Marginal Singularity, and the Benefits of Labels in Covariate-Shift

Samory Kpotufe[*]   SAMORY@PRINCETON.EDU
ORFE, Princeton University

Guillaume Martinet[†]   GGM2@PRINCETON.EDU
ORFE, Princeton University

Abstract

We present new minimax results that concisely capture the relative benefits of source and target labeled data, under covariate-shift. Namely, we show that, in general classification settings, the benefits of target labels are controlled by a *transfer-exponent* γ that encodes how *singular Q is locally* w.r.t. P, and interestingly allows situations where transfer did not seem possible under previous insights. In fact, our new minimax analysis – in terms of γ – reveals a *continuum of regimes* ranging from situations where target labels have little benefit, to regimes where target labels dramatically improve classification. We then show that a recently proposed semi-supervised procedure can be extended to adapt to unknown γ, and therefore requests target labels only when beneficial, while achieving nearly minimax transfer rates.

Keywords: Transfer learning, covariate-shift, nonparametric classification, nearest-neighbors.

1. Introduction

Transfer learning addresses the many practical situations where much labeled data is available from a *source* distribution P, but relatively little labeled data is available from a *target* distribution Q. The aim is to harness source data to improve prediction on the target Q, assuming the source P is informative about Q. Naturally, a main theoretical question is in understanding relations (or divergences) between P and Q that allow information transfer, and in particular, that tightly characterize the relative benefits of source and target labeled samples (towards informing practice).

We focus on nonparametric classification, i.e., predicting labels Y of future X drawn from Q, with minimal assumptions on P and Q. The most common setting is that of *covariate-shift* where $P_{Y|X} = Q_{Y|X}$, but Q_X may differ from P_X. While equal conditionals may seem restrictive, it is well motivated by common applications of transfer (e.g. image, speech, or document classification). The question is then how to express the changes in marginals P_X, Q_X in the context of transfer.

We present new minimax results that concisely capture the relative benefits of source and target labeled data, under covariate-shift. Namely, we show that the benefits of target labels are controlled by a *transfer-exponent* γ that encodes how *singular Q is locally* w.r.t. P, and interestingly allows situations where transfer did not seem possible under previous insights. In fact, our new minimax analysis – in terms of γ – reveals a *continuum of regimes* ranging from situations where target labels have little benefit, to regimes where target labels dramatically improve classification.

The notion of transfer-exponent follows a natural intuition, present in the literature, that transfer is hardest if P_X does not properly cover regions of large Q_X mass. In particular, γ parametrizes

---

the behavior of ball-mass ratios $Q(B(x, r))/P(B(x, r))$ as a function of neighborhood size r (see Definition 3), namely, that these ratios behave like $r^{-\gamma}$. We will see, through both lower and upper-bounds, that transfer is easiest as $\gamma \to 0$ and hardest as $\gamma \to \infty$.

Interestingly, γ is well defined even when Q is singular w.r.t. P – in which case common notions of *density-ratio* and information-theoretic divergences (KL or Renyi) fail to exist, and common extensions of total-variation can be too large to characterize transfer. We note that singularity of Q w.r.t. P can often be the case in practice where high-dimensional data is often very structured, and transfer often involves going from a generic set of data from a domain P to a more structured subdomain Q. Here, our results can directly inform practice: target labels yield greater performance with lower-dimensional Q, but are not necessary; if Q were of higher dimension than P, the benefits of source labels quickly saturate. Now when Q and P are of the same dimension, even sharing the same support, the notion of γ reveals yet a rich set of regimes where transfer is possible at different rates, while usual notions of task-relatedness might indicate otherwise.

As alluded to above, a practical question motivating much of this work, is whether, given a large database of source data, acquiring additional target data might further improve classification; this is usually difficult to test given the costs and unavailability of target data. Here, by capturing the interaction of source and target sample sizes in our rates, in terms of γ, we can sharply characterize those sampling regimes where target or source data are most beneficial. We then show that it is in fact possible to *adapt* to unknown γ, i.e., request target labels only when beneficial, while also attaining nearly optimal classification rates in terms of unknown distributional parameters.

Detailed Results and Related Work

Many interesting notions of divergence have been proposed that successfully capture a general sense of when transfer is possible. In fact, the literature on transfer is by now expansive, and we cannot hope to truly do it justice.

A first line of work considers refinements of total-variation that encode changes in error over the classifiers being used (as defined by a hypothesis class $\mathcal{H}$). The most common such measures are the so-called $d_{\mathcal{A}}$-divergence (Ben-David et al., 2010a,b; Germain et al., 2013) and $\mathcal{Y}$-discrepancy (Mansour et al., 2009a; Mohri and Medina, 2012; Cortes et al.). These notions are the first to capture – through *differences* in mass over space – the intuition that transfer is easiest when P has sufficient mass in regions of substantial Q-mass. Typical excess-error bounds on classifiers learned from source (and some or no target) data are of the form $o_p(1) + C \cdot \text{divergence}(P, Q)$. In other words, transfer seems impossible when these divergences are large; this is certainly the case in very general situations. However, as we show, there are ranges of reasonable situations ($0 \le \gamma < \infty$) where transfer is possible, even at fast rates (while using only source data), yet the above divergences remain large (see Remark 1 of Section 2.3). Also, interestingly, such divergences are symmetric for pairs (P, Q), while our notion of γ is not, attesting to the fact that transfer might be possible from P to Q, while hard from Q to P.

Another prominent line or work, which has led to many practical procedures, considers so-called ratios of densities f_Q/f_P or similarly Radon-Nikodym derivatives dQ/dP as a way to capture the similarity between P and Q (Quionero-Candela et al., 2009; Sugiyama et al., 2012). It is often assumed in such work that dQ/dP is bounded which corresponds to the regime $\gamma = 0$ in our case (see Example 2 of Section 2.3). Typical excess-error bounds are dominated by the estimation rates for dQ/dP (see e.g. rates for α-Hölder dQ/dP, $\alpha \to 0$, in Kpotufe (2017)), which unfortunately could

be arbitrarily higher than the minimax rates we establish for that setting with $\gamma = 0$. Furthermore, as previously mentioned, dQ/dP is inadequate in common scenarios with structured data, or can be unbounded even while γ remains small (see Example 3 of Section 2.3).

Another line of work, instead considers information-theoretic measures such as KL-divergence or Renyi divergence (Sugiyama et al., 2008; Mansour et al., 2009b). In particular, such divergences are closer in spirit to our notion of transfer-exponent γ (viewing it as roughly characterizing the log of ratios between Q_X and P_X), but are also undefined in typical scenarios with structured data.

Our upper-bound are established under the nonparametric classification settings of Audibert and Tsybakov (2007), which parametrize the noise distribution (via smoothness and noise conditions); this allows us to understand the interaction between γ and noise parameters, and capture regimes where classification remains easy despite large γ. Our upper-bounds are established with a generic k-NN classifier defined over the combined source and target sample. In particular, our results imply new convergence rates of independent interest for vanilla k-NN (see Remark 2, Section 3.2).

Our lower-bounds are established over any learner with access to both source and target samples, and interestingly, which also has access to infinite unlabeled source and target data (i.e., is allowed to know P_X and Q_X). In other words, our lower-bound imply that our rates cannot be improved with access to unlabeled data, which is often an important consideration in practice given the cost of target labels (Huang et al., 2007; Ben-David and Urner, 2012).

A related practical consideration, alluded to earlier, are those of *semisupervised* or *active* transfer, where, given unlabeled target data, the goal is to request as few target labels as possible to improve classification over using sourced data alone (Saha et al., 2011; Chen et al., 2011; Chattopadhyay et al., 2013). An early theoretical treatment can be found in (Yang et al., 2013), but which however considers a transfer setting with fixed marginal but varying conditionals (labeling functions). The recent work of Berlind and Urner (2015) gives a nice first theoretical treatment of the problem under similar nonparametric conditions as ours; however their work is less concerned with a minimax understanding of the problem, and mostly concerned with algorithmic strategies towards minimizing label requests. We will show how to extend their procedure to achieve minimax transfer rates in terms of unknown problem parameters, while requesting target labels only when necessary (as controlled by unknown γ).

Paper Outline

We start with definitions and setup in Section 2, followed by an overview of results in Section 3. We discuss the main lines of the analysis in Section 4, followed by detailed proofs in the appendix.

2. Preliminaries

2.1. Basic Distributional Setting

We consider a classification setting where the input variable X belongs to a compact metric space $(\mathcal{X}, \rho)$ of diameter $\Delta_{\mathcal{X}}$, and the label variable Y belongs to $\mathcal{Y} \equiv \{0, 1\}$. We consider a *source* distribution P and a *target* distribution Q over $\mathcal{X} \times \mathcal{Y}$. We'll let P_X and Q_X denote the corresponding marginals over $\mathcal{X}$, and $P_{Y|X}$ and $Q_{Y|X}$ denote the corresponding conditional distributions.

We work under the common *covariate-shift* setting, where marginals might shift from source to target, although conditionals remain the same. This is formalized below.
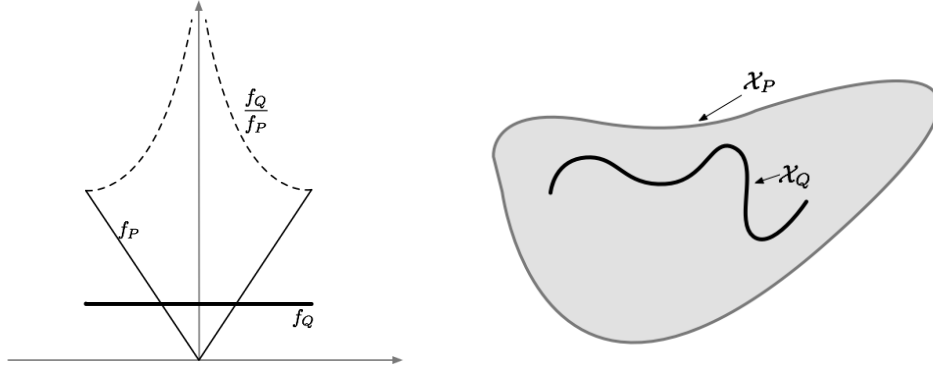
Figure 1: *Some settings with $0 < \gamma < \infty$. Left: the density $f_P \propto |x|^\gamma$ goes fast to 0, while f_Q is uniform on the same support; f_Q/f_P then diverges, but γ is well-defined. Right: Q_X has lower-dimensional support $\mathcal{X}_Q$; γ then captures the difference in dimensions. This last case also illustrates the interesting fact that transfer might be possible from P to Q but not from Q to P ($\gamma = \infty$ when P is the target).*

Definition 1 (Covariate-shift) *There exists a measurable function $\eta : \mathcal{X} \to [0,1]$, called* regression function, *such that $P_{Y|x}(1) = Q_{Y|x}(1) = \eta(x)$ a.s. P_X and Q_X.*

2.2. Classifiers under Transfer

The learner has access labeled source data $(\mathbf{X}, \mathbf{Y})_P \equiv \{(X_i, Y_i)\}_{i=1}^{n_P} \sim P^{n_P}$, and to labeled target data $(\mathbf{X}, \mathbf{Y})_Q \equiv \{(X_i, Y_i)\}_{i=n_P+1}^{n_P+n_Q} \sim Q^{n_Q}$, independent of $(\mathbf{X}, \mathbf{Y})_P$. We write the combined sample as $(\mathbf{X}, \mathbf{Y}) \equiv (\mathbf{X}, \mathbf{Y})_P \cup (\mathbf{X}, \mathbf{Y})_Q$. We only assume that $(n_P \vee n_Q) \geq 1$, although the regime $0 \leq n_Q \leq n_P$ is most meaningful in applications of transfer learning.

For any classifier $h : \mathcal{X} \to \{0,1\}$ learned over $(\mathbf{X}, \mathbf{Y})$, we are interested in the target error $\mathrm{err}_Q(h) \equiv \mathbb{E}_Q \mathbb{1}\{h(X) \neq Y\}$. This is minimized by the *Bayes* classifier $h^*(x) \equiv \mathbb{1}\{\eta(x) \geq 1/2\}$. Our results concern the best error achievable by any classifier h *in excess* over the error of h^*.

Definition 2 *The* **excess error** *of a classifier h, under target distribution Q, is defined as:*

$$\mathcal{E}_Q(h) \equiv \mathrm{err}_Q(h) - \mathrm{err}_Q(h^*) = 2\mathbb{E}_Q \left| \eta(X) - \frac{1}{2} \right| \cdot \mathbb{1}\{h(X) \neq h^*(X)\}. \qquad (1)$$

Our minimax analysis aims to upper and lower-bound $\mathcal{E}_Q(\hat{h})$ – in expectation over $P^{n_P} \times Q^{n_Q}$, over any possible learner $\hat{h}$[1], so as to capture the separate contributions of n_P and n_Q to the rates.

2.3. Transfer-exponent (from P_X to Q_X)

Intuitively, transfer is harder if we are likely to see little data from P_X near typical points $X \sim Q_X$. In other words, for easy transfer from P to Q, we want P_X to give reasonable mass to those regions of non-negligible Q_X mass. We aim to parametrize how much (P, Q) deviates from this ideal.

Let $B(x, r)$ denote the closed ball $\{x' \in \mathcal{X} : \rho(x, x') \leq r\}$. Let $\mathcal{X}_P$ denote the support of P_X, i.e., $\mathcal{X}_P \doteq \{x \in \mathcal{X} : P_X(B(x, r)) > 0, \forall r > 0\}$, and similarly define $\mathcal{X}_Q$ as the support of Q_X. Remark that because $(\mathcal{X}, \rho)$ is compact and hence separable, we have $P_X(\mathcal{X}_P) = Q_X(\mathcal{X}_Q) = 1$.

---

1. We will at times conflate the learner $\hat{h} : (\mathbf{X}, \mathbf{Y}) \mapsto 2^{\mathcal{X}}$ with its output classifier $\hat{h} \in 2^{\mathcal{X}}$ for ease of notation.

Definition 3 *We say that (P,Q) has **transfer-exponent** $\gamma \in \mathbb{R}_+ \cup \{0, \infty\}$, if there exists a constant $C_\gamma \in (0,1]$, and a region $\mathcal{X}_Q^\gamma \subset \mathcal{X}_Q$, $Q_X(\mathcal{X}_Q^\gamma) = 1$, such that:*

$$\forall x \in \mathcal{X}_Q^\gamma, \forall r \in (0, \Delta_\mathcal{X}], \quad P_X(B(x,r)) \geq Q_X(B(x,r)) \cdot C_\gamma \left(\frac{r}{\Delta_\mathcal{X}} \right)^\gamma. \tag{2}$$

First, notice that every pair (P,Q) satisfies the above with at least $\gamma = +\infty$, since the condition in (2) then just defaults to $P_X(B(x,r)) \geq 0$. Second, if (2) holds for some γ, then it holds for any $\gamma' > \gamma$; our results are therefore to be understood as holding for the smallest admissible such transfer-exponent γ. We will see that transfer-learning gets easier with smaller γ, i.e., achievable rates depend more on n_P and less on n_Q as $\gamma \to 0$. In particular for $\gamma = 0$, we need a number of target labels $n_Q \gg n_P$ to get any speedup beyond the rates achievable with $n_Q = 0$ target labels. For $\gamma = \infty$, we have nearly no transfer, i.e., n_P has little effect on achievable rates.

Next, to get a sense of the applicability of the above definition, let's consider some examples of situations with different transfer-exponents, including the boundary cases $\gamma = \infty$ and $\gamma = 0$. As it turns out, these boundary cases encompass much of the usual regimes covered by previous analyses.

Example 1 (Disjoint supports, or higher-dimensional target) *Suppose $\mathcal{X}_Q \setminus \mathcal{X}_P \neq \emptyset$. Then $\gamma = \infty$ since for any $x \in \mathcal{X}_Q \setminus \mathcal{X}_P$, $\exists r > 0$ s.t. $P(B(x,r)) = 0$ while $Q(B(x,r)) > 0$. An important such case in practice is when the support $\mathcal{X}_Q$ is of higher dimension than $\mathcal{X}_P$. As we'll see, source-labeled data have minimal benefits in such cases (beyond improving constants) as discussed above.*

Example 2 (Bounded density ratio dQ_X/dP_X) *Let Q_X be absolutely continuous w.r.t. P_X and therefore admit a density (Radon-Nikodym derivative) dQ_X/dP_X w.r.t. P_X. If $dQ_X/dP_X \leq C$, we then have $\gamma = 0$, since for any ball B we have $Q_X(B) = \int_B \frac{dQ_X}{dP_X} dP_X \leq C \cdot P_X(B)$. Arguably, this is the most studied case in transfer under covariate-shift.*

Example 3 (Unbounded density ratio dQ_X/dP_X) *Again, let Q_X admit a density dQ_X/dP_X w.r.t. P_X. However we now allow dQ_X/dP_X to diverge at some points or regions in space; the speed at which it diverges is then controlled by γ. This is a sense in which we might view γ as encoding a degree of singularity of Q_X w.r.t. P_X. Here is a concrete example (see also Figure 1):*

Let Q_X be uniform on $([-1,1], \rho \doteq |\cdot|)$ (or have bounded Lebesgue density), and let P_X have Lebesgue density $f_P(x) \propto |x|^\gamma$ on $[-1,1]$. Then $dQ_X/dP_X = 1/(2f_P)$ and diverges at $x = 0$. It is immediate that (2) holds for any ball centered at $x = 0$. It is not hard to check however that (2) holds at all $x \in [-1,1]$ since P_X can only assign higher mass away from 0.

Following the above example, we can see that $\gamma = \infty$ happens when dQ_X/dP_X diverges at a rate faster than polynomial (e.g. let $f_P(x) \propto \exp(-1/|x|)$). Such fast divergence in dQ_X/dP_X happens for instance if P_X and Q_X are sufficiently separated Gaussians, in which case transfer can be hard. In fact, the example of two Gaussians was given earlier in (Cortes et al., 2010) where it is shown that this is a case where transfer is hard for importance-sampling approaches; our present results indicate that, in a minimax sense, such situations might be hard for any approach.

Example 4 (Q_X has lower-dimension) *Suppose that Q_X has support $\mathcal{X}_Q$ of dimension d_Q, while P_X has support $\mathcal{X}_P$ of dimension $d_P \geq d_Q$ (Figure 1). In a generic metric space, this would be formalized w.r.t. the mass assigned to balls as $Q_X(B(x,r)) \propto r^{d_Q}$ while $P_X(B(x,r)) \propto r^{d_P}$ for $x \in \mathcal{X}_Q$ (see e.g. Definition 6), following similar intuition for Euclidean spaces. It is then direct that we would have $\gamma = d_P - d_Q$. This is again a sense in which γ encodes the* strength *of singularity of Q_X w.r.t. P_X. We'll then see that the smaller $d_Q \ll d_P$, the more useful target labels are.*

We remark that, in practice, γ might capture any mix of the above examples, but clearly serves to measure the *degree* to which Q_X is close to singular w.r.t. P_X (Example 3) or the *strength* of such singularity (Example 4).

Remark 1 (Divergences can be large) Common notions of dissimilarity used in transfer take the form $d_{\mathcal{A}}(Q_X, P_X) = \sup_{A \in \mathcal{A}} |Q_X(A) - P_X(A)|$, where $A \in \mathcal{A}$ are subsets of $\mathcal{X}$ encoding classification decisions (indicators over classifiers h in a fixed set $\mathcal{H}$, or their symmetric differences). For common families $\mathcal{A}$ we would have $\mathrm{div}(Q_X, P_X) \geq 1/2$ (leading to vacuous transfer rates), while we'll see that nontrivial transfer remains possible ($0 < \gamma < \infty$). This will be the case for instance when $\mathcal{X}_Q$ is of lower-dimension than $\mathcal{X}_P$ as in Example 4 above: suppose for instance that P_X is uniform on a cube $[0,1]^{d_P}$, and Q_X is uniform on a hyperplane through the cube; if $\mathcal{A}$ is all half-spaces (encoding linear separators or their symmetric differences) it's then clear that $\mathrm{div}(Q_X, P_X) \geq 1/2$ while $\gamma = 1$. In fact, even when P and Q have the same dimension and same support (as in Example 2 or 3), we can construct similar situations where $\mathrm{div}(Q_X, P_X)$ is large, simply by assigning different masses to an appropriately chosen $A \in \mathcal{A}$, while allowing small γ. Furthermore, such situations extend to similar notions such as $\mathcal{Y}$-discrepancy, as the main issue is in that they consider *supremum differences* in mass.

Information-theoretic divergences (Renyi or KL) seem related to γ if not only for the fact that γ serves to characterize the behavior of $\log Q_X(B(x,r))/P_X(B(x,r))$ as $r \to 0$. In particular, for the specific choices of distributions in Examples 2, 3 above, it's easy to check that $\mathcal{D}_{\mathrm{kl}}(Q_X|P_X)$ remains small with small γ, and diverges for the examples with $\gamma = \infty$. However, the exact relations between $\mathcal{D}_{\mathrm{kl}}(Q_X|P_X)$ and γ (when dQ_X/dP_X exists) remain unclear and worth further study. Nonetheless, the notion of γ captures more general situations, as it remains well-defined even when Q_X is singular w.r.t. P_X while Renyi or KL divergences fail to exist.

2.4. Classification Regimes

We consider transfer under two nonparametric classification regimes introduced in (Audibert and Tsybakov, 2007). Both regimes similarly parametrize the behavior of $\eta(x) = \mathbb{E}[Y|x]$ near the boundary $1/2$, but differ in their regularity assumptions on Q_X, i.e., in whether Q_X properly covers its support $\mathcal{X}_Q$ or not. These regimes capture the hardness of classification w.r.t. Q_X, while the transfer-exponent γ of earlier, captures the hardness of transfer from P to Q.

SMOOTHNESS OF η AND LOW NOISE CONDITIONS

Definition 4 (Smoothness) *The* regression function η is (C_α, α)–Hölder for $\alpha \in (0, 1]$, $C_\alpha > 0$, if $\forall x, x' \in \mathcal{X}, \quad |\eta(x) - \eta(x')| \leq C_\alpha \cdot \rho(x, x')^\alpha$.

Next, we characterize how likely η is to be close to $1/2$ under Q_X.

Definition 5 (Tsybakov's noise condition for Q) Q *has noise parameters* $\beta, C_\beta > 0$, *if* $\forall t \geq 0$, $Q_X \left(0 < \left| \eta(X) - \frac{1}{2} \right| \leq t \right) \leq C_\beta t^\beta$. *The larger* β, *the easier the classification task.*

Note that the above always hold for any Q with at least $\beta = 0$ and $C_0 = 1$.

