A new similarity measure for covariate shift with applications to nonparametric regression

Reese Pathak[†, ⋆], Cong Ma[◇, ⋆], and Martin J. Wainwright[†, ‡]

[◇] Department of Statistics, University of Chicago
[†] Department of Electrical Engineering and Computer Sciences, UC Berkeley
[‡] Department of Statistics, UC Berkeley
congm@uchicago.edu, {pathakr,wainwrig}@berkeley.edu

Abstract

We study covariate shift in the context of nonparametric regression. We introduce a new measure of distribution mismatch between the source and target distributions that is based on the integrated ratio of probabilities of balls at a given radius. We use the scaling of this measure with respect to the radius to characterize the minimax rate of estimation over a family of Hölder continuous functions under covariate shift. In comparison to the recently proposed notion of transfer exponent, this measure leads to a sharper rate of convergence and is more fine-grained. We accompany our theory with concrete instances of covariate shift that illustrate this sharp difference.

1 Introduction

In the standard formulation of prediction or classification, future data (as represented by a test set) is assumed to be drawn from the same distribution as the training data. This assumption, while theoretically convenient, may fail to hold in many real-world scenarios. For instance, training data might be collected only from a sub-group within a broader population (such as in medical trials), or the environment might change over time as data are collected. Such scenarios result in a distribution mismatch between the training and test data.

In this paper, we study an important case of such distribution mismatch—namely, the covariate shift problem (e.g., [21, 19]). Suppose that a statistician observes covariate-response pairs (X, Y), and wishes to build a prediction rule. In the problem of covariate shift, the distribution of the covariates X is allowed to change between the training and test data, while the posterior distribution of the responses (namely, $Y \mid X$) remains fixed. Compared to the usual i.i.d. setting, this serves as a more accurate model for a variety of real-world applications, including image classification [20], biomedical engineering [13], sentiment analysis [3], and audio processing [8], among many others.

More formally, suppose that the statistician observes n_P covariates $\{X_i\}_{i=1}^{n_P}$ from a *source distribution* P, and n_Q covariates $\{X_i\}_{i=n_P+1}^{n_Q+n_P}$ from a *target distribution* Q. For each observed X_i, she also observes a response Y_i drawn from the same conditional distribution. The *regression function* $f^\star(x) = \mathbf{E}[Y \mid x]$ defined by this conditional distribution is assumed to lie in some function class $\mathcal{F}$. The statistician uses these samples to produce an estimate $\widehat{f}$, which will be evaluated on the target distribution, with a fresh sample $X \sim Q$, yielding the mean-squared error

$$\|\widehat{f} - f^\star\|_{L^2(Q)}^2 := \mathbf{E}\left[\left(\widehat{f}(X) - f^\star(X)\right)^2\right].$$

When there is no covariate shift, the fundamental (minimax) risks for this problem are well-understood [7, 9, 22]. The goal of this paper is to understand how, for nonparametric function classes $\mathcal{F}$, this minimax risk changes as a function of the "amount" of covariate shift between P and Q.

[⋆] RP and CM contributed equally to this paper.

1.1 Our contributions and related work

Let us summarize the main contributions of this paper, and put them in the context of related work.

Our contributions. We introduce a similarity measure[1] ρ_h between two probability measures P, Q on a common metric space $(\mathcal{X}, d)$. For any level $h > 0$, it is defined as

$$\rho_h(P, Q) := \int_{\mathcal{X}} \frac{1}{P\big(\mathsf{B}(x, h)\big)} \, \mathrm{d}Q(x), \tag{1}$$

where $\mathsf{B}(x, h) := \{\, x' \in \mathcal{X} \mid d(x, x') \leqslant h \,\}$ is the closed ball of radius h centered around x. We substantiate the significance of this similarity measure via the following contributions:

(i) For regression functions that are Hölder continuous, we demonstrate a performance guarantee for the Nadaraya-Watson kernel estimator under covariate shift that is fully determined by the scaling of the similarity measure $\rho_h(P, Q)$ with respect to the radius h.

(ii) We complement these upper bounds with matching lower bounds—in a minimax sense—demonstrating that the best achievable rate of estimation in Hölder classes is also determined by the scaling of this similarity measure.

(iii) We show how the similarity measure ρ_h can be controlled based on the metric properties of the space $\mathcal{X}$. In addition, we compare ρ_h with existing notions for covariate shift (e.g., bounded likelihood ratios, transfer exponents), thereby showcasing some of its advantages.

Related work. The problem of covariate shift was studied in the seminal work by Shimodaira [21], who provided asymptotic guarantees for a weighted maximum likelihood estimator under covariate shift. Since then, a plethora of work has analyzed covariate shift, or the general distribution mismatch problem (also referred to as domain adaptation or transfer learning).

For general distribution mismatch, one line of work provides rates that depend on distance metrics between the source-target pair (e.g., [1, 2, 6, 14, 5, 16]). These results hold under fairly general conditions, but do not necessarily guarantee consistency as the sample size n increases. In contrast, our guarantees for covariate shift do guarantee consistency, and moreover, we provide explicit nonasymptotic, optimal nonparametric rates. As pointed out in the paper [11], the distribution mismatch problem is asymmetric in the sense that it may be easier to estimate accurately when dealing with covariate shift from P to Q than from Q to P. Our results also corroborate this intuition. It is worth noting that these prior distance metrics fall short of capturing the inherent asymmetry between P and Q.

Another line of work addresses covariate shift under conditions on the likelihood ratio $\mathrm{d}Q/\mathrm{d}P$. For instance, some authors have obtained results for bounded likelihood ratios [24, 10] or in terms of information-theoretic divergences between the source-target pair [23, 15]. Our work is inspired in part by the work of Kpotufe and Martinet [11], who introduced the notion of the *transfer exponent*. It is a condition that bounds the mass placed by the pair (P, Q) on balls of varying radii; using this notion, they analyzed various problems of nonparametric classification. Our work, focusing instead on nonparametric regression problems and using the measure ρ_h, provides sharper rates than those obtainable by considering the transfer exponent; see Section 3.2 for details. Thus, the similarity measure ρ_h provides a more fine-grained control on the effect of covariate shift on nonparametric regression.

[1] To be clear, this quantity actually serves as a *dis*-similarity measure: as shown in the sequel, source-target pairs (P, Q) with larger values $\rho_h(P, Q)$ lead to "harder" estimation problems in terms of covariate shift.

Finally, it is worth mentioning other recent works that give risk bounds for covariate shift problems, including on linear models [12], as well as linear models and one-layer neural networks [17]. Although these results deal with covariate shift, the rates obtained are parametric ones, and hence not directly comparable to the nonparametric rates that are the focus of our inquiry.

1.2 Notation

Here we collect notation used throughout the paper. We use $\mathbf{R}$ to denote the real numbers. We use $(\mathcal{X}, d)$ to denote a metric space, and we equip it with the usual Borel σ-algebra. We let $\mathsf{B}(x, r) := \{ x' \in \mathcal{X} \mid d(x, x') \leqslant r \}$ be the closed ball of radius r centered at x. We reserve the capital letters X, Y, possibly with subscripts, for a pair of random variables arising from a regression model. Similarly, we reserve P, Q for a pair of two probability measures on $(\mathcal{X}, d)$. For $h > 0$, we denote by $N(h)$ the covering number of $\mathcal{X}$ at resolution h in the metric d. This is the minimal number of balls of radius at most $h > 0$ required to cover the space $\mathcal{X}$.

The remainder of this paper is organized as follows. We begin in Section 2 by setting up the problem more precisely, and stating and discussing our main results on covariate shift: namely, upper bounds in Theorem 1, accompanied by matching lower bounds in Theorem 2. These results establish that the similarity measure (1) provides a useful measure of the "difficulty" of source-target pairs in covariate shift; accordingly, Section 3 is devoted to a comparison and discussion of this measure relevant to concepts from past work, including likelihood ratio bounds and transfer exponents. The proofs of all our results are given in Section 4, and we conclude with a discussion in Section 5.

2 How covariate shift affects nonparametric regression

In this section, we use the similarity measure introduced in equation (1) to characterize how covariate shift can change the minimax risks of estimation for certain classes of nonparametric regression models. We begin in Section 2.1 by setting up the observation model to be considered, along with some associated assumptions on the regression function $f^\star$, the conditional distribution of $Y \mid X$, and the covariate shift. In Section 2.2, we derive an achievable result (Theorem 1) for nonparametric regression in the presence of covariate shift, in particular via a careful analysis of the classical Nadaraya-Watson estimator. Our upper bound in this section is general, and illustrates the key role of the similarity measure ρ_h. In Section 2.3, we introduce the α-families of source-target pairs (P, Q), and use Theorem 1 to derive achievable results for these families. In Section 2.4, we state some complementary lower bounds for α-families (Theorem 2), showing that our achievable results are, in fact, unimprovable.

2.1 Observation model and assumptions

Suppose that we observe covariate-response pairs $\{(X_i, Y_i)\}_{i=1}^n \subset \mathcal{X} \times \mathbf{R}$ that are drawn from nonparametric regression model of the following type. The conditional distribution of $Y \mid X$ is the same for all $i = 1, \ldots, n$, and our goal is to estimate the regression function $f^\star(x) := \mathbf{E}[Y \mid X = x]$. In terms of the "noise" variables, $\xi_i := Y_i - f^\star(X_i)$, the observations can be written in the form

$$Y_i = f^\star(X_i) + \xi_i, \quad i = 1, \ldots, n. \tag{2}$$

In our analysis, we impose three types of regularity conditions: (i) Hölder continuity of the regression function; (ii) the type of covariate shift allowed; and (iii) tail conditions on the noise variables $\{\xi_i\}_{i=1}^n$.

