Estimating divergence functionals and the likelihood ratio by convex risk minimization

XuanLong Nguyen
Department of EECS
University of California, Berkeley
xuanlong@eecs.berkeley.edu

Martin J. Wainwright
Department of Statistics and Department of EECS
University of California, Berkeley
wainwrig@stat.berkeley.edu

Michael I. Jordan
Department of Statistics and Department of EECS
University of California, Berkeley
jordan@stat.berkeley.edu

September 17, 2007

Abstract

We present a novel M-estimation method for the divergence functionals and the density ratios of two probability distributions. Our method is based on a non-asymptotic variational characterization of f-divergences, which turns the problem of estimating divergences to a convex risk optimization. We present an analysis of consistency and convergence for our estimator. Given conditions only on the ratios of densities, we show that our estimators can achieve optimal minimax rates for the likelihood ratio in some regime. Finally, we present an efficient optimization algorithm for our estimator and demonstrate its convergence behavior and practical viability by simulations.[1]

1 Introduction

Given empirical samples from two (multivariate) probability distributions $\mathbb{P}$ and $\mathbb{Q}$, we are interested in estimating a divergence functional between $\mathbb{P}$ and $\mathbb{Q}$. We consider in particular Kullback-Leibler divergence, and then all divergences in the class of Ali-Silvey distance, also known as f-divergences (Ali and Silvey, 1966, Csiszár, 1967). This family of divergence, which shall 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 $d\mathbb{Q}/d\mathbb{P}$.

The divergences have a fundamental role as an objective to optimize in various data analysis and learning tasks. Divergences are used as a measure to distinguish between two hypotheses. In experiment design for binary hypothesis testing and classification applications, the experiments are designed so that the divergence between two underlying hypothesis distributions are maximized. Problems of this type can be seen in signal selection (Kailath, 1967), decentralized detection (Nguyen et al., 2005). An important quantity in information theory, the Shanon mutual information, can be viewed as a KL divergence. Mutual information is often used as a measure of independence to be minimized such as in the problem of independent component

[1] Preliminary versions of this paper have been presented in ISIT (Nguyen et al., 2007b) and NIPS (Nguyen et al., 2007a) conferences.

analysis (Hyvarinen et al., 2001). If the divergences are to be used as objective functional in such tasks, one has to be able to estimate them efficiently from empirical data.

There are two ways in which divergences can be characterized. Taking the KL divergence in particular, in the Neyman-Pearson setting of a binary hypothesis testing problem, the KL divergence emerges as the correct asympotic rate of the probability error, a result known as Stein's lemma. On the other hand, a non-asymptotic view of KL divergence emerges through Fano's lemma, which provide a lower bound for the error probability for decoding/hypothesis test in terms of KL divergence (cf. Cover and Thomas (1991)).

In this paper, we shall present an estimation method that is motivated by a non-asymptotic characterization of f-divergence that was explicated in Nguyen et al. (2005). Roughly speaking, this theorem states that that there is a correspondence between the family of f-divergences and a family of losses such that the minimum 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. While this result deals only with binary hypotheses (as opposed to Fano's lemma) it goes significantly further than Fano's lemma in that it covers a whole class of losses and divergences. This correspondence provides what we shall call a *variational characterization* of divergence: One can write a divergence $D_\phi(\mathbb{P}, \mathbb{Q})$ as the maximum of an Bayes decision problem involving two hypotheses $\mathbb{P}$ and $\mathbb{Q}$. This characterization is stated in Lemma 2.1. As a result, one can estimate $D_\phi(\mathbb{P}, \mathbb{Q})$ by solving the Bayes decision (maximization) problem. Not surprisingly, we show how the problem of estimating f-divergence is intrinsically linked to that of estimating the likelihood ratio $g_0 = d\mathbb{P}/d\mathbb{Q}$. As a result we obtain an M estimator for the likelihood ratio, from which one can obtain an estimation of the divergences by a plug-in procedure.

Our contributions are three-fold:

- We propose a novel M-estimator for the likelihood ratio and the family of f divergences based on a variational characterization of f-divergence as explained above. Our estimation procedure is inherently nonparametric. We make no strong assumption on the form of the densities for $\mathbb{P}$ and $\mathbb{Q}$.

- We provide a consistency and convergence analysis for our estimators. For the analysis, we make assumptions on the boundedness of the *density ratio*, which can be relaxed in some cases. The maximization procedure is cast over a whole function class $\mathcal{G}$ of density ratio, thus our tool is based on results from the theory of empirical processes. Our method of proof is based on the analysis of M-estimation for nonparametric density estimation (van de Geer, 2000, van der Vaart and Wellner, 1996). The key issue essentially hinges on the modulus of continuity of the suprema of two empirical processes (defined on $\mathbb{P}$ and $\mathbb{Q}$ measures) with respect to a metric defined on the class $\mathcal{G}$. This metric turns out to be a surrogate lower bound of a Bregman divergence defined on a pair of density ratios. Our choice of metrics include the Hellinger distance and L_2 norm.

- We also analyze an M-estimator based on penalized convex risk minimization, where the penalty is placed on a complexity measure on $g \in \mathcal{G}$. This method is particularly useful in practice. In particular, we provide an implementation of the estimator by approximating $\mathcal{G}$ by a reproducing kernel Hilbert space given a positive definite kernel function $K(u, v)$ (Saitoh, 1988). The estimation problem is converted into a convex optimization problem, which is then turned into a dual form involving only the Gram matrix $K(u_i, v_j)$, where u_i and v_j are drawn from either $\mathbb{P}$ or $\mathbb{Q}$. This kernel-based method has been widely used in statistical learning tasks (Schölkopf and Smola, 2002, Shawe-Taylor and Cristianini, 2004). Finally, we demonstrate our estimator in a large number of simulation runs on a number of pairs of probability distributions.

Several interesting properties of this estimator is worth highlighting.

- First, in terms of convergence rates. 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 dimensions of the data, our estimation of the likelihood ratio achieves the optimal minimax rate $n^{-\alpha/(2\alpha+d)}$ according to the Hellinger metric, and divergence estimator achieves the same rate. It remains an open question what is the optimal minimax rate for the divergence estimation.

- An obvious alternative approach to our problem would be to separately estimate the densities for $\mathbb{P}$ and $\mathbb{Q}$ and then use an appropriate plug-in estimator for the divergences. As we shall see in our analysis, estimating directly the density ratio has several distinct advantages. Firstly, from computational viewpoint, it is more efficient to perform one estimation procedure instead of two. Comparing to an M-estimator for density estimation (e.g, (Silverman, 1982)), there is no need to enforce the constraint that the estimated function is a valid density. Secondly, from a statistical viewpoint, we achieve the same estimation efficiency without making individual assumptions on each density. Assumptions are made only on the density ratios.

- Finally, if we use small function classes $\mathcal{G}$ that might not include the true likelihood ratio, our estimator of the divergence has the property of being a lower bound of the true divergence. This might provide additional useful information for a task in hand.

Related work. The variational representation of divergences has been derived independently and exploited by several authors (Broniatowski and Keziou, 2004, Keziou, 2003, Nguyen et al., 2005). Broniatowski and Keziou (2004) studied testing and estimation problems based on dual representations of f-divergences, but working in a parametric setting as opposed to the nonparametric framework considered here. Nguyen et al. (2005) established a one-to-one correspondence between the family of f-divergences and the family of surrogate loss functions (Bartlett et al., 2006), through which the (optimum) "surrogate risk" is equal to the negative of an associated f-divergence. Another link is to the problem of estimating integral functionals of a single density, with the Shannon entropy being a well-known example, which has been studied extensively dating back to early work (Ibragimov and Khasminskii, 1978, Levit, 1978) as well as the more recent work (Bickel and Ritov, 1988, Birgé and Massart, 1995, Laurent, 1996). See also Gyorfi and van der Meulen (1987), Joe (1989), Hall and Morton (1993) for the problem of (Shannon) entropy functional estimation. In another branch of related work, Wang et al. (2005) proposed an algorithm for estimating the KL divergence for continuous distributions, which exploits histogram-based estimation of the likelihood ratio by building data-dependent partitions of equivalent (empirical) $\mathbb{Q}$-measure. The estimator was empirically shown to outperform direct plug-in methods, but no theoretical results on its convergence rate were provided.

The paper is organized as follows. In Sec. 2 we describe the variational characterization of f-divergence in general and KL divergence in particular, followed by an M-estimator for the KL divergence and the likelihood ratio. Sec. 3 and Sec. 4 are devoted to the analysis of consistency and convergence rates of our estimators. In Sec. 5 we describe our estimation method and the analysis in a more general light, encompassing virtually all f-divergences. We also consider a general estimation framework based on the delta method, assuming the ϕ is a differentiable function. Sec. 7 describe the optimization in detail. In Sec. 9 we present our simulation results.

2 M-estimators for KL divergence and the density ratio

2.1 Variational characterization of f-divergence

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

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

The KL divergence is a special case of a broader class of divergences known as Ali-Silvey distance, or f-divergence (Csiszár, 1967, Ali and Silvey, 1966):

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

where $\phi : \mathbb{R} \to \mathbb{R}$ 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., (Topsoe, 2000)).

Since ϕ is a convex function, by Legendre-Fenchel convex duality (Rockafellar, 1970) we can write:

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

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

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

where the suppremum is taken over all measurable function $f : \mathcal{X} \to \mathbb{R}$, and $\int f \, d\mathbb{P}$ denotes the expectation of f under distribution $\mathbb{P}$. It is simple to see that equality the supremum is attained at function 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 2.1. *Let $\mathcal{F}$ be any function class $\mathcal{X} \to \mathbb{R}$, 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 whenever $\mathcal{F} \cap \partial\phi(q_0/p_0) \neq \emptyset$.

Remark. There is an interesting connection between Lemma 2.1 and Fano's lower bound in coding theory. Indeed, consider a Bayesian hypothesis testing problem between two distributions $\mathbb{P}$ and $\mathbb{Q}$, which have

equal priors (1/2). Let $-f$ be the loss for incorrectly rejecting $\mathbb{Q}$, and $\phi^*(f)$ the loss for incorrectly rejecting $\mathbb{P}$. Then $-D_\phi(\mathbb{P}, \mathbb{Q})$ is nothing but the lower bound of the risk:

$$\inf_f \int (-f)\ d\mathbb{Q} + \phi^*(f)\ d\mathbb{P} = -D_\phi(\mathbb{P}, \mathbb{Q}).$$

In other words, for each f divergence $D_\phi(\mathbb{P}, \mathbb{Q})$ there exists a binary classification problem with appropriate loss functions whose optimal risk is characterized by the divergence. For a thorough analysis of this correspondence, see Nguyen et al. (2005). It can be seen that for some (appropriately parametrized) choices of f and ϕ so that both loss functions $(-f)$ and $\phi^*(f)$ correspond to the 0-1 loss function, D_ϕ becomes the variational distance (plus a constant). As a result, one obtain a special case of Fano's lemma for binary classification. This connection extends to multiple hypothesis testing, but we shall not pursue further here.