DEFINITIONS OF REGIMES, AND DIMENSION OF $X \sim Q_X$

We now present the two classification regimes. The first regime (DM), ensures that Q_X has near *uniform* mass, and corresponds to the *strong-density* condition of Audibert and Tsybakov (2007).

Definition 6 (Doubling measure) *We say that Q_X is (C_d, d)-doubling, for $C_d \in (0, 1]$ and $d \geq 1$, if $\forall r \in [0, \Delta_{\mathcal{X}}], \forall x \in \mathcal{X}_Q, \quad Q_X(B(x, r)) \geq C_d \left(r/\Delta_{\mathcal{X}}\right)^d.$*

This first regime is thus formalized as follows.

(DM) *The regression function η is (C_α, α)–Hölder, and Q has noise parameters β, C_β. Furthermore, Q_X is (C_d, d)-doubling.*

Classification is easiest in this regime, and so turns out to yield faster transfer rates. The quantity d plays the role of the *dimension* of the input $X \sim Q_X$ (think for instance of $Q_X \equiv \mathcal{U}([0, 1]^d, \ell_\infty)$).

The second regime (BCN), allows arbitrary Q_X, and therefore results in harder classification, and also slower transfer, as we will see. For this regime, the following regularity conditions (and quantity d) serve to capture the *dimension* of the support $\mathcal{X}_Q$. Recall, that the r-covering number of a pre-compact set $\mathcal{X}_Q$, denoted $\mathcal{N}(\mathcal{X}_Q, \rho, r)$, is the smallest number of ρ-balls of diameter r needed to cover $\mathcal{X}_Q$.

Definition 7 (Bounded covering number) *We say that $\mathcal{X}_Q$ has (C_d, d)-bounded covering number, for $d \geq 1$, $C_d \geq 1$, if $\forall r \in (0, \Delta_{\mathcal{X}}], \quad \mathcal{N}(\mathcal{X}_Q, \rho, r) \leq C_d \left(\Delta_{\mathcal{X}}/r\right)^d.$*

The second regime is thus formalized as follows.

(BCN) *The regression function η is (C_α, α)–Hölder, and Q has noise parameters β, C_β. Furthermore, $\mathcal{X}_Q$ has (C_d, d)-bounded covering number.*

The above two parameters, together with the transfer-exponent γ, characterize the classes of distribution tuples (P, Q) considered in this work. We have the following definition.

Definition 8 (Transfer classes) *Fix some parameters $(C_\gamma, \gamma, C_\alpha, \alpha, C_\beta, \beta, C_d, d)$ as in Definitions 3, 4, 5, 6 or 7. We call $\mathcal{T}_{(DM)}$ (resp. $\mathcal{T}_{(BCN)}$) the class of all distribution tuples (P, Q) with transfer parameters (C_γ, γ) and where Q satisfies (DM) (resp. (BCN)) for the fixed parameters.*

3. Results Overview

We start with lower-bound results (Section 3.1), and matching *oracle* upper-bounds (Section 3.2). Our adaptivity results are presented in (Section 3.3).

3.1. Minimax Lower-Bounds

As shown below, the transfer-exponent γ successfully captures the difficulty of transfer. In particular, the rates of transfer get worse with large $\gamma \in \mathbb{R}_+ \cup \{0, \infty\}$. For simplicity, we focus here on the case where d is an integer, $\mathcal{X} = [0, 1]^d$ and $\rho(x, y) = \|x - y\|_\infty$. The results do extend to general metric spaces, however with added technicality that adds little additional insight.

Theorem 1 (Lower-bounds) *Let $(\mathcal{X}, \rho) = ([0,1]^d, \|.\|_\infty)$, for some $d \in \mathbb{N}^*$. Let $\mathcal{T}$ denote either $\mathcal{T}_{(DM)}$ or $\mathcal{T}_{(BCN)}$. For $\mathcal{T} = \mathcal{T}_{(DM)}$ assume further that $\alpha\beta \leq d$. There exists a constant $c = c(\mathcal{T})$ such that, for any classifier $\hat{h}$ learned on $(\mathbf{X}, \mathbf{Y})$ and with knowledge of P_X, Q_X, we have:*

$$\sup_{(P,Q) \in \mathcal{T}} \mathbb{E}_{(\mathbf{X},\mathbf{Y})}[\mathcal{E}_Q(\hat{h})] \geq c \left(n_P^{d_0/(d_0+\gamma/\alpha)} + n_Q \right)^{-(\beta+1)/d_0},$$

where $d_0 = 2 + d/\alpha$ when $\mathcal{T} = \mathcal{T}_{(DM)}$, and $d_0 = 2 + \beta + d/\alpha$ when $\mathcal{T} = \mathcal{T}_{(BCN)}$.

Note that, for $n_P = 0$, we recover known classification lower-bounds of Audibert and Tsybakov (2007). As in that work, our lower-bounds exclude the regime $\alpha\beta > d$ for $\mathcal{T}_{(DM)}$ as it's not possible to construct nontrivial such settings (see discussion and Proposition 3.4 therein).

The main technicality in our transfer lower-bound is in dealing with two sources of randomness (P, Q), along with the constraint of keeping (P, Q) related through γ. Unlike in usual lower-bounds, the learner has access to non-identical samples, in addition to *knowing* both marginals P_X, Q_X. This brings up an interesting point: additional unlabeled data do not improve the rates of transfer.

3.2. Minimax Upper-Bounds

Our oracle upper-bounds are established through a generic k-NN classifier as defined below.

Definition 9 (k-NN) *Pick $1 \leq k \leq n_P \vee n_Q$. Fix $x \in \mathcal{X}$, and let $\{X_{(i)}\}_{i=1}^k$ denote the k nearest neighbors of x in $\mathbf{X}$ (break ties anyhow), with corresponding labels $\{Y_{(i)}\}_{i=1}^k$. Define the regression estimate $\hat{\eta}(x) \equiv \frac{1}{k} \sum_{i=1}^k Y_{(i)}$. The k-NN classifier at x is then given by $\hat{h}_k(x) \equiv \mathbb{1}\{\hat{\eta}(x) \geq 1/2\}$.*

While k-NN methods have received much renewed attention Samworth et al. (2012); Chaudhuri and Dasgupta (2014); Shalev-Shwartz and Ben-David (2014), most results concern the (DM) setting, i.e., assume near-uniform marginals. Our current analysis yields new rates under (BCN).

Remark 2 (New rates for k-NN) Theorem 2 below imply new bounds of independent interest for vanilla k-NN (by setting $n_P = 0$): namely, for (BCN) – which allows general Q_X, Gadat et al. (2014) show that the minimax rates of $n_Q^{-(\beta+1)/(2+d/\alpha+\beta)}$ are reachable by NN methods where k is chosen locally as $k(x)$, assuming a density $dQ_X(x)$ is known. Our results instead states that such optimal rates are reachable by a standard k-NN with fixed k. For context, we note that the original rates of (BCN) in (Audibert and Tsybakov, 2007) are achieved by a non-polynomial time procedure.

Theorem 2 (Upper-bounds) *Let $\mathcal{T}$ denote either $\mathcal{T}_{(DM)}$ or $\mathcal{T}_{(BCN)}$. For $\mathcal{T} = \mathcal{T}_{(BCN)}$ assume further that $\alpha < d$. There exists a constant $C = C(\mathcal{T})$ such that, for a k-NN classifier $\hat{h}_k$ we have*

$$\sup_{(P,Q) \in \mathcal{T}} \mathbb{E}_{(\mathbf{X},\mathbf{Y})}[\mathcal{E}_Q(\hat{h}_k)] \leq C \left(n_P^{d_0/(d_0+\gamma/\alpha)} + n_Q \right)^{-(\beta+1)/d_0},$$

for a choice of $k = \Theta \left(n_P^{d_0/(d_0+\gamma/\alpha)} + n_Q \right)^{2/d_0}$, where $d_0 = 2 + d/\alpha$ when $\mathcal{T} = \mathcal{T}_{(DM)}$, and $d_0 = 2 + \beta + d/\alpha$ when $\mathcal{T} = \mathcal{T}_{(BCN)}$.

For the case $\mathcal{T} = \mathcal{T}_{(BCN)}$ with $\alpha = d$, C above is replaced with $C \cdot \log(2(n_P + n_Q))$.

The bounds match those of Theorem 1 up to constants, apart for the corner case $\mathcal{T}_{\text{(BCN)}}$ with $\alpha = d = 1$ where an additional log term gets introduced. Thus, the transfer-exponent γ indeed captures the relative benefits of source and target samples. Namely, such relation is captured in the sum $(n_P^{d_0/(d_0+\gamma/\alpha)} + n_Q)$, $d_0 = d_0(\mathcal{T})$. In particular, source samples are most beneficial in the sampling-regime $n_P^{d_0/(d_0+\gamma/\alpha)} \gg n_Q$ (the rates are then of order $n_P^{-(\beta+1)/(d_0+\gamma/\alpha)}$), otherwise target samples are most beneficial (the rates then transition to $n_Q^{-(\beta+1)/d_0}$).

Hence, $n_P^{d_0/(d_0+\gamma/\alpha)}$ can be viewed as a threshold beyond which labeled target-sample yields considerable improvement in classification beyond what's possible with just source samples. Notice that this threshold decreases to 1 as $\gamma \to \infty$, in other words, even a small amount n_Q of target labels can considerably improve classification w.r.t. Q. At $\gamma = 0$, P has much information about Q, and we need a large amount $n_Q \geq n_P$ of target-samples to improve over pure transfer. Setting $n_Q = 0$, we see that transfer remains possible in a rich continuum of regimes between $\gamma = 0$ and $\gamma = \infty$ with rates of the form $n_P^{-(\beta+1)/(d_0+\gamma/\alpha)}$, including *fast* rates $o(n^{-1/2})$ for large β (low noise).

Remark 3 (Extended setting) We consider deviations from the above settings of (DM) and (BCN) in Appendix D. In particular, when the supports $\mathcal{X}_Q, \mathcal{X}_P$ don't overlap, or when $P_{Y|X}$ deviates from $Q_{Y|X}$ we can obtain the above rates plus an additive term accounting for such deviation.

3.3. Adaptive Label Requests

Unfortunately we don't know γ in practice, which begs the question of whether we can directly benefit from unknown but small γ in deciding how much target labels to sample in practice.

Setup: We consider a semi-supervised transfer setting where, initially, the learner has access to labeled source data $(\mathbf{X}_P, \mathbf{Y}_P)$ of size n_P, and unlabeled target data $\mathbf{X}_Q$ of size n_Q. The goal is to request few target labels, at most n_Q, and then return a classifier $\hat{h}$ with good target error.

This question has received recent attention (Saha et al., 2011; Chen et al., 2011; Chattopadhyay et al., 2013), with a recent theoretical treatment by Berlind and Urner (2015) which show nontrivial transfer rates in terms of n_Q. Their result does not capture the source sample size n_P, which is explained here by the fact that their conditions hold with $\gamma = \infty$. However, for general $\gamma > 0$, we can build on their approach to achieve the above oracle rates without prior knowledge of distributional parameters. Furthermore, the approach has the interesting guarantee that it only requests target labels *if* they are beneficial, i.e., if the maximum budget n_Q is not too small in terms of unknown γ; the *threshold* for label requests is now a bit larger than that the *oracle* threshold resulting from Theorem 2 (recall that target labels are less useful when $n_Q \ll n_P^{d_0/(d_0+\gamma/\alpha)}$).

The algorithmic approach of Berlind and Urner (2015) builds on the following useful concept.

Definition 10 (k-$2k$ Cover) *Let $1 \leq k \leq (n_P \vee n_Q)/2$, and let $\mathbf{X}_R$ denote samples in $\mathbf{X}$ indexed by $R \subset [n_P + n_Q]$. We say that $\mathbf{X}_R$ is a k-$2k$* **cover** *of $\mathbf{X}$ if, for any $X_i \in \mathbf{X}$, either $X_i \in \mathbf{X}_R$, or its $2k$-NN in $\mathbf{X}$ (including X_i itself) include at least k samples from $\mathbf{X}_R$. If the choice of the $2k$-NN is not unique, at least one of the possible choices must contain k samples from $\mathbf{X}_R$.*

The idea in (Berlind and Urner, 2015) is to request labels only for those points in $\mathbf{X}_R \cap \mathbf{X}_Q$. They present various ways to build such a cover, the obvious way being to start with the labeled samples, i.e., $\mathbf{X}_R = \mathbf{X}_P$, and add in points from $\mathbf{X}_Q \setminus \mathbf{X}_R$ that do not satisfy the conditions. Following this, classification then consists of a k-NN estimate over a labeled sample $(\mathbf{X}_R, \mathbf{Y}_R)$.

We modify the procedure of Berlind and Urner (2015) and construct a *cover* $\mathbf{X}_R$ which is simultaneously a k-$2k$ cover for all k in log-scale $[\log n : n/2], n = n_P \vee n_Q$ (see Algorithm 1, Appendix C). We then propose a NN classifier that automatically chooses an appropriate value of k (in the given range), locally at every query x (see Algorithm 2, Appendix C); such choice of k builds on so-called *Lepski's method* (Lepski and Spokoiny, 1997) for adaptivity.

Definition 11 (Cover-based local NN) *Consider $\mathbf{X}_R \subset \mathbf{X}$, indexed by $R \subset [n_P + n_Q]$, and which is simultaneously a k-$2k$ cover for all values of k in the range $\mathcal{K} \doteq \{2^i k_0\}_{i=1}^{\lfloor \log_2((n_p \vee n_P)/2k_0) \rfloor}$, for some $k_0 = \Theta(\log(n_p + n_Q))$ to be specified. Query labels, once, for all $X_i \in \mathbf{X}_R \cap \mathbf{X}_Q$, and let $(\mathbf{X}, \mathbf{Y})_R$ denote the labeled set. A* cover-based local NN *classifier, denoted $\hat{h}_R$, is a k-NN classifier defined over such $(\mathbf{X}, \mathbf{Y})_R$, and which uses a local choice of $k = k(x) \in \mathcal{K}$ at every query x.*

As in (Berlind and Urner, 2015), our results rely on the following two additional assumptions. The first holds for instance for Euclidean subspaces (with known upper-bounds on $\mathcal{V}_\mathcal{B}$); a simple example of the second assumption is one where Q_X has an upper-bounded Lebesgue density in $\mathbb{R}^d$.

Assumption 1 (Bounded VC) *The family $\mathcal{B}$ of balls in $(\mathcal{X}, \rho)$ has* known *finite VC-dimension $\mathcal{V}_\mathcal{B}$.*

Assumption 2 (Bounded Q-mass) *Let d be the* dimension *parameter in either (DM) or (BCN). Q_X further satisfies: $\forall r \in [0, \Delta_\mathcal{X}], \forall x \in \mathcal{X}_Q, Q_X(B(x,r)) \leq C'_d (r/\Delta_\mathcal{X})^d$, for some $C'_d > 0$.*

The following adaptivity result matches the earlier minimax rates (up to log factors in k_0).

Theorem 3 (Adaptive rates) *Let Assumption 1 hold, and let $\mathcal{T}$ denote $\mathcal{T}_{(DM)}$ or $\mathcal{T}_{(BCN)}$. For $\mathcal{T} = \mathcal{T}_{(BCN)}$ assume further that $\alpha < d$. There exists a cover-based local NN classifier $\hat{h}_R$, defined with $k_0 = \Theta(\log(n_p + n_Q))$, and which, without knowledge of the parameters $(C_\gamma, \gamma, C_\alpha, \alpha, C_\beta, \beta, C_d, d)$, satisfies the following. For a constant $C = C(\mathcal{T})$, let $(n_p + n_Q) \geq C k_0^3 \log^3(n_P + n_Q)$. We have:*

$$\sup_{(P,Q) \in \mathcal{T}} \mathbb{E}_{(\mathbf{X}, \mathbf{Y})}[\mathcal{E}_Q(\hat{h}_R)] \leq C \left(\frac{k_0 \cdot \log(2(n_P + n_Q))}{n_P^{d_0/(d_0 + \gamma/\alpha)} + n_Q} \right)^{(\beta+1)/d_0},$$

where $d_0 = 2 + d/\alpha$ when $\mathcal{T} = \mathcal{T}_{(DM)}$, and $d_0 = 2 + \beta + d/\alpha$ when $\mathcal{T} = \mathcal{T}_{(BCN)}$.
For the case $\mathcal{T} = \mathcal{T}_{(BCN)}$ with $\alpha = d$, C above is replaced with $C \cdot \log(2(n_P + n_Q))$.

Recall from Theorems 1 and 2 that the above is the best rate attainable even if we request n_Q target labels. We in fact request much less whenever target labels have little benefit, i.e., when n_P is already sufficiently large (in terms of unknown γ) w.r.t. the budget n_Q.

Theorem 4 (Labeling threshold) *Let Assumptions 1 and 2 hold. For $0 < \delta < 1$, set $k_0 = \lceil \mathcal{V}_\mathcal{B} \log(2(n_P + n_Q)) + \log(6/\delta) \rceil$, and let $\mathcal{K}$ be the corresponding range of k-values (Definition 9). Then it is possible to construct $\mathbf{X}_R$ (a uniform k-$2k$ cover for $k \in \mathcal{K}$) with the following property. Under both regimes $\mathcal{T}_{(DM)}$ or $\mathcal{T}_{(BCN)}$, there exists a constant $C = C(\gamma, d) > 0$ such that, with probability at least $1 - \delta$, there is* no *label query whenever $n_P^{d/(d+\gamma)} \geq C n_Q$.*

It remains unclear whether a tighter labeling threshold is achievable (as in Theorem 2) without prior knowledge of distributional parameters such as α and β. However the above labeling threshold maintains the ideal property that label queries are unlikely as $\gamma \to 0$, without knowledge of γ.

4. Analysis Overview

We now present our main technical ideas, while the full analysis is provided in the appendix.

4.1. Lower-Bound Analysis

As stated earlier, the main technicality in the proof of Theorem 1 is in the coupling of P and Q, i.e., dealing with classifiers learned on non-identical samples. At a high-level, we call on known extensions of Fano's lemma (Cover and Thomas, 2012) which roughly state the following:

> Let $\{\Pi_h\}$ denote distributions indexed by $h \in \mathcal{H}$. Suppose all h's are *far* from each other under a semi-metric $\bar{\rho}$, but Π_h's are close in KL-divergence (Kullback-Leibler). Then, for any learner $\hat{h}$ of h, there is a sizable Π_h-probability that $\hat{h}$ is $\bar{\rho}$-far from h.

See Proposition 2 for such a statement (due to Tsybakov (2009)). The work of Audibert and Tsybakov (2007) instead uses an approach based on so-called Assouad lemma.

For our purpose, the h indices would stand for Bayes classifiers over possible regression functions η satisfying (DM) or (BCN). Let (P, Q) denote a transfer tuple with corresponding Bayes classifier h; for fixed n_P, n_Q, we let $\Pi_h = P^{n_P} \times Q^{n_Q}$, thus coupling P and Q into a single distribution. Now, while the KL-divergences over the family involve both P and Q, we are free to define $\bar{\rho}$ over Q_X alone, and thus relate it to the target excess error $\mathcal{E}_Q$. Now what's left is to ensure that the various conditions of (DM) or (BCN) are satisfied.

For conditions involving only Q (smoothness, noise, and dimension), we follow closely the original lower-bound construction of Audibert and Tsybakov (2007), apart for some technical details in our choice of smooth basis functions (η is chosen as a linear combination of simple basis functions crossing $1/2$). Now, to ensure that any given transfer-exponent γ holds, we divide up the mass of P_X appropriately over space, following the type of intuitions laid out in Examples 1, 2, 3, 4 of Section 2.3. The rest involves adjusting the construction properly so that h's are sufficiently far in $\bar{\rho} = \bar{\rho}(Q_X)$, while Π_h's (involving both P, Q) remain sufficiently close in KL-divergence.

Finally, we note that the marginals P_X, Q_X remain fixed for our choice family $\{\Pi_h\}$, and thus might be known to the learner (which is allowed to know the family, but not the data's distribution).

4.2. Upper-Bound Analysis

Here we outline the main insights in obtaining Theorem 2. We build on previous insights from work on k-NN methods whenever possible. The two main difficulties are (a), accounting for the noise condition (the parameter β) in the (BCN) setting (without assuming local choices of k of knowledge of Q_X as in Gadat et al. (2014)), and (b), merging this with the fact that $\hat{h}_k$ is defined on two non-identical samples (in particular, merging γ into the bound).

First, a general step in analyses of k-NN (and other plug-in classifiers) is the following inequality which relates classification error for $\hat{h}_k = \mathbb{1}\{\hat{\eta}_k \geq 1/2\}$ to the regression error $|\hat{\eta}_k - \eta|$:

$$\mathcal{E}_Q(\hat{h}_k) \leq 2\mathbb{E}_Q \left| \eta(X) - \frac{1}{2} \right| \cdot \mathbb{1} \left\{ \left| \eta(X) - \frac{1}{2} \right| \leq |\hat{\eta}_k(X) - \eta(X)| \right\}. \tag{3}$$

This is direct from the definition of excess error in equation (1): notice that, for any fixed x, the event $\hat{h}_k(x) \neq h^*(x)$ implies that $|\hat{\eta}_k(x) - \eta(x)| \geq |\eta(x) - 1/2|$.

We first remark that, from (3), $\mathcal{E}_Q(\hat{h}_k)$ is trivially bounded by $2|\hat{\eta}_k(x) - \eta(x)|$, which unfortunately yields a weak bound in terms of α alone (smoothness). The usual approach in accounting for β relies on the following simple insight: suppose a uniform bound $\sup_x |\hat{\eta}_k(x) - \eta(x)| \leq t$ held (at least in high-probability) for some $t = t(k, n_P, n_Q)$, then we would have $\mathcal{E}_Q(\hat{h}_k) \leq C_\beta t^{(\beta+1)}$, using the fact that $\mathbb{E} Z \cdot \mathbb{1}\{Z \leq t\} \leq t \cdot \mathbb{P}(Z \leq t)$, and letting $Z \doteq |\eta(X) - 1/2|$.

Under (DM) such uniform bound on regression error are possible, even in our transfer setting, since the problem is similarly hard everywhere in space. Unfortunately, such uniform bound is not possible under (BCN) where the difficulty changes over space as both P_X, Q_X vary.

