Robust Transfer Learning Against Unreliable Source Data

Jianqing Fan[*] Cheng Gao[*] Jason M. Klusowski[*]

Department of Operations Research and Financial Engineering

Princeton University

Abstract

This paper introduces a novel approach to transfer learning that addresses the challenge of an unreliable source distribution due to ambiguity in Bayes classifiers or signal strength between the target and source distribution. The general convergence theorems establish that the risk of an unreliable source distribution in transfer learning can be controlled by a new quantity called the "ambiguity level", which is a new measurement of the discrepancy between the target and source regression functions. Our proposed model-selected classifier, with a threshold balancing the performance of target and source data, is shown to be both efficient and robust, improving classification while avoiding negative transfer. Moreover, we demonstrate the effectiveness of our approach on nonparametric classification and logistic regression tasks, achieving optimal upper bounds up to logarithmic factors. Simulation studies are conducted to provide numerical evidence that further supports the effectiveness of our proposed classifier. We also provide down-to-earth approaches to bound the excess misclassification rate without the need for specialized knowledge in transfer learning.

1. Introduction

Previous experiences can offer valuable insights for learning new tasks. Human learners often transfer their existing knowledge gained from previous tasks to new and related ones. *Transfer learning* refers to data learning tasks where a portion of the training data is generated from a similar but non-identical distribution to the data distribution for which we seek to make inferences about. The objective is then to transfer knowledge from such source data to gain improvement of learning in the related target task. Such problems, where there is a divergence between the data-generating distributions, arise in a lot of real application problems, including computer vision (Li et al. 2020; Tzeng et al. 2017), natural language processing (Ruder et al. 2019; Wang and Zheng 2015), speech recognition (Huang et al. 2013), and genre classification (Choi et al. 2017). For recent survey papers on transfer learning, see Storkey (2008), Pan and Yang (2010), and Weiss et al. (2016). Also, similar problems have been studied by both statisticians and many other communities possibly under different banners, including label

[*] These authors contributed equally.

noise (Frenay and Verleysen 2014; Scott et al. 2013; Cannings et al. 2020b; Blanchard et al. 2021; Reeve and Kaban 2019; Scott and Zhang 2019), domain adaptation (Scott 2018; Ben-David et al. 2010a,b; Mansour et al. 2009a), multi-task learning (Caruana 1997; Maurer et al. 2016), or distributional robustness (Sinha et al. 2018; Christiansen et al. 2020).

We focus here on the transfer learning setting in the context of binary classification because it is not only essential in statistical learning and has been extensively investigated in diverse contexts, but also a framework that is particularly conducive to algorithms that seek to exploit relationships between the target and source distributions. For transfer learning theory of linear regression models, see Chen et al. (2013); Bastani (2020) under the setting of finite covariate dimensions or Gross and Tibshirani (2016); Ollier and Viallon (2017); Li et al. (2022) under the high-dimensional regime with lasso-based penalties. For transfer learning theory of generalized linear models (GLMs), see Tian and Feng (2022). For transfer learning theory of non-parametric regression models, see Cai and Pu (2022).

To set up the framework, suppose that a labeled data sample (relatively small in size, typically) is drawn from the Q, target distribution we wish to make statistical inferences about. Also, let P be the source distribution from which we transfer knowledge, with a labeled data sample collected additionally. The corresponding random pairs of Q and P distributions are denoted by (X, Y) and (X^P, Y^P) on $\mathbb{R}^d \times \{0, 1\}$ respectively. Let $\eta^Q, \eta^P : \mathbb{R}^d \to [0, 1]$ denote the target and source regression functions, i.e.

$$\eta^Q(x) = Q(Y = 1|X = x) \text{ and } \eta^P(x) = P(Y^P = 1|X^P = x).$$

This the key question is how much information or knowledge could be transferred from P to Q given observations from both distributions. Note that the Bayes classifier $f^*_Q(\cdot) = \mathbf{1}\{\eta^Q(\cdot) \geq \frac{1}{2}\}$ minimizes the misclassification rate $Q(Y \neq f(X))$ over all classifiers. Therefore, we define the excess risk of any empirical classifier $\hat{f}$ as

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

and the key question leads us to the task of constructing an empirical classifier that accelerates the convergence of excess risk to 0 in expectation by utilizing labeled data samples from both Q and P.

1.1. Related Literature

Recent research has aimed to bridge the gap between limited theoretical understanding and significant practical achievements in transfer classification learning. In order to set up the problem, it is necessary to make some assumptions about the similarity between the target and source distributions, which are both useful for theory and practical implementation. Various approaches have been proposed and explored in the literature to measure this similarity, including divergence bounds, covariate shift, and label shift. Some methods concentrate on test error bounds that rely on measures of discrepancy between Q and P, such as modified total-variation or R'enyi divergence, between the target and source distributions (Ben-David et al. 2010a,b; Mansour et al. 2009a,b; Germain et al. 2013; Cortes et al. 2019). This line of works produce distribution-free risk rates, primarily expressed in terms of n_P alone, but the obtained rates do not converge to zero with increasing sample size. In other words, these frameworks cannot yield an excess risk rate converging to zero faster whenever their proposed divergences are non-negligible. Nonetheless, consistent classification is proved achievable regardless of a non-negligible divergence even when $n_Q = 0$, provided that certain additional structures of the target and source distributions are met (Ben-David et al. 2010b).

Two common additional structures include covariate shift and label shift (or posterior shift). Covariate shift (Gretton et al. 2008; Quionero-Candela et al. 2009; Kpotufe and Martinet 2018) basically considers scenarios where the conditional distributions of the response given the covariate are identical across Q and P, i.e. $\eta^Q = \eta^P$. Label shift, on the other hand, assumes identical or similar target and source marginal distributions Q_X and P_X, but the conditional probabilities η^Q and η^P differ.

Most previous works in label shift could be divided into two branches. On one hand, some frameworks do not require identical Bayes classifiers, i.e. $(\eta^P - \frac{1}{2})(\eta^Q - \frac{1}{2}) \geq 0$, but impose very specific relations between η^Q and η^P, such as the literature on *label noise* (Frenay and Verleysen 2014; Scott et al. 2013; Cannings et al. 2020b; Blanchard et al. 2021; Reeve and Kaban 2019; Natarajan et al. 2018; Scott and Zhang 2019). For instance, a common assumption (Reeve and Kaban 2019; Natarajan et al. 2018) is that the $Y^P|X^P = x$ is equal to $Y|X = x$ up to a constant probability of label flipping, i.e. in our terminology

$$P(Y^P = 1|Y = 0, X^P = X = x) = \pi_0 \text{ and } P(Y^P = 0|Y = 1, X^P = X = x) = \pi_1$$

for some constants $\pi_0, \pi_1 \in (0, 1)$. Under this setting, $\eta^P = (1 - \pi_0 - \pi_1)\eta^Q + \pi_0$, which is linear in η^Q. Given knowledge of the specific form of η^P, it is feasible to modify learning algorithms to efficiently infer the target Bayes classifier. In another special type of label shift problem, Maity et al. (2022) assumes that $P_{X|Y} = Q_{X|Y}$, which provides convenience for estimating the joint distribution of (X, Y). However, all their proposed estimators are tailored to fit specific assumptions and may not be applicable to more general relations between η^Q and η^P. On the other hand, recent works such as Cai and Wei (2021) and Hanneke and Kpotufe (2019) have introduced more general label shift settings that impose relatively mild and general conditions on the relation between η^P and η^Q, in addition to assuming identical Bayes classifiers. For instance, Cai and Wei (2021) requires a lower-bounded signal strength of η^P relative to η^Q, besides the assumption of identical Bayes classifiers, i.e.,

$$|\eta^P - \frac{1}{2}| \geq C_\gamma |\eta^Q - \frac{1}{2}|^\gamma$$

for some positive γ and C_γ, and derives a faster risk convergence rate with the transfer exponent γ.

The aforementioned approaches have limitations in that they rely on specific and often untestable assumptions. More importantly, they may not be effective in situations where both ambiguities of the source data hold, i.e. there are no strong relations between η^Q and η^P, but discrepancies between the Bayes classifiers of the target and source domains still exist. The only work of which we are aware that allows such general ambiguous source data is Reeve et al. (2021). Reeve et al. (2021) assumes that η^P can be well approximated by a set of regression functions $g_1(\eta^Q), \cdots, g_{L^*}(\eta^Q)$ that are no less informative than a linear transform of η^Q.

1.2. Main Contribution

In contrast to existing works, this paper considers a scenario where the Bayes classifiers can arbitrarily differ without imposing further conditions on the source and target distributions.

As a crucial concept for capturing the relative information transferred from the source data, we introduce the definition of the signal strength

$$s(x) := \begin{cases} |\eta^P(x) - \frac{1}{2}|, & \operatorname{sgn}\left(\eta^Q(x) - \frac{1}{2}\right) \times \left(\eta^P(x) - \frac{1}{2}\right) \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

Our main assumption involves measuring the ambiguity level of the source data based on the point-wise signal strength:

$$\mathbb{E}_{(X,Y)\sim Q}\left[|\eta^Q(X) - \frac{1}{2}|\mathbf{1}\{s(X) \leq C_\gamma|\eta^Q(X) - \frac{1}{2}|^\gamma \leq C_\gamma z^\gamma\}\right] \leq \varepsilon(z).$$

for some $\gamma, C_\gamma > 0$ and any $z \in [0, \frac{1}{2}]$. This quantity denotes the inevitable risk from hard-to-classify points, which have η^Q values too close to $\frac{1}{2}$ and η^P showing weak signal relative to η^Q. The explicit form of $\varepsilon(z)$ is provided for some special examples in Section 2.3. If $s(x) \geq C_\gamma|\eta^Q(x) - \frac{1}{2}|^\gamma$ covers the entire Ω, the setting is simplified to that of Cai and Wei (2021) with a strong relative signal of η^P across the entire feature space Ω. This quantity denotes the inevitable risk from hard-to-classify points, which have η^Q values too close to $\frac{1}{2}$ and η^P showing weak signal relative to η^Q.

Given the working assumptions above, we propose a simple but effective classifier that can surprisingly adapt to any level of ambiguity in the signal conveyed by η^P measured by $\varepsilon(z)$, named "model-selected classifier":

$$\hat{f}_{MS}(x) = \begin{cases} \mathbf{1}\{\hat{\eta}^Q(x) \geq \frac{1}{2}\}, & \text{if } |\hat{\eta}^Q(x) - \frac{1}{2}| \geq \tau, \\ \hat{f}^P(x), & \text{otherwise.} \end{cases}$$

where η^Q is an estimate of η^Q obtained by the target data and $\hat{f}^P$ is a classifier obtained by the source data. To clarify our decision-making process, we rely on $\hat{\eta}^Q$ as the final prediction when it confidently deviates from $\frac{1}{2}$. Otherwise, we switch to the prediction made by the source data.

In Section 3, we present two related general convergence theorems to show this classifier, with a proper choice of $\tau > 0$, utilizes both the relative signal of η^P and the ambiguity level $\varepsilon(z)$ from the source data. Meanwhile, the target data involved in this classifier is noteworthy for its dual role: not only does it maintains an upper bound of the excess risk if the source data is unreliable, but it also helps to alleviate the risk caused by ambiguity in the source data.

We then apply our general convergence results to non-parametric classification and logistic regression as a special case of parametric classification.

Non-parametric Classification: Suppose that η^Q is β-smooth (Condition 1), margin assumption holds for η^Q with parameter α (Assumption 1), and strong density condition (Condition 2) holds for Q_X and P_X. Let Π^{NP} denote the set of all such distribution pairs (Q, P).

At this moment, we consider the scenario where η^P is sufficiently smooth. Suppose that Π^{NP}_S is the subset of Π^{NP} such that η^P is β_P-smooth with $\beta_P = \gamma\beta$, and we show that the minimax excess risk satisfies

$$\left(n_P^{-\frac{\beta(1+\alpha)}{2\gamma\beta+d}} + \varepsilon(cn_Q^{-\frac{\beta}{2\beta+d}})\right) \wedge n_Q^{-\frac{\beta(1+\alpha)}{2\beta+d}} \lesssim \inf_{\hat{f}} \sup_{(Q,P)\in\Pi^{NP}_S} \mathbb{E}\mathcal{E}_Q(\hat{f})$$
$$\lesssim \left(n_P^{-\frac{\beta(1+\alpha)}{2\gamma\beta+d}} + \varepsilon(c\log(n_Q \vee n_P)n_Q^{-\frac{\beta(1+\alpha)}{2\beta+d}})\right) \wedge \left(\log^{1+\alpha}(n_Q \vee n_P)n_Q^{-\frac{\beta(1+\alpha)}{2\beta+d}}\right).$$

for some constant $c > 0$.

The upper bound we obtained is optimal up to some logarithmic terms of $n_Q \vee n_P$. Our findings indicate that a necessary and sufficient condition for the source data to improve the excess risk rate is a large source data sample size such that $n_P \gg n_Q^{\frac{2\gamma\beta+d}{2\beta+d}}$ paired with a small ambiguity level $\varepsilon(z) \ll z^{1+\alpha}$. In Appendix A.2, we provide a detailed analysis of the general case of Π^{NP}, with a particular focus on the relationship between the smoothness parameters β_P and $\gamma\beta$.

Additionally, we provide the optimal rate for band-like ambiguity, which can be a reasonable potential condition between η^Q and η^P. A particularly interesting and realistic example is to assume that $\sup_{x\in\Omega} |\eta^P(x) - \eta^Q(x)| \le \Delta$, where the minimax optimal excess risk becomes $\left(n_P^{-\frac{\beta(1+\alpha)}{2\beta+d}} + \Delta^{1+\alpha}\right) \wedge n_Q^{-\frac{\beta(1+\alpha)}{2\beta+d}}$ up to some logarithmic terms of $n_Q \vee n_P$. Note that we do not require any smoothness condition on η^P in the band-like ambiguity case.

Logistic Regression: If the target and source logistic regression coefficient pair (β_Q, β_P) belongs to

$$\Theta(s, \Delta) = \{(\beta_Q, \beta_P) : ||\beta_Q||_0 \le s, \langle \beta_Q, \beta_P \rangle \le \Delta\}$$

for some $0 \le \Delta \le \frac{\pi}{2}$, with the standard normal distribution random design of covariates $X, X^P \sim N(0, I_d)$ (See (23) for a precise definition), the minimax optimal excess risk satisfies

$$\left(\frac{s\log d}{n_P} + \Delta^2\right) \wedge \frac{s\log d}{n_Q} \lesssim \inf_{\hat{f}} \sup_{(Q,P)\in\Pi^{LR}} \mathbb{E}\mathcal{E}_Q(\hat{f}) \lesssim \left(\frac{s\log d}{n_P} + \Delta^2\right) \wedge \left(\frac{s\log d}{n_Q} \log^2(n_Q \vee n_P)\right).$$

We see that when $n_P \gg n_Q$ and $\Delta \ll \frac{s\log d}{n_Q}$, the source data help to obtain a faster convergence rate of the excess risk. We claim that our setting assumes a more general "small cone condition" between β_Q and β_P compared with the "small contrast condition" in previous works (Li et al. 2022; Tian and Feng 2022), i.e. $||\beta_Q - \beta_P||_q$ is small for some $q \in [0, 1]$. In addition, unlike the small contrast condition which implicitly assumes sparsity patterns of β_P through l_q norms with $q \le 1$, our constructed parametric space does not impose any sparsity conditions on β_P. In the setting without access to source data, our upper bound $\frac{s\log d}{n_Q}$ obtained in this paper up to logarithmic factors is tighter than the $(\frac{s\log d}{n_Q})^{\frac{2}{3}}$ bound obtained in Theorem 7 of Abramovich and Grinshtein (2019), assuming the same margin parameter $\alpha = 1$. As evidenced by the above two examples, our proposed classifier could maintain the performance of the target data against an unreliable source and enhance the performance with a large source data sample size and relatively small ambiguity, i.e. $n_P \gg n_Q$ and $\Delta \ll \sqrt{\frac{s\log d}{n_Q}}$.

While the target excess risk provided by the target data, $\mathcal{E}_Q(\hat{f}^Q)$, has been extensively studied in the literature, there is a lack of understanding regarding the target excess risk $\mathcal{E}_Q(\hat{f}^P)$ provided by the source data, which is nonetheless one key component in deducing the excess risk upper bound of our proposed classifier. To fill this gap and gain better insight into the contribution of the source data to the overall performance, we present a general result provides a direct bound on $\mathcal{E}_Q(\hat{f}^P)$ in terms of $\mathcal{E}_P(\hat{f}^P)$, which can be obtained through conventional theoretical analysis. Therefore, one can easily obtain an upper bound on the excess risk as long as they have knowledge of some conventional statistical learning, without the need for specialized knowledge in transfer learning. By providing such a general and accessible framework for bounding excess risk, our approach has the potential to make transfer learning more accessible and widely applicable in practice.

1.3. Notation and Organization

We introduce some notations to be used throughout the paper. For any $q \in [0, \infty]$ and a vector $x = (x_1, \cdots, x_d) \in \mathbb{R}^d$, we write $||x||_q$ as its l_q norm. We write $||x|| = ||x||_2$ for the Euclidean norm of x, and, given $r > 0$, we write $B(x, r)$ or $B_d(x, r)$ for the closed Euclidean sphere of radius r centered at x. For two probability measures μ, ν on any general space, if μ is absolutely continuous with respect to ν, we write $\frac{d\mu}{d\nu}$ as the Radon-Nikodyn derivative of μ with respect to ν. Also, let $H(\mu, \nu)$ and $TV(\mu, \nu)$ denote their Hellinger distance and total variation distance, respectively. Write λ as the Lebesgue measure on $\mathbb{R}^d$. Define the topology boundary point of a set A in $\mathbb{R}^d$ as ∂A. Let $a \wedge b$ and $a \vee b$ denote the minimum and maximum of a and b, respectively. Let $a_{n_Q} \lesssim b_{n_Q}$ denote $|a_{n_Q}| \leq c|b_{n_Q}|$ for some constant $c > 0$ when n_Q is large enough. Let $a_{n_Q} \gtrsim b_{n_Q}$ denote $|a_{n_Q}| \geq c|b_{n_Q}|$ for some constant $c > 0$ when n_Q is large enough. Let $a_{n_Q} \asymp b_{n_Q}$ denote $|a_{n_Q}|/|b_{n_Q}| \to c$ for some constant $c > 0$ when n_Q is large enough. Let $a_{n_Q} \overset{n_Q \to \infty}{\longrightarrow} \infty$ denote that a_{n_Q} tends to infinity with n_Q growing to infinity. Let $a_{n_Q} \ll b_{n_Q}$ denote $|a_{n_Q}|/|b_{n_Q}| \to 0$ when n_Q is large enough. Let $a_{n_Q} \gg b_{n_Q}$ denote $|b_{n_Q}|/|a_{n_Q}| \to 0$ when n_Q is large enough. Let $\lfloor a \rfloor$ be the maximum integer that is less equal than a for any real value a. Finally, we assume $0^0 = 0$ for simplicity.

We state our main working assumptions and measurement of ambiguity called "ambiguity risk bound" in Section 2, and provide two general convergence results in Section 3 regarding this measurement. We explicitly formulate the ambiguity risk bound under some additional conditions in Section 2.3. In Sections 4 and 5, we apply our results to non-parametric classification and logistic regression, respectively, with excess risk upper and lower bounds given. In Section 7, we present an approach to bound the signal transfer risk, a crucial part of general convergence results, in terms of the excess risk rate studied in the conventional statistical learning literature.

2. Model

2.1. Problem Formulation

For two Borel-measurable distributions P and Q both taking values in $\mathbb{R}^d \times \{0, 1\}$, we observe two independent random samples, the *source* data $\mathcal{D}_P = \{(X_1^P, Y_1^P), \cdots, (X_{n_P}^P, Y_{n_P}^P)\} \overset{\text{iid}}{\sim} P$ and the *target* data $\mathcal{D}_Q = \{(X_1, Y_1), \cdots, (X_{n_Q}, Y_{n_Q})\} \overset{\text{iid}}{\sim} Q$. Suppose that $n_P \overset{n_Q \to \infty}{\longrightarrow} \infty$. Our goal is to improve the target data empirical classifier by transferring valuable information from the source data. Let P_X and Q_X denote the support set of the marginal probability distributions of X for the P and Q distributions, respectively, where Ω_P and Ω are subsets of $\mathbb{R}^d$. The regression function for the source and target distributions are denoted by η^P and η^Q, respectively, and are defined as follows:

$$\eta^P(x) = P(Y^P = 1 | X^P = x), \quad \eta^Q(x) = Q(Y = 1 | X = x) \tag{1}$$

The goal of a classification model is to forecast the label Y based on the value X. The effectiveness of a decision rule $f : \mathbb{R}^d \to \{0, 1\}$ is evaluated by its misclassification rate with respect to the target distribution, which is defined as follows:

$$R(f) := Q(Y \neq f(X)).$$

The Bayes classifier (or Bayes estimator, Bayes decision rule) $f_Q^*(x) = \mathbf{1}(\eta^Q(x) \geq \frac{1}{2})$ is the minimizer of $R(f)$ over all Borel functions defined on $\mathbb{R}^d$ taking values in $\{0,1\}$. We similarly define the Bayes decision rule for η^P that is $f_P^*(x) = \mathbf{1}(\eta^P(x) \geq \frac{1}{2})$. Since the Bayes decision rule $f_Q^*(x)$ is a minimizer of the misclassification rate $R(f)$, the performance of any empirical classifier $\hat{f} : \mathbb{R}^d \to \{0,1\}$ can be then measured by the excess risk (on the target distribution):

$$\mathcal{E}_Q(\hat{f}) = R(\hat{f}) - R(f_Q^*) = 2\mathbb{E}_{(X,Y)\sim Q}[|\eta_Q(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}(X) \neq f_Q^*(X)\}]. \tag{2}$$

The last equality is the dual representation of the excess risk (Gyorfi 1978). Given the excess risk defined in (2), our objective of transferring useful information from the source data can be reformulated as the task of constructing an empirical classifier that speeds up the convergence rate of the excess risk to zero, which is accomplished by utilizing labeled data samples drawn from both the target and source distributions.

The rate at which the excess risk $\mathcal{E}_Q(\hat{f})$ converges to zero depends on the assumptions made about the target and source distributions (Q, P). In classification, a common assumption is the margin assumption (Audibert and Tsybakov 2007; Mammen and Tsybakov 2004). This assumption is used to measure the behavior of Q_X with respect to the distance between $\eta^Q(X)$ and $\frac{1}{2}$, which is essential for determining the rate of the excess risk. This is because the closer $\eta^Q(x)$ is to $\frac{1}{2}$, the more difficult it becomes to correctly classify the query point x. Note that we exclude the decision boundary set $\{x \in \Omega : \eta^Q(x) = \frac{1}{2}\}$, since it is meaningless to consider points outside the support set Ω and any classifier, performs no better than a random guess for query points on the decision boundary.

Assumption 1 (Margin)**.** There exists some constant $\alpha \geq 0,\ C_\alpha > 0$ such that for any $t > 0$, we have $Q_X(0 < |\eta^Q(X) - \frac{1}{2}| \leq t) \leq C_\alpha t^\alpha$.

Instead of assuming identical Bayes classifiers for η^Q and η^P over $x \in \Omega$, as some existing literature does (Hanneke and Kpotufe 2019; Cai and Wei 2021), we allow them to differ. This is a reasonable relaxation because although the target and source distributions may have a strong correlation and share the same Bayes classifier in a high probability majority area, they could still have slightly different decision boundaries. Therefore, it is crucial to assess the impact of an unreliable source distribution on the optimal rates of classification.

2.2. Source Data Ambiguity

In this subsection, we provide a detailed discussion of the condition that characterizes the ambiguity of an unreliable source distribution. We introduce the concept of the signal strength function, which measures the relative signal of η^P compared with η^Q. This concept is crucial in capturing the efficacy of the source data for the classification task under Q.

Definition 1 (Signal Strength)**.** The signal strength of η^P relative to η^Q is defined as

$$\begin{aligned} s(x) :&= \left\{ \operatorname{sgn}\left(\eta^Q(x) - \frac{1}{2}\right) \times (\eta^P(x) - \frac{1}{2}) \right\} \vee 0 \\ &= \begin{cases} |\eta^P(x) - \frac{1}{2}|, & \operatorname{sgn}\left(\eta^Q(x) - \frac{1}{2}\right) \times (\eta^P(x) - \frac{1}{2}) \geq 0, \\ 0, & \text{otherwise}. \end{cases} \end{aligned}$$

for any $x \in \Omega$.

It is reasonable to consider the signal strength as non-zero only when $(\eta^P(x) - \frac{1}{2})(\eta^Q(x) - \frac{1}{2}) > 0$, indicating that the target and source data provide consistent information about the Bayes classifier. In this case, the signal strength is measured by the distance between $\eta^P(x)$ and $\frac{1}{2}$, and the source data $\eta^P(x)$ is beneficial for classifying the target data. Oppositely, when $(\eta^P(x) - \frac{1}{2})(\eta^Q(x) - \frac{1}{2}) \leq 0$, the source data does not provide useful information for classifying x in the target data, and the signal strength at x is zero.

Next, we present the main working assumption that measures the ambiguity level of the source data based on the point-wise signal strength.

Assumption 2 (Ambiguity Level)**.** There exists some constants $\gamma, C_\gamma > 0$, and a continuous function $\varepsilon(z; \gamma, C_\gamma)$ that is monotone increasing with $z \in [0, \frac{1}{2}]$ such that

$$\mathbb{E}_{(X,Y)\sim Q}\left[|\eta^Q(X) - \frac{1}{2}|\mathbf{1}\{s(X) \leq C_\gamma|\eta^Q(X) - \frac{1}{2}|^\gamma, |\eta^Q(X) - \frac{1}{2}| \leq z\}\right] \leq \varepsilon(z; \gamma, C_\gamma). \tag{3}$$

We abbreviate $\varepsilon(z; \gamma, C_\gamma)$ as $\varepsilon(z)$ when there is no need to specify γ and C_γ.

The expression in the expectation operator of (3) can be divided into two terms. The first term represents the distance between $\eta^Q(X)$ and $\frac{1}{2}$, which corresponds to the dual representation of the excess risk (2).

The second indicator term is crucial for understanding transfer learning ambiguity. On one hand, the constraint $s(X) \leq C_\gamma|\eta^Q(X) - \frac{1}{2}|^\gamma$ indicates a lack of strong signal from the source data relative to the target data. On the other hand, the constraint $|\eta^Q(X) - \frac{1}{2}| \leq z$, indicates the hard-to-classify region by the target distribution, as the data points are too close to the decision boundary. Therefore, our indicator precisely captures the challenging region where classification becomes difficult using either the target or source data.

The ambiguity level $\varepsilon(\cdot)$ allows for the presence of unreliable sources, as it accounts for situations where η^P may not consistently provide a strong signal compared to η^Q. Additionally, $\varepsilon(\cdot)$ controls the ambiguity behavior with respect to the distance from $\frac{1}{2}$. The larger the value of the ambiguity risk is, the harder the classification task. Note that by the margin assumption, a trivial setting of the ambiguity level would be $\varepsilon(z) = C_\alpha z^{1+\alpha}$.

Remark 1. The region of different Bayes classifiers, i.e. $s(x) = 0$, determines the necessary lower bound of the ambiguity bound. Points in this set are hard to classify by the source distribution, as the signal of the source data provides no useful information for the target data. Therefore, a necessary lower bound of $\varepsilon(z)$ is

$$\varepsilon(z) \geq \mathbb{E}_{(X,Y)\sim Q}\left[|\eta^Q(X) - \frac{1}{2}|\mathbf{1}\{s(X) = 0, |\eta^Q(X) - \frac{1}{2}| \leq z\}\right].$$

2.3. On the Ambiguity Level

To clarify the novel idea of the ambiguity level assumption, we provide some explicit formulas for the ambiguity level $\varepsilon(\cdot)$ with a proper choice of γ and C_γ. We list these separate examples separately below. We leave the case of two logistic regression models across the target and source distributions in Section 5.

Example 1 (Perfect Source)**.** Assume the condition of "relative signal exponent" proposed in Cai and Wei (2021), which refers to

$$s(x) \geq C_\gamma |\eta^Q(x) - \frac{1}{2}|^\gamma \quad \forall x \in \Omega.$$

Then Assumption 2 holds with $\varepsilon(z) \equiv 0$.

Specially, if $\eta^P = \eta^Q$, it is straightforward to set $\gamma = 1$ and $\varepsilon(\cdot) = 0$. In this scenario, the Bayes classifiers are identical and η^P gives strong signal compared to η^Q over the whole support Ω, so the ambiguity level is set as zero.

Example 2 (Limited Source Support)**.** Assume a slightly different scenario than Example 1 where the signal strength is strong over the entire source feature space. If

$$s(x) \geq C_\gamma |\eta^Q(x) - \frac{1}{2}|^\gamma \quad \forall x \in \Omega_P,$$

then Assumption 2 holds with

$$\varepsilon(z) = \int_{\Omega/\Omega_P} |\eta^Q(X) - \frac{1}{2}| \mathbf{1}\{0 < |\eta^Q(X) - \frac{1}{2}| < z\} dQ_X,$$

i.e. the ambiguity level is fully depicted by the risk within the complement Ω/Ω_P. Specially, if $\eta^P(x) = \eta^Q(x)$ for any $x \in \Omega_P$, then Assumption 2 holds with $\varepsilon(z) \equiv 0$. This corresponds to the common case where the source data is collected from a subpopulation of one of the target data.

Example 3 (Ambiguity Margin)**.** Suppose that we could control the behavior of the signal ambiguity set with respect to the distance from $\frac{1}{2}$ in a similar way to Assumption 1. There exists some constant $\alpha' \geq 0, C_{\alpha'} > 0$ such that for any $t > 0$, we have

$$Q_X(s(X) \leq C_\gamma |\eta^Q(X) - \frac{1}{2}|^\gamma, 0 < |\eta^Q(X) - \frac{1}{2}| < t) \leq C_{\alpha'} t^{\alpha'}. \tag{4}$$

Then this example can be viewed as Assumption 2 with $\varepsilon(z) = C_{\alpha'} z^{1+\alpha'}$.

We further discuss two special cases under Example 3. In the case where $\alpha' = 0$, (4) reduces to $Q_X(s(X) \leq C_\gamma |\eta^Q(X) - \frac{1}{2}|^\gamma) \leq C_{\alpha'}$, and the ambiguity level is well-defined as

$$\varepsilon(z) = C_{\alpha'} \epsilon.$$

This case corresponds to situations where the only available information about the source ambiguity is the upper bound on the probability of weak signal strength. On the other hand, if $\alpha' \geq \alpha$, then the ambiguity level becomes

$$\varepsilon(z) = C_{\alpha'} z^{1+\alpha'},$$

which is indeed asymptotically less than $C_\alpha z^{1+\alpha}$ when $z \to 0$. This ensures that the ambiguity level is strictly dominated by the one provided by the margin assumption, given that we treat C_α and $C_{\alpha'}$ as fixed constants.

Example 4 (Band-like Ambiguity)**.** A further noteworthy scenario is when the probability distribution η^P concentrates around a "band" that is centered on an informative curve with respect to η^Q, but with some small deviation. A related situation is studied in Reeve et al. (2021), where η^P is approximated by a linear transfer function of η^Q. Suppose that there exists some band error constant $\Delta \geq 0$, which represents the deviation level, such that

$$s(x) \geq C_\gamma|\eta^Q(x) - \frac{1}{2}|^\gamma - \Delta \tag{5}$$

for any $x \in \Omega$. Then Assumption 2 holds with

$$\varepsilon(z; \gamma, C_\gamma/2) = \left(C_\alpha z^{1+\alpha}\right) \wedge \left(2^{\frac{1+\alpha}{\gamma}} C_\alpha C_\gamma^{-\frac{1+\alpha}{\gamma}} \Delta^{\frac{1+\alpha}{\gamma}}\right) \tag{6}$$

Notably, the case of $\Delta = 0$ degenerates into the perfect source scenario in Example 1. For proof of the statement in Example 4, see Lemma E.1. For simplicity, we assume that (5) holds on Ω. To extend the setting that (5) holds only on a subset $\Omega_{BA} \subset \Omega$, one could add the risk term $\int_{\Omega\setminus\Omega_{BA}} |\eta^Q(X) - \frac{1}{2}|\mathbf{1}\{|\eta^Q(X) - \frac{1}{2}| \leq z\}dQ_X$ to the existing risk term part $2^{\frac{1+\alpha}{\gamma}} C_\alpha C_\gamma^{-\frac{1+\alpha}{\gamma}} \Delta^{\frac{1+\alpha}{\gamma}}$. This additional term accounts for the risk due to points in $\Omega_P \setminus \Omega$, which are not covered by the existing term.

Specially, if we have

$$\sup_{x\in\Omega} |\eta^P(x) - \eta^Q(x)| \leq \Delta, \tag{7}$$

then the band-like ambiguity condition 5 holds with $\gamma = C_\gamma = 1$. The condition (7) is common and meaningful in real-world applications where the regression function of the source distribution deviates slightly from η^Q. By (6) , Assumption 2 holds in this special case with

$$\varepsilon(z, 1, \frac{1}{2}) = \left(C_\alpha z^{1+\alpha}\right) \wedge \left(2^{1+\alpha} C_\alpha C_\gamma^{-(1+\alpha)} \Delta^{1+\alpha}\right).$$

Example 5 (Conditional Ambiguity Probability Bound)**.** We could also directly bound the ambiguity level given information of the joint distribution (η^Q, η^P). This approach can be useful when there is prior knowledge about the (random) behavior of η^P given η^Q. Suppose there exists a continuous function $p(\cdot; \gamma, C_\gamma) : [0, 1] \to [0, 1]$ such that for any $\eta \in [0, 1]$,

$$Q(\text{sgn}\left(\eta - \frac{1}{2}\right) \times (\eta^P - \frac{1}{2}) \geq C_\gamma|\eta - \frac{1}{2}|^\gamma | \eta^Q = \eta) \leq p(\eta; \gamma, C_\gamma).$$

Then Assumption 2 holds with

$$\begin{aligned}\varepsilon(z) &= \int_{\frac{1}{2}-z}^{\frac{1}{2}+z} |\eta - \frac{1}{2}|p(\eta; \gamma, C_\gamma)dF_\eta^Q(\eta)\\ &\leq C_\alpha z^{1+\alpha} \times \sup_{\eta\in[\frac{1}{2}-z,\frac{1}{2}+z]} p(\eta; \gamma, C_\gamma)\end{aligned}$$

where F_η^Q is defined as the cumulative distribution function of η^Q w.r.t. Q_X. See Lemma E.2 for the proof.

3. General Convergence Results

Let $\Pi = \Pi(\alpha, C_\alpha, \gamma, C_b, \varepsilon)$ be the set of distributions (Q, P) satisfying that Assumption 1 and 2 hold with parameter $\alpha \geq 0, \gamma, C_\alpha, C_\gamma > 0$ and the ambiguity level $\varepsilon(\cdot)$. We may add more conditions on the joint distribution of both target and source data. Moving forward, our analysis focuses only on the performance of any classifier when the target and source distribution pair (Q, P) belongs to (potentially a subset of) Π. This framework captures the essential information regarding how the source data can accelerate the convergence rate of the excess risk at most.

Suppose $\mathbb{E}_{\mathcal{D}_Q}, \mathbb{E}_{\mathcal{D}_P}, \mathbb{E}_{(\mathcal{D}_Q,\mathcal{D}_P)}$ is the expectation operator taken with respect to the target data $\mathcal{D}_Q$, the source data $\mathcal{D}_P$, and both data $\mathcal{D}_Q$ and $\mathcal{D}_P$, respectively. Define the probability distributions as $\mathbb{P}_{\mathcal{D}_Q}, \mathbb{P}_{\mathcal{D}_P}$ and $\mathbb{P}_{(\mathcal{D}_Q,\mathcal{D}_P)}$ similarly. Plus, we call any quantity that depends on the variable n_Q or n_P except some constant as n_Q-sequence or n_P-sequence, respectively.

In preparation for the general result, it is necessary to introduce one more definition of the risk that can be learned with respect to the source data over the area with strong strength, i.e. $\{x \in \Omega : s(x) \geq C_\gamma |\eta^Q(x) - \frac{1}{2}|^\gamma\}$.

Definition 2 (Signal Transfer Risk)**.** Define the *signal transfer risk* of the classifier f with respect to parameters $\gamma, C_\gamma > 0$ as

$$\varepsilon_T(f; \gamma, C_\gamma) := \mathbb{E}_{(X,Y)\sim Q}[|\eta^Q(X) - \frac{1}{2}| \mathbf{1}\{\hat{f}^P(X) \neq f^*_Q(X), s(X) \geq C_\gamma |\eta^Q(X) - \frac{1}{2}|^\gamma\}]. \tag{8}$$

We abbreviate $\varepsilon_T(f; \gamma, C_\gamma)$ as $\varepsilon_T(f)$ when there is no need to specify γ and C_γ.

The signal transfer risk serves as the risk resulting from the classification of points belonging to the area where $s(x) \geq C_\gamma |\eta^Q(x) - \frac{1}{2}|^\gamma$. Due to the strong signals offered by the source data in relation to the target data within this area, it is expected that the signal transfer risk can be accelerated with the aid of the source data sample $\mathcal{D}_P$.