2.2 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 2.1,

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

In addition, the supremum is attained at $g = p_0/q_0$. This motivates our estimator of the KL divergence as follows: 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, then our estimator has the following form:

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

For the implementation, we shall assume that $\mathcal{G}$ is a convex function class. The above estimator can be posed as a convex optimization problem that can be solved efficiently (see Section X). Suppose that the supremum is attained at $\hat{g}_n$. Then $\hat{g}_n$ is an M-estimator of the density ratio $g_0 = p_0/q_0$.

For the KL divergence estimation, there are two sources of error, namely, approximation error $\mathcal{E}_0(\mathcal{G})$ and estimation error $\mathcal{E}_1(\mathcal{G})$:

$$\mathcal{E}_0(\mathcal{G}) = D_K(\mathbb{P}, \mathbb{Q}) - \sup_{g \in \mathcal{G}} \int (\log g\ d\mathbb{P} - g\ d\mathbb{Q} + 1) \geq 0 \tag{4}$$

$$\mathcal{E}_1(\mathcal{G}) = \sup_{g \in \mathcal{G}} \left| \int \log g\ d(\mathbb{P}_n - \mathbb{P})\ - g d(\mathbb{Q}_n - \mathbb{Q}) \right|. \tag{5}$$

From (2)(3)(4) and (5), it is simple to see that:

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

For the density ratio estimation, $\hat{D}_K - D_K(\mathbb{P}, \mathbb{Q})$ can also be considered as a performance measure. Note that p_0/q_0 can be viewed as a density function with respect to $\mathbb{Q}$ measure. A natural performance measure is the Hellinger distance:

$$h^2_{\mathbb{Q}}(g, g_0) := \frac{1}{2} \int (g^{1/2} - g_0^{1/2})^2\ d\mathbb{Q}. \tag{6}$$

As we shall see, this distance measure is less strong than $\hat{D}_K - D_K(\mathbb{P}, \mathbb{Q})$, but it allows us to obtain convergence rate guarantees with less assumption.

3 Consistency analysis

In this section we shall prove consistency results and obtain convergence rates of our estimators. Throughout the paper, the following assumptions are made with respect to $\mathbb{P}, \mathbb{Q}$ and the function class $\mathcal{G}$.
Assumptions. (i) $D_K(\mathbb{P}, \mathbb{Q}) < \infty$.
(ii) $\mathcal{G}$ is sufficiently rich, i.e., $g_0 \in \mathcal{G}$.

Due to (ii), $\mathcal{E}_0(\mathcal{F}) = 0$. Hence, we shall focus on estimation error $\mathcal{E}_1(\mathcal{G})$ only. Note that if (ii) does not hold, we should obtain instead a lower bound of the KL divergence.

3.1 Preliminary lemmas

Define the following processes:

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

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

We have:

$$\mathcal{E}_1(\mathcal{G}) \leq v_n(\mathcal{G}) + w_n(g_0). \tag{7}$$

Lemma 3.1. $w_n(g_0) \xrightarrow{a.s.} 0$.

Note that in this lemma and other theorems, all almost sure convergence statement can be understood with respect to either $\mathbb{P}$ or $\mathbb{Q}$ because they share the same support.

Proof. This follows immediately from the law of large numbers. We only need to check the condition for which this law applies. Applying the following inequality due to Csiszár (cf. Gyorfi and van der Meulen (1987)):

$$\int p_0 |log(p_0/q_0)| \leq D_K(\mathbb{P}, \mathbb{Q}) + 4\sqrt{D_K(\mathbb{P}, \mathbb{Q})}$$

so that $\log g_0$ is $\mathbb{P}$ integrable. In addition, g_0 is $\mathbb{Q}$ integrable, since $\int g_0 d\mathbb{Q} = \int (p_0/q_0) d\mathbb{Q} = 1$. □

Next, we shall relate $v_n(\mathcal{G})$ to the Hellinger distance. This is done through an intermediate term which is also a (pseudo) distance between g_0 and g:

$$d(g_0, g) \;\; = \;\; \int (g - g_0) d\mathbb{Q} - \log \frac{g}{g_0} d\mathbb{P}. \tag{8}$$

Lemma 3.2. *(i)* $d(g_0, g) \geq 2h^2_{\mathbb{Q}}(g, g_0)$.
(ii) If $\hat{g}_n$ *is an estimate of* g*, then* $d(g_0, \hat{g}_n) \leq v_n(\mathcal{G})$.

Proof. (i) Note that for $x > 0$, $\frac{1}{2} \log x \leq \sqrt{x} - 1$. Thus, $\int \log \frac{g}{g_0} \; d\mathbb{P} \leq 2 \int (g^{1/2} g_0^{-1/2} - 1) \; d\mathbb{P}$. As a result,

$$\begin{aligned} d(g_0, g) \;\; &\geq \;\; \int (g - g_0) \; d\mathbb{Q} - 2 \int (g^{1/2} g_0^{-1/2} - 1) \; d\mathbb{P} \\ &= \;\; \int (g - g_0) \; d\mathbb{Q} - 2 \int (g^{1/2} g_0^{1/2} - g_0) \; d\mathbb{Q} \\ &= \;\; \int (g^{1/2} - g_0^{1/2})^2 d\mathbb{Q}. \end{aligned}$$

(ii) By our estimation procedure, we have $\int \hat{g}_n d\mathbb{Q}_n - \int \log \hat{g}_n d\mathbb{P}_n \leq \int g_0 d\mathbb{Q}_n - \int \log g_0 d\mathbb{P}_n$. It follows that

$$\begin{aligned} d(g_0, \hat{g}_n) &= \int (\hat{g}_n - g_0) d\mathbb{Q} - \int (\log \hat{g}_n - \log g_0) d\mathbb{P} \\ &\leq \int (\hat{g}_n - g_0) d(\mathbb{Q} - \mathbb{Q}_n) - \int (\log \hat{g}_n - \log g_0) d(\mathbb{P} - \mathbb{P}_n) \\ &\leq \sup_{g \in \mathcal{G}} \int \log \frac{g}{g_0} d(\mathbb{P}_n - \mathbb{P}) - \int (g - g_0) d(\mathbb{Q}_n - \mathbb{Q}). \end{aligned}$$

□

We can prove the Hellinger consistency using less assumption. For that we shall need the following lemma using a similar idea of using $(g_0 + g)/2$ due to Birgé and Massart (cf. (van de Geer, 2000)):

Lemma 3.3. *If $\hat{g}_n$ is an estimate of g, then:*

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

Proof. The first inequality is straigthforward. We shall focus on the second. By the definition of our estimator, we have:

$$\int \hat{g}_n d\mathbb{Q}_n - \int \log \hat{g}_n d\mathbb{P}_n \leq \int g_0 d\mathbb{Q}_n - \int \log g_0 d\mathbb{P}_n.$$

Both sides are convex functionals of g. Use the following fact: If F is a convex function and $F(u) \leq F(v)$, then $F((u + v)/2) \leq F(v)$. We obtain:

$$\int \frac{\hat{g}_n + g_0}{2} d\mathbb{Q}_n - \int \log \frac{\hat{g}_n + g_0}{2} d\mathbb{P}_n \leq \int g_0 d\mathbb{Q}_n - \int \log g_0 d\mathbb{P}_n.$$

Rearranging,

$$\begin{aligned} &\int \frac{\hat{g}_n - g_0}{2} d(\mathbb{Q}_n - \mathbb{Q}) - \int \log \frac{\hat{g}_n + g_0}{2g_0} d(\mathbb{P}_n - \mathbb{P}) \leq \int \log \frac{\hat{g}_n + g_0}{2g_0} d\mathbb{P} - \int \frac{\hat{g}_n - g_0}{2} d\mathbb{Q} \\ &= -d(g_0, \frac{g_0 + \hat{g}_n}{2}) \leq -2h^2_{\mathbb{Q}}(g_0, \frac{g_0 + \hat{g}_n}{2}), \end{aligned}$$

where the last inequality is an application of Lemma 3.2.

□

3.2 Consistency results

Our analysis shall rely on results from empirical processes theory. We first introduce several standard notions of *entropy* of a function class (see, e.g., (van der Vaart and Wellner, 1996) for more detail). 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 and centering at these functions. 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 $N^B_\delta(\mathcal{G}, L_r(\mathbb{Q}))$ be the smallest value of N for which there exist pairs of functions $\{g^L_j, g^U_j\}$ such that $\|g^U_j - g^L_j\|_{L_r(\mathbb{Q})} \leq \delta$, and such that for each $g \in \mathcal{G}$ thre is a j

such that $g_j^L \le g \le g_j^L$. 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 3.4. *Assume the envelope conditions*

$$\int G_0 d\mathbb{Q} \;<\; \infty \tag{9a}$$

$$\int G_1 d\mathbb{P} \;<\; \infty \tag{9b}$$

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

$$\frac{1}{n}\mathcal{H}_\delta(\mathcal{G} - g_0, L_1(\mathbb{Q}_n)) \xrightarrow{\mathbb{Q}} 0, \tag{10a}$$

$$\frac{1}{n}\mathcal{H}_\delta(\log \mathcal{G}/g_0, L_1(\mathbb{P}_n)) \xrightarrow{\mathbb{P}} 0. \tag{10b}$$

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

Proof. That $v_n(\mathcal{G}) \xrightarrow{a.s.} 0$ is a direct consequence of Thm 10.1 (see the Appendix). By (7) and Lemma 3.1, $\mathcal{E}_1(\mathcal{G}) \xrightarrow{a.s.} 0$. By Lemma 3.2, this would also imply that $h_\mathbb{Q}(\hat{g}_n, g_0) \xrightarrow{a.s.} 0$, i.e., our estimation of the ratio p_0/g_0 is consistent in Hellinger sense. □

The envelope condition (9a) is satisfied if $\mathcal{G}$ is uniformly bounded from above. The envelope condition (9b) is much more severe. Due to logarithm, this can be satisfied if all functions in $\mathcal{G}$ is bounded from *both* above and below. To ensure the Hellinger consistency of the estimation for g_0, however, we can essentially drop the envelope condition (9b) as well as the entropy condition (10b), which is replaced by a milder entropy condition.

