $u \leq 0$. The convex dual of ϕ is $\phi^*(v) = \sup_u(uv - \phi(u)) = -1 - \log(-v)$ if $u < 0$ and $+\infty$ otherwise. By Lemma 1,

$$\begin{aligned} D_K(\mathbb{P}, \mathbb{Q}) &= \sup_{f<0} \int f\, d\mathbb{Q} - \int (-1 - \log(-f))\, d\mathbb{P} \\ &= \sup_{g>0} \int \log g\, d\mathbb{P} - \int g d\mathbb{Q} + 1. \end{aligned} \tag{2}$$

In addition, the supremum is attained at $g_0 = p_0/q_0$. This motivates the following estimator of the KL divergence: Let $\mathcal{G}$ be a function class of $\mathcal{X} \to \mathbb{R}_+$, and $\int d\mathbb{P}_n$ and $\int d\mathbb{Q}_n$ denote the expectation under empirical measures $\mathbb{P}_n$ and $\mathbb{Q}_n$, respectively, and consider the following optimization problem:

$$\hat{D}_K = \sup_{g \in \mathcal{G}} \int \log g \; d\mathbb{P}_n - \int g d\mathbb{Q}_n + 1. \tag{3}$$

In practice we generally choose $\mathcal{G}$ to be a convex function class. In this case it turns out that the problem can be posed as a convex optimization problem that can be solved efficiently [16]. Suppose that the supremum is attained at $\hat{g}_n$. Then $\hat{g}_n$ is an M-estimator of the density ratio $g_0 = p_0/q_0$.

In the case of KL divergence estimation, we need to analyze the behavior of $|\hat{D}_K - D_K(\mathbb{P}, \mathbb{Q})|$ as $n \to \infty$. In the case of density ratio estimation, we also need a performance measure. Since $g_0 = p_0/q_0$ can be viewed as a density function with respect to $\mathbb{Q}$ measure, a natural metric is the Hellinger distance:

$$h^2_{\mathbb{Q}}(g, g_0) := \frac{1}{2} \int (g^{1/2} - g_0^{1/2})^2 \; d\mathbb{Q}. \tag{4}$$

As we shall see, this distance measure is weaker than $|\hat{D}_K - D_K(\mathbb{P}, \mathbb{Q})|$, with the advantage of allowing us to obtain guarantees under milder assumptions.

III. Consistency and convergence rates

In this section we shall present consistency results and obtain convergence rates for our estimators. Throughout the paper, we impose the following conditions on the distributions $\mathbb{P}, \mathbb{Q}$ and the function class $\mathcal{G}$.

(i) $D_K(\mathbb{P}, \mathbb{Q}) < \infty$; and
(ii) $\mathcal{G}$ is sufficiently rich so that $g_0 \in \mathcal{G}$.

In order to analyze the overall error, we define the *approximation error* $\mathcal{E}_0(\mathcal{G})$ and *estimation error* $\mathcal{E}_1(\mathcal{G})$ as follows:

$$\mathcal{E}_0(\mathcal{G}) = D_K(\mathbb{P}, \mathbb{Q}) - \sup_{g \in \mathcal{G}} \int (\log g \; d\mathbb{P} - g \; d\mathbb{Q} + 1) \geq 0$$

$$\mathcal{E}_1(\mathcal{G}) = \sup_{g \in \mathcal{G}} \left| \int \log g \; d(\mathbb{P}_n - \mathbb{P}) \; - g d(\mathbb{Q}_n - \mathbb{Q}) \right|.$$

Combining with (2) and (3) it is easy to see that:

$$-\mathcal{E}_1(\mathcal{G}) - \mathcal{E}_0(\mathcal{G}) \leq \hat{D}_K - D_K(\mathbb{P}, \mathbb{Q}) \leq \mathcal{E}_1(\mathcal{G}).$$

Since we have imposed condition (ii), the approximation error $\mathcal{E}_0(\mathcal{F})$ vanishes, so that this paper focuses on estimation error $\mathcal{E}_1(\mathcal{G})$ only. Note that if (ii) does not hold, we obtain instead a lower bound on the KL divergence.