In this paper, the classifier derived from the target data is assumed to be a *plug-in rule*, of the form $\mathbf{1}\{\hat{\eta}^Q(x) \geq \frac{1}{2}\}$, where $\hat{\eta}^Q$ is an estimator of the regression function η^Q. By introducing a novel strategy called the *model-selected classifier*, the following result demonstrates a general approach to achieving a faster convergence rate of the excess risk, contingent on the exponential concentration inequality of the estimate of η^Q.

Theorem 1. Let $\hat{\eta}^Q$ be an estimate of the regression functions η^Q and $\hat{f}^P$ be a classifier obtained by $\mathcal{D}_P$. Suppose there exists $\tilde{\Pi} \subset \Pi$ and two n_Q-sequences δ_Q, δ_Q^f such that,

- With probability at least $1 - \delta_Q^f$ w.r.t. the distribution of $X_{1:n_Q} := (X_1, \cdots, X_{n_Q})$, for any $x \in \Omega$ we have $\forall t > 0$,

$$\sup_{(Q,P)\in\tilde{\Pi}} \mathbb{P}_{\mathcal{D}_Q}(|\hat{\eta}^Q(x) - \eta^Q(x)| \geq t | X_{1:n_Q}) \leq C_1 \exp(-(\frac{t}{\delta_Q})^2). \tag{9}$$

  for some constant $C_1 > 0$.

- $\delta_Q^{1+\alpha} \gtrsim n_Q^{-c}$, $\sup_{(Q,P)\in\tilde{\Pi}} \mathbb{E}_{\mathcal{D}_P} \varepsilon_T(\hat{f}^P) \gtrsim n_P^{-c}$ for some constant $c > 0$, Given the choice of $\tau \gtrsim \log(n_Q \vee n_P)\delta_Q$, the *model-selected classifier*

$$\hat{f}_{MS}(x) = \begin{cases} \mathbf{1}\{\hat{\eta}^Q(x) \geq \frac{1}{2}\}, & \text{if } |\hat{\eta}^Q(x) - \frac{1}{2}| \geq \tau, \\ \hat{f}^P(x), & \text{otherwise.} \end{cases} \tag{10}$$

satisfies that

$$\sup_{(Q,P)\in\tilde{\Pi}} \mathbb{E}_{(\mathcal{D}_Q,\mathcal{D}_P)}\mathcal{E}_Q(\hat{f}_{MS}) \lesssim \left(\sup_{(Q,P)\in\tilde{\Pi}} \mathbb{E}_{\mathcal{D}_P}\varepsilon_T(\hat{f}^P) + \varepsilon(2\tau) \right) \wedge \tau^{1+\alpha} + \delta_Q^f. \tag{11}$$

To clarify the decision-making process of our proposed *model-selected* classifier, we keep the prediction made by $\hat{\eta}^Q$ as the final prediction if its estimated value is at least τ far away from $\frac{1}{2}$, indicating high confidence in its accuracy. Conversely, if the estimate $\hat{\eta}^Q$ is close to $\frac{1}{2}$, which implies difficulty in correct classification by the target data, we switch to the prediction made by the source data instead. Now, we would like to illustrate every important term in the upper bound (11).

- $\sup_{(Q,P)\in\tilde{\Pi}} \mathbb{E}_{\mathcal{D}_P}\varepsilon_T(\hat{f}^P)$ represents the risk transferred by the source data with a classifier $\hat{f}^P$ excluding the ambiguity part, typically with a faster convergence rate than $\delta_Q^{1+\alpha}$. It quantifies the benefits obtained from transfer learning.

- The ambiguity risk bound term $\varepsilon(2\tau)$ measures the ambiguity level, or the strength of bias, when η^Q is close to $\frac{1}{2}$. It is worth noting that correct classification within this area is technically given up. Plus, this bias term does not relate to the part of the signal ambiguity set where $|\eta^Q(x) - \frac{1}{2}|$ is large, which means the data points that are easy to classify w.r.t. the target data will not negatively affect the performance of our proposed estimator. In other words, the target data not only serves to conserve an upper bound of the excess risk but also plays a critical role in reducing the risk caused by the ambiguity in the source data.

- The threshold parameter τ in our approach serves to balance the classification ability of the target and source data. Specifically, it filters out the points that are easy to classify using the target data, by requiring that the distance between the estimator $\hat{\eta}^Q$ and $\frac{1}{2}$ is greater than τ while allowing the classifier $\hat{f}^P$ generated by the source data to handle the difficult-to-classify points that cannot be accurately classified using the target data. When $\tau = 0$, the approach corresponds to the setting of Audibert and Tsybakov (2007), and the excess risk is then asymptotically lower than $\delta_Q^{1+\alpha}$ without any source data. On the other hand, when $\tau = 1$, only the source data is used to construct the classifier, which is typically used in the domain adaptation literature when the target data is unavailable. We assert that our selection of τ between the two extreme values allows us to leverage the classification ability of both the target and source data, as it takes the minimum of the risks achieved by each of them.

- The final term in the upper bound, δ_Q^f, represents the probability of the concentration inequality failing due to extreme realizations of the observations in $\mathcal{D}_Q$. Usually it is rather small compared to the main convergence rate. It usually decays exponentially w.r.t. n_Q in non-parametric classification using K-NN estimators (See Lemma 9.1 of Cai and Wei (2021)). We will see that under mild conditions regarding the sample sizes n_Q and n_P, this term is dominated by other terms in the upper bound for non-parametric classification and logistic regression models.

The introduced assumption of $\delta_Q^{1+\alpha} \gtrsim n_Q^{-c}$ and $\sup_{(Q,P)\in\tilde{\Pi}} \mathbb{E}_{\mathcal{D}_P}\varepsilon_T(\hat{f}^P) \gtrsim n_P^{-c}$ are mild and serve to prevent overly rapid convergence rates, such as those that decay exponentially with the sample size. With the optimal choice of $\tau \asymp \log(n_Q \vee n_P)\delta_Q$, our obtained upper bound cannot be worse than the $\delta_Q^{1+\alpha}$ in Audibert and Tsybakov (2007) up to some logarithmic terms.

Sometimes the concentration property may not be in the form of an exponential term as (9). To overcome this limitation, the following theorem generalizes Theorem 1 to allow any type of concentration property of $\hat{\eta}^Q$.

Theorem 2. Let $\hat{\eta}^Q$ be an estimate of the regression functions η^Q and $\hat{f}^P$ be a classifier obtained by $\mathcal{D}_P$. Suppose that for some $\tau > 0, \tilde{\Pi} \subset \Pi$, there exists a function $\delta_Q(\cdot,\cdot)$ such that for any $(Q,P) \in \tilde{\Pi}$, the concentration property

$$\mathbb{P}_{\mathcal{D}_Q}(|\hat{\eta}^Q(x) - \eta^Q(x)| \geq \tau) \leq \delta_Q(n_Q, \tau) \tag{12}$$

holds for any $x \in \Omega^* \subset \Omega$ with $Q(\Omega^*) \geq 1 - \delta(n_Q, \tau)$. Then the *model-selected classifier*

$$\hat{f}_{MS}(x) = \begin{cases} \mathbf{1}\{\hat{\eta}^Q(x) \geq \frac{1}{2}\}, & \text{if } |\hat{\eta}^Q(x) - \frac{1}{2}| \geq \tau, \\ \hat{f}^P(x), & \text{otherwise.} \end{cases} \tag{13}$$

satisfies that

$$\sup_{(Q,P)\in\tilde{\Pi}} \mathbb{E}_{(\mathcal{D}_Q,\mathcal{D}_P)}\mathcal{E}_Q(\hat{f}_{MS}) \lesssim \left(\sup_{(Q,P)\in\tilde{\Pi}} \mathbb{E}_{\mathcal{D}_P}\varepsilon_T(\hat{f}^P) + \varepsilon(2\tau) \right) \wedge \tau^{1+\alpha} + \delta_Q(n_Q,\tau). \tag{14}$$

Theorem 2 is a generalized and more abstract version of Theorem 1. To reduce Theorem 2 to Theorem 1, it suffices to set $\delta_Q(n_Q,\tau) = \delta_Q^f + C_1\exp(-(\frac{\tau}{\delta_Q})^2)$ and $\Omega^* \equiv \Omega$. where δ_Q^f and C_1 are the notations used in Theorem 1.

Plus, readers may be concerned that the exponential concentration in (9) may not hold for all query points over Ω. This is particularly true when the support Ω is not compact, and the concentration inequality may only hold within a major part of Ω with high probability w.r.t. Q_X. Hence, we add the small probability of failure of the exponential concentration, i.e. $1 - Q_X(\Omega^*) \leq \delta(n_Q,\tau)$, to the risk bound given in (14), with only minor modifications to the proof. Note that though the probability of failure $\delta(n_Q,\tau)$ is uniformly bounded and does not depend on the realization of the target data, Ω^* may depend on $\mathcal{D}_Q$.

4. Applications in Non-parametric Classification

In this section, we aim to apply the general result of Theorem 1 to non-parametric classification settings. We specifically design the model-selected classifier by combining the plug-in rules over the target and source data, and obtain the minimax optimal rate under the non-parametric settings. See Audibert and Tsybakov (2007) for a comprehensive overview of theoretical properties of the plug-in rules.

We adopt K-nearest neighbor classifiers as plug-in rules for both $\hat{\eta}^Q$ and $\hat{\eta}^P$. Our analysis in this section then builds on prior work on error rates in K-nearest neighbor classification (e.g. Hall et al. (2008); Samworth (2012); Gadat et al. (2016); Celisse and Mary-Huard (2018); Cannings et al. (2020a)) with both practical and theoretical successes. For a review of early works on the theoretical properties of the K-NN classifier, see Devroye et al. (1997). Also, see Fan and Gijbels (1992) and Fan (1993) for the local polynomial regression as an alternative choice of K-NN methods in the literature of non-parametric classification.

If the classifier $\hat{f}^P(\cdot) = \mathbf{1}\{\hat{\eta}^P(\cdot) \geq \frac{1}{2}\}$ is a general plug-in rule, we also offer an explicit upper bound for the signal transfer risk $\varepsilon_T(\hat{f}^P)$ based on the bound of the point-wise misclassification rate of $\hat{\eta}^P(x)$. Please see Appendix A.1 for the rigorous formulation of this bound and related results.

4.1. Non-parametric Classification Setting

We are now in a position to state the applications of our proposed model-selected classifier in non-parametric classification under the finite dimension regime, i.e. there exists some constant $D > 0$ such that $d \leq D$ for any n_Q. In addition to the margin assumption and the ambiguity risk bound condition, this paper considers the non-parametric classification problem when the following smoothness condition holds.

Condition 1 (Smoothness)**.** For any $\beta \in [0, 1]$, define the (β, C_β)-*Holder* class of functions as the set of functions $g : \mathbb{R}^d \to \mathbb{R}$ satisfying that, for any $x, x' \in \mathbb{R}^d$

$$|g(x) - g(x')| \leq C_\beta ||x - x'||^\beta$$

for some constant $C_\beta > 0$. We denote this class of functions by $\mathcal{H}(\beta, C_\beta)$.

Previous works, including Cai and Wei (2021) and Reeve et al. (2021), do not typically require any smoothness assumption for η^P. In contrast, our approach considers the smoothness of both the target and source regression functions. Specifically, we assume that

$$\eta^Q \in \mathcal{H}(\beta, C_\beta), \quad \eta^P \in \mathcal{H}(\beta_P, C_{\beta_P}).$$

This assumption allows us to obtain a more refined upper bound that depends on the smoothness of both functions.

Our next condition concerns the mass of the source and target distributions in the sense that the density functions w.r.t. Q_X and P_X are bounded below. Note that the identical support condition $\Omega = \Omega_P$ is not necessary, as the discrepancy between Ω and Ω_P is implicitly accounted for by the ambiguity level $\varepsilon(\cdot)$. We require that both Q_X and P_X satisfy the following condition:

Condition 2 (Strong Density)**.** A marginal distribution Q_X is absolutely continuous with respect to the Lebesgue measure λ on its *compact support* (denoted by Ω). Furthermore, we have that

$$\lambda(B(x, r) \cap \Omega) \geq c_\mu \lambda(B(x, r)) \quad \forall 0 < r < r_u, x \in \Omega,$$

$$\mu^- < \frac{dQ_X}{\lambda}(x) < \mu^+, \quad \forall x \in \Omega,$$

Denote the set of such marginal distributions Q_X by $\mathcal{S}(\mu^+,\mu^-,c_\mu,r_\mu)$ with parameter $\mu^+,\mu^-,c_\mu,r_\mu > 0$.

Taking all the conditions above into account, we consider the non-parametric parameter space of Π that satisfies Assumption 1, 2 and Condition 1, 2:

$$\begin{aligned}\Pi^{NP} =&\Pi^{NP}(\alpha,C_\alpha,\gamma,C_\gamma,\varepsilon,\beta,\beta_P,C_\beta,C_{\beta_P},\mu^+,\mu^-,c_\mu,r_\mu)\\ :=&\{(Q,P):(Q,P)\in\Pi(\alpha,C_\alpha,\gamma,C_\gamma,\varepsilon),\eta^Q\in\mathcal{H}(\beta,C_\beta),\\ &\eta^P\in\mathcal{H}(\beta_P,C_{\beta_P}),Q_X,P_X\in\mathcal{S}(\mu^+,\mu^-,c_\mu,r_\mu)\}.\end{aligned}$$

We further impose a mild assumption $\alpha\beta\le d$ to rule out the "super-fast" rates of convergence mentioned in Audibert and Tsybakov (2011). This is guaranteed to hold when η^Q hits $\frac{1}{2}$ at an interior point of Ω (See Proposition 3.4 of Audibert and Tsybakov (2011)).

In the following part of this section, we explore three types of additional conditions on the parametric space Π^{NP}, and analyze their corresponding optimal rate of excess risk.

1. **Band-like Ambiguity:** We consider the scenario of band-like ambiguity described in Example 4. Define the focal parametric space as
$$\Pi^{NP}_{BA} := \{(Q,P)\in\Pi^{NP}: s(x)\ge C_\gamma|\eta^Q(x)-\frac{1}{2}|^\gamma-\Delta,\ \forall x\in\Omega\}.$$
Recall that in Example 4, we have shown that this implies
$$\varepsilon(z;\gamma,C_\gamma/2)=\left(C_\alpha z^{1+\alpha}\right)\wedge\left(2^{\frac{1+\alpha}{\gamma}}C_\alpha C_\gamma^{-\frac{1+\alpha}{\gamma}}\Delta^{\frac{1+\alpha}{\gamma}}\right).$$
Furthermore, in this case, we set $\beta_P=0$, which means that no additional smoothness condition is imposed on η^P. Instead, we allow η^P to arbitrarily fluctuate within a small band whose width is measured by Δ. When $\Delta=0$, our setting covers the one in Cai and Wei (2021).

2. **Smooth Source with Arbitrary Ambiguity:** We consider the additional condition of $\gamma\beta=\beta_P$. Define
$$\Pi^{NP}_S := \Pi^{NP}(\alpha,C_\alpha,\gamma,C_\gamma,\varepsilon,\beta,\gamma\beta,C_\beta,C_{\beta_P},\mu^+,\mu^-,c_\mu,r_\mu).$$
Specially, when η^Q and η^P have the same smoothness degree as β, we have $\gamma=1$. While the ambiguity level $\varepsilon(\cdot)$ is arbitrary, the smoothness condition of η^P makes sure that the ambiguity area $\{x\in\Omega: s(x)\le C_\gamma|\eta^Q(x)-\frac{1}{2}|^\gamma\}$ is smooth enough.

3. **Flipped Strong Signal:** We consider strong signal strength over the entire feature space, while the signal direction may be reversed. Define
$$\Pi^{NP}_F := \{(Q,P)\in\Pi^{NP}: \eta^P \text{ is continuous}, |\eta^P(x)-\frac{1}{2}|\ge C_\gamma|\eta^Q(x)-\frac{1}{2}|^\gamma,\ \forall x\in\Omega\}.$$
It is worth noting that $s(x)\ge C_\gamma|\eta^Q(x)-\frac{1}{2}|^\gamma$ if and only if $(\eta^P(x)-\frac{1}{2})(\eta^Q(x)-\frac{1}{2})\ge 0$. Therefore, the ambiguity level precisely captures the risk caused by different Bayes classifiers between the target and source data. We also let $\beta_P=0$ in this case, but assume that η^P is continuous to makes sure that the area $\{x\in\Omega: s(x)\le C_\gamma|\eta^Q(x)-\frac{1}{2}|^\gamma\}$ is smooth enough.

We will provide a detailed discussion on the conditions of Π^{NP}_S and Π^{NP}_F in Section 4.3, specifically in Remark 3. The detailed analysis of the excess risk rate with respect to Π^{NP} is provided in Appendix A.2, considering the most general and intricate case.

4.2. K-Nearest Neighbor Model-selected Classifier

Given the family of the target and source distributions, we could then provide the excess risk upper bound using the K-NN method under this non-parametric setting.

Precisely, given a query point $x \in \mathbb{R}^d$, we first reorder the target data pairs to be $(X_{(1)}, Y_{(1)}), \ldots, (X_{(n_Q)}, Y_{(n_Q)})$ based on their Euclidean distances to x, i.e.,

$$||X_{(1)} - x||_2 \leq \cdots \leq ||X_{(n_Q)} - x||_2.$$

Then, we define the K-NN estimate $\hat{\eta}_k^Q(x)$ as the simple average of the response values of the k_Q nearest neighbors of x in the target data:

$$\hat{\eta}_{k_Q}^Q = \frac{1}{k_Q} \sum_{i=1}^{k_Q} Y_{(i)}(x).$$

Similarly, we define the K-NN estimate $\hat{\eta}_{k_P}^P(x) := \frac{1}{k_P} \sum_{i=1}^{k_P} Y_{(i)}^P(x)$ for the source data pairs $\mathcal{D}_P$. Finally, we plug these estimates into the model-selected K-NN classifier to obtain

$$\hat{f}_{MS}^{NN}(x) = \begin{cases} \mathbf{1}\{\hat{\eta}_{k_Q}^Q(x) \geq \frac{1}{2}\}, & \text{if } |\hat{\eta}_{k_Q}^Q(x) - \frac{1}{2}| \geq \tau, \\ \mathbf{1}\{\hat{\eta}_{k_P}^P(x) \geq \frac{1}{2}\}, & \text{otherwise.} \end{cases}$$

Parameter Choice: We choose the number of nearest neighbors as follows:

$$k_Q = \lfloor c_Q n_Q^{\frac{2\beta}{2\beta+d}} \rfloor, \; k_P = \lfloor c_P n_P^{\frac{2\gamma\beta}{2\gamma\beta+d}} \rfloor$$

where c_Q and c_P can be any positive constants. This choice is motivated by previous works such as Gadat et al. (2016) and Cannings et al. (2020a), where similar choices are made in the context of nearest-neighbor methods. Notably, the choice of k_P is similarly derived by seeing $\gamma\beta$ as the "smoothness" parameter for the source data, and our choice coincides with the classical optimal choice when $\beta_P = \gamma\beta$,. We assume γ and β are known for convenience. For adaptive and rate-optimal approaches to determining the number of nearest neighbors, see Lepski (1993); Reeve et al. (2021).

As for the threshold in the model-selected classifier $\hat{f}_{MS}^{NN}$, we choose

$$\tau \asymp \log(n_Q \vee n_P) k_Q^{-\frac{1}{2}}.$$

This choice is consistent with the concentration property of $\hat{\eta}_{k_Q}^Q$, of which the "uncertainty" level δ_Q in (9) is proportional to $k_Q^{-\frac{1}{2}}$ since $\hat{\eta}_{k_Q}^Q$ is the average of k_Q random variables. By the definition of k_Q, we see that

$$\tau \asymp \log(n_Q \vee n_P) n_Q^{-\frac{\beta}{2\beta+d}}.$$

4.3. Optimal Rate of Excess Risk

The next theorem gives a provable upper bound on the excess risk of our proposed model-selected K-NN classifier $\hat{f}_{MS}^{NN}$, with the aforementioned proper choices of the nearest neighbors k_Q, k_P and the threshold τ.

Theorem 3 (Non-parametric Classification Upper Bound)**.** Suppose that the asymptotic probability that the concentration inequality of $\hat{\eta}^Q_{k_Q}$ fails, i.e. $n_Q^{\frac{d}{2\beta+d}} \exp(-c_Q n_Q^{\frac{2\beta}{2\beta+d}}) \lesssim n_P^{-\frac{\beta(1+\alpha)}{2\gamma\beta+d}}$. Then the model-selected K-NN classifier $\hat{f}^{NN}_{MS}(x)$ satisfies that

1. **Band-like Ambiguity:**

$$\sup_{(Q,P)\in\Pi^{NP}_{BA}} \mathbb{E}_{(\mathcal{D}_Q,\mathcal{D}_P)}\mathcal{E}_Q(\hat{f}^{NN}_{MS}) \lesssim \left(n_P^{-\frac{\beta(1+\alpha)}{2\gamma\beta+d}} + \Delta^{\frac{1+\alpha}{\gamma}}\right) \wedge \left(\log^{1+\alpha}(n_Q \vee n_P) n_Q^{-\frac{\beta(1+\alpha)}{2\beta+d}}\right). \tag{15}$$

2. **Smooth Source with Arbitrary Ambiguity:**

$$\sup_{(Q,P)\in\Pi^{NP}_{S}} \mathbb{E}_{(\mathcal{D}_Q,\mathcal{D}_P)}\mathcal{E}_Q(\hat{f}^{NN}_{MS}) \lesssim \left(n_P^{-\frac{\beta(1+\alpha)}{2\gamma\beta+d}} + \varepsilon(2\tau)\right) \wedge \left(\log^{1+\alpha}(n_Q \vee n_P) n_Q^{-\frac{\beta(1+\alpha)}{2\beta+d}}\right). \tag{16}$$

3. **Flipped Strong Signal:**

$$\sup_{(Q,P)\in\Pi^{NP}_{F}} \mathbb{E}_{(\mathcal{D}_Q,\mathcal{D}_P)}\mathcal{E}_Q(\hat{f}^{NN}_{MS}) \lesssim \left(n_P^{-\frac{\beta(1+\alpha)}{2\gamma\beta+d}} + \varepsilon(2\tau)\right) \wedge \left(\log^{1+\alpha}(n_Q \vee n_P) n_Q^{-\frac{\beta(1+\alpha)}{2\beta+d}}\right). \tag{17}$$

Theorem 3 is obtained simply by verifying the conditions in Theorem 1 (See also Lemma C.1 and C.2 for the general form of the verification details).

Our obtained asymptotic risk is the minimum of the risks due to the target and source data. It reveals that transfer learning leads to faster convergence rates of excess risk when n_P is large compared to n_Q and the ambiguity level is small, satisfying the conditions

$$\Pi^{NP}_{BA} : n_P \gg n_Q^{\frac{2\gamma\beta+d}{2\beta+d}}, \; \Delta \ll n_Q^{-\frac{\gamma\beta}{2\beta+d}};$$
$$\Pi^{S}_{BA}, \Pi^{F}_{BA} : n_P \gg n_Q^{\frac{2\gamma\beta+d}{2\beta+d}}, \; \varepsilon(2\tau) \ll \tau^{1+\alpha}.$$

On the other hand, if n_P is small compared to n_Q, then the target term $n_Q^{-\frac{\beta(1+\alpha)}{2\beta+d}}$ dominates the upper bound, and we conserve the risk rate in the conventional setting with only target data and the strong density assumption (Audibert and Tsybakov 2007) up to logarithmic factors.

Remark 2. When $\Delta = 0$ in (15), or $\varepsilon(\cdot) \equiv 0$ in (17), our obtained upper bound reduces to Theorem 2 of Cai and Wei (2021) up to logarithmic factors. In (16), our result supplements these previous works (Cai and Wei 2021; Reeve et al. 2021) by allowing the arbitrary type of ambiguity after imposing a smoothness condition on η^P.

Remark 3. Proving the risk upper bound in (16) and (17) requires that the neighborhood of any point with strong strength also has strong signal strength. The condition $\gamma\beta = \beta_P$ ensures that points with strong signal strength have neighboring data points with similar conditional probabilities according to η^P. This guarantees that there are enough neighboring data points to indicate the correct classification with respect to the source distribution.

On the other hand, the condition of strong signal strength in absolute value, combined with the continuity of η^P, ensures that the area of not-so-strong signal strength, i.e. $\{x \in \Omega : s(x) \leq C_\gamma|\eta^Q(x) - \frac{1}{2}|^\gamma\}$, is smooth. In fact, its boundary is part of the decision boundary

$\{x \in \Omega : \eta^Q(x) = \frac{1}{2}\}$ whose smoothness is guaranteed by the smoothness of η^Q (See Corollary 1). This smoothness property ensures the availability of a sufficient number of neighboring data points to accurately determine the classification based on the source distribution. Lacking such conditions will result in an additional risk term related to the boundary of the ambiguity area $\{x \in \Omega : s(x) \leq C_\gamma |\eta^Q(x) - \frac{1}{2}|^\gamma\}$. Please refer to Appendix A.2 for more details.

We meanwhile provide the lower bound result for the parameter spaces, which shows that the model-selected K-NN classifier achieves the optimal rate. We can even require that Ω and Ω_P are identical when proving the optimal lower bound.

Theorem 4 (Non-parametric Classification Lower Bound)**.** *Fix the parameters with $\alpha\beta \leq d$. We have that*

1. **Band-like Ambiguity:**

$$\inf_{\hat{f}} \sup_{\substack{(Q,P)\in\Pi_{BA}^{NP} \\ \Omega=\Omega_P}} \mathbb{E}_{(\mathcal{D}_Q,\mathcal{D}_P)}\mathcal{E}_Q(\hat{f}_{MS}^{NN}) \gtrsim \left(n_P^{-\frac{\beta(1+\alpha)}{2\gamma\beta+d}} + \Delta^{\frac{1+\alpha}{\gamma}}\right) \wedge n_Q^{-\frac{\beta(1+\alpha)}{2\beta+d}}. \tag{18}$$

2. **Smooth Source with Arbitrary Ambiguity:** For some constant $c > 0$,

$$\inf_{\hat{f}} \sup_{\substack{(Q,P)\in\Pi_{S}^{NP} \\ \Omega=\Omega_P}} \mathbb{E}_{(\mathcal{D}_Q,\mathcal{D}_P)}\mathcal{E}_Q(\hat{f}_{MS}^{NN}) \gtrsim \left(n_P^{-\frac{\beta(1+\alpha)}{2\gamma\beta+d}} + \varepsilon(2\tau)\right) \wedge n_Q^{-\frac{\beta(1+\alpha)}{2\beta+d}}. \tag{19}$$

3. **Flipped Strong Signal:** For some constant $c > 0$,

$$\inf_{\hat{f}} \sup_{\substack{(Q,P)\in\Pi_{F}^{NP} \\ \Omega=\Omega_P}} \mathbb{E}_{(\mathcal{D}_Q,\mathcal{D}_P)}\mathcal{E}_Q(\hat{f}_{MS}^{NN}) \gtrsim \left(n_P^{-\frac{\beta(1+\alpha)}{2\gamma\beta+d}} + \varepsilon(2\tau)\right) \wedge n_Q^{-\frac{\beta(1+\alpha)}{2\beta+d}}. \tag{20}$$

In the special case where $\sup_{x\in\Omega} |\eta^Q(x) - \eta^P(x)| \leq \Delta$, we can determine that the minimax optimal excess risk is $\left(n_P^{-\frac{\beta(1+\alpha)}{2\beta+d}} + \Delta^{1+\alpha}\right) \wedge n_Q^{-\frac{\beta(1+\alpha)}{2\beta+d}}$, up to logarithmic terms of $n_Q \vee n_P$. As long as

$$n_P \gg n_Q, \;\; \Delta \ll n_Q^{-\frac{\beta}{2\beta+d}},$$

the convergence rate of the excess risk will benefit from the source data. The proof of Theorem 4 reveals that the condition (7), although slightly stronger than the band-like ambiguity condition (5), remains compatible with the lower bound construction as it ensures $\sup_{x\in\Omega} |\eta^Q(x) - \eta^P(x)| \leq \Delta$.

5. Applications in Logistic Regression

Besides non-parametric classification, we also investigate the use of transfer learning in logistic regression models, which are a commonly used parametric approach in classification. Previous works such as Zheng et al. (2019) have studied the “data enriched model” for logistic regression under a single-source setting, Abramovich and Grinshtein (2019) have explored sparse logistic regression in high-dimensional settings, and Tian and Feng (2022) have considered transfer learning in generalized linear models. Our goal is to reveal how incorporating an additional source logistic regression model with a different linear term coefficient can enhance the convergence of the excess misclassification rate.

Suppose that η^Q and η^P both follow high-dimensional logistic regression formulas. Under the regime of $d \overset{n_Q \to \infty}{\longrightarrow} \infty$, the models are supposed to be

$$\begin{aligned} &\text{Target data model: } \eta^Q(x) = \sigma({\beta_Q}^T x) \\ &\text{Source data model: } \eta^P(x) = \sigma({\beta_P}^T x) \end{aligned} \tag{21}$$

where we observe two independent samples $(X_1, Y_1), \cdots, (X_{n_Q}, Y_{n_Q}) \overset{\text{iid}}{\sim} Q$ and $(X_1^P, Y_1^P), \cdots, (X_{n_P}^P, Y_{n_P}^P) \overset{\text{iid}}{\sim} P$. To simplify the theoretical analysis, we assume that the marginal distributions Q_X and P_X are both just $N(0, I_d)$, the d-dimension standard normal distribution. This simple marginal distribution provides convenience for the restricted strong convexity condition (See Negahban et al. (2009)).

Let $\langle \alpha, \beta \rangle$ be the angle between two vectors α and β. Notably, we view the angle between 0 and any vector as $\frac{\pi}{2}$. Suppose that parametric space of the coefficient pair (β_Q, β_P) we consider is

$$\Theta(s, \Delta) = \{(\beta_Q, \beta_P) : ||\beta_Q||_0 \leq s, \langle \beta_Q, \beta_P \rangle \leq \Delta\} \tag{22}$$

for some $s > 0$ and $\Delta \in [0, \frac{\pi}{2})$. The corresponding family of distribution pairs is then

$$\begin{aligned} \Pi^{LR} = \Pi^{LR}(s, \Delta, M) = \{(Q, P) : X, X^P \sim N(0, I_d), \eta^Q(x) = \sigma(\beta_Q^T x), \\ \eta^P(x) = \sigma(\beta_P^T x), (\beta_Q, \beta_P) \in \Theta(s, \Delta)\}. \end{aligned} \tag{23}$$

To ensure the control of the ambiguity risk bound, we impose an innovative constraint on the angle between β_Q and β_P, which must be smaller than a constant Δ. Though, our constructed parametric space does not impose any sparsity conditions on β_P.

We note that the family of distribution pairs Π^{LR}, which we consider in the context of logistic regression, is a subset of the overall distribution pair space Π. It is worth mentioning that the corresponding ARB for Π^{LR} is bounded by $\varepsilon(z, 1, \frac{m}{\pi}) \lesssim z^2 \wedge \Delta^2$, as shown in the following lemma.

Lemma 1. Assume that $L \leq m||\beta_Q||_2 \leq ||\beta_P||_2 \leq U$ for some constant $L, U > 0$ and $0 < m \leq 1$. We have that $\Pi^{LR}(s, \Delta) \subset \Pi(\alpha, C_\alpha, \gamma, C_\gamma, \varepsilon)$ where

$$\alpha = 1, \; C_\alpha = \frac{16m}{\sqrt{2\pi}L} \vee 4, \; \gamma = 1, \; C_\gamma = \frac{m}{\pi},$$

$$\varepsilon(z; 1, \frac{m}{\pi}) = \left((\frac{16m}{\sqrt{2\pi}L} \vee 4) z^2 \right) \wedge \frac{\sqrt{2}U}{m} \Delta^2.$$

The condition $||\beta_P|| \geq m||\beta_Q||$ ensures that $\eta^P(x)$ gives strong signal strength relative to $\eta^Q(x)$ over the majority of Ω.

For model fitting, we minimize the cross-entropy loss function with lasso regularization terms to obtain $\hat{\beta}_Q$ and $\hat{\beta}_P$, i.e.

$$\begin{aligned}\hat{\beta}_Q &= \underset{\beta\in\mathbb{R}^p}{\arg\min}\,\frac{1}{n_Q}\sum_{i=1}^{n_Q}\left\{\log(1+e^{X_i^T\beta}) - Y_iX_i^T\beta\right\} + \lambda_Q||\beta||_1 \\ \hat{\beta}_P &= \underset{\beta\in\mathbb{R}^p}{\arg\min}\,\frac{1}{n_P}\sum_{i=1}^{n_P}\left\{\log(1+e^{X_i^{PT}\beta}) - Y_i^PX_i^{PT}\beta\right\} + \lambda_P||\beta||_1\end{aligned} \tag{24}$$

where $\lambda_* \asymp \sqrt{\frac{\log d}{n_*}}$ for $* \in \{Q, P\}$. The corresponding target and source plug-in classifiers are then $\mathbf{1}\{\sigma(\hat{\beta}_Q^T x) \geq \frac{1}{2}\} = \mathbf{1}\{\hat{\beta}_Q^T x \geq 0\}$ and $\mathbf{1}\{\sigma(\hat{\beta}_P^T x) \geq \frac{1}{2}\} = \mathbf{1}\{\hat{\beta}_P^T x \geq 0\}$. Hence, the model-selected lasso classifier reads

$$\hat{f}_{MS}^{LR}(x) = \begin{cases}\mathbf{1}\{\hat{\beta}_Q^T x \geq 0\}, & \text{if } |\sigma(\hat{\beta}_Q^T x) - \frac{1}{2}| \geq \tau, \\ \mathbf{1}\{\hat{\beta}_P^T x \geq 0\}, & \text{otherwise.}\end{cases}$$

By setting λ_Q, λ_P and τ properly, the excess risk upper bound we obtain is given by the following theorem:

Theorem 5. Assume that $L \leq m||\beta_Q||_2 \leq ||\beta_P||_2 \leq U$ for some constant $L, U > 0$ and $0 < m \leq 1$. Suppose that for some constant $K > 0$, we have $d^K \gtrsim \frac{n_Q \vee n_P}{s \log d}$, $n_Q \gg \log \frac{n_P}{s \log d}$, and $n_Q \wedge n_P \gg s \log d$. Given $\hat{\beta}_Q, \hat{\beta}_P$ obtained by (24) with

$$\lambda_Q = c_Q\sqrt{\frac{\log d}{n_Q}},\ \lambda_P = c_P\sqrt{\frac{\log d}{n_P}}$$

for some constants $c_Q, c_P \geq \sqrt{(K+1)}$, the model-selected lasso classifier with

$$\tau = c_\tau\sqrt{\frac{s\log d}{n_Q}}\log(n_Q \vee n_P)$$

for some constant $c_\tau > 0$ satisfies that

$$\sup_{(Q,P)\in\Pi^{LR}} \mathbb{E}_{(\mathcal{D}_Q,\mathcal{D}_P)}\mathcal{E}_Q(\hat{f}_{MS}^{LR}) \lesssim (\frac{s\log d}{n_P} + \Delta^2) \wedge \left(\frac{s\log d}{n_Q}\log^2(n_Q \vee n_P)\right). \tag{25}$$

The term $\frac{s \log d}{n_Q}$ is the classical risk term with access to only the target data, and $\frac{s \log d}{n_P}$ is the risk term transferred by the source data with an additional term Δ^2 measuring the discrepancy between β_Q and β_P. From the form of the risk upper bound, we see that the knowledge from the source data could significantly improve the learning performance with n_P being large and Δ being small, satisfying that

$$n_P \gg n_Q, \quad \Delta \ll \sqrt{\frac{s\log d}{n_Q}}.$$

It is worth noting that the small angle condition between β_Q and β_P considered in this paper is a more general assumption than the contrast assumption in Tian and Feng (2022), which requires that the l_q-norm of the difference between β_Q and β_P is small for some $q \in [0, 1]$. This makes the results in this paper applicable to a broader class of parameter spaces, allowing for non-sparse β_P that differ much in norm from β_Q.

The selection of τ is straightforward. We choose this τ because it holds with high probability that $||\hat{\beta}_Q - \beta_Q||_2 \lesssim \sqrt{s}\lambda_Q$. This implies

$$|\sigma(\hat{\beta}_Q^T x) - \sigma(\beta_Q^T x)| \lesssim \sqrt{s}\lambda_Q$$

with high probability with respect to $Q_X = N(0, I_d)$, which is just the desirable concentration bound. The proof of Theorem 5 then comes from verifying every condition in Theorem 2 with the parameters indicated by Lemma 1. The theory holds with $\varepsilon(z) \asymp z^2 \wedge \Delta^2$ and $\sup_{(Q,P)\in\Pi^{LR}} \mathbb{E}_{\mathcal{D}_P}\varepsilon_T(\mathbf{1}\{\hat{\beta}_P^T x \geq 0\}) \asymp \frac{s\log d}{n_P} + \Delta^2$, which indicates that even when β_P is non-sparse, we can obtain a reliable estimate of β_P by incorporating a lasso regularizer if Δ is sufficiently small (See Lemma D.2).

Theorem 6 below shows that the upper bound (25) in Theorem 5 is optimal up to some logarithmic terms of $n_Q \vee n_P$.

