Estimating divergence functionals and the likelihood ratio by convex risk minimization

XuanLong Nguyen
Dept. of Statistical Science
Duke University
xuanlong.nguyen@stat.duke.edu

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

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

September 4, 2008

Abstract

We develop and analyze M-estimation methods for divergence functionals and the likelihood ratios of two probability distributions. Our method is based on a non-asymptotic variational characterization of f-divergences, which allows the problem of estimating divergences to be tackled via convex empirical risk optimization. The resulting estimators are simple to implement, requiring only the solution of standard convex programs. We present an analysis of consistency and convergence for these estimators. Given conditions only on the ratios of densities, we show that our estimators can achieve optimal minimax rates for the likelihood ratio in certain regimes. We derive an efficient optimization algorithm for computing our estimates, and illustrate their convergence behavior and practical viability by simulations.[1]

1   Introduction

Divergences (or pseudo-distances) based on likelihood ratios between pairs of multivariate probability distribution densities play a central role in information theory and statistics. For instance, in the asymptotic analysis of hypothesis testing, the Kullback-Leibler and Chernoff divergences control the decay rates of error probabilities (e.g., see Stein's lemma [7] and its variants). As a particular case of the Kullback-Leibler divergence, the mutual information specifies capacities in

---

[1]Preliminary versions of this work were presented at the International Symposium on Information Theory (2007) [18] and the Neural Information Processing Systems Conference (2007) [17].

channel coding coding and data compression [7]. In statistical machine learning and signal processing, divergences between probability distributions are frequently exploited as metrics to be optimized, such as in independent component analysis [6] and decentralized detection [25].

In all of these settings, an important problem is that of *divergence estimation*: how to estimate the divergence between two multivariate probability distributions, say $\mathbb{P}$ and $\mathbb{Q}$, based on a set of samples from each distribution? A canonical example is estimation of the Kullback-Leibler (KL) divergence from samples. This problem includes as a special case the problem of estimating the mutual information, corresponding to the KL divergence between a joint distribution and the product of its marginals, as well as the problem of estimating the Shannon entropy of a distribution $\mathbb{P}$, which is related to the KL divergence between $\mathbb{P}$ and the uniform distribution. Several researchers have studied the problem of Shannon entropy estimation [9, 13, 10] based on various types of non-parametric techniques. Somewhat more generally, the problem of estimating an integral functional of a single density has been studied extensively, dating back to early work [12, 16] from 1970s, and continuing on in later research [2, 3, 15]. More recent recent work by Wang et al. [29] has developed algorithms for estimating the KL divergence between a pair of continuous distributions $\mathbb{P}$ and $\mathbb{Q}$, based on building data-dependent partitions of equivalent (empirical) $\mathbb{Q}$-measure. Wang et al. [30] also proposed an interesting nonparametric estimator of the KL divergence using 1-nearest neighbor technique. Both estimators were empirically shown to outperform direct plug-in methods, but no theoretical results on convergence rates were provided.

In this paper, we propose methods for estimating divergence functionals as well as likelihood density ratios based on simple M-estimators. Although our primary interest is the Kullback-Leibler divergence, our methodology is more broadly applicable to the class of Ali-Silvey distances, also known as f-divergences [1, 8]. Any divergence in this family, to be defined more 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 $d\mathbb{Q}/d\mathbb{P}$.

Our estimation method is motivated by a non-asymptotic characterization of f-divergence, due independently to several authors [5, 14, 19]. Roughly speaking, the main theorem in [19] 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 for a risk minimization problem. This correspondence provides a *variational characterization*, by which the divergence $D_\phi(\mathbb{P}, \mathbb{Q})$ can be expressed as the maximum of an Bayes decision problem involving two hypotheses $\mathbb{P}$ and $\mathbb{Q}$. In this way, as we show in this paper, estimating the divergence $D_\phi(\mathbb{P}, \mathbb{Q})$ has an equivalent reformulation as solving a certain Bayes decision problem. This reformulation leads to an M-estimation procedure, in which the divergence is estimated by solving the convex program defined by the Bayes decision problem. This approach not only leads to an M-estimation procedure for the divergence but also for the likelihood ratio $d\mathbb{P}/d\mathbb{Q}$.

Our second contribution is to analyze the convergence and consistency properties of our estimators, under certain assumptions on the permitted class $\mathcal{G}$ of density ratios, or logarithms of density ratios. The analysis makes use of some known results in empirical process theory for nonparametric density estimation [26, 28]. The key technical condition is the continuity of the suprema of two empirical processes, induced by $\mathbb{P}$ and $\mathbb{Q}$ distributions respectively, with respect to a metric defined on the class $\mathcal{G}$ of permitted functions. This metric arises as a surrogate lower bound of a Bregman divergence defined on a pair of density ratios.

Our third contribution is to provide an efficient implementation of one version of our estimator,

2

in which the function class $\mathcal{G}$ is approximated by a reproducing kernel Hilbert space (RKHS) [21]. After computing the convex dual, the estimator can be implemented by solving a simple convex program involving only the Gram matrix defined by the kernel associated with the RKHS. Our method thus inherits the simplicity of other kernel-based methods used in statistical machine learning [22, 23]. We illustrate the empirical behavior of our estimator on various instances of KL divergence estimation.

The remainder of this paper is organized as follows. In Section 2, we provide the variational characterization of f-divergences in general, and KL divergence in particular. We then develop an M-estimator for the KL divergence and the likelihood ratio. Section 3 is devoted to the analysis of consistency and convergence rates of these estimators. In Section 4, we develop an efficient kernel-based method for computing our M-estimates, and provide simulation results demonstrating their performance. In Section 5, we discuss our estimation method and its analysis in a more general light, encompassing a broader class of f-divergences. We conclude in Section 6.

Notation: For convenience of the reader, we summarize some notation to be used throughout the paper. Given a probability distribution $\mathbb{P}$ and random variable f measureable with respect to $\mathbb{P}$, we use $\int f d\mathbb{P}$ to denote the expectation of f under $\mathbb{P}$. When $\mathbb{P}$ is absolutely continuous with respect to Lesbesgue measure, say with density p_0, this integral is equivalent to the usual Lebesgue integral $\int f p_0 d\mu = \int f(x) p_0(x) \mu(dx)$. Given samples $X^{(1)}, \ldots, X^{(n)}$ from $\mathbb{P}$, the empirical distribution $\mathbb{P}_n$ is given by $\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X^{(i)}}$, corresponding to a sum of delta functions centered at the data points. We use $\int f d\mathbb{P}_n$ as a convenient short-hand for the empirical expectation $\frac{1}{n} \sum_{i=1}^{n} f(X^{(i)})$.

2 M-estimators for KL divergence and the density ratio

We begin by defining f-divergences, and describing a variational characterization in terms of a Bayes decision problem. We then exploit this variational characterization to develop an M-estimator.

2.1   Variational characterization of f-divergence

Consider two probability distributions $\mathbb{P}$ and $\mathbb{Q}$, with $\mathbb{P}$ absolutely continuous with respect to $\mathbb{Q}$. Assume moreover that both distributions 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 (KL) divergence between $\mathbb{P}$ and $\mathbb{Q}$ is defined by the integral

$$D_K(\mathbb{P}, \mathbb{Q}) \quad := \quad \int p_0 \log \frac{p_0}{q_0} \, d\mu. \tag{1}$$

This divergence is a special case of a broader class of divergences known as Ali-Silvey distances or f-divergence [8, 1], which take the form

$$D_\phi(\mathbb{P}, \mathbb{Q}) \quad := \quad \int p_0 \phi(q_0/p_0) \, d\mu, \tag{2}$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is a convex and lower semi-continuous (l.s.c.) function. Different choices of ϕ result in many divergences that play important roles in information theory and statistics, including not only the KL divergence (1) but also the total variational distance, the Hellinger distance, and so on; see [24] for further examples.

We begin by stating and proving a variational representation for the divergence D_ϕ. In order to do so, we require some basic definitions from convex analysis [20, 11]. The *subdifferential* $\partial\phi(t)$ of the convex function ϕ at a point $t \in \mathbb{R}$ is the set

$$\partial\phi(t) \quad := \quad \{z \in \mathbb{R} \mid \phi(s) \geq \phi(t) + z\,(s - t) \qquad \forall s \in \mathbb{R}\}. \tag{3}$$

As a special case, if ϕ is differentiable at t, then $\partial\phi(t) = \{\phi'(t)\}$. The function ϕ^* is the *conjugate dual function* associated with ϕ, defined as

$$\phi^*(v) \quad := \quad \sup_{u \in \mathbb{R}} \{u\,v - \phi(u)\}. \tag{4}$$

With these definitions, we have:

Lemma 1. *For any class of functions $\mathcal{F}$ mapping from $\mathcal{X}$ to $\mathbb{R}$, we have the lower bound*

$$D_\phi(\mathbb{P}, \mathbb{Q}) \geq \sup_{f \in \mathcal{F}} \int [f\,d\mathbb{Q} - \phi^*(f)\,d\mathbb{P}]. \tag{5}$$

Equality holds if and only if the subdifferential $\partial\phi(q_0/p_0)$ contains an element of $\mathcal{F}$.

