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

**Reese Pathak** [*1] **Cong Ma** [*2] **Martin J. Wainwright** [1 3]

## 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, the 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-population of 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 (Shimodaira, 2000; Quionero-Candela et al., 2009). 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 (Saenko et al., 2010), biomedical engineering (Li et al., 2010), sentiment analysis (Blitzer et al., 2007), and audio processing (Hassan et al., 2013).

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 $\{X_i\}_{i=n_P+1}^{n_P+n_Q}$ 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 (Halász, 1972; Ibragimov & Khas′ minskiĭ, 1980; Stone, 1982; Tsybakov, 2009). 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 $Q$.

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

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

[1]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley [2]Department of Statistics, University of Chicago [3]Department of Statistics, University of California, Berkeley. Correspondence to: Reese Pathak <pathakr@berkeley.edu>.

*Proceedings of the $39^{th}$ International Conference on Machine Learning*, Baltimore, Maryland, USA, PMLR 162, 2022. 

[1]We note that this quantity will actually serve as a *dis*-similarity measure in the sequel. When it is larger, we obtain worse rates of estimation over the nonparametric classes considered in this paper.

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 show the significance of this similarity measure via the following contributions:

(i) For regression functions that are Hölder continuous, we demonstrate that performance guarantee for the Nadaraya-Watson kernel estimator under covariate shift 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 $\rho_h$ can be controlled based on the metric properties of the space $\mathcal{X}$. In addition, we compare $\rho_h$ 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 (Shimodaira, 2000), 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., (Ben-David et al., 2010a;b; Germain et al., 2013; Mansour et al., 2009a; Cortes et al., 2019; Mohri & Medina, 2012)). 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 (Kpotufe & Martinet, 2021), 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 (Sugiyama et al., 2012; Kpotufe, 2017) or in terms of information-theoretic divergences between the source-target pair (Sugiyama et al., 2008; Mansour et al., 2009b). Our work is inspired in part by the work of Kpotufe and Martinet (Kpotufe & Martinet, 2021), 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 $\rho_h$, provides sharper rates than those obtainable by considering the transfer exponent; see Section 3.2 for details. Thus, the similarity measure $\rho_h$ provides a more fine-grained control on the effect of covariate shift on nonparametric regression.

Finally, it is worth mentioning other recent works on covariate shift, including on linear models (Lei et al., 2021), as well as linear models and one-layer neural networks (Mousavi Kalan et al., 2020). Although these results deal with covariate shift, the rates obtained are parametric ones, and hence not directly comparable to the nonparametric rates obtained here.

### 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 $\sigma$-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 $\mathcal{X}$.

## 2. Covariate shift in the context of nonparametric regression

In this section, we use the similarity measure introduced in equation (1) to characterize the rate of estimation for a nonparametric regression model when samples are drawn with covariate shift.

### 2.1. Observation model

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 $f^\star : \mathcal{X} \to \mathbf{R}$ is $(\beta, L)$-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

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

where $n = n_P + n_Q$.

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

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

Note that by construction, the variables $\xi_i$ are (conditionally) centered. Assumption 3 also allows $\xi_i$ 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. It is achieved by using a classical and simple method for nonparametric estimation, namely the Nadaraya-Watson estimator (Nadaraya, 1964; Watson, 1964), 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 $\rho_h$.

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

$$\widehat{f}(x) := \frac{\sum_{i=1}^n Y_i \mathbf{1}\{X_i \in \mathsf{B}(x, h_n)\}}{\sum_{i=1}^n \mathbf{1}\{X_i \in \mathsf{B}(x, h_n)\}}, \quad \text{for } x \in \mathcal{G}_n.$$

For $x \notin \mathcal{G}_n$, we set $\widehat{f}(x) := 0$.

We now state 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*

$$\mathbf{E}\left\|\widehat{f} - f^\star\right\|_{L^2(Q)}^2 \leqslant c_u \Big\{ L^2 h_n^{2\beta} + \frac{\|f^\star\|_\infty^2 + \sigma^2}{n} \rho_{h_n}(\mu_n, Q) \Big\}, \tag{3}$$

*where $c_u > 0$ is a numerical constant.*

See Section 4.1 for a proof of this result.

Note that the bound (3) exhibits a type of bias-variance tradeoff, 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 $\mu_n$ and $Q$.