Theorem 6. Suppose that $\frac{s\log d}{n_Q \vee n_P} \lesssim 1$. We have that

$$\inf_{\hat{f}} \sup_{(Q,P)\in\Pi^{LR}} \mathbb{E}\mathcal{E}_Q(\hat{f}) \gtrsim \left(\frac{s\log d}{n_P} + \Delta^2\right) \wedge \frac{s\log d}{n_Q}.$$

The derivation of the lower bound involves two terms. The term $\frac{s\log d}{n_P} \wedge \frac{s\log d}{n_Q}$ represents the optimal convergence rate when the target and source distributions are identical, and it serves as the convergence rate term under the ideal scenario of $\beta_P = \beta_Q$. The term $\Delta^2 \wedge \frac{s\log d}{n_Q}$ corresponds to the lower bound when $\beta_P = (1, 0, 0, \cdots, 0)$, which imposes sparsity constraints on β_Q within a small cone.

It is worth noting that traditional lower bound results, typically derived from Fano's lemma, only consider the minimax rate over the parametric distance. To address the lower bound of the excess risk $\mathbb{E}\mathcal{E}_Q(\hat{f})$, we introduce a transformation that relates the excess risk to the angle difference of the linear coefficients. By applying Fano's lemma to this transformed quantity, we obtain the desired lower bound. Please refer to Lemma E.4 for a detailed explanation.

6. Simulation Studies

As mentioned earlier, the model-selected classifier offers benefits in scenarios where n_P is large and the ambiguity level is small, while also preventing negative transfer when the source data lacks sufficient information to aid the classification. In this section, we present simulation studies to demonstrate the practical benefits of transfer learning and the model-selected classifier. We separately consider the non-parametric classification and logistic regression settings.

6.1. Non-parametric Classification Setting

The setting we considered is as follows: $d = 2,\ \Omega = \Omega_P = [0, 1]^2,\ Q_X = P_X = \text{Uni}([0, 1]^2)$, the uniform distribution over the feature space. For the target regression function, let $\eta^Q(x) =$

$\eta^Q(x_1, x_2) = \frac{1}{2} + \frac{1}{10}\sin(2\pi(x_1 + x_2))$. We have $\alpha = \beta = 1$ for some constants C_α and C_β. Next, we consider two different non-parametric regression scenarios, by specifying the source regression function η^P with different types of ambiguity.

1. **Band-like Ambiguity:** For $\Delta \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$, $\gamma \in \{0.5, 1\}$:

$$\eta^P(x) = \begin{cases} \frac{1}{2} + 2(\eta^Q(x) - \frac{1}{2})^\gamma - \Delta, & \text{if } \eta^Q(x) \geq \frac{1}{2}, \\ \frac{1}{2} - 2(\frac{1}{2} - \eta^Q(x))^\gamma + \Delta, & \text{if } \eta^Q(x) < \frac{1}{2}. \end{cases}$$

   Here, η^P concentrates around an informative curve with respect to η^Q with some shifting error term Δ. In this case, we have $s(x) \geq |\eta^Q(x) - \frac{1}{2}|^\gamma - \Delta$.

2. **Partially Flipped Sine Functions:** For $r \in \{0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35\}$, $\gamma \in \{0.5, 1\}$:

$$\eta^P(x) = \eta^P(x_1, x_2) = \begin{cases} \frac{1}{2} - \frac{1}{5}\sin(2\pi\frac{\{x_1+x_2\}}{r})^\gamma, & \text{if } \{2x_1 + 2x_2\} \in [0, r], \\ \frac{1}{2} + \frac{1}{5}\sin(2\pi\frac{\{x_1+x_2\}-r}{1-r})^\gamma, & \text{if } \{2x_1 + 2x_2\} \in (r, 1], \end{cases}$$

   where $\{a\} = a - \lfloor a \rfloor$ represents the fraction part of any real value a. The following graph illustrates our setup of η^P. While keeping η^P continuous with $\beta_P = 1$. The ratio parameter r creates an area where the Bayes classifier differs from the target distribution. The Bayes classifiers are identical when $r = 0$ and completely opposite when $r = 1$.

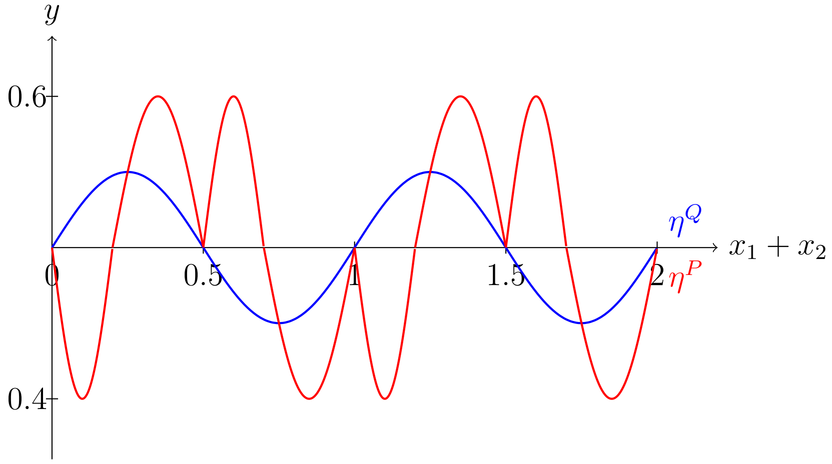

Figure: Illustration of η^P in the second simulation setup. $\gamma = 1$ and $r = 0.4$.

In each scenario, we set $n_Q = 200$ and $n_P = 1000$, as a large n_P is necessary to observe the benefits of transfer learning. Previous studies Cai and Wei (2021); Reeve et al. (2021) have shown that accuracy improves with increasing n_P. We also generate 50000 independent test pairs from Q for obtaining the accuracy. For the model-selected K-NN classifier, we choose $k_Q = \lfloor n_Q^{\frac{2\beta}{2\beta+d}} \rfloor = 31$, $k_P = \lfloor n_Q^{\frac{2\gamma\beta}{2\gamma\beta+d}} \rfloor$, and $\tau = 0.05$.

We choose the simple K-NN classifiers on $Q-$data and P-data as two benchmarks for comparison. Moreover, we add the K-classifier on the pooled data combining both the target and source where we choose the nearest-neighbor parameter k by 5-fold cross-validation.

Figure 1 shows that our model-selected classifier improves the accuracy when Δ is small as K-NN on P-data does, since n_P is large. Furthermore, our model-selected classifier outperforms K-NN on P-data or pooled data by a significant margin, demonstrating its ability to avoid negative transfer when the source data is unreliable. The pooled data algorithm benchmark improves the classification when Δ is small; however, it dramatically breaks down when Δ is large.

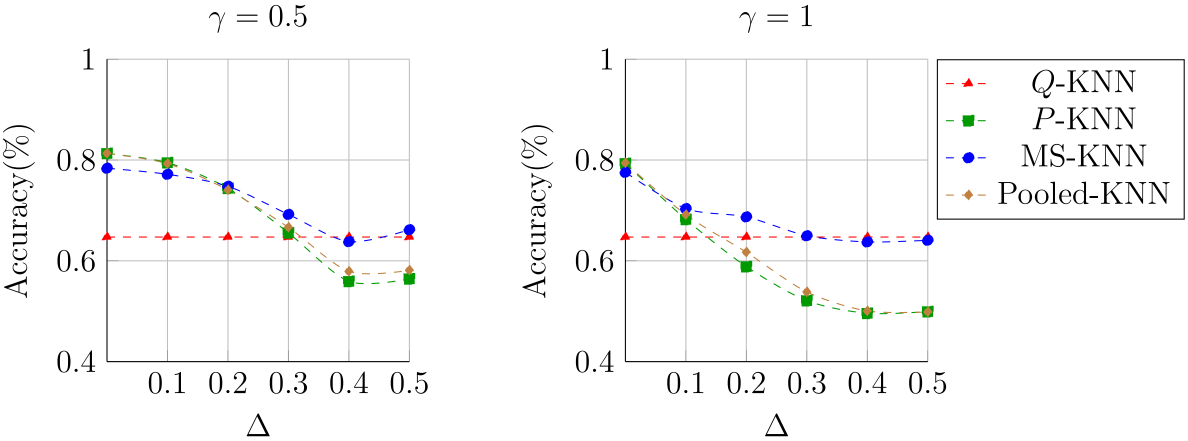

Figure 1: Accuracy of the model-selected K-NN classifiers under the band-like ambiguity scenario. We do experiments on different band error Δ given $\gamma = 0.5$ and 1. Blue: Model-select K-NN classifier. Red: K-NN classifier on only Q-data. Green: K-NN classifier on only P-data. Brown: K-NN classifer on pooled data.

Additionally, Figure 2 shows that our model-selected classifier improves the accuracy when the ambiguity level, depicted by r, is small. Plus, our model-selected classifier outperforms K-NN on P-data or pooled data when the ambiguity level is too large to benefit transfer learning, conserving the classification ability using only Q-data.

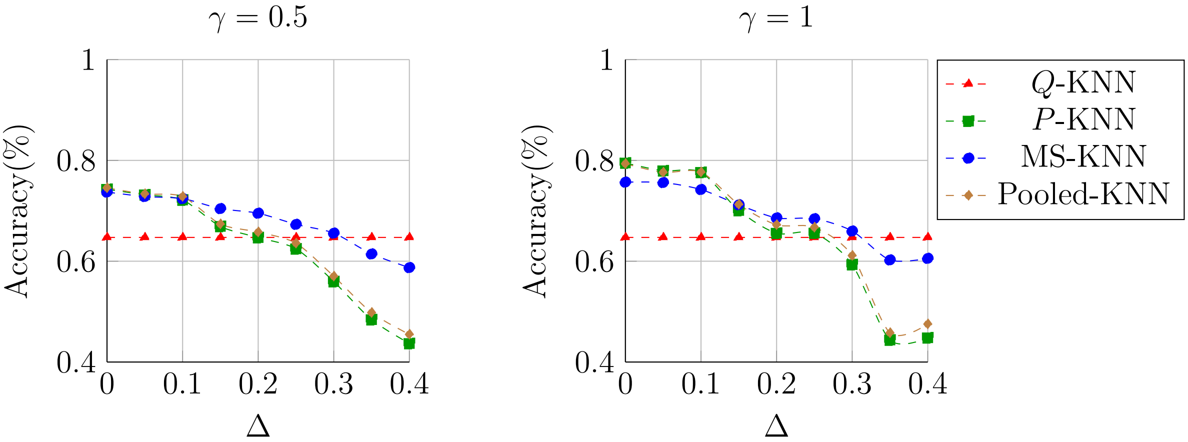

Figure 2: Accuracy of the model-selected K-NN classifiers under the scenario with partially flipped sine functions. We do experiments on the ratio parameter r given $\gamma = 0.5$ and 1. Blue: Model-select K-NN classifier. Red: K-NN classifier on only Q-data. Green: K-NN classifier on only P-data. Brown: K-NN classifer on pooled data.

6.2. Logistic Regression Setting

We meanwhile consider the scenario where both η^Q and η^P follow the logistic models (21). The parameter is given by (24). We set $n_Q = 200$, $n_P = 500$, $Q_X = P_X = N(0, I_d)$, and simulate 50000 test data for calculating accuracy. The linear coefficient is given by

$$\beta_Q = (0.5 \cdot \mathbf{1}_s, \mathbf{0}_{d-s}), \quad \beta_P = (1.5 \cdot \mathbf{1}_s, \frac{||\beta_Q||}{\sqrt{d-s}} \tan \Delta \cdot \mathbf{1}_{d-s}),$$

where $\mathbf{1}_s$ is a vector of all 1s with size s, and $\mathbf{0}_s$ is a vector of all 0s with size $d - s$. Here, we set $s = 10$, which is small compared to d. For the source, β_P could be treated as the rotated version of $3\beta_Q$, with an angle of exactly Δ between them. The range of Δ is chosen in the set $\{0, 0.2, 0.4, 0.6, 0.8, 1, 1.2, 1.4, 1.6\}$, which gradually approaches $\frac{\pi}{2}$. We simulate 50000 test data points for each value of Δ to calculate the accuracy. The lasso regularization parameter is achieved by 5-fold cross-validation and chosen to be the largest λ at which the MSE is within one standard error of the minimum estimation MSE by tuning λ.

In addition to our proposed model-selected classifier, we compare three benchmarks for performance: logistic regression with lasso penalty on Q-data, P-data, and pooled data. Figure 3 demonstrates that our model-selected classifier achieves high accuracy when the angle between β_Q and β_P is small. Interestingly, as evidence of its robustness, our classifier retains some classification ability of the target data even when the angle is large, while the benchmark classifiers based on P-data and pooled data suffer from negative transfer and show much reduced accuracy in such cases.

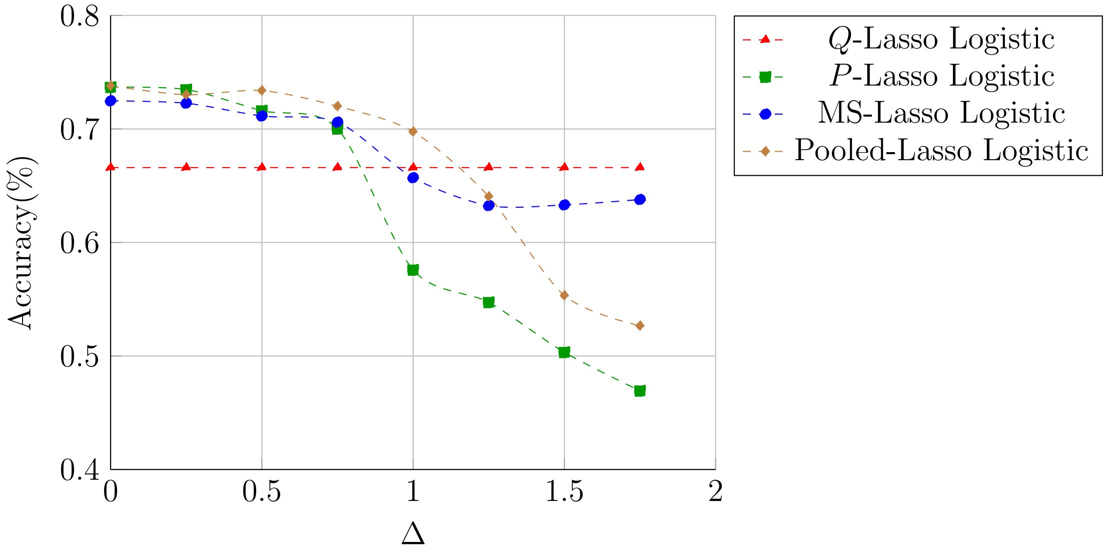

Figure 3: Accuracy of the model-selected logistic classifier with lasso penalty. We do experiments on difference choices of the angle Δ. Blue: Model-select logistic classifiers with lasso penalty. Red: Logistic classifier with lasso penalty on only Q-data. Green: Logistic classifier with lasso penalty on only P-data. Brown: Logistic classifier with lasso penalty on pooled data.

7. From Statistical Learning to Transfer Learning

Recall that the signal transfer risk for $\hat{f}^P$ is defined as

$$\varepsilon_T(f) := \mathbb{E}_{(X,Y)\sim Q}[|\eta^Q(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}^P(X) \neq f_Q^*(X), s(X) \geq C_\gamma|\eta^Q(X) - \frac{1}{2}|^\gamma\}].$$

The majority of the existing literature focuses on bounding the target excess risk using only target data. While the excess risk of a classifier $\hat{f}^P$ with respect to the source distribution, i.e.

$$2\mathbb{E}_{(X,Y)\sim P}[|\eta^P(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}^P(X) \neq f_Q^*(X)\}]$$

is well-studied, its signal transfer risk $\varepsilon_T(\hat{f}^P)$ is less explored. To incorporate the vast literature of traditional statistical learning into our framework, we present a proposition that directly bounds the signal transfer risk $\varepsilon_T(\hat{f}^P)$ in terms of the excess risk of the source distribution. This proposition involves a more refined version of the signal transfer risk, and more significantly, does not rely on any concentration property indicated by plug-in rules.

Some additional work is needed. Firstly, we extend the signal transfer risk in Definition 2 by introducing the restricted signal transfer risk. Next, we propose a slightly stronger version of the general convergence theorem (Theorem 7) to obtain a tighter bound on the signal transfer risk that is necessary for the scenario in this section. Finally, we apply this stronger version to derive the excess risk directly from the source excess risk bound.

Definition 3 (Restricted Signal Transfer Risk)**.** Define the *restricted signal transfer risk* of the classifier f with respect to parameters $\gamma, C_\gamma > 0$ as

$$\begin{aligned}\varepsilon_T(f, z; \gamma, C_\gamma) := \mathbb{E}_{(X,Y)\sim Q}[|\eta^Q(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}^P(X) \neq f_Q^*(X),\\ s(X) \geq C_\gamma|\eta^Q(X) - \frac{1}{2}|^\gamma, |\eta^Q(x) - \frac{1}{2}| \leq z\}].\end{aligned}$$

for any $z \in [0, \frac{1}{2}]$ We abbreviate $\varepsilon_T(f, z; \gamma, C_\gamma)$ as $\varepsilon_T(f, z)$ when there is no need to specify γ and C_γ.

Compared with the signal transfer risk, the restricted signal transfer risk further controls the distance of η^Q from $\frac{1}{2}$. Particularly, the signal transfer risk can be defined as the restricted signal transfer risk with $z = \frac{1}{2}$. Plus, by definition we have

$$\varepsilon_T(f, z; \gamma, C_\gamma) \leq \varepsilon_T(f; \gamma, C_\gamma) \wedge (C_\alpha z^{1+\alpha}).$$

Hence, by replacing the term $\sup_{(Q,P)\in\tilde{\Pi}} \mathbb{E}_{\mathcal{D}_P}\varepsilon_T(\hat{f}^P) \wedge \tau^{1+\alpha}$ with the asymptotically smaller $\sup_{(Q,P)\in\tilde{\Pi}} \mathbb{E}_{\mathcal{D}_P}\varepsilon_T(\hat{f}^P, 2\tau)$, we can obtain stronger versions of the general convergence theorems, which is as follows:

Theorem 7. It holds that

1. by keeping all the notations in Theorem 1, the *model-selected classifier* satisfies that

$$\sup_{(Q,P)\in\tilde{\Pi}} \mathbb{E}_{(\mathcal{D}_Q,\mathcal{D}_P)}\mathcal{E}_Q(\hat{f}_{MS}) \lesssim \sup_{(Q,P)\in\tilde{\Pi}} \mathbb{E}_{\mathcal{D}_P}\varepsilon_T(\hat{f}^P, 2\tau) + \varepsilon(2\tau) + \delta_Q^f. \tag{26}$$

given that $\sup_{(Q,P)\in\tilde{\Pi}} \mathbb{E}_{\mathcal{D}_P}\varepsilon_T(\hat{f}^P, 2\tau) \gtrsim n_Q^{-c} \vee n_P^{-c}$ for some constant $c > 0$.

2. by keeping all the notations in Theorem 2, the *model-selected classifier* satisfies that

$$\sup_{(Q,P)\in\tilde{\Pi}} \mathbb{E}_{(\mathcal{D}_Q,\mathcal{D}_P)}\mathcal{E}_Q(\hat{f}_{MS}) \lesssim \sup_{(Q,P)\in\tilde{\Pi}} \mathbb{E}_{\mathcal{D}_P}\varepsilon_T(\hat{f}^P, 2\tau) + \varepsilon(2\tau) + \delta_Q(n_Q, \tau). \tag{27}$$

Besides the preparation of the restricted signal transfer risk, Condition 3 serves to ensure the boundedness of the Radon-Nikodym derivative $\frac{dQ_X}{dP_X}$.

Condition 3 (Absolutely Continuity)**.** Q_X is absolutely continuous w.r.t. P_X. Moreover, there exists some constant $M > 0$ such that the Radon-Nikodym derivative $\frac{dQ_X}{dP_X}(x) \leq M$ for any $x \in \Omega$. Denote all of such marginal distribution pairs (Q_X, P_X) by $\mathcal{A}(M)$ with parameter $M > 0$.

Condition 3 requires that the Radon-Nikodym derivative $\frac{dQ_X}{dP_X}$ is bounded away from infinity, ensuring that P_X is sufficiently strong to learn every point in Ω relative to Q_X. In other words, the source distribution provides enough coverage over the space Ω to enable accurate learning of the target distribution. It is worth noting that this condition implicitly implies that Ω is a subset of Ω_P.

Now, we are in a position to present the upper bound for the signal transfer risk with respect to the target distribution in terms of the source excess risk.

Proposition 1. Given $\tilde{\Pi} \subset \Pi \cap \mathcal{A}(M)$. Define the source excess risk as

$$\varepsilon_P := \sup_{(Q,P)\in\tilde{\Pi}} \mathbb{E}_{\mathcal{D}_P}\mathbb{E}_{(X,Y)\sim P}[|\eta^P(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}^P(X) \neq f_Q^*(X)\}].$$

Then, for any $z \in (0, \frac{1}{2})$, we have

$$\sup_{(Q,P)\in\tilde{\Pi}} \mathbb{E}_{\mathcal{D}_P}\varepsilon_T(\hat{f}^P, z) \leq \begin{cases} 2M^{\frac{1+\alpha}{\gamma+\alpha}} C_\alpha^{\frac{\gamma-1}{\gamma+\alpha}} C_\gamma^{-\frac{1+\alpha}{\gamma+\alpha}} \varepsilon_P^{\frac{1+\alpha}{\gamma+\alpha}}, \ \gamma \geq 1 \\ MC_\gamma^{-1} z^{1-\gamma}\varepsilon_P, \ \gamma < 1 \end{cases}.$$

Specially, the signal transfer risk is well-defined by choosing

$$\sup_{(Q,P)\in\tilde{\Pi}} \mathbb{E}_{\mathcal{D}_P}\varepsilon_T(\hat{f}^P) \leq \begin{cases} 2M^{\frac{1+\alpha}{\gamma+\alpha}} C_\alpha^{\frac{\gamma-1}{\gamma+\alpha}} C_\gamma^{-\frac{1+\alpha}{\gamma+\alpha}} \varepsilon_P^{\frac{1+\alpha}{\gamma+\alpha}}, \ \gamma \geq 1 \\ 2^{\gamma-1}MC_\gamma^{-1}\varepsilon_P, \ \gamma < 1 \end{cases}.$$

In cases where the information of such source excess risk ε_P is given, Proposition 1 claims that the (restricted) signal transfer risk could be accordingly bounded, and inherits the performance and possible consistency of $\hat{f}^P$ w.r.t. classification on the source data. Plus, we deduce that no matter what the signal transfer exponent γ is, $\sup_{(Q,P)\in\tilde{\Pi}} \mathbb{E}_{\mathcal{D}_P}\varepsilon_T(\hat{f}^P, z)$ converges faster than the rate obtained by only access to the target data by setting $\tau = 0$, i.e. $\delta_Q^{1+\alpha}$, as long as $\varepsilon_P \ll \delta_Q^{\frac{\gamma+\alpha}{1+\alpha}} \lesssim \tau^{1+\alpha}$, which implies that $\hat{f}^P$ could also learn the target distribution well if it learns the source distribution well.

Proposition 1 only requires knowledge of some conventional statistical learning techniques, which are widely studied and well-understood in the literature. This means that researchers and practitioners can easily apply our framework to a wide range of problems, without the need for specialized knowledge or expertise in transfer learning or related fields. Conventional studies of the excess risk rate could be thus applied to obtain the source excess risk by just viewing the source as the target.

8. Conclusion

In this paper, we have proposed a new approach to transfer learning that is robust against an unreliable source distribution with arbitrary ambiguity in the source data. Our work can be viewed as an extension of Cai and Wei (2021) and Reeve et al. (2021). Through the introduction of the ambiguity level, our approach helps people to understand the circumstances under which we can improve the classification given the source data with potential ambiguity. Our proposed model-selected classifier, with a threshold τ balancing the performance of both the target and source data, is shown to be both *efficient* and *robust*, as the excess risk improves for a reliable source distribution and avoids negative transfer with an unreliable source distribution. Plus, we find that the target data could also help alleviate the risk caused by the ambiguity level.

We then demonstrate the power of our approach on specific classification tasks, with a focus on non-parametric classification and logistic regression settings. The upper bounds are shown optimal up to some logarithmic terms and are more general than previous works in the literature on transfer learning. Simulation studies provide numerical evidence for these two classification tasks. Finally, we provide down-to-earth approaches to bound the signal transfer risk, a key component of the excess risk in our general convergence result, in terms of the conventional excess risk extensively studied in the literature of statistical learning.

There are several promising avenues for future research that may build upon the contributions of this paper. One potential extension is to consider an extension of the signal strength and ambiguity level to incorporate a translation parameter, i.e.

$$s_\kappa(x) := \begin{cases} |\eta^P(x) - \frac{1}{2} - \kappa|, & \operatorname{sgn}\left(\eta^Q(x) - \frac{1}{2} - \kappa\right) \times (\eta^P(x) - \frac{1}{2}) \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

$$\mathbb{E}_{(X,Y)\sim Q}\left[|\eta^Q(X) - \frac{1}{2}|\mathbf{1}\{s_\kappa(X) \leq C_\gamma|\eta^Q(X) - \frac{1}{2}|^\gamma \leq C_\gamma z^\gamma\}\right] \leq \varepsilon_\kappa(z; \gamma, C_\gamma).$$

This extension is natural since the decision boundary $\{x \in \Omega : \eta^Q(x) = \frac{1}{2}\}$ may be closer to $\{x \in \Omega : \eta^P(x) = \frac{1}{2} + \kappa\}$ with a translation. We conjecture that an additional estimation error of $(\frac{\log n_Q}{n_Q})^{\frac{1+\alpha}{2+\alpha}}$ may be incurred due to the presence of an unknown κ, and an empirical risk minimization procedure after obtaining $\hat{\eta}^P$ like

$$\hat{\kappa} = \operatorname*{arg\,min}_{\kappa \in [-\frac{1}{2}, \frac{1}{2}]} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left\{\mathbf{1}\{\hat{\eta}^P(X_i) \geq \frac{1}{2} + \kappa\} \neq Y_i\right\}$$

may be required to obtain a corrected version of the estimate $\hat{f}^P$.

Secondly, it is worthy of developing an adaptive procedure for selecting the threshold τ in non-parametric classification, which could incorporate unknown parameters such as α, β in non-parametric classification, and s in logistic regression. While simple ERM-based methods may lead to an additional risk term related to n_Q, Lepski's method (Lepski 1993) may offer a solution in the non-parametric classification for maintaining optimal rates with an adaptive choice of τ. It is an open question whether choosing $\tau \asymp \log(n_Q \vee n_P) k_Q^{-\frac{1}{2}}$, where k_Q is chosen by Lepski's method, is helpful for obtaining an optimal rate. In a more general setting besides non-parametric classification, since the information τ can often be represented by δ_Q, it may be reasonable to develop a general adaptive approach to the estimate of δ_Q, which can be used to determine τ directly.

Lastly, the conditions presented in the non-parametric and logistic model settings could be relaxed to some extent, for instance, by considering non-compact feature spaces with sub-Gaussian conditions or other marginal distribution assumptions (e.g., Assumption A4 in Gadat et al. (2016)). Our proposed model-selected classifier is expected to perform well in these scenarios without a significant modification of the model-selected classifier framework and may offer advantages over other case-to-case carefully-designed estimates in the transfer learning literature.

REFERENCES

Abramovich, F. and Grinshtein, V. (2019). High-dimensional classification by sparse logistic regression. *IEEE Transactions on Information Theory*, 65(5):3068–3079.

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

Audibert, J.-Y. and Tsybakov, A. B. (2011). Fast learning rates for plug-in classifiers under the margin condition.

Bastani, H. (2020). Predicting with proxies: Transfer learning in high dimension. *Management Science*, 67.

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

Ben-David, S., Lu, T., Luu, T., and Pal, D. (2010b). Impossibility theorems for domain adaptation. In Teh, Y. W. and Titterington, M., editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 129–136, Chia Laguna Resort, Sardinia, Italy. PMLR.

Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199 – 227.

Blanchard, G., Deshmukh, A. A., Dogan, U., Lee, G., and Scott, C. (2021). Domain generalization by marginal transfer learning. *J. Mach. Learn. Res.*, 22(1).

Cai, T. T. and Pu, H. (2022). Transfer learning for nonparametric regression: Non-asymptotic minimax analysis and adaptive procedure. Technical report, University of Pennsylvania.

Cai, T. T. and Wei, H. (2021). Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *The Annals of Statistics*, 49:100–128.

Cannings, T. I., Berrett, T. B., and Samworth, R. J. (2020a). Local nearest neighbour classification with applications to semi-supervised learning. *Annals of Statistics*, 48:1789–1814.

Cannings, T. I., Fan, Y., and Samworth, R. J. (2020b). Classification with imperfect training labels. *Biometrika*, 107(2):311–330.

Caruana, R. (1997). Multitask learning. *Machine Learning*, 28.

Celisse, A. and Mary-Huard, T. (2018). Theoretical analysis of cross-validation for estimating the risk of the k-nearest neighbor classifier. *Journal of Machine Learning Research*, 19(58):1–54.

Chen, A., Owen, A., and Shi, M. (2013). Data enriched linear regression. *Electronic Journal of Statistics*, 9.

Choi, K., Fazekas, G., Sandler, M. B., and Cho, K. (2017). Transfer learning for music classification and regression tasks. In *International Society for Music Information Retrieval Conference*.

Christiansen, R., Pfister, N., Jakobsen, M. E., Gnecco, N., and Peters, J. (2020). The difficult task of distribution generalization in nonlinear models. *arXiv: Methodology*.

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

Devroye, L., Györfi, L., and Lugosi, G. (1997). *A Probabilistic Theory of Pattern Recognition*. Stochastic Modelling and Applied Probability. Springer New York.

Fan, J. (1993). Local Linear Regression Smoothers and Their Minimax Efficiencies. *The Annals of Statistics*, 21(1):196 – 216.

Fan, J. and Gijbels, I. (1992). Variable Bandwidth and Local Linear Regression Smoothers. *The Annals of Statistics*, 20(4):2008 – 2036.

Fan, J., Li, R., Zhang, C.-H., and Zou, H. (2020). *Statistical Foundations of Data Science (1st ed.)*. CRC Press.

Frenay, B. and Verleysen, M. (2014). Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869.

Gadat, S., Klein, T., and Marteau, C. (2016). Classification in general finite dimensional spaces with the k-nearest neighbor rule. *Annals of Statistics*, 44(3):982–1009.

Germain, P., Habrard, A., Laviolette, F., and Morvant, E. (2013). A pac-bayesian approach for domain adaptation with specialization to linear classifiers. In Dasgupta, S. and McAllester, D., editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 738–746, Atlanta, Georgia, USA. PMLR.

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., and Schölkopf, B. (2008). 131Covariate Shift by Kernel Mean Matching. In *Dataset Shift in Machine Learning*. The MIT Press.

Gross, S. M. and Tibshirani, R. (2016). Data shared lasso: A novel tool to discover uplift. *Computational Statistics & Data Analysis*, 101:226–235.

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

Hall, P., Park, B. U., and Samworth, R. J. (2008). Choice of neighbor order in nearest-neighbor classification. *The Annals of Statistics*, 36(5):2135 – 2152.

Hanneke, S. and Kpotufe, S. (2019). On the value of target data in transfer learning. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Huang, J.-T., Li, J., Yu, D., Deng, L., and Gong, Y. (2013). 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.

Kpotufe, S. and Martinet, G. (2018). Marginal singularity, and the benefits of labels in covariate-shift. *The Annals of Statistics*, 49.

Lepski, O. V. (1993). Asymptotically minimax adaptive estimation. ii: Schemes without optimal adaptation: Adaptive estimators. *Theory of Probability & Its Applications*, 37:433–448.

Li, S., Cai, T. T., and Li, H. (2022). Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society: Series B*, 84:149–173.

Li, X., Grandvalet, Y., Davoine, F., Cheng, J., Cui, Y., Zhang, H., Belongie, S., Tsai, Y.-H., and Yang, M.-H. (2020). Transfer learning in computer vision tasks: Remember where you come from. *Image and Vision Computing*, 93:103853.

Maity, S., Sun, Y., and Banerjee, M. (2022). Minimax optimal approaches to the label shift problem in non-parametric settings. *Journal of Machine Learning Research*, 23:1–45.

Mammen, E. and Tsybakov, A. B. (2004). Smooth discrimination analysis. *The Annals of Statistics*, 32(5):2340 – 2341.

Mansour, Y., Mohri, M., and Rostamizadeh, A. (2009a). Domain adaptation: Learning bounds and algorithms. In *Proceedings of The 22nd Annual Conference on Learning Theory (COLT 2009)*, Montréal, Canada.

Mansour, Y., Mohri, M., and Rostamizadeh, A. (2009b). Multiple source adaptation and the rényi divergence. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, page 367–374, Arlington, Virginia, USA. AUAI Press.

Maurer, A., Pontil, M., and Romera-Paredes, B. (2016). The benefit of multitask representation learning. *J. Mach. Learn. Res.*, 17(1):2853–2884.

Natarajan, N., Dhillon, I. S., Ravikumar, P., and Tewari, A. (2018). Cost-sensitive learning with noisy labels. *Journal of Machine Learning Research*, 18(155):1–33.

Negahban, S., Ravikumar, P., Wainwright, M. J., and Yu, B. (2009). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, NIPS'09, page 1348–1356, Red Hook, NY, USA. Curran Associates Inc.

Ollier, E. and Viallon, V. (2017). Regression modelling on stratified data with the lasso. *Biometrika*, 104(1):83–96.

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

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

Raskutti, G., Wainwright, M. J., and Yu, B. (2009). Minimax rates of estimation for high-dimensional linear regression over ℓ_q -balls. *IEEE Transactions on Information Theory*, 57:6976–6994.

Reeve, H. and Kaban, A. (2019). Fast rates for a knn classifier robust to unknown asymmetric label noise. *Proceedings of Machine Learning Research*, 97:5401–5409. International Conference on Machine Learning, ICML 2019 ; Conference date: 09-06-2019 Through 15-06-2019.

Reeve, H. W. J., Cannings, T. I., and Samworth, R. J. (2021). Adaptive transfer learning. *Annals of Statistics*, 49(6):3618–3649.

Ruder, S., Peters, M. E., Swayamdipta, S., and Wolf, T. (2019). Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Scott, C., Blanchard, G., and Handy, G. (2013). Classification with asymmetric label noise: Consistency and maximal denoising. In Shalev-Shwartz, S. and Steinwart, I., editors, *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 489–511, Princeton, NJ, USA. PMLR.

Scott, C. and Zhang, J. (2019). Learning from multiple corrupted sources, with application to learning from label proportions. *CoRR*, abs/1910.04665.

Scott, C. D. (2018). A generalized neyman-pearson criterion for optimal domain adaptation. *ArXiv*, abs/1810.01545.

Sinha, A., Namkoong, H., and Duchi, J. C. (2018). Certifying some distributional robustness with principled adversarial training. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Storkey, A. (2008). 23When Training and Test Sets Are Different: Characterizing Learning Transfer. In *Dataset Shift in Machine Learning*. The MIT Press.

Tian, Y. and Feng, Y. (2022). Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association*, 0(0):1–14.

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. (2017). Adversarial discriminative domain adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2962–2971, Los Alamitos, CA, USA. IEEE Computer Society.

Wang, D. and Zheng, T. F. (2015). Transfer learning for speech and language processing. In *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1225–1237. IEEE.

Weiss, K., Khoshgoftaar, T., and Wang, D. (2016). A survey of transfer learning. *Journal of Big Data*, 3.

Zheng, C., Dasgupta, S., Xie, Y., Haris, A., and Chen, Y. Q. (2019). On data enriched logistic regression. *arXiv: Applications*.

A. Supplementary Results

A.1. General Signal Transfer Risk Bound for Plug-in Rules

Suppose that $\hat{f}^P$ can be expressed in the form of a plug-in rule, i.e. $\hat{f}^P(\cdot) = \mathbf{1}\{\hat{\eta}^P(\cdot) \geq \frac{1}{2}\}$, where $\hat{\eta}^P(\cdot)$ is an estimate of the regression function $\eta^P(\cdot)$. For notational simplicity, define the area with strong signal strength as

$$\Omega^+(\gamma, C_\gamma) := \{x \in \Omega : s(x) \geq C_\gamma |\eta^Q(x) - \frac{1}{2}|^\gamma\}.$$

Since $\Omega^+(\gamma, C_\gamma)$ indicate the informative area where η^P provides strong signal relative to η^Q, $\hat{\eta}^P$ could also give indications on η^Q if it approximates η^P well.

In order to obtain a complete convergence rate for the signal transfer risk, it is necessary to impose some conditions on the behavior of $\hat{\eta}^P$. The following result requires that $\hat{\eta}^P$ recovers η^P in the sense that the misclassification rate is upper bounded.