Proof. Since ϕ is convex and lower semi-continuous, Fenchel convex duality [20] guarantees that we can write ϕ in terms of its conjugate dual as $\phi(u) = \sup_{v \in \mathbb{R}} \{uv - \phi^*(v)\}$. Consequently, we have

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

where the supremum is taken over all measurable functions $f : \mathcal{X} \to \mathbb{R}$. It is simple to see that equality in the supremum is attained at a 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 $x \in \mathcal{X}$. □

2.2 M-estimators for the KL divergence and likelihood ratio

We now describe how the variational representation (5) specializes to an M-estimator for the Kullback-Leibler (KL) divergence. As a particular f-divergence, the KL divergence is induced by the convex function

$$\phi(t) \quad = \quad \begin{cases} -\log(t) & \text{for } t > 0 \\ +\infty & \text{for } t \leq 0. \end{cases} \tag{6}$$

A short calculation shows that the conjugate dual takes the form

$$\phi^*(v) \quad = \quad \begin{cases} -1 - \log(-v) & \text{if } u < 0, \text{ and} \\ +\infty & \text{otherwise.} \end{cases} \tag{7}$$

As a consequence of Lemma 1, we obtain the following representation of the Kullback-Leibler divergence: $D_K(\mathbb{P}, \mathbb{Q}) = \sup_{f<0} \int f \, d\mathbb{Q} - \int [-1 - \log(-f)] d\mathbb{P}$. After the substitution $g = -f$, this can be written as

$$D_K(\mathbb{P}, \mathbb{Q}) = \sup_{g>0} \int \log g \, d\mathbb{P} - \int g d\mathbb{Q} + 1, \tag{8}$$

for which the supremum is attained at $g_0 = p_0/q_0$.

We now take a statistical perspective on the variational problem (8), where we assume that the distributions $\mathbb{P}$ and $\mathbb{Q}$ are unknown. We suppose that we are given independent and identically distributed (i.i.d.) samples, say $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$ drawn i.i.d. from $\mathbb{P}$, and $Y^{(1)}, Y^{(2)}, \ldots, Y^{(n)}$ drawn i.i.d. from $\mathbb{Q}$. Denote by $\mathbb{P}_n$ the empirical distribution defined by the samples $\{X^{(1)}, \ldots, X^{(n)}\}$, given explicitly by $\mathbb{P}_n := \frac{1}{n} \sum_{i=1}^n \delta_{X^{(i)}}$, with the empirical distribution $\mathbb{Q}_n$ associated with $\{Y^{(1)}, \ldots, Y^{(n)}\}$ defined analogously. We consider two classes of estimators:

Estimator E1: Given the empirical distributions, we consider the estimator obtained by replacing the true distributions $\mathbb{P}$ and $\mathbb{Q}$ with their empirical versions, and maximizing over some pre-specified class $\mathcal{G}$ of functions $g : \mathcal{X} \to \mathbb{R}_+$, as follows:

$$\widehat{g}_n = \operatorname{argmax}_{g \in \mathcal{G}} \int \log g \, d\mathbb{P}_n - \int g \, d\mathbb{Q}_n, \tag{9a}$$

$$\widehat{D}_K = \int \log \widehat{g}_n \, d\mathbb{P}_n - \int \widehat{g}_n \, d\mathbb{Q}_n + 1. \tag{9b}$$

Assuming that $\mathcal{G}$ is a convex set of functions, the implementation of the estimator requires solving a convex optimization problem, albeit over an (infinite-dimensional) function space. For this estimation method to work, several conditions on $\mathcal{G}$ are required. First, so as to control approximation error, it is natural to require that $\mathcal{G}$ is sufficiently rich so as to contain the true likelihood ratio g_0 in the sense of KL divergence, i.e., there is some $g \in \mathcal{G}$ such that $g = g_0$ a.e.. On the other hand, $\mathcal{G}$ should not be too large, so that estimation is possible. To formalize this condition, let $I(g)$ be a measure of complexity for g, where I is a non-negative functional and $I(g_0) < \infty$. Given some fixed finite constant $M^* \geq I(g_0)$, we then define

$$\mathcal{G} := \mathcal{G}_{M^*} := \{g : I(g) \leq M^*\}. \tag{10}$$

Estimator E2: In practice, the "true" $I(g_0)$ is not known, and hence it is not clear as a practical matter how to choose the fixed M^* defining estimator E1. Thus we also consider an approach that involves an explicit penalty $I(g)$. In this approach, let

$$\mathcal{G} = \cup_{1 \leq M \leq \infty} \mathcal{G}_M. \tag{11}$$

The estimation procedure involves solving the following program:

$$\widehat{g}_n = \operatorname{argmin}_{g \in \mathcal{G}} \int g d\mathbb{Q}_n - \int \log g \, d\mathbb{P}_n + \frac{\lambda_n}{2} I^2(g), \tag{12a}$$

$$\widehat{D}_K = \int \log \widehat{g}_n \, d\mathbb{P}_n - \int \widehat{g}_n \, d\mathbb{Q}_n + 1, \tag{12b}$$

where $\lambda_n > 0$ is a regularization parameter.

As we discuss in Section 4, for function classes defined by reproducing kernel Hilbert spaces, problems (9b) and (12a) can actually be posed as a finite-dimensional convex programs (in n dimensions), and solved efficiently by standard methods. In addition to the estimate $\widehat{D}_K$ of the KL divergence, if the supremum is attained at $\widehat{g}_n$, then $\widehat{g}_n$ is an M-estimator of the density ratio $g_0 := p_0/q_0$.

In the next section, we present results regarding the consistency and convergence rates of both estimators. While these methods have similar convergence behavior, estimator E1 is somewhat simpler to analyze and admits weaker conditions for consistency. On the other hands, estimator E2 seems more practical. Details of algorithmic derivations for estimator E2 are described in Section 4.

3 Consistency and convergence rate analysis

For the KL divergence functional, the difference $|\widehat{D}_K - D_K(\mathbb{P}, \mathbb{Q})|$ is a natural performance measure. For estimating the density ratio function, this difference can also be used, although more direct metrics are customarily preferred. In our analysis, we view $g_0 = p_0/q_0$ as a density function with respect to $\mathbb{Q}$ measure, and adopt the (generalized) Hellinger distance as a performance measure for estimating the likelihood ratio function:

$$h_{\mathbb{Q}}^2(g_0, g) := \frac{1}{2} \int (\sqrt{g_0} - \sqrt{g})^2 \, d\mathbb{Q}. \tag{13}$$

3.1 Consistency of estimator E1

Our analysis of consistency relies on tools from empirical process theory. Let us briefly review the notion of the *metric entropy* of function classes (see, e.g., [28] for further background). For any $r \geq 1$ and distribution function $\mathbb{Q}$, define the empirical L_r metric

$$\|g\|_{L_r(\mathbb{Q})}^r \quad := \quad \int |g|^r d\mathbb{Q},$$

and let $L_r(\mathbb{Q})$ denote the metric space defined by this distance. For any fixed $\delta > 0$, a covering for function class $\mathcal{G}$ using the metric $L_r(\mathbb{Q})$ is a collection of functions which allow $\mathcal{G}$ to be covered using $L_r(\mathbb{Q})$ balls of radius δ centered at these functions. Letting $N_\delta(\mathcal{G}, L_r(\mathbb{Q}))$ 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 the $L_r(\mathbb{Q})$ 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 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:

$$G_0(x) = \sup_{g \in \mathcal{G}} |g(x)|, \qquad \text{and} \qquad G_1(x) = \sup_{g \in \mathcal{G}} |\log \frac{g(x)}{g_0(x)}|. \tag{14}$$

For the estimator E1, we impose the following assumptions on the distributions $\mathbb{P}, \mathbb{Q}$ and the function class $\mathcal{G}$.

Assumptions.

(A). The KL divergence is bounded: $D_K(\mathbb{P}, \mathbb{Q}) < \infty$.

(B). There is some $g \in \mathcal{G}$ such that $g = g_0$ almost surely (a.s.).

In the following theorem, the almost sure statement can be taken with respect to either $\mathbb{P}$ or $\mathbb{Q}$ since they share the same support.

Theorem 1. *Suppose that assumptions (A) and (B) hold.*
(a) Assume the envelope conditions

$$\int G_0 d\mathbb{Q} \quad < \quad \infty \tag{15a}$$

$$\int G_1 d\mathbb{P} \quad < \quad \infty \tag{15b}$$

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

$$\frac{1}{n}\mathcal{H}_\delta(\mathcal{G} - g_0, L_1(\mathbb{Q}_n)) \xrightarrow{\mathbb{Q}} 0, \tag{16a}$$

$$\frac{1}{n}\mathcal{H}_\delta(\log \mathcal{G}/g_0, L_1(\mathbb{P}_n)) \xrightarrow{\mathbb{P}} 0. \tag{16b}$$

Then, $|\widehat{D}_K - D_K(\mathbb{P}, \mathbb{Q})| \xrightarrow{a.s.} 0$, and $h_\mathbb{Q}(g_0, \widehat{g}_n) \xrightarrow{a.s.} 0$.

(b) Suppose only that (15a) and (16a) hold, and

$$\frac{1}{n}\mathcal{H}_\delta(\log \frac{\mathcal{G} + g_0}{2g_0}, L_1(\mathbb{P}_n)) \xrightarrow{\mathbb{P}} 0. \tag{17}$$

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

To provide intuition for the conditions of Theorem 1, note that the envelope condition (15a) is relatively mild, satisfied (for instance) if $\mathcal{G}$ is uniformly bounded from above. On the other hand, the envelope condition (15b) is much more stringent. Due to the logarithm, this can be satisfied if all functions in $\mathcal{G}$ are bounded from *both* above and below. However, as we see in part (b), we do not require boundedness from below; to ensure Hellinger consistency we can drop both the envelope condition (15b) and the entropy condition (16b), replacing them with the milder entropy condition (17).

