Lipschitz Density-Ratios, Structured Data, and Data-driven Tuning

Samory Kpotufe
ORFE, Princeton University

Abstract

Density-ratio estimation (i.e. estimating $f = f_Q/f_P$ for two unknown distributions Q and P) has proved useful in many Machine Learning tasks, e.g., risk-calibration in transfer-learning, two-sample tests, and also useful in common techniques such importance sampling and bias correction. While there are many important analyses of this estimation problem, the present paper derives convergence rates in other practical settings that are less understood, namely, extensions of traditional Lipschitz smoothness conditions, and common high-dimensional settings with structured data (e.g. manifold data, sparse data).

Various interesting facts, which hold in earlier settings, are shown to extend to these settings. Namely, (1) optimal rates depend only on the smoothness of the ratio f, and not on the densities f_Q, f_P, supporting the belief that plugging in estimates for f_Q, f_P is suboptimal; (2) optimal rates depend only on the intrinsic dimension of data, i.e. this problem – unlike density estimation – escapes the curse of dimension.

We further show that near-optimal rates are attainable by estimators tuned from data alone, i.e. with no prior distributional information. This last fact is of special interest in unsupervised settings such as this one, where only *oracle* rates seem to be known, i.e., rates which assume critical distributional information usually unavailable in practice.

1 INTRODUCTION

Density ratios, i.e. the ratio $f = f_Q/f_P$ of two densities f_P and f_Q, are ubiquitous in Machine Learning applications. For instance, they naturally appear in two-sample testing problems [1, 2], outlier detection [3], and estimation of divergence functionals [4, 5, 6]. More recently, they appear as a corner-piece of much work in *transfer-learning* [7, 8] where the goal is to recalibrate a risk functional over some target Q by using data from P. The key idea is that the risk $\mathbb{E}_Q\, l$ (for some loss l) which is to be approximated from sample, is easily rewritten as $\mathbb{E}_P\, l \cdot f$, which is useful if we have more data from P. Similar uses of density-ratios appear more generally in importance sampling and bias correction where an integral $\mathbb{E}_Q\, g$ is to be approximated using samples from some $P \approx Q$ (see e.g. [9, 10]). Thus various estimators of f exist [11, 4, 12, 13, 14], usually derived and analyzed under the assumption that f belongs to a Reproducing Kernel Hilbert Space (RKHS).

This paper aims to yield further insights on the inherent difficulty of density-ratio estimation by considering other practical settings that are currently less understood. In particular, while previous analyses yield important insights for functions in an RKHS, here we consider common Lipschitz conditions and their Hölder extensions, encoding how fast a function varies over its domain. Furthermore, we consider practical situations where high-dimensional data in $\mathbb{R}^D$ actually lies on a structured subspace of lower dimension, e.g. data on a manifold or sparse data; the aim here is to elucidate the effect of data dimension on estimation rates, which manifestly also affects the downstream tasks. Finally, we are particularly interested in which rates are achievable given no prior distributional information, i.e., whether an estimator *tuned from data alone* can still achieve optimal rates (which typically assume optimal tuning). This last question is of special importance in unsupervised (or semisupervised) settings such as this one, where there is little useful information – relative to supervised settings – towards tuning procedures. Data-driven tuning in fact has received much attention in the literature on density-ratios, but with no theoretical guarantees to our knowledge.

Results. Many interesting facts, which hold in the RKHS setting (under various risk measures), are shown to hold more generally in these somewhat more complex settings:

- Optimal rates (under $L_{1,P}$ risk) depend only on the smoothness of the ratio $f = f_Q/f_P$, and not on f_Q nor f_P, supporting the belief that plugging in estimates for the densities f_Q, f_P is suboptimal.

Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54. Copyright 2017 by the author(s).

- Optimal rates depend only on the intrinsic dimension of data, i.e., this problem – unlike density estimation – escapes the curse of dimension when the high-dimensional data is structured.

Unlike in previous work, we further show that near-optimal rates are attainable by estimators tuned from data alone, i.e., with no prior distributional information. This is the most involved part of the analysis. We consider a sample-based tuning approach which relies on a stability criteria akin to so-called Lepski's methods (used in kernel density estimation [15, 16]). The present difficulties have to do with the lack of ground-truth (Q, P, smoothness, intrinsic dimension, are all unknown) and the fact that the risk measures – e.g. $\|\hat{f} - f\|_{1,P}$ – involve ill-bounded ratios that are not empirically stable. The analysis is made possible through a sequence of truncated empirical metrics, of independent technical interest, that form stable approximations to key components of the risks.

Paper outline. We discuss our results in more detail in Section 2, along with relevant prior work. The analysis starts in Section 3 with some preliminary setup and specification of the simple estimator used for upper-bounds. Our $L_{1,P}$ oracles rates are then presented and discussed in Section 4.1, while results on data-driven choice of r are presented in Section 4.2. Analyses of the main theorems are covered in Section 5, with some proofs relegated to the appendix.

2 DETAILED OVERVIEW AND PRIOR WORK

Estimating a density-ratio $f = f_Q/f_P$, is a difficult nonparametric problem which contains density-estimation as a special case (the case $P \equiv$ Lebesgue). Lower-bounds on density estimation therefore automatically apply. Hence, to understand attainable rates, we can restrict attention to upper-bounds (on any suitable estimator), provided these match known lower-bounds for the choice setting.

The work of [4] establishes general rates, in Hellinger distance, in terms of *bracketing entropy* for generic function classes $\mathcal{F}$; the estimator involves penalized empirical risk minimization (P-ERM) over the class $\mathcal{F}$. While P-ERM is not feasible for general $\mathcal{F}$, [4] shows how to instantiate the procedure when $\mathcal{F}$ is an RKHS (via the so-called *kernel trick*). Similar P-ERM over an RKHS is performed in [13, 14], where estimation rates are established in $L_{2,P}$, under different penalization approaches. Thus, so far, upper-bounds for actual estimators seem only known for P-ERM approaches over an RKHS.

While RKHS are quite general, they encode strong smoothness conditions (e.g., Sobolev classes are RKHS only when the smoothness index is higher than $D/2$) which do not capture less smooth functions such as Lipschitz or Hölder [17, 18]. In other words, rates for an RKHS do not carry over to these other smoothness classes, nor are the rates meaningfully comparable (as they involve different characterizations of smoothness). Furthermore, for our upper-bounds for Lipschitz classes, the P-ERM procedure is no longer feasible; however we will see that a simple local averaging procedure (described below) is sufficient to capture the minimax rates under these settings. Finally, as previously discussed, we are particularly interested in common situations where the data is not full-dimensional in $\mathbb{R}^D$; in particular, we assume a generic metric space $(\mathcal{X}, \rho)$ where the *ratio* f remains well-defined by simply letting Q absolutely continuous w.r.t. P (which is in fact also needed in the full-dimensional case, if f is to be well-defined).

In these general settings, we focus on $L_{1,P}$ as it seems natural for density-estimation type problems, for instance in the way it equates to total-variation [19]. Furthermore, it is easily shown that $L_{1,P}$ remains appropriate w.r.t. downstream applications as it yields direct bounds on the target errors in such applications (just as $L_{2,P}$). This is captured in the following simple proposition.

Proposition 1. *Consider the motivating problem of integrating $\mathbb{E}_Q\, g = \mathbb{E}_P\, g \cdot f$, where g is bounded. Let $\hat{f}$ be an estimate of the ratio f. Then the integration error in substituting $\hat{f}$ for f is bounded as:*

$$\mathbb{E}_P\, g \cdot (\hat{f} - f) \leq |g|_{sup} \cdot \mathbb{E}_P |\hat{f} - f| \doteq |g|_{sup} \cdot \|\hat{f} - f\|_{1,P}.$$

Direct estimation. Our upper-bounds are established via a simple local estimator suited to the local-nature of our smoothness conditions. Namely, this estimator $\hat{f}(x)$ is of the form $Q_n(B(x,r))/P_N(B(x,r))$, over a ρ-ball $B(x,r) \subset \mathcal{X}$, where Q_n, P_N denote empirical masses from two samples $\mathbf{X}_Q \sim Q$ and $\mathbf{X}_P \sim P$ of sizes n and N. The rates are in terms of the (first order) smoothness β of the unknown ratio $f = f_Q/f_P$ and do not involve the smoothness of either density f_Q or f_P. In particular, this implies that plugging in two different density estimators (i.e. estimate f as $\hat{f}_Q/\hat{f}_P$) can be suboptimal. To see this, consider for instance the case where $f_Q = f_P$, i.e., f is constant and thus infinitely smooth: estimating $f_Q = f_P$ can be arbitrarily difficult given lack of smoothness, however f is easy to estimate (our rates in this case are of the parametric form $(n \wedge N)^{-1/2}$ as $\beta \to \infty$). This supports a common belief (see e.g. [2]) that *density-ratio estimation is easier than estimating either densities*. Note however that both problems are of the same complexity in a minimax sense, as shown in the present analysis.

Structured data. Under a suitable notion of *intrinsic dimension* d of $(\mathcal{X}, \rho)$ (see Definition 2), our oracle $L_{1,P}$ rates are then of the form $(n \wedge N)^{-\beta/(2\beta+d)}$, where β captures the smoothness of f (e.g., Hölder exponent, see Assumption 1). In particular, when $\mathcal{X}$ is a structured subspace of $\mathbb{R}^D$ with $d \ll D$ (e.g. a manifold, or sparse under some unknown dictionary), the rates are significantly faster than the high-dimensional worst-case $(n \wedge N)^{-\beta/(2\beta+D)}$. In contrast, Lebesgue-density estimation is ill-defined in such favorable settings (distributions supported on structured $\mathcal{X}$ are singular). We note that the estimator of [14] is also shown to attain rates adaptive to d in the particular case where data lives on a d-manifold in $\mathbb{R}^D$. However their rates require knowledge of d in properly setting regularization parameters. Avoiding such requirements is discussed next.

Data driven tuning. We now turn to the more practical question of picking hyperparameters (here r) from data, a largely open problem in unsupervised settings, which we aim to understand better. We note that the procedures of [20, 2, 13, 14] all provide nontrivial data-driven procedures, but which are however not tied into their statistical guarantees.

We derive a data-driven procedure, by extending insights from so-called Lepski's method [15, 16] suited to adapting to unknown smoothness. Here we are interested in the possibility of adapting to both unknown smoothness and, most importantly for structured data, to unknown dimension. The data-driven procedure then consists of properly balancing a sample-based surrogate for variance[1] towards picking an estimate which is *stable* to small changes in bandwidth r. Obtaining high-probability finite sample rates are complicated here by the fact that estimates (being ratios) are potentially unbounded, and therefore might not concentrate. The analysis therefore proceeds by introducing truncated estimators, along with a sequence of truncated (empirical) risk surrogates, which do concentrate. We can then show that the data-driven choice of r results in an estimate whose risk nearly matches the oracle risk order of $(n \wedge N)^{-\beta/(2\beta+d)}$; both smoothness β and intrinsic d are a priori unknown. The main assumptions are that we have access to a rough upper-bound F on f (a mild assumption since F might be obtained from a first pass estimate), and that the base measure P is upper-bounded in a sense that remains general. The approach yields important insights on quantities that most affect tuning choices for this problem.

3 PRELIMINARIES

3.1 Data and Distributions

We have access to two random samples $\mathbf{X}_P \sim P^N$ and $\mathbf{X}_Q \sim Q^n$, where Q and P are probability distributions on $(\mathcal{X}, \rho)$, where ρ is some known metric. Furthermore Q is absolutely continuous with respect to P, and we assume w.l.o.g. that $\text{supp}(P) = \mathcal{X}$. Also we assume for simplicity that $\mathcal{X}$ has diameter $\sup_{x,x'} \rho(x, x') = 1$. We need the following regularity condition for balls on $\mathcal{X}$.