**No covariate shift:** As a sanity check, it is worth checking that the bound (3) 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 (3) 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. Some consequences of Theorem 1

In order to better understand the bias-variance tradeoff in the bound (3) in the presence of covariate shift, it is helpful to derive some explicit consequences for a particular function class $\mathcal{F}$, along with certain families of source-target pairs $(P, Q)$. 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 family

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

The additional constraint $f(0) = 0$ ensures that this class has finite metric entropy.

**$\alpha$-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]

$$\mathcal{D}(\alpha, C) := \Big\{ (P, Q) \mid \sup_{0 < h \leqslant 1} h^{\alpha} \rho_h(P, Q) \leqslant C \Big\}. \quad (4a)$$

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

$$\mathcal{D}'(\alpha, C) := \Big\{ (P, Q) \mid \sup_{0 < h \leqslant 1} h^{\alpha} \rho_h(P, Q) \leqslant C, \sup_{0 < h \leqslant 1} \rho_h(Q, Q) \leqslant C \Big\}, \quad (4b)$$

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 (Stone, 1982; Tsybakov, 2009). 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 $n, n_P, n_Q, \sigma^2$, 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*

$$\inf_{\widehat{f}} \sup_{f^\star \in \mathcal{F}(\beta, L)} \mathbf{E} \big\| \widehat{f} - f^\star \big\|^2_{L^2(Q)} \leqslant c_u' \Big\{ \big(\frac{n_P}{\sigma^2}\big)^{\frac{2\beta+1}{2\beta+\alpha}} + \big(\frac{n_Q}{\sigma^2}\big) \Big\}^{-\frac{2\beta}{2\beta+1}}, \quad (5a)$$

*for any $(P, Q) \in \mathcal{D}(\alpha, C)$.*

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

$$\inf_{\widehat{f}} \sup_{f^\star \in \mathcal{F}(\beta, L)} \mathbf{E} \big\| \widehat{f} - f^\star \big\|^2_{L^2(Q)} \leqslant c_u' \Big\{ \big(\frac{n_P}{\sigma^2}\big)^{\frac{2\beta}{2\beta+\alpha}} + \big(\frac{n_Q}{\sigma^2}\big) \Big\}^{-1} \quad (5b)$$

*for any $(P, Q) \in \mathcal{D}'(\alpha, C)$.*

See Section A.3 for a proof of this corollary.

[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 $C \geqslant 1$.

Let us discuss the bound (5a) 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 $\alpha$ 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 $\alpha$ diverges to $+\infty$.

There are many papers in the literature 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*

$$n_{\text{eff}}(b) := \frac{n_P}{b} + n_Q. \quad (6)$$

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

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

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 (Tsybakov, 2009).

### 2.4. Matching lower bounds

Thus far, we have seen that the similarity measure $\rho_h$ 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 $\rho_h$. To illustrate this, we state our lower bounds for the case $\mathcal{X} = [0, 1]$, along with the usual metric.

More precisely, we prove 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 $\alpha < 1$.

We remark briefly on higher-dimensional lower bounds. Indeed, it is possible to extend our lower bounds to dimensions $d \geqslant 1$; however, our lower bounds will match our upper bounds only in the case $\alpha \geqslant d$. As can be seen from our construction when $d = 1$, we need a separate argument for the case $\alpha < 1$. The reason is that when $\alpha < d$, the form of covariate shift is actually quite favorable, meaning that it is possible to provide rates of estimation which are faster than worst-case rates under no covariate shift. We conjecture that our upper bound is tight for all $\alpha > 0, d = 1$, as we Theorem 2 demonstrates in the case $d = 1$; we leave the determination of minimax lower bounds in the case $\alpha < d, d > 1$ for future work.