It is worth noting that both (16a) and (17) can be deduced from the following single condition: for all $\delta > 0$, the bracketing entropy satisfies

$$\mathcal{H}_\delta^B(\mathcal{G}, L_1(\mathbb{Q})) < \infty. \tag{18}$$

Indeed, given equation (15a) and by the law of large numbers, condition (18) directly implies (16a). To establish condition (17), note that by a Taylor expansion, we have

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

7

so that $\frac{1}{n}\mathcal{H}_\delta(\log\frac{\mathcal{G}+g_0}{2g_0}, L_1(\mathbb{P}_n)) \le \frac{1}{n}\mathcal{H}_\delta(\mathcal{G}/g_0, L_1(\mathbb{P}_n))$. Since $G_0 \in L_1(\mathbb{Q})$, we have $G_0/g_0 \in L_1(\mathbb{P})$. In addition, $\mathcal{H}_\delta^B(\mathcal{G}/g_0, L_1(\mathbb{P})) \le \mathcal{H}_\delta^B(\mathcal{G}, L_1(\mathbb{Q})) < \infty$. By the law of large numbers, the metric entropy $\mathcal{H}_\delta(\mathcal{G}/g_0, L_1(\mathbb{P}_n))$ is bounded in probability, so that condition (17) holds.

In practice, the entropy conditions are satisfied by a variety of function classes. Examples include various types of reproducing kernel Hilbert spaces [21], as described in more detail in Section 4, as well as the Sobolev classes, which we describe in the following example.

Example 1 (Sobolev classes). Let $\kappa = (\kappa_1, \ldots, \kappa_d)$ be a d-dimensional multi-index, where all κ_i are natural numbers. Given a vector $x \in \mathbb{R}^d$, define $x^\kappa := \prod_{i=1}^d x_i^{\kappa_i}$ and $|\kappa| := \sum_{i=1}^d \kappa_i$. For a suitably differentiable function f, let D^κ denote the multivariate differential operator

$$D^\kappa f(x) := \frac{\partial^{|\kappa|}}{\partial x_1^{\kappa_1} \ldots \partial x_d^{\kappa_d}} f(x_1, \ldots, x_d), \tag{19}$$

and define the norm $\|f\|_{L_2^\alpha(\mathcal{X})}^2 := \sum_{|\kappa|=\alpha} \int |D^\kappa f(x)|^2 \, dx$.

With this notation, we define the norm

$$\|f\|_{W_2^\alpha(\mathcal{X})} := \|f\|_{L_2(\mathcal{X})} + \|f\|_{L_2^\alpha(\mathcal{X})}, \tag{20}$$

and the *Sobolev space* $W_2^\alpha(\mathcal{X})$ of functions with finite $\|f\|_{W_2^\alpha(\mathcal{X})}$-norm. Suppose that the domain $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$. Let the complexity measure I be the Sobolev norm—that is, $I(g) := \|g\|_{W_r^\alpha(\mathcal{X})}$. With this choice of complexity measure, it is known [4] that the function class $\mathcal{G}$ defined in equation (10) satisfies, for any $\delta > 0$, the metric entropy bound

$$\mathcal{H}_\delta(\mathcal{G}, L_\infty) = \mathcal{O}\left(\delta^{-d/\alpha}\right) < \infty, \tag{21}$$

for all smoothness indices $\alpha > d/2$. As a result, both conditions (18) and (16b) hold if, for instance, $\mathcal{G}$ is restricted to a subset of smooth functions that are bounded from above, and g_0 is bounded from below. $\diamondsuit$

3.2 Proof of Theorem 1

We now turn to the proof of Theorem 1, beginning with part (a). Define the following quantities associated with the function class $\mathcal{G}$:

$$\mathcal{E}_0(\mathcal{G}) = D_K(\mathbb{P}, \mathbb{Q}) - \sup_{g \in \mathcal{G}} \int (\log g \, d\mathbb{P} - g \, d\mathbb{Q} + 1) \ge 0 \tag{22}$$

$$\mathcal{E}_1(\mathcal{G}) = \sup_{g \in \mathcal{G}} \left| \int \log g \, d(\mathbb{P}_n - \mathbb{P}) - g d(\mathbb{Q}_n - \mathbb{Q}) \right|. \tag{23}$$

The quantity $\mathcal{E}_0$ is the *approximation error*, which measures the bias incurred by limiting the optimization to the class $\mathcal{G}$. The term $\mathcal{E}_1$ is the *estimation error* associated with the class $\mathcal{G}$. Our focus in this paper is the statistical problem associated with the estimation error $\mathcal{E}_1$, and thus we have imposed assumption (B), which implies that the approximation error $\mathcal{E}_0(\mathcal{G}) = 0$. Moreover, from equations (8) and (9b), straightforward algebra yields that

$$|\widehat{D}_K - D_K(\mathbb{P}, \mathbb{Q})| \le \mathcal{E}_0(\mathcal{G}) + \mathcal{E}_1(\mathcal{G}) = \mathcal{E}_1(\mathcal{G}). \tag{24}$$

Accordingly, the remainder of the proof is devoted to analysis of the estimation error $\mathcal{E}_1(\mathcal{G})$.

In order to analyze $\mathcal{E}_1(\mathcal{G})$, define the following processes:

$$v_n(\mathcal{G}) \quad := \quad \sup_{g \in \mathcal{G}} \left| \int \log \frac{g}{g_0} d(\mathbb{P}_n - \mathbb{P}) - \int (g - g_0) d(\mathbb{Q}_n - \mathbb{Q}) \right|, \qquad \text{and} \tag{25a}$$

$$w_n(g_0) \quad := \quad \left| \left| \int \log g_0 \, d(\mathbb{P}_n - \mathbb{P}) \, - g_0 d(\mathbb{Q}_n - \mathbb{Q}) \right| \right|. \tag{25b}$$

Note that we have

$$\mathcal{E}_1(\mathcal{G}) \leq v_n(\mathcal{G}) + w_n(g_0). \tag{26}$$

Note that the quantity $w_n(g_0)$ is the difference between an empirical and a population expectation. Let us verify that the conditions for the strong law of large numbers (SLN) apply. Using the inequality

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

due to Csiszár (cf. [9]), it follows that $\log g_0$ is $\mathbb{P}$ integrable. In addition, the function g_0 is $\mathbb{Q}$ integrable, since $\int g_0 d\mathbb{Q} = \int (p_0/q_0) d\mathbb{Q} = 1$. Thus, the SLN applies, and we conclude that $w_n(g_0) \xrightarrow{a.s.} 0$. By applying Theorem 5 from Appendix G, we conclude that $v_n(\mathcal{G}) \xrightarrow{a.s.} 0$. From the decomposition in equation (26), we conclude that $\mathcal{E}_1(\mathcal{G}) \xrightarrow{a.s.} 0$, so that $|\widehat{D}_K - D_K(\mathbb{P}, \mathbb{Q})| \xrightarrow{a.s.} 0$.

To establish Hellinger consistency of the likelihood ratio, we require the following lemma, whose proof is in Appendix A:

Lemma 2. *Defining the "distance" $d(g_0, g) := \int (g - g_0) d\mathbb{Q} - \log \frac{g}{g_0} d\mathbb{P}$, the following statements hold:*

(i) *For any $g \in \mathcal{G}$, we have $d(g_0, g) \geq 2h_{\mathbb{Q}}^2(g, g_0)$.*

(ii) *For the estimate $\widehat{g}_n$ defined in equation (9a), we have $d(g_0, \widehat{g}_n) \leq v_n(\mathcal{G})$.*

The Hellinger consistency $h_{\mathbb{Q}}(g_0, \widehat{g}_n) \xrightarrow{a.s.} 0$ of Theorem 1(a) is an immediate consequence of this lemma.

Turning now to Theorem 1 (b), we require a more refined lemma relating the Hellinger distance to suprema of empirical processes.

Lemma 3. *If $\widehat{g}_n$ is an estimate of g, then:*

$$\frac{1}{8} h_{\mathbb{Q}}^2(g_0, \widehat{g}_n) \leq 2h_{\mathbb{Q}}^2\left(g_0, \frac{g_0 + \widehat{g}_n}{2}\right) \leq - \int \frac{\widehat{g}_n - g_0}{2} d(\mathbb{Q}_n - \mathbb{Q}) + \int \log \frac{\widehat{g}_n + g_0}{2g_0} d(\mathbb{P}_n - \mathbb{P}).$$

See Appendix B for the proof of this claim. To complete the proof of Theorem 1, define $G_2(x) = \sup_{g \in \mathcal{G}} |\log \frac{g(x) + g_0(x)}{2g_0(x)}|$. Due to Lemma 3 and standard results from empirical process theory (see Appendix G, Theorem 5) it is sufficient to prove that $\int G_2 d\mathbb{P} < \infty$. To establish this claim, note that

$$\int G_2 d\mathbb{P} \quad \leq \quad \int \sup_{g \in \mathcal{G}} \max \left\{ \frac{g(x) + g_0(x)}{2g_0(x)} - 1, \; \log 2 \right\} d\mathbb{P}$$

$$\leq \quad \log 2 + \int \sup_{g \in \mathcal{G}} |g(x) - g_0(x)| d\mathbb{Q} \overset{(a)}{<} \infty,$$

where the last inequality (a) is due to envelope condition (15a).