Theorem 8. Let $\hat{\eta}^P$ be an estimator of the regression function η^P depending on $\mathcal{D}_P$. Suppose that $\tilde{\Pi} \subset \Pi$ and two n_P-sequences δ_P, δ_P^f satisfy that, with probability at least $1 - \delta_P^f$ w.r.t. the distribution of $X_{1:n_P}^P = (X_1^P, \cdots, X_{n_P}^P)$, for some constant $C_2 > 0$,

$$\sup_{(Q,P)\in\tilde{\Pi}} \mathbb{P}_{\mathcal{D}_P}((\hat{\eta}^P(x) - \frac{1}{2})(\eta^Q(x) - \frac{1}{2}) < 0 | X_{1:n_P}^P) \leq C_2 \exp(-(\frac{C_\gamma |\eta^Q(x) - \frac{1}{2}|^\gamma}{\delta_P})^2), \tag{28}$$

holds except at $\Omega^b \subset \Omega^+(\gamma, C_\gamma)$. Then

$$\sup_{(Q,P)\in\tilde{\Pi}} \mathbb{E}_{\mathcal{D}_P} \varepsilon_T(\hat{f}^P) \lesssim \delta_P^{\frac{1+\alpha}{\gamma}} + \delta_P^f + \delta_P^b,$$

where $\delta_P^b = \sup_{(Q,P)\in\tilde{\Pi}} \int_{\Omega^b} |\eta^Q(x) - \frac{1}{2}| dQ_X$.

Theorem 8 does not explicitly require the concentration property for $\hat{\eta}^P$ around the true regression function $\hat{\eta}^P$. Instead, it bounds the misclassification rate for a given query point by $C_2 \exp(-(\frac{|\eta^P(x) - \frac{1}{2}|}{\delta_P})^2)$ which is consistent with the signal provided by η^P relative to η^Q within the signal transfer set.

Similar to δ_Q^f, the quantity δ_P^f denotes the (typically small) probability that the misclassification rate bound (28) fails. To account for the case where the query point is near the boundary of the signal transfer set $\Omega^+(\gamma, C_\gamma)$, and the estimate $\hat{\eta}^P(x)$ may be influenced by its neighborhood outside the set, we introduce a new term δ_P^b in addition to δ_P^f. This allows for the possibility of failure of the misclassification rate bound (28) near the boundary of the signal transfer set. This situation may occur in some non-parametric methods, such as K-NN models, and needs to be considered in theoretical analysis.

The following simple proposition shows that (28) is weaker than the classical exponential concentration inequality.

Proposition 2. Let $\hat{\eta}^P$ be an estimator of the regression function η^P depending on $\mathcal{D}_P$. For any $x \in \Omega^+(\gamma, C_\gamma)$, the event that for any $t > 0$,

$$\sup_{(Q,P)\in\tilde{\Pi}} \mathbb{P}_{\mathcal{D}_P}(|\hat{\eta}^P(x) - \eta^P(x)| \geq t | X_{1:n_P}^P) \leq C_2 \exp(-(\frac{t}{\delta_P})^2)$$

belongs to the event that (28) holds.

Proof. The result is straightforward by observing that

$$\mathbb{P}_{\mathcal{D}_P}((\hat{\eta}^P(x)-\frac{1}{2})(\eta^Q(x)-\frac{1}{2})<0|X_{1:n_P}^P)\le \mathbb{P}_{\mathcal{D}_P}(|\hat{\eta}^P(x)-\eta^P(x)|\ge|\eta^P(x)||X_{1:n_P}^P),$$

of which the latter term is less than $C_2\exp(-(\frac{|\eta^P(x)|}{\delta_P})^2)$. Since we require that $x\in\Omega^+(\gamma,C_\gamma)$, it is also less than $C_2\exp(-(\frac{C_\gamma|\eta^Q(x)-\frac{1}{2}|^\gamma}{\delta_P})^2)$. □

A.2. Non-parametric Classification with General Form of Ambiguity

In this section, we only focus on the general parametric space Π^{NP} under the non-parametric classification scenario. We require one constraint on the parameters α,γ,β and β_P. Assume that $\gamma\beta\ge\beta_P$ to rule out a "super-smooth" source regression function η^P. The motivation behind this assumption is that if $\gamma\beta<\beta_P$, the smoothness of η^P will potentially restrict the family of the target distribution Q. To see this, we define the family of target distributions as follows:

$$\Pi_Q^{NP}:=\{Q:Q\in\mathcal{M}(\alpha,C_\alpha),\eta^Q\in\mathcal{H}(\beta,C_\beta),Q_X\in\mathcal{S}(\mu^+,\mu^-,c_\mu,r_\mu)\},$$

where $\mathcal{M}(\alpha,C_\alpha)$ denotes the set of distributions that satisfies the margin assumption (Assumption 1) with parameters $\alpha\ge 0$ and $C_\alpha>0$. The following proposition shows that when $\gamma\beta<\beta_P$, Π_Q^{NP} is not a subset of the slice of Π^{NP} on Q when setting $\varepsilon(z)\equiv 0$.

Proposition 3. Suppose Π^{NP} satisfies that $\varepsilon(z)\equiv 0$, then

1. If $\gamma\beta\ge\beta_P$, then $\Pi_Q^{NP}=\{Q:(Q,P)\in\Pi^{NP}\}$ for $C_{\beta_P}\ge C_\gamma 2^{-\gamma}\vee 2C_\gamma C_\beta^{\frac{\beta_P}{\beta}}2^{\gamma-\beta_P/\beta}$.
2. If $\gamma\beta<\beta_P$, then $\Pi_Q^{NP}\not\subset\{Q:(Q,P)\in\Pi^{NP}\}$.

Proposition 3 implies that when β_P is sufficiently large, a too-small γ imposes an additional smoothness condition on Q. To avoid this issue, we assume that a large enough $\gamma\ge\beta_P/\beta$.

Define

$$\Omega^+(\gamma,C_\gamma):=\{x\in\Omega:s(x)\ge C_\gamma|\eta^Q(x)-\frac{1}{2}|^\gamma\}$$

as the area with strong signal strength for convenience. The upper bound results is as follows. For the sake of completeness, we also provide the upper bound when $\gamma\beta<\beta_P$, though there is no lower bound result to match the upper bound.

Theorem 9 (Non-parametric Classification Upper Bound, General Case)**.** Suppose that $n_Q^{\frac{d}{2\beta+d}}\exp(-c_Q n_Q^{\frac{2\beta}{2\beta+d}})\lesssim n_P^{-\frac{\beta(1+\alpha)}{2\gamma\beta+d}}$. Then the model-selected K-NN classifier $\hat{f}_{MS}^{NN}(x)$ satisfies that

- $\gamma\beta\ge\beta_P$: For any $(Q,P)\in\Pi^{NP}$,

$$\begin{aligned}\mathbb{E}_{(\mathcal{D}_Q,\mathcal{D}_P)}\mathcal{E}_Q(\hat{f}_{MS}^{NN})&\lesssim\left(n_P^{-\frac{\beta(1+\alpha)}{2\gamma\beta+d}}+\varepsilon_b+\varepsilon(2\tau)\right)\wedge\left(\log^{1+\alpha}(n_Q\vee n_P)n_Q^{-\frac{\beta(1+\alpha)}{2\beta+d}}\right)\\&\lesssim\left(n_P^{-\frac{\beta_P(1+\alpha)/\gamma}{2\gamma\beta+d}}+\varepsilon(2\tau)\right)\wedge\left(\log^{1+\alpha}(n_Q\vee n_P)n_Q^{-\frac{\beta(1+\alpha)}{2\beta+d}}\right)\end{aligned}\qquad(29)$$

where $\varepsilon_b = \int_{\Omega^b} |\eta^Q(X) - \frac{1}{2}| dQ_X$ and the boundary of the signal transfer set is defined as

$$\Omega^b := \{x \in \Omega : |\eta^Q(x) - \frac{1}{2}| \leq c_b n_P^{-\frac{\beta_P/\gamma}{2\gamma\beta+d}}, B(x, c_b n_P^{-\frac{1}{2\gamma\beta+d}}) \cap \Omega \cap \Omega_P \not\subset \Omega^+(\gamma, C_\gamma)\}$$

for some constant $c_b > 0$.

- $\gamma\beta < \beta_P$: If we choose $k_P = \lfloor c_P n_P^{\frac{2\beta_P}{2\beta_P+d}} \rfloor$ for any $c_P > 0$ instead, then

$$\sup_{(Q,P)\in\Pi^{NP}} \mathbb{E}_{(\mathcal{D}_Q,\mathcal{D}_P)} \mathcal{E}_Q(\hat{f}_{MS}^{NN}) \lesssim \left(n_P^{-\frac{\beta_P(1+\alpha)/\gamma}{2\beta_P+d}} + \varepsilon(2\tau) \right) \wedge \left(\log^{1+\alpha}(n_Q \vee n_P) n_Q^{-\frac{\beta(1+\alpha)}{2\beta+d}} \right) \tag{30}$$

When $\gamma\beta \geq \beta_P$, the optimal "smoothness" parameter for the source data should be $\gamma\beta$, provided by $\Omega^+(\gamma, C_\gamma)$, as our upper bound (except the boundary risk term ε_b) and the choice of k_P do not involve β_P. In other words, our upper bound automatically smooths η^P implicitly through the connection between the area with strong signal strength and η^Q.

We meanwhile provide the lower bound result for the parameter space Π^{NP}. The scenario of $\gamma\beta < \beta_P$ causes a "phase transition" transition, where the lower bound does not match the lower bound. How to develop a optimal and robust upper bound methodology under this case is still open.

Theorem 10 (Non-parametric Classification Lower Bound, General Case)**.** (a) Fix the parameters in the definition of Π^{NP} with $\alpha\beta \leq d, \gamma\beta \geq \beta_P$. We have that for some constant $c > 0$,

$$\inf_{\hat{f}} \sup_{\substack{(Q,P)\in\Pi^{NP} \\ \Omega=\Omega_P}} \mathbb{E}\mathcal{E}_Q(\hat{f}) \gtrsim \left(n_P^{-\frac{\beta(1+\alpha)}{2\gamma\beta+d}} + \varepsilon(c n_Q^{-\frac{\beta}{2\beta+d}}) \right) \wedge n_Q^{-\frac{\beta(1+\alpha)}{2\beta+d}}.$$

(b) Fix the parameters in the definition of Π^{NP} with $\max\{\alpha\beta, \alpha\beta_P/\gamma\} \leq d, \gamma\beta < \beta_P$. We have that for some constant $c > 0$,

$$\inf_{\hat{f}} \sup_{\substack{(Q,P)\in\Pi^{NP} \\ \Omega=\Omega_P}} \mathbb{E}\mathcal{E}_Q(\hat{f}) \gtrsim \left(n_P^{-\frac{\beta_P(1+\alpha)/\gamma}{2\beta_P+d}} + \varepsilon(c n_Q^{-\frac{\beta}{2\beta+d}}) \right) \wedge n_Q^{-\frac{\beta_P(1+\alpha)/\gamma}{2\beta_P/\gamma+d}}.$$

Theorem 10 indicates that when $\gamma\beta \geq \beta_P$, as long as the risk term ε_b is dominated in that

$$\varepsilon_b \lesssim n_P^{-\frac{\beta(1+\alpha)}{2\gamma\beta+d}} + \varepsilon(2\tau), \tag{31}$$

our obtained excess risk upper bound is optimal up to some logarithmic factors. This additional risk term ε_b can be attributed to the set of points whose neighboring source data tend to fall outside $\Omega^+(\gamma, C_\gamma)$. We claim that if a point x belonging to $\Omega^+(\gamma, C_\gamma)$ is in close proximity to the boundary of $\Omega^+(\gamma, C_\gamma)$, that is,

$$B(x, r_P) \cap \Omega_P \not\subset \Omega^+(\gamma, C_\gamma)$$

for some small $r_p \asymp n_P^{-\frac{1}{2\gamma\beta+d}}$, it may be difficult to accurately classify this point using $\hat{\eta}^P$. This difficulty arises due to the tendency of the neighboring source data of x to fall outside the signal transfer set. As a result, the small neighborhood of x might not provide enough information to correctly classify $\eta^P(x)$, leading to inaccurate classification. Giving up the correct classification of these points generate such risk term ε_b. We also take the smoothness parameter β_P into consideration since it influences the smoothness of η^P, and thus controls the distortedness of the boundary of the signal transfer set.

A trivial asymptotic upper bound of ε_b is $n_P^{-\frac{\beta_P(1+\alpha)/\gamma}{2\gamma\beta+d}}$. Unfortunately, this trivial bound strictly dominates the risk term $n_P^{-\frac{\beta(1+\alpha)}{2\gamma\beta+d}}$ when $\beta_P < \gamma\beta$. Hence, we need additional conditions on the source distribution to ensure a smaller ε_b.

From the proof of Theorem 3, we show that those three listed cases satisfy (31). The following corollary lists two more special cases to help readers to better understand the upper bound result we obtained. In all of these cases, we can explicitly express ε_b and verify that it satisfies (31), which shows that our upper bound achieves the optimal rate up to logarithmic factors.

Corollary 1. Keep the notations and parameter choices in Theorem 9. Suppose that $\gamma\beta \geq \beta_P$. Consider the following additional conditions added to $(Q, P) \in \Pi^{NP}$, respectively:

1. **Signal Transfer Boundary as part of the Decision Boundary:** If η^Q, η^P are continuous, and
$$\partial\Omega^+(\gamma, C_\gamma) \subset \{x \in \Omega : \eta^Q(x) = \frac{1}{2}\},$$
then $\varepsilon_b \lesssim n_P^{-\frac{\beta(1+\alpha)}{2\gamma\beta+d}}$, and
$$\mathbb{E}_{(\mathcal{D}_Q,\mathcal{D}_P)}\mathcal{E}_Q(\hat{f}^{NN}_{MS}) \lesssim n_P^{-\frac{\beta(1+\alpha)}{2\gamma\beta+d}} \wedge \left(\log^{1+\alpha}(n_Q \vee n_P) n_Q^{-\frac{\beta(1+\alpha)}{2\beta+d}}\right) + \varepsilon(2\tau; \gamma, C_\gamma).$$

2. **Signal Transfer Boundary Margin:** If for any $r > 0$, there exists some constant $\alpha_r \geq 0, r_m > 0, C_r > 0$ such that for any $0 < r < r_m$,
$$Q_X(B(X, r) \cap \Omega \cap \Omega_P \not\subset \Omega^+(\gamma, C_\gamma)) \leq C_r r^{\alpha_r}.$$
Specially, we set $\alpha_r = \infty$ when $\Omega^+(\gamma, C_\gamma)$ can be separated with its compliment set in the sense that
$$\inf\{||x - y|| : x \in \Omega^+(\gamma, C_\gamma), y \in \Omega/\Omega^+(\gamma, C_\gamma)\} \geq r_m$$
for some $r_m > 0$. Then if $\beta_P/\gamma + \alpha_r \geq \beta(1 + \alpha)$, we have $\varepsilon_b \lesssim n_P^{-\frac{\beta(1+\alpha)}{2\gamma\beta+d}}$, and
$$\mathbb{E}_{(\mathcal{D}_Q,\mathcal{D}_P)}\mathcal{E}_Q(\hat{f}^{NN}_{MS}) \lesssim n_P^{-\frac{\beta(1+\alpha)}{2\gamma\beta+d}} \wedge \left(\log^{1+\alpha}(n_Q \vee n_P) n_Q^{-\frac{\beta(1+\alpha)}{2\beta+d}}\right) + \varepsilon(2\tau; \gamma, C_\gamma).$$

B. Proofs of General Convergence Theorems

In this section, we will prove Theorem 1, Theorem 2 and Theorem 7. Suppose $\hat{f}^P$ is a classifier based on $\mathcal{D}_P$ and $\tilde{\Pi} \subset \Pi$. Since by Definition 3, an trivial bound of the restricted signal transfer risk is
$$\varepsilon_T(f, z; \gamma, C_\gamma) \leq \varepsilon_T(f; \gamma, C_\gamma) \wedge (C_\alpha z^{1+\alpha}),$$

it suffices to prove Theorem 7 (2) to show Theorem 2.

We list some additional notations used in the rest of the appendix part. For any set $A \subset \Omega$, define $A^c = \Omega/A$ as the compliment. For any set B, define $|B|$ as its cardinality. Define

$$\Omega^+(\gamma, C_\gamma) := \{x \in \Omega : s(x) \geq C_\gamma|\eta^Q(x) - \frac{1}{2}|^\gamma\},$$

and

$$\Omega^-(\gamma, C_\gamma) := \{x \in \Omega : s(x) < C_\gamma|\eta^Q(x) - \frac{1}{2}|^\gamma\} = (\Omega^+(\gamma, C_\gamma))^c.$$

We abbreviate $(\mathcal{D}_Q, \mathcal{D}_P)$, the combination of the target and source data, as $\mathcal{D}$. Denote the expectation w.r.t. a new target observation, i.e. $\mathbb{E}_{(X,Y)\sim Q}$, by $\mathbb{E}$.

B.1. Proofs of Theorem 2 and Theorem 7 (2)

Proof. Define the following events:

$$E_0 := \{|\eta^Q(X) - \frac{1}{2}| \geq 2\tau\}, \quad E^* := \{X \in \Omega^*\}, \quad E := \{|\hat{\eta}^Q(X) - \eta^Q(X)| \leq \tau\}.$$

For any $(Q, P) \in \tilde{\Pi}$, denote the event on which $\eta^Q(X)$ is far from $\frac{1}{2}$ by 2τ, and the event on which the concentration property holds. Note that E_0 only depends on Q, but E^* and E depends on both Q and $\mathbb{E}_{\mathcal{D}}$. The expected excess risk could be reformulated by the law of total expectation considering $E^c, E_0 \cap E$ and $E_0^c \cap E$:

$$\begin{aligned}\mathcal{E}_Q(\hat{f}_{MS}) = 2Q(E^c)\mathbb{E}[|\eta_Q(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}_{MS}(X) \neq f_Q^*(X)\}|E^c] \quad (32)\\ + 2Q(E_0 \cap E)\mathbb{E}[|\eta_Q(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}_{MS}(X) \neq f_Q^*(X)\}|E_0 \cap E] \quad (33)\\ + 2Q(E_0^c \cap E)\mathbb{E}[|\eta_Q(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}_{MS}(X) \neq f_Q^*(X)\}|E_0^c \cap E]. \quad (34)\end{aligned}$$

Next, we will bound the expectation of the three components (32), (33), (34) with respect to $\mathbb{E}_{\mathcal{D}}$, respectively.

Firstly, a trivial upper bound is

$$2\mathbb{E}[|\eta_Q(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}_{MS}(X) \neq f_Q^*(X)\}|E^c] \leq 1,$$

which deduces that

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}}[2Q(E^c)\mathbb{E}[|\eta_Q(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}_{MS}(X) \neq f_Q^*(X)\}|E^c]]\\ \leq&\mathbb{E}_{\mathcal{D}}[Q(E^c)] = Q(\mathbb{E}_{\mathcal{D}}[E^c)]) \leq 2\delta_Q(n_Q, \tau),\end{aligned} \quad (35)$$

To see the last inequality of (35), the condition of Theorem 2 implies that

$$Q((E^*)^c) = Q(\Omega/\Omega^*) \leq \delta_Q(n_Q, \tau) \Rightarrow \mathbb{E}_{\mathcal{D}}[Q((E^*)^c)] \leq \delta_Q(n_Q, \tau).$$

Plus, the concentration property on Ω^* implies that

$$\mathbb{E}_{\mathcal{D}}[E^c|E^*] \leq \sup_{X\in\Omega^*} \mathbb{E}_{\mathcal{D}}[E^c|X = x] \leq \delta_Q(n_Q, \tau),$$

so by the law of total expectation, we have

$$\begin{aligned}Q(\mathbb{E}_{\mathcal{D}}[Q(E^c)]) &\leq Q(\mathbb{E}_{\mathcal{D}}[E^c|E^*]\mathbb{E}_{\mathcal{D}}[E^*] + \mathbb{E}_{\mathcal{D}}[E^c|(E^*)^c]\mathbb{E}_{\mathcal{D}}[(E^*)^c]) \\ &= Q(\mathbb{E}_{\mathcal{D}}[E^c|E^*]\mathbb{E}_{\mathcal{D}}[E^*]) + Q(\mathbb{E}_{\mathcal{D}}[E^c|(E^*)^c]\mathbb{E}_{\mathcal{D}}[(E^*)^c]) \\ &\leq \delta(n_Q,\tau)Q(\mathbb{E}_{\mathcal{D}}[E^*]) + Q(\mathbb{E}_{\mathcal{D}}[(E^*)^c]) \\ &\leq \delta(n_Q,\tau) + \mathbb{E}_{\mathcal{D}}[Q((E^*)^c)] \\ &\leq 2\delta(n_Q,\tau).\end{aligned}$$

Next, on the event $E_0 \cap E$, we observe that

$$(\hat{\eta}^Q(X) - \frac{1}{2})(\eta^Q(X) - \frac{1}{2}) \geq 0, |\hat{\eta}^Q(X) - \frac{1}{2}| \geq \tau.$$

This fact tells us $\mathbf{1}\{\hat{f}_{MS}(X) \neq f_Q^*(X)\} = 0$. Hence, we have that

$$\mathbb{E}_{\mathcal{D}}[2Q(E_0 \cap E)\mathbb{E}[|\eta_Q(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}_{MS}(X) \neq f_Q^*(X)\}|E_0 \cap E]] = 0 \tag{36}$$

Lastly, let's upper bound the expectation of (34) with respect to $\mathbb{E}_{\mathcal{D}}$. Without loss of generality, suppose that $\eta^Q(X) \neq \frac{1}{2}$, or otherwise the upper bound can be set as just 0. Then, we observe that on the event $E_0^c \cap E$,

$$(\hat{\eta}^Q - \frac{1}{2})(\eta^Q - \frac{1}{2}) < 0 \Rightarrow |\hat{\eta}^Q(X) - \frac{1}{2}| < \tau.$$

This is because, if the plug-in rule $\mathbf{1}\{\hat{\eta}^Q(X) \geq \frac{1}{2}\}$ is different from the Bayes classifier on X, the concentration bound of τ guarantees that

$$|\hat{\eta}^Q(X) - \frac{1}{2}| \leq |\hat{\eta}^Q(X) - \eta^Q(X)| < \tau.$$

Therefore, on the event $E_0^c \cap E$, the indicator of wrong classification could be bounded by

$$\mathbf{1}\{\hat{f}_{MS}(X) \neq f_Q^*(X)\} \leq \mathbf{1}\{\hat{f}^P(X) \neq f_Q^*(X)\},$$

and it holds that

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}}[2Q(E_0^c \cap E)\mathbb{E}[|\eta_Q(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}_{MS}(X) \neq f_Q^*(X)\}|E_0^c \cap E]] \\ =&\mathbb{E}_{\mathcal{D}}[2\mathbb{E}[|\eta_Q(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}_{MS}(X) \neq f_Q^*(X)\}\mathbf{1}\{X \in E_0^c \cap E\}] \\ \leq&\mathbb{E}_{\mathcal{D}}[2\mathbb{E}[|\eta_Q(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}^P(X) \neq f_Q^*(X)\}\mathbf{1}\{X \in E_0^c\}] \\ =&\mathbb{E}_{\mathcal{D}_P}[2\mathbb{E}[|\eta_Q(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}^P(X) \neq f_Q^*(X)\}\mathbf{1}\{X \in E_0^c\}],\end{aligned}$$

where the last equality is due to the fact that the random variable

$$|\eta_Q(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}^P(X) \neq f_Q^*(X)\}\mathbf{1}\{X \in E_0^c\}$$

does not depend on $\mathcal{D}_Q$.

Define $E^+ := E_0^c \cap \{X \in \Omega^+(\gamma, C_\gamma)\}$ and $E^- := E_0^c \cap \{X \in \Omega^-(\gamma, C_\gamma)\}$. Applying the law of total expectation again on E^+ and E^- as a partition of E_0^c, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_P}[2\mathbb{E}[|\eta_Q(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}^P(X) \neq f_Q^*(X)\}\mathbf{1}\{X \in E_0^c\}] \\
\leq &\mathbb{E}_{\mathcal{D}_P}[2\mathbb{E}[|\eta_Q(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}^P(X) \neq f_Q^*(X)\}\mathbf{1}\{X \in E^+\}] \\
+ &\mathbb{E}_{\mathcal{D}_P}[2\mathbb{E}[|\eta_Q(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}^P(X) \neq f_Q^*(X)\}\mathbf{1}\{X \in E^-\}] \\
= &2\mathbb{E}_{\mathcal{D}_P}\int_{\Omega^+(\gamma,C_\gamma)} |\eta^Q(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}^P(X) \neq f_Q^*(X), |\eta^Q(X) - \frac{1}{2}| \leq 2\tau\}dQ_X \\
+ &\mathbb{E}_{\mathcal{D}_P}[2\mathbb{E}[|\eta_Q(X) - \frac{1}{2}|\mathbf{1}\{X \in E^-\}].
\end{aligned}$$

The term $2\mathbb{E}_{\mathcal{D}_P}\int_{\Omega^+(\gamma,C_\gamma)} |\eta^Q(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}^P(X) \neq f_Q^*(X), |\eta^Q(X) - \frac{1}{2}| \leq 2\tau\}dQ_X$ is upper bounded by the restricted signal transfer risk $2\sup_{(Q,P)\in\tilde{\Pi}} \mathbb{E}_{\mathcal{D}_P}\varepsilon_T(\hat{f}^P, 2\tau; \gamma, C_\gamma)$ by the straightforward definition. Moreover, since E^- does not depend on $\mathcal{D}_P$, it holds that

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_P}[2\mathbb{E}[|\eta_Q(X) - \frac{1}{2}|\mathbf{1}\{X \in E^-\}] \\
= &2\int_{\Omega^-(\gamma,C_\gamma)} |\eta^Q(X) - \frac{1}{2}|\mathbf{1}\{0 < |\eta^Q(X) - \frac{1}{2}| < 2\tau\}dQ_X \leq 2\varepsilon(2\tau; \gamma, C_\gamma).
\end{aligned}$$

by the definition of the ambiguity level. Hence, the following bound holds:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}[2Q(E_0^c \cap E)\mathbb{E}[|\eta_Q(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}_{MS}(X) \neq f_Q^*(X)\}|E_0^c \cap E]] \\
\leq &2 \sup_{(Q,P)\in\tilde{\Pi}} \mathbb{E}_{\mathcal{D}_P}\varepsilon_T(\hat{f}^P, 2\tau; \gamma, C_\gamma) + 2\varepsilon(2\tau; \gamma, C_\gamma).
\end{aligned} \tag{37}$$

It is easy to see that Theorem 7 (2) holds by bounding the three components (32), (33), (34) by (35), (36), (37), respectively. This result finishes Theorem 2 as well. □

B.2. Proofs of Proofs of Theorem 1 and Theorem 7 (1)

Suppose that $(Q, P) \in \tilde{\Pi}$. By plugging in $t = \tau$ in (9), we have that with probability at least $1 - \delta_Q^f$ w.r.t. the distribution of $X_{1:n_Q}$, we have for any $x \in \Omega$ we have

$$\mathbb{P}_{\mathcal{D}_Q}(|\hat{\eta}^Q(x) - \eta^Q(x)| \geq \tau | X_{1:n_Q}) \lesssim \exp(-(\frac{\tau}{\delta_Q})^2). \tag{38}$$

In other words, by taking the probability of which the concentration (38) fails into account, for any $x \in \Omega$, the equation

$$|\hat{\eta}^Q(x) - \eta^Q(x)| \leq \tau$$

holds with probability that is asymptotically less than $\delta_Q^f + \exp(-(\frac{\tau}{\delta_Q})^2)$ with respect to $\mathbb{P}_{\mathcal{D}_Q}$. We fix $\Omega^* = \Omega$ in the terminology of Theorem 2 and 7 (2).

Proof of Theorem 7 (1). Suppose that for some $c_0 > 0$, the inequality $\tau \geq c_0 \log(n_Q \vee n_P)\delta_Q$ always holds. We have that

$$\begin{aligned}
\exp(-(\frac{\tau}{\delta_Q})^2) &\leq \exp(-c_0^2 \log^2(n_Q \vee n_P)) \\
&= n_Q^{-c_0^2 \log n_Q} \wedge n_P^{-c_0^2 \log n_P} \\
&\lesssim n_Q^{-c} \wedge n_P^{-c} \\
&\lesssim \sup_{(Q,P)\in\tilde{\Pi}} \mathbb{E}_{\mathcal{D}_P} \varepsilon_T(\hat{f}^P, 2\delta_Q; \gamma, C_\gamma),
\end{aligned}$$

Therefore, the conditions of Theorem 7 (2) holds by setting

$$\delta_Q(n_Q, \tau) = C(\delta_Q^f + \sup_{(Q,P)\in\tilde{\Pi}} \mathbb{E}_{\mathcal{D}_P} \varepsilon_T(\hat{f}^P, 2\delta_Q; \gamma, C_\gamma)),$$

for some $C > 0$ that is large enough. Theorem 7 (1) thus holds. □

Proof of Theorem 1. We follow a similar approach as given in the proof of Theorem 7 (1). Suppose that for some $c_0 > 0$, the inequality $\tau \geq c_0 \log(n_Q \vee n_P)\delta_Q$ always holds. We have that

$$\begin{aligned}
\exp(-(\frac{\tau}{\delta_Q})^2) &\leq \exp(-c_0^2 \log^2(n_Q \vee n_P)) \\
&= n_Q^{-c_0^2 \log n_Q} \wedge n_P^{-c_0^2 \log n_P} \\
&\lesssim n_Q^{-c} \wedge n_P^{-c} \\
&\lesssim \sup_{(Q,P)\in\tilde{\Pi}} \mathbb{E}_{\mathcal{D}_P} \varepsilon_T(\hat{f}^P; \gamma, C_\gamma) \wedge \delta_Q^{1+\alpha},
\end{aligned}$$

Therefore, the conditions of Theorem 2 holds by setting

$$\delta_Q(n_Q, \tau) = C(\delta_Q^f + \sup_{(Q,P)\in\tilde{\Pi}} \mathbb{E}_{\mathcal{D}_P} \varepsilon_T(\hat{f}^P; \gamma, C_\gamma) + \delta_Q^{1+\alpha})$$

for some $C > 0$ that is large enough. Theorem 1 thus holds. □

C. Proofs in Non-parametric Classification

In this section, we only consider the case of $(Q, P) \in \Pi^{NP}$. Since Π^{NP} assumes the compactness of the support sets Ω and Ω_P, we assume for simplicity that Ω, Ω_P is a subset of the d-dimension unit square $[0, 1]^d$.

Our proof of Lemma C.2 partially relies on verifying the conditions in Theorem 8, presented in Appendix A.1. Therefore, we will provide the proof of Theorem 8 first in this section.

C.1. Proof of Theorem 8

Proof. Denote the distribution of $X_{1:n_P}^P$ by $\mathbb{P}_{X_{1:n_P}^P}$, and $\Omega^*(\gamma, C_\gamma) := \Omega^+(\gamma, C_\gamma)/\Omega^b$. Suppose $(Q, P) \in \tilde{\Pi}$. It is easy to see that there exists an event E_P^f related to the distribution of $X_{1:n_P}^P$ such that

- $\mathbb{P}_{X^P_{1:n_P}}(E^f_P) \le \delta^f_P$.
- On the event $(E^f_P)^c$, we have for any $x \in \Omega^*(\gamma, C_\gamma)$,

$$\mathbb{P}_{\mathcal{D}_P}((\hat{\eta}^P(x) - \frac{1}{2})(\eta^Q(x) - \frac{1}{2}) < 0 | X^P_{1:n_P}) \le C_2 \exp(-(\frac{C_\gamma |\eta^Q(x) - \frac{1}{2}|^\gamma}{\delta_P})^2).$$

Define $E^* = \{X \in \Omega^*(\gamma, C_\gamma)\}$. We apply the law of total expectation on

$$\mathbb{E}_{\mathcal{D}_P}[\mathbb{E}[|\eta^Q(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}^P(X) \neq f^*_Q(X)\}\mathbf{1}\{X \in \Omega^+(\gamma, C_\gamma)\}]]$$

by decomposing $\{X \in \Omega^+(\gamma, C_\gamma)\}$ into $E^f_P, (E^f_P)^c \cap E^*$ and $(E^f_P)^c \cap (E^*)^c$. The decomposition reads

$$\mathbb{E}_{\mathcal{D}_P}[\mathbb{E}[|\eta^Q(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}^P(X) \neq f^*_Q(X)\}\mathbf{1}\{X \in \Omega^+(\gamma, C_\gamma)\}]] \tag{39}$$

$$= \mathbb{P}_{X^P_{1:n_P}}(E^f_P)\mathbb{E}_{\mathcal{D}_P}[\mathbb{E}[|\eta^Q(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}^P(X) \neq f^*_Q(X)\}]|E^f_P] \tag{40}$$

$$+ \mathbb{P}_{X^P_{1:n_P}}((E^f_P)^c)\mathbb{E}_{\mathcal{D}_P}[\mathbb{E}[|\eta^Q(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}^P(X) \neq f^*_Q(X)\}\mathbf{1}\{X \in \Omega^*(\gamma, C_\gamma)\}]|(E^f_P)^c] \tag{41}$$

$$+ \mathbb{P}_{X^P_{1:n_P}}((E^f_P)^c)\mathbb{E}_{\mathcal{D}_P}[\mathbb{E}[|\eta^Q(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}^P(X) \neq f^*_Q(X)\}\mathbf{1}\{X \in \Omega^b\}]|(E^f_P)^c] \tag{42}$$

We will then bound the three components (40), (41), (42) respectively. Firstly, we observe that (40) could be bounded by

$$\mathbb{P}_{X^P_{1:n_P}}(E^f_P)\mathbb{E}_{\mathcal{D}_P}[\mathbb{E}[\eta^Q(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}^P(X) \neq f^*_Q(X)\}]|E^f_P] \le \frac{1}{2}\mathbb{P}_{X^P_{1:n_P}}(E^f_P) \le \frac{1}{2}\delta^f_P. \tag{43}$$

Secondly, (41) could be bounded by

$$\begin{aligned}
&\mathbb{P}_{X^P_{1:n_P}}((E^f_P)^c)\mathbb{E}_{\mathcal{D}_P}[\mathbb{E}[|\eta^Q(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}^P(X) \neq f^*_Q(X)\}\mathbf{1}\{X \in \Omega^b\}]|(E^f_P)^c] \\
\le& \mathbb{E}_{\mathcal{D}_P}[\mathbb{E}[|\eta^Q(X) - \frac{1}{2}|\mathbf{1}\{X \in \Omega^b\}\mathbf{1}\{(E^f_P)^c \text{ holds}\}]] \\
\le& \mathbb{E}_{\mathcal{D}_P}[\mathbb{E}[|\eta^Q(X) - \frac{1}{2}|\mathbf{1}\{X \in \Omega^b\}]] \\
=& \mathbb{E}[|\eta^Q(X) - \frac{1}{2}|\mathbf{1}\{X \in \Omega^b\}] \le \delta^b_P,
\end{aligned} \tag{44}$$

where the last equality holds as $\mathbb{E}[\eta^Q(X) - \frac{1}{2}|\mathbf{1}\{X \in \Omega^b\}]$ does not depend on $\mathcal{D}_P$. By plugging in (43) and (44) back into (40) and (42), we see that to finish the proof, it suffices show that

$$\mathbb{E}_{\mathcal{D}_P}[\mathbb{E}[|\eta^Q(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}^P(X) \neq f^*_Q(X)\}\mathbf{1}\{X \in \Omega^*(\gamma, C_\gamma)\}]|(E^f_P)^c] \lesssim \delta_P^{\frac{1+\alpha}{\gamma}}. \tag{45}$$

Consider a partition of $\Omega^*(\gamma, C_\gamma)$, which are $A_j \subset \Omega^*(\gamma, C_\gamma), j = 0, 1, 2, \cdots$ defined as

$$A_0 := \{x \in \Omega^*(\gamma, C_\gamma) : 0 < |\eta^Q(x) - \frac{1}{2}| \le \delta_P^{1/\gamma}\},$$

$$A_j := \{x \in \Omega^*(\gamma, C_\gamma) : 2^{j-1}\delta_P^{1/\gamma} < |\eta^Q(x) - \frac{1}{2}| \leq 2^j \delta_P^{1/\gamma}\},\ j \geq 1.$$

We have that

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_P}[\mathbb{E}[|\eta^Q(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}^P(X) \neq f_Q^*(X)\}\mathbf{1}\{X \in \Omega^*(\gamma, C_\gamma)\}]|(E_P^f)^c] \\
=&\sum_{j \geq 0} \mathbb{E}_{\mathcal{D}_P}[\mathbb{E}[|\eta^Q(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}^P(X) \neq f_Q^*(X)\}\mathbf{1}\{X \in A_j\}]|(E_P^f)^c].
\end{aligned}$$

For $j = 0$, by the margin assumption, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_P}[\mathbb{E}[|\eta^Q(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}^P(X) \neq f_Q^*(X)\}\mathbf{1}\{X \in A_0\}]|(E_P^f)^c] \\
\leq&\mathbb{E}[|\eta^Q(X) - \frac{1}{2}|\mathbf{1}\{X \in A_0\}] \leq \delta_P^{1/\gamma} Q(X \in A_0) \leq C_\alpha \delta_P^{\frac{1+\alpha}{\gamma}}.
\end{aligned} \tag{46}$$

For $j \geq 1$, under the event $X \in A_j$, it holds by (28) that

$$2^{j-1}\delta_P^{1/\gamma} < |\eta^Q(X) - \frac{1}{2}| \leq 2^j \delta_P^{1/\gamma},$$

$$\mathbb{P}_{\mathcal{D}_P}(\hat{f}^P(X) \neq f_Q^*(X)|(E_P^f)^c) = \mathbb{P}_{\mathcal{D}_P}((\hat{\eta}^P(x) - \frac{1}{2})(\eta^Q - \frac{1}{2}) \geq 0|(E_P^f)^c) \leq C\exp(-C_\gamma^2 2^{2\gamma(j-1)})$$

for some constant $C > 0$. Therefore,

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_P}[\mathbb{E}[|\eta^Q(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}^P(X) \neq f_Q^*(X)\}\mathbf{1}\{X \in A_j\}]|(E_P^f)^c] \\
\leq&\mathbb{E}_{\mathcal{D}_P}[\mathbb{E}[2^j\delta_P^{1/\gamma}\mathbf{1}\{\hat{f}^P(X) \neq f_Q^*(X)\}\mathbf{1}\{X \in A_j\}]|(E_P^f)^c] \\
=&\mathbb{E}[2^j\delta_P^{1/\gamma}\mathbb{E}_{\mathcal{D}_P}[\mathbf{1}\{\hat{f}^P(X) \neq f_Q^*(X)\}|(E_P^f)^c]\mathbf{1}\{X \in A_j\}] \\
\leq&C\mathbb{E}[2^j\delta_P^{1/\gamma}\exp(-C_\gamma^2 2^{2\gamma(j-1)})\mathbf{1}\{X \in A_j\}] \\
\leq&C2^j\delta_P^{1/\gamma}\exp(-C_\gamma^2 2^{2\gamma(j-1)})Q(X \in A_j) \\
\leq&C2^j\delta_P^{1/\gamma}\exp(-C_\gamma^2 2^{2\gamma(j-1)})2^{\alpha j}\delta^{\frac{\alpha}{\gamma}} = C2^{(1+\alpha)j}\exp(-C_\gamma^2 2^{2\gamma(j-1)})\delta_P^{\frac{1+\alpha}{\gamma}}.
\end{aligned} \tag{47}$$

where the last inequality is derived by the margin assumption.