Definition 1 (Balls on $\mathcal{X}$)**.** *Let* $B(x, r) \triangleq \{x' \in \mathcal{X} : \rho(x, x') \leq r\}$ *denote a ball under* ρ. *We assume the class* $\mathcal{B} \triangleq \{B(x, r) : x \in \mathcal{X}, r > 0\}$ *of all balls has finite VC dimension at most* $V_{\mathcal{B}}$.

Our goal is to estimate the Radon-Nikodym derivative of Q w.r.t. P. We denote this derivative by f, which by definition satisfies, for all measurable $A \subset \mathcal{X}$:

$$Q(A) = \int_A f\, dP.$$

Remember that if Q and P are both dominated by some base measure σ on $\mathcal{X}$, with respective densities f_Q and f_P w.r.t. σ, then $f =_{\sigma-a.e.} f_Q/f_P$, which justifies the *density-ratio* terminology.

Interestingly, σ needs not be known; this implies in particular that $\mathcal{X}$ might be an unknown subspace of $\mathbb{R}^D$ of lower *dimension* d which we might adapt to. We adopt the following notion of metric dimension.

Definition 2. *The integer d is a* **covering dimension** *of $(\mathcal{X}, \rho)$ if there exists C such that for any $0 < r \leq 1$, $\mathcal{X}$ has an r-cover of size at most Cr^{-d}.*

This simple notion generalizes other notions of intrinsic dimension such as *doubling dimension* common in Machine Learning (see e.g. [21] for a nice overview), and the smallest such d can be shown to tightly capture the dimension of structured data, e.g., low-dimensional manifolds (under curvature conditions), and sparse data (under a bounded-size dictionary) [22].

We will express initial results in terms of the modulus of continuity of the derivative f:

Definition 3 (Modulus of continuity of f)**.** *For any $x \in \mathcal{X}$ and $r > 0$, define*

$$\epsilon_f(x, r) = \sup_{x' \in B(x,r)} |f(x) - f(x')|, \text{ and}$$

$$\epsilon_{P,f}(r) \doteq \mathbb{E}_P [\epsilon_f(X, r)].$$

Parts of the analysis requires more precision, so define

$$\hat{\epsilon}_f(x, r) = \sup_{x' \in B(x,r)} f(x') - f(x), \text{ and}$$

$$\check{\epsilon}_f(x, r) = \sup_{x' \in B(x,r)} f(x) - f(x').$$

We assume $\epsilon_{P,f}$ is bounded. The definitions capture the smoothness of f in some generality which will prove useful. Results under Hölder smoothness can then be obtained as corollaries.

3.2 Basic Estimator

We start with the following basic estimates; the analysis concerns the last estimator $\tilde{f}_r$, while the other estimators are related and serve to yield initial insights into $\tilde{f}_r$. In the second part of the paper we analyze (near-optimal) choices of the *bandwidth* parameter r from data.

[1] "Variance" is used loosely since we're bounding L_1, so "standard deviation" might be more appropriate.

Definition 4 (Estimates.)**.** *Let Q_n, P_N denote resp. empirical distributions w.r.t. $\mathbf{X}_Q$ and $\mathbf{X}_P$. Given a* bandwidth $0 < r \leq 1$, *define the following* basic *estimates:*

$$f_r(x) \triangleq \frac{Q_n(B(x,r))}{P(B(x,r))}, \quad \hat{f}_r(x) \triangleq \frac{Q_n(B(x,r))}{P_N(B(x,r))},$$

$$\text{and } \tilde{f}_r(x) = \hat{f}_r(x) \cdot \mathbb{1}\mathcal{E}_r(x), \text{where}$$

$\mathcal{E}_r(x)$ denotes the event $\{P_N(B(x,r)) \geq 72\alpha_{m,\mathcal{B}}\}$, $m = n \wedge N$, $\alpha_{n_0,\mathcal{B}} \doteq (V_\mathcal{B} \ln 2n_0 + \ln(8/\delta))/n_0$ for any integer n_0 (the quantity $V_\mathcal{B}$ is given in Definition 1).

The estimator $\tilde{f}_r$ simply truncates $\hat{f}_r$ in regions of low-density, while $\hat{f}_r$ differs from f_r only in the normalization by empirical mass P_N rather than by the unknown P. We will proceed in steps by first bounding f_r then $\hat{f}_r$ at a point x. The reason for the truncation will then become apparent as we establish high probability results for $\tilde{f}_r$. Intuitively, we have a *confident* estimate $\hat{f}_r$ whenever $\mathcal{E}_r(x)$ holds, i.e. enough samples contributed to the estimate.

We note that, when $\mathcal{X}$ has general diameter $\Delta_\mathcal{X}$ (rather than 1 as in our simplification), we will just use bandwidths $r = r_0 \cdot \Delta_\mathcal{X}, 0 < r_0 \leq 1$. The analysis trivially extends to general diameter.

4 MAIN RESULTS

4.1 Rates for $\tilde{f}_r$

Our first results (Theorem 1 and Corollary 1) aim to first understand which rates are attainable given potential knowledge of distributional parameters.

We consider an $L_{1,P}$ risk defined for any estimate f' as $\|f' - f\|_{1,P} = \mathbb{E}_P |f'(X) - f(X)|$. The rates for any $\tilde{f}_r, r \in (0,1]$, are first obtained in terms of $\epsilon_{P,f}(r)$.

Theorem 1 ($L_{1,P}$ rates for $\tilde{f}_r$)**.** *Define $\bar{F} \doteq \|f\|_{2,P} \leq \sup_x f(x)$. Let $0 < \delta < 1$. Let $m = n \wedge N$. For any integer n_0, let $c_{n_0,\mathcal{B}} \doteq (V_\mathcal{B} \ln 2n_0 + \ln(8/\delta))$. With probability at least $1 - 2\delta$ over the choice of $\mathbf{X}_P$ and $\mathbf{X}_Q$, for all $r > 0$,*

$$\begin{aligned}\|\tilde{f}_r - f\|_{1,P} \leq & C\Bigg(\bar{F}\sqrt{\frac{c_{N,\mathcal{B}}}{N \cdot r^d}} + \sqrt{\frac{c_{n,\mathcal{B}}}{n \cdot r^d}}\left(1 + \sqrt{\epsilon_{P,f}(r)}\right) \\ & + (\bar{F} \vee 1)\frac{c_{n,\mathcal{B}}}{n \cdot r^d} + \bar{F}\frac{c_{N,\mathcal{B}}}{N \cdot r^d}\Bigg) \\ & + 2\epsilon_{P,f}(r) + \bar{F}\delta,\end{aligned}$$

for some C depending on $\mathcal{X}$.

Let $m = n \wedge N$, the above rate is of the form $\sqrt{1/(m \cdot r^d)} + \epsilon_{P,f}(r)$, and is simply given in enough detail to reflect the contribution of the different samples $\mathbf{X}_P$ and $\mathbf{X}_Q$.

Remark 1. *The confidence parameter might be chosen as $O(n^{-C})$ for some constant C so the last term $\bar{F}\delta$ is in fact of lower-order than other terms. The error term $\bar{F}\delta$ is contributed by* less-confident *estimates $\tilde{f}_r(x)$ (where $\mathbb{1}\mathcal{E}_r(x) = 0$, i.e., $B(x,r)$ is nearly empty), and in fact disappears if we assume P is lower-bounded on $\mathcal{X}$.*

Remark 2. *For the above theorem, we can actually relax $\mathcal{E}_r(x)$ in the definition of $\tilde{f}_r$ to hold when $P_N(B(x,r)) \gtrsim \alpha_{N,\mathcal{B}}$ (rather than when $P_N(B(x,r)) \gtrsim \alpha_{m,\mathcal{B}}$). The stricter threshold is needed in Section 4.2 for sample-driven choices of r, and ensures that the estimate is bounded for small m.*

One is also often interested in the limiting case of $N \to \infty$. This corresponds essentially to rates on f_r (P known, but perhaps not d) and are given in the appendix.

When f is sufficiently smooth, for instance Lipschitz or Hölder, we can minimize the above upper-bound over choices of $r > 0$. This is done next.

Assumption 1 (Smoothness of f)**.** *Let the derivative f satisfy, for some $\lambda, \beta > 0$, the (relaxed) Hölder condition $\epsilon_{P,f}(r) \leq \lambda r^\beta$, $\forall r \in (0,1]$ (or all $r \in (0, r_0]$).*

The condition is clearly satisfied when f is (λ, β)-Hölder, i.e. $|f(x) - f(x')| \leq \lambda \rho(x,x')^\beta$. We note however that this is first-order smoothness (appropriate to our first-order estimates) and therefore is most interesting for $0 < \beta \leq 1$ (larger β hold for piecewise constant f, or atomic P).

We have the following corollary under the above smoothness condition. The rate is simply expressed in terms of $m = n \wedge N$, the smallest sample size, with no requirement on n/N.

Corollary 1 (Oracle rates)**.** *Assume the conditions of Theorem 1. Let the derivative f satisfy Assumption 1, for some $\lambda, \beta > 0$. Let $\tilde{f}_r$ denote the truncated estimator of Theorem 1. Let $m = n \wedge N$. There exists $C_0 = C_0(\mathcal{X}), C = C(\mathcal{X}, \bar{F})$, $m_0 = m_0(\mathcal{X}, \lambda)$ such that the following holds. For all $m \geq m_0$, we have with probability at least $1 - 2\delta$ over the choice of $\mathbf{X}_Q$ and $\mathbf{X}_P$ that, given $r = C_0\left(\log(m^{V_\mathcal{B}}/\delta)/(\lambda^2 m)\right)^{\beta/(2\beta+d)}$,*

$$\begin{aligned}\|\tilde{f}_r - f\|_{1,P} \leq & C\lambda^{d/(2\beta+d)} \cdot \left(\frac{\log(m^{V_\mathcal{B}}/\delta)}{m}\right)^{\beta/(2\beta+d)} \\ & + \bar{F}\delta.\end{aligned}$$

Proof. For m sufficiently large, $\frac{\log(m^{V_\mathcal{B}}/\delta)}{m} \leq 1$ and the above $r \leq 1$. We then have by Theorem 1 that for some $C_1 = C_1(\bar{F})$, with probability at least $1 - 2\delta$,

$$\begin{aligned}\|\tilde{f}_r - f\|_{1,P} \leq & C_1\sqrt{\frac{\log(m^{V_\mathcal{B}}/\delta)}{m \cdot r^d}} + 2\lambda r^\beta \\ & + C_1\sqrt{\frac{\log(m^{V_\mathcal{B}}/\delta)}{m \cdot r^d}}\sqrt{\lambda r^\beta} + \bar{F}\delta \\ \leq & C\lambda^{d/(2\beta+d)} \cdot \left(\frac{\log(m^{V_\mathcal{B}}/\delta)}{m}\right)^{\beta/(2\beta+d)} + \bar{F}\delta.\end{aligned}$$

□

Remark 3 (Rate Optimality)**.** *The above rate matches (up to log terms) known lower-bounds on L_1 density estimation for Hölder classes of densities over $\mathbb{R}^d$ [19], and is therefore tight in this respect since our setting is more general.*

The above results do not involve direct assumptions on densities f_P nor f_Q, as the only relevant complexity is that of f. Also, as previously discussed, if $\mathcal{X}$ were a subspace of $\mathbb{R}^D$, then the dependence of the rates on the intrinsic dimension d rather than on D, allows for fast rates for structured data in high-dimensional settings. This is for instance of interest in *transfer-learning* problems where much of the intended applications involve high-dimensional, but structured data. Examples are robot-control, spam filtering, brain-computer interface, NLP, and more (see e.g. [23, 24] for detailed overviews).