Proposition 3.5. *Assume that* (9a) *and* (10a) *holds, and*

$$\frac{1}{n}\mathcal{H}_\delta(\log \frac{\mathcal{G} + g_0}{2g_0}, L_1(\mathbb{P}_n)) \xrightarrow{\mathbb{P}} 0. \tag{11}$$

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

Proof. Define $G_2(x) = \sup_{g\in\mathcal{G}} |\log \frac{g(x)+g_0(x)}{2g_0(x)}|$. Due to Lemma 3.3(i) and Thm 10.1 (see the Appendix), it is sufficient to prove that

$$\int G_2 d\mathbb{P} < \infty. \tag{12}$$

Indeed,

$$\int G_2 d\mathbb{P} \le \int \sup_{g\in\mathcal{G}} \max\{\frac{g(x) + g_0(x)}{2g_0(x)} - 1, \log 2\} d\mathbb{P} \le \log 2 + \int \sup_{g\in\mathcal{G}} |g(x) - g_0(x)| d\mathbb{Q} < \infty,$$

where the last inequality is due to envelope condition (9a). □

Remark. Let us now turn to a discussion of the entropy conditions. Note that both entropy conditions (10a) and (11) can be deduced from the following single condition: For all $\delta > 0$,

$$\mathcal{H}_\delta^B(\mathcal{G}, L_1(\mathbb{Q})) < \infty. \tag{13}$$

Indeed, that (13) implies (10a) is a direct consequence of the law of large numbers (given (9a)). To show (11), note that (by Taylor's expansion):

$$\left| \log \frac{g_1 + g_0}{2g_0} - \log \frac{g_2 + g_0}{2g_0} \right| \leq \frac{|g_1 - g_2|}{g_0},$$

so $\frac{1}{n}\mathcal{H}_\delta(\log \frac{\mathcal{G}+g_0}{2g_0}, L_1(\mathbb{P}_n)) \leq \frac{1}{n}\mathcal{H}_\delta(\mathcal{G}/g_0, L_1(\mathbb{P}_n))$. Since $G_0 \in L_1(\mathbb{Q})$, we have $G_0/g_0 \in L_1(\mathbb{P})$. In addition, $\mathcal{H}_\delta^B(\mathcal{G}/g_0, L_1(\mathbb{P})) \leq \mathcal{H}_\delta^B(\mathcal{G}, L_1(\mathbb{Q})) < \infty$. By the law of large numbers, $\mathcal{H}_\delta(\mathcal{G}/g_0, L_1(\mathbb{P}_n))$ is bounded in probability, thus (11) holds.

In the remaining of this section, we shall consider an example of smooth function classes for which the conditions of Prop. 3.4 and 3.5 hold.

Sobolev spaces. For $x \in \mathbb{R}^d$, an 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:

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

We use $W_r^\alpha(\mathcal{X})$ to denote the Sobolev space of functions $f : \mathcal{X} \to \mathbb{R}$. The norm in $W_r^\alpha(\mathcal{X})$ is defined by

$$||f||_{W_r^\alpha(\mathcal{X})} = ||f||_{L_r(\mathcal{X})} + ||f||_{L_r^\alpha(\mathcal{X})},$$

where

$$||f||_{L_r^\alpha(\mathcal{X})}^r = \sum_{|\kappa|=\alpha} \int |D^\kappa f(x)|^r \, dx.$$

Suppose, for simplicity, that the domain $\mathcal{X}$ is a compact set such as a cube $[0, h]^d$. Assume that p_0 and q_0 are bounded from above *and* below by some constants (these assumptions shall be relaxed in the next section). As a result, g_0 is bounded from above and below. Suppose that

$$\eta_1 \leq g_0(x) = \frac{p_0(x)}{q_0(x)} \leq \eta_2 \text{ for all } x \in \mathcal{X}. \tag{14}$$

We now restrict our function class to a Sobolev's space of functions that are bounded from above and below:

$$\mathcal{G} = \left\{ g \in W_r^\alpha(\mathcal{X}) \text{ such that } ||g||_{W_r^\alpha(\mathcal{X})} \leq M \right\} \cap \left\{ g : \; K_1 \leq g(x) \leq K_2 \text{ for all } x \in \mathcal{X} \right\}, \tag{15}$$

where K_1 and K_2 are some constants satisfying $K_1 \leq \eta_1 < \eta_2 \leq K_2$. In the algorithmic development and subsequent analysis of our estimator, we typically restrict ourselves to $r = 2$ unless indicated otherwise.

Under the boundedness assumption, the envelope conditions (9) hold trivially. For a function class that is sufficiently smooth, i.e., when $r\alpha > d$, then it was shown (Birman and Solomjak, 1967) that

$$\mathcal{H}_\delta(\mathcal{G}, L_\infty) < c\delta^{-d/\alpha} < \infty,$$

where c is some constant independent of δ. As a result, it is simple to see that the condition (13) holds. The entropy condtion (10b) also holds due to the boundedness condition.

Finally, while the boundedness conditions are rather severe, we can study the rate of convergence under such conditions. Once having the convergence rates for bounded cases, it would be easy to obtain consistency in more general unbounded cases if we have additional knowledge of the tail condition for the densities.

4 Rates of convergence

4.1 Convergence rate of the likelihood ratio in Hellinger metric

In this section, we shall obtain the same convergence rate of the likelihood ratio g using Hellinger metric as a performance measure. Our result is based on Lemma 3.3, in which the Hellinger distance is bounded from above by the suprema of two empirical processes.

We shall need the assumption that

$$\sup_{g\in\mathcal{G}} \|g\|_\infty < K_2. \tag{16}$$

One empirical process in the RHS in Lemma 3.3 involves function class $\mathcal{F} := \log \frac{\mathcal{G}+g_0}{2g_0}$. For each $g \in \mathcal{G}$, let $f_g := \log \frac{g+g_0}{2g_0}$. We endow $\mathcal{F}$ with a new norm, namely, *Bernstein distance*: for a constant $K > 0$,

$$\rho_K(f)^2 := 2K^2 \int (e^{|f|/K} - 1 - |f|/K) d\mathbb{P}.$$

The Bernstein distance is related to the Hellinger distance in several crucial ways (see, e.g., van de Geer (2000), page 97):

- $\rho_1(f_g) \le 4h_{\mathbb{Q}}(g_0, \frac{g+g_0}{2})$.
- The bracket entropy based on Bernstein distance is also related to the bracket entropy based Hellinger distance (i.e., which is the L_2 norm for the square root function):

$$\mathcal{H}^B_{\sqrt{2}\delta}(\mathcal{F}, \rho_1) \le \mathcal{H}^B_\delta(\bar{\mathcal{G}}, L_2(\mathbb{Q})), \tag{17}$$

  where $\bar{\mathcal{G}} := \{((g+g_0)/2)^{1/2}, g \in \mathcal{G}\}$, and $\bar{g} := (g+g_0)/2$.

We shall need an assumption on function class $\bar{\mathcal{G}}$: For some constant $0 < \gamma_{\bar{\mathcal{G}}} < 2$, there holds for any $\delta > 0$,

$$\mathcal{H}^B_\delta(\bar{\mathcal{G}}, L_2(\mathbb{Q})) = O(\delta^{-\gamma_{\bar{\mathcal{G}}}}). \tag{18}$$

Combining this condition with (16), we deduce that for $\mathcal{G}$,

$$\mathcal{H}^B_\delta(\mathcal{G}, L_2(\mathbb{Q})) \le O(\delta^{-\gamma_{\bar{\mathcal{G}}}}).$$

In the following theorem, $O_{\mathbb{P}}$ means "bounded in probability" with respect to $\mathbb{P}$ measure.

Theorem 4.1. *Assume* (16) *and* (18)*, then* $h_{\mathbb{Q}}(g_0, \hat{g}_n) = O_{\mathbb{P}}(n^{-1/(\gamma_{\bar{\mathcal{G}}}+2)})$.

Proof. By Lemma 3.3, for any $\delta > 0$, with respect to $\mathbb{P}$ measure:

$$
\begin{aligned}
&P(h_{\mathbb{Q}}(g_0, \hat{g}_n) > \delta) \leq P(h_{\mathbb{Q}}(g_0, (\hat{g}_n + g_0)/2) > \delta/4) \\
&\leq P\bigg(\sup_{g \in \mathcal{G},\, h_{\mathbb{Q}}(g_0,\bar{g}) > \delta/4} - \int (\bar{g} - g_0) d(\mathbb{Q}_n - \mathbb{Q}) + \int f_g \, d(\mathbb{P}_n - \mathbb{P}) - 2h^2_{\mathbb{Q}}(g_0, \bar{g}) \geq 0 \bigg) \\
&\leq P\bigg(\sup_{g \in \mathcal{G},\, h_{\mathbb{Q}}(g_0,\bar{g}) > \delta/4} - \int (\bar{g} - g_0) d(\mathbb{Q}_n - \mathbb{Q}) - h^2_{\mathbb{Q}}(g_0, \bar{g}) \geq 0 \bigg) + \\
&P\bigg(\sup_{g \in \mathcal{G},\, h_{\mathbb{Q}}(g_0,\bar{g}) > \delta/4} \int f_g \, d(\mathbb{P}_n - \mathbb{P}) - h^2_{\mathbb{Q}}(g_0, \bar{g}) \geq 0 \bigg) := A + B.
\end{aligned}
$$

We need to upper bound the RHS's two quantities A and B, both of which can be handled in a similar manner. Since $\mathcal{H}^B_\delta(\bar{\mathcal{G}}, L_2(\mathbb{Q})) < \infty$ the diameter of $\bar{\mathcal{G}}$ is finite. Let S be the minimum s such that $2^{s+1}\delta/4$ exceeds that diameter. We apply the so-called peeling device: Decompose $\bar{\mathcal{G}}$ into layers of Hellinger balls around g_0 and then applying union bound on the probability of excess. For each layer, one can now apply the modulus of continuity of suprema of an empirical process.