Assumption 1 (Hölder continuity)**.** For some $L > 0$ and $\beta \in (0, 1]$, the function $f^\star \colon \mathcal{X} \to \mathbf{R}$ is (β, L)-Hölder continuous, meaning that

$$\left|f^\star(z) - f^\star(z')\right| \leqslant L\,[d(z, z')]^\beta, \quad \text{for any } z, z' \in \mathcal{X}.$$

We note that in the special case $\beta = 1$, the function $f^\star$ is L-Lipschitz.

Assumption 2 (Covariate shift)**.** The covariates $X_1, \ldots, X_n$ are independent, and drawn as

$$X_1, \ldots, X_{n_P} \overset{\text{i.i.d.}}{\sim} P \quad \text{and} \quad X_{n_P+1}, \ldots X_{n_P+n_Q} \overset{\text{i.i.d.}}{\sim} Q \qquad \text{where } n = n_P + n_Q.$$

Assumption 3 (Noise assumption)**.** The variables $\{\xi_i\}_{i=1}^n$ satisfy the second moment bound

$$\sup_x \mathbf{E}\left[\xi_i^2 \mid X_i = x\right] \leqslant \sigma^2 \qquad \text{for } i = 1, \ldots, n.$$

Note that by construction, the variables ξ_i are (conditionally) centered. Assumption 3 also allows ξ_i to depend on X_i, as long as the variance is uniformly bounded above.

2.2 Achievable performance via the Nadaraya-Watson estimator

We first exhibit an achievable result for the problem of nonparametric regression in the presence of covariate shift. We do so by analyzing a classical and simple method for nonparametric estimation, namely the Nadaraya-Watson estimator [18, 27], or NW for short. The main result of this section is to show that the mean-squared error (MSE) of the NW estimator is upper bounded by a bias-variance decomposition that also involves the similarity measure ρ_h.

We begin by recalling the definition of the NW estimator, focusing here on the version in which the underlying kernel is uniform over a ball of a given bandwidth $h_n > 0$. In particular, define the set

$$\mathcal{G}_n := \bigcup_{i=1}^n \mathsf{B}(X_i, h_n),$$

corresponding to the set of points in $\mathcal{X}$ within distance h_n of the observed covariates. In terms of this set, the *Nadaraya-Watson estimator* $\widehat{f}$ takes the form

$$\widehat{f}(x) := \begin{cases} \dfrac{\sum_{i=1}^n Y_i \mathbf{1}\{X_i \in \mathsf{B}(x, h_n)\}}{\sum_{i=1}^n \mathbf{1}\{X_i \in \mathsf{B}(x, h_n)\}} & \text{for } x \in \mathcal{G}_n \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Our first main result provides an upper bound on the MSE of the NW estimator under covariate shift; this bound exhibits the significance of the similarity measure (1). It involves the distribution $\mu_n := \frac{n_P}{n} P + \frac{n_Q}{n} Q$, which is a convex combination of the source and target distributions weighted by their respective fractions of samples.

Theorem 1. *Suppose that Assumptions 1, 2, and 3 hold. For any $h_n > 0$, the Nadaraya-Watson estimator $\widehat{f}$ with bandwidth h_n has MSE bounded as*

$$\mathbf{E}\left\|\widehat{f} - f^\star\right\|_{L^2(Q)}^2 \leqslant c_u \Big\{ L^2 h_n^{2\beta} + \frac{\|f^\star\|_\infty^2 + \sigma^2}{n} \rho_{h_n}(\mu_n, Q) \Big\}, \tag{4}$$

where $c_u > 0$ is a numerical constant.

See Section 4.1 for a proof of this result.

Note that the bound (4) exhibits a type of bias-variance trade-off, one that controls the optimal choice of bandwidth h_n. The quantity $h_n^{2\beta}$ in the first term is familiar from the classical analysis of the NW estimator; it corresponds to the bias induced by smoothing over balls of radius h_n, and hence is an increasing function of bandwidth. In the second term, the bandwidth appears in the similarity measure $\rho_{h_n}(\mu_n, Q)$, which is a non-increasing function of the bandwidth. The optimal choice of bandwidth arises from optimizing this tradeoff; note that it depends on the pair (P, Q), as well as the sample sizes (n_P, n_Q), via the similarity measure applied to the convex combination μ_n and Q.

No covariate shift: As a sanity check, it is worth checking that the bound (4) recovers known results in the case of no covariate shift ($P = Q$ and hence $\mu_n = Q$). As a concrete example, if Q is uniform on the hypercube $[0,1]^k$, it can be verified that $\rho_h(Q,Q) \asymp h^{-k}$ as $h \to 0^+$. (See Example 2 in the sequel for a more general calculation that implies this fact.) Thus, if we track only the sample size, the optimal bandwidth is given by $h_n^* = n^{-\frac{1}{2\beta+k}}$, and with this choice, the bound (4) implies that the NW estimator has MSE bounded as $n^{-\frac{2\beta}{2\beta+k}}$. Thus, we recover the classical and known results in this special case. As we will see, more interesting tradeoffs arise in the presence of covariate shift, so that $\mu_n \neq Q$.

2.3 Consequences for α-families of source-target pairs

In order to better understand the bias-variance tradeoff in the bound (4) in the presence of covariate shift, it is helpful to derive some explicit consequences of Theorem 1 for a particular function class $\mathcal{F}$, along with certain families of source-target pairs (P, Q). The latter families are indexed by a parameter $\alpha > 0$ that controls the amount of covariate shift; accordingly, we refer to them as α-families.

So as to simplify our presentation, we assume that $\mathcal{X}$ is the unit interval $[0,1]$. For a given pair $\beta \in (0,1]$ and $L > 0$, consider the class of regression functions

$$\mathcal{F}(\beta, L) = \Big\{ f\colon [0,1] \to \mathbf{R} \mid |f(x) - f(x')| \leqslant L|x - x'|^{\beta}, \text{for all } x, x' \in \mathcal{X},\ f(0) = 0 \Big\}.$$

This is a special case of β-Hölder continuous functions when the underlying metric space is the unit interval $[0,1]$ equipped with the absolute value norm. The additional constraint $f(0) = 0$ ensures that this class has finite metric entropy.

Next we introduce some interesting families of source-target pairs.

α-families of (P, Q) pairs: For a given parameter $\alpha \geqslant 1$ and radius $C \geqslant 1$, we define the set of source-target pairs[2]

$$\mathcal{D}(\alpha, C) := \Big\{ (P, Q) \mid \sup_{0<h\leqslant 1} h^{\alpha} \rho_h(P, Q) \leqslant C \Big\}. \tag{5a}$$

In words, these are source target pairs for which the growth of the similarity as $h \to 0^+$ is at most $h^{-\alpha}$. In the case $\alpha \in (0,1]$, we define the related set

$$\mathcal{D}'(\alpha, C) := \Big\{ (P, Q) \mid \sup_{0<h\leqslant \Delta} h^{\alpha} \rho_h(P, Q) \leqslant C,\ \sup_{0<h\leqslant 1} \rho_h(Q, Q) \leqslant C \Big\}, \tag{5b}$$

[2]Note that the restriction of the supremum to $h \in [0,1]$ is necessary, as $\rho_h(P, Q) = 1$ for all $h \geqslant 1$. Note also that since $\rho_1(P, Q) = 1$, one necessarily has $C \geqslant 1$.

where the additional condition is added to address the fact that even without covariate shift, the rate $n^{-2\beta/(2\beta+1)}$ is unimprovable for some distributions [22]. Taking into account the first part of the next corollary, it is necessary to impose some condition on the target distribution in order to obtain significantly faster rates such as $n^{-\frac{2\beta}{2\beta+\alpha}}$, when $\alpha < 1$.

Corollary 1. *Suppose that $\sigma \geqslant L$, and that Assumptions 2 and 3 hold. Then there exists a constant $c'_u > 0$, independent of n, n_P, n_Q, σ^2, and an integer $n_u := n_u(\sigma, \beta, L, \alpha, C)$ such that, provided that $\max\{n_P, n_Q\} \geqslant n_u$:*

(a) *For $\alpha \geqslant 1$ and $C \geqslant 1$, we have*

$$\inf_{\widehat{f}} \sup_{f^\star \in \mathcal{F}(\beta,L)} \mathbf{E}\left\|\widehat{f} - f^\star\right\|^2_{L^2(Q)} \leqslant c'_u \Big\{ \big(\frac{n_P}{\sigma^2}\big)^{\frac{2\beta+1}{2\beta+\alpha}} + \big(\frac{n_Q}{\sigma^2}\big) \Big\}^{-\frac{2\beta}{2\beta+1}} \quad \textit{for any } (P,Q) \in \mathcal{D}(\alpha, C). \tag{6a}$$

(b) *For $\alpha \in (0, 1]$ and $C \geqslant 1$, we have*

$$\inf_{\widehat{f}} \sup_{f^\star \in \mathcal{F}(\beta,L)} \mathbf{E}\left\|\widehat{f} - f^\star\right\|^2_{L^2(Q)} \leqslant c'_u \Big\{ \big(\frac{n_P}{\sigma^2}\big)^{\frac{2\beta}{2\beta+\alpha}} + \big(\frac{n_Q}{\sigma^2}\big) \Big\}^{-1} \quad \textit{for any } (P,Q) \in \mathcal{D}'(\alpha, C). \tag{6b}$$

See Section 4.2 for a proof of this corollary.