**Theorem 2.** *Suppose that Assumptions 2 and 3 hold. Then there is a constant $c_\ell > 0$, independent of $n, n_P, n_Q, \sigma^2$, 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*

$$\inf_{\widehat{f}} \sup_{f^\star \in \mathcal{F}(\beta, L)} \mathbf{E} \left\| \widehat{f} - f^\star \right\|_{L^2(Q)}^2 \geqslant c_\ell \left\{ \left(\frac{n_P}{\sigma^2}\right)^{\frac{2\beta+1}{2\beta+\alpha}} + \left(\frac{n_Q}{\sigma^2}\right) \right\}^{-\frac{2\beta}{2\beta+1}}. \tag{7a}$$

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

$$\inf_{\widehat{f}} \sup_{f^\star \in \mathcal{F}(\beta, L)} \mathbf{E} \left\| \widehat{f} - f^\star \right\|_{(Q)}^2 \geqslant c_\ell \left\{ \left(\frac{n_P}{\sigma^2}\right)^{\frac{2\beta}{2\beta+\alpha}} + \left(\frac{n_Q}{\sigma^2}\right) \right\}^{-1}. \tag{7b}$$

See Sections 4.2 and A.6 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 $\rho_h$.

## 3. Properties of the similarity measure

In the previous sections, we have seen that the similarity measure $\rho_h$ 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 $\rho_h$ via covering numbers

We start with a general way of controlling the similarity measure $\rho_h$, 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 (Wainwright, 2019) 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*

$$P(\mathsf{B}(x, h)) \geqslant \lambda \, Q(\mathsf{B}(x, h)) \quad \textit{for all } x \in \mathcal{X}. \tag{8}$$

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

See Section 4.3 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 (8).

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 numbers (see chapter 5 of (Wainwright, 2019))—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 $\rho_h(P, Q) \leqslant b N(h/2)$.

As noted previously, our work was inspired by the transfer exponent introduced by Kpotufe and Martinet (Kpotufe & Martinet, 2021) 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 $\rho_h$ is a refinement of the transfer exponent. In order to simplify this discussion, we focus here on the special case $\mathcal{X} = [0, 1]$.

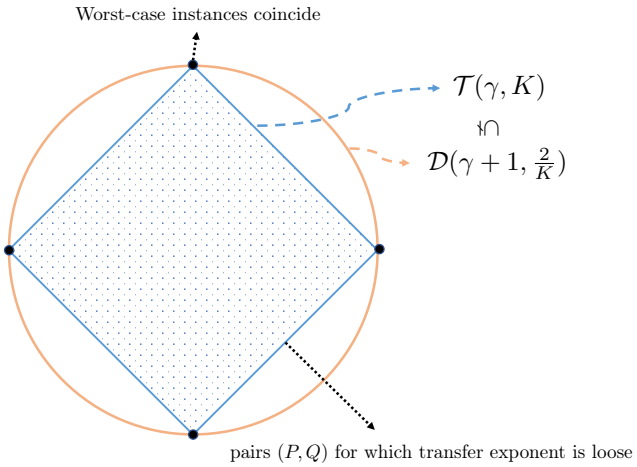


**Figure 1.** The yellow circle depicts the contour for the class $\mathcal{D}(\gamma+1, \frac{2}{K})$, while the blue square plots the contour for the class $\mathcal{T}(\gamma, K)$. It can be seen from Lemma 1 and Example 5 that $\mathcal{T}(\gamma, K)$ is strict subset of $\mathcal{D}(\gamma + 1, \frac{2}{K})$. 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.

We begin by providing the definition of transfer exponent:

**Definition 3.1** (Transfer exponent (Kpotufe & Martinet, 2021))**.** The distributions $(P, Q)$ have transfer exponent $\gamma \geqslant 0$ with constant $K \in (0, 1]$ if

$$P(\mathsf{B}(x, h)) \geqslant Kh^{\gamma} Q(\mathsf{B}(x, h)),$$

for all $x$ 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 $\alpha$-family previously defined in equation (4a). 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*

$$\mathcal{T}(\gamma, K) \subset \mathcal{D}(\gamma + 1, \tfrac{2}{K}). \tag{9}$$

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

From the inclusion (9), we see that any covariate shift instance $(P, Q)$ with finite transfer exponent $\gamma \geqslant 0$ will also have a similarity measure with a polynomial scaling with order $\gamma + 1$. In fact, following a proof similar to that of Theorem 2, we can show that the minimax risk over the class $\mathcal{T}(\gamma, K)$ is $n_P^{-\frac{2\beta}{2\beta+\gamma+1}}$ for a $\beta$-Hölder function, which coincides with that of the class $\mathcal{D}(\gamma+1, \frac{2}{K})$. In other words, from a worst-case point of view, $\mathcal{T}(\gamma, K)$ is equally hard as $\mathcal{D}(\gamma + 1, \frac{2}{K})$ for learning under covariate shift. However, it is worth pointing out that for a wide family of covariate shift instances, the transfer exponent fails to capture the fundamental difficulty of the problem. Let us consider a concrete example to illustrate.