$$
B \;\leq\; \sum_{s=0}^{S} P\bigg(\sup_{g \in \mathcal{G},\, h_{\mathbb{Q}}(g_0,\bar{g}) \leq 2^{s+1}\delta/4} \int f_g \, d(\mathbb{P}_n - \mathbb{P}) \geq 2^{2s}(\delta/4)^2 \bigg).
$$

Note that if $h_{\mathbb{Q}}(g_0, \bar{g}) \leq 2^{s+1}\delta/4$ then $\rho_1(f_g) \leq 2^{s+1}\delta$. Note that for any $s = 1, \ldots, S$, the bracket entropy integral can be bounded as:

$$
\begin{aligned}
&\int_0^{2^{s+1}\delta} \mathcal{H}^B_\epsilon(\mathcal{F} \cap \{h_{\mathbb{Q}}(g_0, \bar{g}) \leq 2^{s+1}\delta/4\}, \rho_1)^{1/2} \, d\epsilon \\
\leq\; &\int_0^{2^{s+1}\delta} \mathcal{H}^B_{\epsilon/\sqrt{2}}(\bar{\mathcal{G}} \cap \{h_{\mathbb{Q}}(g_0, \bar{g}) \leq 2^{s+1}\delta/4\}, L_2(\mathbb{Q}))^{1/2} \, d\epsilon \\
\leq\; &\int_0^{2^{s+1}\delta} C_9(\epsilon/\sqrt{2})^{-\gamma_{\bar{\mathcal{G}}}/2} \, d\epsilon \\
\leq\; &C_8(2^{s+1}\delta)^{1-\gamma_{\bar{\mathcal{G}}}/2},
\end{aligned}
$$

where C_8, C_9 are constants independent of s. Now apply Thm 10.2 (see the Appendix), where $K = 1$, $R = 2^{s+1}\delta$, $a = C_1\sqrt{n}R^2/K = C_1\sqrt{n}2^{2(s+1)}\delta^2$. We need

$$
a \geq C_0 C_8 (2^{s+1}\delta)^{1-\gamma_{\bar{\mathcal{G}}}/2} > C_0 R.
$$

This is satisfied if $\delta = n^{-1/(\gamma_{\bar{\mathcal{G}}}+2)}$, and $C_1 = C_0 C_8$, where C_8 is sufficiently large (independently of s). Finally, $C_0^2 \geq C^2(C_1 + 1) = C^2(C_0 C_8 + 1)$ if $C_0 := 2C^2 C_8 \vee 2C$, where C is some universal constant in Thm 10.2. Applying this theorem, we obtain:

$$
B \;\leq\; \sum_{s=0}^{S} C \exp\bigg[- \frac{C_1^2 n 2^{2(s+1)}\delta^2}{C^2(C_1 + 1)} \bigg] \leq c \exp\bigg[- \frac{n\delta^2}{c^2} \bigg]
$$

for some universal constant c. A similar bound for A, with respect to $\mathbb{Q}$ measure and with $\delta = n^{-1/(\gamma_{\bar{\mathcal{G}}}+2)}$ can be obtained in the same manner. Since p_0/q_0 is bounded from above, this also implies a probability statement with respect to $\mathbb{P}$. Thus, $h_{\mathbb{Q}}(g_0, \hat{g}_n)$ is bounded in $\mathbb{P}$ probability by $n^{-1/(\gamma_{\bar{\mathcal{G}}}+2)}$. □

In the following we note that the rate of convergence with respect to Hellinger metric is also the optimal minimax rate, which is defined as:

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

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 $\mathbb{Q} = \mu$ the Lebesgue measure on $\mathcal{X}$. Our strategy is to reduce this bound to that minimax lower bound for a nonparametric density estimation problem (Yu, 1996). Note a technicality here, in which the space $\mathcal{G}$ ranges over smooth functions that need not to be valid probability density. Therefore, an easy-to-use mimimax lower bound such as that of (Yang and Barron, 1999) is not immediately applicable. Nonetheless, we can still apply the hypercube argument and the Assouad lemma to obtain the right minimax rate. See van der Vaart (1998) (Sec. 24.3) for a proof for the case of one dimension. The proof goes through for general $d \geq 1$.

Proposition 4.2. *For $\mathcal{G}$ defined in* (15)*, $\mathbb{P}, \mathbb{Q}$ satisfy* (14)*, $r_n = \Omega(n^{-1/(\gamma+2)})$, where $\gamma = d/\alpha$.*

4.2 Convergence rate for divergence estimation

In this section we shall obtain the convergence rate of our estimation procedure for the KL divergence, i.e., $\|\hat{D}_K - D_K(\mathbb{P}, \mathbb{Q}).\|$ We shall need the assumption that all functions in $\mathcal{G}$ are bounded from above *and* below:

$$0 < K_1 \leq g \leq K_2 \text{ for all } g \in \mathcal{G}. \tag{19}$$

Theorem 4.3. *Assume* (19) *and* (18)*, then $|\hat{D}_K - D_K(\mathbb{P}, \mathbb{Q})| = O_{\mathbb{P}}(n^{-1/(\gamma_{\bar{\mathcal{G}}}+2)})$.*

Proof. Note that

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

We have $C = O_P(n^{-1/2})$ by the central limit theorem. Using assumption (19),

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

where the last equality is due to Thm 4.1.

Finally, to bound A, we shall apply a modulus of continuity result on the suprema of empirical processes with respect to function $(g - g_0)$ and $(\log g - \log g_0)$. In particular, due to (19), 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 (18). Apply Lemma 5.14 of van de Geer (2000) we obtain that for $\delta_n = n^{-1/(2+\gamma_{\bar{g}})}$, there holds:

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

The overall estimation error is bounded by the upper bound of B. □

5 General methods for estimating f-divergence

In this section, we shall present several general methods for estimating f-divergence, and discuss their properties and limitations.

5.1 M-estimator of D_ϕ and p_0/q_0

It is not difficult to see that our method for estimating the KL divergence can be easily applied to any divergence $D_\phi(p_0, q_0)$. In fact, the method for consistency analysis, while tailored to each specific choice of ϕ, is also very similar in spirit. Assume in this section that ϕ is a differentiable (convex) function. Motivated by Lemma 2.1, our estimator has the following form:

$$\hat{D}_\phi := \sup_{f \in \mathcal{F}} \int f \, d\mathbb{Q}_n - \int \phi^*(f) \, d\mathbb{P}_n. \tag{20}$$

Let $\hat{f}$ be the supremum of the above optimization. $\hat{f}$ is considered an estimator of $f_0 = \phi'(q_0/p_0)$. As before, we define the estimation and approximation error. The latter is assumed to be 0, i.e., $\phi'(q_0/p_0) \in \mathcal{F}$.

$$\mathcal{E}_0^\phi(\mathcal{F}) = D_\phi(\mathbb{P}, \mathbb{Q}) - \sup_{f \in \mathcal{F}} \int (f \, d\mathbb{Q} - \phi^*(f) \, d\mathbb{P}) \geq 0 \tag{21}$$

$$\mathcal{E}_1^\phi(\mathcal{F}) = \sup_{f \in \mathcal{F}} \left| \int f d(\mathbb{Q}_n - \mathbb{Q}) - \phi^*(f) \, d(\mathbb{P}_n - \mathbb{P}) \right|. \tag{22}$$

From (20),(4) and (5), it is simple to see that: $-\mathcal{E}_1^\phi(\mathcal{F}) - \mathcal{E}_0^\phi(\mathcal{F}) \leq \hat{D}_\phi - D_\phi(\mathbb{P}, \mathbb{Q}) \leq \mathcal{E}_1(\mathcal{F})$. Since $\mathcal{E}_0^\phi(\mathcal{F}) = 0$, our main focus is in analysis of $\mathcal{E}_1^\phi(\mathcal{F})$. As before, define:

$$v_n^\phi(\mathcal{F}) = \sup_{f \in \mathcal{F}} \left| \int (\phi^*(f) - \phi^*(f_0)) d(\mathbb{P}_n - \mathbb{P}) - \int (f - f_0) d(\mathbb{Q}_n - \mathbb{Q}) \right|.$$

$$w_n^\phi(f_0) = \left| \int \phi^*(f_0) \, d(\mathbb{P}_n - \mathbb{P}) \, - f_0 d(\mathbb{Q}_n - \mathbb{Q}) \right|.$$

We have $\mathcal{E}_1^\phi(\mathcal{F}) \leq v_n^\phi(\mathcal{F}) + w_n^\phi(f_0)$. Since $w_n(\mathcal{F})$ converges to 0 almost surely under mild assumptions on g_0, to prove consistency of our estimator, it remains to analyze the convergence of $v_n^\phi(\mathcal{F})$. This can be done in the same manner as in Section 3.

To analyze the convergence rate of our estimator, the key idea of our analysis is to exploit the modulus of continuity of the supremum of the empirical processes involved in the definition of $v_n^\phi(\mathcal{F})$ with respect

the a notion of distance between g and f_0:

$$\begin{aligned} d_\phi(f_0, f) &:= D_\phi(\mathbb{P}, \mathbb{Q}) - \int f d\mathbb{Q} - \phi^*(f) d\mathbb{P} && (23) \\ &= \int (\phi^*(f) - \phi^*(f_0)) d\mathbb{P} - (f - f_0) d\mathbb{Q} && (24) \\ &= \int (\phi^*(f) - \phi^*(f_0) - \left.\frac{\partial \phi^*}{\partial f}\right|_{f_0} (f - f_0)\, d\mathbb{P} \geq 0. && (25) \end{aligned}$$

The last line in the above equation shows that d_ϕ is a *Bregman divergence* using convex function ϕ^*. The following lemma is an analogue of Lemma 3.2(ii) whose proof is straightforward:

Lemma 5.1. *If $\hat{f}$ is an estimation of f_0 by solving* (20)*, then $d_\phi(f_0, \hat{f}) \leq v_n^\phi(\mathcal{F})$.*

Since $d_\phi(f_0, f)$ is usually not a proper metric, to apply standard results from empirical process theory one usually needs to replace d_ϕ by a lower bound which is a proper metric (such as L_2 or Hellinger metric). In the case of KL divergence, we have seen that this lower bound is the Hellinger distance.

In the following we shall demonstrate our general method to the estimation yet another f-divergence: the χ-square distance. This divergence is very amenable to the general framework just described. As we shall see, it also plays a special role in another general method for divergence any f-divergence.

