Journal of Machine Learning Research 23 (2022) 1-45 Submitted 12/21; Revised 9/22; Published 11/22

Minimax optimal approaches to the label shift problem in non-parametric settings

Subha Maity SMAITY@UMICH.EDU
Yuekai Sun YUEKAI@UMICH.EDU
Moulinath Banerjee MOULIB@UMICH.EDU
Department of Statistics
University of Michigan
Ann Arbor, MI

Editor: Amos Storkey

Abstract

We study the minimax rates of the label shift problem in non-parametric classification. In addition to the unsupervised setting in which the learner only has access to unlabeled examples from the target domain, we also consider the setting in which a small number of labeled examples from the target domain is available to the learner. Our study reveals a difference in the difficulty of the label shift problem in the two settings, and we attribute this difference to the availability of data from the target domain to estimate the class conditional distributions in the latter setting. We also show that a class proportion estimation approach is minimax rate-optimal in the unsupervised setting.

Keywords: Binary classification, non-parametric classification, semi-supervised classification, transfer learning

1. Introduction

A key feature of general intelligence is the transfer of knowledge from one task to another similar but non-identical task. However, most machine learning (ML) methods are not designed for such out-of-distribution (OOD) generalization. This limitation has led to embarrassing performances of ML models in several high-profile applications (Angwin et al., 2016; Dastin, 2018). Transfer learning attempts to improve the OOD generalization performance of ML models and has attracted attention in a variety of application areas: computer vision (Tzeng et al., 2017; Gong et al., 2012), speech recognition (Huang et al., 2013) and genre classification (Choi et al., 2017). Transfer learning is also known as domain adaptation, and we refer to Pan and Yang (2009); Weiss et al. (2016) for surveys of transfer learning.

Despite its empirical success, there is limited knowledge of the fundamental limits of transfer learning. In this paper, we study the fundamental limits of transfer learning for label shift problem under a binary classification setup. We posit the learner has a (labeled) training dataset from the source distribution/domain P and either a small labeled dataset or an unlabeled dataset from the target domain Q. The learner knows P is similar, but not identical to Q (the differences between P and Q will be made precise later). The learner's

©2022 Subha Maity, Yuekai Sun, and Moulinath Banerjee.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v23/21-1519.html.

task is to combine this knowledge about the similarities between P and Q with data from P for inference in Q.

At a high-level, there are two lines of theoretical work in transfer learning. The first line of work focuses on obtaining bounds on the worst-case performance of ML models in similar target domains (Ben-David et al., 2010; Mansour et al., 2009). Here similarity between the source and target domain is measured by some notion of distance/divergence between probability distributions. Although general, such bounds are usually pessimistic, especially if the learner has access to some data from the target domain.

The second line of work focuses on problems in which the learner has some data from the target domain. In order to leverage the data from the source domain, the learner must make some assumptions on the similarities between the source and target domains. Such assumptions usually take the form of invariances between the source and target domain. To keep things simple, we assume it is possible to partition each sample Z from the source and target domains into two parts: $Z = (U, V)$. We factorize the source (resp. target) distributions as $P(U, V) = P(U \mid V)P(V)$ (resp. $Q(U, V) = Q(U \mid V)Q(V)$). By picking U and V carefully, it is possible to obtain many common transfer learning settings.

If the conditional factor changes between the source and target domains ($P(U \mid V) \neq Q(U \mid V)$) while the marginal factor remains the same ($P(V) = Q(V)$), then there is *conditional drift* between the source and the target domains. A prominent example of conditional drift is posterior drift (Cai and Wei, 2019; Maity et al., 2021), in which the marginal distribution of the features X remains the same between the source and target domains, but the conditional distribution of the label / response given the features changes. On the other hand, if the marginal factor changes between the source and the target domains ($P(V) \neq Q(V)$) while the conditional factor remains the same ($P(U \mid V) = Q(U \mid V)$), then there is *marginal drift* between the source and target domains. Prominent examples of marginal drift include covariate shift (Kpotufe and Martinet, 2018; Zhang et al., 2015) ($U = Y$, $V = X$) and label shift (Storkey, 2009; Saerens et al., 2002; Lipton et al., 2018; Schölkopf et al., 2012; Zhang et al., 2015) ($U = X$, $V = Y$).

In this paper, we focus on the label shift problem: $P(X \mid Y) = Q(X \mid Y)$ but $P(Y) \neq Q(Y)$. This problem arises in many application areas. For example, consider building a pneumonia detector. While the symptoms of pneumonia may not change from month to month, the prevalence of the disease in the population may increase in a month of pandemic (*e.g.* January). There are two version of the label shift problem: the supervised version in which the learner has labeled data from the target domain and the unsupervised version in which the learner only has unlabeled data from the target domain. There has been a flurry of recent work (*e.g.* Lipton et al. (2018); Azizzadenesheli et al. (2019); Garg et al. (2020)) on methods for the label shift problem, especially the unsupervised version. Our work complements this line of work by studying the fundamental limits of the label shift problem in a non-parametric setting.

The organization of rest of the paper follows: we formulate the supervised and unsupervised label shift problems and state the working assumption in Section 2. We study the fundamental limits of the supervised and unsupervised label shift problems in Sections 3 and 4 respectively. In Sections 5 and 6, we present simulation studies that confirm our theoretical results and prove the lower bound for the unsupervised label shift problem. Finally, we wrap up with a brief discussion of the implications of our results in section 7.

2. Setup

In this section, we forumulate the label shift problem and state the working assumptions.

2.1 Notations and definitions

For a random vector $(X, Y) \in [0,1]^d \times \{0,1\}$ with distribution P, we denote the marginal distribution of X by P_X and the marginal probability of the event $\{Y = 1\}$ by π_P. We denote the support of P with $\mathrm{supp}(P)$. We use $\mathbb{1}$ to denote the indicator function taking the value in $\{0,1\}$. We use the $\wedge\vee$ notation for min and max: $a \wedge b \triangleq \min(a,b)$ and $a \vee b \triangleq \max(a,b)$. Finally, $\lambda(\cdot)$ denotes the Lebesgue measure of a set in a Euclidean space, and $B(x,r)$ denotes the d-dimensional (closed) Euclidean ball of radius $r > 0$ with center $x \in \mathbb{R}^d$. For a generic probability distribution μ the notation $X_1, \ldots, X_n \overset{\mathsf{ind}}{\sim} \mu$ implies that $X_1, \ldots X_n$ are independently distributed and each of them have distribution μ.

2.2 Label shift in nonparametric classification

Let P and Q be two distributions on $[0,1]^d \times \{0,1\}$. We consider P as the distribution of the samples from the source domain and Q as that of the samples from the target domain. In the supervised version of the label shift problem, the learner has labeled data from the source and target domains:

$$\mathcal{D}_{\mathrm{L}} \triangleq \Big\{ (X_1^P, Y_1^P), \ldots (X_{n_P}^P, Y_{n_P}^P) \overset{\mathsf{ind}}{\sim} P; \\ (X_1^Q, Y_1^Q), \ldots (X_{n_Q}^Q, Y_{n_Q}^Q) \overset{\mathsf{ind}}{\sim} Q \Big\} \in (\mathcal{X} \times \mathcal{Y})^{(n_P + n_Q)}.$$

On the other hand, in the unsupervised version of the problem, the learner only has unlabeled data from the target domain:

$$\mathcal{D}_{\mathrm{U}} \triangleq \Big\{ (X_1^P, Y_1^P), \ldots (X_{n_P}^P, Y_{n_P}^P) \overset{\mathsf{ind}}{\sim} P;\ X_1^Q, \ldots X_{n_Q}^Q \overset{\mathsf{ind}}{\sim} Q_X \Big\} \\ \in (\mathcal{X} \times \mathcal{Y})^{n_P} \times \mathcal{X}^{n_Q}.$$

In both versions of the problem, the class conditionals in the source and target domains are identical: $P(\cdot|Y) = Q(\cdot|Y)$. However, the (marginal) distributions of the labels differ: $\pi_P \neq \pi_Q$.

Let $G_0 \triangleq P(\cdot \mid Y = 0)$ and $G_1 \triangleq P(\cdot \mid Y = 1)$ be the class conditionals. Note that in light of the equivalence of class conditionals in the source and target domains, we can (equivalently) define $G_0 \triangleq Q(\cdot \mid Y = 0)$ and $G_1 \triangleq Q(\cdot \mid Y = 1)$. The regression functions in the source and target domains are

$$\eta_P(x) \triangleq \begin{cases} P(Y = 1|X = x) & \text{if } x \in \mathrm{supp}(P_X) \\ \frac{1}{2} & \text{otherwise} \end{cases}$$

$$\eta_Q(x) \triangleq \begin{cases} Q(Y = 1|X = x) & \text{if } x \in \mathrm{supp}(Q_X) \\ \frac{1}{2} & \text{otherwise} \end{cases}.$$

In both versions of the label shift problem, the goal of the learner is to correctly classify samples from the target domain: learner wishes to learn a classifier $\hat{f} : [0,1]^d \to \{0,1\}$

from the available data ($\mathcal{D}_{\mathrm{L}}$ in the supervised version and $\mathcal{D}_{\mathrm{U}}$ in the unsupervised version) that minimizes the classification error rate in the target domain $Q(Y \neq \hat{f}(X))$. The Bayes classifier in the target domain is

$$f_Q^*(x) = \begin{cases} 0 & \text{if } \eta_Q(x) \leq \frac{1}{2}, \\ 1 & \text{otherwise;} \end{cases}$$

i.e. $f_Q^* \in \arg\min_{h \in \mathcal{H}} \mathbb{P}_Q(Y \neq h(X))$, where $\mathcal{H}$ is the set of all measurable functions $h : [0,1]^d \to \{0,1\}$. We consider the error rate of the Bayes classifier as a baseline and study the **excess risk** of learned classifier $\hat{f}$:

$$\mathcal{E}_Q(\hat{f}) = Q(Y \neq \hat{f}(X)) - Q(Y \neq f_Q^*(X)).$$

The excess risk is a random quantity depending on the available data through the classifier $\hat{f}$ and is known to have the following representation (Gyorfi, 1978):

$$\mathcal{E}_Q(\hat{f}) = 2\mathbb{E}_Q \left[\left| \eta_Q(X) - \frac{1}{2} \right| \mathbb{1}\{\hat{f}(X) \neq f_Q^*(X)\} \right]. \tag{2.1}$$

To keep things simple, we assume that G_0 and G_1 are absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$. We also assume that the marginal distribution of the features in the target domain $Q_X = \pi_Q G_1 + (1 - \pi_Q) G_0$ satisfies the *strong density condition*. This is a common assumption in non-parametric classification (see (Audibert and Tsybakov, 2007, Definition 2.2)).

Definition 1 (strong density condition) *A distribution P defined on $\mathbb{R}^d$ satisfies the strong density condition with parameters $\mu_-, \mu_+, c_\mu, r_\mu > 0$ if*

1. *P is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$;*
2. *its support is regular: $supp(P)$ is compact and $\lambda\left[supp(P) \cap B(x,r)\right] \geq c_\mu \lambda[B(x,r)]$ for all $0 < r \leq r_\mu$ and $x \in supp(P)$;*
3. *$\mu_- < \frac{dP}{d\lambda}(x) < \mu_+$ for all $x \in supp(P)$.*

We note that we only impose the strong density condition in the target domain because that is the domain in which we wish to study the performance of the classifier. Let g_0 and g_1 be the densities of G_0 and G_1 respectively (with respect to the Lesbegue measure on $\mathbb{R}^d$). In terms of g_0 and g_1, the regression function in the target domain is

$$\eta_Q(x) = \begin{cases} \frac{\pi_Q g_1(x)}{\pi_Q g_1(x) + (1-\pi_Q) g_0(x)} & \text{if } x \in \mathrm{supp}(Q_X) \\ \frac{1}{2} & \text{otherwise.} \end{cases} \tag{2.2}$$

Inspecting (2.2), we see that the main difficulty in estimating η_Q is estimating the class conditional densities g_0 and g_1. The symmetry in the problem suggests that errors in estimating g_0 and g_1 affect the convergence rate of the excess risk equally. If the classes are imbalanced, then the excess risk depends on the estimation error of the rarer class. To measure class imbalance in our problem setup, we assume $\pi_P \in [\epsilon_P, 1 - \epsilon_P]$ for some $\epsilon_P > 0$.

As we shall see, ϵ_P affects the effective sample size from the source domain in the minimax rate.

To quantify the hardness of estimating the class conditional densities g_0 and g_1, we impose standard smoothness conditions on them. For any $s = (s_1, \ldots, s_d) \in \mathbb{N}^d$ and $x = (x_1, \ldots, x_d) \in \text{supp}(Q_X)$, define $|s| \triangleq s_1 + \cdots + s_d$, $s! \triangleq s_1! \ldots s_d!$ and $x^s = x_1^{s_1} \ldots x_d^{s_d}$. Let D^s denote the differential operator

$$D^s = \frac{\partial^{s_1+\cdots+s_d}}{\partial x_1^{s_1} \ldots \partial x_d^{s_d}}.$$

For any $g : \text{supp}(Q_X) \to \mathbb{R}$ that is $\lfloor \beta \rfloor$-times continuously differentiable at a point $x_0 \in \text{supp}(Q_X)$, we denote by $g_{x_0}^{(\beta)}$ its Taylor expansion of degree $\lfloor \beta \rfloor$ at x_0:

$$g_{x_0}^{(\beta)}(x) = \sum_{|s| \le \lfloor \beta \rfloor} \frac{(x - x_0)^s}{s!} D^s g(x_0).$$

Definition 2 (Hölder class) *A function $g : \Omega \to \mathbb{R}$ is called (β, L, r_0)-Hölder smooth (β-Hölder smooth in short) if it is $\lfloor \beta \rfloor$-times continuously differentiable and satisfies*

$$|g(x) - g_y^{(\beta)}(x)| \le L\|x - y\|_2^\beta \qquad \forall x \in B(y, r_0) \cap \Omega,\ y \in \Omega.$$

We denote the set of all such functions as $\Sigma(\beta, L, r_0)$.

As we saw, the difficulty of the label shift problem depends on the hardness of estimating the class conditionals g_0 and g_1, so we assume they are β-Hölder smooth. Prior studies on the fundamental limits of covariate shift (Kpotufe and Martinet, 2018) and posterior drift (Cai and Wei, 2019) have imposed similar smoothness conditions on the regression function instead of the class conditionals. This is because in those problems the Bayes classifier remains the same in the source and target domains:

$$\{x \in [0,1]^d \mid \eta_P(x) \le \tfrac{1}{2}\} = \{x \in [0,1]^d \mid \eta_Q(x) \le \tfrac{1}{2}\}.$$

As the hardness of the transfer learning problem now depends on the hardness of estimating η_P, it is important in both covariate shift and posterior drift to quantify the hardness of estimating η_P.

We further introduce the margin condition to quantify the difficulty of the classification task in the target domain. It was introduced in Tsybakov et al. (2004) and adapted by Audibert and Tsybakov (2007); Cai and Wei (2019); Kpotufe and Martinet (2018) to study the convergence rate of the excess risk in binary classification problems. Intuitively, the condition restricts the probability mass around the Bayes decision boundary (region of the feature space such that $\eta_Q(x) \approx \frac{1}{2}$). In other words, it implies $\eta_Q(X)$ is far from $\frac{1}{2}$ with appreciably high probability.

Definition 3 (margin condition for Q) *The distribution Q satisfies the margin condition with parameter α, if there exist $c_\alpha, C_\alpha > 0$, such that*

$$\text{for all } 0 < t < c_\alpha, \quad Q_X\left(0 < \left|\eta_Q(X) - \frac{1}{2}\right| \le t\right) \le C_\alpha t^\alpha.$$

We note that the condition becomes more stringent as α grows. We also note that if Q_X satisfies the strong density assumption and $\beta(1 \wedge \alpha) > 1$, then there is no distribution Q such that the regression function η_Q crosses $\frac{1}{2}$ in the interior of the support of Q_X (see (Audibert and Tsybakov, 2007, Proposition 3.4)). This leads to a trivial Bayes decision rule. In the rest of this paper, we rule out such settings by only considering target domains such that $\alpha\beta \leq 1$.

Combining all the preceding restrictions, we consider the class $\mathcal{P}$ of source and target distribution pairs (P, Q) in our study of the label shift problem.

Definition 4 *The distribution class $\mathcal{P}(\mu_-, \mu_+, c_\mu, r_\mu, \epsilon_P, \alpha, C_\alpha, \beta, C_\beta)$ (or $\mathcal{P}$ in short) is the set of all distribution pairs (P, Q) which satisfies the following conditions:*

1. *$P(\cdot|Y) = Q(\cdot|Y)$ (label shift);*
2. *Q_X satisfies the strong density condition with parameters $\mu = (\mu_-, \mu_+)$, $c_\mu > 0$, $r_\mu > 0$ (see Definition 1);*
3. *$\epsilon_P \leq \pi_P \leq 1 - \epsilon_P$ for some $\epsilon_P > 0$;*
4. *g_0 and g_1 are locally β-Hölder smooth (see Definition 2);*
5. *η_Q satisfies the margin condition (see Definition 3);*
6. *$\alpha\beta \leq 1$, where α is the parameter of the margin condition and β is the Hölder smoothness parameter.*

The goal of the learner is to learn a decision rule $\hat{f}$ from all the available data (including data from both source and target domains) that has small excess risk in the target domain. To study the difficulty of the label shift problem in both supervised and unsupervised settings, we study their minimax risks as functions of the sample sizes in the source and target domains n_P, n_Q.

3. Supervised label shift

In the supervised version of the label shift problem, the learner has access to a dataset $\mathcal{D}_{\mathrm{L}}$, which includes n_P labeled samples from the source domain and n_Q *labeled* samples from the target domain. We assume the source and target distribution pair (P, Q) is in $\mathcal{P}$ (see Definition (4)).

First, we present an information-theoretic lower bound on the convergence rate of the excess risk in the supervised label shift problem. This is a lower bound on the performance of all learning algorithms which accept data $\mathcal{D}_{\mathrm{L}}$ and return a classifier $f : [0,1]^d \to \{0,1\}$. Formally, such a learning algorithm is a map $\mathcal{A} : \mathcal{S}_{\mathrm{L}} \to \mathcal{H}$, where $\mathcal{S}_{\mathrm{L}} \triangleq (\mathcal{X} \times \mathcal{Y})^{n_P+n_Q}$ is the space of possible datasets in the supervised label shift problem (the subscript 'L' indicates the data from the target domain is labeled) and $\mathcal{H} \triangleq \{h : [0,1]^d \to \{0,1\}\}$ is the set of all possible classifiers on $[0,1]^d$.