Let us discuss the bound (6a) to gain some intuition. The special case of no covariate shift can be captured by setting $n_P = 0$ and $n_Q > 0$, and we recover the familiar $n^{-\frac{2\beta}{2\beta+k}}$ rate previously discussed. At the other extreme, suppose that $n_Q = 0$ so that all of our samples are from the shifted distribution (i.e., $n = n_P$); in this case, the MSE is bounded as $(\sigma^2/n)^{-\frac{2\beta}{2\beta+\alpha}}$. As α increases, our set-up allows for more severe form of covariate shift, and its deleterious effect is witnessed by the exponent $\frac{2\beta}{2\beta+\alpha}$ shrinking towards zero. Thus, the NW estimator—with an appropriate choice of bandwidth—remains consistent but with an arbitrarily slow rate as α diverges to $+\infty$.

There are many papers in the literature (e.g., [24, 10]) that discuss the covariate shift problem when the likelihood ratio is bounded—that is, when Q is absolutely continuous with respect to P and $\sup_{x \in \mathcal{X}} \frac{\mathrm{d}Q}{\mathrm{d}P}(x) \leqslant b$ for some $b \geqslant 1$. We say that the pair (P, Q) are b-bounded in this case.

Example 1 (Bounded likelihood ratio)**.** Suppose that $\mathcal{X} = [0, 1]^k$ with the Euclidean metric, and consider a pair (P, Q) with b-bounded likelihood ratio. In this special case, our general theory yields bounds in terms of the b-weighted *effective sample size*

$$n_{\text{eff}}(b) := \frac{n_P}{b} + n_Q. \tag{7}$$

In particular, it follows from the proof of Corollary 1 that in the regime $\sigma^2 \geqslant L^2$, we have the upper bound

$$\mathbf{E}\left\|\widehat{f} - f^\star\right\|^2_{L^2(Q)} \leqslant c'_u \Big(\frac{\sigma^2}{n_{\text{eff}}(b)} \Big)^{\frac{2\beta}{2\beta+k}},$$

provided that $n_{\text{eff}}(b)$ is large enough. Consequently, the effect of covariate shift with b-bounded pairs is to reduce n_P to n_P/b. Again, we recover the standard rate $(\frac{\sigma^2}{n})^{\frac{2\beta}{2\beta+k}}$ in the case of no covariate shift (or equivalently, when $b = 1$). This recovers a known result and is minimax optimal.

2.4 Matching lower bounds for α-families

Thus far, we have seen that the similarity measure ρ_h plays a central role in determining the estimation error of the NW estimator under covariate shift. However, this is just one of many possible estimators in nonparametric regression. Does this similarity measure play a more fundamental role? In this section, we answer this question in the affirmative by proving minimax lower bounds for covariate shift problems parameterized in terms of bounds on ρ_h. In order to do so, we consider the metric space $\mathcal{X} = [0, 1]$ equipped with the absolute value as the metric.

The main result of this section provides lower bounds on the mean-squared error of any estimator, when measured uniformly over functions in the Hölder class $\mathcal{F}(\beta, L)$, along with target-source pairs (P, Q) belonging to the class $\mathcal{D}(\alpha, C)$ when $\alpha \geqslant 1$ and the class $\mathcal{D}'(\alpha, C)$ when $\alpha < 1$.

Theorem 2. *Suppose that Assumptions 2 and 3 hold. Then there is a constant $c_\ell > 0$, independent of n, n_P, n_Q, σ^2, and an integer $n_\ell := n_\ell(\sigma, L, C, \alpha, \beta)$ such that for all sample sizes $\max\{n_P, n_Q\} \geqslant n_\ell$:*

(a) *For $\alpha > 1$ and $C \geqslant 1$, there is a pair of distributions $(P, Q) \in \mathcal{D}(\alpha, C)$ such that*

$$\inf_{\widehat{f}} \sup_{f^\star \in \mathcal{F}(\beta, L)} \mathbf{E} \left\| \widehat{f} - f^\star \right\|^2_{L^2(Q)} \geqslant c_\ell \Big\{ \big(\frac{n_P}{\sigma^2}\big)^{\frac{2\beta+1}{2\beta+\alpha}} + \big(\frac{n_Q}{\sigma^2}\big) \Big\}^{-\frac{2\beta}{2\beta+1}}. \tag{8a}$$

(b) *For $\alpha \leqslant 1$ and $C \geqslant 1$, there is a pair of distributions $(P, Q) \in \mathcal{D}'(\alpha, C)$ such that*

$$\inf_{\widehat{f}} \sup_{f^\star \in \mathcal{F}(\beta, L)} \mathbf{E} \left\| \widehat{f} - f^\star \right\|^2_{(Q)} \geqslant c_\ell \Big\{ \big(\frac{n_P}{\sigma^2}\big)^{\frac{2\beta}{2\beta+\alpha}} + \big(\frac{n_Q}{\sigma^2}\big) \Big\}^{-1}. \tag{8b}$$

See Sections 4.3 and 4.4 for the proof of this result.

These lower bounds should be compared to Corollary 1. This comparison shows that the MSE bounds achieved by the NW estimator are actually optimal in the minimax sense over families defined by the similarity measure ρ_h.

3 Properties of the similarity measure

In the previous sections, we have seen that the similarity measure ρ_h controls both the behavior of the NW estimator, as well as fundamental (minimax) risks applicable to any estimator. Thus, it is natural to explore the similarity measure in some more detail, and in particular to draw some connections to existing notions in the literature.

3.1 Controlling ρ_h via covering numbers

We start with a general way of controlling the similarity measure ρ_h, which is based on the covering number of the metric space $(\mathcal{X}, d)$. In particular, for any $h > 0$, the *covering number* $N(h)$ is defined to be the smallest number of balls of radius h needed to cover the space $\mathcal{X}$. See Chapter 5 in the book [26] for more background.

Proposition 1 (Covering number bounds for the similarity measure)**.** *Suppose that P, Q are two probability measures on the same metric space $(\mathcal{X}, d)$. Suppose that for some $h > 0$, there is a $\lambda > 0$ such that*

$$P(\mathsf{B}(x, h)) \geqslant \lambda\, Q(\mathsf{B}(x, h)) \quad \textit{for all } x \in \mathcal{X}. \tag{9}$$

Then the similarity at scale h is upper bounded as $\rho_h(P, Q) \leqslant N(\frac{h}{2})/\lambda$.

See Section 4.5 for the proof of this claim.

It is worth emphasizing that—due to the order of quantifiers above—the quantity $\lambda > 0$ is allowed to depend on $h > 0$. We exploit this fact in subsequent uses of the bound (9).

One straightforward application of Proposition 1 is in bounding the similarity measure when there is no covariate shift, as we now discuss.

Example 2 (No covariate shift)**.** Suppose that we compute the similarity measure in the case $P = Q$; intuitively, this models a scenario where there is no covariate shift. In this case, we clearly may apply Proposition 1 with $\lambda = 1$, which reveals that $\rho_h(P, P) \leqslant N(h/2)$. To give one concrete bound, suppose that $\mathcal{X} \subset \mathbf{R}^k$ is a compact set, with diameter D. Then—owing to standard bounds on covering number [26, chap. 5]—we obtain $\rho_h(P, P) \leqslant (1 + \frac{2D}{h})^k$. Note that this bound holds for any metric, so long as the diameter D is computed with the same metric as the balls in the definition of the similarity measure.

We give another application of Proposition 1 in the following subsection.

3.2 Comparison to previous notions of distribution mismatch

Next, we show how the mapping $h \mapsto \rho_h(P, Q)$ can be bounded naturally using previously proposed notions of distribution mismatch for covariate shift. Again, Proposition 1 plays a central role.

Example 3 (Bounded likelihood ratio)**.** Suppose that P, Q are such that $Q \ll P$ and the likelihood ratio $\frac{\mathrm{d}Q}{\mathrm{d}P}(x) \leqslant b$, for all $x \in \mathcal{X}$. Then note that by a simple integration argument $P(\mathsf{B}(x, h)) \geqslant \frac{1}{b} Q(\mathsf{B}(x, h))$. Therefore, we conclude $\rho_h(P, Q) \leqslant bN(h/2)$.

As noted previously, our work was inspired by the transfer exponent introduced by Kpotufe and Martinet [11] in the context of covariate shift for nonparametric regression. It is worth comparing these notions so as to understand in what sense the similarity measure ρ_h is a refinement of the transfer exponent. In order to simplify this discussion, we focus here on the special case $\mathcal{X} = [0, 1]$. We begin by providing the definition of transfer exponent:

Definition 1 (Transfer exponent [11])**.** *The distributions (P, Q) have transfer exponent $\gamma \geqslant 0$ with constant $K \in (0, 1]$ if*

$$P(\mathsf{B}(x, h)) \geqslant K h^\gamma Q(\mathsf{B}(x, h)) \qquad \text{for all } x \text{ in the support of } Q.$$

We denote by $\mathcal{T}(\gamma, K)$ the set of all pairs (P, Q) with this property.

It is natural to ask how the set $\mathcal{T}(\gamma, K)$ is related to the α-family previously defined in equation (5a). The following result establishes an inclusion:

Lemma 1. *For $\mathcal{X} = [0, 1]$ and any $\gamma \geqslant 0$ and $K \in (0, 1]$, we have the inclusion*

$$\mathcal{T}(\gamma, K) \subset \mathcal{D}(\gamma + 1, \tfrac{2}{K}). \tag{10}$$

The proof of this inclusion is given in Section 4.6. At a high level, it exploits Proposition 1 to show that for any $(P, Q) \in \mathcal{T}(\gamma, K)$, we have the bound $\rho_h(P, Q) \leqslant \frac{1}{K h^\gamma} N(h/2)$.