Our approach therefore is to decompose the regression error into various terms, some of which can be bounded uniformly over $x \in \mathcal{X}_Q$. Namely, suppose $|\hat{\eta}_k(x) - \eta(x)| \leq \sum_{i \in [c]} G_i(x)$, then

$$\mathbb{1}\{Z \leq |\hat{\eta}_k(x) - \eta(x)|\} \leq \sum_{i \in [c]} \mathbb{1}\{Z \leq cG_i(x)\}. \tag{4}$$

In other words, if we can bound some such term G_i uniformly over x by some t_i, we can proceed as above to bound $\mathbb{E} Z \cdot \mathbb{1}\{Z \leq cG_i(X)\}$ by $C_\beta(ct_i)^{(\beta+1)}$, and thus account for β in our final bound on the classification error $\mathcal{E}_Q(\hat{h}_k)$. We start our decomposition in a standard way as follows.

Fix any x and let $\{X_{(i)}\}_1^k$ denote its k nearest neighbors in $\mathbf{X} \doteq \mathbf{X}_P \cup \mathbf{X}_Q$. By a triangle inequality and the fact that η is (C_α, α) Hölder, we have:

$$|\hat{\eta}_k(x) - \eta(x)| \leq \frac{1}{k}\left|\sum_{i=1}^k Y_{(i)} - \eta(X_{(i)})\right| + \frac{C_\alpha}{k}\sum_{i=1}^k \rho(X_{(i)}, x)^\alpha. \tag{5}$$

Now, although NN distances $\rho(X_{(i)}, x)$ over $\mathbf{X}$ are trivially bounded by the distance to the k-th NN of x in *either* samples $\mathbf{X}_P$ or $\mathbf{X}_Q$, such a bound would be in terms of only n_P or only n_Q, and therefore would not properly capture the interaction between n_P and n_Q; in particular the effect of the transfer-exponent γ can get lost. However, as it turns out, the interaction between n_P and n_Q (in terms of γ) is easiest to capture when bounding 1-NN distances rather than k-NN distances.

We therefore proceed by first reducing the problem of bounding k-NN distances to that of bounding 1-NN distances by extending a useful trick of Györfi et al. (2006, Section 6.3) to our transfer setting with two samples. Start with the following definition.

Definition 12 (Implicit 1-NNs) *Divide* $(\mathbf{X}, \mathbf{Y})$ *into k disjoint batches each containing $\lfloor \frac{n_P}{k} \rfloor$ samples from $(\mathbf{X}, \mathbf{Y})_P$ and $\lfloor \frac{n_Q}{k} \rfloor$ samples from $(\mathbf{X}, \mathbf{Y})_Q$. Fix $x \in \mathcal{X}$ and define $\{\tilde{X}_i\}_{i=1}^k$ as its 1-NNs in each of the k batches. Let the assignment to each batch consist of picking, without replacement, $\lfloor \frac{n_P}{k} \rfloor$ indices from $[n_P]$ and $\lfloor \frac{n_Q}{k} \rfloor$ indices from $[n_Q]$, so that the $\tilde{X}_i$'s are i.i.d. given x.*

It can then be shown that, for any fixed $x \in \mathcal{X}$ we have (see Lemma 8 in Appendix B) $\sum_{i=1}^k \rho(X_{(i)}, x)^\alpha \leq \sum_{i=1}^k \rho(\tilde{X}_i, x)^\alpha$. Combining this last inequality with (5), it follows that $|\hat{\eta}_k(x) - \eta(x)| \leq \frac{1}{k}\left|\sum_{i=1}^k Y_{(i)} - \eta(X_{(i)})\right| + \frac{C_\alpha}{k}\sum_{i=1}^k \rho(\tilde{X}_i, x)^\alpha$, which equals

$$\underbrace{\frac{1}{k}\left|\sum_{i=1}^k Y_{(i)} - \eta(X_{(i)})\right|}_{G_1(x)} + \underbrace{\frac{C_\alpha}{k}\sum_{i=1}^k \left(\rho(\tilde{X}_i, x)^\alpha - \mathbb{E}_{\tilde{X}_1} \rho(\tilde{X}_1, x)^\alpha\right)}_{G_2(x)} + \underbrace{C_\alpha \mathbb{E}_{\tilde{X}_1} \rho(\tilde{X}_1, x)^\alpha}_{G_3(x)}. \tag{6}$$

The decomposition in (6) serves to further isolate terms that can be bounded uniformly over x, namely G_1 and G_2. The above discussion then leads to the following proposition.

Proposition 1 (Error Decomposition) *Let $1 \leq k \leq n_P \vee n_Q$ and let $\hat{h}_k$ be the k-NN classifier on* $(\mathbf{X}, \mathbf{Y})$. *Consider any $x \in \mathcal{X}$ with k nearest neighbors $\{X_{(i)}\}_1^k$, and implicit 1-NN's $\{\tilde{X}_i\}_1^k$. Let $G_i(x)$, $i = 1, 2, 3$ denote the three terms in inequality* (6), *and for each $G_i(x)$ define the function $\Phi_i(x) \doteq 2\left|\eta(x) - 1/2\right| \cdot \mathbb{1}\{|\eta(x) - 1/2| \leq 3G_i(x)\}$. We have:*

$$\mathbb{E}[\mathcal{E}_Q(\hat{h}_k)] \leq \mathbb{E}[\Phi_1(X)] + \mathbb{E}[\Phi_2(X)] + \mathbb{E}[\Phi_3(X)], \tag{7}$$

where the expectations are taken over $(\mathbf{X}, \mathbf{Y})$ *and* X.

Proof Apply (4) (with $Z \doteq |\eta - 1/2|$) to the decomposition of (6), and conclude using (3). ∎

The first two terms in (7) are of order $(1/\sqrt{k})^{(\beta+1)}$ as shown via a concentration and chaining argument (see Lemma 11, Appendix B). The term $\mathbb{E}[\Phi_3(X)]$ accounts for γ (see Lemmas 12 and 14, resp. for (DM) and (BCN)): for intuition, remark that the 1-NN tail $\mathbb{P}(\rho(\tilde{X}_1, x) > t)$ equals

$$(1 - P_X(B(x,t)))^{\lfloor \frac{n_p}{k} \rfloor}(1 - Q_X(B(x,t)))^{\lfloor \frac{n_Q}{k} \rfloor} \leq (1 - Q_X(B(x,t))C_\gamma t^\gamma)^{\lfloor \frac{n_p}{k} \rfloor}(1 - Q_X(B(x,t)))^{\lfloor \frac{n_Q}{k} \rfloor}.$$

Careful tail-integration reveals further dependence on β while optimally capturing γ. Finally, Theorem 2 is obtained by optimizing over k. All details are given in Appendix B.

4.3. Analysis for Adaptive Labeling

We start with some intuition about the adaptive rates of Theorem 3.

First, $\hat{h}_R$ is defined as $\mathbb{1}\{\hat{\eta} \geq 1/2\}$ for a k-NN regression estimate $\hat{\eta}$, where k is chosen adaptively *at every* x, following Lepski's method (see Algorithm 2). Namely, $\hat{\eta}(x)$ is chosen from a confidence interval on $\eta(x)$ iteratively refined over k-NN regression estimates $\hat{\eta}_k(x)$ of $\eta(x)$ for increasing values of k in the range $\mathcal{K}$. These intervals are of the form $\hat{\eta}_k \pm 1/\sqrt{k}$, accounting for variance in the estimates, and overlap as long as variance dominates bias. The stopping condition is such that, whenever these intervals diverge, we can ensure that the current value of k approximately balances bias and variance terms, and in particular yields a regression bound $|\hat{\eta}(x) - \eta(x)|$ of the same order as that of the optimal regression choice $k^*(x)$ at x (which is itself unknown).

It follows that the regression error for $\hat{\eta}(x)$ cannot be much worse than that for the optimal global classification choice of $k^* = k^*(\gamma, \alpha, \beta, d)$. Thus, suppose for now that the NN estimates were computed over a full sample $(\mathbf{X}, \mathbf{Y})_R = (\mathbf{X}, \mathbf{Y})$. Then, we can conclude using inequality (3) that the classification error of $\hat{h}_R$ is no worse than that of $\hat{h}_{k^*}$ since, pointwise we have

$$\mathbb{1}\{|\eta(x) - 1/2| < |\hat{\eta}(x) - \eta(x)|\} \leq \mathbb{1}\{|\eta(x) - 1/2| < |\hat{\eta}_{k^*}(x) - \eta(x)|\}.$$

However, the NN estimates are actually computed over $(\mathbf{X}, \mathbf{Y})_R \subset (\mathbf{X}, \mathbf{Y})$. We therefore have to argue that the best possible regression error over $(\mathbf{X}, \mathbf{Y})_R$ is of similar order as that over $(\mathbf{X}, \mathbf{Y})$. A first remark is that the variance terms (first term in (5)) remain of the same order $O(1/\sqrt{k})$ uniformly over x. The bias terms (second term of (5)) are in terms of NN distances in $\mathbf{X}_R$. Fortunately, an interesting property of a k-$2k$ cover shown by Berlind and Urner (2015) is that they approximately preserve distances to the k-th NN of any x in the original sample. Here, in order to bring in the effect of γ, we reduce these distances to *implicit* 1-NN distances (see above discussion of Section 4.2); we therefore derive a similar result showing that in fact all i-th NN distances, $i \leq k$, are approximately preserved.

Finally, Theorem 4 follows from a main intuition behind the notion of a k-$2k$ cover. First fix $x = X_i$ for some $X_i \in \mathbf{X}_Q$. Initially $\mathbf{X}_R = \mathbf{X}_P$. Thus, we won't request a label at x if at least k samples *from* $\mathbf{X}_P$ fall in a neighborhood of x. In particular, if n_P is sufficiently large (w.r.t. n_Q as in the result's statement), we can ensure that the smallest ball $B_Q(x)$ containing k samples from $\mathbf{X}_Q$ must also contain k samples from $\mathbf{X}_P$ (this follows from lower-bounding P_X-mass by Q_X-mass using the definition of γ). Now, if the smallest ball $B_{P,Q}(x)$ containing $2k$ samples from $\mathbf{X} = \mathbf{X}_P \cup \mathbf{X}_Q$ contains $B_Q(x)$, we are done; otherwise $B_{P,Q}(x)$ has less than k samples from $\mathbf{X}_Q$ and so must have at least k samples from $\mathbf{X}_P$. The rest follows from concentration arguments.

The detailed proofs can be found in Appendix C.

Final Remarks

The transfer-exponent γ successfully captures the relative benefits of source and target data, as shown through matching upper and lower-bounds. Our results hold for nonparametric classification. However, other interesting transfer problems such as in regression have received considerable attention in the literature (Blitzer et al., 2011; Kuzborskij and Orabona, 2013; Hoffman et al., 2017); it remains unclear whether our present notions tightly capture such problems, especially given the usually stronger structural assumptions on regression functions (e.g. linearity, sparsity).

Finally, while γ appears to tightly capture the complexity of transfer in a minimax sense over all possible procedures, it does not properly capture the interaction between (P, Q) and any fixed family of procedures or hypotheses, as is the goal for instance with earlier notions such as the $d_{\mathcal{A}}$-divergence or $\mathcal{Y}$-discrepancy of Ben-David et al. (2010a) and Mansour et al. (2009a).

References

Jean-Yves Audibert and Alexandre B Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007.

Shai Ben-David and Ruth Urner. On the hardness of domain adaptation and the utility of unlabeled target samples. In *International Conference on Algorithmic Learning Theory*, pages 139–153. Springer, 2012.

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

Shai Ben-David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 129–136, 2010b.

Christopher Berlind and Ruth Urner. Active nearest neighbors in changing environments. In *International Conference on Machine Learning*, pages 1870–1879, 2015.

John Blitzer, Sham Kakade, and Dean Foster. Domain adaptation with coupled subspaces. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 173–181, 2011.

Rita Chattopadhyay, Wei Fan, Ian Davidson, Sethuraman Panchanathan, and Jieping Ye. Joint transfer and batch-mode active learning. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 253–261, 2013.

Kamalika Chaudhuri and Sanjoy Dasgupta. Rates of convergence for nearest neighbor classification. In *Advances in Neural Information Processing Systems*, pages 3437–3445, 2014.

Minmin Chen, Kilian Q Weinberger, and John Blitzer. Co-training for domain adaptation. In *Advances in neural information processing systems*, pages 2456–2464, 2011.

Corinna Cortes, Mehryar Mohri, and Andrés Munoz Medina. Adaptation based on generalized discrepancy. *Machine Learning Research, forthcoming. URL http://www. cs. nyu. edu/~ mohri/pub/daj. pdf.*

Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In *Advances in neural information processing systems*, pages 442–450, 2010.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

Sébastien Gadat, Thierry Klein, and Clément Marteau. Classification with the nearest neighbor rule in general finite dimensional spaces: necessary and sufficient conditions. *arXiv preprint arXiv:1411.0894*, 2014.

Pascal Germain, Amaury Habrard, François Laviolette, and Emilie Morvant. A pac-bayesian approach for domain adaptation with specialization to linear classifiers. In *International Conference on Machine Learning*, pages 738–746, 2013.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

Judy Hoffman, Mehryar Mohri, and Ningshan Zhang. Multiple-source adaptation for regression problems. *arXiv preprint arXiv:1711.05037*, 2017.

Jiayuan Huang, Arthur Gretton, Karsten M Borgwardt, Bernhard Schölkopf, and Alex J Smola. Correcting sample selection bias by unlabeled data. In *Advances in neural information processing systems*, pages 601–608, 2007.

S. Kpotufe. k-NN Regression Adapts to Local Intrinsic Dimension. *NIPS*, 2011.

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

Sanjeev R Kulkarni and Steven E Posner. Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Transactions on Information Theory*, 41(4):1028–1039, 1995.

Ilja Kuzborskij and Francesco Orabona. Stability and hypothesis transfer learning. In *Proceedings of the 30th International Conference on Machine Learning*, pages 942–950, 2013.

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

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

Yishay Mansour, Mehryar Mohri, and Afshin 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, 2009b.

Mehryar Mohri and Andres Munoz Medina. New analysis and algorithm for learning with drifting distributions. In *International Conference on Algorithmic Learning Theory*, pages 124–138. Springer, 2012.

Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009.

Avishek Saha, Piyush Rai, Hal Daumé, Suresh Venkatasubramanian, and Scott L DuVall. Active supervised domain adaptation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 97–112. Springer, 2011.

Richard J Samworth et al. Optimal weighted nearest neighbour classifiers. *The Annals of Statistics*, 40(5):2733–2763, 2012.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

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

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

Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer, 2009.

V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their expectation. *Theory of probability and its applications*, 16:264–280, 1971.

Liu Yang, Steve Hanneke, and Jaime Carbonell. A theory of transfer learning with applications to active learning. *Machine learning*, 90(2):161–189, 2013.

Bin Yu. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

Appendix A. Lower Bound Analysis

Theorem 1 is a consequence of the below Propositions 3, 4 and 5. Our analysis follows the usual intuition behind traditional minimax proofs (i.e. in finding sufficiently large *packings* under appropriate metrics) with the added difficulty that here we deal with tuples (P, Q) of distributions from which two separate samples are drawn. Various basic tools are used in the literature, which often build on Fano's inequality, Assouad, or LeCam's approach (Yu, 1997). In particular, Theorem 2.5 of Tsybakov (2009) is stated in a general enough form to address part of our needs.

Proposition 2 (Thm 2.5 of Tsybakov (2009)) *Let $\{\Pi_h\}_{h \in \mathcal{H}}$ be a family of distributions indexed over a subset $\mathcal{H}$ of a semi-metric $(\mathcal{F}, \bar{\rho})$. Suppose $\exists\, h_0, \dots, h_M \in \mathcal{H}$, for $M \geq 2$, such that:*

(i) $\quad \bar{\rho}(h_i, h_j) \geq 2s > 0, \quad \forall 0 \leq i < j \leq M,$

(ii) $\quad \Pi_{h_i} \ll \Pi_{h_0} \quad \forall i \in [M]$, *and the average KL-divergence to Π_{h_0} satisfies*

$$\frac{1}{M} \sum_{i=1}^{M} \mathcal{D}_{kl}\left(\Pi_{h_i} | \Pi_{h_0}\right) \leq \kappa \log M, \text{ where } 0 < \kappa < 1/8.$$

Let $Z \sim \Pi_h$, and let $\hat{h} : Z \mapsto \mathcal{F}$ denote any improper *learner of $h \in \mathcal{H}$. We have for any $\hat{h}$:*

$$\sup_{h \in \mathcal{H}} \Pi_h \left(\bar{\rho}\left(\hat{h}(Z), h\right) \geq s\right) \geq \frac{\sqrt{M}}{1 + \sqrt{M}} \left(1 - 2\kappa - \sqrt{\frac{2\kappa}{\log(M)}}\right) \geq \frac{3 - 2\sqrt{2}}{8}.$$

We use Proposition 2 on a family of 2^m distribution tuples (P, Q) from $\mathcal{T}$ that can be indexed by their related Bayes classifiers h^*. As these form a subfamily of $\mathcal{T}$, the proven lower bound work also for the full class $\mathcal{T}$ as stated in Theorem 1. All these distribution tuples have the same marginals, that we denote generically by P_X and Q_X, and differ only by their conditional distributions.

Finally, the choice of the $M + 1$ elements h_i in Proposition 2 should be made such that M is as large as possible while maintaining the *packing* (i) and *covering* (ii) conditions of Proposition 2. The following lemma often comes in handy, as it allows us to concentrate on designing a large collection of distributions that satisfy (ii), and then pulling out from it a large enough sub-collection that satisfies (i).

Lemma 5 (Varshamov-Gilbert bound) *Let $m \geq 8$. Then there exists a subset $\{\sigma_0, \dots, \sigma_M\}$ of $\{-1, 1\}^m$ such that $\sigma_0 = (0, \dots, 0)$,*

$$\rho_H(\sigma_i, \sigma_j) \geq \frac{m}{8}, \quad \forall 0 \leq i < j \leq M, \quad \text{and} \quad M \geq 2^{m/8},$$

where $\rho_H(\sigma, \sigma') \doteq card(\{i \in [m] : \sigma(i) \neq \sigma'(i)\})$ is the Hamming distance.

We start by the lower bounds for the family $\mathcal{T}_{(DM)}$.

A.1. Lower Bound for $\mathcal{T} = \mathcal{T}_{(DM)}$ when $\gamma < \infty$

Proposition 3 *Let $(\mathcal{X}, \rho) = ([0,1]^d, \|.\|_\infty)$, for some $d \in \mathbb{N}^*$, and assume that $\alpha\beta \leq d$ and $\gamma < \infty$. There exists a constant $c = c(\mathcal{T}_{(DM)})$ such that, for any classifier $\hat{h}$ learned on $(\mathbf{X}, \mathbf{Y})$ and with knowledge of P_X, Q_X, we have:*

$$\sup_{(P,Q) \in \mathcal{T}_{(DM)}} \mathbb{E}_{(\mathbf{X},\mathbf{Y})}[\mathcal{E}_Q(\hat{h})] \geq c \left(n_P^{d_0/(d_0 + \gamma/\alpha)} + n_Q\right)^{-(\beta+1)/d_0},$$
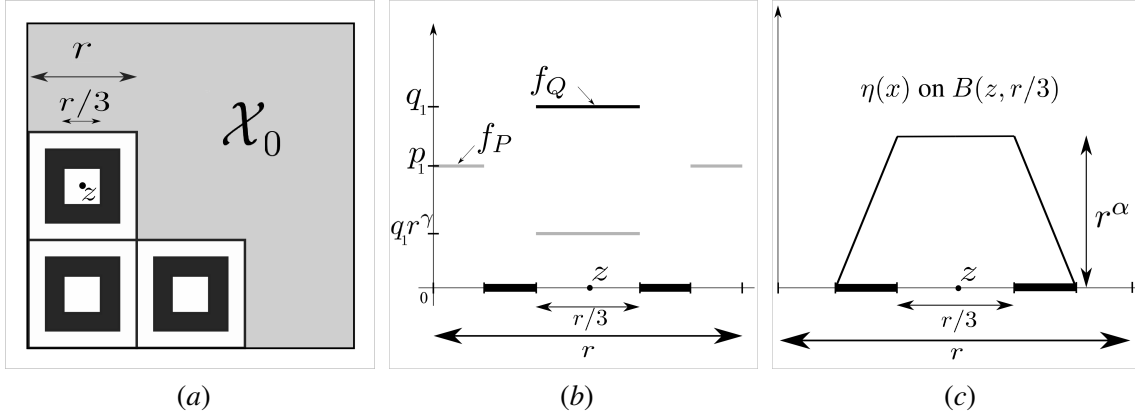
Figure 2: *2(a) illustrates the supports of P_X and Q_X on the hypercubes arising from the subdivision of the space $\mathcal{X}$ into $\mathcal{X}_0$ (which serves to account for missing mass) and $\mathcal{X} \setminus \mathcal{X}_0$ where we make η vary (across hypercubes) so as to make classification difficult, subject to the various distributional conditions. 2(b) shows the profiles of densities f_P, f_Q of P_X and Q_X on a hypercube in $\mathcal{X} \setminus \mathcal{X}_0$, having some center z; we note that the construction here is a simple one that allows a transfer-exponent γ at resolution r, while simplifying the analysis; however other constructions such as ones described in Figure 1 also work, but add technicality with no additional insight. 2(c) displays the profile of the regression function (in the Lipschitz case) on the above-mentioned hypercubes. Notice that η maintains a margin r^α on $B(z, r/6)$, i.e., on the support of Q_X inside the hypercube $B(z, r/2)$.*

where $d_0 = 2 + d/\alpha$.

Proof First, let's fix the following variables, where the constants $c_r = 1/9$, $c_m = 8 \times 9^{\alpha\beta - d}$ and $c_w \in (0, 1]$ depends only on the parameters of $\mathcal{T}_{(DM)}$ and will be given later.

$$r = c_r \left(n_P^{d_0/(d_0 + \gamma/\alpha)} + n_Q \right)^{-1/(\alpha d_0)}, \quad m = \left\lfloor c_m r^{\alpha\beta - d} \right\rfloor, \quad w = c_w r^d.$$

Note that because of the choices of the constants c_r and c_m we have that $8 \le m < \left\lfloor r^{-1} \right\rfloor^d$. Indeed, as $\alpha\beta \le d$ by assumption and $r \le 1/9$ we have $c_m r^{\alpha\beta - d} \ge 8$. Furthermore, in order to prove that $m < \left\lfloor r^{-1} \right\rfloor^d$, first remark that because $r^{-1} \ge 9$ we have that $r^{-1} < 9 \left\lfloor r^{-1} \right\rfloor / 8$. Therefore $c_m r^{\alpha\beta - d} = 8(9r)^{\alpha\beta} (r^{-1}/9)^d < 8^{1-d} \left\lfloor r^{-1} \right\rfloor^d \le \left\lfloor r^{-1} \right\rfloor^d \in \mathbb{N}$, which implies $m < \left\lfloor r^{-1} \right\rfloor^d$ as desired. Notice that this implies also that we have $mw < 1$.