A. Set-up and some basic inequalities

We begin by stating a few basic inequalities used throughout our analysis of consistency and convergence rates. We bound $\mathcal{E}_1(\mathcal{G})$ in terms of the following empirical processes:

$$v_n(\mathcal{G}) = \sup_{g \in \mathcal{G}} \left| \int \log \frac{g}{g_0} d(\mathbb{P}_n - \mathbb{P}) - \int (g - g_0) d(\mathbb{Q}_n - \mathbb{Q}) \right|.$$

$$w_n(g_0) = \left| \int \log g_0 \; d(\mathbb{P}_n - \mathbb{P}) \; - g_0 d(\mathbb{Q}_n - \mathbb{Q}) \right|.$$

Note that by construction, we have:

$$\mathcal{E}_1(\mathcal{G}) \leq v_n(\mathcal{G}) + w_n(g_0). \tag{5}$$

Our first lemma deals with the term w_n:

Lemma 2. *We have the almost-sure convergence* $w_n(g_0) \xrightarrow{a.s.} 0$.

Note that in this lemma and other theorems, all "a.s. convergence" statements can be understood with respect to either $\mathbb{P}$ or $\mathbb{Q}$ because of the mutual absolute continuity. Next, we relate $v_n(\mathcal{G})$ to the Hellinger distance. This link is made via an intermediate term that is also a (pseudo) distance between g_0 and g:

$$d(g_0, g) \quad = \quad \int (g - g_0) d\mathbb{Q} - \int \log \frac{g}{g_0} d\mathbb{P}. \tag{6}$$

Lemma 3. *(i)* $d(g_0, g) \geq 2h^2_{\mathbb{Q}}(g, g_0)$ *for any* $g \in \mathcal{G}$.
(ii) If $\hat{g}_n$ *is the estimate of* g_0*, then* $d(g_0, \hat{g}_n) \leq v_n(\mathcal{G})$.

Lemma 3 asserts that the Hellinger distance between $\hat{g}_n$ and g_0 is bounded by the suprema of empirical processes $v_n(\mathcal{G})$. One difficulty with the function class $\{\log(g/g_0)\}$ is that it can be unbounded when g takes value ∞ or 0. The following lemma borrows an idea due to Birgé and Massart (cf. [19]), considering functions $\log \frac{g_0 + g}{2g_0}$, which are always bounded from below. We have:

Lemma 4. *If* $\hat{g}_n$ *is the estimate of* g_0*, then:*

$$\frac{1}{8} h^2_{\mathbb{Q}}(g_0, \hat{g}_n) \leq 2h^2_{\mathbb{Q}}(g_0, \frac{g_0 + \hat{g}_n}{2}) \leq$$
$$- \int \frac{\hat{g}_n - g_0}{2} d(\mathbb{Q}_n - \mathbb{Q}) + \int \log \frac{\hat{g}_n + g_0}{2g_0} d(\mathbb{P}_n - \mathbb{P}).$$

B. Consistency results

Our analysis relies on results from empirical process theory. We first introduce several standard notions of entropy of a function class [21]. For each $\delta > 0$, a covering for function class $\mathcal{G}$ using metric $L_r(\mathbb{Q})$ is a collection of functions which cover entire $\mathcal{G}$ using $L_r(\mathbb{Q})$ balls of radius δ. Let $N_\delta(\mathcal{G}, L_r(\mathbb{Q}))$ be the smallest

cardinality of such a covering, then $\mathcal{H}_\delta(\mathcal{G}, L_r(\mathbb{Q})) := \log N_\delta(\mathcal{G}, L_r(\mathbb{Q}))$ is called the *entropy* for $\mathcal{G}$ using $L_r(\mathbb{Q})$ metric. A related notion is *entropy with bracketing*. Let $N_\delta^B(\mathcal{G}, L_r(\mathbb{Q}))$ be the smallest value of N for which there exist N pairs of functions $\{g_j^L, g_j^U\}$ such that $\|g_j^U - g_j^L\|_{L_r(\mathbb{Q})} \leq \delta$, and such that for each $g \in \mathcal{G}$ there is a j such that $g_j^L \leq g \leq g_j^L$. Then $\mathcal{H}_\delta^B(\mathcal{G}, L_r(\mathbb{Q})) := \log N_\delta^B(\mathcal{G}, L_r(\mathbb{Q}))$ is called the entropy with bracketing of $\mathcal{G}$. Define the envelope functions:

$$G_0(x) = \sup_{g\in\mathcal{G}} |g(x)|; \;\; G_1(x) = \sup_{g\in\mathcal{G}} |\log \frac{g(x)}{g_0(x)}|.$$