From the inclusion (10), it follows that any covariate shift instance (P, Q) with finite transfer exponent $\gamma \geqslant 0$ belongs to an α-similarity family with $\alpha = \gamma + 1$. In fact, following a proof similar to that of Theorem 2, we can show that for $\gamma \geqslant 0$, there is pair (P, Q) in the class $\mathcal{T}(\gamma, K)$ such that

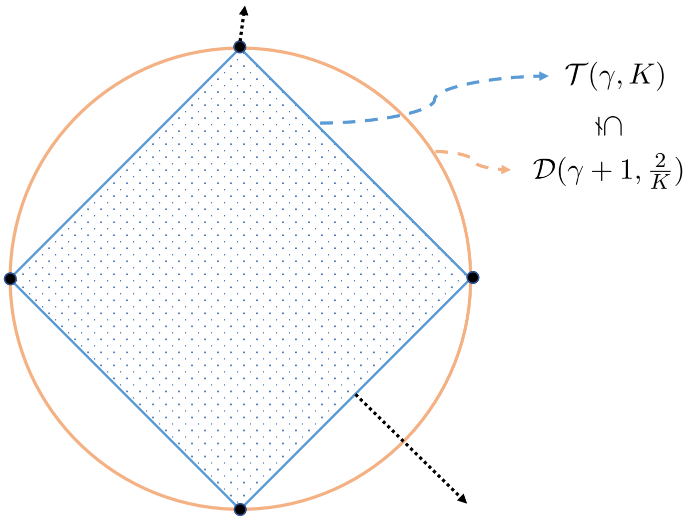

Figure 1. The yellow circle depicts the contour for the class $\mathcal{D}(\gamma+1, \frac{2}{K})$, while the blue square plots the contour for the class $\mathcal{T}(\gamma, K)$. It can be seen from Lemma 1 and Example 5 that $\mathcal{T}(\gamma, K)$ is strict subset of $\mathcal{D}(\gamma+1, \frac{2}{K})$. In addition, our lower bound shows that under covariate shift, the worst-case instances for both classes coincide with each other. However, there exist instances (P, Q) where the characterization using transfer exponent is intrinsically loose.

the minimax risk for β-Hölder-continous functions scales as $n_P^{-\frac{2\beta}{2\beta+\gamma+1}}$. Note that this risk bound coincides with the minimax risk associated with the class $\mathcal{D}(\gamma+1, \frac{2}{K})$. In other words, from a *worst case* point of view, the source-target class $\mathcal{T}(\gamma, K)$ is equally as hard as the class $\mathcal{D}(\gamma+1, \frac{2}{K})$ for nonparametric regression under covariate shift. However, this worst case equivalence does not capture the full picture: there are many covariate shift families for which the transfer exponent provides an overly conservative prediction, and so does not capture the fundamental difficulty of the problem. Let us consider a concrete example to illustrate.

Example 4 (Separation between transfer exponent and ρ_h)**.** Let the target distribution Q be a uniform distribution on the interval $[0, 1]$, and for some $\kappa \geqslant 1$, suppose that the source distribution P has density $p(x) = (\kappa+1)x^\kappa$ for $x \in [0, 1]$. With these definitions, it can be verified that $(P, Q) \in \mathcal{T}(\kappa, K)$ for some constant $K \in (0, 1]$, and moreover, that the quantity κ is the *smallest possible* transfer exponent for this pair. In contrast, another direct computation shows that the pair (P, Q) belongs to the class $\mathcal{D}(\kappa, C')$ for some constant $C' > 0$. These two inclusions establish a separation between the rates predicted by the transfer exponent and the similarity ρ_h. Indeed, as shown by our theory, the difficulty of estimation over $\mathcal{D}(\kappa, C')$ is smaller than that prescribed by $\mathcal{T}(\kappa, K)$. Indeed, if one observe n samples from the source distribution, the worst-case rate indicated by the computation from the transfer exponent is $n^{-\frac{2\beta}{2\beta+\kappa+1}}$, whereas the rate guaranteed by the similarity measure ρ_h is $n^{-\frac{2\beta}{2\beta+\kappa}}$. As an explicit example, Lipschitz functions ($\beta = 1$) and $\kappa = 1$, we obtain the slower rate $n^{-1/2}$ versus the faster rate $n^{-2/3}$, so that the ratio between the two rates diverges as $n^{1/6}$ as the sample size grows.

See also Figure 1 for an illustration of the connections and differences between the similarity measure and the transfer exponent.

4 Proofs

We now turn to the proofs of the results stated in the previous section.

4.1 Proof of Theorem 1

Recall that the estimate $\widehat{f}$ depends on the observations $\{(X_i, Y_i)\}_{i=1}^n$, and so should be understood as a random function. The core of the proof involves proving that, for each $x \in \mathcal{X}$, we have

$$\mathbf{E}\left[\big(\widehat{f}(x) - f^\star(x)\big)^2\right] \leqslant L^2 h_n^{2\beta} + \frac{4\sigma^2 + \|f^\star\|_\infty^2}{n} \frac{1}{\mu_n(\mathsf{B}(x, h_n))}, \tag{11}$$

where the expectation is taking over the observations $\{(X_i, Y_i)\}_{i=1}^n$. Given this inequality, the claim (4) of Theorem 1 follows, since by Fubini's theorem, we can write

$$\mathbf{E}\left[\|\widehat{f} - f^\star\|_{L^2(Q)}^2\right] = \int_{\mathcal{X}} \mathbf{E}\left[\big(\widehat{f}(x) - f^\star(x)\big)^2\right] \,\mathrm{d}Q(x).$$

Applying inequality (11) and recalling the definition of the similarity measure yields the claim (4).

We now focus on establishing the bound (11). Our proof makes use of the conditional expectation of $\widehat{f}$ given the covariates

$$\overline{f}(x) := \mathbf{E}[\widehat{f}(x) \mid X_1, \ldots, X_n], \quad \text{for any } x \in \mathcal{X}.$$

To be explicit, the expectation is taken over $Y_i \mid X_i$, $i = 1, \ldots, n$. With this definition, our first result provides a bound on the conditional bias and variance.

Lemma 2. *For each $x \in \mathcal{X}$ almost surely, the Nadaraya-Watson estimator $\widehat{f}$ satisfies the bounds*

$$\big(\overline{f}(x) - f^\star(x)\big)^2 \leqslant \|f^\star\|_\infty^2 \mathbf{1}\{x \notin \mathcal{G}_n\} + L^2 h_n^{2\beta} \mathbf{1}\{x \in \mathcal{G}_n\} \qquad \textit{and} \tag{12a}$$

$$\mathbf{E}[(\overline{f}(x) - \widehat{f}(x))^2 \mid X_1, \ldots, X_n] \leqslant \frac{\sigma^2}{\sum_{i=1}^n \mathbf{1}\{X_i \in \mathsf{B}(x, h_n)\}} \mathbf{1}\{x \in \mathcal{G}_n\}. \tag{12b}$$

We prove this auxiliary claim at the end of this section.

Taking the results of Lemma 2 as given, we continue our proof of the bound (11). For any fixed $x \in \mathcal{X}$, a conditioning argument yields

$$\mathbf{E}\left[(\widehat{f}(x) - f^\star(x))^2\right] = \mathbf{E}\left[\big(\overline{f}(x) - f^\star(x)\big)^2\right] + \mathbf{E}\Big[\mathbf{E}[(\overline{f}(x) - \widehat{f}(x))^2 \mid X_1, \ldots, X_n]\Big].$$

By applying the bounds (12a) and (12b) to the two terms above, respectively, we arrive at the upper bound $\mathbf{E}\left[(\widehat{f}(x) - f^\star(x))^2\right] \leqslant T_1 + T_2$, where

$$T_1 := \|f^\star\|_\infty^2 \,\mathbf{E}[\mathbf{1}\{x \notin \mathcal{G}_n\}] + L^2 h_n^{2\beta}, \quad \text{and} \quad T_2 := \mathbf{E}\Big[\mathbf{1}\{x \in \mathcal{G}_n\} \frac{\sigma^2}{\sum_{i=1}^n \mathbf{1}\{X_i \in \mathsf{B}(x, h_n)\}}\Big].$$

We bound each of these terms in turn.

Bounding T_1: By definition, the set $\mathcal{G}_n$ involves n independent random variables, so that for any $x \in \mathcal{X}$, we have

$$\mathbf{E}\left[\mathbf{1}\{x \notin \mathcal{G}_n\}\right] = \Big(1 - P\big(\mathsf{B}(x, h_n)\big)\Big)^{n_P} \Big(1 - Q\big(\mathsf{B}(x, h_n)\big)\Big)^{n_Q} \overset{\text{(i)}}{\leqslant} \frac{1}{n\,\mu_n(\mathsf{B}(x, h_n))}, \tag{13}$$

where step (i) follows from the elementary inequality $(1-p)^n(1-q)^m \leqslant \exp(-(np+mq)) \leqslant \frac{1}{np+mq}$, valid for $p, q \in (0,1)$ and nonnegative integers n, m. Consequently, the first term is upper bounded as

$$T_1 \leqslant \|f^\star\|_\infty^2 \frac{1}{n\,\mu_n(\mathsf{B}(x, h_n))} + L^2 h_n^{2\beta}. \tag{14a}$$

Bounding T_2: For a fixed $x \in \mathcal{X}$, and for each $i = 1, \ldots, n$, define the Bernoulli random variable $Z_i = \mathbf{1}[X_i \in \mathsf{B}(x, h_n)] \in \{0, 1\}$, along with the binomial random variables $U = \sum_{i=1}^{n_P} Z_i$ and $V = \sum_{i=n_P+1}^{n} Z_i$. With these definitions, we can write

$$\sum_{i=1}^{n} \mathbf{1}\{X_i \in \mathsf{B}(x, h_n)\} = U + V, \quad \text{and} \quad \mathbf{1}\{x \in \mathcal{G}_n\} = \mathbf{1}\big\{U + V > 0\big\}.$$

Consequently, by an elementary bound for binomial random variables (see Lemma 5), it follows that

$$T_2 = \mathbf{E}\Big[\mathbf{1}\{U + V > 0\}\frac{1}{U + V}\Big] \leqslant \frac{4}{n\,\mu_n(\mathsf{B}(x, h_n))}. \tag{14b}$$

Combining inequalities (14a) and (14b) yields the claim (11).