CONSTRUCTING THE MARGINAL Q_X

Consider a regular subdivision of $\mathcal{X} = [0, 1]^d$ into $\left\lfloor r^{-1} \right\rfloor^d$ smaller hypercubes of side length r (see Figure 2(a)). Call $\mathcal{Z}$ the set of the centers of these hypercubes. Now divide arbitrarily the set $\mathcal{Z}$ into two disjoints subsets $\mathcal{Z}_0$ and $\mathcal{Z}_1$ such that $|\mathcal{Z}_1| = m$ (we represent only the hypercubes centered in $\mathcal{Z}_1$ in Figure 2(a)). Set Q_X to have a uniform density $q_1 > 0$ with respect to the Lebesgue measure on each set $B(z, r/6)$ for $z \in \mathcal{Z}_1$, such that $B(z, r/6) = w$ (see Figure 2(b)). Then, put the remaining weight $1 - mw$ of Q_X uniformly with density $q_0 > 0$ over the remaining hypercubes centered in $\mathcal{Z}_0$, that is over $\mathcal{X}_0 \doteq \cup_{z \in \mathcal{Z}_0} B(z, r/2)$. The rest of the space has probability zero under the target Q_X.

Note that if we choose c_w small enough we have these densities being bounded from below independently of n_P and n_Q:

$$q_0 = \frac{1 - mw}{vol\left(\bigcup_{z \in \mathcal{Z}_0} B(z, r/2)\right)} \geq 1 - c_w c_m > 0, \quad q_1 = \frac{w}{vol\left(B(z, r/6)\right)} \geq r^{-d} w = c_w > 0. \quad (8)$$

Therefore, from the regularity of its support, this implies that Q_X respects the doubling measure assumption for some constant coefficient C_d independent of n_P and n_Q. [2]

CONSTRUCTING THE MARGINAL P_X

Now let's turn to the construction of P_X. The idea is to let P_X be uniformly distributed on each of the sets $B(z, r/6)$ for $z \in \mathcal{Z}_1$, and so that its density is getting smaller and smaller w.r.t. Q_X's density as r goes to zero (when $\gamma > 0$). More precisely, let $p_1 = q_1 r^\gamma$ the density of P_X on the $B(z, r/6)$'s for z in $\mathcal{Z}_1$. Because of the factor $r^\gamma \leq 1$, we have that $P_X(B(z, r/6)) = Q_X(B(z, r/6)) r^\gamma \leq Q_X(B(z, r/6))$. We therefore put the remaining mass of P_X (if any) uniformly on each set $B(z, r/2) \backslash B(z, r/3)$ such that $P_X(B(z, r/2)) = Q_X(B(z, r/2))$, $\forall z \in \mathcal{Z}_1$ (see Figure 2(b)). We let P_X have a uniform density p_0, equal to the one of Q_X (that is $p_0 = q_0$), on the remaining hypercubes $B(z, r/2)$ for $z \in \mathcal{Z}_0$. Hence we have also $P_X(B(z, r/2)) = Q_X(B(z, r/2))$, $\forall z \in \mathcal{Z}_0$.

Recall that the support of Q_X is the union of the sets $B(z, r/6)$ for all $z \in \mathcal{Z}_1$ and $B(z, r/2)$ for $z \in \mathcal{Z}_0$, hence we can check (2) only for points x in these sets. Fix $z \in \mathcal{Z}_1$, we have $\forall x \in B(z, r/6), \forall r' \in [0, r/3]$:

$$P_X(B(x, r')) \geq p_1 vol(B(x, r') \cap B(z, r/6)) = r^\gamma Q_X(B(x, r')) \geq r'^\gamma Q_X(B(x, r')).$$

And for $z \in \mathcal{Z}_0$, and $\forall x \in B(z, r/2), \forall r' \in [0, r/3]$, the inequality is even more direct:

$$P_X(B(x, r')) \geq P_X(B(x, r') \cap (\cup_{z \in \mathcal{Z}_0} B(z, r/2))) = Q_X(B(x, r')) \geq r'^\gamma Q_X(B(x, r')).$$

Therefore we can see that for small values of $r' \leq r/3$, equation (2) from Definition 3 holds (as $C_\gamma \leq 1$). Furthermore, because we set $P_X(B_z) = Q_X(B_z)$, $\forall z \in \mathcal{Z}$, equation (2) holds also for larger r' for a fixed constant $C_\gamma > 0$, that depends only on the dimension d. [3]

CONSTRUCTING THE CONDITIONAL DISTRIBUTIONS

Let $u : \mathbb{R}_+ \to \mathbb{R}_+$ such that:

$$u(x) = \begin{cases} 1 & \text{if } x \leq 1/6, \\ 1 - 6(x - 1/6) & \text{if } x \in (1/6, 1/3], \\ 0 & \text{elsewhere.} \end{cases}$$

It is easy to see that u is 6–Lipschitz. Let $C'_\alpha \doteq \min(C_\alpha 6^{-\alpha}, 1/2)$ (the fact that we take $C'_\alpha \leq 1/2$ will be useful later in our proof). This implies that $C'_\alpha u^\alpha(\|.\|_\infty)$ is (C_α, α)–Hölder, as

---

2. We can in fact start with any coefficient for the family $\mathcal{T}_{(DM)}$ and satisfy the doubling assumption by properly adjusting the size of the hypercube independently of n_P and n_Q. See Remark 6.

3. As in the earlier case for the doubling measure condition, we can start with any constant and appropriately adjust the size of the hypercube, independently of n_P and n_Q. See again Remark 6.

by concavity we have $\forall 0 \leq x \leq y, \ y^\alpha - x^\alpha \leq (y - x)^\alpha$. Therefore, the following functions are (C_α, α)–Hölder:

$$\forall z \in \mathcal{Z}_1, \quad \eta_z(x) \doteq C'_\alpha r^\alpha u^\alpha(\|x - z\|_\infty / r).$$

The profile of these functions on each hypercube $B(z, r/2)$ for $z \in \mathcal{Z}_1$ is represented in Figure 2(c). Now consider the vectors $\sigma \in \{-1, 1\}^m$ that assign values -1 or 1 to each of the m centers z from the set $\mathcal{Z}_1$. And let the following 2^m (C_α, α)–Hölder regression functions, indexed by σ:

$$\eta_\sigma(x) = \begin{cases} (1 + \sigma(z)\eta_z(x))/2 & \text{if } x \in B(z, r/2), \quad z \in \mathcal{Z}_1, \\ 1/2 & \text{elsewhere.} \end{cases}$$

where $\sigma(z) \in \{-1, 1\}$ is the value that σ assigns to z. Note that each of these functions will take constant values $(1 \pm C'_\alpha r^\alpha)/2$ over the balls $B(z, r/6)$ of centers $z \in \mathcal{Z}_1$ and be equal to $1/2$ everywhere else. We therefore define the following 2^m distribution tuples (P^σ, Q^σ), indexed by σ:

$$\forall \sigma \in \{-1, 1\}^m, \quad P_X^\sigma \doteq P_X, \ Q_X^\sigma \doteq Q_X, \ P^\sigma(Y = 1|X) = Q^\sigma(Y = 1|X) \doteq \eta_\sigma(X).$$

And we also define their related full sample distributions:

$$\Pi_\sigma \doteq P^{\sigma \otimes n_P} \otimes Q^{\sigma \otimes n_Q}.$$

CHECKING TSYBAKOV NOISE ASSUMPTION

Now let's check that the Tsybakov low-noise assumption (Definition 5) is satisfied. We have for $t < C'_\alpha r^\alpha / 2$:

$$Q_X(0 < |\eta(X) - 1/2| \leq t) = 0,$$

and for $t \geq C'_\alpha r^\alpha / 2$:

$$Q_X(0 < |\eta(X) - 1/2| \leq t) = mw.$$

Thus, this assumption would be satisfied if we choose c_w small enough such that:

$$mw \leq c_m c_w r^{\alpha\beta} \leq C_\beta \left(\frac{C'_\alpha r^\alpha}{2}\right)^\beta. \tag{9}$$

CHECKING CONDITION (I) OF PROPOSITION 2

First we have to define our semi-metric $\bar{\rho}(.,.)$. Note that for any target measure Q^σ for $\sigma \in \{-1, 1\}^m$, as they all have the same marginal Q_X, we have the following equality for any given classifier h:

$$\mathcal{E}_{Q^\sigma}(h) = 2\mathbb{E}_{Q_X}\left[\left|\eta(X) - \frac{1}{2}\right| \mathbb{1}_{h(X) \neq h_\sigma^*(X)}\right] = C'_\alpha r^\alpha Q_X\left(\{h(X) \neq h_\sigma^*(X)\} \cap \bigcup_{z \in \mathcal{Z}_1} B(z, r/6)\right), \tag{10}$$

where h_σ^* is the Bayes classifier derived from the regression function η_σ. Hence, following the notations of Proposition 2 let $\mathcal{F}$ be the space of all classifiers, that is of all measurable functions from $\mathcal{X}$ to $\{0, 1\}$. We can define the following semi-distance on $\mathcal{F}$:

$$\forall h, h' \in \mathcal{F}, \quad \bar{\rho}(h, h') \doteq C'_\alpha r^\alpha Q_X\left(\{h(X) \neq h'(X)\} \cap \bigcup_{z \in \mathcal{Z}_1} B(z, r/6)\right).$$

Note that we have:

$$\forall \sigma, \sigma' \in \{-1, 1\}^m, \quad \bar{\rho}\left(h_\sigma^*, h_{\sigma'}^*\right) = C_\alpha' r^\alpha w \rho_H\left(\sigma, \sigma'\right),$$

where $\rho_H\left(\sigma, \sigma'\right) \doteq \text{card}(\{z \in \mathcal{Z}_1 : \sigma(z) \neq \sigma'(z)\})$ is the Hamming distance.

Notice that because all of the h_σ^*'s are different from each other, we could actually re-index the sample distributions Π_σ by their respective Bayes classifiers h_σ^* and, as in Proposition 2, we could set $\mathcal{H} = \{h_\sigma^* : \sigma \in \{-1, 1\}^m\}$. Now we would like to select $M + 1$ of the 2^m above sample distributions Π_σ such that the distances of their respective Bayes classifiers, in term of $\bar{\rho}(.,.)$, are as high as possible. We want also M to be as large as possible so that inequality (ii) in Proposition 2 is verified. This dilemma is solved thanks to Lemma 5.

Call $(P^i, Q^i) \doteq (P^{\sigma_i}, Q^{\sigma_i})$, where the σ_i's are the ones deduced from Lemma 5. Let's also define h_i^*, for $0 \le i \le M$, their related Bayes classifiers and their respective full sample distributions:

$$\Pi_i \doteq P^{i \otimes n_P} \otimes Q^{i \otimes n_Q}.$$

Note that for this selected family of distributions, the condition (i) of Proposition 2 is satisfied as follows for some $c > 0$ independent of n_P and n_Q:

$$\forall 0 \le i < j \le M, \quad \bar{\rho}\left(h_i^*, h_j^*\right) \ge C_\alpha' \frac{wmr^\alpha}{8} \doteq 2s \ge 2c\left(n_P^{d_0/(d_0 + \gamma/\alpha)} + n_Q\right)^{-(\beta+1)/d_0}. \quad (11)$$

CHECKING CONDITION (II) OF PROPOSITION 2

Note that by independence we have for $i \in \{1, \dots, M\}$:

$$\mathcal{D}_{\text{kl}}\left(\Pi_i | \Pi_0\right) = n_P \mathcal{D}_{\text{kl}}\left(P^i | P^0\right) + n_Q \mathcal{D}_{\text{kl}}\left(Q^i | Q^0\right).$$

Notice also that because $C_\alpha' \le 1/2$, we have that every regression functions η_{σ_i} are in $[1/4, 3/4]$. This implies that $\forall i$, $P^i \ll P^0$ and $Q^i \ll Q^0$ as all these distributions have the same marginals P_X and Q_X respectively. Hence, we get:

$$
\begin{aligned}
\mathcal{D}_{\text{kl}}\left(P^i | P^0\right) &= \int \log\left(\frac{\mathrm{d}P^i}{\mathrm{d}P^0}\right) \mathrm{d}P^i \\
&= \int \log\left(\frac{\eta_i(x)}{\eta_0(x)}\right) \eta_i(x) + \log\left(\frac{1 - \eta_i(x)}{1 - \eta_0(x)}\right) (1 - \eta_i(x)) \, \mathrm{d}P_X^i(x) \\
&= \sum_{z : \sigma_i(z) \neq \sigma_0(z)} P_X(B(z, r/6)) \left[\log\left(\frac{1 + C_\alpha' r^\alpha}{1 - C_\alpha' r^\alpha}\right) \frac{1 + C_\alpha' r^\alpha}{2} + \log\left(\frac{1 - C_\alpha' r^\alpha}{1 + C_\alpha' r^\alpha}\right) \frac{1 - C_\alpha' r^\alpha}{2}\right] \\
&= w r^\gamma \sum_{z : \sigma_i(z) \neq \sigma_0(z)} \log\left(\frac{1 + C_\alpha' r^\alpha}{1 - C_\alpha' r^\alpha}\right) \frac{1 + C_\alpha' r^\alpha}{2} + \log\left(\frac{1 - C_\alpha' r^\alpha}{1 + C_\alpha' r^\alpha}\right) \frac{1 - C_\alpha' r^\alpha}{2} \\
&= \rho_H\left(\sigma_i, \sigma_0\right) w \log\left(\frac{1 + C_\alpha' r^\alpha}{1 - C_\alpha' r^\alpha}\right) C_\alpha' r^{\alpha+\gamma} \le 2mw C_\alpha'^2 r^{2\alpha+\gamma}/(1 - C_\alpha' r^\alpha) \le 4mw C_\alpha'^2 r^{2\alpha+\gamma},
\end{aligned}
$$

as $C_\alpha' \le 1/2$ and $r \le 1$. On the other hand, following the same steps we get:

$$\mathcal{D}_{\text{kl}}\left(Q^i | Q^0\right) \le 4mw C_\alpha'^2 r^{2\alpha}.$$

21

But as $c_r \leq 1$ we have:

$$r^\gamma = c_r^\gamma (n_P^{d_0/(d_0+\gamma/\alpha)} + n_Q)^{-\gamma/(\alpha d_0)} \leq (n_P^{d_0/(d_0+\gamma/\alpha)})^{-\gamma/(\alpha d_0)} = n_P^{-\gamma/(\alpha d_0+\gamma)}.$$

Therefore we get:

$$
\begin{aligned}
\mathcal{D}_{\mathrm{kl}}\left(\Pi_i | \Pi_0\right) &\leq 4mw C_\alpha'^2 r^{2\alpha}(n_Q + n_P^{1-\gamma/(\alpha d_0+\gamma)}) \\
&= 4mw C_\alpha'^2 r^{2\alpha}(n_Q + n_P^{d_0/(d_0+\gamma/\alpha)}) \\
&= 4m c_w C_\alpha'^2 r^{\alpha d_0}(n_Q + n_P^{d_0/(d_0+\gamma/\alpha)}) \\
&= 4 c_r^{\alpha d_0} c_w C_\alpha'^2 m \leq 2^5 \log(2)^{-1} c_r^{\alpha d_0} c_w C_\alpha'^2 \log(M).
\end{aligned}
\tag{12}
$$

With c_w small enough, the constant in front of $\log(M)$ is below $1/8$, and hence the last condition of Proposition 2 is verified.

CHOOSING THE CONSTANT c_w AND CONCLUSION

To conclude we just need to take c_w as small as necessary so that inequalities (8) and (9) are satisfied and the constant in the R.H.S. of equation (12) is below $1/8$. Note that this value will be independent of n_P and n_Q.

The lower bound is obtained by applying Proposition 2 on the family of sample distributions $\{\Pi_\sigma\}_{\sigma \in \{-1,1\}^m}$, that can be re-indexed by there respective Bayes classifiers h_σ^*. Note that by equality (10), for any classifier h, we can replace $\bar{\rho}(h, h_i^*)$ by the excess error of h. Finally, as this family is a (finite) subfamily of the class $\mathcal{T}_{(DM)}$, the lower bound obtained from Proposition 2 would work also for $\mathcal{T}_{(DM)}$. That is, for any classifier $\hat{h}$ built upon $((\mathbf{X}, \mathbf{Y}))$ we have:

$$
\sup_{(P,Q) \in \mathcal{T}_{(DM)}} \mathbb{P}_{(\mathbf{X},\mathbf{Y})} \left(\mathcal{E}_Q(\hat{h}) \geq s \right) \geq \sup_{\sigma \in \{-1,1\}^m} \Pi_\sigma \left(\mathcal{E}_{Q^\sigma}(h) \geq s \right) \geq \frac{3 - 2\sqrt{2}}{8},
$$

where $s = c \left(n_P^{d_0/(d_0+\gamma/\alpha)} + n_Q \right)^{-(\beta+1)/d_0}$. By Markov inequality, we therefore also get the lower bound in expectation as in the statement of Theorem 1. ∎

Remark 6 *In the above proof, we arrive at a coefficient $C_d > 0$, independent of n_P and n_Q; however, a small technicality, alluded to earlier, is that the class $\mathcal{T}_{(DM)}$ specifies a fixed C_d which might differ from the one arrived at. If the one arrived at is larger, we're fine. Otherwise, to make C_d as large as specified for a given $\mathcal{T}_{(DM)}$, we can easily amend the construction by instead considering a smaller hypercube of side length $D < 1$ included in $\mathcal{X}$ (the remaining space having zero probability under Q_X). We would therefore focus on subdivisions of size $r_D = r \cdot D$; by keeping m and w unchanged (that is, still depending only on r), equation (8) would now read: $q_0 \geq (1 - c_w c_m) D^{-d} > 0$ and $q_1 \geq c_w D^{-d} > 0$. We can then take D small enough (and independent of n_P and n_Q) to appropriately increase the densities so as to satisfy the doubling measure assumption with any given C_d.*

The same approach works towards obtaining any desired coefficient C_γ from Definition 3.

A.2. Lower Bound for $\mathcal{T} = \mathcal{T}_{\textbf{(BCN)}}$ when $\gamma < \infty$

Proposition 4 *Let $(\mathcal{X}, \rho) = ([0,1]^d, \|.\|_\infty)$, for some $d \in \mathbb{N}^*$, and assume that $\gamma < \infty$. There exists a constant $c = c(\mathcal{T}_{(BCN)})$ such that, for any classifier $\hat{h}$ learned on $(\mathbf{X}, \mathbf{Y})$ and with knowledge of P_X, Q_X, we have:*

$$\sup_{(P,Q) \in \mathcal{T}_{(BCN)}} \mathbb{E}_{(\mathbf{X},\mathbf{Y})}[\mathcal{E}_Q(\hat{h})] \geq c \left(n_P^{d_0/(d_0 + \gamma/\alpha)} + n_Q \right)^{-(\beta+1)/d_0},$$

where $d_0 = 2 + \beta + d/\alpha$.

Proof The proof of the lower bound for the (BCN) regime follows almost all the same lines as the above lower bound proof of Proposition 3 for the (DM)regime. The only difference is that we don't have to satisfy the doubling measure assumption for Q_X and hence we don't need the densities to be bounded away from zero independently of n_P and n_Q as in equation (8). In this case, we can set r, m and w as follows:

$$r = c_r \left(n_P^{d_0/(d_0 + \gamma/\alpha)} + n_Q \right)^{-1/(\alpha d_0)}, \quad m = \left\lfloor c_m r^{-d} \right\rfloor, \quad w = c_w r^{d + \alpha\beta},$$

where $c_r = 1/9$, $c_m = (8/9)^d$ and $c_w \in (0,1]$, implying that $8 \leq m < \lfloor r^{-1} \rfloor^d$ and $mw < 1$. After, all the steps are identical to the lower bound proof for (DM), with the difference that $d_0 = 2 + \beta + d/\alpha$ here. In particular, equations (9) and (12) are unchanged, and therefore we just need to take c_w small enough so that the Tsybakov noise assumption (Definition 5) and condition (ii) from Proposition 2 are satisfied. Finally, from equation (11), we have again:

$$\forall 0 \leq i < j \leq M, \quad \bar{\rho}\left(h_i^*, h_j^* \right) \geq C_\alpha' \frac{wmr^\alpha}{8} \doteq 2s \geq 2c \left(n_P^{d_0/(d_0 + \gamma/\alpha)} + n_Q \right)^{-(\beta+1)/d_0},$$

where $c > 0$. Thus, by applying Proposition 2, we get the desired lower bound. [4] ∎

A.3. Lower Bounds when $\gamma = \infty$

Proposition 5 *Let $(\mathcal{X}, \rho) = ([0,1]^d, \|.\|_\infty)$, for some $d \in \mathbb{N}^*$ and consider $\gamma = \infty$. Let $\mathcal{T}$ denote either $\mathcal{T}_{(DM)}$ or $\mathcal{T}_{(BCN)}$. For $\mathcal{T} = \mathcal{T}_{(DM)}$ assume further that $\alpha\beta \leq d$. There exists a constant $c = c(\mathcal{T})$ such that, for any classifier $\hat{h}$ learned on $(\mathbf{X}, \mathbf{Y})$ and with knowledge of P_X, Q_X, we have:*

$$\sup_{(P,Q) \in \mathcal{T}} \mathbb{E}_{(\mathbf{X},\mathbf{Y})}[\mathcal{E}_Q(\hat{h})] \geq c \left(1 + n_Q \right)^{-(\beta+1)/d_0},$$

where $d_0 = 2 + d/\alpha$ when $\mathcal{T} = \mathcal{T}_{(DM)}$, and $d_0 = 2 + \beta + d/\alpha$ when $\mathcal{T} = \mathcal{T}_{(BCN)}$.