Proposition 5. *Assume the envelope conditions*

$$(a) \int G_0 d\mathbb{Q} < \infty, \text{ and } (b) \int G_1 d\mathbb{P} < \infty, \tag{7}$$

and suppose that for all $\delta > 0$ there holds:

$$\frac{1}{n}\mathcal{H}_\delta(\mathcal{G} - g_0, L_1(\mathbb{Q}_n)) \xrightarrow{\mathbb{Q}} 0, \tag{8a}$$

$$\frac{1}{n}\mathcal{H}_\delta(\log \mathcal{G}/g_0, L_1(\mathbb{P}_n)) \xrightarrow{\mathbb{P}} 0. \tag{8b}$$

Then, $v_n(\mathcal{G}) \xrightarrow{a.s.} 0$. As a result, $\mathcal{E}_1(\mathcal{G}) \xrightarrow{a.s.} 0$, and $h_\mathbb{Q}(g_0, \hat{g}_n) \xrightarrow{a.s.} 0$.

Envelope condition (7)(a) is quite severe, because it essentially requires all functions in $\mathcal{G}$ be bounded from *both* above and below. To ensure the Hellinger consistency of the estimation for g_0, however, we can essentially drop envelope condition (7)(a) and replace entropy condition (8)(a) by a milder entropy condition.

Proposition 6. *Assume that* (7)*(b) and* (8a) *hold, and*

$$\frac{1}{n}\mathcal{H}_\delta(\log \frac{\mathcal{G} + g_0}{2g_0}, L_1(\mathbb{P}_n)) \xrightarrow{\mathbb{P}} 0, \tag{9}$$

then $h_\mathbb{Q}(g_0, \hat{g}_n) \xrightarrow{a.s.} 0$.

It can be shown that both entropy conditions (8a) and (9) can be deduced from a single condition—namely, that for all $\delta > 0$, the bracketing entropy is bounded as $\mathcal{H}_\delta^B(\mathcal{G}, L_1(\mathbb{Q})) < \infty$.

As a concrete illustration, let us consider a particular function class:

Example: (Sobolev classes W_2^α) For $x \in \mathbb{R}^d$, and a d-dimensional multi-index $\kappa = (\kappa_1, \ldots, \kappa_d)$ (all κ_i are natural numbers), write $x^\kappa = \prod_{i=1}^d x_i^{\kappa_i}$, and $|\kappa| = \sum_{i=1}^d \kappa_i$. Let D^κ denote the differential operator:

$$D^\kappa g(x) = \frac{\partial^{|\kappa|}}{\partial x_1^{\kappa_1} \ldots \partial x_d^{\kappa_d}} g(x_1, \ldots, x_d).$$

Let $W_2^\alpha(\mathcal{X})$ denote the Sobolev space of functions $f : \mathcal{X} \to \mathbb{R}$, where $||f||^2_{L_2^\alpha(\mathcal{X})} = \sum_{|\kappa|=\alpha} \int |D^\kappa f(x)|^2 \, dx$ is bounded by a constant M^2. If $\mathcal{G}$ is restricted to a subspace of $W_2^\alpha(\mathcal{X})$ such that $\mathcal{G}$ is uniformly bounded from above, then all conditions required in Prop. 6 can be easily shown to hold. If, in addition, all functions in $\mathcal{G}$ is also bounded from below, then all conditions in Prop. 5 hold.

C. Convergence rate of density ratio estimation

We can obtain the convergence rate of our estimator $\hat{g}_n$ using the Hellinger metric. Our result is based on Lemma 4, which bounds the Hellinger metric in terms of the supremum of empirical processes, and the modulus of continuity of this supremum. We shall assume that:

$$\sup_{g\in\mathcal{G}} \|g\|_\infty < K_2, \tag{10}$$

in addition to an entropy condition on the function class $\bar{\mathcal{G}} := \{((g + g_0)/2)^{1/2}, g \in \mathcal{G}\}$: In particular, we assume that for some constant $0 < \gamma_{\bar{\mathcal{G}}} < 2$, there holds for any $\delta > 0$,

$$\mathcal{H}_\delta^B(\bar{\mathcal{G}}, L_2(\mathbb{Q})) = O(\delta^{-\gamma_{\bar{\mathcal{G}}}}). \tag{11}$$

Theorem 7. *Under conditions* (10) *and* (11)*, then $h_\mathbb{Q}(g_0, \hat{g}_n) = O_\mathbb{P}(n^{-1/(\gamma_{\bar{\mathcal{G}}}+2)})$, where $O_\mathbb{P}$ is with respect to $\mathbb{P}$.*