Theorem 5 (lower bound for the supervised label shift problem) *There is $c > 0$ independent of n_P, n_Q and ϵ_P such that*

$$\inf_{\mathcal{A}:\mathcal{S}_L\to\mathcal{H}} \left\{ \sup_{(P,Q)\in\mathcal{P}} \mathbb{E}_{\mathcal{D}_L}\Big[\mathcal{E}_Q(\mathcal{A}(\mathcal{D}_L))\Big] \right\} \geq c\left((n_P\epsilon_P + n_Q)^{-\frac{\beta}{2\beta+d}} + n_Q^{-1/2} \right)^{1+\alpha}.$$

To show that the preceding lower bound is sharp, we consider a simple plug-in classifier whose convergence rate matches the lower bound:

$$\hat{f}(x) \triangleq \begin{cases} 1 & \text{if } \hat{\eta}_Q(x) \geq \frac{1}{2}, \\ 0 & \text{otherwise,} \end{cases} \tag{3.1}$$

where $\hat{\eta}_Q$ is a particular estimator of the regression function. This estimator is constructed by plugging in an estimator of π_Q and kernel-based estimators of g_0 and g_1 in (2.2):

$$\hat{\eta}_Q(x) = \frac{\hat{\pi}_Q \hat{g}_1(x)}{\hat{\pi}_Q \hat{g}_1(x) + (1-\hat{\pi}_Q)\hat{g}_0(x)}. \tag{3.2}$$

Here $\hat{\pi}_Q$ is the fraction of samples from the target domain with label 1, and $\hat{g}_0$ and $\hat{g}_1$ are kernel-based density estimators:

$$\hat{g}_y(x) = \frac{1}{n_y} \sum_{x'\in\mathcal{X}_y} \frac{1}{h_y^d} K\Big(\frac{x-x'}{h_y}\Big), \quad y \in \{0,1\}, \tag{3.3}$$

where $\mathcal{X}_y = \{x : (x,y') \in \mathcal{D}_{\mathrm{L}},\ y' = y\}$, $n_y = |\mathcal{X}_y|$, and $h_y > 0$ is a bandwidth parameter. The kernel K in (3.3) is a β^*-valid kernel (Tsybakov, 2009, §1.2) for some $\beta^* \geq \beta$. Recall K is a β^*-valid kernel iff

$$\int K(x)dx = 1, \quad \int x^l K(x)dx = 0 \text{ for all } l \in [\beta^*]. \tag{3.4}$$

This choice of kernel is motivated by the Hölder smoothness assumption on g_0 and g_1. We note that (3.3) uses samples from both source and target domains to estimate g_0 and g_1; this does not cause bias in $\hat{g}_0$ and $\hat{g}_1$ because in the label shift problem, the class conditionals are identical in the source and target domains.

Theorem 6 (upper bound for the supervised label shift problem) *Let $\hat{f}$ be the plug in classifier defined above with bandwidths $h_y = n_y^{-1/(2\beta+d)}$, $y \in \{0,1\}$, where n_y is the total number of samples in $\mathcal{D}_L$ that has label y. There is $C > 0$ independent of n_P, n_Q and ϵ_P such that*

$$\sup_{(P,Q)\in\mathcal{P}} \mathbb{E}_{\mathcal{D}_L}\Big[\mathcal{E}_Q(\hat{f})\Big] \leq C\left((\epsilon_P n_P + n_Q)^{-\frac{\beta}{2\beta+d}} + n_Q^{-1/2} \right)^{1+\alpha}.$$

Remark 7 *The bandwidths h_0 and h_1 and β^* (in the choice of kernel) in Theorem 5 depend on the smoothness parameter β. In practice, β is usually unknown, so h_0, h_1 and β^* are chosen by cross-validation.*

We defer the proofs of Theorems 5 and 6 to the supplement. Together, the two theorems imply the minimax rate of the excess risk is:

$$\inf_{\mathcal{A}:\mathcal{S}_{\mathrm{L}}\to\mathcal{H}}\left\{\sup_{(P,Q)\in\mathcal{P}}\mathbb{E}_{\mathcal{D}_{\mathrm{L}}}\Big[\mathcal{E}_Q(\mathcal{A}(\mathcal{D}_{\mathrm{L}}))\Big]\right\}\asymp\Big((\epsilon_P n_P+n_Q)^{-\frac{\beta}{2\beta+d}}+n_Q^{-1/2}\Big)^{1+\alpha}. \tag{3.5}$$

The minimax rate shows the benefits of transfer learning, especially when $n_P \gg n_Q$. In the IID setting in which the learner has access to samples from the target domain but not the source domain, the minimax rate simplifies to

$$\inf_{\mathcal{A}:\mathcal{S}_{\mathrm{L}}\to\mathcal{H}}\left\{\sup_{(P,Q)\in\mathcal{P}}\mathbb{E}_{\mathcal{D}\sim Q^{\otimes n_Q}}\left[\mathcal{E}_Q(\mathcal{A}(\mathcal{D}))\right]\right\}\asymp n_Q^{-\frac{\beta(1+\alpha)}{2\beta+d}}$$

(recall $\beta/(2\beta+d)<1/2$). This agrees with known results on the hardness of non-parametric classification in IID settings (Audibert and Tsybakov, 2007). This is also the minimax rate of learners who ignore the data from the source domain, To see the benefits of transferring knowledge from the source domain, let $n_P \gg n_Q$. As long as $\epsilon_P n_P < n_Q^{1+\frac{d}{2\beta}} - n_Q$ ($\frac{d}{2\beta}$ can be large, so this does not conflict with $n_P \gg n_Q$), the minimax rate simplifies to $(\epsilon_P n_P + n_Q)^{-\beta(1+\alpha)/(2\beta+d)}$, which is faster than the minimax rate of learners who ignores the data from the source domain. We wrap up this section with some technical remarks about the minimax rate in (3.5).

Remark 8 *The first term in* (3.5) *depends on the hardness of estimating the class conditional densities g_0 and g_1. This term depends on the* total *sample size $\epsilon_P n_P + n_Q$ from the source and target domains because samples from both domains are equally informative in estimating g_0 and g_1. The astute reader may wonder why ϵ_P affects the total sample size but π_Q does not. This is because the hardest problem instance has π_P close to zero and $\pi_Q = \frac{1}{2}$. Thus class imbalance in the target domain does not affect the minimax rate, while that in the source domain heavily does. An intuitive explanation for $\pi_Q = \frac{1}{2}$ being the hardest case of classification can be understood from the convergence rate of $\hat{\eta}_Q$:*

$$r(n,\pi)=\begin{cases}\sqrt{\dfrac{(1-\pi_Q)\pi_Q}{n_Q}}+\sqrt{\pi_Q}\Big(\pi_P n_P+\pi_Q n_Q\Big)^{-\frac{\beta}{2\beta+d}}\\ +\sqrt{(1-\pi_Q)}\Big((1-\pi_P)n_P+(1-\pi_Q)n_Q\Big)^{-\frac{\beta}{2\beta+d}}\end{cases}.$$

Inspecting $r(n,\pi)$ as a function of π_Q one can see that the function achieves maximum rate when π_Q is close to $\frac{1}{2}$. We refer to the proof of the upper bound in Appendix B.3.1 (especially the discussion around (B.6)*) for more details.*

Remark 9 *The exponent of $n_P\epsilon_P + n_Q$ in* (3.5) *depends on the smoothness of g_0 and g_1; similar exponents arise in the minimax rates of density estimation (Tsybakov, 2009) and density ratio estimation (Kpotufe, 2017). The second term in the minimax rate depends on the hardness of estimating π_Q. Finally, the overall exponent on the outside depends on the noise level, which we measure with the parameters of the margin condition. We wrap up a few additional remarks about the minimax rate in the supervised label shift problem.*

Remark 10 *If the learner only has access to the labels, but the access to the features from the target domain are restricted, then it is possible to adapt the proofs of Theorems 6 and 5 to show that the minimax rate is*

$$\inf_{\mathcal{A}:\mathcal{S}_L\to\mathcal{H}}\left\{\sup_{(P,Q)\in\Pi}\mathbb{E}\left[\mathcal{E}_Q(\hat{f})\right]\right\}\asymp\left(\left(\epsilon_P n_P\right)^{-\frac{\beta}{2\beta+d}}+n_Q^{-1/2}\right)^{1+\alpha}.$$

At a high-level, the absence of features from Q prevents us from using the samples from Q for estimating the class conditional densities (but they are still useful for estimating π_Q).

Remark 11 *Unlike similar results on the fundamental limits of other transfer learning settings (e.g. covariate shift (Kpotufe and Martinet, 2018), posterior drift (Cai and Wei, 2019), etc.), in the label shift problem, there is a sharp jump between the $\pi_P=\pi_Q$ ($P=Q$) regime and the $\pi_P\neq\pi_Q$ ($P\neq Q$) regime. This is most clearly seen by considering the case in which P is known and the only unknown quantity is π_Q, which simplifies the problem to a Bernoulli proportion estimation problem. In general, it is not possible to estimate π_Q at a rate faster than $\frac{1}{\sqrt{n_Q}}$ unless we know $\gamma\triangleq|\pi_P-\pi_Q|\sim o(\frac{1}{\sqrt{n_Q}})$. In this case, (because we know P and hence we also know π_P), we can simply estimate π_Q with $\widehat{\pi}_Q=\pi_P$, and the worst-case risk of this estimator (over the problem class in which $\gamma\sim o(\frac{1}{\sqrt{n_Q}})$) is $o(\frac{1}{\sqrt{n_Q}})$. Consider the other problem class in which $\gamma\sim\Omega(\frac{1}{\sqrt{n_Q}})$. In this case, Le Cam's method shows that the minimax rate is at least $\frac{1}{\sqrt{n_Q}}$ (the two point construction uses two Bernoulli distributions whose means are $\Theta(\frac{1}{\sqrt{n}})$ apart).*

Taking a step back to the full problem (in which P is unknown), the dependence of the minimax rate on $\gamma\triangleq|\pi_P-\pi_Q|$ is (by analogy to the simple case in which P is known)

$$\begin{cases}\frac{1}{\sqrt{n_Q}}+\left(\epsilon_P n_P+n_Q\right)^{-\frac{\beta}{2\beta+d}}, & \gamma\sim\Omega(\frac{1}{\sqrt{n_Q}}),\\ \max\{\gamma,\frac{1}{\sqrt{n_P}}\}+\left(\epsilon_P n_P+n_Q\right)^{-\frac{\beta}{2\beta+d}}, & \gamma\sim o(\frac{1}{\sqrt{n_Q}}).\end{cases}$$

We see that as soon as $\gamma\sim\Omega(\frac{1}{\sqrt{n_Q}})$, the minimax rate no longer depends on γ. However, the problem sub-class in which γ is small consists of source and target distributions that are very similar, so this sub-class is uninteresting from a transfer learning perspective.

In similar studies of minimax rates for covariate shift (Kpotufe and Martinet, 2018) and posterior drift (Cai and Wei, 2019), the minimax rate depends smoothly on a parameter γ that quantifies the allowable difference between the source and target distributions in the problem class, but this does not occur in the minimax rate for the label shift problem. The Bayes decision rule is different *in the source and target domains in the label shift setting, so optimal prediction (in the target domain) entails estimating an* additional *correction term to bridge the gap between the source and target domains. Estimating this correction term basically boils down to estimating π_Q, and the hardness of estimating π_Q does not depend on the difference between the source and target distributions measured in terms of $\gamma\triangleq|\pi_P-\pi_Q|$ (except in a very small problem sub-class in which transfer learning is irrelevant). Thus our minimax rate for the label shift problem does not depend on γ.*

4. Unsupervised label shift

In the unsupervised version of the label shift problem, the learner has access to $\mathcal{D}_{\mathrm{U}}$, which consists of n_P labeled samples from source domain and n_Q unlabeled samples from the target domain. For this version of the label shift problem, we impose an extra separation condition between the class conditional distributions. As we shall see, a preliminary step in solving the unsupervised label shift problem is estimating $\pi_Q = Q(Y = 1)$, and separation between the class conditionals is necessary for its accurate estimation. We consider the L_2-distance D between probability densities:

$$D^2(g_0, g_1) \triangleq \int_{\mathcal{X}} \big(g_1(x) - g_0(x)\big)^2 dx\,. \tag{4.1}$$

For some $0 < C < 1$ we assume g_0 and g_1 satisfy $D(g_0, g_1) \geq C$. In the rest of this section, we work with the following class of source and target distribution pairs (P, Q):

$$\mathcal{P}' \triangleq \{(P, Q) \in \mathcal{P} : D(g_0, g_1) \geq C\}$$

First, we present a lower bound for the convergence rate of the excess risk in the unsupervised label shift problem. The lower bound is valid for any learning algorithm $\mathcal{A} : \mathcal{S}_{\mathrm{U}} \to \mathcal{H}$, where $\mathcal{S}_{\mathrm{U}} \triangleq (\mathcal{X} \times \mathcal{Y})^{n_P} \times \mathcal{X}^{n_Q}$ is the space of possible datasets in the unsupervised label shift problem (the subscript 'U' indicates samples from the target domain are unlabeled) and $\mathcal{H} \triangleq \{h : [0,1]^d \to \{0,1\}\}$ is the set of classifiers on $[0,1]^d$.

Theorem 12 (lower bound for the unsupervised label shift problem) *There is $c > 0$ independent of n_P, n_Q and ϵ_P such that*

$$\inf_{\mathcal{A}:\mathcal{S}_U \to \mathcal{H}} \left\{ \sup_{(P,Q)\in\mathcal{P}'} \mathbb{E}_{\mathcal{D}_U} \left[\mathcal{E}_Q\left(\mathcal{A}\left(\mathcal{D}_U\right)\right)\right] \right\} \geq c\Big((\epsilon_P n_P)^{-\frac{\beta}{2\beta+d}} + n_Q^{-1/2} \Big)^{1+\alpha}.$$

To show that the preceding lower bound is sharp, we design a classifier whose rate of convergence matches the lower bound. As we alluded to earlier, a preliminary step is estimating π_Q. This is known as the class proportion estimation problem, and it is challenging because the samples from the target domain are unlabeled (in the unsupervised label shift problem). There are many ways to solve the class proportion estimation problem (Lipton et al., 2018; Azizzadenesheli et al., 2019; Alexandari et al., 2020; Du Plessis and Sugiyama, 2014; Iyer et al., 2014; Jain et al., 2016). To prove a matching upper bound, we appeal to the method of Iyer et al. (2014), but it is possible to show similar upper bounds with other methods (see Remark 16). Armed with an estimate of π_Q, we consider the same plug-in classifer as (3.2), except that the kernel-based estimators of g_0 and g_1 only depend on the labeled samples from the source domain.

Theorem 13 (upper bound for the unsupervised label shift) *There is $C > 0$ independent of n_P, n_Q and ϵ_P such that*

$$\sup_{(P,Q)\in\mathcal{P}'} \mathbb{E}_{\mathcal{D}_U} \left[\mathcal{E}_Q\left(\hat{f}\right)\right] \leq C\left((\epsilon_P n_P)^{-\frac{\beta}{2\beta+d}} + n_Q^{-1/2} \right)^{1+\alpha}.$$

We defer the proofs of Theorems 12 and 13 to Section 6 and Appendix B. Theorems 12 and 13 together show that the minimax convergence rate of the excess risk in the unsupervised version of the label shift problem is

$$\inf_{\mathcal{A}:\mathcal{S}_{\mathrm{U}}\to\mathcal{H}}\left\{\sup_{(P,Q)\in\mathcal{P}'}\mathbb{E}_{\mathcal{D}_{\mathrm{U}}}\left[\mathcal{E}_Q\left(\mathcal{A}\left(\mathcal{D}_{\mathrm{U}}\right)\right)\right]\right\}\asymp\left(\left(\epsilon_P n_P\right)^{-\frac{\beta}{2\beta+d}}+n_Q^{-1/2}\right)^{1+\alpha}. \tag{4.2}$$

Recall the minimax rate of IID non-parametric classification in the target domain is $n_Q^{-\frac{\beta(1+\alpha)}{2\beta+d}}$ (Audibert and Tsybakov, 2007). Comparing the minimax rates of IID non-parametric classification (in the target domain) and the unsupervised label shift problem, we see that as long as there are enough samples in the target domain (so we are in the regime of the preceding remark), then labeled samples from the source domain are as informative as labeled samples from the target domain. An extremely important practical implication of this observation is if labeled examples are hard to obtain in the target domain, then it is possible to substitute them with labeled examples in a label shifted source domain.

Before moving on, we compare the minimax rates in the supervised (3.5) and unsupervised (4.2) label shift problems. We see that the rates differ in the first term in the parentheses: there is $\epsilon_P n_P$ instead of $\epsilon_P n_P + n_Q$ in the minimax rate of the unsupervised version. Recall this term depends on the hardness of estimating the class conditional densities g_0 and g_1. In the supervised version, the samples from the target domain are labeled, so they can be directly used to estimate g_0 and g_1. In the unsupervised version, the samples from the target domain are unlabeled, so there is no direct way to use them to estimate g_0 and g_1. There may be indirect ways to leverage the samples from the target domain (*e.g.* by imputing their labels), but our results show that such tricks cannot improve the convergence *rate* of the classifier. We wrap up with a few additional remarks about the minimax rate in the unsupervised label shift problem.

Remark 14 *If $n_P \gg n_Q$, then the minimax rate simplifies to*

$$\inf_{\mathcal{A}}\sup_{(P,Q)\in\mathcal{P}'}\mathbb{E}_{\mathcal{D}_U}\left[\mathcal{E}_Q\left(\mathcal{A}\left(\mathcal{D}_U\right)\right)\right]\asymp\begin{cases}\left(\epsilon_P n_P\right)^{-\frac{\beta(1+\alpha)}{2\beta+d}} & \text{if } \epsilon_P n_P \ll n_Q^{1+\frac{d}{2\beta}},\\ n_Q^{-\frac{1+\alpha}{2}} & \text{if } \epsilon_P n_P \gg n_Q^{1+\frac{d}{2\beta}}.\end{cases}$$

Recalling the form of the plug-in classifier, we see that there are two main sources of errors that contribute to the excess risk:

1. *error in the estimation of the class probability π_Q. This leads to the $\mathcal{O}(n_Q^{-\frac{1+\alpha}{2}})$ term in the minimax rate.*

2. *error in the estimation of the class conditional densities g_0 and g_1. This leads to the $\mathcal{O}((\epsilon_P n_P)^{-\frac{\beta(1+\alpha)}{2\beta+d}})$ term in the minimax rate.*

If $\epsilon_P n_P \gg n_Q^{1+\frac{d}{2\beta}}$ then the error in estimation of π_Q dominates the excess risk. In this case, improving the estimates of the class conditional densities (e.g. by increasing n_P) does not improve the overall convergence rate.

Remark 15 *If $\epsilon_P n_P \ll n_Q^{1+\frac{d}{2\beta}}$, then the minimax rate simplifies to*

$$\inf_{\mathcal{A}:\mathcal{S}_U \to \mathcal{H}} \left\{ \sup_{(P,Q)\in\mathcal{P}'} \mathbb{E}_{\mathcal{D}_U} \left[\mathcal{E}_Q \left(\mathcal{A} \left(\mathcal{D}_U \right) \right) \right] \right\} \asymp \left(\epsilon_P n_P \right)^{-\frac{\beta(1+\alpha)}{2\beta+d}},$$

which is the minimax rate of IID non-parametric classification in the source domain. In other words, given enough unlabeled samples from the target distribution, the error in the non-parametric parts of the unsupervised label shift problem dominate. As this is also the essential difficulty in the IID classification problem in the source domain, it is unsurprising that the minimax rates coincide.