9

3.3 Convergence rates

In this section, we describe convergence rates for both estimator E1 and estimator E2. The key technical tool that we use to analyze the convergence rate for the likelihood ratio estimate is Lemma 3, used previously in the proof of Theorem 1. This lemma bounds the Hellinger distance $h_{\mathbb{Q}}(g_0, \widehat{g}_n)$ in terms of the suprema of two empirical processes wrich respect to $\mathbb{P}$ and $\mathbb{Q}$. In a nutshell, the suprema of these two empirical processes can be bounded from above in terms of the Hellinger distance, which allows us to obtain the rates at which the Hellinger distance goes to zero.

3.3.1 Convergence rates for estimator E1

In order to characterize convergence rates for the estimator E1, we require one of the following two conditions:

$$\sup_{g \in \mathcal{G}} \|g\|_{\infty} \; < \; K_2 \tag{27a}$$

$$0 \; < \; K_1 \; \leq \; \inf_x g(x), \quad \text{and} \quad \sup_x g(x) \leq K_2 \quad \text{for all} \quad g \in \mathcal{G}. \tag{27b}$$

We also require the following assumption on function class $\bar{\mathcal{G}} := \{((g+g_0)/2)^{1/2}, g \in \mathcal{G}\}$: for some constant $0 < \gamma_{\bar{\mathcal{G}}} < 2$, there holds for any $\delta > 0$,

$$\mathcal{H}_{\delta}^B(\bar{\mathcal{G}}, L_2(\mathbb{Q})) = O(\delta^{-\gamma_{\bar{g}}}). \tag{28}$$

Combining this metric entropy decay rate with condition (27a), we deduce that for $\mathcal{G}$, the bracketing entropy satisfies

$$\mathcal{H}_{\delta}^B(\mathcal{G}, L_2(\mathbb{Q})) \; = \; \mathcal{O}(\delta^{-\gamma_{\bar{g}}}). \tag{29}$$

With these definitions, we can now state a result characterizing the convergence rate of estimator E1, where the notation $O_{\mathbb{P}}$ means "bounded in probability" with respect to $\mathbb{P}$ measure.

Theorem 2 (Convergence rates for estimator E1). *(a) If conditions* (27a) *and* (28) *hold, then* $h_{\mathbb{Q}}(g_0, \widehat{g}_n) = O_{\mathbb{P}}(n^{-1/(\gamma_{\bar{g}}+2)})$.

(b) If conditions (27b) *and* (28) *hold, then* $|\widehat{D}_K - D_K(\mathbb{P}, \mathbb{Q})| = O_{\mathbb{P}}(n^{-1/(\gamma_{\bar{g}}+2)})$.

Remarks: In order to gain intuition for the convergence rate in part (a), it can be instructive to compare to the *minimax rate*

$$r_n \; := \; \inf_{\widehat{g}_n \in \mathcal{G}} \sup_{\mathbb{P}, \mathbb{Q}} \mathbb{E}_{\mathbb{P}}\left[h_{\mathbb{Q}}(g_0, \widehat{g}_n)\right],$$

where the supremum is taken over all pairs $(\mathbb{P}, \mathbb{Q})$ such that $g_0 \in \mathcal{G}$. As a concrete example, if we take $\mathcal{G}$ as the Sobolev family from Example 1, and if condition (27b) holds, then the minimax rate is $r_n = \Omega(n^{-1/(\gamma+2)})$, where $\gamma = \gamma_{\bar{g}} = d/\alpha$ (see Appendix E). Thus, we see that for the Sobolev classes, the estimator E1 achieves the minimax rate for estimating the likelihood ratio in Hellinger distance.

3.3.2 Convergence rates for estimator E2

We now turn to a discussion of the convergence rates of estimator E2. To analyze this estimator, we assume that

$$I(g_0) < \infty, \tag{30}$$

and moreover we assume that the true likelihood ratio g_0—but *not* necessarily all of $\mathcal{G}$—is bounded from above and below:

$$0 < \eta_0 \le g_0 \le \eta_1 \text{ for some constants } \eta_0, \eta_1. \tag{31}$$

We also assume that the sup-norm over $\mathcal{G}_M$ is Lipschitz with respect to the penalty measure $I(g)$, meaning that there is a constant $c < \infty$ such that for each $M \ge 1$, we have

$$\sup_{g \in \mathcal{G}_M} \|g\|_\infty \le cM. \tag{32}$$

Finally, we assume that the bracketing entropy of $\mathcal{G}$ satisfies, for some $0 < \gamma < 2$, the condition

$$\mathcal{H}_\delta^B(\mathcal{G}_M, L_2(\mathbb{Q})) = \mathcal{O}\left[(M/\delta)^\gamma\right] \quad \text{for any} \quad \delta > 0. \tag{33}$$

Given these assumptions, we then state the following convergence rate result for the estimator E2:

Theorem 3. *(a) Suppose that assumptions (30) through (33) hold, and that the regularization parameter $\lambda_n \to 0$ is chosen such that*

$$\lambda_n^{-1} = \mathcal{O}_\mathbb{P}\left(n^{2/(2+\gamma)}\right).$$

Then under $\mathbb{P}$, we have

$$h_\mathbb{Q}(g_0, \widehat{g}_n) = \mathcal{O}_\mathbb{P}(\lambda_n^{1/2}). \tag{34}$$

(b) Suppose that in addition to assumptions (30) through (33), there holds $\inf_{x \in \mathcal{X}} \inf_{g \in \mathcal{G}} g(x) \ge \eta_0$. Then we have

$$|\widehat{D}_K - D_K(\mathbb{P}, \mathbb{Q})| = \mathcal{O}_\mathbb{P}(\lambda_n^{1/2}). \tag{35}$$

Remarks: Note that with the choice $\lambda_n^{-1} = \mathcal{O}_\mathbb{P}\left(n^{2/(2+\gamma)}\right)$ and the special case of $\mathcal{G}$ as the Sobolev space $W_2^\alpha(\mathcal{X})$ (see Example 1), estimator E2 again achieves the minimax rate for estimating the density ratio in Hellinger distance.

3.4 Proof of convergence theorems

In this section we present a proof of Theorem 3. The proof of Theorem 2 is similar in spirit, and is provided in Appendix D. The key to our analysis of the convergence rate of estimator E2 is the following lemma, which can be viewed as the counterpart of Lemma 3:

Lemma 4. *If $\widehat{g}_n$ is an estimate of g using (12a), then:*

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

See Appendix C for the proof of this lemma. Equipped with this auxiliary result, we can now prove Theorem 3(a). Define $f_g := \log \frac{g+g_0}{2g_0}$, and let $\mathcal{F}_M := \{f_g \mid g \in \mathcal{G}_M\}$. Since f_g is a Lipschitz function of g, conditions (31) and (33) imply that

$$\mathcal{H}_\delta^B(\mathcal{F}_M, L_2(\mathbb{P})) = \mathcal{O}\left\{(M/\delta)^\gamma\right\}. \tag{36}$$

Applying Lemma 5.14 from van de Geer [26] using the distance $d_2(g_0, g) = \|g - g_0\|_{L_2(\mathbb{Q})}$, we have that the following statement holds under $\mathbb{Q}$, and hence holds under $\mathbb{P}$ as well, since $d\mathbb{P}/d\mathbb{Q}$ is bounded from above:

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

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

$$\sup_{g \in \mathcal{G}} \frac{|\int f_g d(\mathbb{P}_n - \mathbb{P})|}{n^{-1/2}d_2(g_0, g)^{1-\gamma/2}(1 + I(g) + I(g_0))^{\gamma/2} \vee n^{-\frac{2}{2+\gamma}}(1 + I(g) + I(g_0))} = \mathcal{O}_\mathbb{P}(1). \tag{38}$$

Now using condition (32), it can be verified that

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

Combining Lemma 4 and equations (38) and (37), we conclude that

$$\frac{1}{4}h_\mathbb{Q}^2(g_0, \widehat{g}_n) + \frac{\lambda_n}{2}I^2(\widehat{g}_n) \le \lambda_n I(g_0)^2/2$$
$$+ O_\mathbb{P}\left(n^{-1/2}h_\mathbb{Q}(g_0, g)^{1-\gamma/2}(1 + I(g) + I(g_0))^{1/2+\gamma/4} \vee n^{-\frac{2}{2+\gamma}}(1 + I(g) + I(g_0))\right). \tag{39}$$

From this point, the proof involves simple algebraic manipulation of (39). To simplify notation, let $\widehat{h} = h_\mathbb{Q}(g_0, \widehat{g}_n)$, $\widehat{I} = I(\widehat{g}_n)$, and $I_0 = I(g_0)$. We break the analysis into four cases, depending on the behavior of $\widehat{h}$ and $\widehat{I}$:

Case A:. In this case, we assume $\widehat{h} \ge n^{-1/(2+\gamma)}(1 + \widehat{I} + I_0)^{1/2}$ and $\widehat{I} \ge 1 + I_0$. From (39), we have either

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

These conditions imply, respectively, either

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

or

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

In either case, we conclude the proof by setting $\lambda_n^{-1} = \mathcal{O}_\mathbb{P}(n^{2/(\gamma+2)}(1 + I_0))$.

Case B. In this case, we assume that $\widehat{h} \ge n^{-1/(2+\gamma)}(1 + \widehat{I} + I_0)^{1/2}$ and $\widehat{I} < 1 + I_0$. From equation (39), we have either

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

These conditions imply, respectively, that

$$\widehat{h} \le (1 + I_0)^{1/2} O_{\mathbb{P}}(n^{-1/(\gamma+2)}), \qquad \text{and} \qquad \widehat{I} \le 1 + I_0$$

or

$$\widehat{h} \le O_{\mathbb{P}}(\lambda_n^{1/2} I_0), \qquad \text{and} \qquad \widehat{I} \le O_{\mathbb{P}}(I_0).$$

In either case, the proof is concluded by setting $\lambda_n^{-1} = \mathcal{O}_{\mathbb{P}}(n^{2/(\gamma+2)}(1 + I_0))$.