Summing all terms above in (46) and (47), we have that

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_P}[\mathbb{E}[|\eta^Q(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}^P(X) \neq f_Q^*(X)\}\mathbf{1}\{X \in \Omega^*(\gamma, C_\gamma)\}]|(E_P^f)^c] \\
\leq&\delta_P^{\frac{1+\alpha}{\gamma}}(C_\alpha + \sum_{j \geq 1} C2^{(1+\alpha)j}\exp(-C_\gamma^2 2^{2\gamma(j-1)})) \lesssim \delta_P^{\frac{1+\alpha}{\gamma}},
\end{aligned}$$

since $\sum_{j \geq 1} C2^{(1+\alpha)j}\exp(-C_\gamma^2 2^{2\gamma(j-1)}) < \infty$. Therefore, (45) holds and we finish the proof. □

C.2. Proof of Theorem 9

Next, we will prove the excess risk upper bound of the model-selected K-NN classifier. Let $\beta_P^* = \max\{\gamma\beta, \beta_P\}$. Define $\Omega^*(\gamma, C_\gamma) := \Omega^+(\gamma, C_\gamma)/\Omega^b$, and $\delta_P := n_P^{-\frac{\beta_P^*}{2\beta_P^*+d}}$. To simplify the analysis, we assume that n_Q is large enough so that

$$\frac{1}{2}c_Q n_Q^{\frac{2\beta}{2\beta+d}} \leq k_Q \leq c_Q n_Q^{\frac{2\beta}{2\beta+d}}, \quad \frac{1}{2}c_P n_P^{\frac{2\beta_P^*}{2\beta_P^*+d}} \leq k_P \leq c_P n_P^{\frac{2\beta_P^*}{2\beta_P^*+d}}, \tag{48}$$

where $k_Q = \lfloor c_Q n_Q^{\frac{2\beta}{2\beta+d}} \rfloor$ and $k_P = \lfloor c_P n_P^{\frac{2\beta_P^*}{2\beta_P^*+d}} \rfloor$ are the chosen number of nearest neighbors in $\mathcal{D}_Q$ and $\mathcal{D}_P$, respectively, and c_Q and c_P are constants. This assumption is valid since our results are shown in the asymptotic regime, and we assume that $n_P \overset{n_Q \to \infty}{\longrightarrow} \infty$.

Given the condition of (48), the K-NN distance bound (Lemma E.3) shows that there exists a constant $c_D > 0$ such that with probability at least $1 - \frac{c_D}{c_Q} n_Q^{\frac{d}{2\beta+d}} \exp(-c_Q n_Q^{\frac{2\beta}{2\beta+d}})$ w.r.t. the distribution of $X_{1:n_Q}$, we have

$$||X_{(k_Q)}(x) - x|| \leq c_D (\frac{k_Q}{n_Q})^{\frac{1}{d}} \leq c_D c_Q^{\frac{1}{d}} n_Q^{-\frac{1}{2\beta+d}}, \quad \forall x \in \Omega. \tag{49}$$

Plus, with probability at least $1 - \frac{c_D}{c_P} n_P^{\frac{d}{2\beta_P^*+d}} \exp(-c_P n_P^{\frac{2\beta_P^*}{2\beta_P^*+d}})$ w.r.t. the distribution of $X_{1:n_P}^P$, we have

$$||X_{(k_P)}^P(x) - x|| \leq c_D (\frac{k_P}{n_P})^{\frac{1}{d}} \leq c_D c_P^{\frac{1}{d}} n_P^{-\frac{1}{2\beta_P^*+d}}, \quad \forall x \in \Omega_P. \tag{50}$$

Denote E_Q and E_P as the event that (49) and (50) hold, respectively. Note that E_Q and E_P depend on the distribution of $X_{1:n_Q}$ and $X_{1:n_P}^P$ respectively. From the probability bound, we have

$$\mathbb{P}_{\mathcal{D}_Q}(E_Q^c) \leq \frac{c_D}{c_Q} n_Q^{\frac{d}{2\beta+d}} \exp(-c_Q n_Q^{\frac{2\beta}{2\beta+d}}), \quad \mathbb{P}_{\mathcal{D}_P}(E_P^c) \leq \frac{c_D}{c_P} n_P^{\frac{d}{2\beta_P^*+d}} \exp(-c_Q n_P^{\frac{2\beta_P^*}{2\beta_P^*+d}}).$$

The proof of Theorem 9 is obtained simply by verifying the conditions in Theorem 1. The next two lemmas aim to prove such conditions:

Lemma C.1. By choosing $k_Q = \lfloor c_Q n_Q^{\frac{2\beta}{2\beta+d}} \rfloor$ for any constant $c_Q > 0$, the regression function estimates $\hat{\eta}_{k_Q}^Q(x)$ satisfies that, there exists constants $c_1, c_2, c_3 > 0$ such that with probability at least $1 - c_1 n_Q^{\frac{d}{2\beta+d}} \exp(-c_Q n_Q^{\frac{2\beta}{2\beta+d}})$ w.r.t. the distribution of $X_{1:n_Q} := (X_1, \cdots, X_{n_Q})$, for any $x \in \Omega$ we have $\forall t > 0$,

$$\sup_{(Q,P)\in\Pi^{NP}} \mathbb{P}_{\mathcal{D}_Q}(|\hat{\eta}^Q(x) - \eta^Q(x)| \geq t | X_{1:n_Q}) \leq c_2 \exp(-(\frac{t}{c_3 n_Q^{-\frac{\beta}{2\beta+d}}})^2). \tag{51}$$

Lemma C.2. Let $\beta_P^* = \max\{\gamma\beta, \beta_P\}$. By choosing $k_P = \lfloor c_P n_P^{\frac{2\beta_P^*}{2\beta_P^*+d}} \rfloor$ for any constant $c_P > 0$, the regression function estimates $\hat{\eta}_{k_P}^P(x)$ satisfies that, there exists constants $c_b > 0$ such that for any $\Pi^{NP,s} \subset \Pi^{NP}$, we have

$$\varepsilon_T(\hat{f}^P; \gamma, C_\gamma, \Pi^{NP,s}) \lesssim n_P^{-\frac{\beta_P^*(1+\alpha)/\gamma}{2\beta_P^*+d}} + \varepsilon_b,$$
$$\varepsilon_b = \sup_{(Q,P)\in\Pi^{NP,s}} \int_{\Omega^b} |\eta^Q(X) - \frac{1}{2}| dQ_X$$

where the boundary of the signal transfer set is defined as

$$\Omega^b := \{x \in \Omega : |\eta^Q(x) - \frac{1}{2}| \leq c_b n_P^{-\frac{\beta_P/\gamma}{2\beta_P^*+d}}, B(x, c_b n_P^{-\frac{1}{2\beta_P^*+d}}) \cap \Omega_P \not\subset \Omega^+(\gamma, C_\gamma)\}$$

for some constant $c_b > 0$.

Proof of Lemma C.1. Define the simple average of the regression functions of the k_Q neighbors of x as

$$\bar{\eta}^Q_{k_Q}(x) := \frac{1}{k_Q}\sum_{i=1}^{k_Q} \eta^Q(Y_{(i)}(x)).$$

It is easy to see that $\bar{\eta}^Q_{k_Q}(x)$ depends on the distribution of $X_{1:n_Q}$, and the conditional expectation

$$\mathbb{E}_{\mathcal{D}_Q}[\hat{\eta}^Q_{k_Q}(x)|X_{1:n_Q}] = \bar{\eta}^Q_{k_Q}(x).$$

Recall that Condition 1 states that for any $x, y \in \mathbb{R}^d$, we have

$$|\eta^Q(x) - \eta^Q(y)| \leq C_\beta ||x - y||^\beta.$$

Therefore, on the event E_Q we have

$$\begin{aligned}
|\bar{\eta}^Q_{k_Q}(x) - \eta^Q(x)| &\leq \frac{1}{k_Q}\sum_{i=1}^{k_Q} |\eta^Q(Y_{(i)}(x)) - \eta^Q(x)| \\
&\leq C_\beta \frac{1}{k_Q}\sum_{i=1}^{k_Q} ||X_{(i)}(x) - x||^\beta \\
&\leq C_\beta c_D^\beta c_Q^{\frac{\beta}{d}} n_Q^{-\frac{\beta}{2\beta+d}}.
\end{aligned}$$

On the other hand, conditioning on $X_{1:n_Q}$, the quantities $\{Y_{(i)}(x) - \eta^Q(Y_{(i)}(x))\}_{i=1,\cdots,k_Q}$ are independent with mean 0. Hence, the Hoeffding's inequality tells that for any $t \geq 0$

$$\mathbb{P}_{\mathcal{D}_Q}(|\hat{\eta}^Q_{k_Q}(x) - \bar{\eta}^Q_{k_Q}(x)| \geq t|X_{1:n_Q}) \leq 2\exp(-\frac{2t^2}{k_Q}) \leq 2\exp(-\frac{2}{c_Q}(\frac{t}{n_Q^{-\frac{\beta}{2\beta+d}}})^2). \tag{52}$$

On the event E_Q, $\bar{\eta}^Q_{k_Q}(x)$ falls into the interval $[\eta^Q(x) - C_\beta c_D^\beta c_Q^{\frac{\beta}{d}} n_Q^{-\frac{\beta}{2\beta+d}}, \eta^Q(x) + C_\beta c_D^\beta c_Q^{\frac{\beta}{d}} n_Q^{-\frac{\beta}{2\beta+d}}]$, so with probability at least $1 - \frac{c_D}{c_Q} n_Q^{\frac{d}{2\beta+d}} \exp(-c_Q n_Q^{\frac{2\beta}{2\beta+d}})$ w.r.t. the distribution of $X_{1:n_Q}$, we have for any $t \geq 2C_\beta c_D^\beta c_Q^{\frac{\beta}{d}} n_Q^{-\frac{\beta}{2\beta+d}}$,

$$\begin{aligned}
\mathbb{P}_{\mathcal{D}_Q}(|\hat{\eta}^Q_{k_Q}(x) - \eta^Q(x)| \geq t|X_{1:n_Q}) &\leq \mathbb{P}_{\mathcal{D}_Q}(|\hat{\eta}^Q_{k_Q}(x) - \bar{\eta}^Q_{k_Q}(x)| \geq t/2|X_{1:n_Q}) \\
&\leq 2\exp(-\frac{1}{2c_Q}(\frac{t}{n_Q^{-\frac{\beta}{2\beta+d}}})^2).
\end{aligned} \tag{53}$$

Plus, for any $t \in [0, 2C_\beta c_D^\beta c_Q^{\frac{\beta}{d}} n_Q^{-\frac{\beta}{2\beta+d}}]$, since $\exp(-\frac{1}{2c_Q}(\frac{t}{n_Q^{-\frac{\beta}{2\beta+d}}})^2)$ are bounded below from 0, we have

$$\mathbb{P}_{\mathcal{D}_Q}(|\hat{\eta}^Q_{k_Q}(x) - \eta^Q(x)| \geq t|X_{1:n_Q}) \leq C\exp(-\frac{1}{2c_Q}(\frac{t}{n_Q^{-\frac{\beta}{2\beta+d}}})^2). \tag{54}$$

for some $C > 0$ large enough. Combining (53) and (54), the first statement holds with $c_1 = \frac{c_D}{c_Q}$ and $c_2 = \max\{C, 2\}$. □

Proof of Lemma C.2. Consider any $(Q, P) \in \Pi^{NP,s}$. Suppose that $x \in \Omega^+(\gamma, C_\gamma)$. Similarly, define the simple average of the regression functions of the k_P neighbors of x as

$$\bar{\eta}^P_{k_P}(x) := \frac{1}{k_P} \sum_{i=1}^{k_P} \eta^Q(Y^P_{(i)}(x)),$$

which depends on the distribution of $X^P_{1:n_P}$, and the conditional expectation

$$\mathbb{E}_{\mathcal{P}_Q}[\hat{\eta}^P_{k_P}(x) | X_{1:n_P}] = \bar{\eta}^P_{k_P}(x).$$

Again, Condition 1 gives that for any $x, y \in \mathbb{R}^d$, we have

$$|\eta^P(x) - \eta^P(y)| \leq C_{\beta_P} ||x - y||^{\beta_P}.$$

Therefore, on the event E_P we have

$$\begin{aligned} |\bar{\eta}^P_{k_P}(x) - \eta^P(x)| &\leq \frac{1}{k_P} \sum_{i=1}^{k_P} |\eta^P(Y^P_{(i)}(x)) - \eta^P(x)| \\ &\leq C_{\beta_P} \frac{1}{k_P} \sum_{i=1}^{k_P} ||X^P_{(i)}(x) - x||^{\beta_P} \\ &\leq C_{\beta_P} c_D^{\beta_P} c_P^{\frac{\beta_P}{d}} n_P^{-\frac{\beta_P}{2\beta^*_P + d}}. \end{aligned} \tag{55}$$

Setting $c_b \geq (2C_{\beta_P} c_D^{\beta_P} c_Q^{\frac{\beta_P}{d}} / C_\gamma)^{\frac{1}{\gamma}}$, (55) implies that on the event E_P, if $|\eta^Q(x) - \frac{1}{2}| \geq c_b n_P^{-\frac{\beta_P/\gamma}{2\beta^*_P + d}}$, we have

$$|\eta^P(x) - \frac{1}{2}| \geq 2C_{\beta_P} c_D^{\beta_P} c_Q^{\frac{\beta_P}{d}} n_P^{-\frac{\beta_P}{2\beta^*_P + d}} \Rightarrow (\bar{\eta}^P_{k_P}(x) - \frac{1}{2})(\eta^P(x) - \frac{1}{2}) \geq 0, |\bar{\eta}^P_{k_P}(x) - \frac{1}{2}| \geq \frac{1}{2}|\eta^P(x) - \frac{1}{2}|.$$

The Hoeffding's inequality then tells that for any $t \geq 0$

$$\mathbb{P}_{\mathcal{D}_P}(|\hat{\eta}^P_{k_P}(x) - \bar{\eta}^P_{k_P}(x)| \geq t | X^P_{1:n_P}) \leq 2\exp(-\frac{2t^2}{k_P}) \leq 2\exp(-\frac{2}{c_P}(\frac{t}{n_P^{-\frac{\beta^*_P}{2\beta^*_P + d}}})^2). \tag{56}$$

Define

$$\Omega_1^+(\gamma, C_\gamma) := \{x \in \Omega^+(\gamma, C_\gamma) : |\eta^Q(x) - \frac{1}{2}| \geq c_b n_P^{-\frac{\beta_P/\gamma}{2\beta^*_P + d}}\}.$$

If $x \in \Omega_1^+(\gamma, C_\gamma)$, then on the event E_P, we have

$$\begin{aligned} \mathbb{P}_{\mathcal{D}_P}((\hat{\eta}^P(x) - \frac{1}{2})(\eta^Q(x) - \frac{1}{2}) < 0 | X^P_{1:n_P}) &\leq \mathbb{P}_{\mathcal{D}_P}(|\hat{\eta}^P_{k_P}(x) - \bar{\eta}^P_{k_P}(x)| \geq \frac{1}{2}|\eta^P(x) - \frac{1}{2}| | X^P_{1:n_P}) \\ &\leq 2\exp(-(\frac{C_\gamma |\eta^Q(x) - \frac{1}{2}|^\gamma}{\sqrt{\frac{c_P}{2}} n_P^{-\frac{\beta^*_P}{2\beta^*_P + d}}})^2) \end{aligned} \tag{57}$$

On the other hand, setting $c_b \geq c_D c_P^{\frac{1}{d}}$, (50) implies that if $B(x, c_b n_P^{-\frac{1}{2\beta_P^*+d}}) \cap \Omega_P \subset \Omega^+(\gamma, C_\gamma)$, on the event E_P we have

$$X_{(i)}^P(x) \in \Omega^+(\gamma, C_\gamma) \Rightarrow |\eta^P(X_{(i)}^P(x)) - \frac{1}{2}| \geq C_\gamma |\eta^Q(X_{(i)}^P(x)) - \frac{1}{2}|^\gamma.$$

Define

$$\Omega_2^+(\gamma, C_\gamma) := \{x \in \Omega^+(\gamma, C_\gamma) : B(x, c_b n_P^{-\frac{1}{2\beta_P^*+d}}) \cap \Omega \cap \Omega_P \subset \Omega^+(\gamma, C_\gamma)\}.$$

If $x \in \Omega_2^+(\gamma, C_\gamma)$ satisfies that $\eta^Q(x) - \frac{1}{2} \geq 2C_\beta c_D^\beta c_P^{\frac{\beta}{d}} n_P^{-\frac{\beta}{2\beta_P^*+d}}$, then on the event E_P, we have

$$\begin{aligned}
\bar{\eta}_{k_P}^P(x) - \frac{1}{2} &= \frac{1}{k_P} \sum_{i=1}^{k_P} (\eta^P(X_{(i)}^P(x)) - \frac{1}{2}) \\
&\geq C_\gamma \frac{1}{k_P} \sum_{i=1}^{k_P} (\eta^Q(X_{(i)}^P(x)) - \frac{1}{2})^\gamma \\
&\geq C_\gamma \frac{1}{k_P} \sum_{i=1}^{k_P} (\eta^Q(x) - \frac{1}{2} - C_\beta ||X_{(i)}^P(x) - x||^\beta)^\gamma \\
&\geq C_\gamma \frac{1}{k_P} \sum_{i=1}^{k_P} (\eta^Q(x) - \frac{1}{2} - C_\beta (c_D c_P^{\frac{1}{d}} n_P^{-\frac{1}{2\beta_P^*+d}})^\beta)^\gamma \\
&\geq C_\gamma \frac{1}{2^\gamma} (\eta^Q(x) - \frac{1}{2})^\gamma.
\end{aligned}$$

The Hoeffding's inequality then tells that

$$\begin{aligned}
\mathbb{P}_{\mathcal{D}_P}((\hat{\eta}^P(x) - \frac{1}{2})(\eta^Q(x) - \frac{1}{2}) < 0 | X_{1:n_P}^P) &\leq \mathbb{P}_{\mathcal{D}_P}(|\hat{\eta}_{k_P}^P(x) - \bar{\eta}_{k_P}^P(x)| \geq C_\gamma \frac{1}{2^\gamma} |\eta^Q(x) - \frac{1}{2}|^\gamma | X_{1:n_P}^P) \\
&\leq 2 \exp(-(\frac{C_\gamma |\eta^Q(x) - \frac{1}{2}|^\gamma}{2^\gamma \sqrt{\frac{c_P}{2}} n_P^{-\frac{\beta_P^*}{2\beta_P^*+d}}})^2)
\end{aligned} \tag{58}$$

(58) holds for $\eta^Q(x) - \frac{1}{2} \leq -2C_\beta c_D^\beta c_P^{\frac{\beta}{d}} n_P^{-\frac{\beta}{2\beta_P^*+d}}$ with the similar argument. To summarize, if $x \in \Omega_2^+(\gamma, C_\gamma)$ satisfies that $|\eta^Q(x) - \frac{1}{2}| \geq 2C_\beta c_D^\beta c_P^{\frac{\beta}{d}} n_P^{-\frac{\beta}{2\beta_P^*+d}}$, then on the event E_P, we have

$$\mathbb{P}_{\mathcal{D}_P}((\hat{\eta}^P(x) - \frac{1}{2})(\eta^Q(x) - \frac{1}{2}) < 0 | X_{1:n_P}^P) \leq 2 \exp(-(\frac{C_\gamma |\eta^Q(x) - \frac{1}{2}|^\gamma}{2^\gamma \sqrt{\frac{c_P}{2}} n_P^{-\frac{\beta_P^*}{2\beta_P^*+d}}})^2).$$

We finish the proof of this lemma divided into two cases:

I. Case of $\gamma\beta \geq \beta_P$: If $x \in \Omega_2^+(\gamma, C_\gamma)$ satisfies that

$$|\eta^Q(x) - \frac{1}{2}| \geq 2C_\beta c_D^\beta c_P^{\frac{\beta}{d}} n_P^{-\frac{\beta}{2\beta_P^*+d}},$$

then $\exp(-(\frac{C_\gamma|\eta^Q(x)-\frac{1}{2}|^\gamma}{2^\gamma\sqrt{\frac{c_P}{2}}n_P^{-\frac{\beta_P^*}{2\beta_P^*+d}}})^2)$ is bounded below from 0, which means on the event E_P,

$$\mathbb{P}_{\mathcal{D}_P}((\hat{\eta}^P(x)-\frac{1}{2})(\eta^Q(x)-\frac{1}{2})<0|X_{1:n_P}^P)\le C\exp(-(\frac{C_\gamma|\eta^Q(x)-\frac{1}{2}|^\gamma}{2^\gamma\sqrt{\frac{c_P}{2}}n_P^{-\frac{\beta_P^*}{2\beta_P^*+d}}})^2)$$

for some constant $C>0$ large enough. In other words, the exponential concentration holds within the whole $\Omega_2^+(\gamma,C_\gamma)$. Combining this concentration bound with (57) and (58), we see that by setting that $c_b=\max\{(2C_{\beta_P}c_D^{\beta_P}c_Q^{\frac{\beta_P}{d}}/C_\gamma)^{\frac{1}{\gamma}},c_Dc_P^{\frac{1}{2}}\}$

$$\mathbb{P}_{\mathcal{D}_P}((\hat{\eta}^P(x)-\frac{1}{2})(\eta^Q(x)-\frac{1}{2})<0|X_{1:n_P}^P)\le\max\{C,2\}\exp(-(\frac{C_\gamma|\eta^Q(x)-\frac{1}{2}|^\gamma}{2^\gamma\sqrt{\frac{c_P}{2}}n_P^{-\frac{\beta_P^*}{2\beta_P^*+d}}})^2)\tag{59}$$

holds with probability at least $1-\frac{c_D}{c_P}n_P^{\frac{d}{2\beta_P^*+d}}\exp(-c_Qn_P^{\frac{2\beta_P^*}{2\beta_P^*+d}})$ w.r.t. the distribution of $X_{1:n_P}^P$ for any $x\in\Omega_1^+(\gamma,C_\gamma)\cup\Omega_2^+(\gamma,C_\gamma)=\Omega^+(\gamma,C_\gamma)/\Omega^b$. The statement then holds by directly applying Theorem 8 with the setting of

$$\delta_P=2^\gamma\sqrt{\frac{c_P}{2}}n_P^{-\frac{\beta_P^*}{2\beta_P^*+d}},\ \delta_P^f=\mathbb{P}_{\mathcal{D}_P}[E_P^c],\ \delta_P^b=\varepsilon_b.\tag{60}$$

II. Case of $\gamma\beta<\beta_P$: Without loss of generality, we suppose that n_Q is large enough such that

$$2C_\beta c_D^\beta c_P^{\frac{\beta}{d}}n_P^{-\frac{\beta}{2\beta_P^*+d}}\ge(2C_{\beta_P}c_D^{\beta_P}c_Q^{\frac{\beta_P}{d}}/C_\gamma)^{\frac{1}{\gamma}}n_P^{-\frac{\beta_P/\gamma}{2\beta_P^*+d}}.$$

Hence, if $x\in\Omega_2^+(\gamma,C_\gamma)$ satisfies that

$$|\eta^Q(x)-\frac{1}{2}|\ge2C_\beta c_D^\beta c_P^{\frac{\beta}{d}}n_P^{-\frac{\beta}{2\beta_P^*+d}},$$

then by setting $c_b=\max\{(2C_{\beta_P}c_D^{\beta_P}c_Q^{\frac{\beta_P}{d}}/C_\gamma)^{\frac{1}{\gamma}},c_Dc_P^{\frac{1}{2}}\}$, it holds that $x\in\Omega_1^+(\gamma,C_\gamma)$. Therefore, (59) holds with probability at least $1-\frac{c_D}{c_P}n_P^{\frac{d}{2\beta_P^*+d}}\exp(-c_Qn_P^{\frac{2\beta_P^*}{2\beta_P^*+d}})$ w.r.t. the distribution of $X_{1:n_P}^P$ for any $x\in\Omega_1^+(\gamma,C_\gamma)\cup\Omega_2^+(\gamma,C_\gamma)=\Omega^+(\gamma,C_\gamma)/\Omega^b$. The statement then holds by directly applying Theorem 8 with the same setting as the one in (60). □

Proof of Theorem 9. Consider any $(Q,P)\in\Pi^{NP}$. If $\gamma\beta\ge\beta_P$, the proof is straightforward by applying the results derived in Lemma C.1 and C.2 to Theorem 1 by setting $\Pi^{NP,s}=\{(Q,P)\}$ and $\beta_P^*=\gamma\beta$.

If $\gamma\beta<\beta_P$, the proof is straightforward by applying the results derived in Lemma C.1 and C.2 to Theorem 1 by setting $\Pi^{NP,s}=\{(Q,P)\}$ and $\beta_P^*=\beta_P$. To illustrate the reason why the risk term ε_b in Lemma C.2 is asymptotically dominated by $n_P^{-\frac{\beta_P(1+\alpha)/\gamma}{2\beta_P+d}}$, by Assumption 1 we have $\varepsilon_b\le c_bn_P^{-\frac{\beta_P/\gamma}{2\beta_P+d}}Q_X(\Omega_b)\le C_\alpha c_b^{1+\alpha}n_P^{-\frac{\beta_P(1+\alpha)/\gamma}{2\beta_P+d}}$, which finishes the proof. □

C.3. Proofs of Theorem 3

For convenience, we repeat the definitions that

$$\varepsilon_b = \int_{\Omega^b} |\eta^Q(X) - \frac{1}{2}| dQ_X$$

and the boundary of the signal transfer set is defined as

$$\Omega^b := \{x \in \Omega : |\eta^Q(x) - \frac{1}{2}| \leq c_b n_P^{-\frac{\beta_P/\gamma}{2\beta_P^*+d}}, B(x, c_b n_P^{-\frac{1}{2\beta_P^*+d}}) \cap \Omega \cap \Omega_P \not\subset \Omega^+(\gamma, C_\gamma)\}$$

for some constant $c_b > 0$. From Theorem 9, which is the upper bound result for the general parametric space Π^{NP}, it suffices to show that $\varepsilon_b \lesssim n_P^{-\frac{\beta}{2\gamma\beta+d}}$ in each case listed in Theorem 3 to prove the results.

Proof. **Band-like Ambiguity:** By Lemma E.1, Assumption 2 holds with

$$\varepsilon(z; \gamma, C_\gamma/2) = \left(C_\alpha z^{1+\alpha}\right) \wedge \left(2^{\frac{1+\alpha}{\gamma}} C_\alpha C_\gamma^{-\frac{1+\alpha}{\gamma}} \Delta^{\frac{1+\alpha}{\gamma}}\right).$$

From Theorem 9, it suffices to show that $\varepsilon_b \lesssim \Delta^{\frac{1+\alpha}{\gamma}} + n_P^{-\frac{\beta(1+\alpha)}{2\gamma\beta+d}}$ to prove the excess risk upper bound.

Following the same proof in Lemma E.1, we have that

$$\{x \in \Omega : |\eta^Q(x) - \frac{1}{2}| \geq (\frac{2\Delta}{C_\gamma})^{\frac{1}{\gamma}}\} \subset \Omega^+(\gamma, C_\gamma/2).$$

Suppose that $x \in \Omega^+(\gamma, C_\gamma/2)$ satisfies that $|\eta^Q(x) - \frac{1}{2}| \geq (\frac{2\Delta}{C_\gamma})^{\frac{1}{\gamma}} + C_\beta c_b^\beta n_P^{-\frac{\beta}{2\gamma\beta+d}}$. For any $x' \in B(x, c_b n_P^{-\frac{1}{2\gamma\beta+d}})$, we have

$$|\eta^Q(x') - \eta^Q(x)| \leq C_\beta ||x' - x||^\beta \leq C_\beta c_b^\beta n_P^{-\frac{\beta}{2\gamma\beta+d}} \Rightarrow |\eta^Q(x') - \frac{1}{2}| \geq (\frac{2\Delta}{C_\gamma})^{\frac{1}{\gamma}},$$

which further implies that $x \in \Omega^+(\gamma, C_\gamma/2)$. In other words,

$$\Omega_b \subset \{x \in \Omega^+(\gamma, C_\gamma/2) : |\eta^Q(x) - \frac{1}{2}| < (\frac{2\Delta}{C_\gamma})^{\frac{1}{\gamma}} + C_\beta c_b^\beta n_P^{-\frac{\beta}{2\gamma\beta+d}}\},$$

so by Assumption 1, it holds that

$$\begin{aligned}
\varepsilon_b &\leq \left((\frac{2\Delta}{C_\gamma})^{\frac{1}{\gamma}} + C_\beta c_b^\beta n_P^{-\frac{\beta}{2\gamma\beta+d}}\right) Q_X(\Omega_b) \\
&\leq C_\alpha \left((\frac{2\Delta}{C_\gamma})^{\frac{1}{\gamma}} + C_\beta c_b^\beta n_P^{-\frac{\beta}{2\gamma\beta+d}}\right)^{1+\alpha} \\
&\lesssim \Delta^{\frac{1+\alpha}{\gamma}} + n_P^{-\frac{\beta(1+\alpha)}{2\gamma\beta+d}}.
\end{aligned}$$

The proof is then finished given the inequality above.

Smooth Source with Arbitrary Ambiguity: Suppose $(Q,P)\in\Pi_S^{NP}$. By Assumption 1, since $\beta_P=\gamma\beta$, we have $\varepsilon_b\le c_b n_P^{-\frac{\beta}{2\gamma\beta+d}}Q_X(\Omega_b)\le C_\alpha c_b^{1+\alpha}n_P^{-\frac{\beta(1+\alpha)}{2\gamma\beta+d}}$, which finishes the proof.
Flipped Strong Signal: Suppose $(Q,P)\in\Pi_F^{NP}$. For any $x\in\Omega, r>0$ such that

$$\eta^Q(x')\neq\frac{1}{2},\quad \forall x'\in B(x,r),$$

i.e. $B(x,r)$ does not intersect with the decision boundary $\{x\in\Omega:\eta^Q(x)=\frac{1}{2}\}$, by continuity of η^Q we see that either $\eta^Q(x')>\frac{1}{2}$ or $\eta^Q(x')<\frac{1}{2}$ for any $x'\in B(x,r)$.

We claim that either

$$B(x,r)\cap\Omega\subset\Omega^+(\gamma,C_\gamma)$$

or

$$B(x,r)\cap\Omega\subset\Omega/\Omega^+(\gamma,C_\gamma).$$

Otherwise, suppose for any two points $x_1,x_2\in B(x,r)\cap\Omega$, we have $x_1\in\Omega^+(\gamma,C_\gamma)$ but $x_2\in\Omega/\Omega^+(\gamma,C_\gamma)$. Since $|\eta^P(x)-\frac{1}{2}|\ge C_\gamma|\eta^Q(x)-\frac{1}{2}|^\gamma$ holds for any $x\in\Omega$, we have

$$\begin{aligned}&(\eta^P(x_1)-\frac{1}{2})(\eta^Q(x_1)-\frac{1}{2})>0,(\eta^P(x_2)-\frac{1}{2})(\eta^Q(x_2)-\frac{1}{2})<0\\ \Rightarrow&(\eta^P(x_1)-\frac{1}{2})(\eta^P(x_2)-\frac{1}{2})<0.\end{aligned}$$

Therefore, there exists some $\lambda\in(0,1)$ such that

$$\eta^P(\lambda x_1+(1-\lambda)x_2)=\frac{1}{2},$$

which contradicts with the facts that $\eta^Q(\lambda x_1+(1-\lambda)x_2)\neq\frac{1}{2}$ and $|\eta^P(x)-\frac{1}{2}|\ge C_\gamma|\eta^Q(x)-\frac{1}{2}|^\gamma$ for any $x\in\Omega$.

Since the choice of $x\in\Omega$ and $r>0$ are arbitrary, we conclude that any sphere in Ω that does not intersect with the decision boundary $\{x\in\Omega:\eta^Q(x)=\frac{1}{2}\}$ is a subset of either $\Omega^+(\gamma,C_\gamma)$ or $\Omega/\Omega^+(\gamma,C_\gamma)$. Therefore, the signal transfer boundary is a part of the decision boundary, i.e.

$$\partial\Omega^+(\gamma,C_\gamma)\subset\{x\in\Omega:\eta^Q(x)=\frac{1}{2}\},$$

which is a sufficient condition for $\varepsilon_b\lesssim n_P^{-\frac{\beta(1+\alpha)}{2\gamma\beta+d}}$ as shown in the first case of Corollary 1.

□

C.4. Proofs of Corollary 1

Proof. **Signal Transfer Boundary as part of the Decision Boundary:** Suppose that $x\in\Omega^+(\gamma,C_\gamma)$ satisfies that $|\eta^Q(x)-\frac{1}{2}|\ge 2C_\beta c_b^\beta n_P^{-\frac{\beta}{2\gamma\beta+d}}$. For any $x'\in B(x,c_b n_P^{-\frac{1}{2\gamma\beta+d}})$, we have

$$|\eta^Q(x')-\eta^Q(x)|\le C_\beta||x'-x||^\beta\le C_\beta c_b^\beta n_P^{-\frac{\beta}{2\gamma\beta+d}}\le\frac{1}{2}|\eta^Q(x)-\frac{1}{2}|,$$

which further deduces that

$$(\eta^Q(x')-\frac{1}{2})(\eta^Q(x)-\frac{1}{2})>0,\quad\forall x'\in B(x,c_b n_P^{-\frac{1}{2\gamma\beta+d}}).\tag{61}$$

Since $\partial\Omega^+(\gamma, C_\gamma) \subset \{x \in \Omega : \eta^Q(x) = \frac{1}{2}\}$, (61) implies that

$$B(x, c_b n_P^{-\frac{1}{2\gamma\beta+d}}) \cap \Omega \subset \Omega^+(\gamma, C_\gamma) \Rightarrow x \notin \Omega_b.$$

In other words,

$$\Omega_b \subset \{x \in \Omega^+(\gamma, C_\gamma) : |\eta^Q(x) - \frac{1}{2}| < 2C_\beta c_b^\beta n_P^{-\frac{\beta}{2\gamma\beta+d}}\},$$

so by Assumption 1, it holds that

$$\varepsilon_b \leq 2C_\beta c_b^\beta n_P^{-\frac{\beta}{2\gamma\beta+d}} Q_X(\Omega_b) \leq C_\alpha(2C_\beta c_b^\beta)^{1+\alpha} n_P^{-\frac{\beta(1+\alpha)}{2\gamma\beta+d}} \lesssim n_P^{-\frac{\beta(1+\alpha)}{2\gamma\beta+d}},$$

which finishes the proof.