Remark 16 *Reviewing the methods for class proportion estimation shows that most methods converge at a $n_Q^{-1/2} + \delta(n_P)$-rate uniformly on $\mathcal{P}'$, where $\delta(n_P)$ is a term that captures the dependence of the rate on n_P. Inspecting the proof of Theorem 13 reveals that as long as $\delta(n_P) \lesssim (\epsilon_P n_P)^{-\beta/(2\beta+d)}$ (which is satisfied by the method of Iyer et al. (2014)), the excess risk of the resulting classifier attains the minimax rate* (4.2)*. Thus, it is possible to prove similar upper bounds with other methods for class proportion estimation as well (Lipton et al., 2018; Azizzadenesheli et al., 2019; Alexandari et al., 2020; Du Plessis and Sugiyama, 2014; Jain et al., 2016).*

5. Simulations

In this section, we present simulations that illustrate the effects of class imbalance in the source domain. The labels in the source domain are distributed as $Y_i \sim \mathsf{Ber}(\pi_P)$, and the labels in the target domain are distributed as $Y_i \sim \mathsf{Ber}(0.75)$. The class conditional distributions (in both source and target domains) are

$$X_i \mid Y_i \sim Y_i * \mathrm{TN}(0,1,-2,2)^{\otimes 3} + (1-Y_i) * \mathrm{TN}\left(2,1,0,4\right)^{\otimes 3},$$

where $\mathrm{TN}(\mu, \sigma^2, a, b)$ is the $N(\mu, \sigma^2)$ distribution truncated to the interval $[a, b]$. We consider three class imbalance settings: $\pi_P = 0.5$ (solid line, solid circle as pointer), $\pi_P \sim 1/\sqrt{n_P}$ (dashed line, star as pointer) and $\pi_P \sim 1/n_P$ (dotted line, plus as pointer). We defer other details of the simulation setup to Appendix A.

Supervised label shift simulations In the supervised setting, we compare the excess risks of the two following methods:

1. a minimax rate optimal plug-in classifier $\mathbb{1}\left\{\hat{\eta}_Q(x) \geq \frac{1}{2}\right\}$, where $\hat{\eta}_Q$ is defined in (3.2). This method is denoted as `Cl-labeled`.

2. a classical classifier designed for IID settings. This is the same classifier as the minimax rate optimal plug-in classifier that only uses data from the *target* domain to estimate the class conditional densities and class probabilities. We considered this classifier in Section 3 when discussing the benefits of transfer learning. This method is denoted as `Cl-classical`.

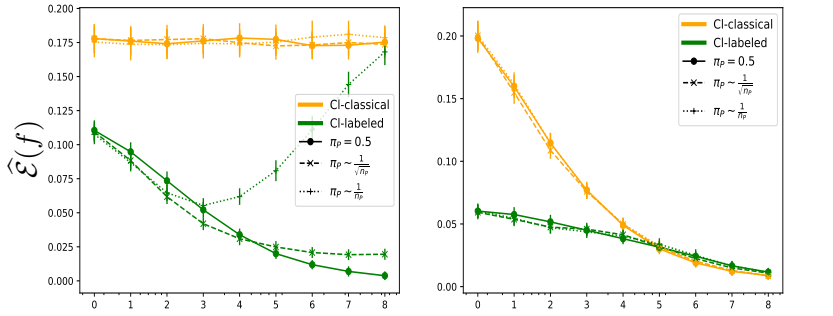

Figure 1: Excess risks in the supervised label shift problem. We see that the minimax optimal approach studied in section 3 is the only approach that learns effectively from both source and target domains.

To estimate the class conditional densities, we use a 2-valid kernel (see (3.4)) with the optimal bandwidth $h_0 = n_0^{-1/7}$ and $h_1 = n_1^{-1/7}$ (see Theorem 6).

We compare the efficacy of the two approaches in two settings. In the first setting (Figure 1, left), $n_P \gg n_Q$ (n_P is growing and n_Q is held fixed), so it is necessary to transfer knowledge from source to target data to reduce the excess risk. As expected, the excess risk of the classical classifier designed for IID settings remains constant as n_P changes as it ignores data from the source domain. On the other hand, the excess risk of the `Cl-labeled` classifier, which transfers knowledge from source to target domain, decreases as n_P increases in the $\pi_P = 0.5$ and $\pi_P \sim \frac{1}{\sqrt{n_P}}$ settings. This shows that `Cl-labeled` leverages data from the source domain to improve classification accuracy in the target domain. We note that effect of class imbalance in the source domain appears in the $\epsilon_P n_P$ term in the minimax rate (3.5). Recall this is the minimum expected per class sample size in the source domain. From (3.5), we expect the convergence rate of the excess risk is slower in the $\pi_P \sim \frac{1}{\sqrt{n_P}}$ setting than in the $\pi_P = 0.5$ setting, and we see that this is indeed the case in Figure 1. We also expect the excess risk to remain bounded away from zero in the $\pi_P \sim \frac{1}{n_P}$ setting because the $\epsilon_P n_P$ term remains bounded away from zero in the minimax rate.

In the second setting (Figure 1, right), we compare the classical classifier and `Cl-labeled` in the $n_Q \gg n_P$ (n_Q is growing and n_P is held fixed) setting. This is an easier setting because it is possible to perform well (have small excess risk) in the target domain without transferring knowledge from the source domain. We expect both approaches to perform well in this setting, and Figure 1 confirms this.

Unsupervised label shift simulations In the unsupervised setting, we compare the excess risks of the two following methods:

1. a minimax rate optimal plug-in classifier where $\hat{\eta}_Q$ is estimated using kernel-based density estimates of g_1 and g_0 and a distribution matching estimator of π_Q Lipton

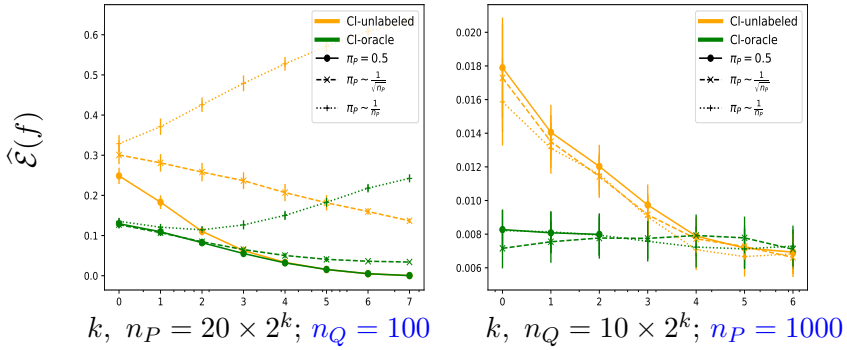

Figure 2: Excess risks in the unsupervised label shift problem. We see that the minimax optimal approach studied in section 4 (eventually) matches the performance of the oracle classifier.

et al. (2018). This method is denoted as `Cl-unlabeled`. We use a different method here to estimate π_Q than in the proof of Theorem 13 to demonstrate the robustness of our results to the choice of estimator (of π_Q).

2. an oracle plugin classifier that uses the exact value of π_Q. We denote this classifier by `Cl-oracle`.

To estimate the class conditional densities, we use a 3-valid kernel (see (3.4)) with the optimal bandwidths $h_0 = \left(n_0'\right)^{-1/7}$, $h_1 = \left(n_1'\right)^{-1/7}$ (see theorem 13). We defer the other details of the simulation setup to Appendix A.

We evaluate the efficacy of both approaches in the two preceding settings, $n_P \gg n_Q$ (see Figure 2, left) and $n_Q \gg n_P$ (see Figure 2, right). In both settings, we fix $\pi_Q = 0.75$ and vary π_P to study the effects of class imbalance. At a high-level, as long as n_Q is large enough, the distributional matching classifier matches the performance of the oracle classifier. This is explained by the fact that if n_Q is large enough, the distributional matching produces a good enough estimator of π_Q, so the error in the estimation of π_Q is no longer the dominant source of error in the excess risk.

We observe that there is always a small gap between the excess risk of the two methods because the estimates of the class probability ratios are not consistent in the simulation settings. Recall that the error incurred by distributional matching is $O(\frac{1}{n_P}) \vee O(\frac{1}{n_Q})$ (see Lipton et al. (2018), Theorem 3). This implies that the estimates of the class probability ratios never converge to their population counterparts in the simulation settings, because one of n_P and n_Q is held fixed by design. Thus, the oracle classifier, which does not suffer from errors in the estimates of the class probability ratios, is always a step ahead of the minimax rate optimal classifier.

6. Proof of Theorem 12

In this section, we prove Theorem 12. The proof of Theorem 5 is similar, and we defer the details to the supplementary materials. At a high-level, the proof has two parts. The first part shows that the minimax rate is at least $(\epsilon_P n_P)^{-\frac{\beta(1+\alpha)}{2\beta+d}}$. This is due to the difficulty arising from the non-parametric part of the label shift problem: estimating the class conditional densities g_0 and g_1. The second part shows that the minimax rate is at least $n_Q^{-\frac{1+\alpha}{2}}$. This stems from the difficulty of the parametric part of the label shift problem: estimating the class probabilities in target domain π_Q.

6.1 Difficulty of the non-parametric part

To study the difficulty of the non-parametric part of the label shift problem, we appeal to the following proposition to obtain lower bounds in the non-parametric regression problems.

Proposition 17 (Theorem 2.5 in 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 there are $h_0, \dots h_M \in \mathcal{H}$, $M \geq 2$, such that:*

1. *$\bar{\rho}(h_i, h_j) \geq 2s > 0$ for all $0 \leq i < j \leq M$,*
2. *$\Pi_{h_i} \ll \Pi_{h_0}$ for all $i \in [M]$, and the average KL-divergence to Π_{h_0} satisfies*

$$\frac{1}{M}\sum_{i=1}^{M} \mathsf{KL}(\Pi_{h_i} \mid \Pi_{h_0}) \leq \kappa \log M \text{ for some } 0 < \kappa < \frac{1}{8}.$$

Let $Z \sim \Pi_h$ and $\hat{f} : Z \to \mathcal{F}$ be any improper learner of $h \in \mathcal{H}$. We have for any $\hat{f}$:

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

The crux of this part is construction of a family of distributions on the source and target domains $\{\Pi_i\}_{i=1}^M$ that satisfies the assumptions of proposition 17. Before delving into the technical details, we describe the intuition behind the construction.

We devote most of our efforts in this part to constructing the class conditional densities because we wish to study (the difficulty of) the non-parametric part of the problem. We partition the sample space $[0, 1]^d$ into small (hyper-)rectangles and divide the rectangles into three groups. The class conditional densities differ on the second and third groups, but they are identical on the first group. The sizes of the two groups are carefully adjusted to satisfy the margin condition in definition 4 and minimize the KL divergence between the Π_i's. Within each rectangle, the class conditional densities are smooth functions with a rise/fall near center of the rectangle and half near the boundaries.

In the proof, we associate the rectangles with the vertices of a hypercube and judiciously pick a subset of rectangles on which the class conditional densities differ, so that the Hamming distance between the associated verticies of the hypercube is maximized. The lower bound on the minimax rate then follows directly from proposition 17.

We assume $\epsilon_P n_P \geq 1$. Let $r = c_r(\epsilon_P n_P)^{-1/(2\beta+d)}, m = \lfloor c_m r^{\alpha\beta-d} \rfloor$, where $\alpha \geq 0$ is the noise condition exponent and $\beta > 0$ is the Hölder smoothness exponent. Also, define $c_r = (1/17),\ c_m = 8 \times 17^{\alpha\beta-d}$ and $c_w \in (0,1)$ is a constant to be picked later. The preceding constants satisfy

$$8 \leq m < \frac{1}{2}\left\lfloor \frac{1}{r} \right\rfloor^d.$$

To see the first inequality, we observe that $r \leq 1/17$ and recall $\alpha\beta \leq d$. Thus

$$m = 8 \times (17r)^{\alpha\beta-d} \geq 8.$$

To see the second inequality, we observe that $r^{-1} \geq 16$, which implies $r^{-1} \leq 17\lfloor r^{-1} \rfloor/16$. Thus

$$m = 8(17r)^{\alpha\beta}\left(\frac{1}{17r}\right)^d \leq 8 \cdot 16^{-d}\lfloor r^{-1} \rfloor^d \leq \frac{1}{2}\lfloor r^{-1} \rfloor^d.$$

We also have $2mw = 2mc_w r^d \leq 2c_m c_w < 1$ for a suitable c_w.

Construction of $\{\Pi_i\}_{i=1}^M$. Let $r_1 = 1/\lfloor 1/(c_w r) \rfloor$ if $\lfloor 1/(c_w r) \rfloor$ is even, otherwise let $r_1 = 1/\left(\lfloor 1/(c_w r) \rfloor + 1\right)$. Let us consider the grid of points

$$\mathcal{Z} = \{(1/2 + i)r_1 : i = 0, 1, \ldots, 1/r_1 - 1\}^d. \tag{6.1}$$

We see that $\mathcal{Z}$ is a grid of equally spaced points of size r_1^{-d}. For a $z \in \mathcal{Z}$ we consider the hyper-cube

$$C(z) = \{x \in [0,1]^d : \|x - z\|_\infty \leq r_1/2\}.$$

Note that, volume of each of these hyper-cubes is r_1^d. Let $\mathcal{Z}_1, \mathcal{Z}_2 \subset \mathcal{Z}$ be subsets of size m. Moreover, we let $\mathcal{Z}_1$ and $\mathcal{Z}_2$ are disjoint. We define a bijection $u : \mathcal{Z}_1 \to \mathcal{Z}_2$ which shall be used to construct the conditional densities. We define $\mathcal{Z}_0 = \mathcal{Z} \backslash (\mathcal{Z}_1 \cup \mathcal{Z}_2)$. Note that, $\mathcal{Z}$ has even number of points and $|\mathcal{Z}_1| = |\mathcal{Z}_2|$. Hence, $\mathcal{Z}_0$ has even number of points. We further divide $\mathcal{Z}_0$ in two sets $\mathcal{Z}_3, \mathcal{Z}_4$ of equal sizes. We shall define a set of distributions parametrized by $\sigma \in \{-1, 1\}^{\mathcal{Z}_1}$.

Conditional densities. For $a > 0$ we define a function v_a supported on on $\mathbb{R}$ which will be used heavily for the construction of conditional densities. Define

$$u_a(x) = \begin{cases} 0 & \text{for } x < 0 \\ \frac{\int_0^x e^{-\frac{1}{at(1-t)}}\,dt}{\int_0^1 e^{-\frac{1}{at(1-t)}}\,dt} & \text{for } 0 \leq x \leq 1 \\ 1 & \text{for } x > 1 \end{cases}$$

and

$$v_a(x) = \begin{cases} \big(1 - u_a(x)\big)^{1/\alpha} & \text{for } \beta < 1, \\ \big(1 - u_a(x)\big) & \text{for } \beta \geq 1. \end{cases} \tag{6.2}$$

According to lemma C.6 we choose a such that $v_a \equiv v$ is (β, C_β) Hölder smooth. Therefore, the following functions are (β, C_β)-Hölder smooth:

$$z \in \mathcal{Z}, \quad \eta_z(x) = \frac{r^\beta}{3} v\left(\frac{2\|x - z\|_\infty}{r_1}\right)$$

and

$$z \in \mathcal{Z}, \quad \xi_z(x) = v\left(\frac{2\|x-z\|_\infty}{br_1} - 2\left(\frac{1}{b}-1\right)\right).$$

For a parameter σ the construction of conditional densities are given below.

$$\begin{cases} g_1^\sigma(x) = \begin{cases} 1+\sigma(z)\sqrt{\epsilon_P}\eta_z(x) & x \in C(z),\ z \in \mathcal{Z}_1, \\ 1-\sigma(z)\sqrt{\epsilon_P}\eta_{f(z)}(x) & x \in C(f(z)),\ z \in \mathcal{Z}_1, \\ 1+\xi_z(x) & x \in C(z),\ z \in \mathcal{Z}_3, \\ 1-\xi_z(x) & x \in C(z),\ z \in \mathcal{Z}_4, \end{cases} \\ g_0^\sigma(x) = \begin{cases} 1-\sigma(z)\sqrt{\epsilon_P}\eta_z(x) & x \in C(z),\ z \in \mathcal{Z}_1, \\ 1+\sigma(z)\sqrt{\epsilon_P}\eta_{f(z)}(x) & x \in C(f(z)),\ z \in \mathcal{Z}_1, \\ 1-\xi_z(x) & x \in C(z),\ z \in \mathcal{Z}_3, \\ 1+\xi_z(x) & x \in C(z),\ z \in \mathcal{Z}_4. \end{cases} \end{cases} \tag{6.3}$$

We also define $\pi_Q^\sigma = 1/2$ and $\pi_P^\sigma = 1/2$. We then define the probabilities

$$P_\sigma(X \in A, Y = y) = \int_A [\pi_P^\sigma g_1^\sigma(x)\mathbb{1}(y=1) + (1-\pi_P^\sigma)g_0^\sigma(x)\mathbb{1}(y=0)]dx$$

and

$$Q_\sigma(X \in A, Y = y) = \int_A [\pi_Q^\sigma g_1^\sigma(x)\mathbb{1}(y=1) + (1-\pi_Q^\sigma)g_0^\sigma(x)\mathbb{1}(y=0)]dx. \tag{6.4}$$

Given the source and target distributions we define the joint distribution of $\mathcal{D}_{\mathrm{U}}$ as

$$\Pi_\sigma = P_\sigma^{\otimes n_P} \otimes Q_{\sigma,X}^{\otimes n_Q} \tag{6.5}$$

Here, $\epsilon_P \le \pi_P^\sigma \le 1-\epsilon_P$. Also, $q_X \equiv 1$ for any $x \in \Omega$. Hence, $\mu_- \le q_X^\sigma(x) \le \mu_+$. Furthermore, $\Omega = [0,1]^d$ is a regular set. Hence, q_X^σ satisfies strong density assumption.

For such a construction, the marginals are

$$p_X^\sigma(x) = q_X^\sigma(x) = \frac{1}{2}g_1^\sigma(x) + \frac{1}{2}g_0^\sigma(x) = 1 \tag{6.6}$$

Furthermore, the regression function η_Q^σ is

$$\eta_Q^\sigma(x) = \frac{\pi_Q^\sigma g_1^\sigma(x)}{q_X^\sigma(x)} = \frac{1}{2}g_1^\sigma(x) \tag{6.7}$$

We refer to lemma 20, where it is shown Q_σ satisfies α-margin condition with constant C_α. Also, the separation assumption **??** is verified in lemma 22.