Remarks: To follow up on our earlier example, if $\mathcal{G}$ is a Sobolev class with smoothness α, and g_0 is bounded from below, then it is known [5] that $\gamma_{\bar{\mathcal{G}}} = d/\alpha$. In this particular case, we obtain the rate $n^{-\alpha/(2\alpha+d)}$. It is worthwhile comparing to the optimal minimax rates with respect to Hellinger metric. More precisely, the minimax rate is defined as:

$$r_n := \inf_{\hat{g}_n \in \mathcal{G}} \sup_{\mathbb{P},\mathbb{Q}} \mathbb{E}_\mathbb{P} h_\mathbb{Q}(g_0, \hat{g}_n).$$

Here the infimum is taken with respect to all estimators $\hat{g}_n \in \mathcal{G}$, where $\mathcal{G}$ is a Sobolev class with smoothness α. First, note that $r_n \geq \inf_{\hat{g}_n\in\mathcal{G}} \sup_\mathbb{P} \mathbb{E} h_\mu(g_0, \hat{g}_n)$, where we have fixed $= \mu$ as the Lebesgue measure on $\mathcal{X}$. Thus, we have lower bounded the minimax rate by that of a nonparametric density estimation problem.[1] Thus, we obtain the following:

Proposition 8. *When the likelihood ratio lies in the Sobolev class of smoothness α, the optimal minimax rate $r_n = \Omega(n^{-\alpha/(2\alpha+d)})$ is achieved by our estimator.*

[1]There is a small technical aspect here: the space $\mathcal{G}$ ranges over smooth functions that need not be valid probability densities. Nonetheless, a standard hypercube argument is still directly applicable. (See [20], Sec. 24.3, for such an argument).

D. Convergence rates of divergence estimation

We now turn to the convergence rate of our estimator for the KL divergence, i.e., $\|\hat{D}_K - D_K(\mathbb{P},\mathbb{Q})\|$. We assume that all functions in $\mathcal{G}$ are bounded from above *and* below:

$$0 < K_1 \leq \|g\|_\infty \leq K_2 \text{ for all } g \in \mathcal{G}. \tag{12}$$

Theorem 9. *Under conditions* (12) *and* (11)*, we have* $\left|\hat{D}_K - D_K(\mathbb{P},\mathbb{Q})\right| = O_{\mathbb{P}}(n^{-1/(\gamma_{\bar{\mathcal{G}}}+2)})$.

Proof: We provide only a sketch here. From equations (2) and (3), we can bound $|\hat{D}_K - D_K(\mathbb{P},\mathbb{Q})|$ from above by the sum $A+B+C$ of three terms, where

$$\begin{aligned} A &:= \left|\int \log \hat{g}_n/g_0 d(\mathbb{P}_n - \mathbb{P}) - \int (\hat{g}_n - g_0) d(\mathbb{Q}_n - \mathbb{Q})\right| \\ B &:= \left|\int \log \hat{g}_n/g_0 d\mathbb{P} - \int (\hat{g}_n - g_0) d\mathbb{Q}\right| \\ C &:= \left|\int \log g_0 d(\mathbb{P}_n - \mathbb{P}) - \int g_0 d(\mathbb{Q}_n - \mathbb{Q})\right|. \end{aligned}$$

We have $C = O_P(n^{-1/2})$ by the central limit theorem. Using assumption (12),

$$\begin{aligned} B &\leq \int |\hat{g}_n - g_0| \frac{K_2}{K_1} d\mathbb{Q}| + \int |\hat{g}_n - g_0| d\mathbb{Q} \\ &\leq (K_2/K_1 + 1)\|\hat{g}_n - g_0\|_{L_2(\mathbb{Q})} \\ &\leq (K_2/K_1 + 1) K_2^{1/2} 4 h_{\mathbb{Q}}(g_0, \hat{g}_n) \\ &\overset{(a)}{=} O_{\mathbb{P}}(n^{-1/(2+\gamma_{\bar{\mathcal{G}}})}), \end{aligned}$$

where equality (a) is due to Theorem 7.

Finally, to bound A, we apply a modulus of continuity result on the suprema of empirical processes with respect to function $(g - g_0)$ and $(\log g - \log g_0)$. From equation (12), the bracket entropy for both function classes $\mathcal{G}$ and $\log \mathcal{G}$ has the same order as that of $\bar{\mathcal{G}}$, as given in (11). Applying Lemma 5.13 from van de Geer [19], we obtain that for $\delta_n = n^{-1/(2+\gamma_{\bar{\mathcal{G}}})}$, there holds:

$$A = O_{\mathbb{P}}(n^{-1/2}\|\hat{g}_n - g_0\|_{L_2(\mathbb{Q})}^{1-\gamma_{\bar{\mathcal{G}}}/2} \vee \delta_n^2) = O_{\mathbb{P}}(n^{-2/(2+\gamma_{\bar{\mathcal{G}}})}).$$

Overall, we have established that the sum $A + B + C$ is upper bounded by $O_{\mathbb{P}}(n^{-2/(2+\gamma_{\bar{\mathcal{G}}})})$.