4.2 Data-driven Choice of r

The main question in this section is whether a procedure that selects r based on the data can (nearly) achieve the oracle rate of Theorem 1, with no a priori knowledge of distributional parameters (d, λ, β). We show that this is the case for the approach described below, which is based on a stability type criteria.

We note that, a different approach, more akin to cross-validation is possible. The main intuition, used for instance in [20, 2, 13, 14] (and also [25] in the case of density-estimation) is to decompose the $L^2_{2,P}$ risk $\|\tilde{f}_r - f\|^2_{2,P}$ into terms $\|f_r\|^2_{2,P}$ and $\|f_r\|_{1,Q}$ independent of f, and estimate these terms from data. However, this estimation error is $O(m^{-1/4})$ on the final $L_{1,P}$ error, at least using common concentration inequalities, and thus is too large for our purpose since it can dominate the rate of Theorem 1 (the problem is in having to bound $L_{1,P}$ by $L_{2,P}$). This is discussed further in the Appendix with a full-analysis of the approach. In contrast, the approach described next is directly designed around the $L_{1,P}$ error.

A stability type criteria

Our approach below extends insights from so-called Lepski's methods [16, 15], and proceeds from small values of r to large, with the added computational benefit of early stopping (with no need to evaluate the full range of r values).

Basic Lepski's methods aim at adapting to the unknown smoothness (λ, β) of a target f, however assuming a known base measure $P \equiv$ Lebesgue, and known dimension d. Such knowledge informs the choice of a variance upper bound (e.g. of the form $1/\sqrt{mr^d}$ of Theorem 1) which can be balanced with bias towards choosing a good r. The main intuition is as follows.

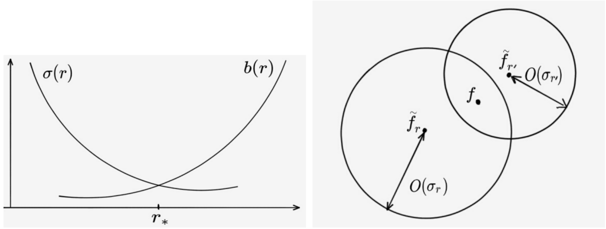

Figure 1: Data-driven choice of r: main intuition. (Left) For $r > 0$, the error $\|\tilde{f}_r - f\|_{1,P}$ is bounded by the sum of variance and bias terms $\sigma(r)$, $b(r)$, depending on d, λ, β unknown. The choice r^* balances such terms. (Right) for any $r < r' < r_*$, the variance term σ dominates so the depicted balls (in $L_{1,P}$) all contain f, therefore must intersect. The balls get smaller as $r, r' \to r^*$, so $\tilde{f}_r$ and $\tilde{f}_{r'}$ must get close (more stable estimates). Setting a proper threshold requires a suitable data-driven surrogate for $\sigma(r)$.

Intuition. Suppose r_*, unknown, balances variance and bias i.e. $\sigma(r^*) \approx b(r^*)$ where $\sigma(r)$ and $b(r)$ are variance and bias terms respectively of the form $1/\sqrt{mr^d}$ and λr^β, d, λ, β unknown. Then for smaller $r \le r_*$, we would have $\|\tilde{f}_r - f\|_{1,P} \lesssim \sigma(r) + b(r) \lesssim 2\sigma(r)$, hence for $r', r \le r_*$,

$$\|\tilde{f}_r - \tilde{f}_{r'}\|_{1,P} \lesssim 2(\sigma(r) + \sigma(r')), \tag{1}$$

decreasing as $r, r' \to r_*$. In other words, estimates $\tilde{f}_r$ get more *stable* to changes in r as we increase r from 0 to r_*. This is depicted in Figure 1.

Thus, we might increase r and check changes between $\tilde{f}_r$ and any $\tilde{f}_{r'}$, $r' \le r$ till (1) no longer holds. At that point we know r has gone past r_* since changes in estimates have increased. We can then return the last r for which (1) held (w.r.t. other $r' \le r$). The corresponding estimate $\tilde{f}_r$ can be argued to not be too far from $\tilde{f}_{r_*}$ which is of optimal rate w.r.t. unknown smoothness (λ, β). This nice idea, inherent in Lepski's methods, however requires we know (1), i.e., that we can approximate $L_{1,P}$ and know the dimension d.

In contrast, here the base measure P is unknown, so are its support $\mathcal{X}$ and dimension d. We therefore need an appropriate surrogate for variance in terms of sample quantities. The analysis provides a clue: for any *confident* estimate $\tilde{f}_r(x)$, i.e. where $\mathcal{E}_r(x)$ holds, the quantity $1/\sqrt{mP(B(x,r))}$ (P might be replaced by P_N) appears to control variance at a point x. However, we have no useful such surrogate when $\mathcal{E}_r(x)$ fails. We will therefore separate integration (in approximating $L_{1,P}$) over confident point-estimates and non-confident ones. This involves introducing a series of pseudo-metrics $\Delta_{M,r}$ (defined on a validation sample of size M) that serve to compare confident estimates, and which can be shown to be bounded by a variance surrogate (truncated for the sake of concentration). The integration over *non-confident* estimates is then handled separately by showing that they occur (under the final choice of r) with probability of order at most our target risk bound.

The procedure requires an upper-bound F on f, which might be obtained via a first pass estimate (see Proposition 2). The bound $V_{\mathcal{B}}$ on VC dimension of $\mathcal{B}$ can be replaced by $O(D)$ for data in $\mathbb{R}^D$.

Procedure 1 (Stability):
SETUP. Let $R \doteq \left\{r_i \doteq 2^{-i}\right\}_{i=0}^{k}$, for some integer k, denote values of r.

Assume a known upper-bound $F \geq \sup_x f(x)$ (in fact F needs only be valid w.h.p.). Let $\delta \in (0,1)$, $m \doteq N \wedge n$, $\alpha_{m,\mathcal{B}} \doteq (V_{\mathcal{B}} \ln 2m + \ln(8,\delta))/m$.

Let $\mathbf{X}'_P \sim P^M$ denote a validation sample of size M independent of $\mathbf{X}_P$. Let P_M denote the empirical distribution w.r.t. $\mathbf{X}'_P$.

Now, let $\mathcal{E}_r(x)$ be the event $P_N(B(x,r)) \geq 72\alpha_{m,\mathcal{B}}$, and for any $g, g' : \mathcal{X} \mapsto \mathbb{R}, r > 0$, define the pseudo-metrics $\Delta_{M,r}\left(g, g'\right) = \|(g - g') \cdot \mathbb{1}\mathcal{E}_r\|_{1,P_M}$.

Variance surrogate: $\hat{e}_{m,r}(x) = \sqrt{\dfrac{24\alpha_{m,\mathcal{B}}}{P_N(B(x,r))}}$ if $\mathcal{E}_r(x)$

else, $\hat{e}_{m,r}(x) = \sqrt{1/3}$.

Define $\epsilon_{M,\delta} \doteq 15F\sqrt{\log(2k^2/\delta)/2M}$. Also define $\gamma(r) \doteq 8F\|\hat{e}_{m,r}\|_{1,P_M} + 2\epsilon_{M,\delta}$.

PROCEDURE. For $i = k-1$ to 0 do:
If $\exists j > i$ s.t. $\Delta_{M,r_j}\left(\tilde{f}_{r_i}, \tilde{f}_{r_j}\right) > \gamma(r_i) + \gamma(r_j)$, return $r = r_{i+1}$, otherwise continue.
(If the loop ends without returning r, return $r = r_0$).

The main theorem of this section requires in addition that P be bounded in the following sense.

Assumption 2. *P is* upper-bounded*, i.e. $\exists C_0$ such that $\forall 0 < r \leq 1, x \in \mathcal{X}$, $P(B(x,r)) \leq C_0 r^d$.*

The above is a rather mild assumption: it holds for instance if P has an upper-bounded Lebesgue density on $\mathbb{R}^d$ (then $P(B(x,r)) \leq c \cdot \text{vol}\left(B(x,r)\right) \propto r^d$), and more generally if P has an upper-bounded density w.r.t. a volume measure on a compact metric $\mathcal{X}$ of dimension d.

Theorem 2. *Let the derivative f satisfy Assumption 1, for some $\lambda, \beta > 0$, and $\sup_x f(x) < \infty$. Suppose Assumption 2 holds for P. Let $m = n \wedge N$ and let $0 < \delta < 1$. Define $\bar{F} = \|f\|_{2,P}$ as before. There exist $m_1 = m_1(\mathcal{X}, f)$, $C_1 = C_1(\mathcal{X}, f)$ such that the following holds with probability at least $1 - 5\delta$ over the choice of $\mathbf{X}_P, \mathbf{X}_Q$ and $\mathbf{X}'_P$.*

Choose $k = \lceil \log m \rceil$, and suppose $m > m_1$. Let r be the value returned by Procedure 1. We have

$$\|\tilde{f}_r - f\|_{1,P} \leq C_1 \lambda^{d/(2\beta+d)} \left(\frac{\log(m^{V_{\mathcal{B}}}/\delta)}{m}\right)^{\beta/(2\beta+d)} + 32\epsilon_{M,\delta} + \bar{F}\delta.$$

Thus, (provided $M = \Omega(m)$) the procedure attains a rate of nearly the same order as the oracle rate of Corollary 1, with no a priori knowledge of (λ, β) nor d. However, if $V_{\mathcal{B}}$ is chosen as $O(D)$, then m needs to be at least linear in D, which is still benign w.r.t. non-adaptive exponential rates in D. Thus, importantly, the analysis reveals key data-dependent quantities that tightly control the choice of r.

As mentioned before, the upper-bound F required by the procedure can be chosen from a first pass estimate. This is stated in the following simple proposition whose proof is in the appendix. Furthermore such an F picked from estimates will not be too large (see Proposition 3). Thus, the overall procedure can be made fully independent of distributional unknowns.

Proposition 2. *Suppose $\sup_x f(x)$ is attained at $x_0 \in \mathcal{X}$, and f is continuous in a neighborhood of x_0. Let $0 < \delta < 1$. Suppose F is picked as $4\max_{x \in \mathbf{X}_P, r \in R, \textit{ s.t. } \mathcal{E}_r(x)} \tilde{f}_r(x)$. Then, for m sufficiently large, w.p. at least $1 - 3\delta$, we have $F \geq \sup_x f(x)$.*

5 ANALYSIS OVERVIEW

In this section we go over important details of the main results. Some proofs are relegated to the appendix. The proof of Theorem 1 and 2 both require the same starting lemmas. These involve deriving pointwise bounds for f_r, then $\hat{f}_r$, which are then properly integrated to obtain Theorem 1, whose proof is outlined in Section 5.3. Theorem 1 is then proved in Section 5.4.

5.1 Pointwise Rates for f_r

Our first lemma analyzes the behavior of the basic estimate f_r at a point x. In other words, if we knew P (assumed by f_r), what rate should be expected.

Lemma 1 (Rate for $f_r(x)$.)**.** *Let $0 < \delta < 1$. Define $c_{n,\mathcal{B}} = V_{\mathcal{B}} \ln 2n + \ln(8/\delta)$, where $\mathcal{B}$ is the set of balls on $\mathcal{X}$. We have w.p. $1 - \delta$, for all $x \in \mathcal{X}$ and $r > 0$,*

$$f_r(x) - f(x) \leq \hat{\epsilon}_f(x,r) + \sqrt{\frac{c_{n,\mathcal{B}} \cdot (f(x) + \hat{\epsilon}_f(x,r))}{n \cdot P(B(x,r))}} + \frac{c_{n,\mathcal{B}}}{n \cdot P(B(x,r))}, \quad \textit{and}$$

$$f(x) - f_r(x) \leq \check{\epsilon}_f(x, r) + \sqrt{\frac{3c_{n,\mathcal{B}} \cdot (f(x) + \hat{\epsilon}_f(x, r))}{n \cdot P(B(x, r))}} + \frac{3c_{n,\mathcal{B}}}{n \cdot P(B(x, r))}.$$

Notice that $\hat{\epsilon}_f$ and $\check{\epsilon}_f$ in the above bounds can be replaced by (the less tight) modulus ϵ_f.