Let $\mathcal{F}$ be the set of all classifier relevant to this classification problem. For $\sigma \in \{-1,1\}^{\mathcal{Z}_1}$ let f_σ be the Bayes classifier corresponding to the probability distribution Q_σ defined as

$f_\sigma(x) = \mathbb{1}\{\eta_{Q_\sigma}(x) \geq 1/2\}$. For $\sigma, \sigma' \in \{-1,1\}^{\mathcal{Z}_1}$ define $\bar{\rho}(\sigma,\sigma') \coloneqq \mathcal{E}_\sigma(f_{\sigma'})$ and $\rho(\sigma,\sigma') = \text{card}\{z \in \mathcal{Z}_1 : \sigma(z) \neq \sigma'(z)\}$ as the Hamming distance. Then

$$\begin{aligned}\bar{\rho}(\sigma,\sigma') &= 2\mathbb{E}_{Q_{\sigma,X}}\left[\left|\eta_Q^\sigma(X) - \frac{1}{2}\right| \mathbf{1}\left(f_\sigma(X) \neq f_{\sigma'}(X)\right)\right] \\ &\geq c_1 r_1^d r^\beta \rho(\sigma,\sigma') \\ &\geq c_1 c_w^d r^{\beta+d} \rho(\sigma,\sigma').\end{aligned}$$

We recall Varshamov-Gilbert bound, which shall be used to construct the probability class.

Lemma 18 (Varshamov-Gilbert bound) *Let $m \geq 8$. Then there exists a subset $\{\sigma_0,\ldots,\sigma_M\} \subset \{-1,1\}^m$ such that $\sigma_0 = (1,\ldots,1)$,*

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

where, ρ_H is the hamming distance.

Let $\{\sigma_0,\ldots,\sigma_M\} \subset \{-1,1\}^m$ be the choice obtained from the lemma 18. Note that for such a choice $\rho(\sigma_i,\sigma_j) \geq m/8$ whenever $i \neq j$.

Then

$$\begin{aligned}\bar{\rho}(\sigma_i,\sigma_j) &\geq c_1 c_w^d r^{\beta+d}\frac{m}{8} \\ &\geq c_1 c_w^d r^{\beta+d} r^{\alpha\beta-d} \\ &\geq c' r^{\beta(1+\alpha)} \\ &= c'(\epsilon_P n_P)^{-\frac{\beta(1+\alpha)}{2\beta+d}} \\ &\triangleq 2s\end{aligned}$$

Now we bound the Kulback-Leibler divergence between the joint distributions Π_{σ_i}. Using lemma 21 we get

$$\begin{aligned}KL(\Pi_{\sigma_i}||\Pi_{\sigma_j}) &\leq c_w^d K(d,\alpha,\beta)\rho(\sigma_i,\sigma_j) \\ &\leq c_w^d K(d,\alpha,\beta) m \\ &\leq \frac{1}{9}\log_2(M)\end{aligned}$$

for suitable $c_w < 1$.

Finally we appeal to proposition 17 (and Markov's inequality) to obtain the minimax rate

$$\begin{aligned}\sup_{(P,Q)\in\Pi} \mathbb{E}\mathcal{E}_Q(\hat{f}) &\geq \sup_{(P,Q)\in\Pi} s\mathbb{P}_\Pi\left(\mathcal{E}_Q(\hat{f}) \geq s\right) \\ &\geq s \sup_{\sigma\in\{-1,1\}^{\mathcal{Z}_1}} \Pi_\sigma\left(\mathcal{E}_{Q^\sigma}(\hat{f}) \geq s\right) \\ &\geq s\frac{3-2\sqrt{2}}{8} \\ &\geq C(\epsilon_P n_P)^{-\frac{\beta(1+\alpha)}{2\beta+d}}.\end{aligned}$$

6.2 Difficulty of the parametric part

To study the difficulty of the parametric part, it is enough to construct two well-separated hypotheses (versus the family of well-separated hypotheses required in our study of the non-parametric part). We refer to the following theorem to establish lower bound. This particular form of LeCam's bound is taken from Tsybakov (2009), Chapter 2. The result directly follows from Equation (2.5) and statement (iii) in Theorem 2.2.

Theorem 19 (LeCam's bound) *Let $\mathcal{P}$ be a set of distributions. For any pair $P_0, P_1 \in \mathcal{P}$,*

$$\inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P\big[d(\hat{\theta}, \theta(P))\big] \geq \frac{\Delta}{8} e^{-KL(P_0||P_1)}$$

where $\Delta = d\big(\theta(P_0), \theta(P_1)\big)$.

The construction will closely follow the non-parametric part. Let $\sigma \in \{-1, 1\},\ c_w < 1,\ r = n_Q^{-\alpha/2}$,

$$r_1 = \begin{cases} 1/\lfloor 1/(c_w r)\rfloor & \text{if } \lfloor 1/(c_w r)\rfloor \text{ is odd} \\ 1/\left(\lfloor 1/(c_w r)\rfloor + 1\right) & \text{if } \lfloor 1/(c_w r)\rfloor \text{ is even} \end{cases}$$

Thus $\frac{1}{r_1}$ is always an odd number. Define $2D + 1 = 1/r_1$. Let us consider the grid of points

$$\mathcal{Z} = \{z_i = (1/2 + i)r_1 : i = 0, 1, \ldots, 2D\}.$$

We again consider the function $v_a \equiv v$ (as defined in (6.2)) which is (β, C_β) Hölder smooth. Hence the following functions are also (β, C_β) Hölder smooth.

$$z \in \mathcal{Z}, \quad \eta_z(x) = v\left(\frac{2\|x - z\|_\infty}{b r_1} - 2\left(\frac{1}{b} - 1\right)\right)$$

. We also define $C(z) = \big\{x \in [0,1]^d : \big|x^{(1)} - z^{(1)}\big| \leq r_1/2\big\}$

$$\begin{cases} g_1^\sigma(x) = \begin{cases} 1 & x \in C(z_0), \\ 1 + \eta_{z_i}(x) & x \in C(z_i),\ \text{for } i = 1, \ldots, D, \\ 1 - \eta_{z_i}(x) & x \in C(z_i),\ \text{for } i = D+1, \ldots, 2D, \end{cases} \\ g_0^\sigma(x) = \begin{cases} 1 & x \in C(z_0), \\ 1 - \eta_{z_i}(x) & x \in C(z_i),\ \text{for } i = 1, \ldots, D, \\ 1 + \eta_{z_i}(x) & x \in C(z_i),\ \text{for } i = D+1, \ldots, 2D \end{cases} \end{cases} \tag{6.8}$$

and $\pi_P^\sigma = 1/2,\ \pi_Q^\sigma = 1/2 + \sigma c_Q n_Q^{-1/2}$. We construct P_σ, Q_σ and define

$$\Pi_\sigma = P_{X,\sigma}^{\otimes n_P} \otimes Q_\sigma^{\otimes n_Q} \tag{6.9}$$

in the similar way. As before, we have $\epsilon_P \leq \pi_P^\sigma \leq 1 - \epsilon_P$ and $q_X \equiv 1$. Hence, $\mu_- \leq q_X \leq \mu_+$ and supported on $[0,1]^d$. This implies Q_X has strong density. Also, we refer to lemma 20 to show Q_σ satisfies α margin condition with constant C_α.

For $\sigma \in \{-1, 1\}$ let η_σ and f_σ be the regression function and Bayes classifier for Q_σ, respectively. Then

$$\begin{aligned}
\bar{\rho}(\Pi_1, \Pi_{-1}) &= \mathcal{E}_{Q_1}(f_{-1}) \\
&= 2\mathbb{E}_{Q_{1,X}}\Big[\big|\eta_Q^\sigma(X) - 1/2\big|\mathbb{1}\{f_1(X) \neq f_{-1}(X)\}\Big] \\
&\geq 4c_Q n_Q^{-1/2} r_1 \\
&\geq c' n_Q^{-\frac{1+\alpha}{2}} = 2s.
\end{aligned}$$

We refer to lemma C.4 in supplement to get the following bound

$$\begin{aligned}
KL(\Pi_1||\Pi_{-1}) &\leq n_Q KL(U_Q^{(1)}||U_Q^{(-1)}) \\
&\leq n_Q(1/2 + c_Q n_Q^{-1/2}) \log\left(\frac{1/2 + c_Q n_Q^{-1/2}}{1/2 - c_Q n_Q^{-1/2}}\right) \\
&\quad + (1/2 - c_Q n_Q^{-1/2}) \log\left(\frac{1/2 - c_Q n_Q^{-1/2}}{1/2 + c_Q n_Q^{-1/2}}\right) \\
&\leq n_Q \frac{2}{4} 3c_Q^2 n_Q^{-1} \\
&\leq c_Q^2
\end{aligned}$$

where $U_Q^\sigma \sim \text{Bernoulli}(\pi_Q^\sigma)$.

Using theorem 19 we conclude

$$\sup_{(P,Q)\in\Pi} \mathbb{E}\big[\mathcal{E}_Q(\hat{f})\big] \geq c n_Q^{-\frac{1+\alpha}{2}}.$$

Finally, we combine the two bounds to get

$$\begin{aligned}
\sup_{(P,Q)\in\Pi} \mathbb{E}\big[\mathcal{E}_Q(\hat{f})\big] &\geq c n_Q^{-\frac{1+\alpha}{2}} \vee c(\epsilon_P n_P)^{-\frac{\beta(1+\alpha)}{2\beta+d}} \\
&\geq c'\left((\epsilon_P n_P)^{-\frac{\beta}{2\beta+d}} + n_Q^{-1/2}\right)^{1+\alpha}.
\end{aligned}$$

Lemma 20 *For any $\sigma \in \{-1, 1\}^{\mathcal{Z}_1}$ Q_σ as defined in 6.4 and 6.9 satisfies α-margin condition with constant C_α.*

Proof We prove the lemma for both 6.4 and 6.9.

Margin condition for non-parametric part We recall the marginal and regression function as in 6.6 and 6.7, respectively. For such a regression function

$$\big|\eta_Q^\sigma(x) - 1/2\big| = \begin{cases} \sqrt{\epsilon_P}\eta_z(x)/2 & x \in C(z),\ z \in \mathcal{Z}_1, \\ \sqrt{\epsilon_P}\eta_{f(z)}(x)/2 & x \in C(f(z)),\ z \in \mathcal{Z}_1, \\ 0 & x \in C(z),\ z \in \mathcal{Z}_0. \end{cases}$$

Let

$$t_0 = \begin{cases} \frac{\sqrt{\epsilon_P} r^\beta}{6} \left(1 - u_a(1/2)\right)^{1/\alpha} & \beta < 1 \\ \frac{\sqrt{\epsilon_P} r^\beta}{6} \left(1 - u_a(1/2)\right) & \beta \geq 1 \end{cases}$$

For $t \leq t_0, \beta < 1$ and $z \in \mathcal{Z}_1 \cup \mathcal{Z}_2$, we see that

$$\int_{1/2}^1 e^{-\frac{1}{a(s-s^2)}} ds \leq \int_0^1 e^{-\frac{1}{a(s-s^2)}} ds \left(\frac{6t}{\sqrt{\epsilon_P} r^\beta}\right)^\alpha. \tag{6.10}$$

Hence,

$$\begin{aligned}
&Q_X\big\{0 < \sqrt{\epsilon_P}\eta_z(x) \leq 2t\big\} \\
&= Q_X\left\{0 < 1 - u\left(\frac{2\|x-z\|_\infty}{r_1}\right) \leq \left(\frac{6t}{\sqrt{\epsilon_P} r^\beta}\right)^\alpha\right\} \\
&= Q_X\left\{0 < \int_{\frac{2\|x-z\|_\infty}{r_1}}^1 e^{-\frac{1}{a(s-s^2)}} ds \leq \int_0^1 e^{-\frac{1}{a(s-s^2)}} ds \left(\frac{6t}{\sqrt{\epsilon_P} r^\beta}\right)^\alpha\right\} \\
&\leq Q_X\left\{0 < e^{-\frac{4}{a}}\left(1 - \frac{2\|x-z\|_\infty}{r_1}\right) \leq \int_0^1 e^{-\frac{1}{a(s-s^2)}} ds \left(\frac{6t}{\sqrt{\epsilon_P} r^\beta}\right)^\alpha\right\} \\
&= Q_X\left\{1 - e^{4/a}\int_0^1 e^{-\frac{1}{a(s-s^2)}} ds \left(\frac{6t}{\sqrt{\epsilon_P} r^\beta}\right)^\alpha < \frac{2\|x-z\|_\infty}{r_1} \leq 1\right\} \\
&= r_1^d\left[1 - \left(1 - e^{4/a}\int_0^1 e^{-\frac{1}{a(s-s^2)}} ds \left(\frac{6t}{\sqrt{\epsilon_P} r^\beta}\right)^\alpha\right)^d\right] \\
&\leq C\epsilon_P^{-\alpha/2} c_w^d \frac{r^d t^\alpha}{r^{\alpha\beta}}
\end{aligned}$$

where the third inequality is true because $\frac{2\|x-z\|_\infty}{r_1} \geq 1/2$ (obtained from inequality 6.10). Similarly, $t \leq t_0, \beta \geq 1$ we have

$$\begin{aligned}
Q_X\big\{0 < \sqrt{\epsilon_P}\eta_z(x) \leq 2t\big\} &\leq C b^d \epsilon_P^{-1/2} c_w^d \frac{r^d t}{r^\beta} \\
&\leq C b^d \epsilon_P^{-1/2} c_w^d \frac{r^d t}{r^\beta} \big(C' t / r^\beta\big)^{\alpha - 1} \\
&\leq C'' \epsilon_P^{-\alpha/2} c_w^d \frac{r^d t^\alpha}{r^{\alpha\beta}}
\end{aligned}$$

because $\alpha \leq 1$ for $\beta \geq 1$. For $z \in \mathcal{Z}_3 \cup \mathcal{Z}_4$

$$Q_X\big\{0 < \sqrt{\epsilon_P}\eta_z(x) \leq 2t\big\} \leq C c_w^d b^d \frac{r^d t^\alpha}{r^{\alpha\beta}}$$

Hence,

$$\begin{aligned}
Q_X\big\{0 < \left|\eta_Q^\sigma(x) - \frac{1}{2}\right| \leq t\big\} &\leq C\epsilon_P^{-\alpha/2} c_w^d \frac{r^d t^\alpha}{r^{\alpha\beta}} 2m + (r_1^{-d} - 2m) C b^d c_w^d \frac{r^d t^\alpha}{r^{\alpha\beta}} \\
&\leq C_\alpha t^\alpha.
\end{aligned}$$

for small enough c_w.

Let $\frac{\sqrt{\epsilon_P} r^\beta}{6} (1 - u_a(1/2))^{1/\alpha} \le t \le \frac{1}{3}$. Then

$$\begin{aligned} Q_X \left\{ 0 < \left| \eta_Q^\sigma(x) - \frac{1}{2} \right| \le t \right\} &\le 2mr_1^d + (r_1^{-d} - 2m)(br_1)^d \\ &\le 2c_m c_w^d r^{\alpha\beta} + c_b^d r^d \\ &\le \left(2c_m c_w^d + c_b^d\right) r^{\alpha\beta} \\ &\le C_\alpha \left(\frac{\sqrt{\epsilon_P} r^\beta}{6} (1 - u_a(1/2))^{1/\alpha} \right)^\alpha \\ &\le C_\alpha t^\alpha \end{aligned}$$

where the second last inequality is true for small c_w and c_b.

Margin condition for parametric part Let $t \le \frac{1}{2} (1 - u_a(1/2))^{1/\alpha}$. For $i \ge 1$

$$Q_X \{ 0 < \sqrt{\epsilon_P} \eta_{z_i}(x) \le 2t \} \le C b c_w r t^\alpha.$$

Hence, for $t < c_Q n_Q^{-1/2}$

$$\begin{aligned} Q_X \left\{ 0 < \left| \eta_Q^\sigma(x) - \frac{1}{2} \right| \le t \right\} &\le 4DCbc_w r t^\alpha \\ &\le C_\alpha t^\alpha. \end{aligned}$$

for small enough c_w. For $c_Q n_Q^{-1/2} \le t \le \frac{1}{2} (1 - u_a(1/2))^{1/\alpha}$

$$\begin{aligned} Q_X \left\{ 0 < \left| \eta_Q^\sigma(x) - \frac{1}{2} \right| \le t \right\} &\le 4DCbc_w r t^\alpha + c_w^\alpha r^\alpha \\ &\le C_\alpha t^\alpha. \end{aligned}$$

for small enough c_w.

■

Lemma 21 *Let* $\left\{ \Pi_\sigma : \sigma \in \{-1, 1\}^{\mathcal{Z}_1} \right\}$ *be the class of joint distributions defined in 6.5. For* $\sigma, \sigma' \in \{-1, 1\}^{\mathcal{Z}_1}$ *we have*

$$KL(\Pi_\sigma || \Pi_{\sigma'}) \le c_w^d K(d, \alpha, \beta) \rho(\sigma, \sigma').$$

Proof We recall from 6.6 $Q_{X,\sigma} \sim \text{Uniform}\left([0,1]^d\right)$ doesn't depend on σ. Hence,

$$KL\left(Q_{X,\sigma} || Q_{X,\sigma'} \right) = 0.$$

We refer to lemma C.4 in supplement to show

$$\begin{aligned} KL\left(P_\sigma || P_{\sigma'} \right) &= \frac{1}{2} \left(g_0^\sigma || g_0^{\sigma'} \right) + \frac{1}{2} \left(g_1^\sigma || g_1^{\sigma'} \right) \\ &\le \epsilon_P c_w^d K'(d, \alpha, \beta) r^{2\beta + d} \rho(\sigma, \sigma'). \end{aligned}$$

Combining them we get

$$\begin{aligned} KL\Big(\Pi_\sigma||\Pi_{\sigma'}\Big) &= n_P KL\Big(P_\sigma||P_{\sigma'}\Big) + n_Q KL\Big(Q_{X,\sigma}||Q_{X,\sigma'}\Big) \\ &\leq n_P \epsilon_P c_w^d K'(d,\alpha,\beta) r^{2\beta+d} \rho(\sigma,\sigma') \\ &\leq c_w^d K'(d,\alpha,\beta) c_r^{2\beta+d} (\epsilon_P n_P)(\epsilon_P n_P)^{-1} \rho(\sigma,\sigma') \\ &\leq c_w^d K(d,\alpha,\beta)\rho(\sigma,\sigma'). \end{aligned}$$

■

Lemma 22 *For suitable choices of $c_w, b > 0$ we have*

$$\int_{\mathcal{X}} (g_1^\sigma(x) - g_0^\sigma(x))^2 dx \geq C^2.$$

Proof Notice that,

$$\xi_z(x) = v\left(\frac{2\|x-z\|_\infty}{br_1} - \frac{2}{r_1}\left(\frac{1}{b} - 1\right)\right) = 1$$

whenever $\|x-z\|_\infty \leq r_1(1-b)$ and in such regions the density differences are ≥ 1. Hence, the integral of the difference squared is $\geq r_1^d(1-b)^d$ for non-parametric case and $\geq r_1(1-b)$ for parametric case. Noticing that there are $(1/r_1^d - 2m)$ (non-parametric) and $2D$ (parametric) many regions we get the following

$$\int_{\mathcal{X}} \big(g_0(x) - g_1(x)\big)^2 dx \geq \begin{cases} (1 - 2c_m c_w^d)(1-b)^d \text{ for non-parametric part,} \\ \Big(1 - \frac{1}{2D+1}\Big)(1-b) \text{ for parametric part.} \end{cases}$$

Here $\frac{1}{2D+1} \sim c_w n_Q^{-\alpha/2}$. The constants c_w, b can be suitably chosen to satisfy the condition in lemma.