The only remaining detail is to prove the auxiliary lemma used in the proof.

Proof of Lemma 2. Recall that by definition, we have

$$\overline{f}(x) = \begin{cases} \dfrac{\sum_{i=1}^{n} f^\star(X_i)\mathbf{1}\{X_i \in \mathsf{B}(x, h_n)\}}{\sum_{i=1}^{n} \mathbf{1}\{X_i \in \mathsf{B}(x, h_n)\}} & x \in \mathcal{G}_n \\ 0 & x \notin \mathcal{G}_n \end{cases}$$

Proof of the bound (12a)**:** By a direct expansion, we have

$$\begin{aligned} \big(\overline{f}(x) - f^\star(x)\big)^2 \mathbf{1}\{x \in \mathcal{G}_n\} &= \Big(\frac{\sum_{i=1}^{n}(f^\star(x) - f^\star(X_i))\mathbf{1}\{X_i \in \mathsf{B}(x, h_n)\}}{\sum_{i=1}^{n} \mathbf{1}\{X_i \in \mathsf{B}(x, h_n)\}}\Big)^2 \mathbf{1}\{x \in \mathcal{G}_n\} \\ &\overset{\text{(i)}}{\leqslant} \frac{\sum_{i=1}^{n}(f^\star(x) - f^\star(X_i))^2\mathbf{1}\{X_i \in \mathsf{B}(x, h_n)\}}{\sum_{i=1}^{n} \mathbf{1}\{X_i \in \mathsf{B}(x, h_n)\}} \mathbf{1}\{x \in \mathcal{G}_n\} \\ &\overset{\text{(ii)}}{\leqslant} L^2 h_n^{2\beta} \mathbf{1}\{x \in \mathcal{G}_n\}, \end{aligned}$$

where step (i) follows from Jensen's inequality; and step (ii) makes use of Assumption 1. The bound (12a) is an immediate consequence.

Proof of the bound (12b)**:** In order to prove this claim, note that by independence among $\{(X_i, \xi_i)\}_{i=1}^n$,

$$
\begin{aligned}
\mathbf{E}[(\overline{f}(x) - \widehat{f}(x))^2 \mid X_1, \ldots, X_n] &= \sum_{i=1}^n \mathbf{E}[\xi_i^2 \mid X_i]\big(\tfrac{\mathbf{1}\{X_i \in \mathsf{B}(x,h_n)\}}{\sum_{i=1}^n \mathbf{1}\{X_i \in \mathsf{B}(x,h_n)\}}\big)^2 \mathbf{1}\{x \in \mathcal{G}_n\} \\
&\overset{\text{(iii)}}{\leqslant} \sigma^2 \sum_{i=1}^n \big(\tfrac{\mathbf{1}\{X_i \in \mathsf{B}(x,h_n)\}}{\sum_{i=1}^n \mathbf{1}\{X_i \in \mathsf{B}(x,h_n)\}}\big)^2 \mathbf{1}\{x \in \mathcal{G}_n\} \\
&= \frac{\sigma^2}{\sum_{i=1}^n \mathbf{1}\{X_i \in \mathsf{B}(x,h_n)\}} \mathbf{1}\{x \in \mathcal{G}_n\},
\end{aligned}
$$

which proves the claim. Here step (iii) is a consequence of Assumption 3. □

4.2 Proof of Corollary 1

Fix some $h \in (0, 1]$, and introduce the indicator variable $\eta = \mathbf{1}\{\alpha \geqslant 1\}$. We then have

$$
\begin{aligned}
\int_{\mathcal{X}} \frac{1}{n_P P(\mathsf{B}(x,h)) + n_Q Q(\mathsf{B}(x,h))}\, \mathrm{d}Q(x) &\leqslant \min\Big\{\frac{1}{n_P}\rho_h(P,Q), \frac{1}{n_Q}\rho_h(Q,Q)\Big\} \\
&\leqslant 3^\eta C \min\Big\{\frac{1}{n_P h^\alpha}, \frac{1}{n_Q h^\eta}\Big\} \\
&\leqslant 2 \cdot 3^\eta C \frac{1}{n_P h^\alpha + n_Q h^\eta}.
\end{aligned}
$$

The last inequality follows from (1) and standard covering number bounds (note $h \leqslant 1$). Thus the final performance bound is

$$
2 \cdot 3^\eta C L^2 \Big\{h^{2\beta} + \frac{L^2 + \sigma^2}{n_P h^\alpha + n_Q h^\eta}\Big\}.
$$

We choose the bandwidth $h^\star$ so as to trade off between two terms in this risk bound; more precisely, we set

$$
h^\star = \Big(\big(\frac{n_Q}{L^2 + \sigma^2}\big) + \big(\frac{n_P}{L^2 + \sigma^2}\big)^{\frac{2\beta+\eta}{2\beta+\alpha}}\Big)^{-\frac{1}{2\beta+\eta}}
$$

This choice is valid, since $\sigma^2 \geqslant L^2$ and $\max\{n_P, n_Q\} \geqslant 4\sigma^2$ by assumption. Substituting this choice of bandwidth into the risk bound (4) yields the claim.

4.3 Proof of Theorem 2(a)

Before giving the complete proof, we outline the main steps involved.

1. We first construct a hard instance $(P, Q) \in \mathcal{D}(\alpha, C)$. This instance is designed such that the integral quantity $\rho_h(P, Q)$ must scale as $Ch^{-\alpha}$.

2. Then we select a family of hard regression functions contained within $\mathcal{F}(\beta, L)$ that guarantees the worst-case expected error for our pair of distributions, (P, Q).

3. Finally, we apply Fano's method over this set of regression functions to show that the expected error must scale as the righthand side of inequality (8a).

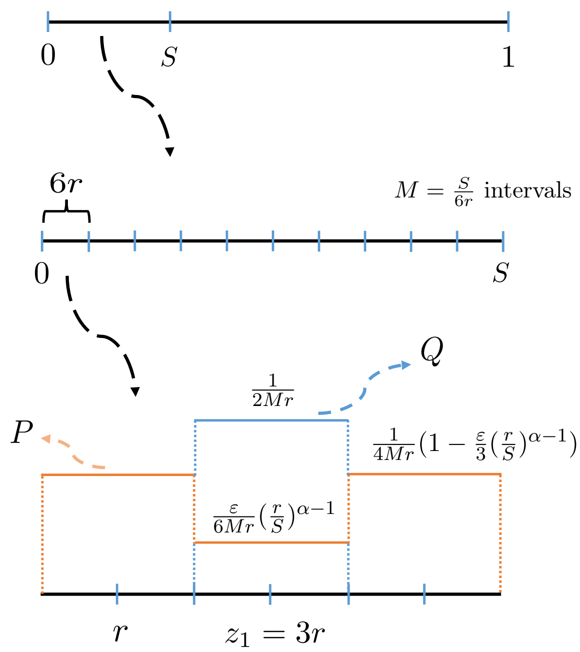

Figure 2. An illustration of the distributions (P, Q) constructed as a hard pair in our lower bound.

It is worth commenting on our proof strategy in relation to past work. On one hand, in the case $\alpha \geqslant 1$, our construction of the distributions (P, Q) is adapted from the lower bound argument introduced by Kpotufe and Martinet [11]. The technical work involves constructing pairs of densities of P, Q, and establishing their membership in the class $\mathcal{D}(\alpha, C)$. As for the case $\alpha \in (0, 1)$, as stated in Theorem 2(b), we use a different construction of the distribution pair (P, Q), one that is new (to the best of our knowledge). We combine these constructions of "hard" source-target pairs, in particular by packing the interval $[0, 1]$ with a variable number of small intervals (e.g., [28, 25, 26]). By adapting the number of intervals (and constructing a packing set of the function class $\mathcal{F}(\beta, L)$ appropriately over these intervals), one can adapt the hardness of the lower bound instance to change with the number of samples. In this case, we are able to do this such that the hardness scales appropriately with the critical parameters that govern the final minimax lower bound: $n_P, n_Q, \sigma, \alpha, \beta$. With this high-level overview in place, we now proceed to the technical content of the proof.

Constructing "hard" source-target pairs: For scalars $S, r \in (0, 1]$, define $M = \frac{S}{6r}$ along with the intervals

$$I_j := (z_j - 3r, z_j + 3r], \quad \text{where} \quad z_j := 6jr - 3r, \qquad j = 1, \ldots, M.$$

We specify P and Q on each interval I_j as follows:

subinterval	density of P	density of Q
$(z_j - 3r, z_j - r]$	$\frac{1}{4Mr}(1 - \frac{\varepsilon}{3}(\frac{r}{S})^{\alpha-1})$	0
$(z_j - r, z_j + r]$	$\frac{\varepsilon}{6Mr}(\frac{r}{S})^{\alpha-1}$	$\frac{1}{2Mr}$
$(z_j + r, z_j + 3r]$	$\frac{1}{4Mr}(1 - \frac{\varepsilon}{3}(\frac{r}{S})^{\alpha-1})$	0

Table 1. Specification of densities for lower bound pair of distributions (P, Q) on the interval I_j.

By construction, both P and Q assign probability $1/M$ to the entire interval I_j. The following proposition verifies that (P, Q) lies in $\mathcal{D}(\alpha, C)$ for proper choices of the ε and S.