Proof As the proof of the lower bound for $\gamma = \infty$ is again quite similar to the previous ones, we treat both regimes (BCN) and (DM) simultaneously, by taking $d_0 = 2 + d/\alpha$ when $\mathcal{T} = \mathcal{T}_{(DM)}$ and $d_0 = 2 + \beta + d/\alpha$ when $\mathcal{T} = \mathcal{T}_{(BCN)}$. Actually, the main difference is the choice the source

---

4. Notice that, Definition 7 might not be satisfied for some pre-selected coefficients C_d that are too low. Again, as discussed in the previous part, this can also be solved by doing the construction of the distributions over a smaller hypercube included in $\mathcal{X}$. See Remark 6.

marginal P_X. Notice that because $\gamma = \infty$ we have no restriction on the choice of such a probability measure. In particular, we could set the density of P_X being equal to zero on $\mathcal{X}_Q$, and the proof would be even more direct. However, we do the proof of the lower bound with $p_0, p_1 > 0$ to show that, indeed, the lower bound even holds for the situations where we have both $\mathcal{X}_Q \subset \mathcal{X}_P$ and $\gamma = \infty$. For $c_r = 1/9$, we set:

$$r = c_r \left(1 + n_Q\right)^{-1/(\alpha d_0)},$$

and w and m are defined as in the previous proofs. The construction of the marginal Q_X remains also the same. Recall that q_1 is the density of Q_X w.r.t. Lebesgue measure on each set $B(z, r/6)$ for $z \in \mathcal{Z}_1$. We define P_X as having density p_1 on these sets as follows:

$$p_1 = \frac{q_1}{n_P}.$$

Furthermore, as before, we set P_X as being uniformly distributed on $B(z, r/2) \backslash B(z, r/3)$ for each $z \in \mathcal{Z}_1$ so that $P_X(B(z, r/2)) = Q_X(B(z, r/2))$, and again we still set P_X to have the same density $p_0 = q_0$ than Q_X on the hypercubes $B(z, r/2)$ for all $z \in \mathcal{Z}_0$. The following lines of the proof remain identical to the ones in the previous proofs, until the computation of the Kullback-Leibler divergence which becomes instead:

$$\forall i \in \{1, \ldots, M\}, \quad \mathcal{D}_{\mathrm{kl}}\left(P^i | P^0\right) = n_P^{-1} \rho_H\left(\sigma_i, \sigma_0\right) w \log\left(\frac{1 + C'_\alpha r^\alpha}{1 - C'_\alpha r^\alpha}\right) C'_\alpha r^\alpha$$

$$\leq 2n_P^{-1} m w C'^2_\alpha r^{2\alpha}/(1 - C'_\alpha r^\alpha) \leq 4 n_P^{-1} m w C'^2_\alpha r^{2\alpha}.$$

Hence, equation (12) becomes:

$$\forall i \in \{1, \ldots, M\}, \quad \mathcal{D}_{\mathrm{kl}}\left(\Pi_i | \Pi_0\right) \leq 4 m w C'^2_\alpha r^{2\alpha}(n_Q + 1)$$

$$= 4 c_r^{\alpha d_0} c_w C'^2_\alpha m \leq 2^5 \log(2)^{-1} c_r^{\alpha d_0} c_w C'^2_\alpha \log(M).$$

Note also that equation (11) is now as follows:

$$\forall 0 \leq i < j \leq M, \quad \bar{\rho}\left(h_i^*, h_j^*\right) \geq C'_\alpha \frac{w m r^\alpha}{8} \doteq 2s \geq 2c\left(1 + n_Q\right)^{-(\beta+1)/d_0}.$$

Thus, by taking again c_w small enough such that all conditions and assumptions are satisfied, we can apply Proposition 2 to get the lower bound of Theorem 1 for the case where $\gamma = \infty$. ∎

Appendix B. Upper Bound Analysis

B.1. Supporting Lemmas

Lemma 7 (Basic inequalities) *We have the following inequalities:*

1. Take $\alpha \geq 1$ and $a, b \geq 0$, then:

$$a^\alpha + b^\alpha \leq (a + b)^\alpha \leq 2^{\alpha-1}(a^\alpha + b^\alpha).$$

2. *Take $\alpha, \alpha' > 0$ and $a, b \geq 0$ such that $a + b > 0$. Then if $\alpha \geq 1$:*

$$\left(\frac{1}{a^{\alpha/\alpha'} + b}\right)^{\frac{1}{\alpha}} \geq \frac{1}{a^{1/\alpha'} + b^{1/\alpha}} \geq \frac{1}{2}\left(\frac{1}{a^{\alpha/\alpha'} + b}\right)^{\frac{1}{\alpha}},$$

and when $\alpha < 1$, we have:

$$2^{1-1/\alpha}\frac{1}{a^{1/\alpha'} + b^{1/\alpha}} \leq \left(\frac{1}{a^{\alpha/\alpha'} + b}\right)^{\frac{1}{\alpha}} \leq \frac{1}{a^{1/\alpha'} + b^{1/\alpha}}.$$

3. *Take $\alpha_1, \alpha_2, \beta_1, \beta_2 > 0$ and $a, b \geq 0$ such that $a + b > 0$ and $\alpha_1\beta_1 \leq 1$. Assume $\alpha_2 - \alpha_1 = \frac{1}{\beta_2} - \frac{1}{\beta_1}$. Then, for $c = \max(a^{\beta_1}, b^{\beta_2})^{-1}$ we have:*

$$\frac{1}{ac^{\alpha_1} + bc^{\alpha_2}} \leq \frac{2}{a^{1-\alpha_1\beta_1} + b^{1-\alpha_2\beta_2}}.$$

Proof Inequalities (a) are well-known and inequalities (b) are direct consequences of the later. So we need just to prove inequality (c). Note that the cases $a = 0$ or $b = 0$ are trivial, so we can restrict ourselves to the situation where both $a > 0$ and $b > 0$. Plugging in the expression of ϵ we get:

$$\frac{1}{a\epsilon^{\alpha_1} + b\epsilon^{\alpha_2}} = \frac{1}{\min(a^{1-\alpha_1\beta_1}, ab^{-\alpha_1\beta_2}) + \min(ba^{-\alpha_2\beta_1}, b^{1-\alpha_2\beta_2})}.$$

Note that:

$$a^{1-\alpha_1\beta_1} \leq ab^{-\alpha_1\beta_2} \Leftrightarrow a^{-\alpha_1\beta_1} \leq b^{-\alpha_1\beta_2} \Leftrightarrow a^{-\beta_1} \leq b^{-\beta_2} \qquad (13)$$
$$\Leftrightarrow a^{-\alpha_2\beta_1} \leq b^{-\alpha_2\beta_2} \Leftrightarrow ba^{-\alpha_2\beta_1} \leq b^{1-\alpha_2\beta_2}.$$

This means that $a^{1-\alpha_1\beta_1}$ is minimum in the left component of the denominator if and only if $ba^{-\alpha_2\beta_1}$ is minimum in the right component. First, assume that it is $a^{1-\alpha_1\beta_1}$ the minimum in the left component. Recall that $\alpha_1\beta_1 \leq 1$ and $\alpha_2 - \alpha_1 = \frac{1}{\beta_2} - \frac{1}{\beta_1} \Leftrightarrow \frac{\beta_2}{\beta_1} - \alpha_1\beta_2 = 1 - \alpha_2\beta_2$. In this case, from equation (13) we have:

$$a^{-\beta_1} \leq b^{-\beta_2} \Leftrightarrow a \geq b^{\beta_2/\beta_1} \Rightarrow a^{1-\alpha_1\beta_1} \geq b^{\frac{\beta_2}{\beta_1} - \alpha_1\beta_2} = b^{1-\alpha_2\beta_2}.$$

Hence, we can notice that $a^{1-\alpha_1\beta_1} \geq \frac{1}{2}a^{1-\alpha_1\beta_1} + \frac{1}{2}b^{1-\alpha_2\beta_2}$. This lead us to the result:

$$\frac{1}{a\epsilon^{\alpha_1} + b\epsilon^{\alpha_2}} \leq \frac{1}{a^{1-\alpha_1\beta_1}} \leq \frac{2}{a^{1-\alpha_1\beta_1} + b^{1-\alpha_2\beta_2}}.$$

Now assume that $b^{1-\alpha_2\beta_2}$ is strictly the minimum in the right component (recall that, by (13), this is equivalent to $a^{-\beta_1} > b^{-\beta_2}$), we have:

$$a^{-\beta_1} > b^{-\beta_2} \Leftrightarrow a < b^{\beta_2/\beta_1} \Rightarrow a^{1-\alpha_1\beta_1} \leq b^{\frac{\beta_2}{\beta_1} - \alpha_1\beta_2} = b^{1-\alpha_2\beta_2}.$$

Therefore, again we have $b^{1-\alpha_2\beta_2} \geq \frac{1}{2}a^{1-\alpha_1\beta_1} + \frac{1}{2}b^{1-\alpha_2\beta_2}$ from which we can conclude:

$$\frac{1}{a\epsilon^{\alpha_1} + b\epsilon^{\alpha_2}} \leq \frac{1}{b^{1-\alpha_2\beta_2}} \leq \frac{2}{a^{1-\alpha_1\beta_1} + b^{1-\alpha_2\beta_2}}.$$

■

Lemma 8 (Bounding the bias using the implicit 1-NNs) *Fix $x \in \mathcal{X}$. Let $\{X_{(i)}\}_{i=1}^{k}$ its k nearest-neighbors as in Definition 9 and $\{\tilde{X}_i\}_{i=1}^{k}$ its k implicit 1-NNs from Definition 12. We have the following inequality:*

$$\sum_{i=1}^{k} \rho(X_{(i)}, x)^\alpha \leq \sum_{i=1}^{k} \rho(\tilde{X}_i, x)^\alpha.$$

Proof Actually the proof of the inequality can be done for any subset of size k of the data $(\mathbf{X}, \mathbf{Y})$. That is, for any $\{X_i'\}_{i=1}^{k} \subset \{X_i\}_{i=1}^{n_P+n_Q}$ we have:

$$\sum_{i=1}^{k} \rho(X_{(i)}, x)^\alpha \leq \sum_{i=1}^{k} \rho(X_i', x)^\alpha. \tag{14}$$

Indeed, assume WLOG that $\rho(X_1', x) \leq \ldots \leq \rho(X_k', x)$. Then, X_i' is in fact the ith nearest neighbor of x from $\{X_i'\}_{i=1}^{k}$, while $X_{(i)}$ is its ith nearest neighbor from $\{X_i\}_{i=1}^{n_P+n_Q}$. As $\{X_i'\}_{i=1}^{k} \subset \{X_i\}_{i=1}^{n_P+n_Q}$, this clearly implies that $\forall i \in \{1, \ldots, k\}$, $\rho(X_{(i)}, x) \leq \rho(X_i', x)$. Inequality (14) is a direct consequence of this.

∎

B.2. Proof of Theorem 2

The main arguments are given here inline, and require bias and variance bounds we establish in subsequent sections.

Recall the bound on the excess error of $\hat{h}_k$ derived in Proposition 1:

$$\mathbb{E}_{(\mathbf{X},\mathbf{Y})}[\mathcal{E}_Q(\hat{h}_k)] \leq \mathbb{E}[\Phi_1(X)] + \mathbb{E}[\Phi_2(X)] + \mathbb{E}[\Phi_3(X)].$$

Under (DM), we bound the terms $\mathbb{E}[\Phi_1(X)] + \mathbb{E}[\Phi_2(X)]$ using Lemma 11 and the term $\mathbb{E}[\Phi_3(X)]$ using Lemma 12. We get the following bound under the (DM) regime:

$$\mathbb{E}_{(\mathbf{X},\mathbf{Y})}[\mathcal{E}_Q(\hat{h}_k)] \leq C_1 \left(\frac{1}{\sqrt{k}}\right)^{\beta+1} + C_2 \left(\left\lfloor\frac{n_P}{k}\right\rfloor^{(d_0-2)/((d_0-2)+\gamma/\alpha)} + \left\lfloor\frac{n_Q}{k}\right\rfloor\right)^{-(\beta+1)/(d_0-2)},$$

where $C_1, C_2 > 0$ are two constants and $d_0 = 2 + d/\alpha$.

Under (BCN), we also bound the terms $\mathbb{E}[\Phi_1(X)] + \mathbb{E}[\Phi_2(X)]$ using Lemma 11, but the term $\mathbb{E}[\Phi_3(X)]$ is now bounded in Lemma 14 using another approach than for the (DM) case. Applying this lemma, we get the following bound under the (BCN) regime:

$$\mathbb{E}_{(\mathbf{X},\mathbf{Y})}[\mathcal{E}_Q(\hat{h}_k)] \leq C_1 \left(\frac{1}{\sqrt{k}}\right)^{\beta+1} + C_2 \left(\left\lfloor\frac{n_P}{k}\right\rfloor^{(d_0-2)/((d_0-2)+\gamma/\alpha)} + \left\lfloor\frac{n_Q}{k}\right\rfloor\right)^{-(\beta+1)/(d_0-2)},$$

where $d_0 = 2+\beta+d/\alpha$, $C_1, C_2 > 0$ are two constants with the exception that $C_2 = C_2 \log(2(n_P + n_Q))$ when $\alpha = d$.

The upperbounds of Theorem 2 are deduced by just plugging in the value of $k = \Theta(n_P^{d_0/(d_0+\gamma/\alpha)} + n_Q)^{2/d_0}$ (where d_0 is defined as in Theorem 2), and where k is also chosen such that $1 \leq k \leq n_P \vee n_Q$. To see that the above setting of k indeed yields the rates of Theorem 2, a bit of nontrivial algebra is required. This is handled in Lemma 9 below.

Lemma 9 (Plugging in the value of k) *Let the exponent $d_0 > 2$ be as defined in Theorem 2, that is, $d_0 = 2 + d/\alpha$ when $\mathcal{T} = \mathcal{T}_{(DM)}$, and $d_0 = 2 + \beta + d/\alpha$ when $\mathcal{T} = \mathcal{T}_{(BCN)}$. Recall that $1 \leq k \leq n_P \vee n_Q$. Suppose that for some constant $C_1 > 0$, k is upper-bounded as*

$$k \leq C_1 \left(n_P^{d_0/(d_0+\gamma/\alpha)} + n_Q \right)^{2/d_0}.$$

Then, for some constant $C_2 > 0$ we have that:

$$\left(\left\lfloor \frac{n_P}{k} \right\rfloor^{(d_0-2)/((d_0-2)+\gamma/\alpha)} + \left\lfloor \frac{n_Q}{k} \right\rfloor \right)^{-1/(d_0-2)} \leq C_2 (n_P^{d_0/(d_0+\gamma/\alpha)} + n_Q)^{-1/d_0}. \quad (15)$$

Proof From result 2 of Lemma 7, note that proving the bound (15) is in fact equivalent to proving that there exists a constant $C_2 > 0$ such that:

$$\left(\left\lfloor \frac{n_P}{k} \right\rfloor^{1/((d_0-2)+\gamma/\alpha)} + \left\lfloor \frac{n_Q}{k} \right\rfloor^{1/(d_0-2)} \right)^{-1} \leq C_2 \left(n_P^{1/(d_0+\gamma/\alpha)} + n_Q^{1/d_0} \right)^{-1}.$$

Notice also that, as $1 \leq k \leq n_P \vee n_Q$, we can find a lower bound on the floor values such that:

$$\left\lfloor \frac{n_P}{k} \right\rfloor^{1/((d_0-2)+\gamma/\alpha)} + \left\lfloor \frac{n_Q}{k} \right\rfloor^{1/(d_0-2)} \geq \frac{1}{3} \left(\left(\frac{n_P}{k} \right)^{1/((d_0-2)+\gamma/\alpha)} + \left(\frac{n_Q}{k} \right)^{1/(d_0-2)} \right),$$

so we just need to bound $((n_P/k)^{1/((d_0-2)+\gamma/\alpha)} + (n_Q/k)^{1/(d_0-2)})^{-1}$ directly. In order to use Lemma 7, remark that we can rewrite the upper bound on k as follows:

$$k \leq 2^{2/d_0} C_1 \max \left(n_P^{2/(d_0+\gamma/\alpha)}, n_Q^{2/d_0} \right).$$

We can use finally result 3 from Lemma 7 by setting: $\alpha_1 = 1/((d_0 - 2) + \gamma/\alpha)$ and $\alpha_2 = 1/(d_0 - 2)$; $\beta_1 = 2(\alpha_1)^{-1}/(d_0 + \gamma/\alpha)$ and $\beta_2 = 2(\alpha_2)^{-1}/d_0$; $a = n_P^{\alpha_1}$ and $b = n_Q^{\alpha_2}$. Now let's check the sufficient conditions to apply result 3, we have:

$$1 - \alpha_1\beta_1 = \frac{(d_0 - 2) + \gamma/\alpha}{d_0 + \gamma/\alpha} = \frac{\alpha_1^{-1}}{d_0 + \gamma/\alpha} \geq 0,$$

$$1 - \alpha_2\beta_2 = \frac{(d_0 - 2)}{d_0} = \frac{\alpha_1^{-1}}{d_0},$$

$$\frac{1}{\beta_2} - \frac{1}{\beta_1} = \frac{d_0}{2}\alpha_2 - \frac{d_0 + \gamma/\alpha}{2}\alpha_1 = \alpha_2 - \alpha_1 + \frac{(d_0 - 2)}{2}\alpha_2 - \frac{(d_0 - 2) + \gamma/\alpha}{2}\alpha_1 = \alpha_2 - \alpha_1.$$

Therefore we conclude by using inequality (3) from Lemma 7. For some constant $C_2 > 0$, we have:

$$\left(\left(\frac{n_P}{k} \right)^{1/((d_0-2)+\gamma/\alpha)} + \left(\frac{n_Q}{k} \right)^{1/(d_0-2)} \right)^{-1} \leq C_2 \left(a^{1-\alpha_1\beta_1} + b^{1-\alpha_2\beta_2} \right)^{-1}$$

$$\leq C_2 \left(n_P^{1/(d_0+\gamma/\alpha)} + n_Q^{1/d_0} \right)^{-1}.$$

∎

B.3. Bounding $\mathbb{E}[\Phi_1(X)]$ and $\mathbb{E}[\Phi_2(X)]$

The following is a generalization of a chaining argument of (Audibert and Tsybakov, 2007, Lemma 3.1) adapted to our setting. In particular, in their result, the counterpart for the function G_k is the *regression error of a generic estimator*; here we essentially extend their techniques to any function G_k depending on k.

Lemma 10 (A generic chaining argument) *Fix Q that has noise parameters $\beta, C_\beta > 0$ (see Definition 6). Let $\{G_k((\mathbf{X}, \mathbf{Y}); X)\}_{k=1}^{n_P \vee n_Q}$ a set of measurable functions of $(\mathbf{X}, \mathbf{Y})$ and X indexed by k, where $X \sim Q_X$ independent of $(\mathbf{X}, \mathbf{Y})$. Suppose that there exist $C, c > 0$, such that:*

$$\forall x \in \mathcal{X}, \forall k \geq 1, \forall \epsilon > 0, \quad \mathbb{P}_{(\mathbf{X}, \mathbf{Y})}(G_k((\mathbf{X}, \mathbf{Y}), x) \geq \epsilon) \leq C \exp(-ck\epsilon^2).$$

Then the below expectation (taken w.r.t. both $(\mathbf{X}, \mathbf{Y})$ and X) is bounded as follows:

$$\mathbb{E}\left[\left|\eta(X) - \frac{1}{2}\right| \mathbb{1}\left\{\left|\eta(X) - \frac{1}{2}\right| \leq G_k((\mathbf{X}, \mathbf{Y}), X)\right\}\right] \leq 3C \cdot C_\beta \left(\frac{\beta + 1}{ck}\right)^{(\beta+1)/2}. \tag{16}$$

Proof By using Fubini theorem and then the bound assumed in the lemma, we have:

$$\mathbb{E}\left[\left|\eta(X) - \frac{1}{2}\right| \mathbb{1}\left\{\left|\eta(X) - \frac{1}{2}\right| \leq G_k((\mathbf{X}, \mathbf{Y}), X)\right\}\right]$$

$$= \mathbb{E}_Q\left[\left|\eta(x) - \frac{1}{2}\right| \mathbb{P}_{(\mathbf{X}, \mathbf{Y})}\left(G_k((\mathbf{X}, \mathbf{Y}), X) \geq \left|\eta(X) - \frac{1}{2}\right|\right)\right]$$

$$\leq C\mathbb{E}_Q\left[\left|\eta(X) - \frac{1}{2}\right| e^{-ck|\eta(X) - 1/2|^2}\right].$$

Let $\delta = \sqrt{\frac{\beta+1}{ck}}$ and $\delta_i = i.\delta$ for $i \geq 0$. Call $A_i = \{x : \left|\eta(x) - \frac{1}{2}\right| \in (\delta_i, \delta_{i+1}]\}$. We can decompose the expectation in the above upperbound over the disjoint sets A_i:

$$\mathbb{E}_Q\left[\left|\eta(X) - \frac{1}{2}\right| e^{-ck|\eta(X) - 1/2|^2}\right] = \sum_{i \geq 0} \mathbb{E}_Q\left[\left|\eta(X) - \frac{1}{2}\right| e^{-ck|\eta(X) - 1/2|^2} \mathbb{1}\{X \in A_i\}\right]. \tag{17}$$

Now each term in the above sum can be bounded this way:

$$\mathbb{E}_Q\left[\left|\eta(X) - \frac{1}{2}\right| e^{-ck|\eta(X) - 1/2|^2} \mathbb{1}\{X \in A_i\}\right] \quad \leq \quad \delta_{i+1} e^{-ck\delta_i^2} Q_X\left(\delta_i < |\eta(x) - 1/2| \leq \delta_{i+1}\right)$$

$$\leq \delta(i+1) e^{-ck\delta^2 i^2} Q_X\left(0 < |\eta(x) - 1/2| \leq \delta_{i+1}\right) \quad \leq \quad C_\beta \delta^{\beta+1}(i+1)^{\beta+1} e^{-ck\delta^2 i^2}$$

$$\leq \quad C_\beta \left(\frac{\beta+1}{ck}\right)^{(\beta+1)/2} (i+1)^{\beta+1} e^{-(\beta+1)i^2}, \tag{18}$$

where we used Definition 5 in the third inequality, and replaced δ by its value in the last one. Now, we have:

$$\sum_{i \geq 0} (i+1)^{\beta+1} e^{-(\beta+1)i^2} = \sum_{i \geq 0} e^{-(\beta+1)i^2 + (\beta+1)\log(i+1)} \leq \sum_{i \geq 0} e^{-(\beta+1)i^2 + (\beta+1)i}$$

$$\leq \sum_{i \geq 0} e^{-(\beta+1)i(i-1)} = 1 + \sum_{i \geq 1} e^{-(\beta+1)i(i-1)} \leq 1 + \frac{1}{1 - \exp(-(\beta+1))} \leq 3. \tag{19}$$

Therefore, from equation (17) and using the inequalities from (18) and (19), we finally get inequality (16) of the lemma.