■

7. Summary and discussion

We studied the hardness of the label shift problem in two settings, one in which the learner has access to labeled training examples from the target domain, and another in which the learner only has unlabeled training examples from the target domain. We showed that there is a difference between the hardness of the label shift problem in the two settings. In the former setting (in which the learner has access to labeled training examples from the target domain), the minimax rate is $O\big((\epsilon_P n_P + n_Q)^{-\beta/(2\beta+d)} + n_Q^{-1/2}\big)^{1+\alpha}$, while in the latter setting, the minimax rate is $O\big((\epsilon_P n_P)^{-\beta/(2\beta+d)} + n_Q^{-1/2}\big)^{1+\alpha}$. We attribute this difference in rates is due to the availability of data from the target domain to estimate the the class conditional distributions in the former setting.

Although we studied the hardness of the label shift problem with non-parametric model classes, we expect our results to generalize to more restrictive model classes. Inspecting the minimax rates, we see that they consist of two terms: a term that depends on the hardness of estimating the (ratio of) class conditional distributions, and a term that depends on the hardness of estimating the ratio of class probabilities in the source and target domains. In non-parametric classification, the hardness of estimating the class conditionals determines the hardness of the non-parametric classification problem in the IID setting (Kpotufe, 2017). This observation leads us to interpret the first term in the minimax rate as the hardness of finding the optimal classifier, and we expect this term to change with the model class. Thus, for more restrictive model classes, we expect the first term in the minimax rates to improve (vanish faster) in a way that depends on the (reduced) complexity of the model class.

To wrap up, we mention two possible extensions of our work. First, it is natural to consider the label shift problem in high dimension. To keep the problem tractable, we must impose stronger parametric assumptions on the regression function. Such assumptions may also be phrased as assumptions on the class conditional densities because the regression function is (up to a monotone transform) the ratio of the class conditional densities. In the supervised label shift problem, we expect the minimax rate to depend on the hardness of estimating the regression function under the additional parametric assumptions. In the unsupervised label shift problem, we expect the distributional matching approach to provide good estimates of the class probability ratios (because the class probability ratios are low-dimensional), so we also expect the minimax rate to depend on the hardness of estimating the regression function.

Second, it is natural to consider the possibility of achieving the minimax rate with a classifier that adapts to the smoothness of the regression function and the noise level in the labels. Kpotufe and Martinet (2018) and Cai and Wei (2019) developed adaptive classifiers that attain the minimax rate in the covariate shift and posterior drift problems with Lepski's method. We expect Lepski's method will lead to an adaptive classifier in the label shift problem as well. However, the main goal of this paper is investigating the hardness of the label shift problem, and we defer such methodological questions to future work.

References

A. Alexandari, A. Kundaje, and A. Shrikumar. EM with Bias-Corrected Calibration is Hard-To-Beat at Label Shift Adaptation. *arXiv:1901.06852 [cs, stat]*, Jan. 2020.

J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine Bias. www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, May 2016.

J.-Y. Audibert and A. B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of statistics*, 35(2):608–633, 2007.

K. Azizzadenesheli, A. Liu, F. Yang, and A. Anandkumar. Regularized Learning for Domain Adaptation under Label Shifts. *arXiv:1903.09734 [cs, stat]*, Mar. 2019.

S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, May 2010. ISSN 0885-6125, 1573-0565. doi: 10.1007/s10994-009-5152-4.

T. T. Cai and H. Wei. Transfer Learning for Nonparametric Classification: Minimax Rate and Adaptive Classifier. *arXiv:1906.02903 [cs, math, stat]*, June 2019.

K. Choi, G. Fazekas, M. Sandler, and K. Cho. Transfer learning for music classification and regression tasks. *arXiv preprint arXiv:1703.09179*, 2017.

J. Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*, Oct. 2018.

M. C. Du Plessis and M. Sugiyama. Semi-supervised learning of class balance under class-prior change by distribution matching. *Neural Networks*, 50:110–119, 2014.

S. Garg, Y. Wu, S. Balakrishnan, and Z. C. Lipton. A Unified View of Label Shift Estimation. *arXiv:2003.07554 [cs, stat]*, Mar. 2020.

B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2066–2073. IEEE, 2012.

L. Gyorfi. On the rate of convergence of nearest neighbor rules (corresp.). *IEEE Transactions on Information Theory*, 24(4):509–512, 1978.

J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7304–7308. IEEE, 2013.

A. Iyer, S. Nath, and S. Sarawagi. Maximum mean discrepancy for class ratio estimation: Convergence bounds and kernel selection. In *International Conference on Machine Learning*, pages 530–538. PMLR, 2014.

S. Jain, M. White, and P. Radivojac. Estimating the class prior and posterior from noisy positives and unlabeled data. *Advances in neural information processing systems*, 29: 2693–2701, 2016.

S. Kpotufe. Lipschitz Density-Ratios, Structured Data, and Data-driven Tuning. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, page 14, Fort Lauderdale, Florida, 2017.

S. Kpotufe and G. Martinet. Marginal Singularity, and the Benefits of Labels in Covariate-Shift. *arXiv:1803.01833 [cs, stat]*, Mar. 2018.

Z. C. Lipton, Y.-X. Wang, and A. Smola. Detecting and Correcting for Label Shift with Black Box Predictors. *arXiv:1802.03916 [cs, stat]*, July 2018.

S. Maity, D. Dutta, J. Terhorst, Y. Sun, and M. Banerjee. A linear adjustment based approach to posterior drift in transfer learning. *arXiv preprint arXiv:2111.10841*, 2021.

Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain Adaptation: Learning Bounds and Algorithms. *arXiv:0902.3430 [cs]*, Feb. 2009.

S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

P. Rigollet and R. Vert. Optimal rates for plug-in estimators of density level sets. *Bernoulli*, 15(4):1154–1178, 2009.

M. Saerens, P. Latinne, and C. Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural computation*, 14(1):21–41, 2002.

B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*, 2012.

A. Storkey. When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning*, pages 3–28, 2009.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York ; London, 2009. ISBN 978-0-387-79051-0 978-0-387-79052-7. OCLC: ocn300399286.

A. B. Tsybakov et al. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176, 2017.

K. Weiss, T. M. Khoshgoftaar, and D. Wang. A survey of transfer learning. *Journal of Big data*, 3(1):9, 2016.

K. Zhang, M. Gong, and B. Schölkopf. Multi-source domain adaptation: A causal view. In *AAAI*, volume 1, pages 3150–3157, 2015.

Appendix A. Simulation details

The codes and simple demonstrations are provided in https://github.com/smaityumich/label-shift.

Data generating process We start by describing the data generating process $\mathcal{D}(n, \pi)$. Let μ_X denote the probability distribution of random variable X. For $a < b$ we define the $\text{TN}(\mu, \sigma^2, a, b)$ as the $N(\mu, \sigma^2)$ distribution truncated to the interval $[a, b]$. Given inputs sample size n and class probability π for class 1, $\mathcal{D}(n, \pi)$ returns a pair $(\mathbf{x}, \mathbf{y})$ where,

- $\mathbf{y}$ is a n dimensional random vector with IID $\mathsf{Ber}(1, \pi)$ components.
- $\mathbf{x} = [x_1, \ldots, x_n]^T$ is a $n \times 3$ random matrix with independent rows. The distribution of the i-th row is

$$x_i \mid y_i \sim y_i * \mu^{\otimes 3}_{\text{TN}(0,1,-2,2)} + (1 - y_i) * \mu^{\otimes 3}_{\text{TN}(2,1,0,4)}.$$

  We observe that the features are supported on the hypercube $[-2, 4]^4$

Given the data generating procedure $\mathcal{D}(n, \pi)$ we generate the following synthetic data:

– $(\mathbf{x}_P, \mathbf{y}_P) = \mathcal{D}(n_P, 0.5)$ is the data from source population.

– $(\mathbf{x}_Q, \mathbf{y}_Q) = \mathcal{D}(n_Q, 0.75)$ is the data from target population.

– $(\mathbf{x}_{\text{test}}, \mathbf{y}_{\text{test}}) = \mathcal{D}(n_{\text{test}}, 0.75)$ is the data for evaluating the performance of the classifiers, which shall also be referred as test data. Note the distribution of test data is same as the target distribution.

Other classifiers Next we describe the classifiers that we shall consider for our comparative study:

- `Labeled-Classifier` is a function that takes the data $(\mathbf{x}_P, \mathbf{y}_P)$ from source, $(\mathbf{x}_Q, \mathbf{y}_Q)$ from target distribution and bandwidth parameters $h_0, h_0 > 0$ as inputs, and returns the classifier

$$\texttt{CL-labeled} \triangleq \texttt{Labeled-Classifier}(\mathbf{x}_P, \mathbf{y}_P, \mathbf{x}_Q, \mathbf{y}_Q, h_0, h_1)$$

  as defined in section 3, equation 3.1 and 3.3. Throughout our simulation study we use β^*-valid kernel (definition 23, Tsybakov (2009), definition 1.2 and section 1.2.2) with β^* as 3. Since the densities are infinitely differentiable on the interior of support, we expect to realize a rate of convergence with $\beta = 3$. In that regard, we fix the bandwidth parameter $h_0 = n_0^{-\frac{1}{10}},\ h_1 = n_1^{-\frac{1}{10}}$.

- `Classical-Classifier` is a function that takes the target data $(\mathbf{x}_Q, \mathbf{y}_Q)$ and a bandwidth parameter $h > 0$ as input and returns a classifier

$$\texttt{CL-classical} \triangleq \texttt{Classical-Classifier}(\mathbf{x}_Q, \mathbf{y}_Q, h)$$

  where

$$\texttt{CL-classical}(x) = \mathbb{1}\left\{\frac{\sum_{i=1}^{n_Q} Y_i^Q K_h(x - X_i^Q)}{\sum_{i=1}^{n_Q} K_h(x - X_i^Q)} \geq \frac{1}{2}\right\}.$$

  We fix $h_Q = \frac{1}{2} n_Q^{-\frac{1}{6}}$.

- `Unlabeled-Classifier` takes the data $(\mathbf{x}_P, \mathbf{y}_P)$ from source, x_Q from target distribution, a classifier g fitted on the source distribution and bandwidth parameters $h_0, h_1 > 0$ as inputs, and returns the classifier

$$\texttt{CL-unlabeled} \triangleq \texttt{Unlabeled-Classifier}(x_P, y_P, x_Q, g, h)$$

. It first estimates π_Q using distribution matching approach Lipton et al. (2018); Azizzadenesheli et al. (2019); Alexandari et al. (2020). For this particular simulation study we use Lipton et al. (2018). The classifier g is a non-parametric classifier Audibert and Tsybakov (2007) fitted on $x_P,\ y_P$ with same kernel and bandwidths. The final classifier is obtained by an appropriate re-weighting of the P-samples. We fix $h_0 = (n_0')^{-1/7}, h_1 = (n_1')^{-1/7}$.

- `Oracle-Classifier` takes the source data $x_P,\ y_P$, and π_Q and bandwidths $h_0, h_1 > 0$ as inputs, and returns a classifier

$$\texttt{CL-oracle} \triangleq \texttt{Oracle-Classifier}(x_P, y_P, w_0, w_1, h)$$

exactly same as in `Unlabeled-Classifier` with actual value π_Q used for data generating purpose.Here, we use same kernel and bandwidths.

Appendix B. Proof of Theorem 6 and 13

B.1 Definitions

In this subsection we define β-valid kernel, convergence rates and parametric convergence rates.

Definition 23 (β-valid kernel) *Let K be a real-valued function on $\mathbb{R}^d$, with support $[-1,1]^d$. For fixed $\beta > 0$, the function $K(\cdot)$ is said to be a β-valid kernel if it satisfies $\int K = 1,\ \int |K|^p < \infty$ for any $p \geq 1,\ \int \|t\|^\beta |K(t)|dt < \infty$ and, in the case $\lfloor \beta \rfloor \geq 1$, it satisfies $\int t^s K(t)dt = 0$ for any $s = (s_1, \ldots, s_d) \in \mathbb{N}^d$ such that $1 \leq |s| \leq \lfloor \beta \rfloor$.*

We refer to Tsybakov (2009) section 1.2.2 for construction of such kernels for 1 dimensional data. Kernel for d-dimensional data can be constructed as $K'(x_1, \ldots, x_d) = K(x_1) \ldots K(x_d)$.

The proof of upper bound is broken into some technical lemmas. These lemmas are states in terms of general rate of convergence for parameter π_Q and densities g_0 and g_1. Formal definitions of these rates are given later. We denote the source-target sample size pair (n_P, n_Q) by n, *i.e.* $n \equiv (n_P, n_Q)$.

Definition 24 (Parameter estimation rate) *For $(P, Q) \in \Pi$ let $\hat{\theta}_n$ is an estimator of the parameter $\theta = \theta(P, Q) \in \mathbb{R}$. For non-increasing sequences (φ_n) and (ψ_n) of positive numbers we say $\hat{\theta}_n$ converges to θ at a (φ_n, ψ_n)-rate uniformly on $\mathcal{P}$ if there exists positive numbers c_1, c_2, c_ψ for any $\delta > 0$*

$$\sup_{(P,Q)\in\mathcal{P}} \mathbb{P}\big(|\hat{\theta}_n - \theta| > \delta\big) \leq c_1 \exp\big(-c_2(\delta/\varphi_n)^2\big), \quad \delta \leq c_\psi \psi_n.$$

Definition 25 (Function estimation rate) *For a pair $(P,Q)\in\mathcal{P}$ let $\hat{p}_n$ be an estimator for $p\equiv p(P,Q):\mathcal{X}\to\mathbb{R}$. Let (φ_n) be a sequence of non-increasing positive numbers. We say $\hat{p}_n$ converges pointwise to p at a φ_n-rate for (P,Q) if there exists positive constants c_1,c_2,Δ and c_φ such that for Q_X almost surely all $x\in\mathcal{X}$ we have*

$$\mathbb{P}\big(|\hat{p}_n(x)-p(x)|>\delta\big)\le c_1\mathsf{exp}\big(-c_2(\delta/\varphi_n)^2\big),\quad c_\varphi\varphi_n<\delta<\Delta.$$

We say $\hat{p}_n$ converges pointwise to p at a φ_n-rate uniformly on $\mathcal{P}$, if the above happens for all $(P,Q)\in\mathcal{P}$ for some constants c_1,c_2,Δ and c_φ independent of (P,Q).

B.2 Required Lemmas

The proof of upper bound is broken into three main lemmas, which are presented in this subsection.

Lemma 26 (Concentration of $\hat{\eta}_Q$) *Let $\hat{\pi}_Q^{(n)}$ converges to π_Q at a (φ_n,ψ_n)-rate and for $i\in\{0,1\}$ let $\hat{g}_i^{(n)}$ converges pointwise to g_i at a $\tau_n^{(i)}$ rate, uniformly on $\mathcal{P}$. Then there exists positive constants c_0,c_1,c_2 such that $\hat{\eta}_Q$ converges pointwise to η_Q at a $\big(c_0(1-\pi_Q+\psi_n)\tau_n^{(0)}+c_1(\pi_Q+\psi_n)\tau_n^{(1)}+c_2\varphi_n\big)$-rate uniformly over $\mathcal{P}$.*

Proof We break the proof in several steps.
Step 1: Upper bound for $\hat{\eta}_Q$. Note that both $\hat{\eta}_Q(x)$ and $\eta_Q(x)$ can be expresses as

$$\eta_Q(x)=\frac{\pi_Q g_1(x)}{\pi_Q g_1(x)+(1-\pi_Q)g_0(x)}$$
$$\hat{\eta}_Q(x)=\frac{\hat{\pi}_Q\hat{g}_1(x)}{\hat{\pi}_Q\hat{g}_1(x)+(1-\hat{\pi}_Q)\hat{g}_0(x)}.$$

For the ease of notation, let us define $u(x)=\pi_Q g_1(x),\ v(x)=(1-\pi_Q)g_0(x),\ \hat{u}(x)=\hat{\pi}_Q\hat{g}_1(x)$ and $\hat{v}(x)=(1-\hat{\pi}_Q)\hat{g}_0(x)$. Then

$$\begin{aligned}
|\hat{\eta}_Q(x)-\eta_Q(x)| &= \left|\frac{\hat{u}(x)}{\hat{u}(x)+\hat{v}(x)}-\frac{u(x)}{u(x)+v(x)}\right| \\
&= \frac{|\hat{u}(x)v(x)-u(x)\hat{v}(x)|}{(\hat{u}(x)+\hat{v}(x))(u(x)+v(x))} \\
&= \frac{|\hat{u}(x)v(x)-\hat{u}(x)\hat{v}(x)+\hat{u}(x)\hat{v}(x)-u(x)\hat{v}(x)|}{(\hat{u}(x)+\hat{v}(x))(u(x)+v(x))} \\
&= \frac{|\hat{u}(x)[v(x)-\hat{v}(x)]+\hat{v}(x)[\hat{u}(x)-u(x)]|}{(\hat{u}(x)+\hat{v}(x))(u(x)+v(x))} \\
&\le \frac{\hat{u}(x)|v(x)-\hat{v}(x)|+\hat{v}(x)|\hat{u}(x)-u(x)|}{(\hat{u}(x)+\hat{v}(x))(u(x)+v(x))} \\
&\le \frac{|\hat{u}(x)-u(x)|+|\hat{v}(x)-v(x)|}{q_X(x)}.
\end{aligned}$$

Step 2: Upper bound of $\hat{u}(x)$ and $\hat{v}(x)$. Since, $\hat{\pi}_Q^{(n)}\equiv\hat{\pi}_Q$ converges to π_Q at a (φ_n,ψ_n)-rate, there exists $c_1^{(\pi)},c_2^{(\pi)},c_\pi>0$ such that for any $\delta>0$

$$\sup_{(P,Q)\in\Pi}\mathbb{P}\Big(\big|\hat{\pi}_Q^{(n)}-\pi_Q\big|>\delta\Big)\le c_1^{(\pi)}\mathsf{exp}\Big(-c_2^{(\pi)}(\delta/\varphi_n)^2\Big),\quad \delta\le c_\pi\psi_n$$

The above inequality can be rewritten as

$$\sup_{(P,Q)\in\Pi} \mathbb{P}\Big(\big|\hat{\pi}_Q^{(n)} - \pi_Q\big| > \delta\varphi_n/\sqrt{c_2^{(\pi)}}\Big) \le c_1^{(\pi)}\mathsf{exp}\big(-\delta^2\big), \quad \delta \le c_\pi\psi_n\sqrt{c_2^{(\pi)}}/\varphi_n. \tag{B.1}$$