5.2 Pointwise Rates for $\hat{f}_r$

The following lemma relate the estimate $\hat{f}_r$, given two samples from P and Q respectively, to the estimate f_r which assumes knowledge of P. Rates for $\hat{f}_r$ are then easily obtained from the rates for f_r established in Lemma 1.

Lemma 2 ($\hat{f}_r(x)$ vs $f_r(x)$)**.** *Let $0 < \delta < 1$. Fix the sample $\mathbf{X}_Q$. Define $c_{N,\mathcal{B}} = V_{\mathcal{B}} \ln 2N + \ln(8/\delta)$, where $\mathcal{B}$ is the set of balls on $\mathcal{X}$. The following holds w.p. at least $1 - \delta$ (over the choice of $\mathbf{X}_P$), uniformly for all $x \in \mathcal{X}$ and $r > 0$.*

If r satisfies $P_N(B(x, r)) \geq 72c_{N,\mathcal{B}}/N$, we have

$$\hat{f}_r(x) \leq f_r(x) \cdot \left(1 + 2\sqrt{\frac{3c_{N,\mathcal{B}}}{N \cdot P(B(x, r))}} + 2\frac{c_{N,\mathcal{B}}}{N \cdot P(B(x, r))}\right).$$

For any $r > 0$, we have

$$\hat{f}_r(x) \geq f_r(x) \cdot \left(1 - \sqrt{\frac{c_{N,\mathcal{B}}}{N \cdot P(B(x, r))}} - \frac{c_{N,\mathcal{B}}}{N \cdot P(B(x, r))}\right).$$

5.3 Integrated Rates

As discussed earlier, the bounds on $L_{1,P}$ error for $\hat{f}_r$ are best when $\mathcal{X} \subset \mathbb{R}^D$ is of unknown lower-dimension d. The pointwise errors from earlier lemmas contain $1/P$ ratios. These ratios integrate out (via a covering argument) in terms of the unknown intrinsic dimension d of $\mathcal{X} \equiv$ supp(P).

Lemma 3 (Integrating $(1/P)$ on $\mathcal{X}$)**.** *Let d denote the* covering dimension *of $\mathcal{X}$, and suppose $\mathcal{X}$ is bounded with diameter 1. Let $0 < r \leq 1$, we have:*

$$\mathbb{E}_P \left[\frac{1}{P(B(X, r))}\right] \leq Cr^{-d}, \text{ for a constant } C = C(\mathcal{X}).$$

Theorem 1 is then established by combining the above results with additional VC concentration bounds. The full proof is given in the appendix. We outline the main ideas below.

Proof outline for Theorem 1. Define $e_{n,r}(x) = \sqrt{\frac{3c_{n,\mathcal{B}}}{n \cdot P(B(x,r))}}$, and $e_{N,r}(x) = \sqrt{\frac{24c_{N,\mathcal{B}}}{N \cdot P(B(x,r))}}$. Using Lemmas 1 and 2, we can show that, w.p. at least $1 - 2\delta$, the following holds for all $x \in \mathcal{X}$ satisfying $\mathcal{E}_r(x)$:

$$\left|\hat{f}_r(x) - f(x)\right| \leq e_{N,r}(x) \cdot f(x) + 2\left(e_{n,r}(x) \cdot \sqrt{(f(x) + \epsilon_f(x, r))} + e^2_{n,r}(x) + \epsilon_f(x, r)\right). \tag{2}$$

Integrating over $x \in \mathcal{X}$ and using Lemma 3, we obtain that $\mathbb{E}_P \left|\hat{f}_r(X) - f(X)\right| \cdot \mathbb{1}\mathcal{E}_r(x)$ is at most

$$C \cdot \bar{F} \cdot \sqrt{\frac{c_{N,\mathcal{B}}}{N \cdot r^d}} + 2\left(C\sqrt{\frac{3c_{n,\mathcal{B}}}{n \cdot r^d}} \left(1 + \sqrt{\epsilon_{P,f}(r)}\right) + C\frac{3c_{n,\mathcal{B}}}{n \cdot r^d} + \epsilon_{P,f}(r)\right),$$

for some C depending on $\mathcal{X}$. On the other hand, the integral $\mathbb{E}_P \left|\tilde{f}_r(X) - f(X)\right| \cdot \mathbb{1}\mathcal{E}_r^{\complement}(X)$ can be shown to be upper-bounded over an $r/2$-cover $\mathcal{X}_r$ as:

$$\begin{aligned}\mathbb{E}_P f(X) \cdot \mathbb{1}\mathcal{E}_r^{\complement}(X) &\leq \bar{F} \cdot \left(\sum_{x \in \mathcal{X}_r} 82\alpha_{m,\mathcal{B}}\right) + \bar{F}\delta \\ &\leq C\bar{F} \cdot r^{-d}\alpha_{m,\mathcal{B}} + \bar{F}\delta,\end{aligned} \tag{3}$$

for some C depending on $\mathcal{X}$. Combining the two parts of the integration yields the result. □

5.4 Data-driven Choice of r

The following proposition is needed in justifying the form of $\epsilon_{M,\delta}$ in the procedure.

Proposition 3. *Suppose* $\sup_x f(x) \leq F$. *Let* $0 < \delta < 1$. *With probability at least* $1 - 2\delta$ *over* $\mathbf{X}_Q$ *and* $\mathbf{X}_P$*, we have* $\max_{i \in [k]} \sup_x \tilde{f}_{r_i}(x) \leq 15F$.

The lemma establishes that our variance surrogate is bracketed by functions of r of similar order.

Lemma 4 (Bracketing $\hat{e}_{m,r}$ w.h.p.)**.** *Suppose Assumption 2 holds with some C_0.*
Define $\sigma_\flat(r) \doteq \min\left\{\sqrt{1/3}, \sqrt{8\alpha_{m,\mathcal{B}}/C_0 r^d}\right\}$, and $\sigma_\sharp(r) \doteq \sqrt{72C\alpha_{m,\mathcal{B}}/r^d}$, where C is as defined in Lemma 3. With probability at least $1 - 2\delta$ over $\mathbf{X}_P, \mathbf{X}'_P$, we have $\forall r \in (0, 1]$,

$$\sigma_\flat(r) \leq \|\hat{e}_{m,r}\|_{1,P_M} \leq \sigma_\sharp(r) + \sqrt{3\log(2/\delta)/M}.$$

We are now ready to show the main result on the data-driven choice of r.

Proof Outline for Theorem 2. By Lemma 4 above, w.h.p., the variance surrogate $\|\hat{e}_{m,r}\|_{1,P_M}$ behaves as $\sigma_\flat(r) \approx \sigma_\sharp(r) = O(1/\sqrt{mr^d})$; from the analysis of Theorem 1 it is evident that any r s.t. $1/\sqrt{mr^d} \approx \lambda r^\beta$ would yield an estimate of near-minimax order. Thus, the main idea is to show that the value of r returned properly balances $\|\hat{e}_{m,r}\|_{1,P_M}$ with the bias upper-bound λr^β.

Let $\hat{r}$ denote the largest $r \in R$ such that $\|\hat{e}_{m,r}\|_{1,P_M} \geq 2\lambda r^\beta$. It can be shown that $r_k \leq \hat{r} < r_0$. Furthermore, for any $r_i < r_j \leq \hat{r} \in R$, we would argue (using in particular (2)) that, w.h.p.,

$$\Delta_{M,r_i}\left(\tilde{f}_{r_i}, \tilde{f}_{r_j}\right) \leq \Delta_{M,r_i}\left(\tilde{f}_{r_i}, f\right) + \Delta_{M,r_j}\left(\tilde{f}_{r_j}, f\right) \\ \leq \gamma(r_i) + \gamma(r_j),$$

in other words, let r be returned by the procedure, we necessarily have $r \geq \hat{r}$, hence $\gamma(r) \leq \gamma(\hat{r})$. Now, the return condition did not hold at r, so $\Delta_{M,\hat{r}}\left(\tilde{f}_r, \tilde{f}_{\hat{r}}\right) \leq \gamma(r) + \gamma(\hat{r}) \leq 2\gamma(\hat{r}) = O(\sigma_\sharp(\hat{r}))$. Now, the risk of $\tilde{f}_r$ can be integrated over two subsets of $\mathcal{X}$ defined by $\hat{r}$, that is:

$$\|\tilde{f}_r - f\|_{1,P} = \|(\tilde{f}_r - f)\mathbb{1}\mathcal{E}_{\hat{r}}\|_{1,P} + \|(\tilde{f}_r - f)\mathbb{1}\mathcal{E}_{\hat{r}}^{\complement}\|_{1,P} \tag{4}$$

The first term is close w.h.p. to

$$\Delta_{M,\hat{r}}\left(\tilde{f}_r, f\right) \leq \Delta_{M,\hat{r}}\left(\tilde{f}_r, \tilde{f}_{\hat{r}}\right) + \Delta_{M,\hat{r}}\left(\tilde{f}_{\hat{r}}, f\right) \leq O(\sigma_\sharp(\hat{r})).$$

The second term of (4) is $O(\alpha_{m,\delta} \cdot \hat{r}^{-d}) = O(\sigma_\sharp(\hat{r})^2)$ by (3). Finally we bound $\sigma_\sharp(\hat{r}) \lesssim m^{-\beta/(2\beta+d)}$, by showing that, w.h.p., $\hat{r}$ is close to an explicit value $\tilde{r}$ for which $\sigma_\flat(\tilde{r}) = 2\lambda\tilde{r}^\beta$. □

6 FINAL REMARKS

We have shown that important differences between density-estimation and density-ratio estimation hold in general practical settings. In particular, density-ratio estimation gets considerably easier for structured data of low-intrinsic dimension, and depends only on the smoothness of the ratio rather than on the densities themselves. More general notions of smoothness are possible, for instance higher-order Hölder classes carefully defined over low-dimensional structures $\mathcal{X}$; this would likely require more sophisticated estimators and is left to further investigation.

As in density-estimation, oracle rates are nearly attainable through careful data-driven choice of hyperparameters (bandwidth r), i.e., with no distributional knowledge. While the data-driven procedure employed to establish this final result is of a technical nature, it is implementable and yields insights on important sample quantities involved in good choices of r, namely the empirical P_N-mass of balls on the metric seem quite important for smoothing estimators such as the one considered. Simulations on controlled data (see Figure 2) also reveal that the procedure is quite sensitive to initial estimates of an upper-bound F on $\sup_x f(x)$. While we show a simple way of doing so in theory, deriving proper initial estimates require further attention, especially in smaller sample regimes.