Case C. In this case, we assume that $\widehat{h} \le n^{-1/(2+\gamma)}(1 + \widehat{I} + I_0)^{1/2}$ and $\widehat{I} \ge 1 + I_0$. From equation (39), we have

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

which implies that $\widehat{h} \le O_{\mathbb{P}}(n^{-1/(2+\gamma)})\widehat{I}^{1/2}$ and $\widehat{I} \le \lambda_n^{-1} O_{\mathbb{P}}(n^{-2/(2+\gamma)})$. Consequently, by setting $\lambda_n^{-1} = \mathcal{O}_{\mathbb{P}}(n^{2/(2+\gamma)})(1 + I_0)$, we obtain

$$\widehat{h} \le O_{\mathbb{P}}(\lambda_n^{1/2})(1 + I_0), \qquad \text{and} \qquad \widehat{I} \le \mathcal{O}_{\mathbb{P}}(1 + I_0).$$

Case D. In this final case, we assume that $\widehat{h} \le n^{-1/(2+\gamma)}(1 + \widehat{I} + I_0)^{1/2}$ and $\widehat{I} \le 1 + I_0$, and the claim of Theorem 3(a) follows.

We now proceed to the proof of Theorem 3(b). Note that part (a) and equation (32) imply that $\|\widehat{g}_n\|_\infty = \mathcal{O}_{\mathbb{P}}(1 + I(g_0))$. Without loss of generality, assume that $0 < \eta_0 \le g_0(x)$ and $\widehat{g}_n(x) \le \eta_1$ for all $x \in \mathcal{X}$. Then we have

$$
\begin{aligned}
|\widehat{D}_K - D_K(\mathbb{P}, \mathbb{Q})| &= \left| \int \log \widehat{g}_n d\mathbb{P}_n - \int \widehat{g}_n d\mathbb{Q}_n - \left(\int \log g_0 d\mathbb{P} - \int g_0 d\mathbb{Q} \right) \right| \\
&\le \left| \int \log \widehat{g}_n/g_0 d(\mathbb{P}_n - \mathbb{P}) - \int (\widehat{g}_n - g_0) d(\mathbb{Q}_n - \mathbb{Q}) \right| \\
&\quad + \left| \int \log \widehat{g}_n/g_0 d\mathbb{P} - \int (\widehat{g}_n - g_0) d\mathbb{Q} \right| \\
&\quad + \left| \int \log g_0 d(\mathbb{P}_n - \mathbb{P}) - \int g_0 d(\mathbb{Q}_n - \mathbb{Q}) \right| := T_1 + T_2 + T_3.
\end{aligned}
$$

We have $T_3 = \mathcal{O}_{\mathbb{P}}(n^{-1/2})$ by the central limit theorem. In addition,

$$
\begin{aligned}
T_2 &\le \int |\widehat{g}_n - g_0| \frac{\eta_1}{\eta_0} d\mathbb{Q}| + \int |\widehat{g}_n - g_0| d\mathbb{Q} \\
&\le (\eta_1/\eta_0 + 1)\|\widehat{g}_n - g_0\|_{L_2(\mathbb{Q})} \\
&\le (\eta_1/\eta_0 + 1)\left(\int 4\eta_1 (\widehat{g}_n^{1/2} - g_0^{1/2})^2 d\mathbb{Q} \right)^{1/2} \\
&\le (\eta_1/\eta_0 + 1)\eta_1^{1/2} 4h_{\mathbb{Q}}(g_0, \widehat{g}_n) \\
&= \mathcal{O}_{\mathbb{P}}(n^{-1/(2+\gamma)}),
\end{aligned}
$$

where the last equality is due to Theorem 3(a).

Finally, to bound T_1, we apply a modulus of continuity result on the suprema of empirical processes with respect to function $(g - g_0)$ and $(\log g - \log g_0)$, where g is restricted to smooth

functions bounded from below (by η_0) and above (by η_1). As a result, the bracketing entropy for both function classes $\mathcal{G}$ and $\log \mathcal{G}$ has the same order $O(1/\delta)^\gamma$ as given in (33). Apply Lemma 5.13 of [26] (page 79), to obtain that for $\delta_n = n^{-1/(2+\gamma)}$, there holds:

$$T_1 = \mathcal{O}_{\mathbb{P}}(n^{-1/2}\|\widehat{g}_n - g_0\|_{L_2(\mathbb{Q})}^{1-\gamma/2} \vee \delta_n^2) \; = \; \mathcal{O}_{\mathbb{P}}(n^{-2/(2+\gamma)}),$$

which completes the proof.

4 Algorithmic implementation and simulation examples

In this section, we turn to the more practical issue of implementation, focusing in particular on estimator E2. When $\mathcal{G}$ has a kernel-based structure, we show how, via conversion to its dual form, the computation of the estimator E2 reduces to the solution of an n-dimensional convex program. We illustrate the performance of this practical estimator with a variety of simulations.

4.1 Algorithms for kernel-based function classes

We develop two versions of estimator E2: in the first, we assume that $\mathcal{G}$ is a reproducing kernel Hilbert space (RKHS), whereas in the second, we assume that $\log \mathcal{G}$ forms an RKHS. In both cases, we focus on the Hilbert space induced by a Gaussian kernel. This choice is appropriate as it is sufficiently rich, but also amenable to efficient optimization procedures [22].

We begin with some background on reproducing kernel Hilbert spaces; see the books [21, 22] for further details. Consider a positive definite function K mapping the Cartesian product $\mathcal{X} \times \mathcal{X}$ to the non-negative reals. By Mercer's theorem, any such kernel function K can be associated with a feature map $\Phi : \mathcal{X} \to \mathcal{H}$, where $\mathcal{H}$ is a Hilbert space with inner product $\langle \cdot, \cdot \rangle$. Moreover, for all $x, x' \in \mathcal{X}$, the inner product in this Hilbert space is linked to the kernel via the relation $K(x, x') = \langle \Phi(x), \Phi(x') \rangle$. 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 $\|g\|_{\mathcal{H}} = \|w\|_{\mathcal{H}}$. The kernel used in our simulations is the Gaussian kernel:

$$K(x, y) \quad := \quad \exp\left\{-\|x - y\|^2/\sigma\right\}, \tag{40}$$

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

4.1.1 Imposing RKHS structure of $\mathcal{G}$

Suppose that the function class $\mathcal{G}$ in estimator E2 is the Gaussian RKHS space $\mathcal{H}$, and let the complexity measure be the Hilbert space norm $I(g) = \|g\|_{\mathcal{H}}$. With these choices, equation (12a) becomes:

$$f^* \;=\; \min_w J := \min_w \frac{1}{n} \sum_{i=1}^n \langle w, \Phi(x_i) \rangle - \frac{1}{n} \sum_{j=1}^n \log\langle w, \Phi(y_j) \rangle + \frac{\lambda_n}{2}\|w\|_{\mathcal{H}}^2, \tag{41}$$

where the samples $\{x_i\}$ and $\{y_j\}$ are i.i.d. from $\mathbb{Q}$ and $\mathbb{P}$, respectively. The log function is extended to take the value $-\infty$ for negative arguments.

14

Lemma 5. *The primal value f^* has the equivalent dual form:*

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

Moreover, the optimal dual solution $\widehat{\alpha}$ is linked to the optimal primal solution $\widehat{w}$ via the relation

$$\widehat{w} = \frac{1}{\lambda_n}(\sum_{j=1}^{n}\widehat{\alpha}_j\Phi(y_j)-\frac{1}{n}\sum_{i=1}^{n}\Phi(x_i)). \tag{42}$$

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

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

where the last line is due to the inf-convolution theorem [20]. Simple calculations yield:

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

Thus, we conclude that $f^* = -\min_{\alpha_i}\sum_{j=1}^{n}(-\frac{1}{n}-\frac{1}{n}\log n\alpha_j)+\frac{1}{2\lambda_n}\|\sum_{j=1}^{n}\alpha_j\Phi(y_j)-\frac{1}{n}\sum_{i=1}^{n}\Phi(x_i)\|_{\mathcal{H}}^2$, from which the claim follows. The primal-dual relation (42) also follows from these calculations. □

For an RKHS based on a Gaussian kernel, the entropy condition (33) holds for any $\gamma>0$ (cf. Zhou [33]). Furthermore, condition (32) holds since, via the Cauchy-Schwarz inequality, we have

$$|g(x)| = |\langle w,\,\Phi(x)\rangle| \le \|w\|_{\mathcal{H}}\|\Phi(x)\|_{\mathcal{H}} \le I(g)\sqrt{K(x,x)} \le I(g).$$

Thus, by Theorem 3(a), we have $\|\widehat{w}\|_{\mathcal{H}} = \|\widehat{g}_n\|_{\mathcal{H}} = \mathcal{O}_{\mathbb{P}}(\|g_0\|_{\mathcal{H}})$, so the penalty term $\lambda_n\|\widehat{w}\|^2$ vanishes at the same rate as λ_n. Thus, we obtain the following estimator for the KL divergence:

$$\widehat{D}_K = 1+\sum_{j=1}^{n}(-\frac{1}{n}-\frac{1}{n}\log n\widehat{\alpha}_j) = \sum_{j=1}^{n}-\frac{1}{n}\log n\widehat{\alpha}_j. \tag{43}$$

4.1.2 Imposing RKHS structure on $\log\mathcal{G}$

An alternative starting point is to posit that the function class $\log\mathcal{G}$ has an RKHS structure. In this case, we consider functions of the form $g(x)=\exp\langle w,\,\Phi(x)\rangle$, and use the complexity measure $I(g)=\|\log g\|_{\mathcal{H}}=\|w\|_{\mathcal{H}}$. Unfortunately, Theorem 3 does not apply directly because condition (32) no longer holds, but this choice nonetheless seems reasonable and worth investigating from an empirical viewpoint.