Fix $i \in \{0,1\}$. Since, $\hat{g}_i^{(n)}$ converges pointwise to g_i at a $\tau_n^{(i)}$-rate, there exists positive constants $c_{1,i}, c_{2,i}, \Delta_i, c_{\tau,i}$ such that for any $(P,Q) \in \mathcal{P}$ for Q_X almost surely on $\mathcal{X}$ we have

$$\mathbb{P}\Big(\big|\hat{g}_i^{(n)}(x) - g_i(x)\big| > \delta\Big) \le c_{1,i}\mathsf{exp}\big(-c_{2,i}(\delta/\tau_n^{(i)})^2\big) \qquad c_{\tau,i}\tau_n^{(i)} < \delta < \Delta_i.$$

The above inequality can be rewritten as

$$\mathbb{P}\Big(\big|\hat{g}_i^{(n)}(x) - g_i(x)\big| > \delta\tau_n^{(i)}/\sqrt{c_{2,i}}\Big) \le c_{1,i}\mathsf{exp}\big(-\delta^2\big), \qquad c_{\tau,i} < \delta < \Delta_i/\tau_n^{(i)}. \tag{B.2}$$

Using union bound, for Q_X almost surely $x \in \mathcal{X}$, with probability at least $1 - (c_{1,0} + c_{1,1} + c_1^{(\pi)})e^{-\delta^2} = 1 - c_1' e^{-\delta^2}$ we have the following

$$\big|\hat{\pi}_Q^{(n)} - \pi_Q\big| \le \delta\varphi_n/\sqrt{c_2^{(\pi)}} \tag{B.3}$$

$$\big|\hat{g}_0^{(n)}(x) - g_0(x)\big| \le \delta\tau_n^{(0)}/\sqrt{c_{2,0}} \tag{B.4}$$

$$\big|\hat{g}_1^{(n)}(x) - g_1(x)\big| \le \delta\tau_n^{(1)}/\sqrt{c_{2,1}} \tag{B.5}$$

for $c_{\tau,0} \vee c_{\tau,1} < \delta < (\Delta_0/\tau_n^{(0)}) \wedge (\Delta_1/\tau_n^{(1)}) \wedge (c_\pi\psi_n\sqrt{c_2^{(\pi)}}/\varphi_n)$. From the above inequalities, we get

$$\begin{aligned}
|\hat{u}(x) - u(x)| &= |\hat{\pi}_Q^{(n)}\hat{g}_1^{(n)}(x) - \pi_Q g_1(x)| \\
&= |\hat{\pi}_Q^{(n)}\hat{g}_1^{(n)}(x) - \hat{\pi}_Q^{(n)} g_1(x) + \hat{\pi}_Q^{(n)} g_1(x) - \pi_Q g_1(x)| \\
&\le \hat{\pi}_Q^{(n)}|\hat{g}_1^{(n)}(x) - g_1(x)| + g_1(x)|\hat{\pi}_Q^{(n)} - \pi_Q| \\
&\le \big(|\hat{\pi}_Q^{(n)} - \pi_Q| + \pi_Q\big)|\hat{g}_1^{(n)}(x) - g_1(x)| + g_1(x)|\hat{\pi}_Q^{(n)} - \pi_Q| \\
&\le \big(\pi_Q + \delta\varphi_n/\sqrt{c_2^{(\pi)}}\big)\delta\tau_n^{(1)}/\sqrt{c_{2,1}} + L^*\delta\varphi_n/\sqrt{c_2^{(\pi)}} \\
&\le \big(\pi_Q + c_\pi\psi_n\big)\delta\tau_n^{(1)}/\sqrt{c_{2,1}} + L^*\delta\varphi_n/\sqrt{c_2^{(\pi)}},
\end{aligned}$$

and similarly

$$|\hat{v}(x) - v(x)| \le \big(1 - \pi_Q + c_\pi\psi_n\big)\delta\tau_n^{(0)}/\sqrt{c_{2,0}} + L^*\delta\varphi_n/\sqrt{c_2^{(\pi)}}.$$

Step 3: Concentration of η_Q. Under strong density assumption, for Q_X almost surely $x \in \mathcal{X}$, with probability at least $1 - c_1' e^{-\delta^2}$ we have

$$\begin{aligned}
|\hat{\eta}_Q(x) - \eta_Q(x)| \le& \frac{\delta}{\mu_+}\Bigg[\big(\pi_Q + c_\pi\psi_n\big)\tau_n^{(1)}/\sqrt{c_{2,1}} \\
&+ \big(1 - \pi_Q + c_\pi\psi_n\big)\tau_n^{(0)}/\sqrt{c_{2,0}} + 2L^*\varphi_n/\sqrt{c_2^{(\pi)}}\Bigg] \\
=& \delta r_n
\end{aligned}$$

for $c_{\tau,0} \vee c_{\tau,1} < \delta < (\Delta_0/\tau_n^{(0)}) \wedge (\Delta_1/\tau_n^{(1)}) \wedge (c_\pi \psi_n \sqrt{c_2^{(\pi)}}/\varphi_n)$. We can rewrite this as for Q_X almost surely $x \in \mathcal{X}$

$$\sup_{(P,Q)\in\mathcal{P}} \mathbb{P}\Big(|\hat{\eta}_Q(x) - \eta_Q(x)| > \delta\Big) \leq c_1' \mathsf{exp}\big(-(\delta/r_n)^2\big)$$

for $\big(c_{\tau,0}\vee c_{\tau,1}\big) r_n < \delta < r_n\Big[(\Delta_0/\tau_n^{(0)}) \wedge (\Delta_1/\tau_n^{(1)}) \wedge (c_\pi \psi_n \sqrt{c_2^{(\pi)}}/\varphi_n)\Big]$. Let $c_r = c_{\tau,0} \vee c_{\tau,1}$. Since, $r_n/\tau_n^{(0)} \geq 1/\sqrt{c_{2,0}}$, $r_n/\tau_n^{(1)} \geq 1/\sqrt{c_{2,1}}$ and $r_n\psi_n/\varphi_n \geq c_\pi \sqrt{c_2^{(\pi)}}$ letting $\Delta = (\Delta_0/\sqrt{c_{2,0}}) \wedge (\Delta_1/\sqrt{c_{2,1}}) \wedge (c_\pi \sqrt{c_2^{(\pi)}})$ we get, for Q_X almost surely $x \in \mathcal{X}$

$$\sup_{(P,Q)\in\mathcal{P}} \mathbb{P}\Big(|\hat{\eta}_Q(x) - \eta_Q(x)| > \delta\Big) \leq c_1' \mathsf{exp}\big(-(\delta/r_n)^2\big), \quad c_r r_n < \delta < \Delta.$$

■

Lemma 27 (Bound on $\mathbb{E}\mathcal{E}_Q(\hat{f})$) *Suppose an estimate $\hat{\eta}_Q$ of the regression function η_Q converges pointwise at a r_n-rate uniformly on $\mathcal{P}$. Then under α-margin condition there exists a positive constant C such that*

$$\sup_{(P,Q)\in\mathcal{P}} \mathbb{E}\Big[\mathcal{E}_Q(\hat{f})\Big] \leq C r_n^{1+\alpha}.$$

.

Proof Since, $\hat{\eta}_Q$ converges pointwise to η_Q at a r_n-rate, there exist positive constants c_1, c_2, Δ, c_r such that for Q_X almost surely all $x \in \mathcal{X}$

$$\sup_{(P,Q)\in\mathcal{P}} \mathbb{P}\Big(|\hat{\eta}_Q(x) - \eta_Q(x)| > \delta\Big) \leq c_1 \mathsf{exp}\big(-c_2(\delta/r_n)^2\big), \quad c_r r_n < \delta < \Delta.$$

Recall, under α-margin condition there exists $c_\alpha > 0$ such that

$$Q_X\big(|\eta_Q(x) - 1/2| > \delta\big) \leq c_\alpha \delta^\alpha.$$

We replace c_α by $c_\alpha(\Delta/2)^{-\alpha} \vee 1$ so that $c_\alpha(\Delta/2)^\alpha \geq 1$.

We define the following events.

$$A_0 = \Big\{x \in \mathbb{R}^d : 0 < |\eta_Q(x) - 1/2| < \delta\Big\}$$

and for $j \geq 1$,

$$A_j = \Big\{x \in \mathbb{R}^d : 2^{(j-1)}\delta < |\eta_Q(x) - 1/2| < 2^j\delta\Big\}$$

Now,

$$\begin{aligned}\mathcal{E}_Q(\hat{f}) =& 2\mathbb{E}_X\left(|\eta_Q(X)-1/2|\,\mathbb{1}_{\{\hat{f}(X)\neq f^*(X)\}}\right)\\ =& 2\sum_{j=0}^{\infty}\mathbb{E}_X\left(|\eta_Q(X)-1/2|\,\mathbb{1}_{\{\hat{f}(X)\neq f^*(X)\}}\mathbb{1}_{\{X\in A_j\}}\right)\\ \leq& 2\delta\mathbb{E}_X\left(0<\left|\eta_Q(X)-\frac{1}{2}\right|<\delta\right)\\ &+2\sum_{j=1}^{\infty}\mathbb{P}_X\left(\left|\eta_Q(X)-\frac{1}{2}\right|\mathbb{1}_{\{\hat{f}(X)\neq f^*(X)\}}\mathbb{1}_{\{X\in A_j\}}\right)\end{aligned}$$

Let $\delta = c_r r_n$. Then $2^{j-1}\delta \geq \Delta/2$ if $j \geq \log_2(\Delta/\delta)$.

On the event $\{\hat{f} \neq f^*\}$ we have $\left|\eta_Q - \frac{1}{2}\right| \leq |\hat{\eta}-\eta|$. So, for any $1 \leq j < \log_2(\Delta/\delta)$ we get

$$\begin{aligned}&2\mathbb{E}_X\mathbb{E}\left(\left|\eta_Q(X)-\frac{1}{2}\right|\mathbb{1}_{\{\hat{f}(X)\neq f^*(X)\}}\mathbb{1}_{\{X\in A_j\}}\right)\\ &\leq 2^{j+1}\delta\mathbb{E}_X\mathbb{E}\left(\mathbb{1}_{\{|\hat{\eta}_Q(X)-\eta_Q(X)|\geq 2^{j-1}\delta\}}\mathbb{1}_{\{0<|\eta_Q(X)-1/2|<2^j\delta\}}\right)\\ &= 2^{j+1}\delta\mathbb{E}_X\left[\mathbb{P}\left(\mathbb{1}_{\{|\hat{\eta}_Q(X)-\eta_Q(X)|\geq 2^{j-1}\delta\}}\right)\mathbb{1}_{\{0<|\eta_Q(X)-1/2|<2^j\delta\}}\right]\\ &\leq 2^{j+1}\delta\mathsf{exp}\left(-(2^{j-1}\delta/r_n)^2\right)\mathbb{P}_X(0<|\eta_Q(X)-1/2|<2^j\delta)\\ &\leq 2C_\alpha 2^{j(1+\alpha)}\delta^{1+\alpha}\mathsf{exp}\left(-(2^{j-1}\delta/r_n)^2\right).\end{aligned}$$

For $j \geq \log_2(\Delta/\delta)$ we have $\mathbb{P}\big(|\eta_Q(x)-1/2| \geq 2^{j-1}\delta\big) = 1$ and hence

$$\mathbb{P}\Big(2^{j-1}\delta < |\eta_Q(x)-1/2| < 2^j\delta\Big) = 0.$$

This means

$$\begin{aligned}&2\mathbb{E}_X\mathbb{E}\left(\left|\eta_Q(X)-\frac{1}{2}\right|\mathbb{1}_{\{\hat{f}(X)\neq f^*(X)\}}\mathbb{1}_{\{X\in A_j\}}\right)\\ &\leq 2^{j+1}\delta\mathbb{E}_X\mathbb{E}\left(\mathbb{1}_{\{|\hat{\eta}_Q(X)-\eta_Q(X)|\geq 2^{j-1}\delta\}}\mathbb{1}_{\{2^{j-1}\delta<|\eta_Q(X)-1/2|<2^j\delta\}}\right)\\ &= 2^{j+1}\delta\mathbb{E}_X\left[\mathbb{P}\left(\mathbb{1}_{\{|\hat{\eta}_Q(X)-\eta_Q(X)|\geq 2^{j-1}\delta\}}\right)\mathbb{1}_{\{2^{j-1}\delta<|\eta_Q(X)-1/2|<2^j\delta\}}\right]\\ &= 0.\end{aligned}$$

Finally, we get

$$\begin{aligned}\sup_{\mathbb{P}\in\mathcal{P}}\mathbb{E}\Big[\mathcal{E}_Q(\hat{f})\Big] &\leq 2C_\alpha\left(\delta^{1+\alpha}+\sum_{j\geq 1}2^{j(1+\alpha)}\delta^{1+\alpha}\mathsf{exp}\left(-(2^{j-1}\delta/r_n)^2\right)\right)\\ &\leq Cr_n^{1+\alpha}.\end{aligned}$$

■

Now we provide a rate of convergence for the density estimator $\hat{g}_1^{(n_1)}$. Later we only provide the statement for the rate of convergence for the density estimator $\hat{g}_0^{(n_1)}$. The proof will be similar.

Lemma 28 (Rate of convergence for conditional density estimates) *Let $(P, Q) \in \mathcal{P}$. For $i = 0, 1$ let $m_i(n) \equiv m_i(n, \pi) = \mathbb{E}[n_i]$. Then for $h_i = n_i^{-1/(2\beta+d)}$ the density estimator $\hat{g}_i^{(n_i)}$ converges pointwise to g_i at a $m_i(n)^{-\beta/(2\beta+d)}$-rate.*

Proof We only prove the result for g_1. The proof for g_0 will be similar. Let n_1 be the number of sample points with label 1 (which is sum of independent Bernoulli variables). Let $v_1(n) = \mathrm{Var}(n_1)$. Clearly, $v_1(n) \leq m_1(n)$.

Using Bernsteins's inequality, for any $t > 0$ we get

$$\begin{aligned}
\mathbb{P}\left(|n_1 - m_1(n)| > t\right) \leq & 2\mathsf{exp}\left(-\frac{t^2/2}{v_1(n) + t/3}\right) \\
\leq & 2\mathsf{exp}\left(-\frac{t^2/2}{2v_1(n)}\right) \text{ for } t \leq 3v_1(n), \\
\leq & 2\mathsf{exp}\left(-\frac{t^2}{4m_1(n)}\right) \text{ for } t \leq 3v_1(n).
\end{aligned}$$

Letting $t = \delta m_1(n)^{(2\beta+d/2)/(2\beta+d)}$ we get

$$\mathbb{P}\left(|n_1 - m_1(n)| > \delta m_1(n)^{\frac{2\beta+d/2}{2\beta+d}}\right) \leq 2\mathsf{exp}\left(-\frac{\delta^2}{4} m_1(n)^{\frac{2\beta}{2\beta+d}}\right),$$

for $\delta \leq \frac{3v_1(n)}{m_1(n)^{\frac{2\beta+d/2}{2\beta+d}}}$.

Using (Rigollet and Vert, 2009, Lemma 4.1) we get positive constants c_1, c_2, c', Δ such that for Q_X almost sure all $x \in \mathcal{X}$

$$\sup_{(P,Q)\in\mathcal{P}} \mathbb{P}\Big(\big|\hat{g}_1^{(n_1)}(x) - g_1(x)\big| > \delta \Big| n_1\Big) \leq c_1 \mathsf{exp}\big(-c_2 n_1 h_1^d \delta^2\big), \quad c' h_1^\beta < \delta < \Delta.$$

Letting $h_1 = n_1^{-\frac{1}{2\beta+d}}$ we get

$$\sup_{(P,Q)\in\mathcal{P}} \mathbb{P}\Big(\big|\hat{g}_1^{(n_1)}(x) - g_1(x)\big| > \delta \Big| n_1\Big) \leq c_1 \mathsf{exp}\big(-c_2 n_1^{\frac{2\beta}{2\beta+d}} \delta^2\big), \quad c' n_1^{-\frac{\beta}{2\beta+d}} < \delta < \Delta.$$

Let $n^{(0)} = (n_P^{(0)}, n_Q^{(0)})$ such that $\Delta \leq \left(\frac{3v_1(n^{(0)})}{m_1(n^{(0)})^{\frac{(2\beta+d/2)}{(2\beta+d)}}}\right) \bigwedge \left(\frac{1}{2} m_1(n^{(0)})^{\frac{d/2}{2\beta+d}}\right)$. We say $n \geq n^{(0)}$ if $n_P \geq n_P^{(0)}$ and $n_Q \geq n_Q^{(0)}$. For any $n \geq n^{(0)}$ we have $\delta m(n)^{\frac{2\beta+d/2}{2\beta+d}} \leq m(n)/2$.

Now, for $n \geq n^{(0)}$, and Q_X almost surely all $x \in \mathcal{X}$

$$\begin{aligned}
&\mathbb{P}\Big(\big|\hat{g}_1^{(n_1)}(x) - g_1(x)\big| > \delta\Big) \\
\leq &\mathbb{P}\Big(\big|\hat{g}_1^{(n_1)}(x) - g_1(x)\big| > \delta, |n_1 - m_1(n)| \leq \delta m_1(n)^{\frac{2\beta+d/2}{2\beta+d}}\Big) \\
&+ \mathbb{P}\left(|n_1 - m_1(n)| > \delta m_1(n)^{\frac{2\beta+d/2}{2\beta+d}}\right) \\
\leq &\mathbb{P}\Big(\big|\hat{g}_1^{(n_1)}(x) - g_1(x)\big| > \delta, |n_1 - m_1(n)| < m_1(n)/2\Big) \\
&+ 2\mathsf{exp}\left(-\frac{\delta^2}{4} m_1(n)^{\frac{2\beta}{2\beta+d}}\right) \\
\leq &\mathbb{E}\left[\mathbb{P}\Big(\big|\hat{g}_1^{(n_1)}(x) - g_1(x)\big| > \delta \Big| n_1\Big) \mathbb{1}_{m_1(n)/2 \leq n_1 \leq 3m_1(n)/2}\right] \\
&+ 2\mathsf{exp}\left(-\frac{\delta^2}{4} m_1(n)^{\frac{2\beta}{2\beta+d}}\right) \\
\leq &c_1 \mathsf{exp}\left(-c_2 \left(\frac{m_1(n)}{2}\right)^{\frac{2\beta}{2\beta+d}} \delta^2\right) \\
&+ 2\mathsf{exp}\left(-\frac{\delta^2}{4} m_1(n)^{\frac{2\beta}{2\beta+d}}\right), \text{for } c' \left(\frac{m_1(n)}{2}\right)^{-\frac{\beta}{2\beta+d}} < \delta < \Delta.
\end{aligned}$$

Hence, there exists c_1', c_2', c', Δ such that for Q_X almost surely all $x \in \mathcal{X}$

$$\mathbb{P}\left(|\hat{g}_1^{(n_1)}(x) - g_1(x)| > \delta\right) \leq c_1' \mathsf{exp}\left(-c_2' m_1(n)^{\frac{2\beta}{2\beta+d}} \delta^2\right),$$

for $c' m_1(n)^{-\frac{\beta}{2\beta+d}} < \delta < \Delta$. ■

B.3 Upper Bounds

B.3.1 Proof of Theorem 6

Proof Since, $\hat{\pi}_Q^{(n)} = \frac{1}{n_Q} \sum_{i=1}^{n_Q} Y_i^{(Q)}$ using Bernstein's inequality

$$\mathbb{P}\left(\big|\hat{\pi}_Q^{(n)} - \pi_Q\big| > t\right) \leq 2\mathsf{exp}\left(-\frac{t^2/2}{\pi_Q(1-\pi_Q)/n_Q + t/(3n_Q)}\right).$$

Letting $t \leq 3\pi_Q(1-\pi_Q)$ we have

$$\mathbb{P}\left(\big|\hat{\pi}_Q^{(n)} - \pi_Q\big| > t\right) \leq 2\mathsf{exp}\left(-\frac{n_Q t^2}{4\pi_Q(1-\pi_Q)}\right).$$

Hence, $\hat{\pi}_Q^{(n)}$ converges to π_Q at a $((\pi_Q(1-\pi_Q))/n_Q)^{1/2}, \pi_Q(1-\pi_Q))$-rate uniformly over $\mathcal{P}$.