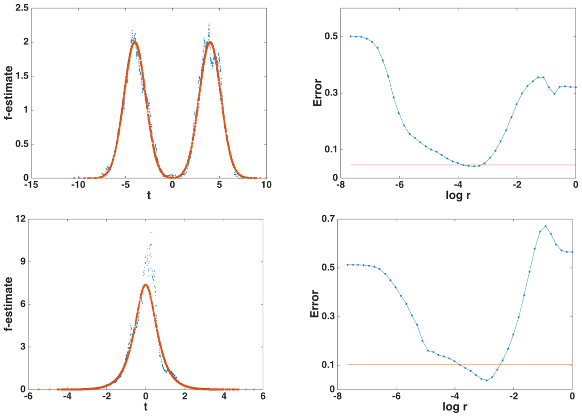

Figure 2: Simulations on 2 sets of controlled data. In both cases $X = t \cdot v$, where $t \in \mathbb{R}$ and $v \propto (1, 2, \ldots, D)$ is a fixed vector in $\mathbb{R}^D$, $N, n = 1000$, test-size (from P) $= 2000$. In each case we show (1) a qualitative plot of estimates (dotted) against true f (red line), and (2) the errors of every $r \in R$ (dotted blue) against the error for the data-driven choice of r (brown line). In implementation we use $M = N/2$, and $\epsilon_{M,\delta} = \sqrt{1/M}$; $\hat{e}_{m,r} = 1/\sqrt{mP(B_r)}$; for F we simply use the average estimated f out of a first pass (where we use $F = 1$). This actually makes a difference in the quality of results, while the setting of $\epsilon_{M,\delta}$ does not seem to matter much. The reported plots show typical results in these controlled settings. We note that the estimates are however poor (for any r) whenever Q is far from dominated by P, for instance as simulated by 2 Gaussians (Q and P) with far apart means. The data used above is as follows: (Top 2 plots.) $D = 20$; for Q, $t \sim 0.5(\mathcal{N}(-4,1) + N(4,1))$, while for P, $t \sim 0.5(\mathcal{N}(-4,4) + \mathcal{N}(4,4))$. (Bottom 2 plots) $D = 30$, for Q, $t \sim \mathcal{N}(0,1)$ while for P, $t \sim 0.5(\mathcal{N}(-2,1) + \mathcal{N}(2,1))$.

References

[1] Jing Qin. Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, 85(3):619–630, 1998.

[2] Masashi Sugiyama, Taiji Suzuki, Yuta Itoh, Takafumi Kanamori, and Manabu Kimura. Least-squares two-sample test. *Neural Networks*, 24(7):735–751, 2011.

[3] Shohei Hido, Yuta Tsuboi, Hisashi Kashima, Masashi Sugiyama, and Takafumi Kanamori. Statistical outlier detection using direct density ratio estimation. *Knowledge and information systems*, 26(2):309–336, 2011.

[4] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In *NIPS*, pages 1089–1096, 2007.

[5] Akshay Krishnamurthy, Kirthevasan Kandasamy, Barnabas Poczos, and Larry A Wasserman. Nonparametric estimation of renyi divergence and friends. In *ICML*, pages 919–927, 2014.

[6] Shashank Singh and Barnabás Póczos. Exponential concentration of a density functional estimator. In

Advances in Neural Information Processing Systems, pages 3032–3040, 2014.

[7] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

[8] Shai Ben-David, Shai Shalev-Shwartz, and Ruth Urner. Domain adaptation–can quantity compensate for quality? *ISAIM*, 2012.

[9] Miroslav Dudík, Steven J Phillips, and Robert E Schapire. Correcting sample selection bias in maximum entropy density estimation. In *Advances in neural information processing systems*, pages 323–330, 2005.

[10] Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning for differing training and test distributions. In *Proceedings of the 24th international conference on Machine learning*, pages 81–88. ACM, 2007.

[11] Jiayuan Huang, Arthur Gretton, Karsten M Borgwardt, Bernhard Schölkopf, and Alex J Smola. Correcting sample selection bias by unlabeled data. In *Advances in neural information processing systems*, pages 601–608, 2006.

[12] Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. Efficient direct density ratio estimation for non-stationarity adaptation and outlier detection. In *Advances in neural information processing systems*, pages 809–816, 2009.

[13] Takafumi Kanamori, Taiji Suzuki, and Masashi Sugiyama. Statistical analysis of kernel-based least-squares density-ratio estimation. *Machine Learning*, 86(3):335–367, 2012.

[14] Qichao Que and Mikhail Belkin. Inverse density as an inverse problem: The fredholm equation approach. In *Advances in Neural Information Processing Systems*, pages 1484–1492, 2013.

[15] Oleg V Lepski and VG Spokoiny. Optimal pointwise adaptive methods in nonparametric estimation. *The Annals of Statistics*, pages 2512–2546, 1997.

[16] Alexander Goldenshluger and Oleg Lepski. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *The Annals of Statistics*, pages 1608–1632, 2011.

[17] Holger Wendland. *Scattered data approximation*, volume 17. Cambridge university press, 2004.

[18] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[19] Luc Devroye and Lászlò Györfi. *Nonparametric density estimation: the L1 view*, volume 119. John Wiley & Sons Inc, 1985.

[20] Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul V Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in neural information processing systems*, pages 1433–1440, 2008.

[21] Kenneth L Clarkson. Nearest-neighbor searching and metric space dimensions. *Nearest-neighbor methods for learning and vision: theory and practice*, pages 15–59, 2006.

[22] S. Dasgupta and Y. Freund. Random projection trees for vector quantization. *IEEE Transactions on Information Theory*, 55(7):3229–3242, 2009.

[23] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

[24] Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

[25] Larry Wasserman. *All of nonparametric statistics*. Springer Science & Business Media, 2006.

[26] O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to statistical learning theory. *Lecture Notes in Artificial Intelligence*, 3176:169–207, 2004.

A POINTWISE RESULTS

This section provides the proofs of the various pointwise results needed for Theorems 1.

We start with the following concentration results.

Lemma 5 (Relative VC bounds [26]). *Consider a collection $\mathcal{A}$ of measurable subsets of $\mathcal{X}$, of finite VC dimension $V_{\mathcal{A}}$. Let $0 < \delta < 1$. Suppose a sample of size n is drawn i.i.d. from a distribution ν over $\mathcal{X}$. For $A \in \mathcal{A}$, let ν_A denote the mass of A under the distribution, and let $\nu_{n,A}$ denote its empirical mass. Define $\alpha_{n,\mathcal{A}} = \left(V_{\mathcal{A}} \ln 2n + \ln(8/\delta)\right)/n$. Then with probability at least $1-\delta$ over the sampling, all $A \in \mathcal{A}$ satisfy*

$$\nu_A \le \nu_{n,A} + \sqrt{\nu_{n,A} \cdot \alpha_{n,\mathcal{A}}} + \alpha_{n,\mathcal{A}}, \text{ and}$$
$$\nu_{n,A} \le \nu_A + \sqrt{\nu_A \cdot \alpha_{n,\mathcal{A}}} + \alpha_{n,\mathcal{A}}.$$

We will often be interested in subsets A of $\mathcal{X}$ containing at least a few points (of either $\mathbf{X}_P$ or $\mathbf{X}_Q$). The following corollary concerns such subsets and will prove useful in much of the analysis.

Corollary 2. *Let $0 < \delta < 1$. Under the same conditions as Lemma 5, the following holds with probability at least $1-\delta$. For any $A \subset \mathcal{A}$ satisfying $\nu_{n,A} \ge 3\alpha_{n,\mathcal{A}}$, we have*

$$\nu_A - (\sqrt{3\nu_A \cdot \alpha_{n,\mathcal{A}}} + \alpha_{n,\mathcal{A}}) \le \nu_{n,A} \le 3\nu_A.$$

Proof of Corollary 2. By the second inequality in Lemma 5, whenever $\nu_A < \alpha_{n,\mathcal{A}}$ we have w.p. $\ge 1-\delta$ that $\nu_{n,A} < 3\alpha_{n,\mathcal{A}}$. Thus for any A satisfying $\nu_{n,A} \ge 3\alpha_{n,\mathcal{A}}$, we have $\nu_A \ge \alpha_{n,\mathcal{A}}$. This in turn implies by the same inequality that $\nu_{n,A} \le 3\nu_A$. It then follows from the first inequality of Lemma 5 that, under the same event for all such large $A \in \mathcal{A}$, $\nu_A \le \nu_{n,A} + \sqrt{3\nu_A \cdot \alpha_{n,\mathcal{A}}} + \alpha_{n,\mathcal{A}}$. □

The above corollary and the relative concentration bounds of Lemma 5 (rather than the more common $1/\sqrt{n}$ VC bounds) allows us to require mild conditions on the balls $B(x,r)$ involved in the estimates f_r and $\hat{f}_r$, namely that the balls contain just $\Omega(\ln n)$ points rather than order $\sqrt{n}$ points as would seem to be required through the use of $1/\sqrt{n}$ bounds. We are now ready to prove the first pointwise rate for f_r and then for $\hat{f}_r$.

Proof of Lemma 1. Let $\alpha_{n,\mathcal{B}}$ as given in Definition 4. By the first inequality of Lemma 5 we have, with probability at least $1-\delta$ that, for all $x \in \mathcal{X}$ and $r > 0$, $f_r(x)$ is at most

$$\begin{aligned}
&\frac{1}{P(B(x,r))}\left(Q(B(x,r)) + \sqrt{Q(B(x,r)) \cdot \alpha_{n,\mathcal{B}}} + \alpha_{n,\mathcal{B}}\right)\\
&= \frac{Q(B(x,r))}{P(B(x,r))} + \sqrt{\frac{Q(B(x,r))}{P(B(x,r))}} \cdot \sqrt{\frac{\alpha_{n,\mathcal{B}}}{P(B(x,r))}}\\
&\quad + \frac{\alpha_{n,\mathcal{B}}}{P(B(x,r))}\\
&\le (f(x) + \hat{\epsilon}_f(x,r)) + \sqrt{(f(x) + \hat{\epsilon}_f(x,r))} \cdot \sqrt{\frac{\alpha_{n,\mathcal{B}}}{P(B(x,r))}}\\
&\quad + \frac{\alpha_{n,\mathcal{B}}}{P(B(x,r))},
\end{aligned}$$

where we used the fact that $Q(B(x,r))$ is given as $\int_{B(x,r)} f\, dP \le (f(x) + \hat{\epsilon}_f(x,r)) \cdot P(B(x,r))$.

Now, by Corollary 2, we have w.p. at leat $1-\delta$ under the same event as above, that for any r satisfying $Q_n(B(x,r)) \ge 3\alpha_{n,B}$, $f_r(x)$ is at least

$$\begin{aligned}
&\frac{1}{P(B(x,r))}\left(Q(B(x,r)) - \sqrt{3Q(B(x,r)) \cdot \alpha_{n,\mathcal{B}}} - \alpha_{n,\mathcal{B}}\right)\\
&\ge (f(x) - \check{\epsilon}_f(x,r)) - \sqrt{(f(x) + \hat{\epsilon}_f(x,r))} \cdot \sqrt{\frac{3\alpha_{n,\mathcal{B}}}{P(B(x,r))}}\\
&\quad - \frac{\alpha_{n,\mathcal{B}}}{P(B(x,r))},
\end{aligned}$$

where we use both inequalities

$$\begin{aligned}
(f(x) + \hat{\epsilon}_f(x,r))P(B(x,r)) \ge Q(B(x,r)) = \int_{B(x,r)} f\, dP\\
\ge (f(x) - \check{\epsilon}_f(x,r))P(B(x,r)).
\end{aligned}$$

Finally, if $Q_n(B(x,r)) < 3\alpha_{n,B}$, then, under the same event as above, we have by Lemma 5, that $f_r(x)$ is at least

$$\begin{aligned}
&\frac{1}{P(B(x,r))}\left(Q(B(x,r)) - \sqrt{Q_n(B(x,r)) \cdot \alpha_{n,\mathcal{B}}} - \alpha_{n,\mathcal{B}}\right)\\
&\ge (f(x) - \check{\epsilon}_f(x,r)) - 3\frac{\alpha_{n,\mathcal{B}}}{P(B(x,r))}.
\end{aligned}$$

Combining these various statements yields the lemma. □

Proof of Lemma 2. Let $\alpha_{N,\mathcal{B}}$ as given in Definition 4. Define $e_{N,r}(x) \triangleq \sqrt{3P(B(x,r)) \cdot \alpha_{N,\mathcal{B}}} + \alpha_{N,\mathcal{B}}$, and suppose $P_N(B(x,r)) \ge 72\alpha_{N,\mathcal{B}}$. By Corollary 2, since $P_N(B(x,r)) \ge 72\alpha_{N,\mathcal{B}}$, we have w.p. at least $1-\delta$ that

$$\begin{aligned}
\hat{f}_r(x) &\le \frac{Q_n(B(x,r))}{P(B(x,r)) - e_{N,r}(x)}\\
&= f_r(x) \cdot \left(1 - \frac{e_{N,r}(x)}{P(B(x,r))}\right)^{-1}\\
&\le f_r(x) \cdot \left(1 + 2\frac{e_{N,r}(x)}{P(B(x,r))}\right),
\end{aligned}$$

where the last inequality holds since, again by Corollary 2, $P(B(x,r)) \geq 24\alpha_{N,\mathcal{B}}$, hence satisfies the quadratic equation $P(B(x,r)) \geq 2\left(6P(B(x,r)) \cdot \alpha_{N,\mathcal{B}} + 2\alpha_{N,\mathcal{B}}^2\right)^{1/2} \geq 2e_{N,r}(x)$.