The χ-square divergence is defined as $D_\chi(\mathbb{P}, \mathbb{Q}) = \int p_0^2/q_0 d\mu$. It is a f-divergence with $\phi(u) = 1/u$. We have $\phi^*(v) = -2\sqrt{-v}$ if $v < 0$ and $+\infty$ otherwise. As a result, we only need to restrict $\mathcal{F}$ to the subset for which $f < 0$ for any $f \in \mathcal{F}$. Let $g := \sqrt{-f}$ and $\mathcal{G} = \sqrt{-\mathcal{F}}$. $\mathcal{G}$ is a function class of positive functions. We have $g_0 := \sqrt{-f_0} = \sqrt{-\phi'(q_0/p_0)} = p_0/q_0$. We shall also replace notation $d_\phi(f_0, f)$ by $d_\phi(g_0, g)$. For our choice of ϕ, we have:

$$\begin{aligned} d_\chi(g_0, g) &= d_\chi(f_0, f) = \int (-2\sqrt{-f} + 2\sqrt{-f_0}) d\mathbb{P} - (f - f_0) d\mathbb{Q} \\ &= \int (g_0 - g)(2p_0/q_0 - g_0 - g) d\mathbb{Q} \\ &= \int (g - g_0)^2 d\mathbb{Q} \\ v_n^\chi(\mathcal{G}) &= v_n^\chi(\mathcal{F}) = \sup_{g \in \mathcal{G}} \left| \int 2(g^2 - g_0^2) d(\mathbb{Q}_n - \mathbb{Q}) - \int (g - g_0) d(\mathbb{P}_n - \mathbb{P}) \right|. \end{aligned}$$

Assume moreover that for some constant $0 < \gamma < 2$,

$$\mathcal{H}_\delta^B(\mathcal{G}, L_2(\mathbb{Q})) \leq A_{\mathcal{G}} \delta^{-\gamma} \tag{26}$$

The following theorem is a parallel of Thm 4.1; the proof is essentially the same (if not simpler) and therefore not included herein:

Theorem 5.2. *Assume* (16) *and* (26)*, and $\hat{g}_n$ be our estimator of g_0 then: $d_\chi(g_0, \hat{g}_n) = O_{\mathbb{P}}(n^{-2/(\gamma+2)})$.*

Remark. Comparing with Thm 4.1 the assumption on the $L_2(\mathbb{Q})$-based entropy and the assertion on the $L_2(\mathbb{Q})$ metric are both weaker, because the $L_2(\mathbb{Q})$ metric is less strong than the Hellinger distance.

Finally, it is straightward to show that our method is applicable to a broader class of functional of the following form:

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

where $\psi : \mathcal{X} \to \mathbb{R}_+$ is a known positive function that is also bounded from both above and below (away from 0). All analysis goes through, with the insertion of ψ in all integrals involved. We also obtain the same convergence rate as when $\psi = 1$.

5.2 Plug-in estimator based on Taylor expansion

In this section we shall present an estimator based on functional delta method. This idea was also used by Joe (1989), Birgé and Massart (1995) to estimate integral functional of a density function. While $D_\phi(\mathbb{P}, \mathbb{Q})$ is a functional of two densities, we can exploit its special structure and our method of estimating the density ratio to achieve an estimator of with similar effects. Indeed, we can write

$$D_\phi(\mathbb{P}, \mathbb{Q}) = \int (p_0/q_0)\phi(q_0/p_0)\, d\mathbb{Q} = \int g_0\phi(1/g_0)\, d\mathbb{Q}.$$

Thus, D_ϕ can be viewed as an integral functional of $g_0 = p_0/q_0$. Of course, the difference here is that the integration is with respect to unknown $\mathbb{Q}$.

Suppose that $\phi : \mathbb{R}_+ \to \mathbb{R}$ is a differentiable convex function up to the third order, $\mathcal{G}$ is a smooth function class bounded from both above and below as in (15) (with smooth parameter α). Suppose that $\hat{g}_n$ is an estimator of g_0 such as the one described in the previous section, i.e., $\|\hat{g}_n - g_0\|_{L_2(\mathbb{Q})} = d_\chi(g_0, \hat{g}_n) = O_P(n^{-\alpha/(2\alpha+d)})$. Using a Taylor expansion around $\hat{g}_n$, we obtain:

$$\begin{aligned} g\phi(1/g) &= \hat{g}_n\phi(1/\hat{g}_n) + (g - \hat{g}_n)(\phi(1/\hat{g}_n) - \phi'(1/\hat{g}_n)/\hat{g}_n) + (g - \hat{g}_n)^2\phi''(1/\hat{g}_n)/\hat{g}_n^3 + \\ &\quad O((g - \hat{g}_n)^3) \\ &= \phi'(1/\hat{g}_n) + \phi''(1/\hat{g}_n)/\hat{g}_n + g(\phi(1/\hat{g}_n) - \phi'(1/\hat{g}_n)/\hat{g}_n - 2\phi''(1/\hat{g}_n)/\hat{g}_n^2) + \\ &\quad g^2\phi''(1/\hat{g}_n)/\hat{g}_n^3 + O((g - \hat{g}_n)^3). \end{aligned}$$

We arrive at

$$\begin{aligned} D_\phi(\mathbb{P}, \mathbb{Q}) &= \int g\phi(1/g) d\mathbb{Q} \\ &= \int \phi'(1/\hat{g}_n) + \phi''(1/\hat{g}_n)/\hat{g}_n\, d\mathbb{Q} + \\ &\quad \int (\phi(1/\hat{g}_n) - \phi'(1/\hat{g}_n)/\hat{g}_n - 2\phi''(1/\hat{g}_n)/\hat{g}_n^2)\, d\mathbb{P} + \\ &\quad \int p_0^2/q_0\phi''(1/\hat{g}_n)/\hat{g}_n^3\, d\mu + O(\|g_0 - \hat{g}_n\|_3^3). \end{aligned}$$

In the above expression, the first two integrals can be estimated from (other) sets of empirical data drawn from $\mathbb{P}$ and $\mathbb{Q}$. Because of the boundedness assumption, these estimations have at most $O_P(n^{-1/2})$ error. The error of our Taylor approximation is $O(\|g_0 - \hat{g}_n\|_3^3) = O_P(n^{-3\alpha/(2\alpha+d)})$. This rate is less than $O(n^{-1/2})$ for $\alpha \geq d/4$. Thus when $\alpha \geq d/4$, the optimal rate of convergence for estimating D_ϕ hinges on the rate of estimating the integral of the form $\int p_0^2/q_0\psi\, d\mu$.

Before ending this section, it is informative to return to the case of KL divergence, i.e., $\phi(u) = -\log u$. If we use Taylor approximation up to first order (thus guaranteeing an error rate of $O_P(n^{-2\alpha/(2\alpha+d)})$, the estimator has the following form:

$$\begin{aligned} \hat{D}_\phi &= \int (\phi(1/\hat{g}_n) - \phi'(1/\hat{g}_n)/\hat{g}_n)\, d\mathbb{P}_n + \int \phi'(1/\hat{g}_n)\, d\mathbb{Q}_n \\ &= \int \log \hat{g}_n + 1 d\mathbb{P}_n - \hat{g}_n d\mathbb{Q}_n, \end{aligned}$$

which has exactly the same form as our original estimator (3), except that here $\hat{g}_n$ can be any estimator of the density ratio. The estimator (3) achieves simultaneously both goals (i) estimating the density ratio and (ii) estimating the divergence. While our method for (i) achieves the optimal minimax bound, our method for (ii) can be viewed as only a first-order Taylor expansion based plug-in estimator. As discussed in the previous paragraph, it seems that one might obtain a better rate by using Taylor expansion up to second order. This is, of course, possible only if we can obtain a better rate for estimating the integral of the form $\int p_0^2/q_0\psi \, d\mu$.

6 M-estimation with penalties

In practice, the "true" size of $\mathcal{G}$ is not known. Accordingly, our approach in this paper is an alternative approach based on controlling the size of $\mathcal{G}$ by using penalties. More precisely, let $I(g)$ be a measure of complexity for g. Assume that I is a non-negative functional and $I(g_0) < \infty$. We decompose the function class $\mathcal{G}$ as follows:

$$\mathcal{G} = \cup_{1 \le M \le \infty} \mathcal{G}_M, \tag{27}$$

where $\mathcal{G}_M := \{g \mid I(g) \le M\}$ is a ball determined by $I(\cdot)$.

The estimation procedure involves solving the following program:

$$\hat{g}_n = \operatorname{argmin}_{g \in \mathcal{G}} \int g d\mathbb{Q}_n - \int \log g \, d\mathbb{P}_n + \frac{\lambda_n}{2} I^2(g), \tag{28}$$

where $\lambda_n > 0$ is a regularization parameter. The minimizing argument $\hat{g}_n$ is plugged into (3) to obtain an estimate of the KL divergence D_K.

For the KL divergence, the difference $|\hat{D}_K - D_K(\mathbb{P}, \mathbb{Q})|$ is a natural performance measure. For estimating the density ratio, various metrics are possible. Viewing $g_0 = p_0/q_0$ as a density function with respect to $\mathbb{Q}$ measure, one useful metric is the (generalized) Hellinger distance:

$$h^2_{\mathbb{Q}}(g_0, g) := \frac{1}{2} \int (g_0^{1/2} - g^{1/2})^2 \, d\mathbb{Q}. \tag{29}$$

For the analysis, several assumptions are in order. First, assume that g_0 (*not* all of $\mathcal{G}$) is bounded from above and below:

$$0 < \eta_0 \le g_0 \le \eta_1 \text{ for some constants } \eta_0, \eta_1. \tag{30}$$

Next, the uniform norm of $\mathcal{G}_M$ is Lipchitz with respect to the penalty measure $I(g)$, i.e.:

$$\sup_{g \in \mathcal{G}_M} |g|_\infty \le cM \quad \text{for any } M \ge 1. \tag{31}$$

Finally, on the bracket entropy of $\mathcal{G}$ van der Vaart and Wellner (1996): For some $0 < \gamma < 2$,

$$\mathcal{H}^B_\delta(\mathcal{G}_M, L_2(\mathbb{Q})) = O(M/\delta)^\gamma \quad \text{for any } \delta > 0. \tag{32}$$

The following is our main theoretical result, whose proof is given in Section 8:

Theorem 6.1. *(a) Under assumptions* (30) (31) (32)*, and set* $\lambda_n \to 0$ *so that:*

$$\lambda_n^{-1} = O_{\mathbb{P}}(n^{2/(2+\gamma)})(1 + I(g_0)),$$

then under $\mathbb{P}$*:*

$$h_{\mathbb{Q}}(g_0, \hat{g}_n) = O_{\mathbb{P}}(\lambda_n^{1/2})(1 + I(g_0)), \quad I(\hat{g}_n) = O_{\mathbb{P}}(1 + I(g_0)).$$

(b) If, in addition to (30) (31) (32)*, there holds* $\inf_{g \in \mathcal{G}} g(x) \ge \eta_0$ *for any* $x \in \mathcal{X}$*, then*

$$|\hat{D}_K - D_K(\mathbb{P}, \mathbb{Q})| = O_{\mathbb{P}}(\lambda_n^{1/2})(1 + I(g_0)). \tag{33}$$

7 Algorithm: Optimization and dual formulation

$\mathcal{G}$ is an RKHS. Our algorithm involves solving program (28), for some choice of function class $\mathcal{G}$. In our implementation, relevant function classes are taken to be a reproducing kernel Hilbert space induced by a Gaussian kernel. The RKHS's are chosen because they are sufficiently rich Saitoh (1988), and as in many learning tasks they are quite amenable to efficient optimization procedures Schölkopf and Smola (2002).

Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a Mercer kernel function Saitoh (1988). Thus, K is associated with a feature map $\Phi : \mathcal{X} \to \mathcal{H}$, where $\mathcal{H}$ is a Hilbert space with inner product $\langle ., . \rangle$ and for all $x, x' \in \mathcal{X}$, $K(x, x') = \langle \Phi(x), \Phi(x') \rangle$. As a reproducing kernel Hilbert space, any function $g \in \mathcal{H}$ can be expressed as an inner product $g(x) = \langle w, \Phi(x) \rangle$, where $\|g\|_{\mathcal{H}} = \|w\|_{\mathcal{H}}$. A kernel used in our simulation is the Gaussian kernel:

$$K(x, y) := e^{-\|x-y\|^2/\sigma},$$

where $\|.\|$ is the Euclidean metric in $\mathbb{R}^d$, and $\sigma > 0$ is a parameter for the function class.

Let $\mathcal{G} := \mathcal{H}$, and let the complexity measure be $I(g) = \|g\|_{\mathcal{H}}$. Thus, Eq. (28) becomes:

$$\min_w J := \min_w \frac{1}{n} \sum_{i=1}^{n} \langle w, \Phi(x_i) \rangle - \frac{1}{n} \sum_{j=1}^{n} \log \langle w, \Phi(y_j) \rangle + \frac{\lambda_n}{2} \|w\|_{\mathcal{H}}^2, \tag{34}$$

where $\{x_i\}$ and $\{y_j\}$ are realizations of empirical data drawn from $\mathbb{Q}$ and $\mathbb{P}$, respectively. The $\log$ function is extended take value $-\infty$ for negative arguments.

Lemma 7.1. *$\min_w J$ has the following dual form:*

$$-\min_{\alpha > 0} \sum_{j=1}^{n} -\frac{1}{n} - \frac{1}{n} \log n\alpha_j + \frac{1}{2\lambda_n} \sum_{i,j} \alpha_i \alpha_j K(y_i, y_j) + \frac{1}{2\lambda_n n^2} \sum_{i,j} K(x_i, x_j) - \frac{1}{\lambda_n n} \sum_{i,j} \alpha_j K(x_i, y_j).$$

Proof. Let $\psi_i(w) := \frac{1}{n} \langle w, \Phi(x_i) \rangle$, $\varphi_j(w) := -\frac{1}{n} \log \langle w, \Phi(y_j) \rangle$, and $\Omega(w) = \frac{\lambda_n}{2} \|w\|_{\mathcal{H}}^2$. We have

$$\begin{aligned} \min_w J &= -\max_w (\langle 0, w \rangle - J(w)) = -J^*(0) \\ &= -\min_{u_i, v_j} \sum_{i=1}^{n} \psi_i^*(u_i) + \sum_{j=1}^{n} \varphi_j^*(v_j) + \Omega^*(-\sum_{i=1}^{n} u_i - \sum_{j=1}^{n} v_j), \end{aligned}$$

where the last line is due to the inf-convolution theorem Rockafellar (1970). Simple calculations yield:

$$\begin{aligned} \varphi_j^*(v) &= -\frac{1}{n} - \frac{1}{n} \log n\alpha_j \text{ if } v = -\alpha_j \Phi(y_j) \text{ and } +\infty \text{ otherwise} \\ \psi_i^*(u) &= 0 \text{ if } u = \frac{1}{n} \Phi(x_i) \text{ and } +\infty \text{ otherwise} \\ \Omega^*(v) &= \frac{1}{2\lambda_n} \|v\|_{\mathcal{H}}^2. \end{aligned}$$

So, $\min_w J = -\min_{\alpha_i} \sum_{j=1}^{n} (-\frac{1}{n} - \frac{1}{n} \log n\alpha_j) + \frac{1}{2\lambda_n} \| \sum_{j=1}^{n} \alpha_j \Phi(y_j) - \frac{1}{n} \sum_{i=1}^{n} \Phi(x_i) \|_{\mathcal{H}}^2$, which implies the lemma immediately. □

If $\hat{\alpha}$ is solution of the dual formulation, it is not difficult to show that the optimal $\hat{w}$ is attained at $\hat{w} = \frac{1}{\lambda_n}(\sum_{j=1}^n \hat{\alpha}_j \Phi(y_j) - \frac{1}{n}\sum_{i=1}^n \Phi(x_i))$.

For an RKHS based on a Gaussian kernel, the entropy condition (32) holds for any $\gamma > 0$ Zhou (2002). Furthermore, (31) trivially holds via the Cauchy-Schwarz inequality: $|g(x)| = |\langle w,\, \Phi(x)\rangle| \leq \|w\|_{\mathcal{H}}\|\Phi(x)\|_{\mathcal{H}} \leq I(g)\sqrt{K(x,x)} \leq I(g)$. Thus, by Theorem 6.1(a), $\|\hat{w}\|_{\mathcal{H}} = \|\hat{g}_n\|_{\mathcal{H}} = O_{\mathbb{P}}(\|g_0\|_{\mathcal{H}})$, so the penalty term $\lambda_n\|\hat{w}\|^2$ vanishes at the same rate as λ_n. We have arrived at the following estimator for the KL divergence:

$$\hat{D}_K = 1 + \sum_{j=1}^n (-\frac{1}{n} - \frac{1}{n}\log n\hat{\alpha}_j) = \sum_{j=1}^n -\frac{1}{n}\log n\hat{\alpha}_j.$$

$\log\mathcal{G}$ is an RKHS. Alternatively, we could set $\log\mathcal{G}$ to be the RKHS, letting $g(x) = \exp\langle w,\, \Phi(x)\rangle$, and letting $I(g) = \|\log g\|_{\mathcal{H}} = \|w\|_{\mathcal{H}}$. Theorem 6.1 is not applicable in this case, because condition (31) no longer holds, but this choice nonetheless seems reasonable and worth investigating, because in effect we have a far richer function class which might improve the bias of our estimator when the density ratio is not very smooth.

A derivation similar to the previous case yields the following convex program:

$$\begin{aligned}\min_w J &:= \min_w \frac{1}{n}\sum_{i=1}^n e^{\langle w,\, \Phi(x_i)\rangle} - \frac{1}{n}\sum_{j=1}^n \langle w,\, \Phi(y_j)\rangle + \frac{\lambda_n}{2}\|w\|_{\mathcal{H}}^2 \\ &= -\min_{\alpha>0} \sum_{i=1}^n \alpha_i \log(n\alpha_i) - \alpha_i + \frac{1}{2\lambda_n}\|\sum_{i=1}^n \alpha_i\Phi(x_i) - \frac{1}{n}\sum_{j=1}^n \Phi(y_j)\|_{\mathcal{H}}^2.\end{aligned}$$

Letting $\hat{\alpha}$ be the solution of the above convex program, the KL divergence can be estimated by:

$$\hat{D}_K = 1 + \sum_{i=1}^n \hat{\alpha}_i \log\hat{\alpha}_i + \hat{\alpha}_i \log\frac{n}{e}.$$

8 Proof of Theorem 6.1

We now sketch out the proof of the main theorem. The key to our analysis is the following lemma:

Lemma 8.1. *If $\hat{g}_n$ is an estimate of g using* (28)*, then:*

$$\frac{1}{4}h_{\mathbb{Q}}^2(g_0,\hat{g}_n) + \frac{\lambda_n}{2}I^2(\hat{g}_n) \leq -\int(\hat{g}_n - g_0)d(\mathbb{Q}_n - \mathbb{Q}) + \int 2\log\frac{\hat{g}_n + g_0}{2g_0}d(\mathbb{P}_n - \mathbb{P}) + \frac{\lambda_n}{2}I^2(g_0).$$

Proof. Define $d_l(g_0, g) = \int(g - g_0)d\mathbb{Q} - \log\frac{g}{g_0}d\mathbb{P}$. Note that for $x > 0$, $\frac{1}{2}\log x \leq \sqrt{x} - 1$. Thus,

$$\int \log\frac{g}{g_0}\, d\mathbb{P} \leq 2\int(g^{1/2}g_0^{-1/2} - 1)\, d\mathbb{P}.$$

As a result, for any g, d_l is related to $h_{\mathbb{Q}}$ as follows:

$$\begin{aligned}d_l(g_0,g) &\geq \int(g - g_0)\, d\mathbb{Q} - 2\int(g^{1/2}g_0^{-1/2} - 1)\, d\mathbb{P} \\ &= \int(g - g_0)\, d\mathbb{Q} - 2\int(g^{1/2}g_0^{1/2} - g_0)\, d\mathbb{Q} = \int(g^{1/2} - g_0^{1/2})^2 d\mathbb{Q} \\ &= 2h_{\mathbb{Q}}^2(g_0,g).\end{aligned}$$

By the definition (28) of our estimator, we have:

$$\int \hat{g}_n d\mathbb{Q}_n - \int \log \hat{g}_n d\mathbb{P}_n + \frac{\lambda_n}{2} I^2(\hat{g}_n) \leq \int g_0 d\mathbb{Q}_n - \int \log g_0 d\mathbb{P}_n + \frac{\lambda_n}{2} I^2(g_0).$$