∎

Lemma 11 (Bounding $\mathbb{E}(\Phi_1(X))$ and $\mathbb{E}(\Phi_2(X))$) *Consider Φ_1 and Φ_2 as defined in Proposition 1. Under both (DM) and (BCN) distributional regimes, there exists a constant $C > 0$ such that:*

$$\mathbb{E}[\Phi_1(X)] + \mathbb{E}[\Phi_2(X)] \leq C \left(\frac{1}{\sqrt{k}}\right)^{\beta+1}.$$

Proof First let's start with Φ_1. Let $A_k(x) = A((\mathbf{X}, \mathbf{Y}), x) \doteq \frac{3}{k} \left|\sum_{i=1}^{k} Y_{(i)} - \eta(X_{(i)})\right|$. By the tower property, followed by a Hoeffding inequality, we have that $\forall k \geq 1, \forall x \in \mathcal{X}, \forall \epsilon > 0$:

$$\mathbb{P}_{(\mathbf{X},\mathbf{Y})}(A_k(x) \geq \epsilon) = \mathbb{E}_{\mathbf{X}}\left[\mathbb{P}_{\mathbf{Y}|\mathbf{X}}\left(\frac{3}{k}\left|\sum_{i=1}^{k} Y_{(i)} - \eta(X_{(i)})\right| \geq \epsilon\right)\right] \leq 2\exp\left(-\frac{2}{9}k\epsilon^2\right).$$

Now let $B_k(x) = B((\mathbf{X}, \mathbf{Y}), x)$ denotes the quantity $\frac{3C_\alpha}{k}\sum_{i=1}^{k}\left(\rho(\tilde{X}_i, x)^\alpha - \mathbb{E}_{\tilde{X}_1}\left[\rho(\tilde{X}_1, x)^\alpha\right]\right)$. Again, using a Hoeffding inequality we have that $\forall k \geq 1, \forall x \in \mathcal{X}, \forall \epsilon > 0$:

$$\mathbb{P}_{(\mathbf{X},\mathbf{Y})}(B_k(x) \geq \epsilon) \leq 2\exp\left(-\frac{2k\epsilon^2}{9C_\alpha^2\Delta_{\mathcal{X}}^{2\alpha}}\right).$$

Hence using Lemma 10 for either Φ_1, Φ_2, we get

$$\mathbb{E}[\Phi_1(X)] \leq 12C_\beta \left(\frac{9(\beta+1)}{2k}\right)^{(\beta+1)/2}, \text{ and } \mathbb{E}[\Phi_2(X)] \leq 12C_\beta \left(\frac{9C_\alpha^2\Delta_{\mathcal{X}}^{2\alpha}(\beta+1)}{2k}\right)^{(\beta+1)/2}.$$

∎

B.4. Bounding $\mathbb{E}[\Phi_3(X)]$ under (DM)

Lemma 12 (Bounding $\mathbb{E}(\Phi_3(X))$ under (DM)) *Consider Φ_3 as defined in Proposition 1. Under (DM), there exists a constant $C > 0$ such that*

$$\mathbb{E}[\Phi_3(X)] \leq C \left(\left\lfloor\frac{n_P}{k}\right\rfloor^{\frac{d}{d+\gamma}} + \left\lfloor\frac{n_Q}{k}\right\rfloor\right)^{-\alpha(\beta+1)/d}.$$

Proof Recall that $\Phi_3(x) \doteq 2\left|\eta(x) - \frac{1}{2}\right|\mathbb{1}\left\{\left|\eta(x) - \frac{1}{2}\right| \leq 3C_\alpha\mathbb{E}_{\tilde{X}_1}\left[\rho(\tilde{X}_1, x)^\alpha\right]\right\}$. Take $x \in \mathcal{X}_Q$, we start by rewriting the expectation inside the indicator function as follows:

$$\mathbb{E}_{\tilde{X}_1}\left[\rho(\tilde{X}_1, x)^\alpha\right] = \int_0^{\Delta_{\mathcal{X}}^\alpha} \mathbb{P}_{\tilde{X}}\left(\rho(\tilde{X}_1, x)^\alpha > t\right) dt = \int_0^{\Delta_{\mathcal{X}}^\alpha} \mathbb{P}_{\tilde{X}}\left(\rho(\tilde{X}_1, x) > t^{1/\alpha}\right) dt$$

$$= \int_0^{\Delta_{\mathcal{X}}^\alpha} \left(1 - P_X(B(x, t^{1/\alpha}))\right)^{\left\lfloor\frac{n_P}{k}\right\rfloor} \left(1 - Q_X(B(x, t^{1/\alpha}))\right)^{\left\lfloor\frac{n_Q}{k}\right\rfloor} dt.$$

Now let's recall that Q_X is doubling (see (DM) and Definition 6), that is:

$$\forall x \in \mathcal{X}_Q, \forall r \in (0, \Delta_{\mathcal{X}}], \quad Q_X(B(x,r)) \geq C_d \left(\frac{r}{\Delta_{\mathcal{X}}}\right)^d.$$

Combining it with equation (2) from Definition 3, we have:

$$\forall x \in \mathcal{X}_Q, \forall r \in (0, \Delta_{\mathcal{X}}], \quad P_X(B(x,r)) \geq Q_X(B(x,r))C_\gamma \left(\frac{r}{\Delta_{\mathcal{X}}}\right)^\gamma \geq C_d C_\gamma \left(\frac{r}{\Delta_{\mathcal{X}}}\right)^{d+\gamma}.$$

Thus, note that for any γ (and in particular for $\gamma = \infty$), using only the fact that Q_X is doubling, we can bound the expectation by:

$$\forall x \in \mathcal{X}_Q, \quad \mathbb{E}_{\tilde{X}_1}\left[\rho(\tilde{X}_1, x)^\alpha\right] \leq \int_0^{\Delta_{\mathcal{X}}^\alpha} \left(1 - C_d \left(t\Delta_{\mathcal{X}}^{-\alpha}\right)^{d/\alpha}\right)^{\left\lfloor \frac{n_Q}{k} \right\rfloor} dt. \tag{20}$$

We use this bound to treat the case $\gamma = \infty$. But for the moment assume $\gamma < \infty$. Recall that $C_d, C_\gamma \leq 1$. We get for any $x \in \mathcal{X}_Q$:

$$\mathbb{E}_{\tilde{X}_1}\left[\rho(\tilde{X}_1, x)^\alpha\right] \leq \int_0^{\Delta_{\mathcal{X}}^\alpha} \left(1 - C_d C_\gamma \left(t\Delta_{\mathcal{X}}^{-\alpha}\right)^{(d+\gamma)/\alpha}\right)^{\left\lfloor \frac{n_P}{k} \right\rfloor} \left(1 - C_d \left(t\Delta_{\mathcal{X}}^{-\alpha}\right)^{d/\alpha}\right)^{\left\lfloor \frac{n_Q}{k} \right\rfloor} dt$$

$$\leq \int_0^{\Delta_{\mathcal{X}}^\alpha} \exp\left(-C_d C_\gamma \left(t\Delta_{\mathcal{X}}^{-\alpha}\right)^{(d+\gamma)/\alpha} \left\lfloor \frac{n_P}{k} \right\rfloor - C_d \left(t\Delta_{\mathcal{X}}^{-\alpha}\right)^{d/\alpha} \left\lfloor \frac{n_Q}{k} \right\rfloor\right) dt$$

$$\leq \int_0^{\Delta_{\mathcal{X}}^\alpha} \exp\left(-\left(t(C_d C_\gamma)^{\alpha/d}\Delta_{\mathcal{X}}^{-\alpha}\right)^{(d+\gamma)/\alpha} \left\lfloor \frac{n_P}{k} \right\rfloor - \left(t(C_d C_\gamma)^{\alpha/d}\Delta_{\mathcal{X}}^{-\alpha}\right)^{d/\alpha} \left\lfloor \frac{n_Q}{k} \right\rfloor\right) dt$$

$$\leq \frac{\Delta_{\mathcal{X}}^\alpha}{(C_d C_\gamma)^{\alpha/d}} \int_0^\infty \exp\left(-\left\lfloor \frac{n_P}{k} \right\rfloor t^{(d+\gamma)/\alpha} - \left\lfloor \frac{n_Q}{k} \right\rfloor t^{d/\alpha}\right) dt.$$

To bound this integral we are going to use a chaining argument, that is we are going to discretize this integral. Set $c = \left(\max(\lfloor \frac{n_P}{k} \rfloor^{\alpha/(d+\gamma)}, \lfloor \frac{n_Q}{k} \rfloor^{\alpha/d})\right)^{-1}$. Let's use inequality (3) from Lemma 7. Following the notations of the lemma, set $\alpha_1 = \frac{d+\gamma}{\alpha} = \frac{1}{\beta_1}$ and $\alpha_2 = \frac{d}{\alpha} = \frac{1}{\beta_2}$. Notice that this implies $\alpha_2 - \alpha_1 = \frac{1}{\beta_2} - \frac{1}{\beta_1}$ and $\alpha_1\beta_1 = 1 = \alpha_2\beta_2$. Finally let $a = \lfloor \frac{n_P}{k} \rfloor$ and $b = \lfloor \frac{n_Q}{k} \rfloor$. Recall that $k \leq n_P \vee n_Q$ and hence $a + b > 0$. Applying inequality (3) from Lemma 7:

$$\left\lfloor \frac{n_P}{k} \right\rfloor c^{(d+\gamma)/\alpha} + \left\lfloor \frac{n_Q}{k} \right\rfloor c^{d/\alpha} = ac^{\alpha_1} + bc^{\alpha_2} \geq \frac{1}{2}(a^{1-\alpha_1\beta_1} + b^{1-\alpha_2\beta_2}) \geq \frac{1}{2}. \tag{21}$$

Now let $c_i = i.c$ for $i \geq 0$. We have therefore:

$$\int_0^{\Delta_{\mathcal{X}}^\alpha} \exp\left(-\left\lfloor \frac{n_P}{k} \right\rfloor t^{(d+\gamma)/\alpha} - \left\lfloor \frac{n_Q}{k} \right\rfloor t^{d/\alpha}\right) dt = \sum_{i \geq 0} \int_{c_i}^{c_{i+1}} e^{-\left\lfloor \frac{n_P}{k} \right\rfloor t^{(d+\gamma)/\alpha} - \left\lfloor \frac{n_Q}{k} \right\rfloor t^{d/\alpha}} dt$$

$$\leq c \sum_{i \geq 0} \exp\left(-\left\lfloor \frac{n_P}{k} \right\rfloor c_i^{(d+\gamma)/\alpha} - \left\lfloor \frac{n_Q}{k} \right\rfloor c_i^{d/\alpha}\right) \leq c \sum_{i \geq 0} \exp\left(-\frac{i^{d/\alpha}}{2}\right),$$

where we used equation (21) in the last inequality. As $d > 0$, there exists a constant $C > 0$ such that:

$$\sum_{i \geq 0} \exp\left(-\frac{i^{d/\alpha}}{2}\right) \leq C.$$

Now, for any $x \in \mathcal{X}_Q$ we can bound the expectation as follows:

$$\mathbb{E}_{\tilde{X}_1}\left[\rho(\tilde{X}_1, x)^\alpha\right] \leq \frac{C\Delta_{\mathcal{X}}^\alpha}{(C_d C_\gamma)^{\alpha/d}} c = \frac{C\Delta_{\mathcal{X}}^\alpha}{(C_d C_\gamma)^{\alpha/d}} \left(\max\left(\left\lfloor \frac{n_P}{k} \right\rfloor^{d/(d+\gamma)}, \left\lfloor \frac{n_Q}{k} \right\rfloor\right)\right)^{-\alpha/d}$$

$$\leq \frac{C\Delta_{\mathcal{X}}^\alpha 2^{\alpha/d}}{(C_d C_\gamma)^{\alpha/d}} \left(\left\lfloor \frac{n_P}{k} \right\rfloor^{\frac{d}{d+\gamma}} + \left\lfloor \frac{n_Q}{k} \right\rfloor\right)^{-\alpha/d} \doteq G_k'(n_P, n_Q, x),$$

where we used result (2) from Lemma 7 in the last inequality. Therefore, by using Definition 5, we get:

$$\mathbb{E}_Q[\Phi_3(X)] \leq 2\mathbb{E}_Q\left[\left|\eta(x) - \frac{1}{2}\right| \mathbb{1}\left\{\left|\eta(x) - \frac{1}{2}\right| \leq 3C_\alpha G_k'(n_P, n_Q, x)\right\}\right]$$

$$\leq 2C_\beta \left(\frac{3CC_\alpha \Delta_{\mathcal{X}}^\alpha 2^{\alpha/d}}{(C_d C_\gamma)^{\alpha/d}}\right)^{\beta+1} \left(\left\lfloor \frac{n_P}{k} \right\rfloor^{\frac{d}{d+\gamma}} + \left\lfloor \frac{n_Q}{k} \right\rfloor\right)^{-\alpha(\beta+1)/d}.$$

Finally, the case $\gamma = \infty$ is proved by starting from equation (20) and by following the same steps (and even simpler ones) as above. ∎

B.5. Bounding $\mathbb{E}[\Phi_3(X)]$ under (BCN)

Recall that the main difficulty under (BCN)is that nearest neighbor distances are not uniformly bounded over x. However, they can be bounded on average using the fact that domain $\mathcal{X}_Q$ admits finite covers; such intuition is used for instance in the proof of Lemma 6.4 in Györfi et al. (2006, section 6.3), itself being a special case of Theorem 1 from Kulkarni and Posner (1995). Here, we have to be careful as we consider nearest neighbor distances over a combined sample from two separate distributions. We start with the following result concerning the *tail* (as defined by some ϵ parameter) of such distances.

Lemma 13 (Bounding 1-NN bias) *Assume* $(P, Q) \in \mathcal{T}_{(BCN)}$. *Take* $\epsilon \in (0, 1]$ *and define*

$$A(\epsilon, x) \doteq \int_\epsilon^{\Delta_{\mathcal{X}}^\alpha} \left(1 - P_X(B(x, t^{1/\alpha}))\right)^{\left\lfloor \frac{n_P}{k} \right\rfloor} \left(1 - Q_X(B(x, t^{1/\alpha}))\right)^{\left\lfloor \frac{n_Q}{k} \right\rfloor} dt.$$

Then, there exist two constants $C_1, C_2 > 0$ *such that, when* $\gamma < \infty$*:*

$$\mathbb{E}_Q[A(\epsilon, X)] \leq \begin{cases} C_1 \left(\left\lfloor \frac{n_P}{k} \right\rfloor \epsilon^{(\gamma+d)/\alpha-1} + \left\lfloor \frac{n_Q}{k} \right\rfloor \epsilon^{d/\alpha-1}\right)^{-1}, & \text{for } \alpha < d, \\ C_1(\log(1/\epsilon) + C_2) \left(\left\lfloor \frac{n_P}{k} \right\rfloor \epsilon^{\gamma/\alpha} + \left\lfloor \frac{n_Q}{k} \right\rfloor\right)^{-1}, & \text{for } \alpha = d, \end{cases}$$

and when $\gamma = \infty$*:*

$$\mathbb{E}_Q[A(\epsilon, X)] \leq \begin{cases} C_1 \left\lfloor \frac{n_Q}{k} \right\rfloor^{-1} \epsilon^{-d/\alpha+1}, & \text{for } \alpha < d, \\ C_1(\log(1/\epsilon) + C_2) \left\lfloor \frac{n_Q}{k} \right\rfloor^{-1}, & \text{for } \alpha = d. \end{cases}$$

Proof We can use Fubini theorem to get:

$$\mathbb{E}_Q[A(\epsilon, X)] = \int_{\mathcal{X}_Q} \int_{\epsilon}^{\Delta_{\mathcal{X}}^{\alpha}} \left(1 - P_X(B(x, t^{1/\alpha}))\right)^{\left\lfloor \frac{n_P}{k} \right\rfloor} \left(1 - Q_X(B(x, t^{1/\alpha}))\right)^{\left\lfloor \frac{n_Q}{k} \right\rfloor} \, \mathrm{d}t \, \mathrm{d}Q_X(x)$$

$$= \int_{\epsilon}^{\Delta_{\mathcal{X}}^{\alpha}} \int_{\mathcal{X}_Q} \left(1 - P_X(B(x, t^{1/\alpha}))\right)^{\left\lfloor \frac{n_P}{k} \right\rfloor} \left(1 - Q_X(B(x, t^{1/\alpha}))\right)^{\left\lfloor \frac{n_Q}{k} \right\rfloor} \, \mathrm{d}Q_X(x) \, \mathrm{d}t.$$

Now let's focus on the inner integral. Take $t \in [\epsilon, \Delta_{\mathcal{X}}^{\alpha}]$ and consider a cover of $\mathcal{X}_Q$ composed of balls $(B_i)_{i \in I}$ of diameter $t^{1/\alpha}$ indexed by some set I of size $\mathcal{N}(\mathcal{X}_Q, \rho, t^{1/\alpha})$. We can therefore decompose the above mentioned integral as follows:

$$\int_{\mathcal{X}_Q} \left(1 - P_X(B(x, t^{1/\alpha}))\right)^{\left\lfloor \frac{n_P}{k} \right\rfloor} \left(1 - Q_X(B(x, t^{1/\alpha}))\right)^{\left\lfloor \frac{n_Q}{k} \right\rfloor} \, \mathrm{d}Q_X(x)$$

$$\leq \sum_{i \in I} \int_{\mathcal{X}_Q} \left(1 - P_X(B(x, t^{1/\alpha}))\right)^{\left\lfloor \frac{n_P}{k} \right\rfloor} \left(1 - Q_X(B(x, t^{1/\alpha}))\right)^{\left\lfloor \frac{n_Q}{k} \right\rfloor} \mathbb{1}\{x \in B_i\} \, \mathrm{d}Q_X(x)$$

$$\leq \sum_{i \in I} (1 - P_X(B_i))^{\left\lfloor \frac{n_P}{k} \right\rfloor} (1 - Q_X(B_i))^{\left\lfloor \frac{n_Q}{k} \right\rfloor} Q_X(B_i)$$

$$\leq \sum_{i \in I} \exp\left(-P_X(B_i) \left\lfloor \frac{n_P}{k} \right\rfloor - Q_X(B_i) \left\lfloor \frac{n_Q}{k} \right\rfloor\right) Q_X(B_i)$$

$$\leq \sum_{i \in I} \exp\left(-Q_X(B_i) C_\gamma \left(t^{1/\alpha}/(2\Delta_{\mathcal{X}})\right)^\gamma \left\lfloor \frac{n_P}{k} \right\rfloor - Q_X(B_i) \left\lfloor \frac{n_Q}{k} \right\rfloor\right) Q_X(B_i) \qquad (22)$$

$$\leq \mathcal{N}(\mathcal{X}_Q, \rho, t^{1/\alpha}) e^{-1} \left(C_\gamma \left(t^{1/\alpha}/(2\Delta_{\mathcal{X}})\right)^\gamma \left\lfloor \frac{n_P}{k} \right\rfloor + \left\lfloor \frac{n_Q}{k} \right\rfloor\right)^{-1}$$

$$\leq C_d t^{-d/\alpha} e^{-1} \left(C_\gamma \left(t^{1/\alpha}/(2\Delta_{\mathcal{X}})\right)^\gamma \left\lfloor \frac{n_P}{k} \right\rfloor + \left\lfloor \frac{n_Q}{k} \right\rfloor\right)^{-1},$$

where we used equation (2) from Definition 3 in inequality (22), the fact that for $a > 0$, we have $e^{-ax} x \leq a^{-1} e^{-1}$, $\forall x \geq 0$ in the next inequality and Definition 7 in the last inequality. Note also that when $\gamma = \infty$ inequality (22) can be replaced by:

$$\sum_{i \in I} \exp\left(- \left\lfloor \frac{n_Q}{k} \right\rfloor Q_X(B_i)\right) Q_X(B_i). \qquad (23)$$

But for now, let's assume $\gamma < \infty$. Assuming that $\left\lfloor \frac{n_P}{k} \right\rfloor, \left\lfloor \frac{n_Q}{k} \right\rfloor > 0$ (recall also that $C_\gamma \leq 1$), we get:

$$\mathbb{E}_Q[A(\epsilon, X)] \leq C_d e^{-1} \int_{\epsilon}^{\Delta_{\mathcal{X}}^{\alpha}} \left(C_\gamma t^{(d+\gamma)/\alpha}/(2\Delta_{\mathcal{X}})^\gamma \left\lfloor \frac{n_P}{k} \right\rfloor + t^{d/\alpha} \left\lfloor \frac{n_Q}{k} \right\rfloor\right)^{-1} \, \mathrm{d}t$$

$$\leq C_d e^{-1} \int_{\epsilon}^{\Delta_{\mathcal{X}}^{\alpha}} \min\left(t^{-(d+\gamma)/\alpha} \left(\left\lfloor \frac{n_P}{k} \right\rfloor C_\gamma/(2\Delta_{\mathcal{X}})^\gamma\right)^{-1}, t^{-d/\alpha} \left\lfloor \frac{n_Q}{k} \right\rfloor^{-1}\right) \, \mathrm{d}t.$$

Now assume $d > \alpha$. By switching the integral with the min sign, we get:

$$\mathbb{E}_Q[A(\epsilon, X)] \leq C_d e^{-1} \min\left(\frac{\alpha}{\gamma + d - \alpha} \left(\left\lfloor \frac{n_P}{k} \right\rfloor C_\gamma / (2\Delta_{\mathcal{X}})^\gamma \right)^{-1} \epsilon^{-(d+\gamma)/\alpha+1} , \right. \tag{24}$$

$$\left. \frac{\alpha}{d - \alpha} \left\lfloor \frac{n_Q}{k} \right\rfloor^{-1} \epsilon^{-d/\alpha+1} \right)$$

$$\leq 2 \frac{C_d}{C_\gamma} e^{-1} \frac{\alpha}{d - \alpha} ((2\Delta_{\mathcal{X}})^\gamma \vee 1) \left(\left\lfloor \frac{n_P}{k} \right\rfloor \epsilon^{(d+\gamma)/\alpha-1} + \left\lfloor \frac{n_Q}{k} \right\rfloor \epsilon^{d/\alpha-1} \right)^{-1}.$$