**Example 4** (Separation between transfer exponent and $\rho_h$)**.** Let the target distribution $Q$ be a uniform distribution on the interval $[0, 1]$, and for some $\nu \geqslant 1$, suppose that the source distribution $P$ has density $p(x) = (\nu + 1)x^{\nu}$ for $x \in [0, 1]$. With these definitions, it can be verified that $(P, Q) \in \mathcal{T}(\nu, K)$ for some constant $K \in (0, 1]$, and moreover, that the quantity $\nu$ is the *smallest possible* transfer exponent for this pair. In contrast, another direct computation shows that the pair $(P, Q)$ belongs to the class $\mathcal{D}(\nu, C')$ for some constant $C' > 0$; note that as shown by our theory, the difficulty of estimation over $\mathcal{D}(\nu, C')$ is much smaller than that prescribed by $\mathcal{T}(\nu, K)$. Indeed, if one observe $n$ samples from the source distribution, the worst-case rate indicated by the computation from the transfer exponent is $n^{-\frac{2\beta}{2\beta+\nu+1}}$, whereas our rate—induced by computing the more fine-grained similarity measure $\rho_h$—is $n^{-\frac{2\beta}{2\beta+\nu}}$, which is significantly smaller when $n$ is large.

Taking Lemma 1 and the example above collectively, one sees that the new similarity measure is a strictly better characterization than the transfer exponent. See also Figure 1 for an illustration of the connections and differences between the new similarity measure and the transfer exponent.

## 4. Proofs

In this section, we collect proofs for the main results in this paper.

### 4.1. Proof of Theorem 1

Our proof makes use of the conditional expectation of the estimator given the covariates

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

Above, the expectation is over $Y_i \mid X_i$, $i = 1, \ldots, n$. In particular, we have the following result which bounds the conditional bias and variance.

**Lemma 2.** *The Nadaraya-Watson estimator $\widehat{f}$ satisfies the following guarantees for each $x \in \mathcal{X}$, almost surely:*

(a) $\left(\overline{f}(x) - f^{\star}(x)\right)^2 \leqslant \|f^{\star}\|_{\infty}^2 \mathbf{1}\{x \notin \mathcal{G}_n\} + L^2 h_n^{2\beta}$

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

We can now prove Theorem 1. Fix $x \in \mathcal{X}$. Note that by a conditioning argument, and after applying bounds (i) and (ii) from Lemma 2, we obtain

$$\mathbf{E}\left[(\widehat{f}(x) - f^\star(x))^2\right] \leqslant \left\{\|f^\star\|_\infty^2 \, \mathbf{E}[\mathbf{1}\{x \notin \mathcal{G}_n\}] + L^2 h_n^{2\beta}\right\} + \mathbf{E}\left[\mathbf{1}\{x \in \mathcal{G}_n\} \frac{\sigma^2}{\sum_{i=1}^n \mathbf{1}\{X_i \in \mathsf{B}(x,h_n)\}}\right] \quad (10)$$

By independence, note that for any $x \in \mathcal{X}$,[3]

$$\mathbf{E}\left[\mathbf{1}\{x \notin \mathcal{G}_n\}\right] = \Big(1 - P\big(\mathsf{B}(x,h_n)\big)\Big)^{n_P} \Big(1 - Q\big(\mathsf{B}(x,h_n)\big)\Big)^{n_Q} \leqslant \frac{1}{n\,\mu_n(\mathsf{B}(x,h_n))}. \quad (11)$$

Applying Lemma 6, it follows that for $x \in \mathcal{X}$,

$$\mathbf{E}\left[\frac{\mathbf{1}\{x \in \mathcal{G}_n\}}{\sum_{i=1}^n \mathbf{1}\{X_i \in \mathsf{B}(x,h_n)\}}\right] \leqslant \frac{4}{n\,\mu_n(\mathsf{B}(x,h_n))}. \quad (12)$$

Applying inequalities (11) and (12) in bound (10), we obtain

$$\mathbf{E}\left[\big(\widehat{f}(x) - f^\star(x)\big)^2\right] \leqslant L^2 h_n^{2\beta} + \frac{4\sigma^2 + L^2}{n} \frac{1}{\mu_n(\mathsf{B}(x,h_n))}. \quad (13)$$

Applying Fubini's theorem, we obtain

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

Applying inequality (13) to the integrand above yields the result. The proof of the pointwise bounds of Lemma 2 are given in section A.4.