Proposition 2. *Let $\alpha \geqslant 1$ and $C \geqslant 1$. Define P and Q as in Table 1, with the following choice of parameters ε, S:*

(a) *if $C > 6$, set $\varepsilon = 6/C$, and $S = 1/4$;*

(b) *if $1 \leqslant C \leqslant 6$, set $\varepsilon = 1$, and $S = \frac{1}{4}(C/6)^{1/\alpha}$.*

Then for any choice of $M, r > 0$ satisfying $S = 6Mr$, the pair (P, Q) lies in $\mathcal{D}(\alpha, C)$.

See Section 4.3.1 for the proof of this claim.

Construction of "hard" regression functions. Next we construct a packing of the function class of $\mathcal{F}(\beta, L)$. We do so by summing together scaled and shifted copies of base function $\Psi\colon [-1, 1] \to \mathbf{R}$ that satisfies the boundary conditions $\Psi(-1) = \Psi(1) = 0$, along with

$$\big|\Psi(x) - \Psi(y)\big| \leqslant |x - y|^{\beta}, \quad \text{for all } x, y \in [-1, 1], \quad \text{and,} \tag{15a}$$

$$\int_{-1}^{1} \Psi^2(x)\, \mathrm{d}x =: C_{\Psi}^2 > 0. \tag{15b}$$

There are many possible choices of Ψ; see Chapter 2 in the book [25] for details. For our proof, we also require the bound $C_{\Psi}^2 \leqslant 1/6$, so that we make the explicit choice

$$\Psi(x) := e^{-1/(1-x^2)} \mathbf{1}\{|x| \leqslant 1\}.$$

We now form a class of functions using sums of the form

$$f_b(x) := \sum_{j=1}^{M} b_j \phi_j(x), \quad \text{where} \quad \phi_j(x) := L r^{\beta} \Psi\Big(\frac{x - z_j}{r}\Big),$$

and $b = (b_1, \ldots, b_M) \in \{0, 1\}^M$ is a Boolean sequence. Our construction makes use of the Gilbert-Varshamov lemma (e.g. [25, Lemma 2.9]), which for $M \geqslant 8$, guarantees the existence of a subset $\mathcal{B} \subset \{0, 1\}^M$ of cardinality at least $2^{M/8}$ such that

$$\|b - b'\|_1 \geqslant M/8 \qquad \text{for all distinct } b, b' \in \mathcal{B}. \tag{16}$$

Lemma 3. *The function class $\mathcal{H} := \Big\{ f_b \mid b \in \mathcal{B} \Big\}$ has the following properties:*

(a) *It is contained within the Hölder class—$\mathcal{H} \subset \mathcal{F}(\beta, L)$.*

(b) *Pairs of functions are well-separated: for each distinct $f, g \in \mathcal{H}$, we have*

$$\|f - g\|_{L^2(Q)}^2 \geqslant \tfrac{C_{\Psi}^2}{16} L^2 r^{2\beta}.$$

(c) *Its elements satisfy the following $L^2(P)$ and $L^2(Q)$ bounds:*

$$\|f\|_{L^2(Q)}^2 \leqslant \frac{C_{\Psi}^2 M}{2S} L^2 r^{2\beta+1} \quad \textit{and} \quad \|f\|_{L^2(P)}^2 \leqslant \frac{\varepsilon C_{\Psi}^2 M}{6 S^{\alpha}} L^2 r^{2\beta+\alpha},$$

for all $f \in \mathcal{H}$.

Applying Fano's method. We now combine the preceding constructions with a Fano argument to complete the proof of the lower bound. For any function $f \in \mathcal{H}$, let ν_f be the distribution $\{(X_i, Y_i)\}_{i=1}^n$ where (X, Y) pairs are related by our nonparametric regression model (2) with $f = f^\star$. For proving our lower bound, it suffices to consider Gaussian noise: in particular, $\xi_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, \sigma^2)$ for $i = 1, \ldots, n$. These variables satisfy Assumption 3.

With these choices, Kullback-Leibler divergence between any given pair (ν_f, ν_g) can be bounded as

$$D_{\text{kl}}(\nu_f \parallel \nu_g) = \frac{1}{2\sigma^2}\Big(n_P\|f-g\|_{L^2(P)}^2 + n_Q\|f-g\|_{L^2(Q)}^2\Big) \leqslant \frac{2}{\sigma^2}\Big(n_P \max_{f\in\mathcal{H}}\|f\|_{L^2(P)}^2 + n_Q \max_{f\in\mathcal{H}}\|f\|_{L^2(Q)}^2\Big).$$

Now applying part (c) of Lemma 3 yields

$$\begin{aligned} D_{\text{kl}}(\nu_f \parallel \nu_g) &\leqslant MC_\Psi^2\Big\{n_P\frac{L^2}{3\sigma^2}\frac{\varepsilon}{S^\alpha}r^{2\beta+\alpha} + n_Q\frac{L^2}{\sigma^2}\frac{1}{S}r^{2\beta+1}\Big\} \\ &\leqslant M\Big\{\frac{4^\alpha}{C}\frac{L^2}{\sigma^2}n_P r^{2\beta+\alpha} + \frac{4^\alpha}{C}\frac{L^2}{\sigma^2}n_Q r^{2\beta+1}\Big\} \end{aligned}$$

The final inequality arises by using $C_\Psi^2 \leqslant 1/6$. Suppose we take

$$r = \Big(\big(64\frac{4^\alpha}{C}\frac{L^2 n_P}{\sigma^2}\big)^{\frac{2\beta+1}{2\beta+\alpha}} + \big(64\frac{4^\alpha}{C}\frac{L^2 n_Q}{\sigma^2}\big)\Big)^{-\frac{1}{2\beta+1}}$$

Then for any distinct $f, g \in \mathcal{H}$, we obtain

$$D_{\text{kl}}(\nu_f \parallel \nu_g) \leqslant M/32. \tag{17}$$

By a standard reduction to hypothesis testing [26, chap. 15] along with part (a),

$$\inf_{\widehat{f}}\sup_{f^\star\mathcal{F}(\beta,L)} \mathbf{E}\Big[\|\widehat{f}-f^\star\|_{L^2(Q)}^2\Big] \geqslant \frac{\min_{(f,g)\in\binom{\mathcal{H}}{2}}\|f-g\|_{L^2(Q)}^2}{4}\Big\{1 - \frac{\log 2 + \max_{(f,g)\in\binom{\mathcal{H}}{2}} D_{\text{kl}}(\nu_f \parallel \nu_g)}{\log|\mathcal{H}|}\Big\}$$

Thus, after applying part (b) of Lemma 3, we obtain

$$\inf_{\widehat{f}}\sup_{f^\star\mathcal{F}(\beta,L)} \mathbf{E}\Big[\|\widehat{f}-f^\star\|_{L^2(Q)}^2\Big] \geqslant \frac{C_\Psi^2}{64}L^2 r^{2\beta}\Big(1-\frac{8}{M}-\frac{1}{4}\Big) \geqslant \frac{C_\Psi^2 L^2}{256}\Big(\big(64\frac{4^\alpha}{C}\frac{L^2 n_P}{\sigma^2}\big)^{\frac{2\beta+1}{2\beta+\alpha}} + \big(64\frac{4^\alpha}{C}\frac{L^2 n_Q}{64\sigma^2}\big)\Big)^{-\frac{2\beta}{2\beta+1}},$$

provided that $M \geqslant 32$. Equivalently, $r \leqslant S/192$. It suffices that $r \leqslant \frac{1}{4608}$, this is ensured by having

$$\max\{n_P, n_Q\} \geqslant \Big(72\frac{\sigma^2}{L^2}\frac{C}{4^\alpha}\Big)^{2\beta+\alpha}.$$

4.3.1 Proof of Proposition 2

We will show that for a general choice of $\varepsilon, S \in (0, 1]$, the following holds:

$$P\big(\mathsf{B}(x,h)\big) \geqslant \frac{\varepsilon}{3}\big(\frac{h}{4S}\big)^{\alpha-1}Q\big(\mathsf{B}(x,h)\big), \quad \text{for all } x \in \text{supp}(Q), \text{ and any } h > 0. \tag{18}$$

For the moment let us take this bound as given. By Lemma 1, note that bound (18) implies that $(P, Q) \in \mathcal{D}(\alpha, \mathcal{C}(\varepsilon, S))$, with $\mathcal{C}(\varepsilon, S) = \frac{6}{\varepsilon}(4S)^{\alpha-1}$, for any $\varepsilon, S \in (0, 1]$. Note that the parameter choices given in the statement of the result ensure that $\varepsilon, S \in (0, 1]$. When $C \geqslant 6$, we have

$\mathcal{C}(\varepsilon, S) = 6(C/6)^{1-1/\alpha} = C(6/C)^{1/\alpha} \leqslant 6 \leqslant C$. Otherwise $C \leqslant 6$ and $\mathcal{C}(\varepsilon, S) = C$. Therefore, checking the two cases $C > 6$ and $C \leqslant 6$ verifies $\mathcal{C}(\varepsilon, S) = C$ in both regimes, which furnishes the claim.

We now turn to establish bound (18). Let $h > 0$. First observe that the support of Q is the disjoint union of intervals $\cup_{j=1}^{M}(z_j - r, z_j + r]$. Thus, fix x in the support of Q, and let z_j denote the center of the interval to which x belongs. Suppose that $h \in [0, 4r]$, in which case, we have the inclusion $\mathsf{B}(x, h) \subset I_j$, whence the lower bound

$$\begin{aligned} P(\mathsf{B}(x,h)) &\geqslant P\big(\mathsf{B}(x,h) \cap \mathsf{B}(z_j, r)\big) \\ &\overset{\text{(i)}}{=} \frac{\varepsilon}{3}\Big(\frac{r}{S}\Big)^{\alpha-1} Q\big(\mathsf{B}(x,h) \cap \mathsf{B}(z_j, r)\big) \\ &\overset{\text{(ii)}}{\geqslant} \frac{\varepsilon}{3}\Big(\frac{h}{4S}\Big)^{\alpha-1} Q\big(\mathsf{B}(x,h) \cap \mathsf{B}(z_j, r)\big) \\ &\overset{\text{(iii)}}{=} \frac{\varepsilon}{3}\Big(\frac{h}{4S}\Big)^{\alpha-1} Q\big(\mathsf{B}(x,h)\big) \end{aligned} \tag{19}$$

Above, step (i) follows from the construction of P, Q; step (ii) follows from $h \leqslant 4r$, whereas step (iii) follows since $\mathsf{B}(x, h) \subset I_j$ and Q assigns no mass to the set $I_j \setminus \mathsf{B}(z_j, r)$.