Signal Transfer Boundary Margin: Suppose that n_P is large enough such that

$$c_b n_P^{-\frac{1}{2\gamma\beta+d}} \leq r_m.$$

It is obvious that

$$\varepsilon_b \leq c_b n_P^{-\frac{\beta_P/\gamma}{2\gamma\beta+d}} Q_X(B(X, c_b n_P^{-\frac{1}{2\gamma\beta+d}}) \cap \Omega \cap \Omega_P \not\subset \Omega^+(\gamma, C_\gamma)) \leq C_r c_b^{1+\alpha_r} n_P^{-\frac{\beta_P/\gamma+\alpha_r}{2\gamma\beta+d}},$$

of which the last term is asymptotically less than $n_P^{-\frac{\beta(1+\alpha)}{2\gamma\beta+d}}$ given that $\beta_P/\gamma+\alpha_r \geq \beta(1+\alpha)$. □

C.5. Proof of Theorem 4 and 10

Before proving the minimax lower bound in Theorem 4 and 10 over

$$\Pi^{NP}(\alpha, C_\alpha, \gamma, C_\gamma, \varepsilon, \beta, \beta_P, C_\beta, C_{\beta_P}, \mu^+, \mu^-, c_\mu, r_\mu)$$

and its subsets $\Pi_{BA}^{NP}, \Pi_S^{NP}, \Pi_F^{NP}$, we first prove the excess risk minimax lower bound over a special subset of Π^{NP} and Π_{BA}^{NP} that satisfies $\varepsilon(z; \gamma, C_\gamma) \equiv 0$ or $\Delta = 0$. The parameter space is rigorously defined as

$$\Pi_0^{NP} := \Pi^{NP}(\alpha, C_\alpha, \gamma, C_\gamma, 0, \beta, \beta_P, C_\beta, C_{\beta_P}, \mu^+, \mu^-, c_\mu, r_\mu) \cap \{(Q, P) : \Omega^+(\gamma, C_\gamma) = \Omega = \Omega_P\}.$$

It is trivial to see that $\Pi_0^{NP} \subset \Pi_F^{NP} \subset \Pi^{NP}$ and $\Pi_0^{NP} \subset \Pi_{BA}^{NP}$ since the "perfect source" setting can be viewed as a special case of those with different types of ambiguity. Plus, when $\gamma\beta = \beta_P$, we have $\Pi_0^{NP} \subset \Pi_S^{NP}$. Therefore, the minimax lower over Π_0^{NP} must be the minimax lower bound over the desired parametric spaces, and our first goal is to see the minimax lower bound over Π_0^{NP}. The propositions are as follows:

Proposition 4. *Fix the parameters in the definition of Π^{NP} with $\alpha\beta \leq d, \gamma\beta \geq \beta_P$. We have that*

$$\inf_{\hat{f}} \sup_{(Q,P)\in\Pi_0^{NP}} \mathbb{E}\mathcal{E}_Q(\hat{f}) \gtrsim n_P^{-\frac{\beta(1+\alpha)}{2\gamma\beta+d}} \wedge n_Q^{-\frac{\beta(1+\alpha)}{2\beta+d}}.$$

Proposition 5. *Fix the parameters in the definition of Π^{NP} with $\max\{\alpha\beta, \alpha\beta_P/\gamma\} \leq d, \gamma\beta < \beta_P$. We have that*

$$\inf_{\hat{f}} \sup_{(Q,P)\in\Pi_0^{NP}} \mathbb{E}\mathcal{E}_Q(\hat{f}) \gtrsim n_P^{-\frac{\beta_P(1+\alpha)/\gamma}{2\beta_P+d}} \wedge n_Q^{-\frac{\beta_P(1+\alpha)/\gamma}{2\beta_P/\gamma+d}}.$$

Proposition 4 and 5 serve as the minimax lower bound under the regimes of $\gamma\beta \geq \beta_P$ and $\gamma\beta < \beta_P$ given that there is no ambiguity in the signal provided by η^P relative to η^Q. Furthermore, we see that Proposition 4 is an extension of Theorem 3.2 of Cai and Wei (2021) as we impose the condition $\eta^P = 0$ as well.

The idea of the proofs are based on the application of Assouad's lemma on the family $\mathbb{P}^{\sigma}_{\mathcal{D}_Q} \times \mathbb{P}^{\sigma}_{\mathcal{D}_P}$. Here $\mathbb{P}^{\sigma}_{\mathcal{D}_Q}$ and $\mathbb{P}^{\sigma}_{\mathcal{D}_P}$ are defined as the product probability measure with respect to the target and source data, corresponding to the distribution pairs (Q_σ, P_σ), $\sigma \in \{1, -1\}^m$. See Section E.3 and E.4 for the proof details.

Proofs of Theorem 4 (2), (3) and Theorem 10. We claim that it suffices to prove

$$\inf_{\hat{f}} \sup_{\substack{(Q,P)\in\Pi_S^{NP}\\ \Omega=\Omega_P}} \mathbb{E}\mathcal{E}_Q(\hat{f}) \gtrsim \varepsilon(cn_Q^{-\frac{\beta}{2\beta+d}}; \gamma, C_\gamma). \tag{62}$$

for some constant $c > 0$ given any setting of γ, β and $\beta_P = \gamma\beta$ to obtain Theorem 4 (2). To clarify, since $\Pi_0^{NP} \subset \Pi^{NP} \cap \{(Q,P) : \Omega = \Omega_P\}$, the minimax lower over Π_0^{NP} must be the minimax lower bound over $\Pi^{NP} \cap \{(Q,P) : \Omega = \Omega_P\}$. Hence,

$$\inf_{\hat{f}} \sup_{\substack{(Q,P)\in\Pi_S^{NP}\\ \Omega=\Omega_P}} \mathbb{E}\mathcal{E}_Q(\hat{f}) \gtrsim n_P^{-\frac{\beta(1+\alpha)}{2\gamma\beta+d}} \wedge n_Q^{-\frac{\beta(1+\alpha)}{2\beta+d}}.$$

Since the minimax lower bounds we desire is just $\varepsilon(cn_Q^{-\frac{\beta}{2\beta+d}}; \gamma, C_\gamma)$ added to the lower bounds above, it suffices to prove (62) for both cases to finish the proof. Similarly, it suffices to prove

$$\inf_{\hat{f}} \sup_{\substack{(Q,P)\in\Pi_F^{NP}\\ \Omega=\Omega_P}} \mathbb{E}\mathcal{E}_Q(\hat{f}) \gtrsim \varepsilon(cn_Q^{-\frac{\beta}{2\beta+d}}; \gamma, C_\gamma)$$

for some constant $c > 0$ given any setting of γ, β to obtain Theorem 4 (3), and

$$\inf_{\hat{f}} \sup_{\substack{(Q,P)\in\Pi^{NP}\\ \Omega=\Omega_P}} \mathbb{E}\mathcal{E}_Q(\hat{f}) \gtrsim \varepsilon(cn_Q^{-\frac{\beta}{2\beta+d}}; \gamma, C_\gamma)$$

for some constant $c > 0$ given any setting of γ, β, and β_P to obtain Theorem 10.

We only prove the case of Π_S^{NP}. The cases of Π_F^{NP} and Π^{NP}, no matter what γ, β and β_P are, are the same because it is easy to check that our construction of distribution pairs satisfy all conditions, i.e. $\{(Q_\sigma, P_\sigma),\ \sigma \in \{1,-1\}^m\} \subset \Pi_S^{NP} \cap \Pi_F^{NP} \cap \Pi^{NP} \cap \{(Q,P) : \Omega = \Omega_P\}$ for any γ, β and β_P.

We would like to illustrate our proof idea as follows. In order to satisfy the ambiguity level condition, we need to select a smaller number m of spheres with positive density than the classical case of no source data, as stated in Theorem 4.1 of Audibert and Tsybakov (2007). To achieve this, we set $\eta^P \equiv 1$ and select the maximum possible value of m while still satisfying the given ambiguity level $\varepsilon(z; \gamma, C_\gamma)$.

Fix $\eta^P \equiv 1$. Since η^P is a constant, it is trivial that $\eta^P \in \mathcal{H}(\beta_P, C_{\beta_P})$ for any β_P, C_{β_P}. Next, define the quantities

$$r = c_r n_Q^{-\frac{1}{2\beta+d}}, \quad w = c_w r^d, \quad m = \lceil \frac{c_m \varepsilon(C_\beta c_r^\beta n_Q^{-\frac{\beta}{2\beta+d}}; \gamma, C_\gamma)}{C_\beta w r^\beta} \rceil,$$

where the constants c_r, c_w, c_m will be specified later in the proof. Here, $\lceil a \rceil$ is the minimum integer that is greater equal than a for any real value a. Plus, suppose n_Q is large enough such that

$$m \leq \frac{2c_m \varepsilon(C_\beta c_r^\beta n_Q^{-\frac{\beta}{2\beta+d}}; \gamma, C_\gamma)}{C_\beta w r^\beta}.$$

Suppose that $\varepsilon(C_\beta c_r^\beta n_Q^{-\frac{\beta}{2\beta+d}}; \gamma, C_\gamma) > 0$, or otherwise the lower bound is 0 and the proof is trivial.

We consider a packing $\{x_k\}_{k=1,\cdots,m}$ with radius $2r$ in $[0,1]^d$. For an example of such constuction, divide $[0,1]^d$ into uniform small cubes with side length as $6r$, which forms a grid with at least $\lfloor (6r)^{-1} \rfloor^d$ small cubes. Since

$$\varepsilon(C_\beta c_r^\beta n_Q^{-\frac{\beta}{2\beta+d}}; \gamma, C_\gamma) \leq C_\alpha (C_\beta c_r^\beta)^{1+\alpha} n_Q^{-\frac{\beta(1+\alpha)}{2\beta+d}} = C_\alpha c_r^{-\beta(1+\alpha)} (C_\beta c_r^\beta)^{1+\alpha} r^{\beta(1+\alpha)},$$

we have

$$\begin{aligned} m &\leq \frac{2c_m \varepsilon(2n_Q^{-\frac{\beta}{2\beta+d}}; \gamma, C_\gamma)}{C_\beta w r^\beta} \\ &\leq \frac{2c_m C_\alpha c_r^{-\beta(1+\alpha)} (C_\beta c_r^\beta)^{1+\alpha} r^{\beta(1+\alpha)}}{C_\beta w r^\beta} \\ &= \frac{2c_m C_\alpha c_r^{-\beta(1+\alpha)} (C_\beta c_r^\beta)^{1+\alpha}}{C_\beta c_w} r^{\alpha\beta - d} \\ &\ll \lfloor (6r)^{-1} \rfloor^d. \end{aligned}$$

Therefore, we could suppose that c_m is small enough such that $m < \lfloor (6r)^{-1} \rfloor^d$. Therefore, we could assign exactly one sphere with radius $2r$ in one cube without intersection, which forms a packing of $\{x_k\}_{k=1,\cdots,m}$ with radius $2r$ in $[0,1]^d$. For simplicity of notations, we denote $\frac{2c_m C_\alpha c_r^{-\beta(1+\alpha)} (C_\beta c_r^\beta)^{1+\alpha}}{C_\beta c_w}$ by c_m' that converges to 0 when $c_m \to 0$. Also, define B_c as the compliment of these m balls, i.e. $B_c := [0,1]^d / \left(\bigcup_{k=1}^m B(x_k, 2r) \right)$.

For any $\sigma \in \{1, -1\}^m$, we consider the regression function $\eta_\sigma^Q(x)$ defined as follows:

$$\eta_\sigma^Q(x) = \begin{cases} \frac{1}{2} + \sigma_k C_\beta r^\beta g^\beta(\frac{\|x - x_k\|}{r}) & \text{if } x \in B(x_k, 2r) \text{ for some } k = 1, \cdots, m \\ \frac{1}{2} & \text{otherwise,} \end{cases}$$

where the function $g(\cdot)$ is defined as $g(x) = \min\{1, 2 - x\}$ on $x \in [0, 2]$.

The construction of the marginal distributions $Q_{\sigma,X}$ is as follows. Define

$$r_0 = \left(\frac{c_m \varepsilon(C_\beta c_r^\beta n_Q^{-\frac{\beta}{2\beta+d}}; \gamma, C_\gamma) - (m-1) C_\beta w r^\beta}{C_\beta c_w} \right)^{\frac{1}{\beta+d}}, \quad w_0 = c_w r_0^d.$$

By the definition of m, we have that

$$r_0 \leq (\frac{C_\beta w r^\beta}{C_\beta c_w})^{\frac{1}{\beta+d}} = r.$$

Let $Q_{\sigma,X}$ have the density function $\mu(\cdot)$ defined as

$$\mu(x) = \begin{cases} \frac{w}{\lambda[B(x_k,r)]} & \text{if } x \in B(x_k, r) \text{ for some } k = 1, \cdots, m-1 \\ \frac{w_0}{\lambda[B(x_m,r_0)]} & \text{if } x \in B(x_m, r_0) \\ \frac{1-(m-1)w-w_0}{1-m\lambda[B(x_k,2r)]} & \text{if } x \in B_c \\ 0 & \text{otherwise.} \end{cases}$$

Since the density function $\mu(\cdot)$ does not depend on the specific choice of σ, we fix $P_X = Q_{\sigma,X}$ for any $\sigma \in \{1, -1\}^m$. Given the the construction of Q_σ, we next verify that (Q_σ, P) belongs to Π_S^{NP} for any $\sigma \in \{1, -1\}^m$.

Verify Margin Assumption: We have that

$$\begin{aligned} Q_\sigma(0 < |\eta_\sigma^Q - \frac{1}{2}| < t) &= (m-1)Q_\sigma(0 < C_\beta r^\beta g^\beta(\frac{||X - x_1||}{r}) \leq t) \\ &+ Q_\sigma(0 < C_\beta r^\beta g^\beta(\frac{||X - x_m||}{r}) \leq t) \\ &= (m-1)Q_\sigma(0 < g(\frac{||X - x_1||}{r}) \leq (\frac{t}{C_\beta r^\beta})^{\frac{1}{\beta}}) \\ &+ Q_\sigma(0 < g(\frac{||X - x_m||}{r}) \leq (\frac{t}{C_\beta r^\beta})^{\frac{1}{\beta}}) \\ &= ((m-1)w + w_0)\mathbf{1}\{t \geq C_\beta r^\beta\} \\ &\leq mw\mathbf{1}\{t \geq C_\beta r^\beta\} \\ &\leq c_m' c_w r^{\alpha\beta}\mathbf{1}\{t \geq C_\beta r^\beta\} \\ &\leq C_\alpha t^\alpha. \end{aligned}$$

given that c_m is small enough since $c_m' \xrightarrow{c_m \to 0} 0$. Therefore, it holds that $Q_\sigma \in \mathcal{M}(\alpha, C_\alpha)$.

Verify Smoothness Assumption: It is easy to see that for any $a, b \in [0, 2]$ we have

$$|g^\beta(a) - g^\beta(b)| \leq |a - b|^\beta.$$

Thus, for any $x, x' \in B(x_k, 2r)$, we obtain from the triangular inequality $||x-x_k|| - ||x'-x_k|| \leq ||x - x'||$ that

$$|r^\beta g^\beta(||x - x_k||/r)) - r^\beta g^\beta(||x' - x_k||/r))| \leq ||x - x'||^\beta, \tag{63}$$

Therefore, by the definition of η_σ^Q and (63), we see that

$$|\eta_\sigma^Q(x) - \eta_\sigma^Q(x')| \leq C_\beta ||x - x'||^\beta, \quad \forall x, x \in [0, 1]^d.$$

η^P is smooth with any smoothness parameter β_P.

Verify Strong Density Condition: If $x \in B(x_k, r)$ for some $k = 1, \cdots, m-1$, we have

$$\mu(x) = \frac{w}{\lambda[B(x_k, r)]} = c_w/\pi_d.$$

If $x \in B(x_m, r_0)$ we have

$$\mu(x) = \frac{w_0}{\lambda[B(x_m, r_0)]} = c_w/\pi_d.$$

If $x \in B_c$, we have

$$(m-1)w + w_0 \leq mw \leq c_m' r^{\alpha\beta} \overset{n_Q \to \infty}{\longrightarrow} 0,$$

and

$$m\lambda[B(x_k, 2r)] \leq c_m' 2^d \pi_d r^{\alpha\beta} \overset{n_Q \to \infty}{\longrightarrow} 0.$$

Hence,

$$\mu(x) = \frac{1-(m-1)w - w_0}{1 - m\lambda[B(x_k, 2r)]} \overset{n_Q \to \infty}{\longrightarrow} 1.$$

Combining all cases above, we could set $c_w \in [\pi_d \mu^-, \pi_d \mu^+]$ to satisfy the condition $(Q_\sigma, P) \in \mathcal{S}(\mu^+, \mu^-, c_\mu, r_\mu)$.

Verify Ambiguity Level: Since $\Omega = \Omega_P$, we trivially have that

$$\Omega^-(\gamma, C_\gamma) \subset \{x \in \Omega : \eta^Q(x) \neq \frac{1}{2}\} =: \Omega^-.$$

Then,

$$\begin{aligned}
A(z) :=& \int_{\Omega^-(\gamma, C_\gamma)} |\eta^Q(X) - \frac{1}{2}| \mathbf{1}\{0 < |\eta^Q(X) - \frac{1}{2}| \leq z\} dQ_X \\
\leq& \int_{\Omega^-} |\eta^Q(X) - \frac{1}{2}| \mathbf{1}\{0 < |\eta^Q(X) - \frac{1}{2}| \leq z\} dQ_X \\
=& ((m-1)w + w_0) C_\beta r^\beta \mathbf{1}\{z \geq C_\beta r^\beta\}
\end{aligned}$$

If $z < C_\beta r^\beta$, we have

$$A(z) = 0 \leq \varepsilon(z; \gamma, C_\gamma).$$

If $z \geq C_\beta r^\beta$, we have

$$A(z) \leq C_\beta m w r^\beta \leq 2c_m \varepsilon(C_\beta c_r^\beta n_Q^{-\frac{\beta}{2\beta+d}}; \gamma, C_\gamma) \leq \varepsilon(z; \gamma, C_\gamma).$$

Therefore, the given ambiguity level $\varepsilon(z; \gamma, C_\gamma)$ is well-defined provided that $c_m \leq \frac{1}{2}$.

Putting all verification steps above together, we conclude that $(Q_\sigma, P) \in \Pi_S^{NP}$. We finish the proof of this part by applying Assouad's lemma to (Q_σ, P), $\forall \sigma \in \{1, -1\}^m$.

Suppose that $\sigma, \sigma' \in \{1, -1\}^m$ differ only at one coordinate, i.e.

$$\sigma_k = -\sigma_k', \quad \sigma_l = \sigma_l' \; (\forall l \neq k).$$

If $k \neq m$, we have the Hellinger distance bound as

$$\begin{aligned}
H^2(Q_\sigma, Q_{\sigma'}) &= \frac{1}{2} \int \left(\sqrt{\eta_\sigma^Q(X)} - \sqrt{\eta_{\sigma'}^Q(X)}\right)^2 + \left(\sqrt{1 - \eta_\sigma^Q(X)} - \sqrt{1 - \eta_{\sigma'}^Q(X)}\right)^2 dQ_X \\
&= \int_{B(x_k, r)} \frac{w}{\lambda[B(x_k, r)]} \left(\sqrt{\frac{1}{2} + C_\beta r^\beta} - \sqrt{\frac{1}{2} - C_\beta r^\beta}\right)^2 dx \\
&= \frac{1}{2} w (1 - \sqrt{1 - 2C_\beta^2 r^{2\beta}}) \\
&\leq C_\beta^2 w r^{2\beta}
\end{aligned}$$

Similarly, if $k = m$, the Hellinger distance bound should be

$$H^2(Q_\sigma, Q_{\sigma'}) \leq C_\beta^2 w_0 r^{2\beta} \leq C_\beta^2 w r^{2\beta}.$$

Recall that $r = c_r n_Q^{-\frac{1}{2\beta+d}}$. By the property of Hellinger distance, we have

$$H^2(\mathbb{P}^{\sigma}_{\mathcal{D}_Q}, \mathbb{P}^{\sigma'}_{\mathcal{D}_Q}) \leq n_Q H^2(Q_\sigma, Q_{\sigma'}) \leq n_Q C_\beta^2 w r^{2\beta} \leq C_\beta^2 c_r^{2\beta+d} c_w \leq \frac{\sqrt{2}}{4}$$

provided that c_r is small enough. This further indicates that

$$TV(\mathbb{P}^{\sigma}_{\mathcal{D}_Q}, \mathbb{P}^{\sigma'}_{\mathcal{D}_Q}) \leq \sqrt{2} H^2(\mathbb{P}^{\sigma}_{\mathcal{D}_Q}, \mathbb{P}^{\sigma'}_{\mathcal{D}_Q}) \leq \frac{1}{2}. \tag{64}$$

For any empirical classifier $\hat{f}$, if $k \neq m$, we have

$$\begin{aligned}
\mathcal{E}_{Q_\sigma}(\hat{f}) + \mathcal{E}_{Q_{\sigma'}}(\hat{f}) &= 2\mathbb{E}_{Q_\sigma}[|\eta^{Q_\sigma}(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}(X) \neq f^*_{Q_\sigma}(X)\}] \\
&\quad + 2\mathbb{E}_{Q_{\sigma'}}[|\eta^{Q_{\sigma'}}(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}(X) \neq f^*_{Q_{\sigma'}}(X)\}] \\
&\geq 2\int_{B(x_k,r)} \frac{w}{\lambda[B(x_k,r)]} C_\beta r^\beta (\mathbf{1}\{\hat{f}(X) \neq f^*_{Q_\sigma}(X)\} + \mathbf{1}\{\hat{f}(X) \neq f^*_{Q_{\sigma'}}(X)\})dx \\
&= 2C_\beta w r^\beta.
\end{aligned}$$

If $k = m$, we have

$$\begin{aligned}
\mathcal{E}_{Q_\sigma}(\hat{f}) + \mathcal{E}_{Q_{\sigma'}}(\hat{f}) &= 2\mathbb{E}_{Q_\sigma}[|\eta^{Q_\sigma}(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}(X) \neq f^*_{Q_\sigma}(X)\}] \\
&\quad + 2\mathbb{E}_{Q_{\sigma'}}[|\eta^{Q_{\sigma'}}(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}(X) \neq f^*_{Q_{\sigma'}}(X)\}] \\
&\geq 2\int_{B(x_m,r_0)} \frac{w_0}{\lambda[B(x_m,r_0)]} C_\beta r^\beta (\mathbf{1}\{\hat{f}(X) \neq f^*_{Q_\sigma}(X)\} + \mathbf{1}\{\hat{f}(X) \neq f^*_{Q_{\sigma'}}(X)\})dx \\
&= 2C_\beta w_0 r^\beta.
\end{aligned}$$

Combining this lower bound with (64), the Assouad's lemma shows that

$$\sup_{(Q,P)\in\Pi_S^{NP}} \mathbb{E}\mathcal{E}_Q(\hat{f}) \geq \sup_{\substack{(Q_\sigma,P) \\ \sigma\in\{1,-1\}^m}} \mathbb{E}\mathcal{E}_{Q_\sigma}(\hat{f}) \geq \frac{1}{2} C_\beta((m-1)w + w_0)r^\beta.$$

On one hand, if $m \geq 2$, we further have

$$\sup_{(Q,P)\in\Pi_S^{NP}} \mathbb{E}\mathcal{E}_Q(\hat{f}) \geq \frac{1}{2} C_\beta((m-1)w + w_0)r^\beta \geq \frac{1}{4} C_\beta m w r^\beta \gtrsim \varepsilon(C_\beta c_r^\beta n_Q^{-\frac{\beta}{2\beta+d}}; \gamma, C_\gamma).$$

On the other hand, if $m = 1$, we have

$$r_0 = \left(\frac{c_m \varepsilon(C_\beta c_r^\beta n_Q^{-\frac{\beta}{2\beta+d}}; \gamma, C_\gamma)}{C_\beta c_w} \right)^{\frac{1}{\beta+d}}, \quad w_0 = c_w r_0^d.$$

Therefore,

$$\sup_{(Q,P)\in\Pi_S^{NP}} \mathbb{E}\mathcal{E}_Q(\hat{f}) \geq \frac{1}{2}C_\beta w_0 r^\beta \geq \frac{1}{2}C_\beta c_w r_0^{\beta+d} \gtrsim \varepsilon(C_\beta c_r^\beta n_Q^{-\frac{\beta}{2\beta+d}}; \gamma, C_\gamma).$$

We conclude that it always holds that $\sup_{(Q,P)\in\Pi_{BA}^{NP}} \mathbb{E}\mathcal{E}_Q(\hat{f}) \gtrsim \varepsilon(C_\beta c_r^\beta n_Q^{-\frac{\beta}{2\beta+d}}; \gamma, C_\gamma)$, which finishes the proof. □

Proofs of Theorem 4 (1). Following a similar argument with the ones in the proofs of Theorem 4 (2), (3) and Theorem 10, it suffices to show that

$$\inf_{\hat{f}} \sup_{(Q,P)\in\Pi_{BA}^{NP}} \mathbb{E}\mathcal{E}_Q(\hat{f}) \gtrsim \Delta^{\frac{1+\alpha}{\gamma}} \wedge n_Q^{-\frac{\beta(1+\alpha)}{2\beta+d}}. \tag{65}$$

for any setting of γ and β.

Fix $\eta^P \equiv \frac{1}{2}$. Since η^P is a constant, it is trivial that $\eta^P \in \mathcal{H}(\beta_P, C_{\beta_P})$ for any β_P, C_{β_P}. Next, define the quantities

$$r = (c_r n_Q^{-\frac{1}{2\beta+d}}) \wedge (C_\beta^{-\frac{1}{\beta}}(\frac{\Delta}{C_\gamma})^{\frac{1}{\gamma\beta}}), \quad w = c_w r^d, \quad m = \lfloor c_m r^{\alpha\beta-d} \rfloor,$$

where the constants c_r, c_w, c_m will be specified later in the proof.

We consider a packing $\{x_k\}_{k=1,\cdots,m}$ with radius $2r$ in $[0,1]^d$. We repeat the example of such construction given in Proposition 4. Divide $[0,1]^d$ into uniform small cubes with side length as $6r$, which forms a grid with at least $\lfloor (6r)^{-1} \rfloor^d$ small cubes. Since $m \ll \lfloor (6r)^{-1} \rfloor^d$, we suppose that c_m is small enough such that $m < \lfloor (6r)^{-1} \rfloor^d$. Therefore, we could assign exactly one sphere with radius $2r$ in one cube without intersection, which forms a packing of $\{x_k\}_{k=1,\cdots,m}$ with radius $2r$ in $[0,1]^d$. Also, define B_c as the compliment of these m balls, i.e. $B_c := [0,1]^d / \left(\bigcup_{k=1}^m B(x_k, 2r)\right)$.

For any $\sigma \in \{1,-1\}^m$, we consider the regression function $\eta_\sigma^Q(x)$ defined as follows:

$$\eta_\sigma^Q(x) = \begin{cases} \frac{1}{2} + \sigma_k C_\beta r^\beta g^\beta(\frac{||x-x_k||}{r}) & \text{if } x \in B(x_k, 2r) \text{ for some } k = 1, \cdots, m \\ \frac{1}{2} & \text{otherwise,} \end{cases}$$

where the function $g(\cdot)$ is defined as $g(x) = \min\{1, 2-x\}$ on $x \in [0,2]$.

The construction of the marginal distributions $Q_{\sigma,X}$ is as follows. Let $Q_{\sigma,X}$ have the density function $\mu(\cdot)$ defined as

$$\mu(x) = \begin{cases} \frac{w}{\lambda[B(x_k,r)]} & \text{if } x \in B(x_k, r) \text{ for some } k = 1, \cdots, m \\ \frac{1-mw}{1-m\lambda[B(x_k,2r)]} & \text{if } x \in B_c \\ 0 & \text{otherwise.} \end{cases}$$

Since the density function $\mu(\cdot)$ does not depend on the specific choice of σ, we fix $P_X = Q_{\sigma,X}$ for any $\sigma \in \{1,-1\}^m$. Given the the construction of Q_σ, we next verify that (Q_σ, P) belongs to Π_{BA}^{NP} for any $\sigma \in \{1,-1\}^m$.

Verify Margin Assumption: We have that

$$\begin{aligned}Q_\sigma(0 < |\eta_\sigma^Q - \frac{1}{2}| < t) &= mQ_\sigma(0 < C_\beta r^\beta g^\beta(\frac{||X - x_1||}{r}) \le t)\\ &= mQ_\sigma(0 < g(\frac{||X - x_1||}{r}) \le (\frac{t}{C_\beta r^\beta})^{\frac{1}{\beta}})\\ &= mw\mathbf{1}\{t \ge C_\beta r^\beta\}\\ &\le c_m c_w r^{\alpha\beta}\mathbf{1}\{t \ge C_\beta r^\beta\}\\ &\le C_\alpha t^\alpha.\end{aligned}$$

given that c_m is small enough. Therefore, $Q_\sigma \in \mathcal{M}(\alpha, C_\alpha)$.

Verify Smoothness Assumption: It is easy to see that for any $a, b \in [0, 2]$ we have

$$|g^\beta(a) - g^\beta(b)| \le |a - b|^\beta.$$

Thus, for any $x, x' \in B(x_k, 2r)$, we obtain from the triangular inequality $||x-x_k||-||x'-x_k|| \le ||x - x'||$ that

$$|r^\beta g^\beta(||x - x_k||/r)) - r^\beta g^\beta(||x' - x_k||/r))| \le ||x - x'||^\beta. \tag{66}$$

Therefore, by the definition of η_σ^Q and (66), we see that

$$|\eta_\sigma^Q(x) - \eta_\sigma^Q(x')| \le C_\beta ||x - x'||^\beta, \quad \forall x, x \in [0, 1]^d.$$

Verify Strong Density Condition: If $x \in B(x_k, r)$ for some $k = 1, \cdots, m$, we have

$$\mu(x) = \frac{w}{\lambda[B(x_k, r)]} = c_w/\pi_d.$$

If $x \in B_c$, we have

$$\mu(x) = \frac{1 - c_w\lfloor c_m r^{\alpha\beta-d}\rfloor r^d}{1 - 2^d\pi_d\lfloor c_m r^{\alpha\beta-d}\rfloor r^d} \overset{n_Q\to\infty}{\longrightarrow} 1.$$

Therefore, we could set $c_w \in [\pi_d\mu^-, \pi_d\mu^+]$ to satisfy the condition $(Q_\sigma, P) \in \mathcal{S}(\mu^+, \mu^-, c_\mu, r_\mu)$.

Verify Band-like Ambiguity: Since $\eta^P \equiv \frac{1}{2}$, by Definition 1 we have $s(x) = 0$ for any $x \in \Omega$. Plus, by definition of η_σ^Q and r, we have that for any $x \in \Omega$,

$$|\eta_\sigma^Q - \frac{1}{2}| \le C_\beta r^\beta \le (\frac{\Delta}{C_\gamma})^{\frac{1}{\gamma}} \Rightarrow C_\gamma|\eta_\sigma^Q - \frac{1}{2}|^\gamma \le \Delta.$$

Therefore, we have

$$s(x) = 0 \ge C_\gamma|\eta_\sigma^Q - \frac{1}{2}|^\gamma - \Delta$$

for any $x \in \Omega$.

Putting all verification steps above together, we conclude that $(Q_\sigma, P) \in \Pi_{BA}^{NP}$. We finish the proof of this part by applying Assouad's lemma to (Q_σ, P), $\forall\sigma \in \{1, -1\}^m$.

If $\sigma, \sigma' \in \{1, -1\}^m$ differ only at one coordinate, i.e.

$$\sigma_k = -\sigma_k', \quad \sigma_l = \sigma_l' \ (\forall l \ne k),$$

we have the Hellinger distance bound as

$$\begin{aligned} H^2(Q_\sigma, Q_{\sigma'}) &= \frac{1}{2}\int \left(\sqrt{\eta^Q_\sigma(X)} - \sqrt{\eta^Q_{\sigma'}(X)}\right)^2 + \left(\sqrt{1-\eta^Q_\sigma(X)} - \sqrt{1-\eta^Q_{\sigma'}(X)}\right)^2 dQ_X \\ &= \int_{B(x_k,r)} \frac{w}{\lambda[B(x_k,r)]}\left(\sqrt{\frac{1}{2}+C_\beta r^\beta} - \sqrt{\frac{1}{2}-C_\beta r^\beta}\right)^2 dx \\ &= \frac{1}{2}w(1-\sqrt{1-2C_\beta^2 r^{2\beta}}) \\ &\le C_\beta^2 w r^{2\beta} \end{aligned}$$

Recall that $r = (c_r n_Q^{-\frac{1}{2\beta+d}}) \wedge (C_\beta^{-\frac{1}{\beta}}(\frac{\Delta}{C_\gamma})^{\frac{1}{\gamma\beta}})$. By the property of Hellinger distance, we have

$$H^2(\mathbb{P}^\sigma_{\mathcal{D}_Q}, \mathbb{P}^{\sigma'}_{\mathcal{D}_Q}) \le n_Q H^2(Q_\sigma, Q_{\sigma'}) \le C_\beta^2 w n_Q r^{2\beta} \le C_\beta^2 c_r^{2\beta+d} c_w \le \frac{\sqrt{2}}{4}$$

provided that c_r is small enough. This further indicates that

$$TV(\mathbb{P}^\sigma_{\mathcal{D}_Q}, \mathbb{P}^{\sigma'}_{\mathcal{D}_Q}) \le \sqrt{2}H^2(\mathbb{P}^\sigma_{\mathcal{D}_Q}, \mathbb{P}^{\sigma'}_{\mathcal{D}_Q}) \le \frac{1}{2}. \tag{67}$$

For any empirical classifier $\hat{f}$, we have

$$\begin{aligned} \mathcal{E}_{Q_\sigma}(\hat{f}) + \mathcal{E}_{Q_{\sigma'}}(\hat{f}) &= 2\mathbb{E}_{Q_\sigma}[|\eta^{Q_\sigma}(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}(X) \neq f^*_{Q_\sigma}(X)\}] \\ &\quad + 2\mathbb{E}_{Q_{\sigma'}}[|\eta^{Q_{\sigma'}}(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}(X) \neq f^*_{Q_{\sigma'}}(X)\}] \\ &\ge 2\int_{B(x_k,r)} \frac{w}{\lambda[B(x_k,r)]} C_\beta r^\beta(\mathbf{1}\{\hat{f}(X) \neq f^*_{Q_\sigma}(X)\} + \mathbf{1}\{\hat{f}(X) \neq f^*_{Q_{\sigma'}}(X)\})dx \\ &= 2C_\beta w r^\beta. \end{aligned}$$

Combining this lower bound with (67), the Assouad's lemma shows that

$$\sup_{(Q,P)\in\Pi^{NP}_{BA}} \mathbb{E}\mathcal{E}_Q(\hat{f}) \ge \sup_{\substack{(Q_\sigma,P) \\ \sigma\in\{1,-1\}^m}} \mathbb{E}\mathcal{E}_{Q_\sigma}(\hat{f}) \ge \frac{1}{2}C_\beta m w r^\beta \gtrsim r^{\beta(1+\alpha)} \gtrsim \Delta^{\frac{1+\alpha}{\gamma}} \wedge n_Q^{-\frac{\beta(1+\alpha)}{2\beta+d}}.$$

□

D. Proofs in Parametric Classification

We first list some definitions that are helpful for the proofs in this section. Denote the target and source distribution with certain linear coefficients β, w as

$$\begin{aligned} Q_\beta &:= \{Q : X \sim N(0, I_d), \eta^Q(x) = \sigma(\beta^T x)\}, \\ P_w &:= \{Q : X \sim N(0, I_d), \eta^Q(x) = \sigma(w^T x)\}, \end{aligned}$$

for any $\beta, w \in \mathbb{R}^d$. The family of distribution pairs then becomes

$$\Pi^{LR} = \Pi^{LR}(s, \Pi, M) = \{(Q_\beta, P_w) : (\beta, w) \in \Theta(s, \Delta)\},$$

where $\Theta(s, \Delta)$ is defined in (22). Moreover, we abbreviate $\mathbb{E}_\beta$ as $\mathbb{E}_{(X,Y)\sim Q_\beta}$, the expectation with respect to a new observation drawn from the distribution Q_β. Similarly, define $\mathcal{E}_{Q_\beta}(f)$ as the excess risk with respect to the target distribution Q_β.

D.1. Proof of Lemma 1

Proof. For any $x \in B(0,1)$, we have

$$\sigma'(\beta_Q^T x) = \sigma(\beta_Q^T x)(1-\sigma(\beta_Q^T x))\beta_Q,$$

$$\sigma'(\beta_P^T x) = \sigma(\beta_Q^T x)(1-\sigma(\beta_Q^T x))\beta_P.$$

Define e_Q and e_P as the unit vectors that have the same directions with β_Q and β_P, respectively.

Verify Margin Assumption: From the simple equality that $\log(1+\frac{4t}{1-2t}) \leq 8t$ for any $t \in [0, \frac{1}{4}]$ and $||\beta_Q|| \geq \frac{L}{m}$ we have

$$\begin{aligned} Q_X(0 < |\sigma(\beta_Q^T X) - \frac{1}{2}| < t) &= Q_X(0 < |\beta_Q^T X| \leq \log(1+\frac{4t}{1-2t})) \\ &\leq Q_X(0 < |\beta_Q^T X| \leq 8t) \\ &\leq Q_X(0 < |e_Q^T X| \leq \frac{8m}{L} t) \\ &= Q_X(|N(0,1)| \leq \frac{8m}{L} t) \\ &\leq 16\frac{m}{L} t\phi(0) = \frac{16m}{\sqrt{2\pi}L} t \end{aligned}$$

where $\phi(\cdot)$ is the density function of the univariate standard normal distribution $N(0,1)$. For $t \in (\frac{1}{4}, 1]$, we trivially have $Q_X(0 < |\sigma(\beta_Q^T X) - \frac{1}{2}| < t) \leq 4t$ since Q_X is a probability measure. Therefore, Assumption holds with $\alpha = 1, C_\alpha = \frac{16m}{\sqrt{2\pi}L} \vee 4$.