### 4.2. 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 (7a).

[3] We use the elementary inequalities $(1-p)^n(1-q)^m \leqslant \exp(-(np+mq)) \leqslant \frac{1}{np+mq}$, valid for $p, q \in (0,1)$ and nonnegative integers $n, m$.

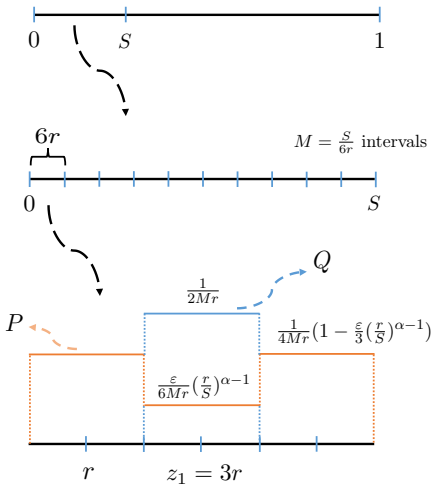


**Figure 2.** An illustration of the distributions $(P, Q)$ constructed as a hard pair in our lower bound.

**Construction of hard pair of distributions.** Let $S, r \in (0, 1]$ Let $M = \frac{S}{6r}$. Define the intervals for $j \in [M]$,

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

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 $\varepsilon$ 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 $\varepsilon, S$:*

(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 A.1 for the proof of this claim.

**Construction of hard regression functions.** Now we move on to construct a packing set of $\mathcal{F}(\beta, L)$. Let $\Psi\colon [-1, 1] \to \mathbf{R}$ be such that $\Psi(-1) = \Psi(1) = 0$ and

$$\big|\Psi(x) - \Psi(y)\big| \leqslant |x-y|^\beta, \quad \text{for all } x, y \in [-1,1], \quad \text{and,} \tag{14a}$$

$$\int_{-1}^{1} \Psi^2(x)\, \mathrm{d}x =: C_\Psi^2 > 0. \tag{14b}$$

Many choices of $\Psi$ are possible above (Tsybakov, 2009); we require $C_\Psi^2 \leqslant 1/6$, which is possible by taking $\Psi(x) = e^{-1/(1-x^2)}\mathbf{1}\{|x| \leqslant 1\}$. For a sequence $b = (b_1, \ldots, b_M) \in \{0,1\}^M$, we define

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

We will take

$$\mathcal{H} := \Big\{ f_b \mid b \in \mathcal{B} \Big\}.$$

Above, $\mathcal{B}$ is a packing set of the discrete cube $\{0,1\}^M$, originally constructed by Gilbert (Gilbert, 1952) and Varshamov (Varshamov, 1957). The following result records the main property of this set.

**Lemma 3** (Gilbert-Varshamov (Tsybakov, 2009))**.** *Let $M \geqslant 8$. There is a subset $\mathcal{B} \subset \{0,1\}^M$ such that $\|b - b'\|_1 \geqslant M/8$ for all distinct $b, b' \in \mathcal{B}$, and $|\mathcal{B}| \geqslant 2^{M/8}$.*

The next result summarizes the important properties of the hard set of regression functions, $\mathcal{H}$.

**Lemma 4.** *The set $\mathcal{H}$ has the following properties:*

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

(b) *it has the following separation: for each distinct $f, g \in \mathcal{H}$, $\|f - g\|_{L^2(Q)}^2 \geqslant \frac{C_\Psi^2}{16} L^2 r^{2\beta}$;*