Now, let $e'_{N,r}(x) \triangleq \sqrt{P(B(x,r)) \cdot \alpha_{N,\mathcal{B}}} + \alpha_{N,\mathcal{B}}$. By Lemma 5, w.p. at least $1-\delta$ (under the same event of Corollary 2), we have

$$\begin{aligned}\hat{f}_r(x) &\geq \frac{Q_n(B(x,r))}{P(B(x,r)) + e'_{N,r}(x)} \\ &= f_r(x) \cdot \left(1 + \frac{e'_{N,r}(x)}{P(B(x,r))}\right)^{-1} \\ &\geq f_r(x) \cdot \left(1 - \frac{e'_{N,r}(x)}{P(B(x,r))}\right).\end{aligned}$$

□

Proof of Lemma 3. Let $\mathcal{X}_r$ denote an $(r/2)$-cover of $\mathcal{X}$, of size $|\mathcal{X}_r| \leq Cr^{-d}$. For any $x \in \mathcal{X}_r$, let and any $x' \in X$ where $\Delta_{M,x}(x', \leq) r/2$, we have $B(\mathcal{X}_r(x), r/2) \subset B(x,r)$. Thus for $\alpha = 1$, $\mathbb{E}_P\left[\frac{1}{P(B(X,r))}\right]$ is at most

$$\begin{aligned}&\sum_{x \in \mathcal{X}_r} \mathbb{E}_P\left[\frac{1}{P(B(X,r))} \cdot \mathbb{1}(X \in B(x,r/2))\right] \\ &\leq \sum_{x \in \mathcal{X}_r} \frac{1}{P(B(x,r/2))} \cdot \mathbb{E}_P \mathbb{1}(X \in B(x,r/2)) \\ &= |\mathcal{X}_r| \leq Cr^{-d}.\end{aligned}$$

□

B PROOF OF THEOREM 1

The full proof of the theorem is given below.

Proof of Theorem 1. Define $e_{n,r}(x) = \sqrt{\frac{3c_{n,\mathcal{B}}}{n \cdot P(B(x,r))}}$, and $e_{N,r}(x) = \sqrt{\frac{24c_{N,\mathcal{B}}}{N \cdot P(B(x,r))}}$. Note that, by Corollary 2, we have w.p. at least $1-\delta$ that under event $\mathcal{E}_r(x)$, $P(B(x,r)) \geq 24\alpha_{N,\mathcal{B}}$. Thus we have $e_{N,r}(x) \leq 1$, and furthermore the bound in Lemma 2 yields

$$f_r(x) \cdot (1 - e_{N,r}(x)) \leq \hat{f}_r(x) \leq f_r(x) \cdot (1 + e_{N,r}(x)).$$

Rearranging Lemmas 1 and 2, we have w.p. at least $1-2\delta$ that, $\forall x \in \mathcal{X}$ s.t. $\mathcal{E}_r(x)$ holds,

$$\begin{aligned}&\left|\hat{f}_r(x) - f(x)\right| \leq e_{N,r}(x) \cdot f(x) + \\ &\quad 2\left(e_{n,r}(x) \cdot \sqrt{(f(x) + \epsilon_f(x,r))} + e_{n,r}^2(x) + \epsilon_f(x,r)\right),\end{aligned} \tag{5}$$

therefore, using Cauchy-Schwartz on product terms, $\mathbb{E}_P\left|\tilde{f}_r(X) - f(X)\right| \cdot \mathbb{1}\mathcal{E}_r(X)$ is at most

$$\begin{aligned}&\bar{F} \cdot \|e_{N,r}\|_{2,P} + 2\Big(\|e_{n,r}\|_{2,P} \cdot \sqrt{\|f + \epsilon_f(\cdot, r)\|_{1,P}} \\ &\qquad + \|e_{n,r}\|_{2,P}^2 + \epsilon_{P,f}(r)\Big) \\ &\leq \bar{F} \cdot \|e_{N,r}\|_{2,P} + 2\Big(\|e_{n,r}\|_{2,P}\left(1 + \sqrt{\epsilon_{P,f}(r)}\right) \\ &\qquad + \|e_{n,r}\|_{2,P}^2 + \epsilon_{P,f}(r)\Big)\end{aligned}$$

Both $\|e_{N,r}\|_{2,P}$ and $\|e_{n,r}\|_{2,P}$ are bounded via Lemma 3 to obtain that $\mathbb{E}_P\left|\tilde{f}_r(X) - f(X)\right| \cdot \mathbb{1}\mathcal{E}_r(x)$ is at most, for some C depending on $\mathcal{X}$,

$$\begin{aligned}&C \cdot \bar{F} \cdot \sqrt{\frac{c_{N,\mathcal{B}}}{N \cdot r^d}} + 2\Big(C\sqrt{\frac{3c_{n,\mathcal{B}}}{n \cdot r^d}}\left(1 + \sqrt{\epsilon_{P,f}(r)}\right) \\ &\qquad + C\frac{3c_{n,\mathcal{B}}}{n \cdot r^d} + \epsilon_{P,f}(r)\Big).\end{aligned}$$

Now, to handle the case of $\mathcal{E}_r^{\mathsf{C}}(X)$ we further consider the event $\mathcal{E}(\mathbf{X}_P)$ that, for all balls $B \in \mathcal{B}$, $P(B) \leq P_N(B) + \sqrt{P_N(B)\alpha_{N,\mathcal{B}}} + \alpha_{N,\mathcal{B}}$. Under the event $\mathcal{E}(\mathbf{X}_P)$, for any X such that $\mathcal{E}_r^{\mathsf{C}}(X)$, we have $P(B(X,r)) \leq 82\alpha_{m,\mathcal{B}}$. Similar to the proof of Lemma 3, consider an $(r/2)$-cover $\mathcal{X}_r$ of $\mathcal{X}$. For any $x \in \mathcal{X}_r$, and $X \in B(x,r/2)$, clearly $P(B(x,r/2)) \leq P(B(X,r))$.

The integral $\mathbb{E}_P\left|\tilde{f}_r(X) - f(X)\right| \cdot \mathbb{1}\mathcal{E}_r^{\mathsf{C}}(X)$ can then be upper-bounded via Cauchy-Schwartz (first inequality), and Lemma 5 (third inequality), and the above facts on $\mathcal{X}_r$ (last inequality):

$$\begin{aligned}&\mathbb{E}_P f(X) \cdot \mathbb{1}\mathcal{E}_r^{\mathsf{C}}(X) \leq \bar{F} \cdot \mathbb{E}_P \mathbb{1}\mathcal{E}_r^{\mathsf{C}}(X) \\ &\quad \leq \bar{F} \cdot \left(\mathbb{E}_P \mathbb{1}[\mathcal{E}_r^{\mathsf{C}}(X), \mathcal{E}(\mathbf{X}_P)] + \mathbb{E}_P \mathbb{1}\mathcal{E}^{\mathsf{C}}(\mathbf{X}_P)\right) \\ &\quad \leq \bar{F} \cdot \sum_{x \in \mathcal{X}_r} \int_{B(x,r/2)} \mathbb{1}[\mathcal{E}_r^{\mathsf{C}}(x'), \mathcal{E}(\mathbf{X}_P)]\, dP(x') + \bar{F} \cdot \delta \\ &\quad \leq \bar{F} \cdot \left(\sum_{x \in \mathcal{X}_r} 82\alpha_{m,\mathcal{B}}\right) + \bar{F}\delta \leq C\bar{F} \cdot r^{-d}\alpha_{m,\mathcal{B}} + \bar{F}\delta,\end{aligned} \tag{6}$$

for some C depending on $\mathcal{X}$. □

C DATA-DRIVEN CHOICE OF r

In this section gives the proofs of the additional supporting results for Theorem 2.

First, note that, since the base P is a probability measure, the upper-bound F on f is greater than 1 ($1 \leq \int f\, dP \leq F$).

Proof of Proposition 3. Let $m = n \wedge N$. By definition of the tresholded estimate $\tilde{f}_r(x)$, for every x we only need to consider r such that $P_N(B(x,r)) \geq 72\alpha_{m,\mathcal{B}}$, in which case we also have $P(B(x,r)) \geq 24\alpha_{m,\mathcal{B}}$ with probability at least $1-\delta$ (by Corollary 2). We then have by Lemma 1 that w.p. at least $1-2\delta$, $f_r(x) \leq 3F$ (since $f(x) + \hat{\epsilon}_f(x,r) \leq F$). By Lemma 2 (under the same events), we have $\tilde{f}_r(x) \leq 5f_r(x)$. □

Proof of Lemma 4. By Corollary 2 and Assumption 2, with probability at least $1-\delta$, $P_N(B(x,r)) \leq 3P(B(x,r)) \leq 3C_0r^d$ for all x s.t. $\mathcal{E}_r(x)$. It follows that $\|\hat{e}_{m,r}\|_{1,P_M} \geq \min\left\{\sqrt{1/3}, \sqrt{8\alpha_{m,\mathcal{B}}/C_0r^d}\right\} \doteq \sigma_\flat(r)$ since $\hat{e}_{m,r}$ is so bounded pointwise.

For the upper-bound, let $\tilde{e}_{m,r}(x) = \sqrt{72\alpha_{m,\mathcal{B}}/P(B(x,r))}$ if $\mathcal{E}_r(x)$, and $\tilde{e}_{m,r}(x) = \sqrt{1/3}$ otherwise. By Lemma 5, w.p. at least $1-\delta$ (under the same event as above), $P(B(x,r)) \leq 3P_N(B(x,r))$ if $\mathcal{E}_r(x)$; we therefore have $\|\hat{e}_{m,r}\|_{1,P_M} \leq \|\tilde{e}_{m,r}\|_{1,P_M}$. Furthermore, by Corollary 2, $P(B(x,r)) \geq 24\alpha_{m,\mathcal{B}}$ if $\mathcal{E}_r(x)$ so $\tilde{e}_{m,r}(x) \leq \sqrt{3}$. Thus with probability at least $1-2\delta$, by a Hoeffding bound,

$$\|\tilde{e}_{m,r}\|_{1,P_M} \leq \|\tilde{e}_{m,r}\|_{1,P} + \sqrt{3\log(2/\delta)/M}.$$

Finally we relate $\|\tilde{e}_{m,r}\|_{1,P}$ to $\sigma_\sharp(r)$ as follows. By Lemma 5, w.p. at least $1-\delta$ (under the same events as above), $P(B(x,r)) \leq 82\alpha_{m,\mathcal{B}}$ for all x s.t. $\mathcal{E}_r^{\complement}(x)$. Therefore, $\tilde{e}_{m,r}(x) \leq \sqrt{72\alpha_{m,\mathcal{B}}/P(B(x,r))}$ for all x. Combining with Jensen's followed by Lemma 3, we have

$$\begin{aligned}\|\tilde{e}_{m,r}\|_{1,P} &\leq \sqrt{72\alpha_{m,\mathcal{B}}} \cdot \sqrt{\mathbb{E}_P \frac{1}{P(B(X,r))}} \\ &\leq \sqrt{72\alpha_{m,\mathcal{B}}} \cdot \sqrt{Cr^d},\end{aligned}$$

Combine with the above and conclude. □

We are now ready to prove the main result on stability-based choice of bandwidth r.