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

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

$$= \;-\min_{\alpha>0}\sum_{i=1}^n \alpha_i \log(n\alpha_i) - \alpha_i + \frac{1}{2\lambda_n}\|\sum_{i=1}^n \alpha_i \Phi(x_i) - \frac{1}{n}\sum_{j=1}^n \Phi(y_j)\|_{\mathcal{H}}^2.$$

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

$$\widehat{D}_K = 1 + \sum_{i=1}^n \widehat{\alpha}_i \log \widehat{\alpha}_i + \widehat{\alpha}_i \log \frac{n}{e}. \tag{44}$$

4.2 Simulation results

In this section, we describe the results of various simulations that demonstrate the practical viability of the estimators (43) and (44), 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 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-dimensional data, and choose larger σ for higher dimensions. We use M1 and M2 to denote the estimators (43) and (44), respectively. We compare these methods to algorithm A in Wang et al [29], which was shown empirically to be one of the best methods in the literature. This 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 four plots in Figure 4.2 present results with univariate distributions. We see that the estimator M2 generally exhibits the best convergence rate among the estimators considered. The WKV estimator performs somewhat less well, and shows sensitivity to the choice of partition size s, with the ranking of the different WKV estimators changing over the experiments. The performance of estimator M1 is comparable to that of the WKV estimator, although clearly better in the first plot. In Figure 4.2 we present the results with two- and three-dimensional data. Again, estimator M2 has the best convergence rates in all examples. The M1 estimator does not converge in the last example, suggesting that the underlying function class exhibits very strong bias. In these examples, the WKV estimator again shows sensitivity to the choice of partition size; moreover, its performance is noticeably degraded in the case of three-dimensional data (the lower two plots).

It is worth noting that as one increases the number of dimensions, histogram-based methods such as WKV become increasingly difficult to implement, whereas increased dimension has only a mild effect on the complexity of implementation of our method.
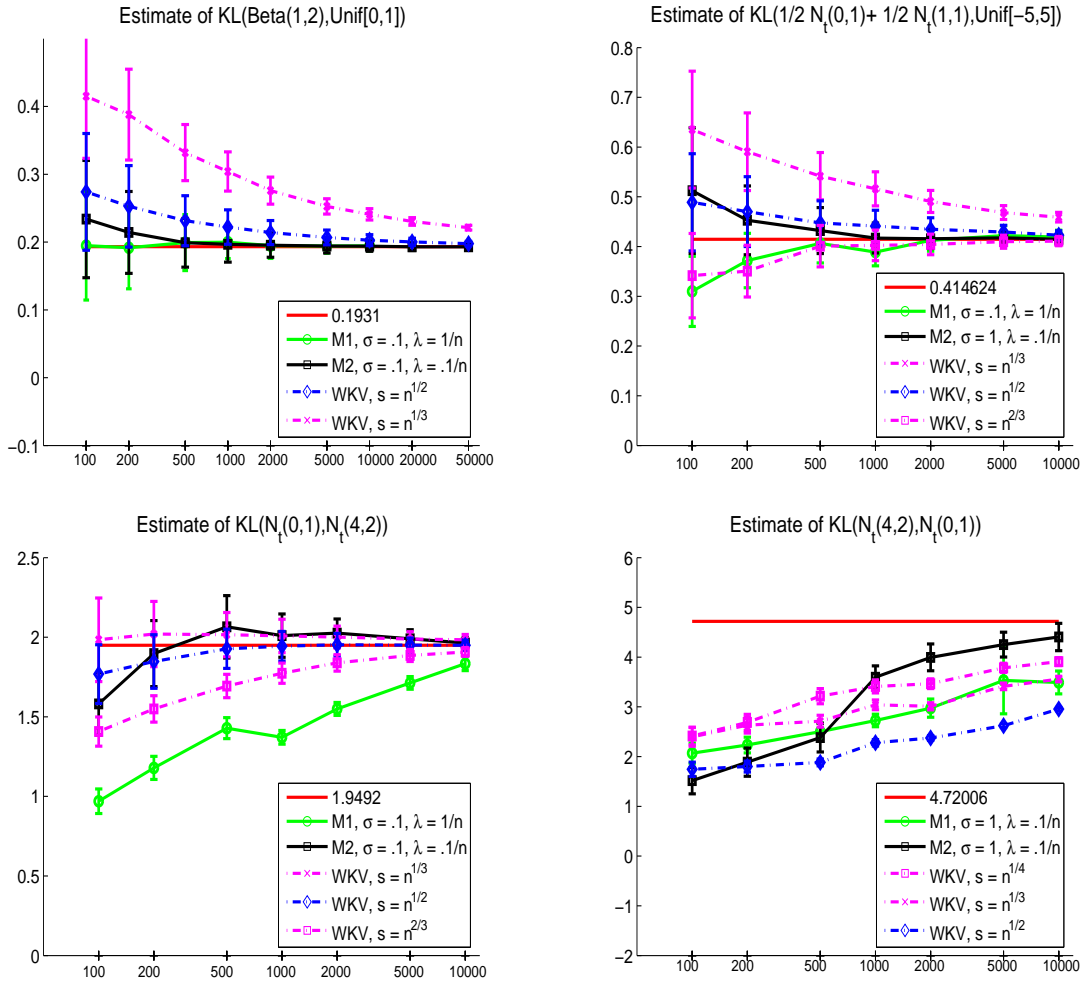
Figure 1. Results of estimating KL divergences for various choices of probability distributions. In all plots, the X-axis is the number of data points plotted on a log scale, and the Y-axis is the estimated value. The error bar is obtained by replicating the simulation 250 times. $N_t(a, I_k)$ denotes a truncated normal distribution of k dimensions with mean $(a, \ldots, a)$ and identity covariance matrix.
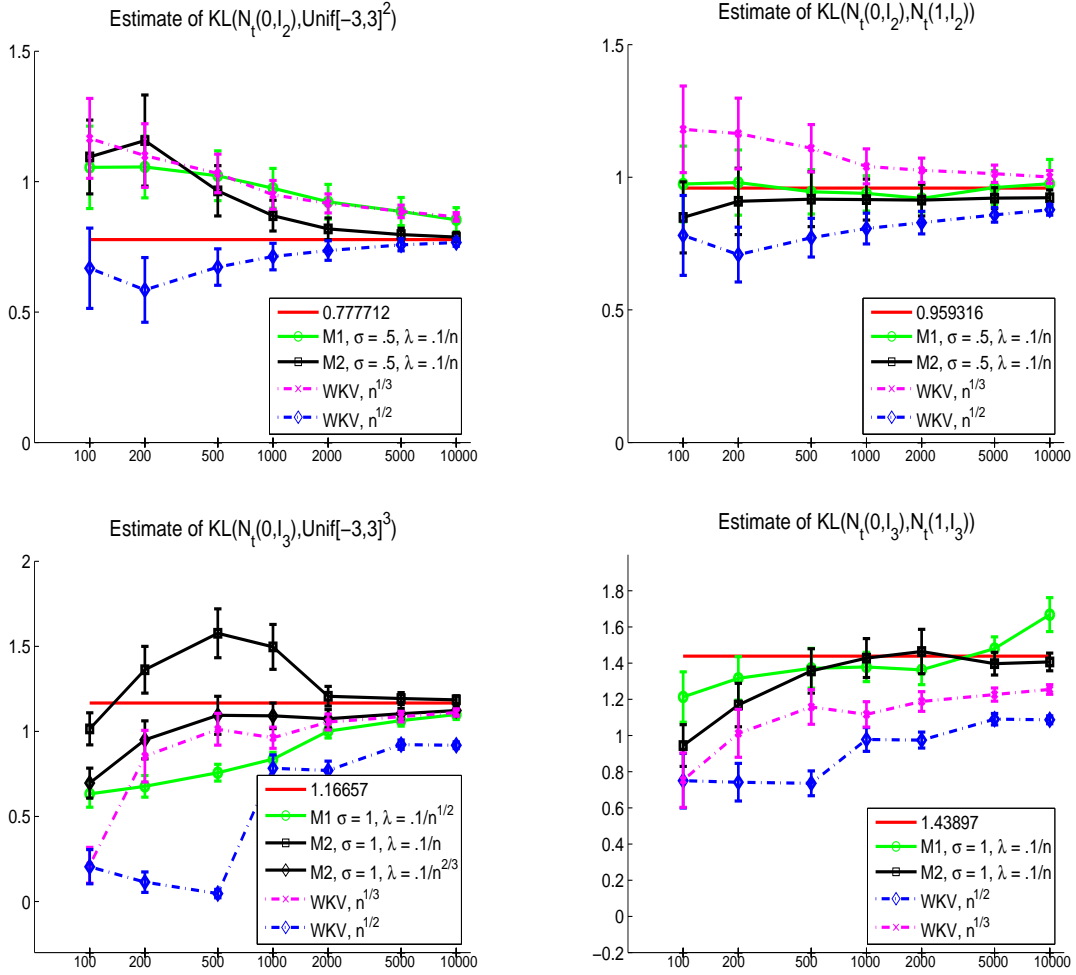
Figure 2. Results of estimating KL divergences for various choices of probability distributions. In all plots, the X-axis is the number of data points plotted on a log scale, and the Y-axis is the estimated value. The error bar is obtained by replicating the simulation 250 times. $N_t(a, I_k)$ denotes a truncated normal distribution of k dimensions with mean $(a, \ldots, a)$ and identity covariance matrix.