(c) *it satisfies the following $L^2(P)$ and $L^2(Q)$ bounds:*

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

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

**Applying Fano's method.** To prove the claim we first introduce some notation. We denote, for a function $f \in \mathcal{H}$, by $\mu_f$ the distribution of the sequence $\{(X_i, Y_i)\}_{i=1}^n$. For the lower bound, we restrict to Gaussian noises, and select $\xi_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, \sigma^2)$. Note that this satisfies Assumption 3. Given this, note that using standard properties of the KL divergence, we obtain

$$\begin{aligned} D_{\mathrm{kl}}(\mu_f \parallel \mu_g) &= \frac{1}{2\sigma^2}\Big(n_P \|f - g\|_{L^2(P)}^2 + n_Q \|f - g\|_{L^2(Q)}^2\Big) \\ &\leqslant \frac{2}{\sigma^2}\Big(n_P \max_{f \in \mathcal{H}} \|f\|_{L^2(P)}^2 + n_Q \max_{f \in \mathcal{H}} \|f\|_{L^2(Q)}^2\Big). \end{aligned}$$

After applying part (c) of Lemma 4, we obtain

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

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

$$D_{\mathrm{kl}}(\mu_f \parallel \mu_g) \leqslant M/32. \tag{15}$$

A standard reduction to hypothesis testing (see Chapter 15 of (Wainwright, 2019)) and part (a), gives the lower bound

$$\frac{\min_{f \neq g \in \mathcal{H}} \|f - g\|_{L^2(Q)}^2}{4}\Big\{1 - \frac{\log 2 + \max_{f,g \in \mathcal{H}} D_{\mathrm{kl}}(\mu_f \parallel \mu_g)}{\log |\mathcal{H}|}\Big\}$$

Thus, after applying part (b) of Lemma 4, we obtain

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

provided that $r \leqslant \frac{1}{4608} \leqslant S/192$, which happens if

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

### 4.3. Proof of Proposition 1

Set $N := N(h/2)$. There is a collection $\{z^1, \ldots, z^N\} \subset \mathcal{X}$ such that we have

$$\mathcal{X} \subset \bigcup_{j=1}^{N} \mathsf{B}(z^j, \frac{h}{2}). \tag{16}$$

Therefore, we observe that

$$\begin{aligned} \int_{\mathcal{X}} \frac{1}{P(\mathsf{B}(x,h))}\, \mathrm{d}Q(x) &\leqslant \frac{1}{\lambda}\int_{\mathcal{X}} \frac{1}{Q(\mathsf{B}(x,h))}\, \mathrm{d}Q(x) \\ &\leqslant \frac{1}{\lambda}\sum_{j=1}^{N} \int_{\mathsf{B}(z_j, h/2)} \frac{1}{Q(\mathsf{B}(x,h))}\, \mathrm{d}Q(x). \end{aligned} \tag{17}$$

Above, the final inequality follows by the inclusion (16). 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 implies that

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

for each $j \in [N]$. Applying the display above to the bound (17), we obtain the result.

## 5. Discussion

In this paper, we introduced a new similarity measure $\rho_h$, and used it to characterize the minimax rate of estimation for a standard Hölder classes in the context of nonparametric regression and covariate shift. These results are always as good as what one can obtain from the previously suggested notion of transfer exponent, and recover standard rates for standard settings (such as the uniform distribution on $[0,1]^k$). It should also be noted that our similarity measure can be used to refine existing results for a classification setting under covariate shift with an analagous Hölder condition on the conditional class probabilities of the outcome variable given the covariate variates. Specifically, our similarity measure can be used to provide bounds that refine the results of (Kpotufe & Martinet, 2021); we mention this only in passing and do not develop those results here for the sake of brevity.

There are certainly many interesting directions to be developed in the future. For instance, one can ask whether a completely instance-dependent characterization of the minimax rate of estimation is possible. For instance, is the upper bound of Theorem 1 always optimal? Or are there instances of covariate shift for which Nadaraya-Watson is suboptimal for some Hölder continuous function? In general this may be a difficult question: even without covariate shift, there are few results that give distribution-dependent results for nonparametric regression outside of the uniform distribution and fixed-design problems.

## References


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

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

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

Chao, M. T. and Strawderman, W. E. Negative moments of positive random variables. *Journal of the American Statistical Association*, 67(338):429–431, 1972. doi: 10.1080/01621459.1972.10482404.

Cortes, C., Mohri, M., and Medina, A. M. Adaptation based on generalized discrepancy. *The Journal of Machine Learning Research*, 20(1):1–30, 2019.

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

Gilbert, E. A comparison of signalling alphabets. *Bell Systems Technical Journal*, 31:504–522, 1952.