Notice that, by following nearly identical steps, we get the same inequality (or an even tighter one) in the case where either $\frac{n_Q}{k} = 0$ or $\frac{n_P}{k} = 0$. Thus, we arrive at the desired inequality for the case $d > \alpha$ and $\gamma < \infty$. When $d = \alpha$, the r.h.s. of (24) can be bounded as:

$$\mathbb{E}_{X \sim Q_X}[A(\epsilon, X)] \leq C \left(1 \vee (\log(\Delta_{\mathcal{X}}^\alpha) - \log(\epsilon)) \right) \min\left(\left\lfloor \frac{n_P}{k} \right\rfloor^{-1} \epsilon^{-\gamma/\alpha}, \left\lfloor \frac{n_Q}{k} \right\rfloor^{-1} \right)$$

$$\leq 2C \left(1 \vee (\log(\Delta_{\mathcal{X}}^\alpha) - \log(\epsilon)) \right) \left(\left\lfloor \frac{n_P}{k} \right\rfloor \epsilon^{\gamma/\alpha} + \left\lfloor \frac{n_Q}{k} \right\rfloor \right)^{-1},$$

where $C = \frac{C_d}{C_\gamma} e^{-1} ((2\Delta_{\mathcal{X}})^\gamma \vee 1) \left(1 + \frac{\alpha}{\gamma} \mathbb{1}\{\gamma > 0\} \right)$. Again the case where either $\frac{n_Q}{k} = 0$ or $\frac{n_P}{k} = 0$ is handled similarly. Now for $\gamma = \infty$, note that by equations (22) and (23), and following the above intermediary steps, we get that:

$$\mathbb{E}_Q[A(\epsilon, X)] \leq C_d e^{-1} \int_\epsilon^{\Delta_{\mathcal{X}}^\alpha} t^{-d/\alpha} \left\lfloor \frac{n_Q}{k} \right\rfloor^{-1} \mathrm{d}t$$

$$\leq \begin{cases} C_d e^{-1} \frac{\alpha}{d - \alpha} \left\lfloor \frac{n_Q}{k} \right\rfloor^{-1} \epsilon^{-d/\alpha+1}, & \text{for } \alpha < d, \\ C_d e^{-1} (\log(\Delta_{\mathcal{X}}^\alpha) - \log(\epsilon)) \left\lfloor \frac{n_Q}{k} \right\rfloor^{-1}, & \text{for } \alpha = d. \end{cases}$$

$\blacksquare$

Lemma 14 (Bounding $\mathbb{E}(\Phi_3(X))$ under (BCN)) *Consider Φ_3 as defined in Proposition 1. We work under (BCN) regime. Assume $\gamma < \infty$, there exist two constants $C_1, C_2 > 0$ such that, for $d > \alpha$:*

$$\mathbb{E}[\Phi_3(X)] \leq C_1 \left(\left\lfloor \frac{n_P}{k} \right\rfloor^{\frac{(d_0-2)}{(d_0-2)+\gamma/\alpha}} + \left\lfloor \frac{n_Q}{k} \right\rfloor \right)^{-(\beta+1)/(d_0-2)},$$

where $d_0 \doteq d/\alpha + \beta + 2$. And for $d = \alpha$:

$$\mathbb{E}[\Phi_3(X)] \leq C_1 \left(\log\left(\left\lfloor \frac{n_P}{k} \right\rfloor + \left\lfloor \frac{n_Q}{k} \right\rfloor \right) + C_2 \right) \left(\left\lfloor \frac{n_P}{k} \right\rfloor^{\frac{(d_0-2)}{(d_0-2)+\gamma/\alpha}} + \left\lfloor \frac{n_Q}{k} \right\rfloor \right)^{-(\beta+1)/(d_0-2)}.$$

The case $\gamma = \infty$ has the same bounds where one just plug-in $\gamma = \infty$ in these (that is, they will depend only on k and n_Q).

Proof Here we have to be careful about the order of integration as we at times deal with integration under indicator functions (see e.g. equation (25) below). We start with the following decomposition over small nearest neighbor distances and larger ones over a fixed *tail* parameter $\epsilon > 0$:

$$
\begin{aligned}
\mathbb{E}_{\tilde{X}_1}[\rho(\tilde{X}_1, x)^\alpha] &= \int_0^{\Delta_\mathcal{X}^\alpha} \mathbb{P}_{\tilde{X}_1}\left(\rho(\tilde{X}_1, x)^\alpha > t\right) \, \mathrm{d}t \\
&= \int_0^{\Delta_\mathcal{X}^\alpha} \mathbb{P}_{\tilde{X}_1}\left(\rho(\tilde{X}_1, x) > t^{1/\alpha}\right) \, \mathrm{d}t \\
&= \int_0^{\Delta_\mathcal{X}^\alpha} \left(1 - P_X(B(x, t^{1/\alpha}))\right)^{\left\lfloor \frac{n_P}{k} \right\rfloor} \left(1 - Q_X(B(x, t^{1/\alpha}))\right)^{\left\lfloor \frac{n_Q}{k} \right\rfloor} \, \mathrm{d}t \\
&\leq \epsilon + \int_\epsilon^{\Delta_\mathcal{X}^\alpha} \left(1 - P_X(B(x, t^{1/\alpha}))\right)^{\left\lfloor \frac{n_P}{k} \right\rfloor} \left(1 - Q_X(B(x, t^{1/\alpha}))\right)^{\left\lfloor \frac{n_Q}{k} \right\rfloor} \, \mathrm{d}t \\
&\leq \epsilon + A(\epsilon, x),
\end{aligned}
$$

where $A(\epsilon, x)$ denotes the integral in r.h.s. of the previous inequality. We can now use this inequality to bound $\mathbb{E}[\Phi_3(X)]$ as follows:

$$
\begin{aligned}
\mathbb{E}[\Phi_3(X)] &\leq 2\mathbb{E}_Q\left[\left|\eta(X) - \frac{1}{2}\right| \mathbb{1}\left\{\left|\eta(X) - \frac{1}{2}\right| \leq 3C_\alpha(\epsilon + A(\epsilon, X))\right\}\right] \\
&\leq 2\mathbb{E}_Q\left[\left|\eta(X) - \frac{1}{2}\right| \mathbb{1}\left\{\left|\eta(X) - \frac{1}{2}\right| \leq 6C_\alpha\epsilon\right\}\right] \\
&\quad + 2\mathbb{E}_Q\left[\left|\eta(X) - \frac{1}{2}\right| \mathbb{1}\left\{\left|\eta(X) - \frac{1}{2}\right| \leq 6C_\alpha A(\epsilon, X)\right\}\right] \\
&\leq 2C_\beta(6C_\alpha\epsilon)^{\beta+1} + 12C_\alpha\mathbb{E}_Q[A(\epsilon, X)], \tag{25}
\end{aligned}
$$

where we used equation (2) from Definition 3 in the last inequality. We now use Lemma 13 to bound $\mathbb{E}_Q[A(\epsilon, X)]$ for $\epsilon \leq 1$. Assume $\gamma < \infty$. The case $\gamma = \infty$ will be omitted as it follows the same lines (and is in fact more direct). Recall that $\alpha \leq d$ and take:

$$
\epsilon \doteq \left(\max\left(\left\lfloor \frac{n_P}{k} \right\rfloor^{\alpha/(d+\gamma+\beta\alpha)}, \left\lfloor \frac{n_Q}{k} \right\rfloor^{\alpha/(d+\beta\alpha)}\right)\right)^{-1} \leq 1.
$$

Let $a = \left\lfloor \frac{n_P}{k} \right\rfloor$ and $b = \left\lfloor \frac{n_Q}{k} \right\rfloor$. Set $\alpha_1 = \frac{\gamma+d-\alpha}{\alpha}$, $\alpha_2 = \frac{d-\alpha}{\alpha}$, $\beta_1 = \frac{\alpha}{d+\gamma+\beta\alpha}$, $\beta_2 = \frac{\alpha}{d+\beta\alpha}$. Notice that: $\alpha_1\beta_1 = \frac{\gamma+d-\alpha}{\gamma+d+\beta\alpha} \leq 1$ and $\alpha_2 - \alpha_1 = \frac{\gamma}{\alpha} = \frac{1}{\beta_2} - \frac{1}{\beta_1}$. Therefore we can apply Lemma 7 inequality (3) to get:

$$
\begin{aligned}
(a\epsilon^{\alpha_1} + b\epsilon^{\alpha_2})^{-1} &= \left(\left\lfloor \frac{n_P}{k} \right\rfloor \epsilon^{\frac{\gamma+d}{\alpha}-1} + \left\lfloor \frac{n_Q}{k} \right\rfloor \epsilon^{\frac{d}{\alpha}-1}\right)^{-1} \\
&\leq 2\left(\left\lfloor \frac{n_P}{k} \right\rfloor^{\frac{\beta+1}{(d_0-2)+\gamma/\alpha}} + \left\lfloor \frac{n_Q}{k} \right\rfloor^{\frac{\beta+1}{(d_0-2)}}\right)^{-1} = 2\left(a^{1-\alpha_1\beta_1} + b^{1-\alpha_2\beta_2}\right)^{-1}. \tag{26}
\end{aligned}
$$

Therefore, for instance for $d > \alpha$, from equation (25) and by using Lemma 13 and inequality (26), we obtain that there exist $C, C_1, C_2 > 0$ constants such that:

$$\mathbb{E}[\Phi_3(X)] \leq C_1 \left(\max \left(\left\lfloor \frac{n_P}{k} \right\rfloor^{\frac{\beta+1}{(d_0-2)+\gamma/\alpha}}, \left\lfloor \frac{n_Q}{k} \right\rfloor^{\frac{\beta+1}{(d_0-2)}} \right) \right)^{-1} + C_2 \left(\left(\left\lfloor \frac{n_P}{k} \right\rfloor^{\frac{\beta+1}{(d_0-2)+\gamma/\alpha}} + \left\lfloor \frac{n_Q}{k} \right\rfloor^{\frac{\beta+1}{(d_0-2)}} \right) \right)^{-1}$$

$$\leq C \left(\left(\left\lfloor \frac{n_P}{k} \right\rfloor^{\frac{(d_0-2)}{(d_0-2)+\gamma/\alpha}} + \left\lfloor \frac{n_Q}{k} \right\rfloor \right) \right)^{\frac{\beta+1}{(d_0-2)}},$$

where we used result (2) from Lemma 7 in the last inequality.

The case $d = \alpha$ is treated the same way.

$\blacksquare$

Appendix C. Adaptive Labeling Results

C.1. Algorithms

We now present the procedures used for adaptive labeling and choice of k in detail.

Algorithm 1 below builds a set $\mathbf{X}_R \subset \mathbf{X}$ so that $\mathbf{X}_R$ is a k-$2k$ cover of $\mathbf{X}$ simultaneously for several values of k.

Algorithm 1: Build simultaneous k-$2k$ covers over a log-scaled set of k's

---

Input: Source $(\mathbf{X}_P, \mathbf{Y}_P)$ of size n_P, target $\mathbf{X}_Q$ of size n_Q, and confidence parameter $0 < \delta < 1$

Start with indices $R \leftarrow [n_P]$, and set $k_0 = \lceil \mathcal{V}_{\mathcal{B}} \log(2(n_P + n_Q)) + \log(6/\delta) \rceil$

for $i = 0$ **to** $\lfloor \log_2((n_P \vee n_Q)/2k_0) \rfloor$ **do**

$\quad$ Let $k \leftarrow 2^i k_0$

$\quad$ /* Ensure that R is a k–$2k$ cover of $\mathbf{X}$ $\qquad\qquad\qquad\qquad\qquad$ */

$\quad$ $R \leftarrow R \cup \{i \in (n_P, n_P + n_Q] : X_i$ has less than k NNs from $\mathbf{X}_R$ amongst its $2k$ NNs from $\mathbf{X}\}$

end

return $\mathbf{X}_R$

Algorithm 2 below does not require knowledge of distributional parameters, yet returns a choice of k that yields near-optimal convergence rates in terms of unknown γ, α, β, d. Building on Lepski's approach (for adapting to unknown α, but known d, in the passive case), it works by considering the intersection of confidence sets on the regression function $\eta(x)$ for increasing values of k, and stops when confidence sets no longer intersect; this is an indication of having reached a good choice of k that approximately balances regression bias and variance at a point x. While the basic Lepski's approach usually appears in the literature for adaptation to smoothness α – as applied to *kernel* regression type procedures, we will show that we also automatically get adaptation to γ, β, d in our classification setting under transfer.

C.2. Supporting Lemmas for Theorem 3

The first lemma is an adaptation of Lemma 1 from Berlind and Urner (2015) and gives a bound on the distance to nearest neighbors from a $k - 2k$ cover. Our version of the lemma differs from the initial one in the sense that the below bound is derived for any $i \in [k]$, and not only for $i = k$.

Algorithm 2: Adaptive NN classification estimate

Input: A labeled sample $(\mathbf{X}, \mathbf{Y})$ of size n, a query point x and an integer $k_0 \geq 1$

For the input x and for any k call $\hat{\eta}_k(x) \doteq \frac{1}{k} \sum_{i=1}^{k} X_{(i)}$ the k-NN regression estimate using

  $(\mathbf{X}, \mathbf{Y})$, as defined in Definition 9.

Let $k = k_0$, $\hat{\eta}^- = \hat{\eta}_k(x) - \sqrt{\frac{\mathcal{V}_\mathcal{B}}{k}} \log n$, $\hat{\eta}^+ = \hat{\eta}_k(x) + \sqrt{\frac{\mathcal{V}_\mathcal{B}}{k}} \log n$ and $\hat{\eta} = \hat{\eta}_k(x)$

while $\hat{\eta}^- \leq 1/2$ *and* $\hat{\eta}^+ \geq 1/2$ *and* $k \leq n/2$ **do**

    $k \leftarrow 2k$

    $\hat{\eta}^- \leftarrow (\hat{\eta}_k(x) - \sqrt{\frac{\mathcal{V}_\mathcal{B}}{k}} \log n) \vee \hat{\eta}^-$

    $\hat{\eta}^+ \leftarrow (\hat{\eta}_k(x) + \sqrt{\frac{\mathcal{V}_\mathcal{B}}{k}} \log n) \wedge \hat{\eta}^+$

    if $\hat{\eta}^+ < \hat{\eta}^-$ **then**

        break

    $\hat{\eta} = (\hat{\eta}^+ + \hat{\eta}^-)/2$

end

return Classification estimate $\hat{h}(x) \leftarrow \mathbb{1}\{\hat{\eta} \geq 1/2\}$

Lemma 15 (Relating NN distances) *Let $x \in \mathcal{X}$, $1 \leq k \leq (n_P + n_Q)/2$ and consider $R \subset [n_P + n_Q]$ such that $\mathbf{X}_R$ is a k-$2k$ cover of $\mathbf{X}$. Call $X_{(i)}^R$ the i-th nearest neighbor of x from $\mathbf{X}_R$. We still call $X_{(i)}$ the i-th nearest neighbor of x from $\mathbf{X}$. Then:*

$$\forall i \in [k], \quad \rho(X_{(i)}^R, x) \leq 3\rho(X_{(i+k)}, x).$$

Proof We do the proof by contradiction. Assume that for some $i \in [k]$, we have:

$$\rho(X_{(i)}^R, x) > 3\rho(X_{(i+k)}, x).$$

It means that in the ball $B(x, 3\rho(X_{(i+k)}, x))$ there are strictly less than i observations from $\mathbf{X}_R$ and in the ball $B(x, \rho(X_{(i+k)}, x))$ there are at least $i + k$ observations from $\mathbf{X}$. Therefore, there must be a $x' \in \mathbf{X} \backslash \mathbf{X}_R$ such that $x' \in B(x, \rho(x, X_{(i+k)}))$. As:

$$B(x, \rho(X_{(i+k)}, x)) \subset B(x', 2\rho(X_{(i+k)}, x)) \subset B(x, 3\rho(X_{(i+k)}, x)).$$

We have therefore that $B(x', 2\rho(X_{(i+k)}, x))$ contains strictly less than i elements from $\mathbf{X}_R$ but at least $k + i$ elements from $\mathbf{X}$, meaning that it contains at least $k + 1$ elements from $\mathbf{X} \backslash \mathbf{X}_R$ while having less than k elements from $\mathbf{X}_R$.

Therefore, among the $2k$ nearest neighbors of x' from $\mathbf{X}$, there are strictly less than k elements from $\mathbf{X}_R$, this is in contradiction with the definition of a $k - 2k$ cover (see Definition 10). ∎

We now present, without proof, a second lemma due to Kpotufe (2011) which bounds in high probability the error of regression of a k-NN estimator. We adapt it to our situation of classification under the covariate-shift setting.

Lemma 16 (Lemma 3 from Kpotufe (2011)) *Assume that the VC-dimension $\mathcal{V}_\mathcal{B}$ of $\mathcal{B}$, the class of all the balls in $(\mathcal{X}, \rho)$, is finite. For any independently distributed sample $\{(X_i, Y_i)\}_{i=1}^n$ (not necessarily identically distributed), with at least the same conditional distribution $\eta(x) = \mathbb{P}(Y_i = 1|X_i = x)$, $\forall i$. Define the k-NN regression estimate $\hat{\eta}(x) \doteq \frac{1}{k}\sum_{i=1}^k Y_{(i)}$ where $X_{(i)}$ is the i-th nearest neighbor of x from $\{(X_i, Y_i)\}_{i=1}^n$ and $Y_{(i)}$ its label. Then for $\delta > 0$, we have with probability at least $1 - \delta$:*

$$\forall x \in \mathcal{X}, \forall k \in [n], \quad |\hat{\eta}(x) - \eta(x)| \leq \sqrt{\frac{\mathcal{V}_\mathcal{B} ln(2n/\delta) + 8}{k}} + \frac{C_\alpha}{k}\sum_{i=1}^k \rho^\alpha(X_{(i)}, x).$$

C.3. Proof of Theorem 3

In this proof we use the following notations. For the variables of Algorithm 2, we respectively call $\hat{\eta}_k^-$, $\hat{\eta}_k^+$, $\hat{\eta}_k$ the respective values of the variables $\hat{\eta}^-, \hat{\eta}^+, \hat{\eta}$ at iteration k. Remind that we call Algorithm 2 with, as input, a labeled sample $(\mathbf{X}, \mathbf{Y})_R$ such that $\mathbf{X}_R$ is a k-$2k$ cover of $\mathbf{X}$ for all $k \in \mathcal{K} \doteq \{2^i k_0 : i \in \{0, \ldots, \lfloor\log_2((n_P \vee n_Q)/2k_0)\rfloor\}\}$ for $k_0 \doteq \lceil\mathcal{V}_\mathcal{B}\log(2(n_P+n_Q))+\log(6/\delta)\rceil$. Let $n_R = |R|$, we have therefore $n_P + n_Q \geq n_R \geq (n_P \vee n_Q)/4 \geq (n_P + n_Q)/8$, as for $i = \lfloor\log_2((n_P \vee n_Q)/2k_0)\rfloor$ we have $2^i k_0 \geq (n_P \vee n_Q)/4$. Define $X_{(i)}^R$ the i-th nearest neighbor of x from $\mathbf{X}_R$ and $Y_{(i)}^R$ its label. Finally, for all k and x, we call $\hat{\eta}_{k,R}(x) \doteq \frac{1}{k}\sum_{i=1}^k Y_{(i)}^R$ the k-NN regression estimate built on the labeled data $(\mathbf{X}, \mathbf{Y})_R$ (see Definition 9).

For $\delta' \in (0, 1)$, consider the event:

$$A_{\delta'} = \left\{\forall x \in \mathcal{X}, \forall k \in \mathcal{K}, \quad |\hat{\eta}_{k,R}(x) - \eta(x)| \leq \sqrt{\frac{\mathcal{V}_\mathcal{B}\log(2n_R/\delta') + 8}{k}} + \frac{C_\alpha}{k}\sum_{i=1}^k \rho^\alpha(X_{(i)}^R, x)\right\}.$$

Note that the variables R and n_R are random as they are characteristics of the cover built by Algorithm 1. However, the construction of this cover, before querying labels, depends only on the feature samples $\mathbf{X}$. Thus, by conditioning on $\mathbf{X}$ and using Lemma 16 we get that $\mathbb{P}(A_{\delta'}|\mathbf{X}) \geq 1 - \delta'$. By tower property, we have that event $A_{\delta'}$ has (unconditional) probability at least $1 - \delta'$.

Now, let $\delta' = (n_P + n_Q)^{-(\beta+1)/d_0}$. We can see that there exists a constant $N_1 = N_1(\mathcal{T})$ such that $\forall n_P, n_Q$ s.t. $n_P + n_Q \geq N_1$:

$$\sqrt{\mathcal{V}_\mathcal{B}\log(2n_R/\delta') + 8} \leq \frac{1}{2}\sqrt{\mathcal{V}_\mathcal{B}}\log n_R.$$

Moreover, as an optimal choice of k (see for instance Theorem 2), let:

$$k(n_P, n_Q) \doteq \left\lceil\frac{k_0}{5}\log(n_P + n_Q)(n_P^{d_0/(d_0+\gamma/\alpha)} + n_Q)^{2/d_0}\right\rceil.$$

Note that, as $\mathcal{V}_\mathcal{B} \geq 1$, we have therefore that $\exists N_2 = N_2(\mathcal{T}), n_P + n_Q \geq N_2$:

$$k_0 \leq k(n_P, n_Q) \leq \frac{n_P + n_Q}{8},$$

where the right inequality is obtained from our assumption $(n_p + n_Q) \geq Ck_0^3\log^3(n_P + n_Q)$ for an appropriately chosen universal constant C. Indeed, as $d_0 \geq$ we have:

$$(n_p + n_Q) \geq Ck_0^3\log^3(n_P + n_Q) \implies (n_p + n_Q) \geq C^{1/3}k_0\log(n_P + n_Q)(n_P + n_Q)^{2/3}$$
$$\implies (n_P + n_Q) \geq C^{1/3}k_0\log(n_P + n_Q)(n_P + n_Q)^{2/d_0}.$$

The right inequality is deduced from this fact.