5   Some extensions

In this section, we discuss some extensions and related estimators, all based on the same basic variational principle.

5.1   Estimation of likelihood ratio functions

Suppose that we are primarily interested in estimating the likelihood ratio function p_0/q_0, as opposed to the Kullback-Leibler divergence. In this case, we may consider any divergence functional $D_\phi(p_0, q_0)$, where ϕ is a convex function on $\mathbb{R}_+$, possibly different than the logarithm leading to KL divergence. Again applying Lemma 1, choosing a different divergence leads to the following alternative estimator of the likelihood ratio:

$$\widehat{f}_n \quad := \quad \mathrm{argmax}_{f \in \mathcal{F}} \int f \, d\mathbb{Q}_n - \int \phi^*(f) \, d\mathbb{P}_n \tag{45}$$

$$\widehat{D}_\phi \quad := \quad \int \widehat{f}_n \, d\mathbb{Q}_n - \int \phi^*(\widehat{f}_n) \, d\mathbb{P}_n. \tag{46}$$

The quantity $\widehat{f}_n$ is an estimate of the quantity $f_0 = \partial\phi(q_0/p_0)$, whereas $\widehat{D}_\phi$ is an estimate of the divergence $D_\phi(p_0, q_0)$ (of secondary interest for the moment).

We make the following observations:

- If ϕ is a differentiable and strictly convex function, i.e., $\phi'' > 0$, then the likelihood ratio function p_0/q_0 can be recovered by applying the inverse of ϕ' to $\widehat{f}_n$. Thus, we obtain a family of estimation methods for the likelihood ratio function by simply ranging over choices of ϕ.

- If (on the other hand) the function ϕ is chosen to be non-differentiable, we cannot directly invert the mapping $\partial\phi$, but we can nonetheless obtain estimators for other interesting objects. For instance, suppose that ϕ has the piecewise-linear form

$$\phi(u) \quad := \quad \begin{cases} |u - 1| \text{ if } u \geq 0 \\ +\infty \text{ otherwise,} \end{cases} \quad ,$$

  so that D_ϕ is the variational distance. Noting that $\partial\phi(u) = \mathrm{sign}(u - 1)$ for any $u > 0$, we see that the quantity $\widehat{f}_n$ in equation (45) provides an estimate of the thresholded likelihood ratio.[2]

---

[2]In fact, there is strong relationship between variational distance and a threshold function of the likelihood ratio. Note that the conjugate dual for ϕ has the form:

$$\phi^*(v) = \begin{cases} -1 \text{ if } v \geq -1 \\ v \text{ if } v \in [-1, 1] \\ +\infty \text{ otherwise,} \end{cases}$$

which is related to a hinge loss in the literature of binary classification in machine learning. Indeed, a binary classification problem can be viewed as estimating the threshold function of the likelihood ratio. See [19] for a discussion of divergences and surrogate losses from this viewpoint.

5.2 Extensions to different ϕ

Let us assume that ϕ is chosen to be differentiable and strictly convex, so that we can estimate the likelihood ratio g_0 by applying $(\phi')^{-1}$ to $\widehat{f}_n$. Since there are many such ϕ, it is natural to ask how the properties of ϕ affect the quality of the estimate of g_0. The analysis provided in the preceding sections can be adapted to other choices of ϕ, as we describe here.

In order to describe these extensions, we first define a distance between f_0 and f:

$$d_\phi(f_0, f) \;\; := \;\; D_\phi(\mathbb{P}, \mathbb{Q}) - \int f d\mathbb{Q} - \phi^*(f) d\mathbb{P} \tag{47}$$

Note that this distance is simply the generalization of the quantity $d(g_0, g)$ previously defined in Lem 2). For future reference, we note the equivalence

$$
\begin{aligned}
d_\phi(f_0, f) &= \int (\phi^*(f) - \phi^*(f_0)) d\mathbb{P} - (f - f_0) d\mathbb{Q} \\
&= \int \left(\phi^*(f) - \phi^*(f_0) - \frac{\partial \phi^*}{\partial f}(f_0)(f - f_0) \right) d\mathbb{P},
\end{aligned}
$$

where the final line uses the facts that $\frac{\partial \phi^*}{\partial f}(f_0) = q_0/p_0$ and $\phi'(q_0/p_0) = f_0$. This expression shows that d_ϕ is the *Bregman divergence* defined by the convex function ϕ^*.

Recall that the key ingredient in our earlier analysis was the relation between the empirical processes defined by equation (25a) and the "distance" $d(g_0, g)$ (see Lemma 2). Similarly, the key technical ingredient in the extension to general ϕ involves relating the quantity

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

to the distance $d_\phi(f_0, f)$ defined in equation (47). In particular, we can state the following analog of Lemma 2 and Lemma 3:

Lemma 6. *Let $\widehat{f}_n$ be the estimate of f_0 obtained by solving the problem (45). Then*

$$d_\phi(f_0, \widehat{f}_n) \;\; \leq \;\; v_n^\phi(\mathcal{F}). \tag{48}$$

Under suitable technical conditions, we have $v_n^\phi(\mathcal{F}) \xrightarrow{a.s.} 0$, so that Lemma 6 implies that $\widehat{f}_n$ is a consistent estimator for f_0 in the sense of d_ϕ. This lemma also provides the technical means to derive convergence rates in the same manner as in the previous sections. Note that $d_\phi(f_0, f)$ is usually not a proper metric. To apply standard results from empirical process theory, the trick is that one can 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 (via Lemma 2(i)).

Let us illustrate this idea by stating a result about likelihood ratio estimation in terms of the χ-*square divergence*, which is defined by

$$D_\chi(\mathbb{P}, \mathbb{Q}) \;\; := \;\; \int p_0^2/q_0 d\mu. \tag{49}$$

Note that this divergence is an f-divergence with $\phi(u) = 1/u$; a short calculation (see Appendix F) shows that the associated "distance" d_χ is given by $d_\chi(g, g_0) = \int (g - g_0)^2 d\mathbb{Q}$, which is simply the $L_2(\mathbb{Q})$ metric. With this set-up, the following theorem is an analog of Theorem 2 (with an almost identical proof):

Theorem 4. *Assume that for some constant $0 < \gamma < 2$,*

$$\mathcal{H}_\delta^B(\mathcal{G}, L_2(\mathbb{Q})) \ \leq \ A_\mathcal{G} \delta^{-\gamma}, \tag{50}$$

and moreover that condition (27a) *holds. Then the estimator $\widehat{g}_n$ obtained from the χ-square diver-gence is consistent with rate $d_\chi(g_0, \widehat{g}_n) = O_\mathbb{P}(n^{-2/(\gamma+2)})$.*

Remark: Comparing with Theorem 2, we see that the conditions of Theorem 4 are weaker. Indeed, the $L_2(\mathbb{Q})$ metric is dominated by the Hellinger metric, so that imposing bounds on $L_2(\mathbb{Q})$-metric and its induced entropy are milder conditions.

5.3   Estimation of the divergence functional D_ϕ

Suppose that we are primarily interested in estimating the divergence functional D_ϕ, given that we have already obtained an optimal estimator $\widehat{g}_n$ of the likelihood ratio function $g_0 = p_0/q_0$ (such as the one defined by (9a) or (12a), or more generally (45)). We have demonstrated that D_ϕ can be estimated by (9b) and (12b), or more generally by (46). Note that D_ϕ can be viewed as an integral of the likelihood ratio under the distribution $\mathbb{Q}$. Indeed, we can write

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

Although D_ϕ is an integral functional of $g_0 = p_0/q_0$, an interesting feature here is that the integra-tion is with respect to *unknown* $\mathbb{Q}$. In this section, we show that estimators such as (9b) and (12b) for the KL divergence can be viewed as a first-order Taylor expansion of the integral functional around the estimate $\widehat{g}_n$ of the likelihood ratio. This discussion is motivated by a line of work on the estimation of integral functional of a single density function (cf. [13, 3]), and also leads to an open question.

Suppose that $\phi : \mathbb{R}_+ \to \mathbb{R}$ is a convex function differentiable up to third order, $\mathcal{G}$ is a smooth function class bounded from both above and below as in Example 1 (with smoothness parameter α). Suppose that $\widehat{g}_n$ is an estimator of g_0 such as the one defined by (45), so that $\|\widehat{g}_n - g_0\|_{L_2(\mathbb{Q})} = O_P(n^{-\alpha/(2\alpha+d)})$. Using a Taylor expansion around $\widehat{g}_n$, we obtain:

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

We arrive at

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

In the above expression, the first two integrals can be estimated from two n-samples 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 $O(\|g_0 - \widehat{g}_n\|_3^3) = O_P(n^{-3\alpha/(2\alpha+d)})$. 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 convergence rate for estimating the integral of the form $\int p_0^2/q_0 \psi \, d\mu$. This is interesting because we have reduced the problem of estimating any f-divergence to a particular integral of two densities $\int p_0^2/q_0 \psi \, d\mu$, where ψ is a known function.