Proof of Theorem 2. Fix $\mathbf{X}_P$ and $\mathbf{X}_Q$. We first consider only those $x \in \mathcal{X}$ for which $\mathcal{E}_r(x)$ holds. We start with a few simple concentration statements.

Define $\Delta_{P,r}(g,g') \doteq \|(g-g') \cdot \mathbb{1}\mathcal{E}_r\|_{1,P}$. Assume the conclusion of Proposition 3, so that the quantity $\left|\tilde{f}_{r_i} - f\right| \cdot \mathbb{1}\mathcal{E}_{r_j}, i,j \in [k]$ is appropriately bounded by $15F$. By Hoeffding, followed by a union bound on all pairs $i \in [k]$, we have w.p. at least $1-\delta$ over the choice of $\mathbf{X}'_P$ that

$$\sup_{i,j\in[k]} \left|\Delta_{M,r_j}\left(\tilde{f}_{r_i}, f\right) - \Delta_{P,r_j}\left(\tilde{f}_{r_i}, f\right)\right| \leq \epsilon_{M,\delta}. \tag{7}$$

By Corollary 2, w.p. at least $1-\delta$, for any $r > 0$, for all x such that $\mathcal{E}_r(x)$ holds (this fact will be used repeatedly)

$$P(B(x,r)) \geq P_N(B(x,r))/3. \tag{8}$$

Let $e_{m,r}(x) = \sqrt{\frac{24\alpha_{m,\mathcal{B}}}{P(B(x,r))}}$. By (8), w.p. at least $1-\delta$, under $\mathcal{E}_r(x)$, $P(B(x,r)) \geq 24\alpha_{m,\mathcal{B}}$ so $e_{m,r}(x)\mathbb{1}\mathcal{E}_r(x) \leq 1$. Hence, by a Hoeffding bound we have w.p. at least $1-\delta$ over $\mathbf{X}'_P$ that

$$\left|\|e_{m,r} \cdot \mathbb{1}\mathcal{E}_r\|_{1,P} - \|e_{m,r} \cdot \mathbb{1}\mathcal{E}_r\|_{1,P_M}\right| \leq \sqrt{\frac{\log(2/\delta)}{2M}}. \tag{9}$$

By equation (5) (in proving Theorem 1), we have w.p. at least $1-2\delta$ that when $\mathcal{E}_r(x)$ holds

$$\begin{aligned}\left|\hat{f}_r(x) - f(x)\right| &\leq 2Fe_{m,r}(x) + 2e_{m,r}^2(x) + 2\epsilon_f(x,r) \\ &\leq 4Fe_{m,r}(x) + 2\epsilon_f(x,r). \end{aligned} \tag{10}$$

Combining (9) and (10), we have w.p. at least $1-3\delta$

$$\begin{aligned}\|(\tilde{f}_r - f) \cdot \mathbb{1}\mathcal{E}_r\|_{1,P} &\leq 4F\|e_{m,r} \cdot \mathbb{1}\mathcal{E}_r\|_{1,P} + 2\epsilon_{P,f}(r) \\ &\leq 4F\|e_{m,r} \cdot \mathbb{1}\mathcal{E}_r\|_{1,P_M} + 2\lambda r^\beta + \epsilon_{M,\delta} \\ &\leq 7F\|\hat{e}_{m,r}\|_{1,P_M} + 2\lambda r^\beta + \epsilon_{M,\delta}\end{aligned} \tag{11}$$

where the last inequality is again obtained by (8).

We next compare $\|\hat{e}_{m,r}\|_{1,P_M}$ and $2\lambda r^\beta$ (as functions of r). From Lemma 4, $\sigma_\flat(r) \leq \|\hat{e}_{m,r}\|_{1,P_M}$ w.p. at least $1-2\delta$. For m sufficiently large, $\sigma_\flat(r_k) = \sqrt{1/3} \geq 2\lambda(1/m)^\beta \geq 2\lambda r_k^\beta$. Thus for large enough m, $\|\hat{e}_{m,r}\|_{1,P_M}$ is non-increasing in r, larger than $2\lambda r^\beta$ at r_k, and lower than $2\lambda r^\beta$ at $r_0 = 1$ (since $\|\hat{e}_{m,r_0}\|_{1,P_M} = \sqrt{24\alpha_{m,\mathcal{B}}} < \lambda$).

Therefore let $\hat{r}$ denote the largest $r \in R$ such that $\|\hat{e}_{m,r}\|_{1,P_M} \geq 2\lambda r^\beta$. We have $r_k \leq \hat{r} < r_0$. By definition, for any $r \leq \hat{r}$ we have by (11) and (7) that w.p. at least $1-5\delta$

$$\begin{aligned}\Delta_{M,r}\left(\tilde{f}_r, f\right) &\doteq \|(\tilde{f}_r - f) \cdot \mathbb{1}\mathcal{E}_r\|_{1,P_M} \\ &\leq 8F\|\hat{e}_{m,r}\|_{1,P_M} + 2\epsilon_{M,\delta} = \gamma(r).\end{aligned}$$

It follows that, for any $r_i < r_j \leq \hat{r} \in R$, we have

$$\begin{aligned}\Delta_{M,r_i}\left(\tilde{f}_{r_i}, \tilde{f}_{r_j}\right) &\leq \Delta_{M,r_i}\left(\tilde{f}_{r_i}, f\right) + \Delta_{M,r_i}\left(\tilde{f}_{r_j}, f\right) \\ &\leq \Delta_{M,r_i}\left(\tilde{f}_{r_i}, f\right) + \Delta_{M,r_j}\left(\tilde{f}_{r_j}, f\right) \\ &\leq \gamma(r_i) + \gamma(r_j).\end{aligned}$$

In other words, let r be returned by the procedure, we necessarily have $r \geq \hat{r}$. Since the return condition did not hold at r, we must have $\Delta_{M,\hat{r}}\left(\tilde{f}_r, \tilde{f}_{\hat{r}}\right) \leq \gamma(r) + \gamma(\hat{r}) \leq 2\gamma(\hat{r})$. Hence, by (7) and another triangle inequality we have

$$\begin{aligned}\Delta_{P,\hat{r}}\left(\tilde{f}_r, f\right) &\leq \Delta_{M,\hat{r}}\left(\tilde{f}_r, \tilde{f}_{\hat{r}}\right) + \Delta_{M,\hat{r}}\left(\tilde{f}_{\hat{r}}, f\right) + \epsilon_{M,\delta} \\ &\leq 3\gamma(\hat{r}) + \epsilon_{M,\delta} \\ &= 24F\|\hat{e}_{m,\hat{r}}\|_{1,P_M} + 7\epsilon_{M,\delta} \\ &\leq 24F\sigma_\sharp(\hat{r}) + 31\epsilon_{M,\delta},\end{aligned}$$

where $\sigma_\sharp(r)$ is as defined in Lemma 4. Using (6) (see Theorem 1), we therefore have w.p. at least $1 - 5\delta$

$$\begin{aligned} \|\tilde{f}_r - f\|_{1,P} &= \Delta_{P,\hat{r}}\left(\tilde{f}_r, f\right) + \|(\tilde{f}_r - f)\mathbb{1}\mathcal{E}_{\hat{r}}^{\mathsf{C}}\|_{1,P} \\ &\leq \Delta_{P,\hat{r}}\left(\tilde{f}_r, f\right) + 15CF_\infty \cdot \hat{r}^{-d}\alpha_{m,\mathcal{B}} + \bar{F}\delta \\ &\leq C_1(\sigma_\sharp(\hat{r}) + \sigma_\sharp^2(\hat{r})) + 31\epsilon_{M,\delta} + \bar{F}\delta, \end{aligned} \tag{12}$$

for some $C_1 = C_1(F_\infty, \mathcal{X})$.

Finally we bound $\sigma_\sharp(\hat{r})$ as follows. Let $\tilde{r}$ satisfy $\sigma_\flat(\tilde{r}) = 2\lambda\tilde{r}^\beta$. Now suppose $2\hat{r} < \tilde{r}$. Then

$$\|\hat{e}_{m,2\hat{r}}\|_{1,P_M} \geq \sigma_\flat(2\hat{r}) \geq \sigma_\flat(\tilde{r}) = 2\lambda\tilde{r}^\beta > 2\lambda(2\hat{r})^\beta,$$

which is not possible by definition of $\hat{r}$. Thus $2\hat{r} \geq \tilde{r}$, and it follows that $\sigma_\sharp(\hat{r}) \leq \sigma_\sharp(\tilde{r}/2)$.

To get an exact form for $\tilde{r}$, consider $r_0 = (24\alpha_{m,\mathcal{B}}/C_0)^{1/d}$, the largest value or r so that $\sigma_\flat(r) = \sqrt{1/3}$. We have $\sigma_\flat(r_0) = \sqrt{1/3} > 2\lambda r_0^\beta$ for m sufficiently large. Thus $\tilde{r} > r_0$ (otherwise $\sigma_\flat(\tilde{r}) > 2\lambda\tilde{r}^\beta$). It follows, by definition of r_0, that $\tilde{r} = (2\alpha_{m,\mathcal{B}}/C_0\lambda^2)^{1/(2\beta+d)}$. Thus $\sigma_\sharp(\hat{r}) \leq C_2\lambda^{d/(2\beta+d)}\alpha_{m,\mathcal{B}}^{\beta/(2\beta+d)}$, which is at most 1 for large m. Now, combine with (12) and conclude. □

C.1 Data-driven Upper-Bound on $\sup f$

The upper-bound of Proposition 2 is obtained as follows. We denote the weak law of large numbers by LLN in what follows.

Proof of Proposition 2. Let's denote $f(x_0)$ by f_0, and let m sufficiently large so we can pick $2r_0 \in R$ such that $\epsilon_f(x_0, 2r_0) \leq f_0/6$. Also, by LLN, for m sufficiently large, we have with probability at least $1 - \delta$, that $B(x_0, r_0) \cap \mathbf{X}_P$ contains a sample x. It is clear that we then have $B(x_0, r_0) \subset B(x, 2r_0)$.

Let $\mathcal{E}_r(x)$, as defined in Procedure 2, be the event that $P_N(B(x,r)) \geq 72\alpha_{m,\mathcal{B}}$; then, again by LLN, we have for large enough m than $\mathcal{E}_{r_0}(x_0)$ holds w.p. at least $1 - \delta$, and so $\mathcal{E}_{2r_0}(x)$ also holds. Therefore, by Lemma 2, $\tilde{f}_{2r_0}(x) \geq f_{2r_0}(x)/2$ for large m. Now, by Lemma 1, we have for such an x that $f_{2r_0}(x) \geq f(x) - \epsilon_f(x, 2r_0) - f_0/6 \geq 5f_0/6 - \epsilon_f(x, 2r_0) - f_0/6$. Since f is assumed continuous in a neighborhood of x_0, we can pick r_0 small enough so that $\epsilon_f(x, 2r_0) < f_0/6$ and we then get $f_{2r_0}(x) \geq f_0/2$ so that $4\tilde{f}_{2r_0}(x) \geq f_0$. □

D CROSS-VALIDATION APPROACH

The main problem in finding a good parameter r is that the objective $\|\tilde{f}_r - f\|_{1,P}$ is in terms of the unknown f. The approach considered here, also used in [20, 2, 13, 14], relies on the following insights (also described in [25] for the case of density-estimation).