Halász, G. Statistical interpolation. *Mat. Lapok*, 23:71–87 (1973), 1972. ISSN 0025-519X.

Hassan, A., Damper, R., and Niranjan, M. On acoustic emotion recognition: Compensating for covariate shift. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(7):1458–1468, 2013. doi: 10.1109/TASL.2013.2255278.

Ibragimov, I. A. and Khas′ minskiĭ, R. Z. Nonparametric regression estimation. *Dokl. Akad. Nauk SSSR*, 252(4):780–784, 1980. ISSN 0002-3264.

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

Kpotufe, S. and Martinet, G. Marginal singularity and the benefits of labels in covariate-shift. *Ann. Statist.*, 49(6):3299–3323, 2021. ISSN 0090-5364. doi: 10.1214/21-aos2084. URL https://doi.org/10.1214/21-aos2084.

Lei, Q., Hu, W., and Lee, J. Near-optimal linear regression under distribution shift. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6164–6174. PMLR, 18–24 Jul 2021.

Li, Y., Kambara, H., Koike, Y., and Sugiyama, M. Application of covariate shift adaptation techniques in brain–computer interfaces. *IEEE Transactions on Biomedical Engineering*, 57(6):1318–1324, 2010. doi: 10.1109/TBME.2009.2039997.

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

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

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

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

Nadaraya, E. A. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964.

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. *Dataset Shift in Machine Learning*. The MIT Press, 2009. ISBN 0262170051.

Saenko, K., Kulis, B., Fritz, M., and Darrell, T. Adapting visual category models to new domains. In Daniilidis, K., Maragos, P., and Paragios, N. (eds.), *Computer Vision – ECCV 2010*, pp. 213–226, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-15561-1.

Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *J. Statist. Plann. Inference*, 90(2):227–244, 2000. ISSN 0378-3758. doi: 10.1016/S0378-3758(00)00115-4. URL https://doi.org/10.1016/S0378-3758(00)00115-4.

Stone, C. J. Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, 10(4):1040–1053, 1982. ISSN 0090-5364.

Sugiyama, M., Nakajima, S., Kashima, H., Buenau, P. V., and Kawanabe, M. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in neural information processing systems*, pp. 1433–1440, 2008.

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

Tsybakov, A. B. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009. ISBN 978-0-387-79051-0. doi: 10.1007/b13794. URL https://doi.org/10.1007/b13794. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

Varshamov, R. Estimate of the number of signals in error correcting codes. *Doklady Akad. Nauk SSSR*, 117:739–741, 1957.

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

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

# A. Deferred results

## A.1. Proof of Proposition 2

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

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

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 that $x$ belongs to. Suppose that $0 \leqslant h \leqslant 4r$. In this case, $\mathsf{B}(x,h) \subset I_j$, and consequently,

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

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

On the other hand, now suppose $4r \leqslant h \leqslant S$. Then, we have $\mathsf{B}(x,h) \supset I_j$. 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,

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

Above (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 $Q$. The second inequality (iv) follows since $P(I_j) = Q(I_j)$, and $x \mapsto \frac{x}{x+2}$ is increasing on $\{x \geqslant 1\}$.

Therefore, combining inequalities (19) and (20), we conclude that for every $x$ in the support of $Q$,

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

The final inequality follows since $\alpha \geqslant 1$. Since $h > 0$ was arbitrary, this establishes bound (18) and completes the proof.

## A.2. Proof of Lemma 1

Note that by assumption

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

Note that there exists a collection of $N := \lceil 1/h \rceil$ balls with centers $z^1, \ldots, z^N$ of radius $h/2$ that cover the interval $[0,1]$. Therefore,

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

The final inequality is due to the inclusion $\mathsf{B}(x,h) \supset \mathsf{B}(z_j,h/2)$. Define $g(t) := \lceil t \rceil / t$. Since $g(t) \leqslant 2$ whenever $t \geqslant 1$, we obtain

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

Passing to the supremum over $h \in (0,1]$, we obtain the result.

### A.3. Proof of Corollary 1

Let $\xi = \mathbf{1}\{\alpha \geqslant 1\}$. Consider $h \in (0,1]$.

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

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

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

Then we can take

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

This is valid, since $\sigma^2 \geqslant L^2$ and we can have $\max\{n_P, n_Q\} \geqslant 4\sigma^2$. Evaluating this in our risk bound, we obtain the result.