Otherwise, we may assume that $h \in [4r, S]$, in which case we have the inclusion $\mathsf{B}(x, h) \supset I_j$. Denote by $N \geqslant 1$ the number of intervals of the form I_j that are included within $\mathsf{B}(x, h)$. Note that since $\mathsf{B}(x, h)$ is connected, it is always contained in at most $N + 2$ intervals (by considering partial intervals on the left and right). Thus,

$$\frac{P(\mathsf{B}(x,h))}{Q(\mathsf{B}(x,h))} \overset{\text{(iii)}}{\geqslant} \frac{N \cdot P(I_j)}{(N+2) \cdot Q(I_j)} \overset{\text{(iv)}}{\geqslant} \frac{1}{3}. \tag{20}$$

Here step (iii) follows since $\mathsf{B}(x, h)$ is contained in a collection of at most $(N + 2)$ intervals and contains at least N intervals, and the intervals are disjoint and have the same mass under both P and Q. On the other hand, step (iv) uses the equivalence $P(I_j) = Q(I_j)$, along with the fact that the function $x \mapsto \frac{x}{x+2}$ is increasing on the set $\{x \geqslant 1\}$.

Therefore, combining inequalities (19) and (20), we conclude that

$$P(\mathsf{B}(x,h)) \geqslant \frac{1}{3}\Big[\varepsilon\Big(\frac{h}{4S}\Big)^{\alpha-1} \wedge 1\Big] Q(\mathsf{B}(x,h)) \geqslant \frac{\varepsilon}{3}\big(\frac{h}{4S}\big)^{\alpha-1} Q(\mathsf{B}(x,h))$$

for every x in the support of Q, the final inequality follows since $\alpha \geqslant 1$. Since $h > 0$ was arbitrary, this establishes bound (18) and completes the proof.

4.3.2 Proof of Lemma 3

We prove each of the three parts in turn.

Proof of part (a): Fix a Boolean vector $b \in \{0, 1\}^M$. Note that the function ϕ_j is supported on the interval I_j, which is disjoint from any other interval $I_k, k \neq j$. Since Ψ satisfies the continuity condition (15a), it follows that ϕ_j is (β, L)-Hölder. Finally, we have $f_\varepsilon(0) = 0$ by definition. Taking these properties together, we have shown that $f_\varepsilon \in \mathcal{F}(\beta, L)$, as required.

Proof of part (b): For any distinct pair $b, b' \in \mathcal{B}$, we have

$$\begin{aligned}\int_0^1 (f_b(x) - f_{b'}(x))^2 \, \mathrm{d}Q(x) &= \int_0^1 \Big(\sum_{j=1}^M (b_j - b_j')\phi_j(x)\Big)^2 \, \mathrm{d}Q(x) \\ &\overset{\text{(i)}}{=} \frac{1}{2Mr} \sum_{j=1}^M (b_j - b_j')^2 \int_{z_j - 3r}^{z_j+3r} \phi_j^2(x) \, \mathrm{d}x \\ &\overset{\text{(ii)}}{=} \frac{C_\Psi^2}{2M} L^2 r^{2\beta} \|b - b'\|_1 \\ &\overset{\text{(iii)}}{\geqslant} \frac{C_\Psi^2}{16} L^2 r^{2\beta}.\end{aligned}$$

Here step (i) follows from the definition of Q along with the disjointedness of the supports of ϕ_j. Step (ii) follows from equation (15b) and the fact that $b, b' \in \mathcal{B} \subset \{0,1\}^M$. Finally, step (iii) follows from the Gilbert-Varshamov separation (16).

Proof of part (c): For any $b \in \mathcal{B}$, by following the calculations above, for $\mu \in \{P, Q\}$, we have by symmetry

$$\int_0^1 f_b^2(x) \, \mathrm{d}\mu(x) = \sum_{j=1}^M b_j^2 \int_{I_j} \phi_j^2(x) \, \mathrm{d}\mu(x) \leqslant M \int_{I_1} \phi_1^2(x) \, \mathrm{d}\mu(x).$$

Now observe that $\int_0^{6r} \phi_1^2(x) \, \mathrm{d}Q(x) = \frac{C_\Psi^2}{2M} L^2 r^{2\beta}$, and consequently, $\|f_b\|_{L^2(Q)}^2 \leqslant L^2 r^{2\beta} C_\Psi^2/2$. Additionally, we can compute

$$\int_0^{6r} \phi_1^2(x) \, \mathrm{d}P(x) = \frac{\varepsilon}{6rM^\alpha} \int_{2r}^{4r} \phi_1^2(x) \, \mathrm{d}x = \frac{\varepsilon}{6S^\alpha} L^2 r^{2\beta+\alpha} C_\Psi^2.$$

Thus, we have established the upper bound $\|f_b\|_{L^2(P)}^2 \leqslant \varepsilon L^2 r^{2\beta+\alpha-1}/(6S^{\alpha-1})$.

4.4 Proof of Theorem 2(b)

Given the inclusion $\mathcal{D}'(\alpha, 1) \subset \mathcal{D}'(\alpha, C)$, it suffices to prove a lower bound for $C = 1$.

Construction of "hard" distributions. Let $Q = \delta_1$, and let P_α be the distribution supported on $[0,1]$ with density $p_\alpha(x) := \alpha(1-x)^{\alpha-1}\mathbf{1}\{x \in [0,1]\}$. By construction, we then have

$$\rho_h(P_\alpha, Q) = \frac{1}{P_\alpha(B(1,h))} = h^{-\alpha} \quad \text{for all } h \in (0,1],$$

which implies that $(P_\alpha, Q) \in \mathcal{D}'(\alpha, 1)$. From herein, we adopt the shorthand $P := P_\alpha$ so as to lighten notation.

Construction of two point alternative. If the regression function is f, we denote the resulting joint distribution of $\{(X_i, Y_i)\}_{i=1}^n$ by ν_f. We consider the two point alternatives $\{f_t, g\}$ with $g \equiv 0$ and $f_t(x) := L(x-t)_+^\beta$. The next result demonstrates the validity of this choice:

Lemma 4. *For any $t \in [0,1]$, the function f_t belongs to $\mathcal{F}(\beta, L)$.*

See Section 4.4.1 for the proof.

Moreover, by straightforward calculations, we find that $\|f_t\|_{L^2(Q)}^2 = L^2(1-t)^{2\beta}$, and

$$\begin{aligned}\|f_t\|_{L^2(P)}^2 &= L^2 \int_t^1 \alpha(1-x)^{\alpha-1}(x-t)^{2\beta}\,\mathrm{d}x \\ &\leqslant L^2(1-t)^{2\beta}\int_0^{1-t}\alpha s^{\alpha-1}\,\mathrm{d}s = L^2(1-t)^{2\beta+\alpha}.\end{aligned}$$

Applying Le Cam's method. We are now equipped to apply Le Cam's two point bound. In particular, we have

$$\inf_{\widehat{f}}\sup_{f^\star\in\mathcal{F}(\beta,L)} \mathbf{E}\Big[\|\widehat{f}-f^\star\|_{L^2(Q)}^2\Big] \geqslant \frac{L^2(1-t)^{2\beta}}{16}\exp\big(-D_{\mathrm{kl}}(\nu_{f_t}\parallel\nu_g)\big)$$

By standard KL calculations (using $\mathsf{N}(0,\sigma^2)$ noises)

$$D_{\mathrm{kl}}(\nu_{f_t}\parallel\nu_g) = \frac{L^2}{2\sigma^2}\Big\{n_P(1-t)^{2\beta+\alpha} + n_Q(1-t)^{2\beta}\Big\}$$

Finally, we make the

$$1-t = \bigg(\Big(\frac{L^2 n_P}{2\sigma^2}\Big)^{\frac{1}{2\beta+\alpha}} + \Big(\frac{L^2 n_Q}{2\sigma^2}\Big)^{\frac{1}{2\beta}}\bigg)^{-1}$$

A little bit of algebra shows that this choice guarantees that $D_{\mathrm{kl}}(\nu_{f_t}\parallel\nu_g)\leqslant 2$, which completes the proof.

4.4.1 Proof of Lemma 4

We begin by observing that $f_t(0)=0$. Thus, in order to prove the claim, it suffices to show that

$$f_t(y) - f_t(x) \leqslant L(y-x)^\beta \quad \text{for any pair } x,y \text{ such that } 0\leqslant t<x<y\leqslant 1.$$

In order to prove this bound, consider an arbitrary point $x\in(t,1)$, and define the function

$$\phi_x(y) := L(y^\beta - x^\beta) - L(y-x)^\beta \qquad \text{for } y\in[x,1].$$

We can compute the derivative $\phi_x'(y) = L\beta(y^{\beta-1}-(y-x)^{\beta-1})$. Since $y\geqslant y-x>0$ and $\beta\leqslant 1$, we have $y^{\beta-1}\leqslant(y-x)^{\beta-1}$, and hence $\phi_x'(y)\leqslant 0$. Consequently, the function ϕ_x is non-increasing, and since $y>x$, it follows that $\phi_x(y)\leqslant\phi_x(x)=0$. Putting together the pieces completes the proof.