Since $n_1 = \sum_{i=1}^{n_P} Y_i^{(P)} + \sum_{i=1}^{n_Q} Y_i^{(Q)}$ from lemma 28 we see that $\hat{g}_1^{(n_1)}$ converges pointwise to g_1 at a $m(n)^{-\beta/(2\beta+d)}$-rate uniformly on $\mathcal{P}$. Similarly, $\hat{g}_0^{(n_1)}$ converges pointwise to g_0 at

a $\big(n_P + n_Q - m(n)\big)^{-\beta/(2\beta+d)}$-rate. We refer to lemma 26 to conclude $\hat{\eta}_Q$ converges to η_Q at a rate

$$r(n,\pi) = \begin{cases} \sqrt{\dfrac{(1-\pi_Q)\pi_Q}{n_Q}} + \sqrt{\pi_Q}\Big(\pi_P n_P + \pi_Q n_Q\Big)^{-\frac{\beta}{2\beta+d}} \\ + \sqrt{(1-\pi_Q)}\Big((1-\pi_P)n_P + (1-\pi_Q)n_Q\Big)^{-\frac{\beta}{2\beta+d}} \end{cases} = (A) + (B) + (C) \qquad \text{(B.6)}$$

uniformly on $\mathcal{P}$.

To complete the proof, we derive the worst possible rate over the function class. First, we seek the worst rate with respect to π_Q. Considering them term by term we see that $(A) \leq \frac{1}{2\sqrt{n_Q}}$, (B) is an increasing function of π_Q and hence

$$(B) \leq \big(\pi_P n_P + n_Q\big)^{-\frac{\beta}{2\beta+d}}.$$

By similar logic we see that

$$(C) \leq \big((1-\pi_P)n_P + n_Q\big)^{-\frac{\beta}{2\beta+d}}.$$

Hence,

$$\begin{aligned} r(n,\pi) &\leq \frac{1}{2\sqrt{n_Q}} + \big(\pi_P n_P + n_Q\big)^{-\frac{\beta}{2\beta+d}} + \big((1-\pi_P)n_P + n_Q\big)^{-\frac{\beta}{2\beta+d}} \\ &\leq \frac{1}{\sqrt{n_Q}} + \big(\pi_P n_P + n_Q\big)^{-\frac{\beta}{2\beta+d}} + \big((1-\pi_P)n_P + n_Q\big)^{-\frac{\beta}{2\beta+d}} \end{aligned}.$$

But a concern regarding the above calculation is whether this dominating rate is achievable? The following argument shows that it is achieved by $\pi_Q = 1/2$.

$$\begin{aligned} r(n,\pi;\pi_Q = 1/2) &= \frac{1}{2\sqrt{n_Q}} + \frac{1}{\sqrt{2}}\big(\pi_P n_P + n_Q/2\big)^{-\frac{\beta}{2\beta+d}} \\ &\quad + \frac{1}{\sqrt{2}}\big((1-\pi_P)n_P + n_Q/2\big)^{-\frac{\beta}{2\beta+d}} \\ &\geq \frac{1}{4\sqrt{n_Q}} + \frac{1}{4}\big(\pi_P n_P + n_Q\big)^{-\frac{\beta}{2\beta+d}} \\ &\quad + \frac{1}{4}\big((1-\pi_P)n_P + n_Q\big)^{-\frac{\beta}{2\beta+d}} \\ &= \frac{1}{4}\Bigg[\frac{1}{\sqrt{n_Q}} + \big(\pi_P n_P + n_Q\big)^{-\frac{\beta}{2\beta+d}} \\ &\quad + \big((1-\pi_P)n_P + n_Q\big)^{-\frac{\beta}{2\beta+d}}\Bigg] \end{aligned}$$

This explains why $\pi_Q = 1/2$ exhibits the worst behavior.

Next, denoting $f(\pi_P) = \big(\pi_P n_P + n_Q\big)^{-\frac{\beta}{2\beta+d}}$ we note that

$$\max\Big\{f(\pi_P), f(1-\pi_P)\Big\} \leq f(\pi_P) + f(1-\pi_P) \leq 2\max\Big\{f(\pi_P), f(1-\pi_P)\Big\},$$

which implies $\max\Big\{f(\pi_P), f(1-\pi_P)\Big\}$ and $f(\pi_P)+f(1-\pi_P)$ have same rate of convergence. Furthermore, using the fact that f is a decreasing function we get

$$\max\Big\{f(\pi_P), f(1-\pi_P)\Big\} = f(\min\{\pi_P, 1-\pi_P\}) \leq f(\epsilon_P),$$

where the last inequality is an equality in the worst possible case ($\pi_P = \epsilon_P$ or $1-\epsilon_P$). Hence, we finally get the worst possible rate when $\pi_Q = 1/2$ and $\pi_P = \epsilon_P$ or $1-\epsilon_P$:

$$n_Q^{-1/2} + \big(\epsilon_P n_P + n_Q\big)^{-\frac{\beta}{2\beta+d}}.$$

Finally, under α-margin condition lemma 27 we have

$$\sup_{(P,Q)\in\mathcal{P}} \mathbb{E}\mathcal{E}_Q(\hat{f}) \leq C\left(n_Q^{-1/2} + (\epsilon_P n_P + n_Q)^{-\frac{\beta}{2\beta+d}}\right)^{1+\alpha}.$$

■

B.3.2 Proof of Theorem 13

Proof Let $n_1 = \sum_{i=1}^{n_P} Y_i^{(P)}$ and $n_0 = n_P - n_1$.

We use the method by Iyer et al. (2014) to estimate π_Q with a Gaussian kernel $K_c(x,y) = A\mathsf{exp}(-c\|x-y\|_2^2)$, where $A > 0$ is suitably chosen to satisfy $\int_{\mathcal{X}^2} K_c(x,y)dxdy = 1$. This ensures K_c to be a joint density on $\mathcal{X}^2$. Furthermore, at the limit $c \to \infty$ the joint probability distribution corresponding to K_c converges to an uniform distribution on the line $x = y$. Since $\big(g_0(x) - g_1(x)\big)\big(g_0(y) - g_1(y)\big)$ is a continuous function on a compact support $\mathcal{X}^2$, this is bounded. We apply bounded convergence theorem to conclude

$$\lim_{c\to\infty} \int_{\mathcal{X}^2} \big(g_0(x) - g_1(x)\big)\big(g_0(y) - g_1(y)\big)K_c(x,y)dxdy = \int_{\mathcal{X}} \big(g_0(x) - g_1(x)\big)^2 dx.$$

From the Assumption **??** we have

$$\int_{\mathcal{X}} \big(g_0(x) - g_1(x)\big)^2 dx \geq C^2,$$

which implies

$$\int_{\mathcal{X}^2} \big(g_0(x) - g_1(x)\big)\big(g_0(y) - g_1(y)\big)K_c(x,y)dxdy \geq \frac{C^2}{2},$$

for a large enough $c > 0$. Denoting the corresponding feature vector (of the kernel K_c) as Φ_c we see that $\bar{A}$ in Iyer et al. (2014), Equation (2) is merely $\int \Phi_c(x)(g_1(x) - g_0(x))dx$. Hence we have

$$\begin{aligned}
\bar{A}^\top \bar{A} &= \left\langle \int \Phi_c(x)(g_1(x) - g_0(x))dx, \int \Phi_c(y)(g_1(y) - g_0(y))dy \right\rangle \\
&= \int_{\mathcal{X}^2} \langle \Phi_c(x), \Phi_c(y)\rangle \big(g_0(x) - g_1(x)\big)\big(g_0(y) - g_1(y)\big)dxdy \\
&= \int_{\mathcal{X}^2} K_c(x,y)\big(g_0(x) - g_1(x)\big)\big(g_0(y) - g_1(y)\big)dxdy \geq \frac{C^2}{2}
\end{aligned}$$

Using the statement right after Lemma 2 in Iyer et al. (2014) we get

$$2(\hat{\pi}_Q - \pi_Q)^2 \leq \frac{R^2\left(\frac{c^2+2c+2}{n_Q} + \frac{2}{n_0} + \frac{2}{n_1}\right)\left(1+\sqrt{\log(4/\delta)}\right)^2}{\bar{A}^\top \bar{A} - 8R^2\left(\frac{1}{n_0}+\frac{1}{n_1}\right)\sqrt{\left(\frac{c^2}{n_0}+\frac{1}{n_1}\right)\log(2/\delta)}}$$

with probability at least $1-\delta$. Here $c=1$, $R = \max_{x\in\mathcal{X}} \|\Phi_c\|$, n_0 is the number of observations with $Y=0$ in source and similarly n_1. We note that the numbers R and $\bar{A}^\top\bar{A} \geq \frac{C_k^2}{2}$ are fixed and n_0, n_1, n_Q are growing. Hence, for sufficiently large n_0, n_1 and n_Q we get

$$|\hat{\pi}_Q - \pi_Q| \leq C'\left(\sqrt{\frac{1}{n_Q}+\frac{1}{n_0}+\frac{1}{n_1}}\right)\left(1+\sqrt{\log(4/\delta)}\right)$$

with probability at least $1-\delta$. In other words, there is a $c>0$ such that for any n_1 with probability (conditioned over n_1) $\geq 1-e^{-t^2}$ the following holds

$$|\hat{\pi}_Q - \pi_Q| \leq ct\sqrt{\frac{1}{n_Q}+\frac{1}{n_0}+\frac{1}{n_1}}\,.$$

Recall from the Bernstein inequality in 28 we have

$$\pi_P n_P - c't\sqrt{n_P\pi_P(1-\pi_P)} \leq n_1 \leq \pi_P n_P + c't\sqrt{n_P\pi_P(1-\pi_P)}$$

with probability $\geq 1-e^{-t^2}$ for $t \leq 2\sqrt{n_P\pi_P(1-\pi_P)}$ for some c'. Hence, for $t \leq 1\sqrt{2c'n_P\pi_P(1-\pi_P)}$ with probability $\geq 1-2e^{-t^2}$ we have

$$\begin{aligned}
&|\hat{\pi}_Q - \pi_Q| \\
&\leq ct\sqrt{\frac{1}{n_Q}+\frac{1}{n_0}+\frac{1}{n_1}} \\
&\leq ct\sqrt{\frac{1}{n_Q}+\frac{1}{(1-\pi_P)n_P - c't\sqrt{n_P\pi_P(1-\pi_P)}}+\frac{1}{\pi_P n_P - c't\sqrt{n_P\pi_P(1-\pi_P)}}} \\
&\leq c_1 t\sqrt{\frac{1}{n_Q}+\frac{1}{\epsilon_P n_P - c't\sqrt{\epsilon_P n_P}}} \\
&\leq c_2 t\sqrt{\frac{1}{n_Q}+\frac{1}{\epsilon_P n_P}} \\
&\leq c_3\left(\frac{1}{\sqrt{n_Q}}+\frac{1}{\sqrt{n_P\epsilon_P}}\right)
\end{aligned}$$

for suitably chosen constants. This implies $\hat{\pi}_Q$ converges to π_Q at a $\frac{1}{\sqrt{n_Q}}+\frac{1}{\sqrt{n_P\epsilon_P}}$ rate.

From lemma 28 we see that $\hat{g}_1^{(n_1)}$ converges pointwise to g_1 at a $(n_P\pi_P)^{-\beta/(2\beta+d)}$-rate and $\hat{g}_0^{(n_0)}$ converges pointwise to g_0 at a $\left(n_P(1-\pi_P)\right)^{-\beta/(2\beta+d)}$-rate. Rest of the proof is same as in the proof of B.3.1. ■

Appendix C. Additional detail for proof of lower bounds

C.1 Proof of Theorem 5

The proof has two parts. The first part shows that the minimax rate is at least $(\epsilon_P n_P + n_Q)^{-\beta(1+\alpha)/(2\beta+d)}$. This arises from the difficulties of estimating the non-parametric parts g_0 and g_1. The second part shows that the minimax rate is at least $n_Q^{-(1+\alpha)/2}$. This part is relatively easy to work with and mainly stems from the estimation of parametric part π_Q.

C.1.1 Difficulty of the non-parametric part

The crux of this part lies in construction of a family of distribution $\{\Pi_i\}_{i=1}^M$.

Let $r = c_r(\epsilon_P n_P + n_Q)^{-1/(2\beta+d)}, m = \lfloor c_m r^{\alpha\beta-d} \rfloor$, where $\alpha \geq 0$ is the noise condition exponent and $\beta > 0$ is the Hölder smoothness exponent.

Also, define $c_r = (1/17),\ c_m = 8 \times 17^{\alpha\beta-d}$ and $c_w \in (0,1)$ is a constant to be picked later. The preceding constants satisfy

$$8 \leq m < \frac{1}{2}\left\lfloor \frac{1}{r} \right\rfloor^d.$$

To see the first inequality, we observe that $r \leq 1/17$ and recall $\alpha\beta \leq 1 \leq d$. Thus

$$m = 8 \times (17r)^{\alpha\beta-d} \geq 8.$$

To see the second inequality, we observe that $r^{-1} \geq 16$, which implies $r^{-1} \leq 17\lfloor r^{-1} \rfloor/16$. Thus

$$m = 8(17r)^{\alpha\beta}\left(\frac{1}{17r}\right)^d \leq 8 \cdot 16^{-d}\lfloor r^{-1} \rfloor^d \leq \frac{1}{2}\lfloor r^{-1} \rfloor^d.$$

We also have $2mw = 2mc_w r^d \leq 2c_m c_w < 1$ for a suitable c_w.

Construction of $\{\Pi_i\}_{i=1}^M$. Let $r_1 = 1/\lfloor 1/(c_w r) \rfloor$ if $\lfloor 1/(c_w r) \rfloor$ is even, otherwise let $r_1 = 1/\left(\lfloor 1/(c_w r) \rfloor + 1\right)$. Let us consider the grid of points

$$\mathcal{Z} = \{(1/2 + i)r_1 : i = 0, 1, \ldots, 1/r_1 - 1\}^d. \tag{C.1}$$

We see that $\mathcal{Z}$ is a grid of equally spaced points of size r_1^{-d}. For a $z \in \mathcal{Z}$ we consider the hyper-cube

$$C(z) = \{x \in [0,1]^d : \|x - z\|_\infty \leq r_1/2\}.$$

Note that, volume of each of these hyper-cubes is r_1^d. Let $\mathcal{Z}_1, \mathcal{Z}_2 \subset \mathcal{Z}$ be subsets of size m. Moreover, we let $\mathcal{Z}_1$ and $\mathcal{Z}_2$ are disjoint. We define a bijection $u : \mathcal{Z}_1 \to \mathcal{Z}_2$ which shall be used to construct the conditional densities. We define $\mathcal{Z}_0 = \mathcal{Z} \backslash (\mathcal{Z}_1 \cup \mathcal{Z}_2)$. Note that, $\mathcal{Z}$ has even number of points and $|\mathcal{Z}_1| = |\mathcal{Z}_2|$. Hence, $\mathcal{Z}_0$ has even number of points. We further divide $\mathcal{Z}_0$ in two sets $\mathcal{Z}_3, \mathcal{Z}_4$ of equal sizes. We shall define a set of distributions parametrized by $\sigma \in \{-1, 1\}^{\mathcal{Z}_1}$.

Conditional densities. For $a > 0$ we define a function v_a supported on on $\mathbb{R}$ which will be used heavily for the construction of conditional densities.

Define

$$u_a(x) = \begin{cases} 0 & \text{for } x < 0 \\ \frac{\int_0^x e^{-\frac{1}{at(1-t)}}\,dt}{\int_0^1 e^{-\frac{1}{at(1-t)}}\,dt} & \text{for } 0 \le x \le 1 \\ 1 & \text{for } x > 1 \end{cases}$$

and

$$v_a(x) = \begin{cases} \big(1 - u_a(x)\big)^{1/\alpha} & \text{for } \beta < 1, \\ \big(1 - u_a(x)\big) & \text{for } \beta \ge 1. \end{cases} \tag{C.2}$$

According to lemma 34 we choose a such that $v_a \equiv v$ is (β, C_β) Hölder smooth. Therefore, the following functions are (β, C_β)-Hölder smooth:

$$z \in \mathcal{Z}, \quad \eta_z(x) = \frac{\mu_\Delta r^\beta}{3} v\left(\frac{2\|x - z\|_\infty}{r_1}\right)$$

and

$$z \in \mathcal{Z}, \quad \xi_z(x) = v\left(\frac{2\|x - z\|_\infty}{b r_1} - \frac{2}{r_1}\left(\frac{1}{b} - 1\right)\right).$$

Here $\mu_\Delta < 1$ is chosen later. For a parameter σ the construction of conditional densities are given below.

$$\begin{cases} g_1^\sigma(x) = \begin{cases} 1 + \sigma(z)\sqrt{\epsilon_P}\eta_z(x) & x \in C(z),\ z \in \mathcal{Z}_1, \\ 1 - \sigma(z)\sqrt{\epsilon_P}\eta_{f(z)}(x) & x \in C(f(z)),\ z \in \mathcal{Z}_1, \\ 1 + \xi_z(x) & x \in C(z),\ z \in \mathcal{Z}_3, \\ 1 - \xi_z(x) & x \in C(z),\ z \in \mathcal{Z}_4, \end{cases} \\ g_0^\sigma(x) = \begin{cases} 1 - \sigma(z)\eta_z(x) & x \in C(z),\ z \in \mathcal{Z}_1, \\ 1 + \sigma(z)\eta_{f(z)}(x) & x \in C(f(z)),\ z \in \mathcal{Z}_1, \\ 1 - \xi_z(x) & x \in C(z),\ z \in \mathcal{Z}_3, \\ 1 + \xi_z(x) & x \in C(z),\ z \in \mathcal{Z}_4. \end{cases} \end{cases} \tag{C.3}$$

We also define $\pi_Q^\sigma = 1/2$ and $\pi_P^\sigma = 1 - \epsilon_P$. We then define the probabilities

$$P_\sigma(X \in A, Y = y) = \int_A [\pi_P^\sigma g_1^\sigma(x)\mathbb{1}(y = 1) + (1 - \pi_P^\sigma) g_0^\sigma(x)\mathbb{1}(y = 0)]dx$$

and

$$Q_\sigma(X \in A, Y = y) = \int_A [\pi_Q^\sigma g_1^\sigma(x)\mathbb{1}(y = 1) + (1 - \pi_Q^\sigma) g_0^\sigma(x)\mathbb{1}(y = 0)]dx. \tag{C.4}$$

Given the source and target distributions we define the joint distribution of $\mathcal{D}_{\mathrm{U}}$ as

$$\Pi_\sigma = P_\sigma^{\otimes n_P} \otimes Q_{\sigma,X}^{\otimes n_Q} \tag{C.5}$$

Here, $\epsilon_P \leq \pi_P^\sigma \leq 1 - \epsilon_P$. Also, for any $x \in \Omega$. Hence, $\mu_- \leq q_X^\sigma(x) \leq \mu_+$ for a suitable μ_Δ. Furthermore, $\Omega = [0,1]^d$ is a regular set. Hence, q_X^σ satisfies strong density assumption.