### A.4. Proof of conditional pointwise bounds

We give the proof of Lemma 2, used to establish the performance guarantee for the NW estimator.

*Proof of Lemma 2.* By definition, we have

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

This implies that

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

Bound (a) now follows immediately. Above, (i) follows from Jensen's inequality and (ii) makes use of Assumption 1. For bound (b), 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. Above, inequality (iii) follows from Assumption 3. □

### A.5. Proof of Lemma 4

Fix $b \in \{0,1\}^M$. Note that each $\phi_j$ has disjoint support (specifically, the interval $I_j$). Additionally since $\Psi$ satisfies the continuity condition (14a), it follows that $\phi_j$ is $(\beta, L)$-Hölder. Finally, as $f_\varepsilon(0) = 0$, it follows that $f_\varepsilon \in \mathcal{F}(\beta, L)$, as required. To prove (b), let $b, b' \in \mathcal{B}$ be distinct. Note that

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

Above, (i) follows from the definition of $Q$ along with the disjointedness of the supports of $\phi_j$. The relation (ii) follows from (14b) and the fact that $b, b' \in \mathcal{B} \subset \{0,1\}^M$. Finally (iii) follows from Lemma 3. Lastly, to prove (c), fix $b \in \mathcal{B}$. Following the manipulations above, for $\mu \in \{P, Q\}$, we have by symmetry

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

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

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

Thus $\|f_b\|_{L^2(P)}^2 \leqslant \varepsilon L^2 r^{2\beta+\alpha-1}/(6S^{\alpha-1})$.

### A.6. Proof of Theorem 2(a)

Since $\mathcal{D}'(\alpha, 1) \subset \mathcal{D}'(\alpha, C)$, we use the case $C = 1$ to prove the result.

**Construction of hard distributions.** Let $Q = \delta_1$, and we set $P_\alpha$ to have Lebesgue density

$$p_\alpha(x) := \alpha(1-x)^{\alpha-1} \mathbf{1}\{x \in [0,1]\}.$$

Then, for $h \in (0, 1]$,

$$\rho_h(P_\alpha, Q) = \frac{1}{P_\alpha(B(1,h))} = h^{-\alpha}.$$

This implies $(P_\alpha, Q) \in \mathcal{D}'(\alpha, 1)$. In the rest of the argument we denote $P := P_\alpha$ to lighten notation.

**Construction of two point alternative.** If the regression function is $f$, we denote the resulting distribution of $\{(X_i, Y_i)\}_{i=1}^n$ by $\mu_f$. We consider the two point alternatives $\{f_t, g\}$ with $g \equiv 0$ and $f_t(x) := L(x-t)_+^\beta$. The next result, whose proof is found in section A.7, demonstrates the validity of this choice of alternatives.

**Lemma 5.** *The function $f_t$ lies in $\mathcal{F}(\beta, L)$ for any $t \in [0,1]$.*

Note that for any $t \in [0,1]$, $\|f_t\|_{L^2(Q)}^2 = L^2(1-t)^{2\beta}$. Additionally,

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

**Applying Le Cam's method.** By Le Cam's two point method (Tsybakov, 2009), we have

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

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

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

We will take

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

This assures $D_{\mathrm{kl}}(\mu_{f_t} \parallel \mu_g) \leqslant 2$, and thus we obtain the result.

### A.7. Proof of Lemma 5

Clearly $f_t(0) = 0$. To prove the claim it suffices to show that

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

To show this, fix $0 \leqslant t < x \leqslant 1$. Define

$$\phi_x(y) := L(y^\beta - x^\beta) - L(y-x)^\beta.$$

Differentiating, $\phi_x'(y) = L\beta(y^{\beta-1} - (y-x)^{\beta-1})$. For $y \geqslant x$, it follows that $\phi_x'(y) \leqslant 0$ (since $\beta \leqslant 1$), and thus $\phi_x(y) \leqslant \phi_x(x) = 0$. This proves the claim.

**Lemma 6.** *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.* Note first that on $\{U+V>0\}$, we certainly have

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

Consequently,

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

The penultimate inequality is a consequence of known results for Binomial random variables (see equation (3.4) in (Chao & Strawderman, 1972)). □