4.5 Proof of Proposition 1

Starting with the assumed bound (9), we have

$$\int_{\mathcal{X}}\frac{1}{P(\mathsf{B}(x,h))}\,\mathrm{d}Q(x) \leqslant \frac{1}{\lambda}\int_{\mathcal{X}}\frac{1}{Q(\mathsf{B}(x,h))}\,\mathrm{d}Q(x). \tag{21}$$

By definition of the covering number $N := N(h/2)$, there is a collection $\{z^j\}_{j=1}^N$ such that the set $\mathcal{X}$ is contained within the union $\bigcup_{j=1}^N \mathsf{B}(z^j,\frac{h}{2})$. This fact, combined with our previous bound (21), implies that

$$\int_{\mathcal{X}}\frac{1}{P(\mathsf{B}(x,h))}\,\mathrm{d}Q(x) \leqslant \frac{1}{\lambda}\sum_{j=1}^N\int_{\mathsf{B}(z_j,h/2)}\frac{1}{Q(\mathsf{B}(x,h))}\,\mathrm{d}Q(x). \tag{22}$$

Note by the triangle inequality, for each $j \in [N]$ and $x \in \mathsf{B}(z_j, h/2)$, we have $\mathsf{B}(z_j, h/2) \subset \mathsf{B}(x, h)$. This inclusion implies that

$$\int_{\mathsf{B}(z_j,h/2)} \frac{1}{Q(\mathsf{B}(x,h))} \, \mathrm{d}Q(x) \leqslant \int_{\mathsf{B}(z_j,h/2)} \frac{1}{Q(\mathsf{B}(z_j,h/2))} \, \mathrm{d}Q(x) = 1,$$

for each $j \in [N]$. Combining this inequality with the bound (22) yields the claim.

4.6 Proof of Lemma 1

By assumption, we have the upper bound

$$\int_0^1 \frac{1}{P(\mathsf{B}(x,h))} \, \mathrm{d}Q(x) \leqslant \frac{1}{Kh^\gamma} \int_0^1 \frac{1}{Q(\mathsf{B}(x,h))} \, \mathrm{d}Q(x)$$

Moreover, we can find a collection of $N := \lceil 1/h \rceil$ balls with centers $\{z_j\}_{j=1}^N$ of radius $h/2$ that cover the interval $[0, 1]$, whence

$$\int_0^1 \frac{1}{Q(\mathsf{B}(x,h))} \, \mathrm{d}Q(x) \leqslant \sum_{j=1}^N \int_{x \in \mathsf{B}(z_j,h/2)} \frac{1}{Q(\mathsf{B}(x,h))} \, \mathrm{d}Q(x) \leqslant N.$$

The final inequality follows from the inclusion $\mathsf{B}(x, h) \supset \mathsf{B}(z_j, h/2)$.

Now define the function $g(t) := \lceil t \rceil / t$, and observe that $g(t) \leqslant 2$ whenever $t \geqslant 1$. Consequently, we can write

$$h^{\gamma+1} \rho_h(P,Q) \leqslant \frac{1}{K} g(1/h) \leqslant \frac{2}{K}, \quad \text{for any } h \leqslant 1.$$

Passing to the supremum over $h \in (0, 1]$ yields the claim.

5 Discussion

In this paper, we have studied the problem of covariate shift in the context of nonparametric regression. We have shown that a measure of (dis)-similarity ρ_h between the source and target distributions, as defined in equation (1), can be used to characterize how minimax risks change as the source-target pair are varied. In particular, we proved upper bounds on the Nadaraya-Watson estimator over Hölder classes that are an explicit function of the similarity ρ_h, and also established matching lower bounds over classes constrained in terms of the similarity. We also discussed how the measure ρ_h is related to other characterizations of covariate shift from past work, including likelihood ratio bounds and transfer exponents. Our work shows that similarity measure ρ_h provides a more fine-grained characterization of how covariate shift changes the difficulty of non-parametric regression.

Our work leaves open a number of open questions. First, our lower bounds for covariate shift (cf. Theorem 2) are obtained within a global minimax framework, which involves worst-case assessments over a certain function class. These lower bounds match our upper bound on the NW estimator (cf. Theorem 1) for certain source-target pairs (P, Q). But the upper bound actually depends explicitly on the source-target pair. Is this upper bound always optimal? Or are there instances of covariate shift for which Nadaraya-Watson is suboptimal for some Hölder continuous function? In general, this question appears non-trivial: even without the (interesting) complication of covariate shift, there are few results that give distribution-dependent results for nonparametric regression outside of the uniform distribution and fixed-design problems.

Acknowledgements

We would like to thank Samory Kpotufe for helpful email exchanges. This work was partially supported by NSF-DMS grant 2015454, NSF-IIS grant 1909365, NSF-FODSI grant 202350, and DOD-ONR Office of Naval Research N00014-21-1-2842 to MJW.

A Elementary bound for binomial variables

In this section, we state and prove an elementary bound for binomial random variables, used in the proof of Theorem 1.

Lemma 5. *Let n, m be positive integers and $p, q \in (0, 1)$. Suppose that $U \sim \mathsf{Bin}(n, p)$ and $V \sim \mathsf{Bin}(m, q)$. Then*

$$\mathbf{E}\left[\frac{1}{U+V}\mathbf{1}\{U+V>0\}\right] \leqslant \frac{4}{np+mq}.$$

Proof. We begin by observing that conditionally on the event $\{U + V > 0\}$, we have the lower bound

$$U+V \geqslant \frac{U+V+1}{2} \geqslant \frac{U+1}{2} \vee \frac{V+1}{2}.$$

These lower bounds allow us to write

$$\mathbf{E}\left[\frac{1}{U+V}\mathbf{1}\{U+V>0\}\right] \leqslant \mathbf{E}\,\frac{2}{U+1} \wedge \mathbf{E}\,\frac{2}{V+1} \leqslant \frac{2}{(np \vee mq)} \leqslant \frac{4}{np+mq}.$$

Here the penultimate inequality is a consequence of known results for binomial random variables [4, equation (3.4)]. □

References

[1] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

[2] S. Ben-David, T. Lu, T. Luu, and D. Pál. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 129–136, 2010.

[3] J. Blitzer, M. Dredze, and F. Pereira. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[4] M. T. Chao and W. E. Strawderman. Negative moments of positive random variables. *J. Amer. Statist. Assoc.*, 67(338):429–431, 1972.

[5] C. Cortes, M. Mohri, and A. M. Medina. Adaptation based on generalized discrepancy. *J. Mach. Learn. Res.*, 20(1):1–30, 2019.

[6] P. Germain, A. Habrard, F. Laviolette, and E. Morvant. A pac-bayesian approach for domain adaptation with specialization to linear classifiers. In *International Conference on Machine Learning*, pages 738–746, 2013.

[7] G. Halász. Statistical interpolation. *Mat. Lapok*, 23:71–87 (1973), 1972.

[8] A. Hassan, R. Damper, and M. Niranjan. On acoustic emotion recognition: Compensating for covariate shift. *IEEE Trans. Audio Speech Lang. Process.*, 21(7):1458–1468, 2013.

[9] I. A. Ibragimov and R. Z. Khas′ minskiĭ. Nonparametric regression estimation. *Dokl. Akad. Nauk SSSR*, 252(4):780–784, 1980.

[10] S. Kpotufe. Lipschitz density-ratios, structured data, and data-driven tuning. In *Artificial Intelligence and Statistics*, pages 1320–1328, 2017.

[11] S. Kpotufe and G. Martinet. Marginal singularity and the benefits of labels in covariate-shift. *Ann. Statist.*, 49(6):3299–3323, 2021.

[12] Q. Lei, W. Hu, and J. Lee. Near-optimal linear regression under distribution shift. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proc. Mach. Learn. Res.*, pages 6164–6174. PMLR, 18–24 Jul 2021.

[13] Y. Li, H. Kambara, Y. Koike, and M. Sugiyama. Application of covariate shift adaptation techniques in brain–computer interfaces. *IEEE Trans. Biomed. Eng.*, 57(6):1318–1324, 2010.

[14] Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

[15] Y. Mansour, M. Mohri, and A. Rostamizadeh. Multiple source adaptation and the rényi divergence. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 367–374. AUAI Press, 2009.

[16] M. Mohri and A. M. Medina. New analysis and algorithm for learning with drifting distributions. In *International Conference on Algorithmic Learning Theory*, pages 124–138. Springer, 2012.

[17] M. Mousavi Kalan, Z. Fabian, S. Avestimehr, and M. Soltanolkotabi. Minimax lower bounds for transfer learning with linear and one-hidden layer neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1959–1969. Curran Associates, Inc., 2020.

[18] E. A. Nadaraya. On estimating regression. *Theory Probab. Appl.*, 9(1):141–142, 1964.

[19] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset Shift in Machine Learning*. The MIT Press, 2009.

[20] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *Computer Vision – ECCV 2010*, pages 213–226, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[21] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *J. Statist. Plann. Inference*, 90(2):227–244, 2000.

[22] C. J. Stone. Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, 10(4):1040–1053, 1982.

[23] M. Sugiyama, S. Nakajima, H. Kashima, P. V. Buenau, and M. 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.

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

[25] A. B. Tsybakov. *Introduction to nonparametric estimation.* Springer Series in Statistics. Springer, New York, 2009. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

[26] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint.* Cambridge University Press, Cambridge, UK, 2019.

[27] G. S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372, 1964.

[28] B. Yu. Assouad, Fano and Le Cam. In *Festschrift in Honor of L. Le Cam on his 70th Birthday.* 1993.