For such a construction we refer to lemma 6.4, where it is shown Q_σ satisfies α-margin condition with constant C_α.

Let $\mathcal{F}$ be the set of all classifier relevant to this classification problem. For $\sigma \in \{-1,1\}^{\mathcal{Z}_1}$ let f_σ be the Bayes classifier corresponding to the probability distribution Q_σ defined as $f_\sigma(x) = \mathbb{1}\{\eta_{Q_\sigma}(x) \geq 1/2\}$. For $\sigma, \sigma' \in \{-1,1\}^{\mathcal{Z}_1}$ define $\bar{\rho}(\sigma, \sigma') \coloneqq \mathcal{E}_\sigma(f_{\sigma'})$ and $\rho(\sigma, \sigma') = \text{card}\{z \in \mathcal{Z}_1 : \sigma(z) \neq \sigma'(z)\}$ as the Hamming distance. Then

$$\begin{aligned}
\bar{\rho}(\sigma, \sigma') &= 2\mathbb{E}_{Q_{\sigma,X}}\left[\left|\eta_Q^\sigma(X) - \frac{1}{2}\right| \mathbf{1}\left(f_\sigma(X) \neq f_{\sigma'}(X)\right)\right] \\
&\geq c_1 r_1^d r^\beta \rho(\sigma, \sigma') \\
&\geq c_1 c_w^d r^{\beta+d} \rho(\sigma, \sigma').
\end{aligned}$$

We recall Varshamov-Gilbert bound, which shall be used to construct the probability class.

Let $\{\sigma_0, \ldots, \sigma_M\} \subset \{-1,1\}^m$ be the choice obtained from the lemma 18. Note that for such a choice $\rho(\sigma_i, \sigma_j) \geq m/8$ whenever $i \neq j$.

Then

$$\begin{aligned}
\bar{\rho}(\sigma_i, \sigma_j) &\geq c_1 c_w^d r^{\beta+d} \frac{m}{8} \\
&\geq c_1 c_w^d r^{\beta+d} r^{\alpha\beta-d} \\
&\geq c' r^{\beta(1+\alpha)} \\
&= c' (\epsilon_P n_P)^{-\frac{\beta(1+\alpha)}{2\beta+d}} \\
&\triangleq 2s
\end{aligned}$$

Now we bound the Kulback-Leibler divergence between the joint distributions Π_{σ_i}. Using lemma 33 we get

$$\begin{aligned}
KL(\Pi_{\sigma_i} || \Pi_{\sigma_j}) &\leq c_w^d K(d, \alpha, \beta) \rho(\sigma_i, \sigma_j) \\
&\leq c_w^d K(d, \alpha, \beta) m \\
&\leq \frac{1}{9} \log_2(M)
\end{aligned}$$

for suitable $c_w < 1$.

Finally we appeal to proposition 6.1 (and Markov's inequality) to obtain the minimax rate

$$\begin{aligned}\sup_{(P,Q)\in\Pi}\mathbb{E}\mathcal{E}_Q(\hat{f}) &\geq \sup_{(P,Q)\in\Pi} s\mathbb{P}_\Pi\left(\mathcal{E}_Q(\hat{f})\geq s\right)\\ &\geq s\sup_{\sigma\in\{-1,1\}^{\mathcal{Z}_1}}\Pi_\sigma\left(\mathcal{E}_{Q^\sigma}(\hat{f})\geq s\right)\\ &\geq s\frac{3-2\sqrt{2}}{8}\\ &\geq C(\epsilon_P n_P)^{-\frac{\beta(1+\alpha)}{2\beta+d}}.\end{aligned}$$

Proof of the parametric part will be exactly same as in the proof of theorem 4.1, where we get the lower bound

$$\sup_{(P,Q)\in\Pi}\mathbb{E}\big[\mathcal{E}_Q(\hat{f})\big]\geq cn_Q^{-\frac{1+\alpha}{2}}.$$

Finally, we combine the two bounds to get

$$\begin{aligned}\sup_{(P,Q)\in\Pi}\mathbb{E}\big[\mathcal{E}_Q(\hat{f})\big] &\geq cn_Q^{-\frac{1+\alpha}{2}}\vee c(\epsilon_P n_P+n_Q)^{-\frac{\beta(1+\alpha)}{2\beta+d}}\\ &\geq c'\left((\epsilon_P n_P+n_Q)^{-\frac{\beta}{2\beta+d}}+\frac{1}{\sqrt{n_Q}}\right)^{1+\alpha}.\end{aligned}$$

C.2 Additional Lemmas

Lemma 29 *For any $0\leq x\leq 1/3$ we have*

$$\log\left(\frac{1+x}{1-x}\right)\leq 3x.$$

Proof Note that for $0\leq x\leq 1/3$ we have $x-3x^2\geq 0$. Hence,

$$\begin{aligned}1+x &\leq 1+2x-3x^2\\ &\leq (1-x)(1+3x)\\ &\leq (1-x)e^{3x}.\end{aligned}$$

Taking logarithm in both sides we have the result. ■

Lemma 30 *Let $|e|\leq 1$. For $\sigma\in\{-1,1\}^{\mathcal{Z}_1}$ let p_σ be a probability density function*

$$p_\sigma(x)=\begin{cases}1+\sigma(z)e\eta_z(x) & x\in C(z),\ z\in\mathcal{Z}_1,\\ 1-\sigma(z)e\eta_{f(z)}(x) & x\in C(f(z)),\ z\in\mathcal{Z}_1,\\ 1+\xi_z(x) & x\in C(z),\ z\in\mathcal{Z}_3,\\ 1-\xi_z(x) & x\in C(z),\ z\in\mathcal{Z}_4,\end{cases},$$

as defined in C.3 and Section 6, in main document, equation 6.3. For $\sigma, \sigma' \in \{-1,1\}^{\mathcal{Z}_1}$ let us define the Hamming distance as $\rho(\sigma,\sigma') = \sum_{z\in\mathcal{Z}_1} \mathbb{1}_{\{\sigma(z)\neq\sigma'(z)\}}$. Then for $\sigma, \sigma' \in \{-1,1\}^{\mathcal{Z}_1}$

$$KL(p_\sigma||p_{\sigma'}) \leq e^2 c_w^d K(d,\alpha,\beta) r^{2\beta+d} \rho(\sigma,\sigma'), \tag{C.6}$$

for some constant $K(d,\alpha,\beta)$ only being dependent on d,α and beta.

Proof

$$\begin{aligned}
KL(p_\sigma||p_{\sigma'}) &= \int \log\left(\frac{p_\sigma(x)}{p_{\sigma'}(x)}\right) p_\sigma(x)dx \\
&= \sum_{z\in\mathcal{Z}_1} \int_{C(z)} \log\left(\frac{1+\sigma(z)e\eta_z(x)}{1+\sigma'(z)e\eta_z(x)}\right)(1+\sigma(z)e\eta_z(x))dx \\
&+ \sum_{z\in\mathcal{Z}_1} \int_{C(f(z))} \log\left(\frac{1-\sigma(z)e\eta_{f(z)}(x)}{1-\sigma'(z)e\eta_{f(z)}(x)}\right)(1-\sigma(z)e\eta_{f(z)}(x))dx \\
&= \sum_{z\in\mathcal{Z}_1} \int_{C(z)} \log\left(\frac{1+\sigma(z)e\eta_z(x)}{1+\sigma'(z)e\eta_z(x)}\right)(1+\sigma(z)e\eta_z(x))dx \\
&\quad + \sum_{z\in\mathcal{Z}_1} \int_{C(z)} \log\left(\frac{1-\sigma(z)e\eta_z(x)}{1-\sigma'(z)e\eta_z(x)}\right)(1-\sigma(z)e\eta_z(x))dx \\
&= \sum_{\sigma(z)\neq\sigma'(z)} \int_{C(z)} \log\left(\frac{1+\sigma(z)e\eta_z(x)}{1+\sigma'(z)e\eta_z(x)}\right)(1+\sigma(z)e\eta_z(x))dx \\
&\quad + \sum_{\sigma(z)\neq\sigma'(z)} \int_{C(z)} \log\left(\frac{1-\sigma(z)e\eta_z(x)}{1-\sigma'(z)e\eta_z(x)}\right)(1-\sigma(z)e\eta_z(x))dx
\end{aligned}$$

It's easy to see that above expression is invariant with respect to the sign of b. So, we assume $b > 0$.

$$\begin{aligned}
KL(p_\sigma||p_{\sigma'}) &= \sum_{\sigma(z)\neq\sigma'(z)} \int_{C(z)} \log\left(\frac{1+\sigma(z)e\eta_z(x)}{1-\sigma(z)e\eta_z(x)}\right)(1+\sigma(z)e\eta_z(x))dx \\
&\quad + \sum_{\sigma(z)\neq\sigma'(z)} \int_{C(z)} \log\left(\frac{1-\sigma(z)e\eta_z(x)}{1+\sigma(z)e\eta_z(x)}\right)(1-\sigma(z)e\eta_z(x))dx \\
&= \sum_{\sigma(z)\neq\sigma'(z)} \int_{C(z)} \log\left(\frac{1+\sigma(z)e\eta_z(x)}{1-\sigma(z)e\eta_z(x)}\right) 2\sigma(z)e\eta_z(x)dx \\
&= \rho(\sigma,\sigma') \int_{C(z)} \log\left(\frac{1+e\eta_z(x)}{1-e\eta_z(x)}\right) 2e\eta_z(x)dx \\
&\leq \frac{2}{3}\rho(\sigma,\sigma') \int_{C(z)} e^2\eta_z^2(x)dx, \quad \text{using 29,} \\
&= K(d,\alpha,\beta)e^2 r^{2\beta} r_1^d \rho(\sigma,\sigma') \\
&= K(d,\alpha,\beta)e^2 c_w^d r^{2\beta+d}\rho(\sigma,\sigma').
\end{aligned}$$

■

Lemma 31 *Let P, Q be two probability distributions defined on $\Omega \times \{0,1\}$ defined as*

$$P(X \in A, Y = y) = \int_A [\pi_P p_1(x)\mathbb{1}(y=1) + (1-\pi_P)p_0(x)\mathbb{1}(y=0)]dx$$

and

$$Q(X \in A, Y = y) = \int_A [\pi_Q q_1(x)\mathbb{1}(y=1) + (1-\pi_Q)q_0(x)\mathbb{1}(y=0)]dx$$

for any $y \in \{0,1\}$ and Borel subset A of Ω, where $0 \leq \pi_P, \pi_Q \leq 1$ and p_0, p_1, q_0, q_1 are probability densities defined on Ω. Let $U \sim Ber(\pi_P)$ and $V \sim Ber(\pi_Q)$. Then

$$KL(P||Q) = KL(U||V) + \pi_P KL(p_1||q_1) + (1-\pi_P)KL(p_0||q_0).$$

Proof Let (X, Y) be a generic pair following the distributions P and Q. Then the lemma directly follows from the decomposition

$$KL(P||Q) = KL(P_Y||Q_Y) + E_{P_Y}\big[KL(P_{X|Y}||Q_{X|Y})\big].$$

■

Lemma 32 *For $\sigma \in \{-1,1\}^{\mathcal{Z}_1}$ let Q_σ be the distribution defined in C.4. Then for any σ, Q_σ satisfies α-margin condition with constant C_α.*

Proof

We recall the conditional densities C.3

$$g_1^\sigma(x) = \begin{cases} 1 + \sigma(z)\sqrt{\epsilon_P}\eta_z(x) & x \in C(z),\ z \in \mathcal{Z}_1, \\ 1 - \sigma(z)\sqrt{\epsilon_P}\eta_{f(z)}(x) & x \in C(f(z)),\ z \in \mathcal{Z}_1, \\ 1 + \xi_z(x) & x \in C(z),\ z \in \mathcal{Z}_3, \\ 1 - \xi_z(x) & x \in C(z),\ z \in \mathcal{Z}_4, \end{cases}$$

and

$$g_0^\sigma(x) = \begin{cases} 1 - \sigma(z)\sqrt{\epsilon_P}\eta_z(x) & x \in C(z),\ z \in \mathcal{Z}_1, \\ 1 + \sigma(z)\sqrt{\epsilon_P}\eta_{f(z)}(x) & x \in C(f(z)),\ z \in \mathcal{Z}_1, \\ 1 - \xi_z(x) & x \in C(z),\ z \in \mathcal{Z}_3, \\ 1 + \xi_z(x) & x \in C(z),\ z \in \mathcal{Z}_4, \end{cases}$$

From

$$\eta_Q^\sigma(x) - \frac{1}{2} = \frac{\pi_Q g_1^\sigma(x) - (1-\pi_Q^\sigma)g_0^\sigma(x)}{2\pi_Q g_1^\sigma(x) + 2(1-\pi_Q^\sigma)g_0^\sigma(x)} = \frac{g_1^\sigma(x) - g_0^\sigma(x)}{2(g_1^\sigma(x) + g_0^\sigma(x))}$$

we get

$$\eta_Q^\sigma(x) - 1/2 = \begin{cases} \frac{(1+\sqrt{\epsilon_P})\sigma(z)\eta_z(x)}{2-(1-\sqrt{\epsilon_P})\sigma(z)\eta_z(x)} & x \in C(z),\ z \in \mathcal{Z}_1, \\ \frac{-(1+\sqrt{\epsilon_P})\sigma(z)\eta_{f(z)}(x)}{2+(1-\sqrt{\epsilon_P})\sigma(z)\eta_{f(z)}(x)} & x \in C(f(z)),\ z \in \mathcal{Z}_1, \\ \frac{1}{2}\xi_z(x) & x \in C(z),\ z \in \mathcal{Z}_3, \\ -\frac{1}{2}\xi_z(x) & x \in C(z),\ z \in \mathcal{Z}_4. \end{cases}$$

Note that

$$\left|\eta_Q^\sigma(x) - 1/2\right| \geq \begin{cases} \frac{1}{4}\eta_z(x) & x \in C(z),\ z \in \mathcal{Z}_1 \cup \mathcal{Z}_2, \\ \frac{1}{4}\xi_z(x) & x \in C(z),\ z \in \mathcal{Z}_0. \end{cases}$$

Rest of the proof is similar as in the proof of lemma 6.4. ■

Lemma 33 *Let Π_σ be the probability distribution as defined in equation C.5. For $\sigma, \sigma' \in \{-1,1\}^{\mathcal{Z}_1}$ we have*

$$KL(\Pi_\sigma || \Pi_{\sigma'}) \leq 2c_w^d K(d,\alpha,\beta) c_r^{2\beta+d} \rho(\sigma,\sigma').$$

Proof From lemmas 31 and 30 we get

$$\begin{aligned} KL(Q_\sigma || Q_{\sigma'}) &= \frac{1}{2} KL(g_1^\sigma || g_1^{\sigma'}) + \frac{1}{2} KL(g_0^\sigma || g_0^{\sigma'}) \\ &\leq \frac{1}{2}(\epsilon_P + 1) c_w^d K(d,\alpha,\beta) r^{2\beta+d} \rho(\sigma,\sigma') \\ &\leq c_w^d K(d,\alpha,\beta) r^{2\beta+d} \rho(\sigma,\sigma') \end{aligned}$$

and

$$\begin{aligned} KL(P_\sigma || P_{\sigma'}) &= (1-\epsilon_P) KL(g_1^\sigma || g_1^{\sigma'}) + \epsilon_P KL(g_0^\sigma || g_0^{\sigma'}) \\ &\leq (\epsilon_P(1-\epsilon_P) + \epsilon_P) c_w^d K(d,\alpha,\beta) r^{2\beta+d} \rho(\sigma,\sigma') \\ &\leq 2\epsilon_P c_w^d K(d,\alpha,\beta) r^{2\beta+d} \rho(\sigma,\sigma'). \end{aligned}$$

Hence

$$\begin{aligned} KL(\Pi_\sigma || \Pi_{\sigma'}) &= n_P KL(P_\sigma || P_{\sigma'}) + n_Q KL(Q_\sigma || Q_{\sigma'}) \\ &\leq 2(\epsilon_P n_P + n_Q) c_w^d K(d,\alpha,\beta) r^{2\beta+d} \rho(\sigma,\sigma') \\ &\leq 2c_w^d K(d,\alpha,\beta)(\epsilon_P n_P + n_Q)(\epsilon_P n_P + n_Q)^{-\frac{2\beta+d}{2\beta+d}} \rho(\sigma,\sigma') \\ &\leq 2c_w^d K(d,\alpha,\beta) c_r^{2\beta+d}(\epsilon_P n_P + n_Q)(\epsilon_P n_P + n_Q)^{-\frac{2\beta+d}{2\beta+d}} \rho(\sigma,\sigma') \\ &\leq 2c_w^d K(d,\alpha,\beta) c_r^{2\beta+d} \rho(\sigma,\sigma'). \end{aligned}$$

■

Lemma 34 *For any pair (β, C_β) of positive numbers there exists an $M > 0$ such that the function v_a (defined in equation (C.2)) is β smooth with constant C_β.*

Proof If $\beta < 1$ then

$$\begin{aligned}
&\left|(1-u_a)^{1/\alpha}(x) - \big((1-u_a)^{1/\alpha}\big)(y)\right| \\
&\leq \max_{y\geq 0} \frac{d}{dy}\big((1-u_a)^{1/\alpha}\big)(y)|x-x_0| \\
&\leq \frac{L}{a}|x-x_0|^{\beta}.
\end{aligned}$$

If $\beta \geq 1$ using Taylor's approximation theorem we get

$$\begin{aligned}
&\left|(1-u_a)(x) - (1-u_a)_{x_0}^{(\beta)}\right| \\
&\leq \max_{y\geq 0} \frac{(1-u_a)^{(\lfloor\beta\rfloor+1)}(y)}{(\lfloor\beta\rfloor+1)!}|x-x_0|^{(\lfloor\beta\rfloor+1)} \\
&\leq \frac{L}{a}|x-x_0|^{\beta}.
\end{aligned}$$

for some $L > 0$, if $|x - x_0| \leq 1$. Hence, for a suitable a, v_a is $(\beta, C_\beta, 1)$ smooth. ■