Both sides are convex functionals of g. By Jensen's inequality, if F is a convex function, then $F((u+v)/2) - F(v) \leq (F(u) - F(v))/2$. We obtain:

$$\int \frac{\hat{g}_n + g_0}{2} d\mathbb{Q}_n - \int \log \frac{\hat{g}_n + g_0}{2} d\mathbb{P}_n + \frac{\lambda_n}{4} I^2(\hat{g}_n) \leq \int g_0 d\mathbb{Q}_n - \int \log g_0 d\mathbb{P}_n + \frac{\lambda_n}{4} I^2(g_0).$$

Rearranging, $\int \frac{\hat{g}_n - g_0}{2} d(\mathbb{Q}_n - \mathbb{Q}) - \int \log \frac{\hat{g}_n + g_0}{2g_0} d(\mathbb{P}_n - \mathbb{P}) + \frac{\lambda_n}{4} I^2(\hat{g}_n) \leq$

$$\begin{aligned}\int \log \frac{\hat{g}_n + g_0}{2g_0} d\mathbb{P} - \int \frac{\hat{g}_n - g_0}{2} d\mathbb{Q} + \frac{\lambda_n}{4} I^2(g_0) &= -d_l(g_0, \frac{g_0 + \hat{g}_n}{2}) + \frac{\lambda_n}{4} I^2(g_0) \\ &\leq -2h^2_{\mathbb{Q}}(g_0, \frac{g_0 + \hat{g}_n}{2}) + \frac{\lambda_n}{4} I^2(g_0) \leq -\frac{1}{8} h^2_{\mathbb{Q}}(g_0, \hat{g}_n) + \frac{\lambda_n}{4} I^2(g_0),\end{aligned}$$

where the last inequality is a standard result for the (generalized) Hellinger distance (cf. van de Geer (2000)). □

Let us now proceed to part (a) of the theorem. Define $f_g := \log \frac{g+g_0}{2g_0}$, and let $\mathcal{F}_M := \{f_g | g \in \mathcal{G}_M\}$. Since f_g is a Lipschitz function of g, conditions (30) and (32) imply that

$$\mathcal{H}^B_\delta(\mathcal{F}_M, L_2(\mathbb{P})) = O(M/\delta)^\gamma. \tag{35}$$

Apply Lemma 5.14 of van de Geer (2000) using distance metric $d_2(g_0, g) = \|g - g_0\|_{L_2(\mathbb{Q})}$, the following is true under $\mathbb{Q}$ (and so true under $\mathbb{P}$ as well, since $d\mathbb{P}/d\mathbb{Q}$ is bounded from above),

$$\sup_{g \in \mathcal{G}} \frac{|\int (g - g_0) d(\mathbb{Q}_n - \mathbb{Q})|}{n^{-1/2} d_2(g_0, g)^{1-\gamma/2} (1 + I(g) + I(g_0))^{\gamma/2} \vee n^{-\frac{2}{2+\gamma}} (1 + I(g) + I(g_0))} = O_{\mathbb{P}}(1). \tag{36}$$

In the same vein, we obtain that under $\mathbb{P}$ measure:

$$\sup_{g \in \mathcal{G}} \frac{|\int f_g d(\mathbb{P}_n - \mathbb{P})|}{n^{-1/2} d_2(g_0, g)^{1-\gamma/2} (1 + I(g) + I(g_0))^{\gamma/2} \vee n^{-\frac{2}{2+\gamma}} (1 + I(g) + I(g_0))} = O_{\mathbb{P}}(1) \tag{37}$$

By condition (31), it is easy to see that:

$$d_2(g_0, g) = \|g - g_0\|_{L_2(\mathbb{Q})} \leq 2c^{1/2} (1 + I(g) + I(g_0))^{1/2} h_{\mathbb{Q}}(g_0, g).$$

Combining Lemma 8.1 and Eqs. (37), (36), we obtain the following:

$$\begin{aligned}\frac{1}{4} h^2_{\mathbb{Q}}(g_0, \hat{g}_n) + \frac{\lambda_n}{2} I^2(\hat{g}_n) \leq \lambda_n I(g_0)^2/2 + \\ O_{\mathbb{P}}\Big(n^{-1/2} h_{\mathbb{Q}}(g_0, g)^{1-\gamma/2} (1 + I(g) + I(g_0))^{1/2+\gamma/4} \ \vee \ n^{-\frac{2}{2+\gamma}} (1 + I(g) + I(g_0)) \Big).\end{aligned} \tag{38}$$

From this point, the proof involves simple algebraic manipulation of (38). To simplify notation, let $\hat{h} = h_{\mathbb{Q}}(g_0, \hat{g}_n)$, $\hat{I} = I(\hat{g}_n)$, and $I_0 = I(g_0)$. There are four possibilities:

Case a. $\hat{h} \geq n^{-1/(2+\gamma)}(1+\hat{I}+I_0)^{1/2}$ and $\hat{I} \geq 1+I_0$. From (38), either

$$\hat{h}^2/4 + \lambda_n \hat{I}^2/2 \leq O_{\mathbb{P}}(n^{-1/2})\hat{h}^{1-\gamma/2}\hat{I}^{1/2+\gamma/4} \text{ or } \hat{h}^2/4 + \lambda_n \hat{I}^2/2 \leq \lambda_n I_0^2/2,$$

which implies, respectively, either

$$\hat{h} \leq \lambda_n^{-1/2} O_{\mathbb{P}}(n^{-2/(2+\gamma)}), \quad \hat{I} \leq \lambda_n^{-1} O_{\mathbb{P}}(n^{-2/(2+\gamma)}).$$

or

$$\hat{h} \leq O_{\mathbb{P}}(\lambda_n^{1/2} I_0), \quad \hat{I} \leq O_{\mathbb{P}}(I_0).$$

Both scenarios conclude the proof if we set $\lambda_n^{-1} = O_{\mathbb{P}}(n^{2/(\gamma+2)}(1+I_0))$.
Case b. $\hat{h} \geq n^{-1/(2+\gamma)}(1+\hat{I}+I_0)^{1/2}$ and $\hat{I} < 1+I_0$. From (38), either

$$\hat{h}^2/4 + \lambda_n \hat{I}^2/2 \leq O_{\mathbb{P}}(n^{-1/2})\hat{h}^{1-\gamma/2}(1+I_0)^{1/2+\gamma/4} \text{ or } \hat{h}^2/4 + \lambda_n \hat{I}^2/2 \leq \lambda_n I_0^2/2,$$

which implies, respectively, either

$$\hat{h} \leq (1+I_0)^{1/2} O_{\mathbb{P}}(n^{-1/(\gamma+2)}), \quad \hat{I} \leq 1+I_0$$

or

$$\hat{h} \leq O_{\mathbb{P}}(\lambda_n^{1/2} I_0), \quad \hat{I} \leq O_{\mathbb{P}}(I_0).$$

Both scenarios conclude the proof if we set $\lambda_n^{-1} = O_{\mathbb{P}}(n^{2/(\gamma+2)}(1+I_0))$.
Case c. $\hat{h} \leq n^{-1/(2+\gamma)}(1+\hat{I}+I_0)^{1/2}$ and $\hat{I} \geq 1+I_0$. From (38)

$$\hat{h}^2/4 + \lambda_n \hat{I}^2/2 \leq O_{\mathbb{P}}(n^{-2/(2+\gamma)})\hat{I},$$

which implies that $\hat{h} \leq O_{\mathbb{P}}(n^{-1/(2+\gamma)})\hat{I}^{1/2}$ and $\hat{I} \leq \lambda_n^{-1} O_{\mathbb{P}}(n^{-2/(2+\gamma)})$. This means that

$$\hat{h} \leq O_{\mathbb{P}}(\lambda_n^{1/2})(1+I_0), \quad \hat{I} \leq O_{\mathbb{P}}(1+I_0)$$

if we set $\lambda_n^{-1} = O_{\mathbb{P}}(n^{2/(2+\gamma)})(1+I_0)$.
Case d. $\hat{h} \leq n^{-1/(2+\gamma)}(1+\hat{I}+I_0)^{1/2}$ and $\hat{I} \leq 1+I_0$. Part (a) of the theorem is immediate.

Finally, part (b) is a simple consequence of part (a) using the same argument as in Thm 4.3.

9 Simulation results

In this section, we describe the results of various simulations that demonstrate the practical viability of our estimators, as well as their convergence behavior. We experimented with our estimators using various choices of $\mathbb{P}$ and $\mathbb{Q}$, including Gaussian, beta, mixture of Gaussians, and multivariate Gaussian distributions. Here we report results in terms of KL estimation error. For each of the eight estimation problems described here, we experiment with increasing sample sizes (the sample size, n, ranges from 100 to 10^4 or more). Error bars are obtained by replicating each set-up 250 times.

For all simulations, we report our estimator's performance using the the simple fixed rate $\lambda_n \sim 1/n$, noting that this may be a suboptimal rate. We set the kernel width to be relatively small ($\sigma = .1$) for one-dimension data, and larger σ for higher dimensions. We use M1 to denote the method in which $\mathcal{G}$ is the RKHS, and M2 for the method in which $\log \mathcal{G}$ is the RKHS. Our methods are compared to algorithm A in Wang et al Wang et al. (2005), which was shown empirically to be one of the best methods in the