Verify Ambiguity Level: A trivial well-defined ambiguity level by the margin assumption is

$$\varepsilon(z;\gamma,\frac{m}{\pi}) = C_\alpha z^{1+\alpha} = (\frac{16m}{\sqrt{2\pi}L} \vee 4) z^2.$$

Hence, it suffices to show that a well-defined ambiguity level is

$$\varepsilon(z;1,\frac{m}{\pi}) = \frac{\sqrt{2}U}{m}\Delta^2.$$

Without loss of generality and due to the symmetry property of of $N(0, I_d)$, we could rotate β_Q and β_P at the same time so that

$$\beta_Q = (||\beta_Q||, 0, 0, \cdots, 0), \quad \beta_P = (||\beta_P|| \cos\langle \beta_Q, \beta_P\rangle, ||\beta_P|| \sin\langle \beta_Q, \beta_P\rangle, 0, 0, \cdots, 0).$$

Therefore, we assume that only the first coordinate of β_Q and the first and second coordinates of β_P can be non-zero.

For any $x \in \mathbb{R}^d$, we define $d_Q(x)$ and $d_P(x)$ as the angle between $(x_1, x_2, 0, \cdots, 0)$ and the normal planes of β_Q and β_P, respectively. Note that $d_Q(x), d_P(x) \in [0, \pi/2]$. Define the area

$$D := \{x \in \mathbb{R}^d : (\beta_Q^T x)(\beta_P^T x) < 0\} \cup \{x \in \mathbb{R}^d : (\beta_Q^T x)(\beta_P^T x) \geq 0, d_P(x) < \frac{1}{2} d_Q(x)\}$$

as the region where η^P does not give strong signal relative to η^Q. We see that D is a cone centered at origin with angle between the two normal planes equal to $2\langle\beta_Q,\beta_P\rangle \leq 2\Delta$. Plus, for any $x \in D$, the norm of projection of x onto the normal plane of β_Q is less than

$$(x_1^2+x_2^2)^{\frac{1}{2}}\sin 2\langle\beta_Q,\beta_P\rangle \leq 2(x_1^2+x_2^2)^{\frac{1}{2}}\Delta,$$

where the simple inequality $\sin 2\langle\beta_Q,\beta_P\rangle \leq 2\langle\beta_Q,\beta_P\rangle$ is used. Therefore,

$$|\eta^Q(x)-\frac{1}{2}| \leq 2\max\{\sigma'(z)\}_{z\in\mathbb{R}}||\beta_Q||(x_1^2+x_2^2)^{\frac{1}{2}}\Delta \leq \frac{U}{2m}(x_1^2+x_2^2)^{\frac{1}{2}}\Delta \quad \forall x \in D,$$

and

$$\begin{aligned}\int_D|\eta^Q(X)-\frac{1}{2}|dQ_X &\leq \frac{U}{2m}\Delta\int_D(X_1^2+X_2^2)^{\frac{1}{2}}dQ_X\\ &= \frac{U}{2m}\Delta(2\Delta)\int_{\mathbb{R}^d}(X_1^2+X_2^2)^{\frac{1}{2}}dQ_X\\ &\leq \frac{U}{m}\Delta^2\int_{\mathbb{R}^d}(X_1^2+X_2^2)^{\frac{1}{2}}dQ_X\\ &\leq \frac{\sqrt{2}U}{m}\Delta^2.\end{aligned}$$

since $\Delta \in [0,\frac{\pi}{2}]$ and $\mathbb{E}_{Q_X}[(X_1^2+X_2^2)^{\frac{1}{2}}] \leq \mathbb{E}_{Q_X}[X_1^2+X_2^2]^{\frac{1}{2}} = \sqrt{2}$.

On the other hand, if $x \notin D$, we have

$$\begin{aligned}|\eta^P(x)-\frac{1}{2}| &= |\sigma(\beta_P^T x)-\frac{1}{2}|\\ &= |\sigma(||\beta_P||(x_1^2+x_2^2)^{\frac{1}{2}}\sin(d_P(x))-\frac{1}{2}|\\ &\geq |\sigma(m||\beta_Q||(x_1^2+x_2^2)^{\frac{1}{2}}\sin(\frac{1}{2}d_Q(x))-\frac{1}{2}|\\ &\geq |\sigma(\frac{m}{\pi}||\beta_Q||(x_1^2+x_2^2)^{\frac{1}{2}}\sin(d_Q(x))-\frac{1}{2}|\\ &\geq |\sigma(\frac{m}{\pi}\beta_Q^T x)-\frac{1}{2}|\\ &\geq \frac{m}{\pi}|\sigma(\beta_Q^T x)-\frac{1}{2}|\end{aligned}$$

given the fact that $\sin(\frac{1}{2}t) \geq \frac{\sin t}{\pi}$ on $t \in [0,\frac{\pi}{2}]$ and that $|\sigma(t)-\frac{1}{2}|/x$ in monotone decreasing on $t \geq 0$. Therefore, a well-defined ambiguity level is $\varepsilon(z;1,\frac{m}{\pi}) = \frac{\sqrt{2}U}{m}\Delta^2$. □

D.2. Proof of Theorem 5

Define the logistic link function $\psi(x) := \log(1+e^x)$. It is obvious from straightforward calculation that $\psi'(x) = \frac{e^x}{e^x+1}$, $\psi''(x) = \frac{e^x}{(e^x+1)^2} \leq \frac{1}{4}$. Prior to the proof, we present two lemmas that are helpful for deriving the key conditions in Theorem 2. They serve as the parametric bound of the estimator obtained by the loss function optimization in (21). See Section E.5 for the proof of the two lemmas.

Lemma D.1 (Parameter Bound for $\hat{\beta}_Q$)**.** Under the notations and conditions of Theorem 5, there exists some constant $C_Q, \kappa_Q > 0$ such that

$$\mathbb{P}_{\mathcal{D}}(||\hat{\beta}_Q - \beta_Q|| \geq \kappa_Q \sqrt{\frac{s \log d}{n_Q}}) \leq C_Q \left(\frac{s \log d}{n_Q} \wedge \frac{s \log d}{n_P} \right). \tag{68}$$

Lemma D.2 (Parameter Bound for $\hat{\beta}_P$)**.** Under the notations and conditions of Theorem 5, there exists some constant $C_P, \kappa_P > 0$ such that

$$\mathbb{P}_{\mathcal{D}}(||\hat{\beta}_P - \beta_P|| \geq \kappa_P (\sqrt{\frac{s \log d}{n_P}} + \Delta)) \leq C_P \left(\frac{s \log d}{n_Q} \wedge \frac{s \log d}{n_P} \right) \tag{69}$$

when $\Delta \leq c_P$ for some constant $c_P > 0$.

Proof of Theorem 5. By Lemma D.1, there exist constants $C_Q, \kappa_Q > 0$ such that

$$\mathbb{P}_{\mathcal{D}_Q}(E^c) \leq C_Q (\frac{s \log d}{n_Q} \wedge \frac{s \log d}{n_P}).$$

where we define the event $E := \{||\hat{\beta}_Q - \beta_Q|| \leq \kappa_Q \sqrt{\frac{s \log d}{n_Q}}\}$. Define $\hat{e}$ as the unit vector with the same direction as $\hat{\beta}_Q - \beta_Q$. We define

$$\Omega^* := \{x \in \Omega : |\hat{e}^T x| \leq \sqrt{2}\tau / (\frac{\kappa_Q}{2\sqrt{2}} \sqrt{\frac{s \log d}{n_Q}})\}.$$

Since $\hat{e}^T X \sim N(0, 1)$ with respect to Q_X, the tail bound of normal distributions that

$$Q(\Omega^*) \geq 1 - 2 \exp(- \left(\tau / (\frac{\kappa_Q}{2\sqrt{2}} \sqrt{\frac{s \log d}{n_Q}}) \right)^2).$$

Since $\sigma(\cdot)$ is Lipschitz-continuous with the Lipschitz constant as $\frac{1}{4}$, on the event E, we further have

$$|\sigma(\beta_Q^T x) - \sigma(\hat{\beta}_Q^T x)| \leq \frac{1}{4} |(\beta_Q^T - \hat{\beta}_Q^T) x| \leq ||\hat{\beta}_Q - \beta_Q|| |\hat{e}^T x| \leq \tau. \tag{70}$$

for any $x \in \Omega^*$.

Now we could directly apply Theorem 2 where the parameters are

$$\tau = c_\tau \sqrt{\frac{s \log d}{n_Q}} \log(n_Q \vee n_P),$$

and

$$\delta_Q(n_Q, \tau) = 2 \exp(- \left(\tau / (\frac{\kappa_Q}{2\sqrt{2}} \sqrt{\frac{s \log d}{n_Q}}) \right)^2) + C_Q (\frac{s \log d}{n_Q} \wedge \frac{s \log d}{n_P})$$

with $c_\tau \geq \frac{\kappa_Q}{2\sqrt{2}}$. Note that $\delta_Q(n_Q, \tau) \lesssim \frac{s\log d}{n_Q} \wedge \frac{s\log d}{n_P}$. By Theorem 2, the model-selected classifier $\hat{f}^{LR}_{MS}$ satisfies that

$$
\begin{aligned}
\hat{f}^{LR}_{MS} &\lesssim \sup_{(Q,P)\in\Pi^{LR}} \mathbb{E}_{\mathcal{D}_P}\varepsilon_T(\mathbf{1}\{\hat{\beta}_P^T x \geq 0\}; 1, \frac{m}{\pi}) \wedge \tau^2 + \varepsilon(2\tau; 1, \frac{m}{\pi}) + \delta_Q(n_Q, \tau)\\
&\lesssim \sup_{(Q,P)\in\Pi^{LR}} \mathbb{E}_{\mathcal{D}_P}\varepsilon_T(\mathbf{1}\{\hat{\beta}_P^T x \geq 0\}; 1, \frac{m}{\pi}) \wedge (\log^2(n_Q \vee n_P)\frac{s\log d}{n_Q})\\
&+ (\log^2(n_Q \vee n_P)\frac{s\log d}{n_Q}) \wedge \Delta^2\\
&+ \frac{s\log d}{n_Q} \wedge \frac{s\log d}{n_P}.
\end{aligned}
$$

Therefore, it suffices to show that

$$
\mathbb{E}_{\mathcal{D}_P}\varepsilon_T(\mathbf{1}\{\hat{\beta}_P^T x \geq 0\}; 1, \frac{m}{\pi}) \lesssim \frac{s\log d}{n_P} + \Delta^2
$$

for any $(Q, P) \in \Pi^{LR}$ to prove Theorem 5. If $\Delta > c_p$ where c_p is the constant constraint of Δ in Lemma D.2, this bound is trivial as $\mathbb{E}_{\mathcal{D}_P}\varepsilon_T(\mathbf{1}\{\hat{\beta}_P^T x \geq 0\}; 1, \frac{m}{\pi}) \leq 1$. Hence, we WLOG assume that $\Delta \leq c_p$ to apply Lemma D.2. Plus, we may as well assume that c_P is small enough such that $\kappa_P c_P \leq \frac{L}{4}$.

By Lemma D.2, there exist constants $C_P, \kappa_P > 0$ such that

$$
\mathbb{P}_{\mathcal{D}_P}(E_2^c) \leq C_P(\frac{s\log d}{n_Q} \wedge \frac{s\log d}{n_P}).
$$

where we define the event $E_2 := \{||\hat{\beta}_P - \beta_P|| \leq \kappa_P(\sqrt{\frac{s\log d}{n_P}} + \Delta)\}$. Denote $\kappa_P(\sqrt{\frac{s\log d}{n_P}} + \Delta)$ by Λ by simplicity. By the theorem setting, we could assume that n_P is large enough such that $\kappa_P\sqrt{\frac{s\log d}{n_P}} \leq \frac{L}{4}$, so we have $\Lambda \leq \frac{L}{2} \leq ||\beta_P||_2$. Therefore, on the event E_2 we have $\langle\hat{\beta}_P, \beta_P\rangle \leq \frac{\pi}{2}$ and

$$
\sin\langle\hat{\beta}_P, \beta_P\rangle / ||\hat{\beta}_P - \beta_P|| \leq 1/||\beta_P||
$$

by the law of sines. Hence,

$$
\langle\hat{\beta}_P, \beta_P\rangle \leq \frac{\pi}{2}\sin\langle\hat{\beta}_P, \beta_P\rangle \leq ||\hat{\beta}_P - \beta_P||/||\beta_P|| \leq \frac{\Lambda}{L}.
$$

Without loss of generality and due to the symmetry property of of $N(0, I_d)$, we could rotate β_P and $\hat{\beta}_P$ at the same time so that

$$
\beta_P = (||\beta_P||, 0, 0, \cdots, 0), \quad \hat{\beta}_P = (||\hat{\beta}_P||\cos\langle\hat{\beta}_P, \beta_P\rangle, ||\hat{\beta}_P||\sin\langle\hat{\beta}_P, \beta_P\rangle, 0, 0, \cdots, 0).
$$

Therefore, we assume that only the first coordinate of β_Q and the first and second coordinates of $\hat{\beta}_P$ can be non-zero.

Define $D := \{x \in \mathbb{R}^d : (\beta_P^T x)(\hat{\beta}_P^T x) < 0\}$, which is a cone centered at the origin with angle size bounded by $\frac{\Lambda}{L}$. For any $x \in D$, the norm of projection of x onto the normal plane of β_Q is less than

$$
(x_1^2 + x_2^2)^{\frac{1}{2}}\sin\langle\hat{\beta}_P, \beta_P\rangle \leq (x_1^2 + x_2^2)^{\frac{1}{2}}\frac{\Lambda}{L},
$$

where the simple inequality $\sin\langle\hat{\beta}_P, \beta_P\rangle \leq \langle\hat{\beta}_P, \beta_P\rangle$ is used. Therefore,

$$|\eta^Q(x) - \frac{1}{2}| \leq \max\{\sigma'(z)\}_{z\in\mathbb{R}}||\beta_P||(x_1^2 + x_2^2)^{\frac{1}{2}}\frac{\Lambda}{L} \leq \frac{U}{4}(x_1^2 + x_2^2)^{\frac{1}{2}}\frac{\Lambda}{L} \quad \forall x \in D,$$

and

$$\begin{aligned}\int_D |\eta^Q(X) - \frac{1}{2}|dQ_X &\leq \frac{U}{4}\frac{\Lambda}{L}\int_D (X_1^2 + X_2^2)^{\frac{1}{2}} dQ_X \\ &= \frac{U}{4}(\frac{\Lambda}{L})^2 \int_{\mathbb{R}^d} (X_1^2 + X_2^2)^{\frac{1}{2}} dQ_X \\ &\leq \frac{\sqrt{2}U}{4L^2}\Lambda^2.\end{aligned}$$

since $\frac{\Lambda}{L} \in [0, \frac{\pi}{2}]$ and $\mathbb{E}_{Q_X}[(X_1^2 + X_2^2)^{\frac{1}{2}}] \leq \mathbb{E}_{Q_X}[X_1^2 + X_2^2]^{\frac{1}{2}} = \sqrt{2}$. Hence, we obtain that

$$\begin{aligned}\varepsilon_T(\mathbf{1}\{\hat{\beta}_P^T x \geq 0\}; 1, \frac{m}{\pi}) &= 2\mathbb{E}_{X\sim N(0,I_d)}[|\eta^P - \frac{1}{2}|\mathbf{1}\{X \in D\}] \\ &\leq \frac{\sqrt{2}U}{2L^2}\Lambda^2 \\ &= \frac{\sqrt{2}U}{2L^2}\kappa_P^2(\sqrt{\frac{s\log d}{n_P}} + \Delta)^2\end{aligned}$$

To summarize, we have the bounds$\varepsilon_T(\mathbf{1}\{\hat{\beta}_P^T x \geq 0\}; 1, \frac{m}{\pi}) \leq \frac{\Lambda}{L}$ on the event E_2 and $\varepsilon_T(\mathbf{1}\{\hat{\beta}_P^T x \geq 0\}; 1, \frac{m}{\pi}) \leq 1$ on the event E_2^c. As a result,

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}_P}\varepsilon_T(\mathbf{1}\{\hat{\beta}_P^T x \geq 0\}; 1, \frac{m}{\pi}) \\ =&\mathbb{E}_{\mathcal{D}_P}[\varepsilon_T(\mathbf{1}\{\hat{\beta}_P^T x \geq 0\}; 1, \frac{m}{\pi})|E_2]\mathbb{P}_{\mathcal{D}_P}(E_2) + \mathbb{E}_{\mathcal{D}_P}[\varepsilon_T(\mathbf{1}\{\hat{\beta}_P^T x \geq 0\}; 1, \frac{m}{\pi})|E_2^c]\mathbb{P}_{\mathcal{D}_P}(E_2^c) \\ \leq&\frac{\sqrt{2}U}{2L^2}\kappa_P^2(\sqrt{\frac{s\log d}{n_P}} + \Delta)^2 \times 1 + 1 \times C_P(\frac{s\log d}{n_Q} \wedge \frac{s\log d}{n_P}) \\ =&\frac{\sqrt{2}U}{2L^2}\kappa_P^2(\sqrt{\frac{s\log d}{n_P}} + \Delta)^2 + C_P(\frac{s\log d}{n_Q} \wedge \frac{s\log d}{n_P}) \\ \lesssim&\frac{s\log d}{n_P} + \Delta^2.\end{aligned}$$

□

D.3. Proof of Theorem 6

Given an arbitrary classifier $\hat{f}$ based on $\mathcal{D}_Q$ and $\mathcal{D}_P$. Our goal is to show that

$$\sup_{(Q,P)\in\Pi^{LR}} \mathbb{E}\mathcal{E}_Q(\hat{f}) \gtrsim \left(\frac{s\log d}{n_P} + \Delta^2\right) \wedge \frac{s\log d}{n_Q}.$$

First part of Proof. As part of the proof, we would like to first show that

$$\sup_{(Q,P)\in\Pi^{LR}} \mathbb{E}\mathcal{E}_Q(\hat{f}) \gtrsim \frac{s\log d}{n_P} \wedge \frac{s\log d}{n_Q}.$$

The idea is to suppose

$$\beta_P = \beta_Q = \beta, \quad P = Q = Q_\beta$$

which implies $\langle \beta_Q, \beta_P \rangle = 0$, and then reveal the lower bound provided by the combined sample $\mathcal{D}_Q \cup \mathcal{D}_P$. The proof is based on Fano's lemma on the vector angle.

We consider a class of $\beta \in \mathbb{R}^d$ such that

$$||\beta||_0 \leq s, \;\; \beta_1 = 1, \;\; |\beta_j| \in \{0, C\sqrt{\log d/(n_Q \vee n_P)}, -C\sqrt{\log d/(n_Q \vee n_P)}\} \; (\forall 2 \leq j \leq d),$$

where the constant $C > 0$ will be specified later. Denote such a class of β as $\mathcal{H}$. The construction of the class of β is inspired and similar to Raskutti et al. (2009). By Lemma 5 of Raskutti et al. (2009), we see that there exists a subset of $\mathcal{H}$, named $\tilde{\mathcal{H}}$, such that $|\tilde{\mathcal{H}}| \geq \exp(\frac{s-1}{2} \log \frac{d-s}{(s-1)/2})$ and

$$\sum_{j=2}^{d} \mathbf{1}\{\beta_j \neq \beta'_j\} \geq \frac{s-1}{2}, \quad \forall \beta, \beta' \in \tilde{\mathcal{H}}.$$

We provide the following two observations on the elements of $\tilde{\mathcal{H}}$:

- For any $\beta \in \tilde{\mathcal{H}}$, its norm is bounded by the following inequality:

$$1 \leq ||\beta||_2 \leq 1 + C\sqrt{s \log d/(n_Q \vee n_P)} \leq M, \tag{71}$$

  for some $M > 0$ large enough since we supposed that $s \log d/(n_Q \vee n_P) \lesssim 1$.

- For any $\beta \neq \beta' \in \tilde{\mathcal{H}}$, their angle could be bounded by

$$\begin{aligned} \langle \beta, \beta' \rangle \geq \sin\langle \beta, \beta' \rangle \geq & \frac{(\sum_{j=2}^{d} \mathbf{1}\{\beta_j \neq \beta'_j\})^{\frac{1}{2}} C\sqrt{\log d/(n_Q \vee n_P)}}{||\beta||} \\ \geq & \frac{C\sqrt{s \log d/(n_Q \vee n_P)}}{2M}. \end{aligned} \tag{72}$$

  This inequality holds since as long as $\beta_j \neq \beta'_j$ for some $2 \leq j \leq d$, based on the construction of $\tilde{\mathcal{H}}$ we have $|\beta_j - \beta'_j| \geq C\sqrt{\log d/(n_Q \vee n_P)}$.

Next, we define the random variable Z as the $\beta \in \tilde{\mathcal{H}}$ that attains the minimum of $\mathcal{E}_{Q_\beta}(\hat{f})$, i.e.

$$Z := \arg\min_{\beta \in \tilde{\mathcal{H}}} \mathcal{E}_{Q_\beta}(\hat{f}).$$

For any $\beta' \neq Z, \beta' \in \tilde{\mathcal{H}}$, by Lemma E.4, we have that

$$\mathcal{E}_{Q_Z}(\hat{f}) + \mathcal{E}_{Q_{\beta'}}(\hat{f}) \geq \frac{\sigma(M)(1-\sigma(M))}{20\pi} \langle Z, \beta' \rangle^2.$$

by (71). The choice of Z further indicates that

$$\begin{aligned} \mathcal{E}_{Q_{\beta'}}(\hat{f}) \geq & \frac{\sigma(M)(1-\sigma(M))}{40\pi} \langle Z, \beta' \rangle^2 \\ \geq & \frac{C\sigma(M)(1-\sigma(M))}{80\pi M} \sqrt{s \log d/(n_Q \vee n_P)}. \end{aligned}$$

We denote the constant term $\frac{C\sigma(M)(1-\sigma(M))}{80\pi M}$ by C_M for simplicity. Hence,

$$\mathbb{E}_{\mathcal{D}_{\beta'}}\mathcal{E}_{Q_{\beta'}}(\hat{f}) \geq C_M^2 s\log d/(n_Q \vee n_P) \times \mathbb{P}_{\mathcal{D}_{\beta'}}(Z \neq \beta'),$$

and so

$$\begin{aligned}\sup_{(Q,P)\in\Pi} \mathbb{E}\mathcal{E}_Q(\hat{f}) \geq & \max_{\beta_Q=\beta_P=\beta'\in\tilde{\mathcal{H}}} \mathbb{E}_{\mathcal{D}_{\beta'}}\mathcal{E}_{Q_{\beta'}}(\hat{f}) \\ \geq & C_M^2 s\log d/(n_Q \vee n_P) \max_{\beta'\in\tilde{\mathcal{H}}} \mathbb{P}_{\mathcal{D}_{\beta'}}(Z \neq \beta') \\ \geq & C_M^2 s\log d/(n_Q \vee n_P) \frac{1}{|\tilde{\mathcal{H}}|} \sum_{\beta'\in\tilde{\mathcal{H}}} \mathbb{P}_{\mathcal{D}_{\beta'}}(Z \neq \beta').\end{aligned} \tag{73}$$

The remaining work is to lower bound $\frac{1}{|\tilde{\mathcal{H}}|}\sum_{\beta'\in\tilde{\mathcal{H}}} \mathbb{P}_{\mathcal{D}_{\beta'}}(Z \neq \beta')$ by some positive constant. By Fano's lemma where the total sample size is $n_Q + n_P$ (Note that we suppose $P = Q$), we have

$$\begin{aligned}\frac{1}{|\tilde{\mathcal{H}}|} \sum_{\beta'\in\tilde{\mathcal{H}}} \mathbb{P}_{\mathcal{D}_{\beta'}}(Z \neq \beta') \geq & 1 - \frac{\log 2 + (n_Q + n_P)\max_{\beta\neq\beta'\in\tilde{\mathcal{H}}} KL(Q_\beta, Q_{\beta'})}{\log|\tilde{\mathcal{H}}|} \\ \geq & 1 - \frac{\log 2 + (n_Q + n_P)\max_{\beta\neq\beta'\in\tilde{\mathcal{H}}} KL(Q_\beta, Q_{\beta'})}{c \cdot s\log d}\end{aligned} \tag{74}$$

since $\log|\tilde{\mathcal{H}}| \geq \frac{s-1}{2}\log\frac{d-s}{(s-1)/2} \geq c \cdot s\log d$ for some $c > 0$ small enough. For any $\beta \neq \beta' \in \tilde{\mathcal{H}}$, their Kullback-Leibler divergence satisfies

$$\begin{aligned}KL(Q_\beta, Q_{\beta'}) &= \mathbb{E}_{Q_\beta}[\psi''(X^T(t\beta + (1-t)\beta'))(X^T(\beta - \beta'))^2] \\ &\leq C_\psi \mathbb{E}_{Q_\beta}[(X^T(\beta - \beta'))^2] \\ &\leq C_\psi ||\beta - \beta'||^2 \leq C_\psi C^2(s\log d)/(n_Q \vee n_P),\end{aligned}$$

for the logistic regression link function $\psi(u) = \log(1 + e^u)$, some constant $t \in [0, 1]$ and

$$C_\psi := ||\psi''||_\infty.$$

Therefore,

$$\frac{\log 2 + (n_Q + n_P)\max_{\beta\neq\beta'\in\tilde{\mathcal{H}}} KL(Q_\beta, Q_{\beta'})}{c \cdot s\log d} \leq \frac{\log 2 + C_\psi C^2(s\log d)\frac{(n_Q+n_P)}{n_Q\vee n_P}}{c \cdot s\log d} \leq \frac{1}{2}$$

by choosing C to be small enough. With such a choice of C, we see that (73) further reduces to

$$\begin{aligned}\sup_{(Q,P)\in\Pi} \mathbb{E}\mathcal{E}_Q(\hat{f}) \geq & C_M^2 s\log d/(n_Q \vee n_P) \frac{1}{|\tilde{\mathcal{H}}|} \sum_{\beta'\in\tilde{\mathcal{H}}} \mathbb{P}_{\mathcal{D}_{\beta'}}(Z \neq \beta') \\ \geq & \frac{1}{2} C_M^2 s\log d/(n_Q \vee n_P) \\ = & \frac{1}{2} C_M^2 \frac{s\log d}{n_Q} \wedge \frac{s\log d}{n_P},\end{aligned}$$

which is just our desired result. □

Second part of Proof. Next, we would like to show that

$$\sup_{(Q,P)\in\Pi^{LR}} \mathbb{E}\mathcal{E}_Q(\hat{f}) \gtrsim \Delta^2 \wedge \frac{s\log d}{n_Q}.$$

First, we fix $\beta_P = h$ in the sense that

$$h_1 = 1, \quad h_j = 0 \quad (\forall 2 \leq j \leq d).$$

Next, we consider a class of β such that

$$||\beta||_0 \leq s,\ \beta_1 = 1,\ |\beta_j| \in \{0, C_\Delta\sqrt{\log d/n_Q} \wedge \frac{\sin\Delta}{\sqrt{s}}, -C_\Delta\sqrt{\log d/n_Q} \wedge \frac{\sin\Delta}{\sqrt{s}}\}\ (\forall 2 \leq j \leq d),$$

where the constant $C_\Delta \geq 0$ will be specified later. Denote such a class of β as $\mathcal{H}_\Delta$.

We want to verify that for any $\beta \in \mathcal{H}_\Delta$, $(Q_\beta, P_h) \in \Pi^{LR}$. Since it is obvious that $||\beta||_0 \leq s$ by definition, it suffices to show that $\langle\beta, h\rangle \leq \Delta$ from the calculation of the angle sine. It holds that

$$\sin\langle\beta, h\rangle \leq \frac{(\sum_{j=2}^{d}\mathbf{1}\{\beta \neq h\})^{\frac{1}{2}}(\frac{\sin\Delta}{\sqrt{s}})}{||h||} \leq \sin\Delta, \tag{75}$$

since $\sum_{j=2}^{d}\mathbf{1}\{\beta \neq h\} \leq s, ||h|| = 1$, and $|\beta - h| \leq \frac{\sin\Delta}{\sqrt{s}}$ if $\beta_j \neq h_j$. Also, the dot product satisfies

$$\beta^T h = 1 > 0, \tag{76}$$

which indicates that $\langle\beta, h\rangle < \frac{\pi}{2}$. Based on (75) and (76), we conclude that

$$\langle\beta, h\rangle \leq \Delta, \quad \forall\beta \in \mathcal{H}_\Delta.$$

By following the same idea in the first part of the proof, Lemma 5 of Raskutti et al. (2009) shows that there exists a subset of $\mathcal{H}_\Delta$, named $\tilde{\mathcal{H}}_\Delta$, such that $|\tilde{\mathcal{H}}_\Delta| \geq \exp(\frac{s-1}{2}\log\frac{d-s}{(s-1)/2})$ and

$$\sum_{j=2}^{d}\mathbf{1}\{\beta_j \neq \beta'_j\} \geq \frac{s-1}{2}, \quad \forall\beta, \beta' \in \tilde{\mathcal{H}}_\Delta.$$

Following the same procedure of (71) and (72), we claim that

- For any $\beta \in \tilde{\mathcal{H}}_\Delta$, its norm is bounded by the following inequality:

$$1 \leq ||\beta||_2 \leq 1 + C_\Delta\sin\Delta \leq 1 + C_\Delta. \tag{77}$$

- For any $\beta \neq \beta' \in \tilde{\mathcal{H}}_\Delta$, their angle sine could be bounded by

$$\begin{aligned}\langle\beta, \beta'\rangle \geq \sin\langle\beta, \beta'\rangle \geq & \frac{(\sum_{j=2}^{d}\mathbf{1}\{\beta_j \neq \beta'_j\})^{\frac{1}{2}}(C_\Delta\sqrt{\log d/n_Q} \wedge \frac{\sin\Delta}{\sqrt{s}})}{||\beta||} \\ \geq & \frac{C_\Delta\sqrt{s\log d/n_Q} \wedge \sin\Delta}{2(1 + C_\Delta)}.\end{aligned} \tag{78}$$

This inequality holds since as long as $\beta_j \neq \beta'_j$ for some $2 \leq j \leq d$, based on the construction of $\tilde{\mathcal{H}}_\Delta$ we have $|\beta_j - \beta'_j| \geq C_\Delta\sqrt{\log d/n_Q} \wedge \frac{\sin\Delta}{\sqrt{s}}$.

Define the random variable Z as the $\beta \in \tilde{\mathcal{H}}_\Delta$ that attains the minimum of $\mathcal{E}_{Q_\beta}(\hat{f})$, i.e.

$$Z := \arg\min_{\beta \in \tilde{\mathcal{H}}_\Delta} \mathcal{E}_{Q_\beta}(\hat{f}).$$

For any $\beta' \neq Z, \beta' \in \tilde{\mathcal{H}}_\Delta$, by Lemma E.4, we have that

$$\mathcal{E}_{Q_Z}(\hat{f}) + \mathcal{E}_{Q_{\beta'}}(\hat{f}) \geq \frac{\sigma(1+C_\Delta)(1-\sigma(1+C_\Delta))}{20\pi}\langle Z, \beta'\rangle^2.$$

by (77). The choice of Z further indicates that

$$\begin{aligned}\mathcal{E}_{Q_{\beta'}}(\hat{f}) \geq & \frac{\sigma(1+C_\Delta)(1-\sigma(1+C_\Delta))}{40\pi}\langle Z, \beta'\rangle^2 \\ \geq & \frac{\sigma(1+C_\Delta)(1-\sigma(1+C_\Delta))}{80\pi(1+C_\Delta)}(C_\Delta^2 s\log d/n_Q) \wedge \sin^2\Delta.\end{aligned}$$

We denote the constant term $\frac{\sigma(1+C_\Delta)(1-\sigma(1+C_\Delta))}{80\pi}$ by C'_Δ for simplicity. Hence,

$$\mathbb{E}_{\mathcal{D}_{\beta'}}\mathcal{E}_{Q_{\beta'}}(\hat{f}) \geq C'_\Delta((C_\Delta^2 s\log d/n_Q) \wedge \sin^2\Delta) \times \mathbb{P}_{\mathcal{D}_{\beta'}}(Z \neq \beta'),$$

and so

$$\begin{aligned}\sup_{(Q,P)\in\Pi} \mathbb{E}\mathcal{E}_Q(\hat{f}) \geq & \max_{\beta'\in\tilde{\mathcal{H}}_\Delta, \beta_P = h} \mathbb{E}_{\mathcal{D}_{\beta'}}\mathcal{E}_{Q_{\beta'}}(\hat{f}) \\ \geq & C'_\Delta((C_\Delta^2 s\log d/n_Q) \wedge \sin^2\Delta) \max_{\beta'\in\tilde{\mathcal{H}}_\Delta} \mathbb{P}_{\mathcal{D}_{\beta'}}(Z \neq \beta') \\ \geq & C'_\Delta((C_\Delta^2 s\log d/n_Q) \wedge \sin^2\Delta)\frac{1}{|\tilde{\mathcal{H}}_\Delta|}\sum_{\beta'\in\tilde{\mathcal{H}}_\Delta} \mathbb{P}_{\mathcal{D}_{\beta'}}(Z \neq \beta').\end{aligned} \tag{79}$$

By Fano's lemma where the total sample size is n_Q, we have

$$\begin{aligned}\frac{1}{|\tilde{\mathcal{H}}_\Delta|}\sum_{\beta'\in\tilde{\mathcal{H}}_\Delta} \mathbb{P}_{\mathcal{D}_{\beta'}}(Z \neq \beta') \geq & 1 - \frac{\log 2 + n_Q \max_{\beta\neq\beta'\in\tilde{\mathcal{H}}_\Delta} KL(Q_\beta, Q_{\beta'})}{\log|\tilde{\mathcal{H}}|} \\ \geq & 1 - \frac{\log 2 + n_Q \max_{\beta\neq\beta'} KL(Q_\beta, Q_{\beta'})}{c \cdot s\log d}\end{aligned} \tag{80}$$

since $\log|\tilde{\mathcal{H}}_\Delta| \geq \frac{s-1}{2}\log\frac{d-s}{(s-1)/2} \geq c \cdot s\log d$ for some $c > 0$ small enough. For any $\beta \neq \beta' \in \tilde{\mathcal{H}}_\Delta$, their Kullback-Leibler divergence satisfies

$$\begin{aligned}KL(Q_\beta, Q_{\beta'}) &= \mathbb{E}_{Q_\beta}[\psi''(X^T(t\beta + (1-t)\beta'))(X^T(\beta-\beta'))^2] \\ &\leq C_\psi \mathbb{E}_{Q_\beta}[(X^T(\beta-\beta'))^2] \\ &\leq C_\psi||\beta-\beta'||^2 \leq C_\psi C_\Delta^2\frac{s\log d}{n_Q},\end{aligned}$$

for the logistic regression link function $\psi(u) = \log(1+e^u)$ and some constants $t \in [0,1]$. Therefore,

$$\frac{\log 2 + (n_Q + n_P)\max_{\beta\neq\beta'} KL(Q_\beta, Q_{\beta'})}{c \cdot s\log d} \leq \frac{\log 2 + C_\psi C_\Delta^2 s\log d}{c \cdot s\log d} \leq \frac{1}{2}$$

by choosing C_Δ to be small enough. With such a choice of C_Δ, we see that (73) further reduces to

$$\begin{aligned}\sup_{(Q,P)\in\Pi} \mathbb{E}\mathcal{E}_Q(\hat{f}) \geq & C'_\Delta((C_\Delta^2 s\log d/n_Q)\wedge \sin^2\Delta) \max_{\beta'\in\tilde{\mathcal{H}}_\Delta} \mathbb{P}_{\mathcal{D}_{\beta'}}(Z\neq\beta') \\ \geq & \frac{1}{2}C'_\Delta((C_\Delta^2 s\log d/n_Q)\wedge \sin^2\Delta) \\ = & \frac{1}{2}C'_\Delta C_\Delta^2 \frac{s\log d}{n_Q} \wedge C'_\Delta \sin^2\Delta \\ \geq & \left(\frac{1}{2}C'_\Delta C_\Delta^2 \frac{s\log d}{n_Q}\right)\wedge\left(\frac{4C'_\Delta}{\pi^2}\Delta^2\right),\end{aligned}$$

which is just our desired result.

□

Theorem 6 is achieved simply by combining the first and second parts of the proof above.