IV. Other results

In [16], we provide several further results that cannot be presented here due to space limitations. At a high level, these results include the following:

M-estimation of D_ϕ and $\partial\phi(q_0/p_0)$: Although we have focused primarily here on the KL divergence, our approach is applicable to the estimation of any f-divergence D_ϕ and subgradient $\partial\phi(q_0/p_0)$. In general, the estimator of D_ϕ takes the following form:

$$\hat{D}_\phi := \sup_{f \in \mathcal{F}} \int f \, d\mathbb{Q}_n - \int \phi^*(f) \, d\mathbb{P}_n,$$

where the supremum is attained at an estimate of $\partial\phi(q_0/p_0)$. The convergence analysis hinges on the modulus of continuity of the suprema of appropriate empirical processes with respect to the following Bregman divergence (a special case of which was defined in Eq. (6)):

$$d_\phi(f_0, f) = \int (\phi^*(f) - \phi^*(f_0) - \left.\frac{\partial \phi^*}{\partial f}\right|_{f_0} (f - f_0) \, d\mathbb{P}.$$

Implementation and experimental results: In practice, we implement our estimator by taking $\mathcal{G}$ to be the reproducing kernel Hilbert space induced by a Gaussian kernel. Doing so yields a convex optimization problem that can be solved efficiently [16]. Our experiment results [16] demonstrate the practical viability of such estimators, which complements the theoretical analysis of consistency and convergence rates that we have reported here.

References

[1] S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *J. Royal Stat. Soc. Series B*, 28:131–142, 1966.

[2] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101:138–156, 2006.

[3] P. Bickel and Y. Ritov. Estimating integrated squared density derivatives: Sharp best order of convergence estimates. *Sankhyā Ser. A*, 50:381–393, 1988.

[4] L. Birgé and P. Massart. Estimation of integral functionals of a density. *Annals of Statistics*, 23(1):11–29, 1995.

[5] M. S. Birman and M. Z. Solomjak. Piecewise-polynomial approximations of functions of the classes W_p^α. *Math. USSR-Sbornik*, 2(3):295–317, 1967.

[6] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley and Sons, New York, 1991.

[7] I. Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Sci. Math. Hungar*, 2:299–318, 1967.

[8] L. Gyorfi and E.C. van der Meulen. Density-free convergence properties of various estimators of entropy. *Computational Statistics and Data Analysis*, 5:425–436, 1987.

[9] P. Hall and S. Morton. On estimation of entropy. *Ann. Inst. Statist. Math.*, 45(1):69–88, 1993.

[10] I. A. Ibragimov and R. Z. Khasminskii. On the nonparametric estimation of functionals. In *Symposium in Asymptotic Statistics*, pages 41–52, 1978.

[11] H. Joe. Estimation of entropy and other functionals of a multivariate density. *Ann. Inst. Statist. Math.*, 41:683–697, 1989.

[12] T. Kailath. The divergence and Bhattacharyya distance measures in signal selection. *IEEE Trans. on Communication Technology*, 15(1):52–60, 1967.

[13] B. Laurent. Efficient estimation of integral functionals of a density. *Annals of Statistics*, 24(2):659–681, 1996.

[14] B. Ya. Levit. Asymptotically efficient estimation of nonlinear functionals. *Problems Inform. Transmission*, 14:204–209, 1978.

[15] X. Nguyen, M. J. Wainwright, and M. I. Jordan. On divergences, surrogate losses and decentralized detection. Technical Report 695, Department of Statistics, UC Berkeley, October 2005.

[16] X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. Technical report, Department of Statistics, UC Berkeley, January 2007.

[17] G. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970.

[18] F. Topsoe. Some inequalities for information divergence and related measures of discrimination. *IEEE Transactions on Information Theory*, 46:1602–1609, 2000.

[19] S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.

[20] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

[21] A. W. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, NY, 1996.

[22] Q. Wang, S. R. Kulkarni, and S. Verdú. Divergence estimation of continuous distributions based on data-dependent partitions. *IEEE Transactions on Information Theory*, 51(9):3064–3074, 2005.