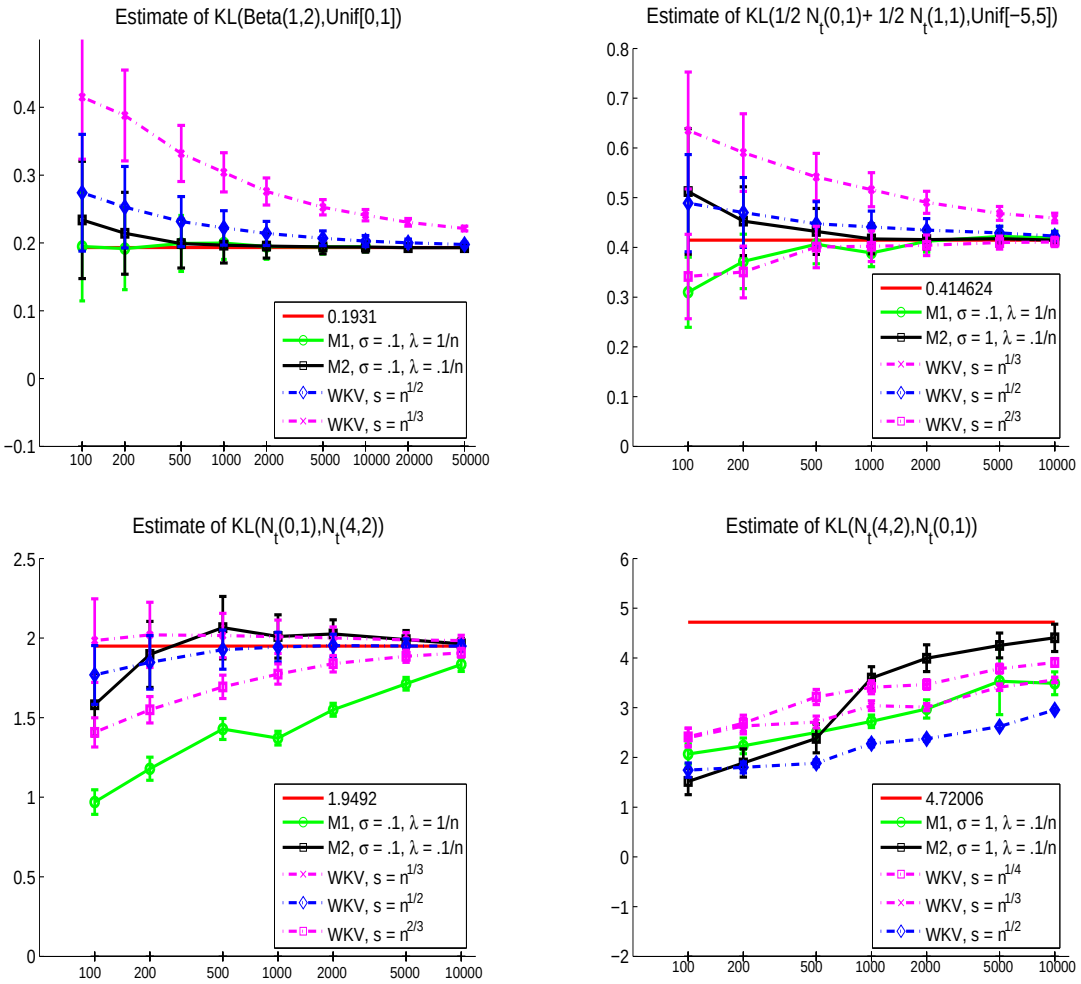

Figure 1. Results of estimating KL divergences for various choices of probability distributions. In all plots, the X-axis is the number of data points plotted on a log scale, and the Y-axis is the estimated value. The error bar is obtained by replicating the experiment 250 times. $N_t(a, I_k)$ denotes a truncated normal distribution of k dimensions with mean $(a, \ldots, a)$ and identity covariance matrix.

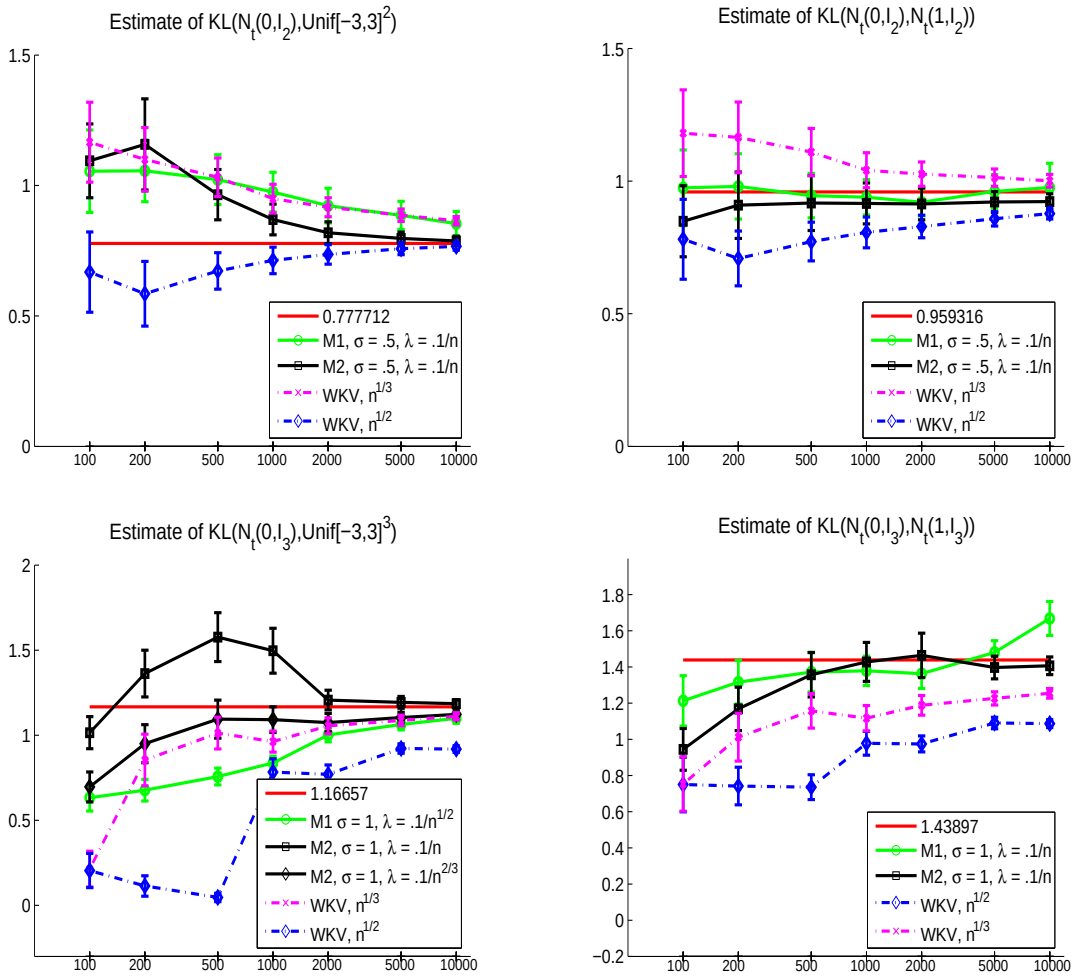

Figure 2. Results of estimating KL divergences for various choices of probability distributions. In all plots, the X-axis is the number of data points plotted on a log scale, and the Y-axis is the estimated value. The error bar is obtained by replicating the experiment 250 times. $N_t(a, I_k)$ denotes a truncated normal distribution of k dimensions with mean $(a, \ldots, a)$ and identity covariance matrix.

literature. Their method, to be denoted by WKV, is based on data-dependent partitioning of the covariate space. Naturally, the performance of WKV is critically dependent on the amount s of data allocated to each partition; here we report results with $s \sim n^{\gamma}$, where $\gamma = 1/3, 1/2, 2/3$.

The first four plots present results with univariate distributions. In the first two, our estimators $M1$ and $M2$ appear to have faster convergence rate than WKK. The WKV estimator performs very well in the third example, but rather badly in the fourth example. The next four plots present results with two and three dimensional data. Again, M1 has the best convergence rates in all examples. The M2 estimator does not converge in the last example, suggesting that the underlying function class exhibits very strong bias. WKV have weak convergence rates despite different choices of the partition sizes. It is worth noting that as one increases the number of dimensions, histogram based methods such as WKV become increasingly difficult to implement, whereas increasing dimension has only a mild effect on our method.

References

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.

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.

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

L. Birgé and P. Massart. Estimation of integral functionals of a density. *Ann. Statist.*, 23(1):11–29, 1995.

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.

M. Broniatowski and A. Keziou. Parametric estimation and tests through divergences. Technical report, LSTA, Université Pierre et Marie Curie, 2004.

T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, 1991.

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

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.

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

A. Hyvarinen, J. Karhunen, and E. Oja. *Independent component analysis*. John Wiley & Sons, Inc, 2001.

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

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

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

A. Keziou. Dual representation of ϕ-divergences and applications. *C. R. Acad. Sci. Paris, Ser. I 336*, pages 857–862, 2003.

B. Laurent. Efficient estimation of integral functionals of a density. *Ann. Statist.*, 24(2):659–681, 1996.

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

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

X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In *Advances in Neural Information Processing Systems 20 (NIPS)*, 2007a.

X. Nguyen, M. J. Wainwright, and M. I. Jordan. Nonparametric estimation of the likelihood ratio and divergence functionals. In *International Symposium on Information Theory (ISIT)*, 2007b.

G. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970.

S. Saitoh. *Theory of Reproducing Kernels and its Applications*. Longman, Harlow, UK, 1988.

B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

J. Shawe-Taylor and N. Cristianini. *Kernel methods for pattern analysis*. Cambridge Univ Press, 2004.

B. W. Silverman. On the estimation of a probability density function by the maximum penalized likelihood method. *Annals of Statistics*, 10:795–810, 1982.

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

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

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

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

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.

Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27(5):1564–1599, 1999.

B. Yu. Assouad, Fano and Le Cam. *Research Papers in Probability and Statistics: Festschrift in Honor of Lucien Le Cam*, pages 423–435, 1996.

D. X. Zhou. The covering number in learning theory. *Journal of Complexity*, 18:739–767, 2002.

10 Appendix

Two theorems (Thm 3.7 and 5.11) in van de Geer (2000) that were invoked frequently in this paper:

Theorem 10.1. *Let G be the envelope function for $\mathcal{G}$. Assume that $\int G d\mathbb{P} < \infty$, and suppose that for any $\delta > 0$, $\frac{1}{n}\mathcal{H}_\delta(\mathcal{G}, L_1(\mathbb{P}_n)) \xrightarrow{\mathbb{P}} 0$, then $\sup_{g\in\mathcal{G}} \int g d(\mathbb{P}_n - \mathbb{P}) \xrightarrow{a.s.} 0$.*

Theorem 10.2. *Let K, R be some constants, $\mathcal{G}$ satisfy $\sup_{g\in\mathcal{G}} \rho_K(g) \leq R$. If there hold, for some sufficiently large universal constant C:*

$$\begin{aligned} a &\leq C_1\sqrt{n}R^2/K \\ a &\geq C_0\bigg(\int_0^R \mathcal{H}_u^B(\mathcal{G}, \rho_K)^{1/2} du \vee R\bigg) \\ C_0^2 &\geq C^2(C_1+1), \end{aligned}$$

then

$$P\bigg(\sup_{g\in\mathcal{G}} |\sqrt{n}\int g d(\mathbb{P}_n - \mathbb{P})| \geq a\bigg) \leq C\exp\bigg[-\frac{a^2}{C^2(C_1+1)R^2}\bigg].$$