Let us 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 $O_P(n^{-\alpha/(2\alpha+d)})$), the estimator has the following form:

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

which has exactly the same form as our original estimator (9b), except that here $\widehat{g}_n$ can be any (optimal) estimator of the likelihood ratio. Thus, a better convergence rate for estimating KL divergence (and D_ϕ in general) is certainly achievable by using Taylor expansion up to second order, *assuming* that a separate method exists to achieve a superior rate for the integral $\int p_0^2/q_0 \psi \, d\mu$. Thus, one interesting open question is whether $\int p_0^2/q_0 \psi \, d\mu$ can be estimated at a rate better than the $O_P(n^{-\alpha/(2\alpha+d)})$ rate that one can achieve with the general method defined by (46).

6   Conclusions

We have developed and analyzed M-estimation methods for both the likelihood ratio and f-divergence functionals of two unknown multivariate probability distributions by exploiting a variational characterization of f-divergence functionals. The methods are shown to be amenable to efficient computational algorithms for optimization in high-dimensional function spaces. We have also described our method in the general context of estimating integral functionals of the likelihood ratio of two unknown densities, and discussed directions for future work suggested by our results.

Acknowledgements

A   Proof of Lemma 2

(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 $\int \widehat{g}_n d\mathbb{Q}_n - \int \log \widehat{g}_n d\mathbb{P}_n \leq \int g_0 d\mathbb{Q}_n - \int \log g_0 d\mathbb{P}_n$. It follows that

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

B   Proof of Lemma 3

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

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

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:

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

Rearranging,

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

where the last inequality is an application of Lemma 2.

C   Proof of Lemma 4

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:

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

By the definition (12a) of our estimator, we have:

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

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:

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

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

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

$$\leq -2h_{\mathbb{Q}}^2(g_0, \frac{g_0 + \widehat{g}_n}{2}) + \frac{\lambda_n}{4} I^2(g_0) \leq -\frac{1}{8} h_{\mathbb{Q}}^2(g_0, \widehat{g}_n) + \frac{\lambda_n}{4} I^2(g_0),$$

where the last inequality is a standard result for the (generalized) Hellinger distance (cf. [26]).

D   Proof of Theorem 2

(a) One of the empirical processes on the right-hand side of Lemma 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 "norm," namely, *Bernstein distance*. This is defined as follows: 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., [26], page 97):

- $\rho_1(f_g) \leq 4h_{\mathbb{Q}}(g_0, \frac{g + g_0}{2})$.

- The bracketing entropy based on Bernstein distance is also related to the bracketing entropy based Hellinger distance (i.e., which is the L_2 norm for the square root function):

$$\mathcal{H}_{\sqrt{2}\delta}^B(\mathcal{F}, \rho_1) \leq \mathcal{H}_{\delta}^B(\bar{\mathcal{G}}, L_2(\mathbb{Q})), \tag{51}$$

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

By Lemma 3, for any $\delta > 0$, with respect to $\mathbb{P}$ measure:

$$P(h_{\mathbb{Q}}(g_0, \widehat{g}_n) > \delta) \leq P(h_{\mathbb{Q}}(g_0, (\widehat{g}_n + g_0)/2) > \delta/4)$$

$$\leq P\left(\sup_{g \in \mathcal{G}, \, h_{\mathbb{Q}}(g_0, \bar{g}) > \delta/4} - \int (\bar{g} - g_0) d(\mathbb{Q}_n - \mathbb{Q}) + \int f_g \, d(\mathbb{P}_n - \mathbb{P}) - 2h_{\mathbb{Q}}^2(g_0, \bar{g}) \geq 0 \right)$$

$$\leq P\left(\sup_{g \in \mathcal{G}, \, h_{\mathbb{Q}}(g_0, \bar{g}) > \delta/4} - \int (\bar{g} - g_0) d(\mathbb{Q}_n - \mathbb{Q}) - h_{\mathbb{Q}}^2(g_0, \bar{g}) \geq 0 \right)$$

$$+ P\left(\sup_{g \in \mathcal{G}, \, h_{\mathbb{Q}}(g_0, \bar{g}) > \delta/4} \int f_g \, d(\mathbb{P}_n - \mathbb{P}) - h_{\mathbb{Q}}^2(g_0, \bar{g}) \geq 0 \right) := A + B.$$

We need to upper bound the two quantities A and B on the right-hand side of this equation. These can be handled in a similar manner. Since $\mathcal{H}_\delta^B(\bar{\mathcal{G}}, L_2(\mathbb{Q})) < \infty$ 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 apply the union bound on the probability of excess. For each layer, one can now apply the modulus of continuity of suprema of an empirical process.

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

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 bracketing entropy integral can be bounded as:

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

where C_8, C_9 are constants independent of s. Now apply Theorem 6 (see Appendix G), 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 Theorem 6. Applying this theorem, we obtain:

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

for some universal constant c. A similar bound can be obtained for A, with respect to $\mathbb{Q}$ measure and with $\delta = n^{-1/(\gamma_{\bar{\mathcal{G}}}+2)}$. 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 $n^{-1/(\gamma_{\bar{\mathcal{G}}}+2)}$.

(b) The proof is similar to Theorem 3(b) and is omitted.

E Comparison of the rate in Lemma 2 to the minimax rate

Recall that the minimax rate is defined as

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

where the supremum is taken over all pairs $(\mathbb{P}, \mathbb{Q})$ such that $g_0 \in \mathcal{G}$. 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 $\mathcal{X}$. We can reduce this bound to the minimax

lower bound for a nonparametric density estimation problem [32]. This reduction is not entirely straightforward, however, because the space $\mathcal{G}$ ranges over smooth functions that need not be valid probability densities. Therefore, an easy-to-use minimax lower bound such as that of [31] is not immediately applicable. Nonetheless, we can still apply the hypercube argument and the Assouad lemma to obtain the right minimax rate. See §24.3 of van der Vaart [27] for a proof for the case of one dimension. This proof goes through for general $d \geq 1$.

F   Some calculations for Theorem 4

Note that the conjugate dual of $\phi(u) = 1/u$ takes the form

$$
\phi^*(v) \;=\; \begin{cases} -2\sqrt{-v} & \text{if } v < 0, \text{ and} \\ +\infty & \text{otherwise.} \end{cases}
$$

Consequently, we can 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 $g_0 := \sqrt{-f_0} = \sqrt{-\phi'(q_0/p_0)} = p_0/q_0$. Define $\widehat{g}_n := \sqrt{-\widehat{f}_n}$. We also replace notation $d_\phi(f_0, f)$ by $d_\phi(g_0, g)$. For our choice of ϕ, we have:

$$
\begin{aligned}
d_\chi(g_0, g) \;&=\; d_\chi(f_0, f) = \int (-2\sqrt{-f} + 2\sqrt{-f_0})d\mathbb{P} - (f - f_0)d\mathbb{Q} \\
&=\; \int (g_0 - g)(2p_0/q_0 - g_0 - g)d\mathbb{Q} \\
&=\; \int (g - g_0)^2 d\mathbb{Q},
\end{aligned}
$$

as claimed. Moreover, we have

$$
v_n^\chi(\mathcal{G}) \;=\; v_n^\chi(\mathcal{F}) = \sup_{g \in \mathcal{G}} \left| \int 2(g^2 - g_0^2)d(\mathbb{Q}_n - \mathbb{Q}) - \int (g - g_0)d(\mathbb{P}_n - \mathbb{P}) \right|.
$$

G   Results from empirical process theory

For completeness, we state here two standard results from empirical process theory that are needed in the paper. These results are versions of Theorems 3.7 and 5.11 from van de Geer [26], respectively:

Theorem 5. *Let $G(x) = \sup_{g \in \mathcal{G}} |g(x)|$ be the envelope function for a function $\mathcal{G}$. Assume that $\int G d\mathbb{P} < \infty$, and suppose moreover 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 6. *Suppose that the function class $\mathcal{G}$ satisfies $\sup_{g \in \mathcal{G}} \rho_K(g) \leq R$ for some constants K and R. Given $a > 0$, suppose that for some constants C and C_1, there holds*

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

Then the empirical process is bounded as

$$
\mathbb{P}\left[\sup_{g \in \mathcal{G}} \left| \sqrt{n} \int g d(\mathbb{P}_n - \mathbb{P}) \right| \geq a \right] \;\leq\; C \exp\left[-\frac{a^2}{C^2(C_1+1)R^2} \right]. \tag{52}
$$

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

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

[4] 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.

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

[6] P Common. Independent component analysis, a new concept? *Signal Process.*, 36:387–314, 1994.

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

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

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

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

[11] J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms*, volume 1. Springer-Verlag, New York, 1993.

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

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

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

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

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

[17] 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 (NIPS)*, 2007.

[18] 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)*, 2007.

[19] X. Nguyen, M. J. Wainwright, and M. I. Jordan. On surrogate losses and f-divergences. *Annals of Statistics*, 2008. To appear.

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

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

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

[23] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, UK, 2004.

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

[25] J. N. Tsitsiklis. Decentralized detection. In *Advances in Statistical Signal Processing*, pages 297–344. JAI Press, 1993.

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

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

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

[29] 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.

[30] Q. Wang, S. R. Kulkarni, and S. Verdú. A nearest-neighbor approach to estimating divergence between continuous random vectors. In *IEEE Symposium on Information Theory*, 2006.

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

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

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