If f is upper-bounded, $L_{1,P}$ rates and $L_{2,P}$ rates differ in these settings only by constants. It is therefore reasonable to choose r to minimize $L_{2,P}$: for any value of r,

$$\begin{aligned} \|\tilde{f}_r - f\|_{2,P}^2 &= \int f^2\, dP + \int \tilde{f}_r^2\, dP - 2\int \tilde{f}_r \cdot f\, dP \\ &= \|f\|_{2,P}^2 + \|\tilde{f}_r\|_{2,P}^2 - 2\|\tilde{f}_r\|_{1,Q}, \end{aligned} \tag{13}$$

so we only need to minimize $\|\tilde{f}_r\|_{2,P}^2 - 2\|\tilde{f}_r\|_{1,Q}$, which we might approximate from samples from P and Q. This is formalized in Procedure 1 below.

Procedure 2 (Cross-validation):
SETUP: Let $R \doteq \left\{r_i \doteq 2^{-i}\right\}_{i=0}^k$, for some integer k, denote values of r.

Let $\mathbf{X}'_P \sim P^M$ and $\mathbf{X}'_Q \sim Q^M$ denote validation samples of size M independent of $\mathbf{X}_P$ and $\mathbf{X}_Q$. Let P_M and Q_M denote resp. the empirical distributions w.r.t. $\mathbf{X}'_P$ and $\mathbf{X}'_Q$.

PROCEDURE: Return $r = \arg\min_{r\in R} \|\tilde{f}_r\|_{2,P_M}^2 - 2\|\tilde{f}_r\|_{1,Q_M}$.

From the above intuition, this approach should yield nearly optimal rates provided the empirical norms concentrate around their expectations, i.e., the corresponding norms under P and Q. However, there is a difficulty: usual concentration inequalities do not apply directly, e.g., Chernoff bounds require the integrands ($\tilde{f}_r^2(x)$ and $\tilde{f}_r(x)$) be bounded, however direct bounds on $\tilde{f}_r(x)$ are $O(m)$ (since truncation only happens when $P(B(x,r)) \lesssim O(1/m)$, $m = n \wedge N$). In other words, we will get a discrepancy of $O(m^2/\sqrt{M})$ between $\|\tilde{f}_r\|_{2,P_M}^2$ and $\|\tilde{f}_r\|_{2,P}^2$, which is vacuous unless $M = \Omega(m^4)$, i.e., unless the validation sets are unreasonably large. It is desirable that the validation sets be of the same size as the training, i.e. $M = O(m)$.

Fortunately, if f is upper-bounded by some F, we can show that, with high probability, the integrand $\tilde{f}_r(x)$ are bounded by a multiple of F (Proposition 3). Combining with some of the intermediary results used to establish Theorem 1, we get the following result for the cross-validation choice of r.

Theorem 3. *Let the derivative f satisfy Assumption 1, for some $\lambda, \beta > 0$, and assume $\sup_x f(x) \leq F < \infty$. Let $m = n \wedge N$ and let $0 < \delta < 1$. There exist $m_1 = m_1(\mathcal{X}, f)$, $C_1 = C_1(\mathcal{X}, f)$ such that the following holds with probability at least $1 - 5\delta$ over the choice of $\mathbf{X}_P, \mathbf{X}_Q$ and $\mathbf{X}'_P$.*

Choose $k = \lceil \log m \rceil$, and suppose $m > m_1$. Let r be the *value returned by Procedure 2. We have*

$$\begin{aligned}\|\tilde{f}_r - f\|_{1,P} \le{}& C_1\lambda^{d/(2\beta+d)}\left(\frac{\log(m^{V_\mathcal{B}}/\delta)}{m}\right)^{\beta/(2\beta+d)}\\ &+ 30F\left(\frac{\log(2k/\delta)}{2M}\right)^{1/4} + F\delta.\end{aligned}$$

Remark 4. *The second term (of order $M^{-1/4}$) on the r.h.s. above dominates whenever $d < 2\beta$, and $M = \Theta(m)$. This seems unavoidable (for this cross-validation approach) since the second term tightly captures the concentration of the empirical norms $\|\tilde{f}_r\|_{2,P_M}^2$ and $\|\tilde{f}_r\|_{1,Q_M}$. However this is not a problem for sufficiently large d, and the cross-validation approach works quite well in practice[2] (see e.g. [20, 2]).*

We first need to establish an intermediary (oracle) $L_{2,P}$ bound, under the additional assumption that f is bounded (this was not needed for the $L_{1,P}$ bound of Theorem 1).

Theorem 4. *Suppose f is Hölder, i.e., $\sup_x \epsilon_f(x, r) \le \lambda r^\beta$ for some $\lambda, \beta > 0$, and furthermore that $\sup_x f(x) = F < \infty$. Let $0 < \delta < 1$. Let $m = n \wedge N$. There exists C depending on $\mathcal{X}$ such that, for m sufficiently large, we have w.p. at least $1 - 2\delta$, that for any $r > 0$,*

$$\|\tilde{f} - r\|_{2,P}^2 \le CF^2\sqrt{\frac{3c_{m,\mathcal{B}}}{m \cdot r^d}} + 8\lambda^2 r^{2\beta} + F^2\delta.$$

It follows that for some C_0, C depending on $\mathcal{X}$ and F, picking $r = C_0\left(\log(m^{V_\mathcal{B}}/\delta)/(\lambda^2 m)\right)^{\beta/(2\beta+d)}$,

$$\begin{aligned}\|\tilde{f}_r - f\|_{2,P}^2 \le{}& C\lambda^{2d/(2\beta+d)} \cdot \left(\frac{\log(m^{V_\mathcal{B}}/\delta)}{m}\right)^{2\beta/(2\beta+d)}\\ &+ F^2\delta.\end{aligned}$$

Proof. The arguments, similar to that for Theorem 1, also build on Lemmas 1 and 2. We will therefore refer back to some of the earlier results of Theorem 1.

Define $e_{m,r}(x) = \sqrt{\frac{24c_{m,\mathcal{B}}}{m \cdot P(B(x,r))}}$. By Corollary 2, we have w.p. at least $1 - \delta$ that under event $\mathcal{E}_r(x)$, $P(B(x, r)) \ge 24\alpha_{m,\mathcal{B}}$, implying $e_{m,r}(x) \le 1$.

By (5), for m sufficiently large, we have w.p. at least $1 - 2\delta$ that, $\forall x \in \mathcal{X}$ s.t. $\mathcal{E}_r(x)$ holds,

$$\begin{aligned}\left|\hat{f}_r(x) - f(x)\right|^2 &\le (5e_{m,r}(x) \cdot F + 2\epsilon_f(x, r))^2\\ &\le 50F^2 \cdot e_{m,r}^2(x) + 8\epsilon_f(x, r)^2,\\ &\le 50F^2 \cdot e_{m,r}^2(x) + 8\lambda^2 r^{2\beta}\end{aligned}$$

therefore, bounding $\|e_{m,r}\|_{2,P}$ via Lemma 3, we have

$$\begin{aligned}\mathbb{E}_P\left|\tilde{f}_r(X) - f(X)\right|^2 \cdot \mathbb{1}\mathcal{E}_r(X) &\le 50F^2\|e_{m,r}(x)\|_{2,P}^2\\ &\quad + 8\lambda^2 r^{2\beta}\\ &\le CF^2\sqrt{\frac{3c_{m,\mathcal{B}}}{m \cdot r^d}} + 8\lambda^2 r^{2\beta}, \quad (14)\end{aligned}$$

for some C depending on $\mathcal{X}$.

Now, the case of $\mathcal{E}_r^{\complement}(X)$ is handled using the same arguments as for (6). We have that the integral $\mathbb{E}_P\left|\tilde{f}_r(X) - f(X)\right|^2 \cdot \mathbb{1}\mathcal{E}_r^{\complement}(X)$ can be upper-bounded via Lemma 5 (third inequality), and the above facts on $\mathcal{X}_r$ (last inequality):

$$\begin{aligned}&\mathbb{E}_P f^2(X) \cdot \mathbb{1}\mathcal{E}_r^{\complement}(X) \le F^2 \cdot \mathbb{E}_P\mathbb{1}\mathcal{E}_r^{\complement}(X)\\ &\quad\le F^2 \cdot \left(\mathbb{E}_P\mathbb{1}\mathcal{E}_r^{\complement}(X), \mathcal{E}(\mathbf{X}_P) + \mathbb{E}_P\mathbb{1}\mathcal{E}^{\complement}(\mathbf{X}_P)\right)\\ &\quad\le F^2\left(\sum_{x\in\mathcal{X}_r}\int_{B(x,r/2)}\mathbb{1}\mathcal{E}_r^{\complement}(x'), \mathcal{E}(\mathbf{X}_P)\,dP(x') + \delta\right)\\ &\quad\le F^2\left(\left(\sum_{x\in\mathcal{X}_r} 82\alpha_{m,\mathcal{B}}\right) + \delta\right) \qquad (15)\\ &\quad\le CF^2 \cdot r^{-d}\alpha_{m,\mathcal{B}} + F^2\delta, \qquad (16)\end{aligned}$$

for some C depending on $\mathcal{X}$. Now combine (14) and (16) to conclude. □

We are now ready to prove the main result on the cross-validation choice of bandwidth r.

Proof of Theorem 3. By Proposition 3, w.p. at least $1 - 2\delta$ we have for all $r \in R$ that $\sup_x \tilde{f}(x) \le 15F$. It follows by Chernoff bounds and a union bound over R that, w.p. at least $1 - 3\delta$,

$$\begin{aligned}&\left|\left(\|\tilde{f}_r\|_{2,P_M}^2 - 2\|\tilde{f}_r\|_{1,Q_M}\right) - \left(\|\tilde{f}_r\|_{2,P}^2 - 2\|\tilde{f}_r\|_{1,Q}\right)\right|\\ &\le 450F^2\sqrt{\frac{\log(2k/\delta)}{2M}}.\end{aligned}$$

It follows by (13) that, for the value of $r \in R$ returned by the procedure, we have under the same event

$$\begin{aligned}\|\tilde{f}_r - f\|_{1,P}^2 &\le \|\tilde{f}_r - f\|_{2,P}^2\\ &\le \inf_{r'\in R}\|\tilde{f}_{r'} - f\|_{2,P}^2 + 900F^2\sqrt{\frac{\log(2k/\delta)}{2M}}.\end{aligned}$$

Now, by Theorem 4, we can pick $r_0 \in R$, $r_0 = C_0\left(\log(m^{V_\mathcal{B}}/\delta)/(\lambda^2 m)\right)^{\beta/(2\beta+d)}$, such that w.p. $\ge 1 - 2\delta$

$$\begin{aligned}\|\tilde{f}_{r_0} - f\|_{2,P}^2 \le{}& C\lambda^{2d/(2\beta+d)} \cdot \left(\frac{\log(m^{V_\mathcal{B}}/\delta)}{m}\right)^{2\beta/(2\beta+d)}\\ &+ F^2\delta.\end{aligned}$$

Combine this last inequality with the previous one and conclude. □

[2] The cross-validation approach is applied to different estimators in the cited work. The main arguments of our analysis easily extend to any *bounded* estimator with bounded $L_{2,P}$ risk; the second term remains.