Notice that we just need to prove the upper bound of the theorem only for n_P and n_Q big enough, as indeed we can then increase the constant C so that the inequality of the theorem remains true for lower values of n_P and n_Q. So, from now on, let $\delta' = (n_P + n_Q)^{-(\beta+1)/d_0}$ and assume that $n_P + n_Q \geq N_1 \vee N_2$.

First, let's suppose in our analysis that we are under the event $A_{\delta'}$. Hence we have:

$$\forall x \in \mathcal{X}, \forall k \in \mathcal{K}, \quad |\hat{\eta}_{k,R}(x) - \eta(x)| \leq \max\left(\log(n_R)\sqrt{\frac{\mathcal{V}_\mathcal{B}}{k}}, \frac{2C_\alpha}{k}\sum_{i=1}^{k}\rho^\alpha(X_{(i)}^R, x)\right).$$

Notice that the function $k \mapsto k^{-1/2}$ is decreasing while $k \mapsto k^{-1}\sum_{i=1}^{k}\rho^\alpha(X_{(i)}^R, x)$ is non-decreasing. Therefore, it makes sense to define:

$$k^* \doteq \max\left\{k \in \mathcal{K} \cap [n_R] : \log(n_R)\sqrt{\frac{\mathcal{V}_\mathcal{B}}{k}} \geq \frac{2C_\alpha}{k}\sum_{i=1}^{k}\rho^\alpha(X_{(i)}^R, x)\right\}.$$

Note that for $n_P + n_Q \geq N_3$ for some $N_3 = N_3(\mathcal{T})$, we have $k^* \neq -\infty$ as $\rho(X_{(i)}^R, x) \leq \Delta_\mathcal{X}$ for any i and x. We can assume again $n_P + n_Q \geq N_3$, so that $k^* \neq -\infty$. We have therefore:

$$\forall k \leq k^*, \quad \eta(x) \in [\hat{\eta}_k^-, \hat{\eta}_k^+] = \bigcap_{k'=1}^{k}\left[\hat{\eta}_{k',R}(x) - \log(n_R)\sqrt{\frac{\mathcal{V}_\mathcal{B}}{k}}, \hat{\eta}_{k',R}(x) + \log(n_R)\sqrt{\frac{\mathcal{V}_\mathcal{B}}{k}}\right].$$

Now, call k_{stop} the iteration k at which the algorithm stops, that is:

$$k_{stop} \doteq \min\left\{k \in \mathcal{K} : \quad k > n_R/2 \text{ or } \hat{\eta}_{2k}^- > \hat{\eta}_{2k}^+ \text{ or } \hat{\eta}_k^+ < \frac{1}{2} \text{ or } \hat{\eta}_k^- > \frac{1}{2}\right\}.$$

There are two cases:

<u>1st Case:</u> $k_{stop} < k^*$

Obviously this implies that $k_{stop} \leq n_R/2$ and also, as $\forall k \leq k^*$, $\eta(x) \in [\hat{\eta}_k^-, \hat{\eta}_k^+]$ we cannot have that $\hat{\eta}_{2k_{stop}}^- > \hat{\eta}_{2k_{stop}}^+$. Therefore, if $k_{stop} < k^*$, it only means that $\hat{\eta}_{k_{stop}}^+ < 1/2$ or $\hat{\eta}_{k_{stop}}^- > 1/2$. In this case as $\eta(x) \in [\hat{\eta}_{k_{stop}}^-, \hat{\eta}_{k_{stop}}^+]$, this implies:

$$h^*(x) \doteq \{\eta(x) \geq 1/2\} = \{\hat{\eta}_{k_{stop}}(x) \geq 1/2\} = \hat{h}_R(x).$$

Therefore the excess error at x is equal to 0. That is:

$$2\left|\eta(x) - \frac{1}{2}\right|\mathbb{1}_{h^*(x)\neq\hat{h}_R(x)} = 0.$$

<u>2nd Case:</u> $k_{stop} \geq k^*$

Recall again that $k \mapsto k^{-1/2}$ is decreasing and $k \mapsto k^{-1}\sum_{i=1}^{k}\rho^\alpha(X_{(i)}^R)$ non-decreasing and hence the following minimum:

$$\mu \doteq \min_{k\in\mathcal{K}}\left(\max\left(\log(n_R)\sqrt{\frac{\mathcal{V}_\mathcal{B}}{k}}, \frac{2C_\alpha}{k}\sum_{i=1}^{k}\rho^\alpha(X_{(i)}^R, x)\right)\right).$$

38

is reached at either k^* or $2k^*$. This means:

$$\log(n_R)\sqrt{\frac{\mathcal{V_B}}{k^*}} \leq \sqrt{2}\mu. \tag{27}$$

Indeed this inequality is obvious when the minimum μ is reached at k^*. And if it is reached at $2k^*$, it means that:

$$\log(n_R)\sqrt{\frac{\mathcal{V_B}}{2k^*}} \leq \frac{2C_\alpha}{2k^*}\sum_{i=1}^{2k^*}\rho^\alpha(X_{(i)}^R, x) = \mu.$$

and then we deduce inequality (27) again.

Finally note that $[\hat{\eta}_{k_{stop}}^-, \hat{\eta}_{k_{stop}}^+] \subset [\hat{\eta}_{k^*}^-, \hat{\eta}_{k^*}^+]$ and $\eta(x) \in [\hat{\eta}_{k^*}^-, \hat{\eta}_{k^*}^+]$, hence we have:

$$|\hat{\eta}_{k_{stop}} - \eta(x)| \leq 2\log(n_R)\sqrt{\frac{\mathcal{V_B}}{k}} \leq 2\sqrt{2}\mu.$$

Note also that there is a $k \in \mathcal{K}$ such that $k \leq k(n_P, n_Q) \leq 2k$. Hence we have:

$$\mu \leq \log(n_R)\sqrt{\frac{\mathcal{V_B}}{k}} + \frac{2C_\alpha}{k}\sum_{i=1}^{k}\rho^\alpha(X_{(i)}^R, x) \leq \log(n_R)\sqrt{\frac{2\mathcal{V_B}}{k(n_P, n_Q)}} + \frac{2C_\alpha}{k(n_P, n_Q)}\sum_{i=1}^{k(n_P, n_Q)}\rho^\alpha(X_{(i)}^R, x).$$

So we can bound the excess error at x as follows:

$$2\left|\eta(x) - \frac{1}{2}\right|\mathbb{1}\{h^*(x) \neq \hat{h}_R(x)\}$$

$$\leq 2\left|\eta(x) - \frac{1}{2}\right|\mathbb{1}\left\{\left|\eta(x) - \frac{1}{2}\right| \neq 2\sqrt{2}\mu\right\}$$

$$\leq 2\left|\eta(x) - \frac{1}{2}\right|\mathbb{1}\left\{\left|\eta(x) - \frac{1}{2}\right| \neq 2\sqrt{2}\log(n_R)\sqrt{\mathcal{V_B}k(n_P, n_Q)} + \frac{\sqrt{2}4C_\alpha}{k(n_P, n_Q)}\sum_{i=1}^{k(n_P, n_Q)}\rho^\alpha(X_{(i)}^R, x)\right\}.$$

We recognize a bias-variance bound that is similar to the one we obtained in our proofs of Theorem 2 and Proposition 1. Except the log terms and some additional constant factors, the only difference is that the bias term depends on a cover $\mathbf{X}_R$ instead of the full sample $\mathbf{X}$. We use Lemma 15 to show that it is almost the same thing. Indeed we have:

$$\frac{1}{k(n_P, n_Q)}\sum_{i=1}^{k(n_P, n_Q)}\rho^\alpha(X_{(i)}^R, x) \leq \frac{1}{2k}\sum_{i=1}^{2k}\rho^\alpha(X_{(i)}^R, x) \leq \frac{1}{2k}\sum_{i=1}^{2k}\rho^\alpha(X_{(i+2k)}, x)$$

$$\leq \frac{1}{2k}\sum_{i=1}^{4k}\rho^\alpha(X_{(i)}, x) \leq \frac{1}{2k(n_P, n_Q)}\sum_{i=1}^{4k(n_P, n_Q)}\rho^\alpha(X_{(i)}, x).$$

The rest of the proof is therefore done following similar lines as for the upper bound proof (see Theorem 2 and Proposition 1 and the proofs in Appendix B), as the optimal $k(n_P, n_Q)$ we chose is indeed the one of Theorem 2. Finally, recall that we did all these lines under the event $A_{\delta'}$. As its complement has probability at most $(n_P + n_Q)^{-(\beta+1)/d_0}$ that is of the order of the convergence rate we seek, this ends the proof of the theorem.

C.4. Proof of Theorem 4

For the proof of Theorem 4 we use Lemma 1 from Kpotufe (2011). It is the direct consequence of some known result in VC-theory (see Vapnik and Chervonenkis (1971)). We restate it below without proof.

Lemma 17 (Lemma 1 from Kpotufe (2011)) *Let $\mathcal{B}$ denote the class of all the balls in $(\mathcal{X}, \rho)$, and let D be a distribution over $\mathcal{X}$. Let $\hat{D}$ be the empirical distribution of D from n i.i.d. realizations of D. For $\delta \in (0,1)$, define $\alpha_n = (\mathcal{V}_\mathcal{B} \log(2n) + \log(6/\delta))/n$. With probability at least $1 - \delta$ over the n i.i.d. samples drawn from D, we have simultaneously $\forall B \in \mathcal{B}, \forall a \geq \alpha_n$:*

$$\hat{D}(B) \geq 3a \implies D(B) \geq a, \quad D(B) \geq 3a \implies \hat{D}(B) \geq a.$$

We now turn to the proof of Theorem 4. The proof is based on a similar intuition as used by Berlind and Urner (2015) in their Theorem 2: namely that there is no label request at a target sample $X_i \sim Q_X$ if the distances to its nearest neighbor in $\mathbf{X}_P$ is of similar order as the distance to its nearest neighbor in $\mathbf{X}_P$. However, their theorem is only shown for a fixed k, while we need this result to hold simultaneously for several values of k for our iterative construction of the cover in Algorithm 1. We therefore present a new analysis below that simultaneously considers multiple values of k, and also manages to remove some extraneous log-terms present in their earlier result.

Proof [Theorem 4] Let $k_0 \doteq \lceil \mathcal{V}_\mathcal{B} \log(2(n_P + n_Q)) + \log(6/\delta) \rceil$ as defined in Algorithm 1 and $\mathcal{K} \doteq \{2^i k_0 : i \in \{0, \ldots, \lfloor \log_2((n_P \vee n_Q)/2k_0) \rfloor\}\}$. Call respectively $\hat{P}_X$ and $\hat{Q}_X$ the empirical distributions of P_X and Q_X from their samples $\mathbf{X}_P$ and $\mathbf{X}_Q$. Define also $X_{(i)}^P$ and $X_{(i)}^Q$ respectively the i-th NN of x from $\mathbf{X}_P$ and $\mathbf{X}_Q$. Obviously we have $\rho(x, X_{(2k)}) \geq \min(\rho(x, X_{(k)}^P), \rho(x, X_{(k)}^Q))$. Hence in order to prove that a point $x \in \mathcal{X}_Q$ won't have its label queried, we just to prove that $\rho(x, X_{(k)}^P) \leq \rho(x, X_{(k)}^Q)$, so that among its $2k$-NN from $\mathbf{X}$ it has at least k NN from $\mathbf{X}_P$.

From Lemma 17, and by continuity of any probability measure, we have with probability at least $1 - \delta$,

$$\forall x \in \mathcal{X}_Q, \forall k \in \mathcal{K}, \quad k \leq n_Q \hat{Q}_X(B(x, \rho(x, X_{(k)}^Q))) \leq 3n_Q Q_X(B(x, \rho(x, X_{(k)}^Q))).$$

As by assumption we have $Q_X(B(x, \rho(x, X_{(k)}^Q))) \leq C_d'(\rho(x, X_{(k)}^Q)/\Delta_\mathcal{X})^d$, this implies with probability at least $1 - \delta$:

$$1 \leq k \leq 3n_Q C_d'(\rho(x, X_{(k)}^Q)/\Delta_\mathcal{X})^d \implies \rho(x, X_{(k)}^Q) \geq \Delta_\mathcal{X}(3n_Q C_d')^{-1/d}.$$

Assume that $3^{2+\gamma/d} C_d'^{\gamma/d} n_Q^{(d+\gamma)/d} n_P^{-1} C_\gamma \leq 1$. By using a second time Lemma 17 (on P_X this time) and Definition 3 introducing the transfer exponent, we get with probability at least $1 - 2\delta$:

$$\forall x \in \mathcal{X}_Q, \forall k \in \mathcal{K}, \quad k \leq 3n_Q Q_X(B(x, \rho(x, X_{(k)}^Q)))$$

$$\leq 3\frac{n_Q}{n_P} C_\gamma n_P P_X(B(x, \rho(x, X_{(k)}^Q))) \left(\frac{\rho(x, X_{(k)}^Q)}{\Delta_\mathcal{X}}\right)^{-\gamma}$$

$$\leq (3^{1+\gamma/d} C_d'^{\gamma/d} n_Q^{(\gamma+d)/d} n_P^{-1} C_\gamma) n_P P_X(B(x, \rho(x, X_{(k)}^Q)))$$

$$\leq (3^{2+\gamma/d} C_d'^{\gamma/d} n_Q^{(\gamma+d)/d} n_P^{-1} C_\gamma) n_P \hat{P}_X(B(x, \rho(x, X_{(k)}^Q)))$$

$$\leq n_P \hat{P}_X(B(x, \rho(x, X_{(k)}^Q))).$$

The above lines prove that $n_P \hat{P}_X(B(x, \rho(x, X_{(k)}^Q))) \geq k$ that is the ball $B(x, \rho(x, X_{(k)}^Q))$ contains k elements of $\mathbf{X}_P$. Hence, we have $\rho(x, X_{(k)}^P) \leq \rho(x, X_{(k)}^Q)$. Thus when

$$n_P^{d/(d+\gamma)} \geq (3^{2+\gamma/d} C_d'^{\gamma/d} C_\gamma)^{d/(d+\gamma)} n_Q,$$

we have with probability at least $1-2\delta$ that Algorithm 1 won't make any query as $\forall x \in \mathcal{X}_Q, \forall k \in \mathcal{K}$, x has at least k elements of $\mathbf{X}_P$ in its $2k$-NN from $\mathbf{X}$. ∎

Appendix D. Extensions

We give in this section a couple of extensions of our results to more general settings. The below proposition gives what we should expect as rates of convergence in the situation where the support of P doesn't include the one of Q, but are in some sense close to each other, allowing some amount of transfer.

Proposition 6 (Generalized transfer exponent) *Let $\epsilon \in (0, 3/4]$. Assume that the region $\mathcal{X}_Q^\gamma$ from Definition 3 is such that $Q_X(\mathcal{X}_Q^\gamma) \geq 1 - \epsilon$, instead of $Q_X(\mathcal{X}_Q^\gamma) = 1$. Then the optimal minimax rates are reached by a mixture of k-NN classifiers $\hat{h}(x) \doteq \mathbb{1}\{x \in \mathcal{X}_Q^\gamma\} \hat{h}_{k_1}(x) + \mathbb{1}\{x \notin \mathcal{X}_Q^\gamma\} \hat{h}_{k_2}(x)$, where $k_1 = \Theta(n_P^{d_0/(d_0+\gamma/\alpha)} + n_Q)^{2/d_0}$ and $k_2 = \Theta(1 + n_Q)^{2/d_0}$, where d_0 is defined below. These rates are as follows. Let $\mathcal{T}$ denote either $\mathcal{T}_{(DM)}$ or $\mathcal{T}_{(BCN)}$ and for $\mathcal{T} = \mathcal{T}_{(BCN)}$ assume further that $\alpha < d$. There exist constants C_1, C_2 depending only on $\mathcal{T}$, such that:*

$$\sup_{(P,Q)\in\mathcal{T}} \mathbb{E}_{(\mathbf{X},\mathbf{Y})}[\mathcal{E}_Q(\hat{h})] \leq C_1 \left(n_P^{d_0/(d_0+\gamma/\alpha)} + n_Q \right)^{-(\beta+1)/d_0} + \epsilon \wedge \left(C_2 (1+n_Q)^{-(\beta+1)/d_0} \right),$$

where $d_0 = 2 + d/\alpha$ when $\mathcal{T} = \mathcal{T}_{(DM)}$, and $d_0 = 2 + \beta + d/\alpha$ when $\mathcal{T} = \mathcal{T}_{(BCN)}$.

For the case $\mathcal{T} = \mathcal{T}_{(BCN)}$ with $\alpha = d$, C_1 is replaced with $C_1 \cdot \log(2(n_P + n_Q))$, and C_2 with $C_2 \cdot \log(2(1 + n_Q))$

Proof [Outline] The proof is simply done by dividing the expected excess error into two parts $\mathbb{E}_{(\mathbf{X},\mathbf{Y})}[\mathcal{E}_Q(\hat{h})] = \mathbb{E}_{(\mathbf{X},\mathbf{Y}),X}[\mathcal{E}_Q(\hat{h}_{k_1})(X)\mathbb{1}\{X \in \mathcal{X}_Q^\gamma\}] + \mathbb{E}_{(\mathbf{X},\mathbf{Y}),X}[\mathcal{E}_Q(\hat{h}_{k_2})(X)\mathbb{1}\{X \notin \mathcal{X}_Q^\gamma\}]$ where $X \sim Q_X$ is independent of the data $(\mathbf{X}, \mathbf{Y})$ and $\mathcal{E}_Q(\hat{h})(x) \doteq |\eta(x) - \frac{1}{2}| \cdot \mathbb{1}\{\hat{h}(x) \neq h^*(x)\}$ is the excess error at point x. The L.H.S. rates are obtained by bounding $\mathbb{E}_{(\mathbf{X},\mathbf{Y}),X}[\mathcal{E}_Q(\hat{h}_{k_1})(X)\mathbb{1}\{X \in \mathcal{X}_Q^\gamma\}]$ following similar lines as in our proof of the upper bounds in Theorem 2 (see Proposition 1 and subsequent lemmas and proofs in Appendix B). Indeed, we can redo all theses proofs by restricting the integral to the set $\mathcal{X}_Q^\gamma$ as we just need in this case that the condition of Definition 3 to be satisfied only on this subset for some γ. Finally, the R.H.S. rates are simply obtained because $\mathbb{E}_{(\mathbf{X},\mathbf{Y}),X}[\mathcal{E}_Q(\hat{h}_{k_2})(X)\mathbb{1}\{X \notin \mathcal{X}_Q^\gamma\}]$ is simultaneously bounded by ϵ and by the rate of convergence in the worst case scenario of $\gamma = \infty$. ∎

The last proposition treats the case where the covariate-shift assumption is not verified, that is, there are two different regression functions η_P and η_Q though close to each other.

Proposition 7 (Minimax rates without covariate-shift) *Assume that P and Q have respective regression functions η_P and η_Q such that $\|\eta_P - \eta_Q\|_\infty \le \epsilon$, for $\epsilon \in [0,1]$. Let $\mathcal{T}$ denote either $\mathcal{T}_{(DM)}$ or $\mathcal{T}_{(BCN)}$, where here we added the previous assumption on η_P and η_Q to the definitions of these classes of distribution tuples. For $\mathcal{T} = \mathcal{T}_{(BCN)}$ assume further that $\alpha < d$. There exists a constant $C = C(\mathcal{T})$ such that, for a k-NN classifier $\hat{h}_k$ we have*

$$\sup_{(P,Q)\in\mathcal{T}} \mathbb{E}_{(\mathbf{X},\mathbf{Y})}[\mathcal{E}_Q(\hat{h}_k)] \le C \left(n_P^{d_0/(d_0+\gamma/\alpha)} + n_Q \right)^{-(\beta+1)/d_0} + 2C_\beta(2\epsilon)^{\beta+1},$$

for a choice of $k = \Theta \left(n_P^{d_0/(d_0+\gamma/\alpha)} + n_Q \right)^{2/d_0}$, where $d_0 = 2 + d/\alpha$ when $\mathcal{T} = \mathcal{T}_{(DM)}$, and $d_0 = 2 + \beta + d/\alpha$ when $\mathcal{T} = \mathcal{T}_{(BCN)}$.

For the case $\mathcal{T} = \mathcal{T}_{(BCN)}$ with $\alpha = d$, C above is replaced with $C \cdot \log(2(n_P + n_Q))$.

Proof Note that in this setting we would have:

$$\mathcal{E}_Q(\hat{h}_k) \le 2\mathbb{E}_Q \left| \eta_Q(X) - \frac{1}{2} \right| \cdot \mathbb{1} \left\{ \left| \eta_Q(X) - \frac{1}{2} \right| \le |\hat{\eta}_k(X) - \eta_Q(X)| \right\}.$$

In this case we obtain this bound on the regression error:

$$|\hat{\eta}_k(x) - \eta(x)| \le \frac{1}{k} \left| \sum_{i=1}^{k} Y_{(i)} - \mathbb{E}[Y_{(i)}|X_{(i)}] \right| + \frac{1}{k} \left| \eta_Q(X_{(i)}) - \mathbb{E}[Y_{(i)}|X_{(i)}] \right| + \frac{C_\alpha}{k} \sum_{i=1}^{k} \rho(X_{(i)}, x)^\alpha.$$

Note that by assumption the middle term is bounded as follows:

$$\frac{1}{k} \left| \eta_Q(X_{(i)}) - \mathbb{E}[Y_{(i)}|X_{(i)}] \right| \le \epsilon.$$

Hence, by using $\mathbb{1}\{x \le a + b\} \le \mathbb{1}\{x \le 2a\} + \mathbb{1}\{x \le 2b\}$, and using low noise assumption, we get:

$$\mathcal{E}_Q(\hat{h}_k) \le 2\mathbb{E}_Q \left| \eta_Q(X) - \frac{1}{2} \right| \cdot \mathbb{1} \left\{ \left| \eta_Q(X) - \frac{1}{2} \right| \le 2A \right\} + 2C_\beta(2\epsilon)^{\beta+1},$$

where $A = \frac{1}{k} \left| \sum_{i=1}^{k} Y_{(i)} - \mathbb{E}[Y_{(i)}|X_{(i)}] \right| + \frac{C_\alpha}{k} \sum_{i=1}^{k} \rho(X_{(i)}, x)^\alpha$.

To bound A, we just follow the lines of our previous upper-bound analysis. ∎