E. Auxiliary Results

E.1. Proof of Proposition 1

Proof. Suppose that $(Q,P)\in\Pi\cap\mathcal{A}(M)$. For any $t\geq 0$, define

$$\Omega(t) := \{x\in\Omega : |\eta^Q(x)-\frac{1}{2}|\leq t\}.$$

For any $z\in(0,\frac{1}{2})$, define

$$\begin{aligned}& A(z) := \mathbb{E}_{\mathcal{D}_P}\int_{\Omega^+(\gamma,C_\gamma)\cap\Omega(z)} |\eta^Q(X)-\frac{1}{2}|\mathbf{1}\{\hat{f}^P(X)\neq f_Q^*(X)\}dQ_X, \\ & A := A(\frac{1}{2}),\ B := \mathbb{E}_{\mathcal{D}_P}\mathbb{E}_{(X,Y)\sim P}[|\eta^P(X)-\frac{1}{2}|\mathbf{1}\{\hat{f}^P(X)\neq f_P^*(X)\}].\end{aligned}$$

I. Case of $\gamma\geq 1$: By definition, ε_P is an upper bound of B, and A is an upper bound of $A(z)$, so it suffices to prove

$$A\leq 2M^{\frac{1+\alpha}{\gamma+\alpha}}C_\alpha^{\frac{\gamma-1}{\gamma+\alpha}}C_\gamma^{-\frac{1+\alpha}{\gamma+\alpha}}B^{\frac{1+\alpha}{\gamma+\alpha}}.$$

By decomposing $\Omega^+(\gamma,C_\gamma)$ into $\Omega^+(\gamma,C_\gamma)\cap\Omega(t)$ and $\Omega^+(\gamma,C_\gamma)\cap\Omega(t)^c$, we have

$$\begin{aligned}A/M \leq & \frac{1}{M}\mathbb{E}_{\mathcal{D}_P}\int_{\Omega^+(\gamma,C_\gamma)\cap\Omega(t)} |\eta^Q(X)-\frac{1}{2}|\mathbf{1}\{\hat{f}^P(X)\neq f_Q^*(X)\}dQ_X \\ & +\mathbb{E}_{\mathcal{D}_P}\int_{\Omega^+(\gamma,C_\gamma)\cap\Omega(t)^c} |\eta^Q(X)-\frac{1}{2}|\mathbf{1}\{\hat{f}^P(X)\neq f_Q^*(X)\}dP_X\end{aligned}$$

since $dQ_X\leq MdP_X$. First, by upper bounding $\mathbf{1}\{\hat{f}^P(X)\neq f_Q^*(X)\}$ by 1, we have

$$\begin{aligned}& \frac{1}{M}\mathbb{E}_{\mathcal{D}_P}\int_{\Omega^+(\gamma,C_\gamma)\cap\Omega(t)} |\eta^Q(X)-\frac{1}{2}|\mathbf{1}\{\hat{f}^P(X)\neq f_Q^*(X)\}dQ_X \\ \leq & \frac{1}{M}\int_{\Omega^+(\gamma,C_\gamma)\cap\Omega(t)} |\eta^Q(X)-\frac{1}{2}|dQ_X \\ \leq & \frac{C_\alpha}{M}tQ_X(\Omega^+(\gamma,C_\gamma)\cap\Omega(t)) \leq \frac{C_\alpha}{M}t^{1+\alpha}.\end{aligned}$$

by the margin assumption. Secondly, when $x \in \Omega^+(\gamma, C_\gamma) \cap \Omega(t)^c$, we have

$$\begin{aligned}
&t^{\gamma-1}|\eta^Q(x) - \frac{1}{2}| \leq |\eta^Q(x) - \frac{1}{2}|^\gamma \leq |\eta^P(x) - \frac{1}{2}|/C_\gamma \\
\Rightarrow&|\eta^Q(x) - \frac{1}{2}| \leq t^{1-\gamma}|\eta^P(x) - \frac{1}{2}|/C_\gamma.
\end{aligned}$$

Applying this inequality to the risk term with respect to $\Omega^+(\gamma, C_\gamma) \cap \Omega(t)^c$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_P} \int_{\Omega^+(\gamma,C_\gamma)\cap\Omega(t)} |\eta^Q(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}^P(X) \neq f_Q^*(X)\}dP_X \\
\leq&\frac{t^{1-\gamma}}{C_\gamma}\mathbb{E}_{\mathcal{D}_P} \int_{\Omega^+(\gamma,C_\gamma)\cap\Omega(t)} |\eta^P(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}^P(X) \neq f_Q^*(X)\}dP_X \\
\leq&\frac{t^{1-\gamma}}{C_\gamma}B.
\end{aligned}$$

By plugging in the two inequalities above, we have that for any $t \geq 0$,

$$A/M \leq \frac{C_\alpha}{M}t^{1+\alpha} + \frac{t^{1-\gamma}}{C_\gamma}B.$$

Since the choice of t is arbitrary, we choose $t = (\frac{MB}{C_\alpha C_\gamma})^{\gamma+\alpha}$, the bound becomes

$$A \leq 2M^{\frac{1+\alpha}{\gamma+\alpha}}C_\alpha^{\frac{\gamma-1}{\gamma+\alpha}}C_\gamma^{-\frac{1+\alpha}{\gamma+\alpha}}B^{\frac{1+\alpha}{\gamma+\alpha}}.$$

II. Case of $\gamma < 1$: By definition, ε_P is an upper bound of B, so it suffices to prove

$$A(z) \leq 2^{\gamma-1}MC_\gamma^{-1}B.$$

Note that when $x \in \Omega^+(\gamma, C_\gamma) \cap \Omega(z)$, we have

$$\begin{aligned}
&z^{\gamma-1}|\eta^Q(x) - \frac{1}{2}| \leq |\eta^Q(x) - \frac{1}{2}|^\gamma \leq |\eta^P(x) - \frac{1}{2}|/C_\gamma \\
\Rightarrow&|\eta^Q(x) - \frac{1}{2}| \leq z^{1-\gamma}|\eta^P(x) - \frac{1}{2}|/C_\gamma.
\end{aligned}$$

Therefore,

$$\begin{aligned}
A(z)/M &\leq \frac{1}{M}\mathbb{E}_{\mathcal{D}_P} \int_{\Omega^+(\gamma,C_\gamma)\cap\Omega(z)} |\eta^Q(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}^P(X) \neq f_Q^*(X)\}dQ_X \\
&\leq C_\gamma^{-1}z^{1-\gamma}\mathbb{E}_{\mathcal{D}_P} \int_{\Omega^+(\gamma,C_\gamma)\cap\Omega(z)} |\eta^P(x) - \frac{1}{2}|\mathbf{1}\{\hat{f}^P(X) \neq f_Q^*(X)\}dP_X \\
&= C_\gamma^{-1}z^{1-\gamma}B,
\end{aligned}$$

which finishes the proof. □

E.2. Proof of Proposition 3

In this subsection, $\varepsilon(z;\gamma,C_\gamma)\equiv 0$ in the parameter setting of Π^{NP}.

Proof of the first statement. The direction of $\{Q:(Q,P)\in\Pi^{NP}\}\subset\Pi^{NP}_Q$ is obvious from the definitions. Suppose $Q\in\Pi^{NP}_Q$. We always assume that $P_X=Q_X$ in this part of proof. It is then obvious that $P_X=Q_X\in\mathcal{S}(\mu^+,\mu^-,c_\mu,r_\mu)$.

If $\beta=0$, then $\beta_P=0$, and η^Q,η^P can be noncontinuous. Simply setting

$$\eta^P=\frac{1}{2}+\frac{C_\gamma}{2^\gamma}\mathrm{sgn}\left(\eta^Q-\frac{1}{2}\right)$$

satisfies that $(Q,P)\in\Pi^{NP}$ provided that $C_{\beta_P}\geq\frac{C_\gamma}{2^\gamma}$.

If $\beta>0$, and η^Q does not hit $\frac{1}{2}$ on Ω, then the continuity of η^Q indicates that $\mathrm{sgn}(\eta^Q-\frac{1}{2})$ is a fixed constant. In this case, either setting $\eta^P\equiv 1$ or $\eta^P\equiv 0$ satisfies that $(Q,P)\in\Pi^{NP}$.

At last, we prove the first statement under the conditions that $\beta>0$ and η^Q hits $\frac{1}{2}$ on Ω. If $\beta_P=0$, then we can again set $\eta^P=\frac{1}{2}+\frac{C_\gamma}{2^\gamma}\mathrm{sgn}\left(\eta^Q-\frac{1}{2}\right)$ to satisfy that $(Q,P)\in\Pi^{NP}$, so we consider $\beta_P>0$ below.

For any $x\in\Omega$, define $d(x)$ as the point $x'\in\Omega,\eta^Q(x')=\frac{1}{2}$ that is closest to x, i.e.

$$d(x):=\mathop{\arg\min}_{x'\in\Omega,\eta^Q(x')=\frac{1}{2}}||x-x'||.$$

This function is well-defined since η^Q hits $\frac{1}{2}$ on Ω. Define η^P as

$$\eta^P=\frac{1}{2}+C\mathrm{sgn}\left(\eta^Q-\frac{1}{2}\right)||x-d(x)||^{\beta_P},$$

where the constant $C>0$ will be specified later.

By the definition of P, we have $(\eta^P-\frac{1}{2})(\eta^Q-\frac{1}{2})\geq 0$. Moreover, we have

$$\begin{aligned}|\eta^P(x)-\frac{1}{2}|&\geq C||x-d(x)||^{\beta_P}\\&\geq CC_\beta^{-\frac{\beta_P}{\beta}}\left(C_\beta||x-d(x)||^\beta\right)^{\frac{\beta_P}{\beta}}\\&\geq CC_\beta^{-\frac{\beta_P}{\beta}}|\eta^Q(x)-\frac{1}{2}|^{\frac{\beta_P}{\beta}}\\&\geq CC_\beta^{-\frac{\beta_P}{\beta}}2^{\beta_P/\beta-\gamma}|\eta^Q(x)-\frac{1}{2}|^\gamma\\&\geq C_\gamma|\eta^Q(x)-\frac{1}{2}|^\gamma.\end{aligned}$$

we conclude that $\Omega^+(\gamma,C_\gamma)=\Omega$ provided that $C\geq C_\gamma C_\beta^{\frac{\beta_P}{\beta}}2^{\gamma-\beta_P/\beta}$. Hence, it suffices to obtain that

$$\eta^P\in\mathcal{H}(\beta_P,C_{\beta_P})$$

to show $(Q,P)\in\Pi^{NP}(\alpha,C_\alpha,\gamma,C_\gamma,0,\beta,\beta_P,C_\beta,C_{\beta_P},\mu^+,\mu^-,c_\mu,r_\mu)$. For any $x,x'\in\mathbb{R}^d$, we assume that $||x-d(x)||\geq||x'-d(x')||$ without loss of generality. If $(\eta^Q(x)-\frac{1}{2})(\eta^Q(x')-\frac{1}{2})\geq 0$, we have

$$\begin{aligned}|\eta^P(x)-\eta^P(x')|&\leq C|||x-d(x)||^{\beta_P}-||x'-d(x')||^{\beta_P}|\\&\leq C|||x-d(x)||-||x'-d(x')|||^{\beta_P}\\&\leq C||x-x'||^{\beta_P},\end{aligned}$$

so $\eta^P \in \mathcal{H}(\beta_P, C_{\beta_P})$ if $C_{\beta_P} \geq C$. If $(\eta^Q(x) - \frac{1}{2})(\eta^Q(x') - \frac{1}{2}) < 0$, since η^P is continuous, there exists $\lambda \in (0,1)$ such that

$$\eta^Q(\lambda x + (1-\lambda)x') = \frac{1}{2}.$$

We have

$$\begin{aligned}|\eta^P(x) - \eta^P(x')| &\leq 2C||x - d(x)||^{\beta_P} \\ &\leq 2C||x - (\lambda x + (1-\lambda)x')||^{\beta_P} \\ &\leq 2C||x - x'||^{\beta_P},\end{aligned}$$

so $\eta^P \in \mathcal{H}(\beta_P, C_{\beta_P})$ if $C_{\beta_P} \geq 2C$. Combining all cases above, we conclude that for any $Q \in \Pi_Q^{NP}$, there exists some P such that $(Q,P) \in \Pi^{NP}$ provided that

$$C_{\beta_P} \geq \max\{C_\gamma 2^{-\gamma}, 2C_\gamma C_\beta^{\frac{\beta_P}{\beta}} 2^{\gamma - \beta_P/\beta}\}.$$

□

Proof of the second statement. Let $\Omega = [0,1]^d$. Since $\beta_P > \gamma\beta \geq 0$, η^P is continuous. Define

$$\eta^Q = \frac{1}{2} + C_\beta(x_1 - \frac{1}{2})^\beta.$$

When $x_1 = \frac{1}{2}$ for any $x \in \Omega$, it holds that $\eta^Q(x) = \frac{1}{2}$. Define $\Omega_1 = \{x \in \Omega : x_1 = \frac{1}{2}\}$. We claim $\eta^P = \frac{1}{2}$ on $x \in \Omega_1$. Otherwise, it is trivial to see that there will be a small ball in Ω on which $(\eta^P - \frac{1}{2})(\eta^Q - \frac{1}{2}) < 0$, which contradicts with the fact that $\varepsilon(z; \gamma, C_\gamma) \equiv 0$.

Define the unit vector on the first coordinate as $e_1 := (1, 0, 0, \cdots, 0)$. We see that for any $t \in [0,1]$,

$$\eta^Q(e_1) = \frac{1}{2}, \quad \eta^Q(te_1) = \frac{1}{2} + C_\beta(t - \frac{1}{2})^\beta.$$

$$\eta^P(e_1) = \frac{1}{2}, \quad |\eta^P(te_1) - \frac{1}{2}| \geq C_\gamma|\eta^Q(te_1) - \frac{1}{2}|^\gamma \geq C_\gamma C_\beta^\gamma |t - \frac{1}{2}|^{\gamma\beta}. \tag{81}$$

However, since $\eta^P \in \mathcal{H}(\beta_P, C_{\beta_P})$, we have

$$|\eta^P(te_1) - \frac{1}{2}| = |\eta^P(te_1) - \eta^P(e_1)| \leq C_{\beta_P} t^{\beta_P},$$

which contradicts with (81) by choosing $t > 0$ to be small enough. □

E.3. Proof of Proposition 4

Proof. The proof is similar to the proof of Theorem 3.2 in Cai and Wei (2021), which shows that

$$\inf_{\hat{f}} \sup_{\substack{(Q,P)\in\Pi_0^{NP} \\ \Omega=\Omega_P, \beta_P=0, C_{\beta_P}=1}} \mathbb{E}\mathcal{E}_Q(\hat{f}) \gtrsim n_P^{-\frac{\beta(1+\alpha)}{2\gamma\beta+d}} \wedge n_Q^{-\frac{\beta(1+\alpha)}{2\beta+d}}. \tag{82}$$

without imposing any smoothness condition on η^P. Our objective is to show that the minimax lower bound (82) could apply to a broader class:

$$\inf_{\hat{f}} \sup_{\substack{(Q,P)\in\Pi_0^{NP} \\ \gamma\beta\geq\beta_P}} \mathbb{E}\mathcal{E}_Q(\hat{f}) \gtrsim n_P^{-\frac{\beta(1+\alpha)}{2\gamma\beta+d}} \wedge n_Q^{-\frac{\beta(1+\alpha)}{2\beta+d}}$$

with some modification on the original proof. The only extra work is to verify that our construction $P_\sigma,\ \sigma \in \{0,1\}^m$ also satisfies the smoothness condition for η^P with parameters (β_P, C_{β_P}), i.e., $(Q,P) \in \mathcal{H}(\beta, \beta_P, C_\beta, C_{\beta_P})$, given that $\beta_P < \gamma\beta$.

Define the quantities

$$r = c_r(n_P^{-\frac{1}{2\gamma\beta+d}} \wedge n_Q^{-\frac{1}{2\beta+d}}), \quad w = c_w r^d, \quad m = \lfloor c_m r^{\alpha\beta - d} \rfloor,$$

where the constants c_r, c_w, c_m will be specified later in the proof.

We consider a packing $\{x_k\}_{k=1,\cdots,m}$ with radius $2r$ in $[0,1]^d$. For an example of such construction, we divide $[0,1]^d$ into uniform small cubes with side length as $6r$, which forms a grid with at least $\lfloor (6r)^{-1} \rfloor^d$ small cubes. Since $m \ll \lfloor (6r)^{-1} \rfloor^d$, we suppose that c_m is small enough such that $m < \lfloor (6r)^{-1} \rfloor^d$. Therefore, we could assign exactly one sphere with radius $2r$ in one cube without intersection, which forms a packing of $\{x_k\}_{k=1,\cdots,m}$ with radius $2r$ in $[0,1]^d$. Also, define B_c as the compliment of these m balls, i.e. $B_c := [0,1]^d / \left(\bigcup_{k=1}^m B(x_k, 2r)\right)$.

For any $\sigma \in \{1,-1\}^m$, we consider the regression functions $\eta_\sigma^Q(x)$ and $\eta_\sigma^P(x)$ defined as follows:

$$\eta_\sigma^Q(x) = \begin{cases} \frac{1}{2} + \sigma_k C_\beta r^\beta g^\beta(\frac{||x-x_k||}{r}) & \text{if } x \in B(x_k, 2r) \text{ for some } k = 1, \cdots, m \\ \frac{1}{2} & \text{otherwise,} \end{cases}$$

$$\eta_\sigma^P(x) = \begin{cases} \frac{1}{2} + \sigma_k (C_{\beta_P} \vee C_\gamma C_\beta^\gamma) r^{\gamma\beta} g^{\gamma\beta}(\frac{||x-x_k||}{r}) & \text{if } x \in B(x_k, 2r) \text{ for some } k = 1, \cdots, m \\ \frac{1}{2} & \text{otherwise,} \end{cases}$$

where the function $g(\cdot)$ is defined as $g(x) = \min\{1, 2-x\}$ on $x \in [0,2]$.

The construction of the marginal distributions $Q_{\sigma,X}$ and $P_{\sigma,X}$ is as follows. Let $Q_{\sigma,X} = P_{\sigma,X}$, both of which have the density function $\mu(\cdot)$ defined as

$$\mu(x) = \begin{cases} \frac{w}{\lambda[B(x_k,r)]} & \text{if } x \in B(x_k, r) \text{ for some } k = 1, \cdots, m \\ \frac{1-mw}{1-m\lambda[B(x_k,2r)]} & \text{if } x \in B_c \\ 0 & \text{otherwise.} \end{cases}$$

Given the the construction of (Q_σ, P_σ), we next verify that (Q_σ, P_σ) belongs to Π_0^{NP} for any $\sigma \in \{1,-1\}^m$. The fact that $\Omega^+(\gamma, C_\gamma) = \Omega$ is trivial by the construction of the regression functions $\eta_\sigma^Q(x)$ and $\eta_\sigma^P(x)$, so we omit its verification.

Verify Margin Assumption: We have that

$$\begin{aligned} Q_\sigma(0 < |\eta_\sigma^Q - \frac{1}{2}| < t) &= m Q_\sigma(0 < C_\beta r^\beta g^\beta(\frac{||X - x_1||}{r}) \leq t) \\ &= m Q_\sigma(0 < g(\frac{||X - x_1||}{r}) \leq (\frac{t}{C_\beta r^\beta})^{\frac{1}{\beta}}) \\ &= m w \mathbf{1}\{t \geq C_\beta r^\beta\} \\ &\leq c_m c_w r^{\alpha\beta} \mathbf{1}\{t \geq C_\beta r^\beta\} \\ &\leq C_\alpha t^\alpha. \end{aligned}$$

given that c_m is small enough. Therefore, $Q_\sigma \in \mathcal{M}(\alpha, C_\alpha)$.

Verify Smoothness Assumption: It is easy to see that for any $a, b \in [0, 2]$ we have

$$|g^{\beta}(a) - g^{\beta}(b)| \leq |a-b|^{\beta}$$

and

$$|g^{\gamma\beta}(a) - g^{\gamma\beta}(b)| \leq \begin{cases} |a-b|^{\gamma\beta} & \gamma\beta \leq 1 \\ 1-(1-|a-b|)^{\gamma\beta} \leq \gamma\beta|a-b| & \gamma\beta > 1. \end{cases}$$

Thus, for any $x, x' \in B(x_k, 2r)$, we obtain from the triangular inequality $||x-x_k|| - ||x'-x_k|| \leq ||x-x'||$ that

$$|r^{\beta} g^{\beta}(||x-x_k||/r)) - r^{\beta} g^{\beta}(||x'-x_k||/r))| \leq ||x-x'||^{\beta} \tag{83}$$

and

$$|r^{\gamma\beta} g^{\gamma\beta}(||x-x_k||/r)) - r^{\gamma\beta} g^{\gamma\beta}(||x'-x_k||/r))| \leq \begin{cases} ||x-x'||^{\gamma\beta} & \gamma\beta \leq 1 \\ \gamma\beta \cdot r^{\gamma\beta}||x-x'|| & \gamma\beta > 1. \end{cases} \tag{84}$$

Suppose that c_r is small enough such that for any $n_Q, n_P \geq 1$, we have

$$\max\{\gamma\beta \cdot r^{\gamma\beta}(4r)^{1-\beta_P}, (4r)^{\gamma\beta-\beta_P}\} \leq 1,$$

then we further deduces that

$$\begin{aligned} ||x-x'||^{\gamma\beta} \leq ||x-x'||^{\beta_P}(4r)^{\gamma\beta-\beta_P} \leq ||x-x'||^{\beta_P} \\ \gamma\beta \cdot r^{\gamma\beta}||x-x'|| \leq \gamma\beta \cdot r^{\gamma\beta}(4r)^{1-\beta_P}||x-x'||^{\beta_P} \leq ||x-x'||^{\beta_P}. \end{aligned} \tag{85}$$

On one hand, by the definition of η_{σ}^{Q} and (83), we see that

$$|\eta_{\sigma}^{Q}(x) - \eta_{\sigma}^{Q}(x')| \leq C_{\beta}||x-x'||^{\beta}, \quad \forall x, x \in [0,1]^d.$$

One the other hand, by the definition of η_{σ}^{P}, (84) and (85), we see that

$$|\eta_{\sigma}^{P}(x) - \eta_{\sigma}^{P}(x')| \leq C_{\beta_P}||x-x'||^{\gamma\beta} \leq C_{\beta_P}||x-x'||^{\beta_P}, \quad \forall x, x \in [0,1]^d.$$

Hence, $(Q, P) \in \mathcal{H}(\beta, \beta_P, C_{\beta}, C_{\beta_P})$.

Verify Strong Density Condition: If $x \in B(x_k, r)$ for some $k = 1, \cdots, m$, we have

$$\mu(x) = \frac{w}{\lambda[B(x_k, r)]} = c_w/\pi_d.$$

If $x \in B_c$, we have

$$\mu(x) = \frac{1 - c_w\lfloor c_m r^{\alpha\beta-d}\rfloor r^d}{1 - 2^d\pi_d\lfloor c_m r^{\alpha\beta-d}\rfloor r^d} \overset{n_Q\to\infty}{\longrightarrow} 1.$$

Therefore, we could set $c_w \in [\pi_d\mu^-, \pi_d\mu^+]$ to satisfy the condition $(Q_\sigma, P_\sigma) \in \mathcal{S}(\mu^+, \mu^-, c_\mu, r_\mu)$.

Putting all verification steps above together, we conclude that $(Q_\sigma, P_\sigma) \in \Pi_0^{NP}$. We finish the proof of this part by applying Assouad's lemma to (Q_σ, P_σ), $\forall \sigma \in \{1, -1\}^m$.

If $\sigma, \sigma' \in \{1, -1\}^m$ differ only at one coordinate, i.e.

$$\sigma_k = -\sigma'_k, \quad \sigma_l = \sigma'_l \ (\forall l \neq k),$$

we have the Hellinger distance bound as

$$\begin{aligned}
H^2(Q_\sigma, Q_{\sigma'}) &= \frac{1}{2}\int \left(\sqrt{\eta^Q_\sigma(X)} - \sqrt{\eta^Q_{\sigma'}(X)}\right)^2 + \left(\sqrt{1-\eta^Q_\sigma(X)} - \sqrt{1-\eta^Q_{\sigma'}(X)}\right)^2 dQ_X \\
&= \int_{B(x_k,r)} \frac{w}{\lambda[B(x_k,r)]}\left(\sqrt{\frac{1}{2}+C_\beta r^\beta} - \sqrt{\frac{1}{2}-C_\beta r^\beta}\right)^2 dx \\
&= \frac{1}{2}w(1-\sqrt{1-2C_\beta^2 r^{2\beta}}) \\
&\le C_\beta^2 w r^{2\beta}
\end{aligned}$$

$$\begin{aligned}
H^2(P_\sigma, P_{\sigma'}) &= \frac{1}{2}\int \left(\sqrt{\eta^P_\sigma(X)} - \sqrt{\eta^P_{\sigma'}(X)}\right)^2 + \left(\sqrt{1-\eta^P_\sigma(X)} - \sqrt{1-\eta^P_{\sigma'}(X)}\right)^2 dP_X \\
&= \int_{B(x_k,r)} \frac{w}{\lambda[B(x_k,r)]}\left(\sqrt{\frac{1}{2}+(C_{\beta_P}\vee C_\gamma C_\beta^\gamma)r^{\gamma\beta}} - \sqrt{\frac{1}{2}-(C_{\beta_P}\vee C_\gamma C_\beta^\gamma)r^{\gamma\beta}}\right)^2 dx \\
&= \frac{1}{2}w(1-\sqrt{1-2(C_{\beta_P}^2\vee C_\gamma^2 C_\beta^{2\gamma})r^{2\gamma\beta}}) \\
&\le (C_{\beta_P}^2\vee C_\gamma^2 C_\beta^{2\gamma})r^{2\gamma\beta}
\end{aligned}$$

Recall that $r = c_r(n_P^{-\frac{1}{2\gamma\beta+d}} \wedge n_Q^{-\frac{1}{2\beta+d}})$. By the property of Hellinger distance, we have

$$\begin{aligned}
H^2(\mathbb{P}^\sigma_{\mathcal{D}_Q}\times\mathbb{P}^\sigma_{\mathcal{D}_P}, \mathbb{P}^{\sigma'}_{\mathcal{D}_Q}\times\mathbb{P}^{\sigma'}_{\mathcal{D}_P}) \le& n_Q H^2(Q_\sigma, Q_{\sigma'}) + n_P H^2(P_\sigma, P_{\sigma'}) \\
\le& C_\beta^2 w n_Q r^{2\beta} + (C_{\beta_P}^2\vee C_\gamma^2 C_\beta^{2\gamma}) n_P r^{2\gamma\beta} \\
\le& C_\beta^2 c_r^{2\beta+d} c_w + (C_{\beta_P}^2\vee C_\gamma^2 C_\beta^{2\gamma}) c_r^{2\gamma\beta+d} c_w \\
\le& \frac{\sqrt{2}}{4}
\end{aligned}$$

provided that c_r is small enough. This further indicates that

$$TV(\mathbb{P}^\sigma_{\mathcal{D}_Q}\times\mathbb{P}^\sigma_{\mathcal{D}_P}, \mathbb{P}^{\sigma'}_{\mathcal{D}_Q}\times\mathbb{P}^{\sigma'}_{\mathcal{D}_P}) \le \sqrt{2}H^2(\mathbb{P}^\sigma_{\mathcal{D}_Q}\times\mathbb{P}^\sigma_{\mathcal{D}_P}, \mathbb{P}^{\sigma'}_{\mathcal{D}_Q}\times\mathbb{P}^{\sigma'}_{\mathcal{D}_P}) \le \frac{1}{2}. \tag{86}$$

For any empirical classifier $\hat{f}$, we have

$$\begin{aligned}
\mathcal{E}_{Q_\sigma}(\hat{f}) + \mathcal{E}_{Q_{\sigma'}}(\hat{f}) &= 2\mathbb{E}_{Q_\sigma}[|\eta^{Q_\sigma}(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}(X) \ne f^*_{Q_\sigma}(X)\}] \\
&\quad + 2\mathbb{E}_{Q_{\sigma'}}[|\eta^{Q_{\sigma'}}(X) - \frac{1}{2}|\mathbf{1}\{\hat{f}(X) \ne f^*_{Q_{\sigma'}}(X)\}] \\
&\ge 2\int_{B(x_k,r)} \frac{w}{\lambda[B(x_k,r)]} C_\beta r^\beta(\mathbf{1}\{\hat{f}(X) \ne f^*_{Q_\sigma}(X)\} + \mathbf{1}\{\hat{f}(X) \ne f^*_{Q_{\sigma'}}(X)\})dx \\
&= 2C_\beta w r^\beta.
\end{aligned}$$

Combining this lower bound with (86), the Assouad's lemma shows that

$$\sup_{\substack{(Q,P)\in\Pi_0^{NP} \\ \Omega=\Omega_P, \gamma\beta\ge\beta_P}} \mathbb{E}\mathcal{E}_Q(\hat{f}) \ge \sup_{\substack{(Q_\sigma,P_\sigma) \\ \sigma\in\{1,-1\}^m}} \mathbb{E}\mathcal{E}_{Q_\sigma}(\hat{f}) \ge \frac{1}{2}C_\beta m w r^\beta \gtrsim r^{\beta(1+\alpha)} \gtrsim n_P^{-\frac{\beta(1+\alpha)}{2\gamma\beta+d}} \wedge n_Q^{-\frac{\beta(1+\alpha)}{2\beta+d}}.$$

□

E.4. Proof of Proposition 5

Proof. Our objective is to show that the minimax lower bound satisfies

$$\inf_{\hat{f}} \sup_{\substack{(Q,P)\in\Pi_0^{NP}\\ \gamma\beta<\beta_P}} \mathbb{E}\mathcal{E}_Q(\hat{f}) \gtrsim n_P^{-\frac{\beta_P(1+\alpha)/\gamma}{2\beta_P+d}} \wedge n_Q^{-\frac{\beta_P(1+\alpha)/\gamma}{2\beta_P/\gamma+d}}.$$

Define the quantities

$$r_Q = c_r(n_P^{-\frac{\beta_P/\gamma\beta}{2\beta_P+d}} \wedge n_Q^{-\frac{1}{2\beta+d\gamma\beta/\beta_P}}), \quad r_P = r_Q^{\gamma\beta/\beta_P} = c_r^{\gamma\beta/\beta_P}(n_P^{-\frac{1}{2\beta_P+d}} \wedge n_Q^{-\frac{1}{2\beta_P/\gamma+d}}),$$

$$w = c_w r_Q^d, \quad m_0 = \lfloor c_{m_0}\frac{r_P^d}{r_Q^d}\rfloor, \quad m = \lfloor c_m r_P^{\alpha\beta_P/\gamma-d}\rfloor,$$

where the constants c_r, c_w, c_{m_0}, c_m will be specified later in the proof.

Similar to the proof of the case of $\gamma\beta \geq \beta_P$, we consider a packing $\{x_k\}_{k=1,\cdots,m}$ with radius $2r_P$ in $[0,1]^d$. Define B_c as the compliment of these m balls, i.e.

$$B_c := [0,1]^d / \left(\bigcup_{k=1}^m B(x_k, 2r_P)\right).$$

Next, we further consider a packing $\{x_{k,l}\}_{l=1,\cdots,m_0}$ for any $k = 1,\cdots,m$ with radius $2r_Q$. By scaling the radius of the balls, the feasibility of our considered packing reduces to finding a packing $\{x_l\}_{l=1,\cdots,m_0}$ of $B(0,1)$ with radius r_Q/r_P. We denote r_Q/r_P by R, and m_0 then becomes $\lfloor c_{m_0}R^{-d}\rfloor$.

We provide an example of the reduced form of packing as follows. Consider the inscribed cube of $B(0,1)$ with diagonal length 2 and side length $\frac{2}{\sqrt{d}}$. We divide this inscribed cube into uniform small cubes with side length as $6R$, which forms a grid with at least

$$\lfloor\frac{2}{\sqrt{d}}(6R)^{-1}\rfloor^d \geq \lfloor c\frac{2}{\sqrt{d}}(6R)^{-d}\rfloor$$

small cubes for some constant $c > 0$. Therefore, provided that $c_{m_0} \leq \frac{2c}{6^d\sqrt{d}}$ is small enough, we could assign exactly one sphere with radius $2R$ in one cube without intersection, which forms a packing $\{x_l\}_{l=1,\cdots,m_0}$ with radius $2R$. Hence, we can find a packing $\{x_{k,l}\}_{l=1,\cdots,m_0}$ for any $k = 1,\cdots,m$ with radius $2r_Q$.

For any $\sigma \in \{1,-1\}^m$, we consider the regression functions $\eta_\sigma^Q(x)$ and $\eta_\sigma^P(x)$ defined as follows:

$$\eta_\sigma^Q(x) = \begin{cases} \frac{1}{2}\left(1+\sigma_k(C_\beta \wedge (\frac{C_{\beta_P}}{C_\gamma})^{\frac{1}{\gamma}})r_Q^\beta g^\beta(\frac{||x-x_{k,l}||}{r_Q})\right) & \text{if } x \in B(x_{k,l}, 2r_Q) \text{ for some } k,l \\ \frac{1}{2} & \text{otherwise,} \end{cases} \tag{87}$$

$$\eta_\sigma^P(x) = \begin{cases} \frac{1}{2}\left(1+\sigma_k C_{\beta_P} r_P^{\beta_P} g^{\beta_P}(\frac{||x-x_k||}{r_P})\right) & \text{if } x \in B(x_k, 2r_P) \text{ for some } k \\ \frac{1}{2} & \text{otherwise,} \end{cases} \tag{88}$$

where the function $g(\cdot)$ is defined as $g(x) = \min\{1, 2-x\}$ on $x \in [0,2]$.

The construction of the marginal distributions $Q_{\sigma,X}$ and $P_{\sigma,X}$ is as follows. Let $Q_{\sigma,X} = P_{\sigma,X}$, both of which have the density function $\mu(\cdot)$ defined as

$$\mu(x) = \begin{cases} \frac{w}{\lambda[B(x_k,r_Q)]} & \text{if } x \in B(x_{k,l}, r) \text{ for some } k, l \\ \frac{1-mm_0w}{1-m\lambda[B(x_k,2r_P)]} & \text{if } x \in B_c \\ 0 & \text{otherwise.} \end{cases}$$

Given the the construction of (Q_σ, P_σ), we next verify that (Q_σ, P_σ) belongs to Π_0^{NP} for any $\sigma \in \{1, -1\}^m$.

Verify Margin Assumption: We have that

$$\begin{aligned} Q_\sigma(0 < |\eta_\sigma^Q - \frac{1}{2}| < t) &= mm_0 Q_\sigma(0 < (C_\beta \wedge (\frac{C_{\beta_P}}{C_\gamma})^{\frac{1}{\gamma}}) r_Q^\beta g^\beta(\frac{||X - x_{1,1}||}{r_Q}) \leq 2t) \\ &= mm_0 Q_\sigma(0 < g(\frac{||X - x_1||}{r_Q}) \leq (\frac{2t}{(C_\beta \wedge (\frac{C_{\beta_P}}{C_\gamma})^{\frac{1}{\gamma}}) r_Q^\beta})^{\frac{1}{\beta}}) \\ &= mm_0 w \mathbf{1}\{t \geq \frac{1}{2}(C_\beta \wedge (\frac{C_{\beta_P}}{C_\gamma})^{\frac{1}{\gamma}}) r_Q^\beta\} \\ &\leq c_m c_{m_0} c_w r_P^{\alpha\beta_P/\gamma} \mathbf{1}\{t \geq \frac{1}{2}(C_\beta \wedge (\frac{C_{\beta_P}}{C_\gamma})^{\frac{1}{\gamma}}) r_Q^\beta\} \\ &= c_m c_{m_0} c_w r_Q^{\alpha\beta} \mathbf{1}\{t \geq \frac{1}{2}(C_\beta \wedge (\frac{C_{\beta_P}}{C_\gamma})^{\frac{1}{\gamma}}) r_Q^\beta\} \\ &\leq C_\alpha t^\alpha. \end{aligned}$$

given that c_m is small enough. Therefore, $Q_\sigma \in \mathcal{M}(\alpha, C_\alpha)$.

Verify Smoothness Assumption: It is easy to see that for any $a, b \in [0, 2]$ we have

$$|g^\beta(a) - g^\beta(b)| \leq |a - b|^\beta, \quad |g^{\beta_P}(a) - g^{\beta_P}(b)| \leq |a - b|^{\beta_P}$$

Thus, for any $x, x' \in B(x_{k,l}, 2r_Q)$, we obtain from the triangular inequality $||x - x_k|| - ||x' - x_k|| \leq ||x - x'||$ that

$$|r_Q^\beta g^\beta(||x - x_{k,l}||/r_Q)) - r_Q^\beta g^\beta(||x' - x_{k,l}||/r_Q))| \leq ||x - x'||^\beta. \tag{89}$$

By the definition of η_σ^Q and (89), we see that

$$|\eta_\sigma^Q(x) - \eta_\sigma^Q(x')| \leq C_\beta ||x - x'||^\beta, \quad \forall x, x \in [0, 1]^d.$$

Plus, for any $x, x' \in B(x_k, 2r_P)$, we obtain from the triangular inequality $||x - x_k|| - ||x' - x_k|| \leq ||x - x'||$ that

$$|r_P^{\beta_P} g^{\beta_P}(||x - x_k||/r_P)) - r_P^{\beta_P} g^{\beta_P}(||x' - x_k||/r_P))| \leq ||x - x'||^{\beta_P}. \tag{90}$$

By the definition of η_σ^P and (90), we see that

$$|\eta_\sigma^P(x) - \eta_\sigma^P(x')| \leq C_{\beta_P} ||x - x'||^{\beta_P}, \quad \forall x, x \in [0, 1]^d.$$

Hence, $(Q,P) \in \mathcal{H}(\beta, \beta_P, C_\beta, C_{\beta_P})$.

Verify Strong Density Condition: If $x \in B(x_{k,l}, r)$ for some $k = 1, \cdots, m$ and $l = 1, \cdots, m_0$, we have

$$\mu(x) = \frac{w}{\lambda[B(x_{k,l}, r)]} = c_w/\pi_d.$$

If $x \in B_c$, we have

$$\begin{aligned}\mu(x) =& \frac{1 - c_w \lfloor c_{m_0} \frac{r_P^d}{r_Q^d} \rfloor \lfloor c_m r_P^{\alpha\beta_P/\gamma - d} \rfloor r_Q^d}{1 - 2^d \pi_d \lfloor c_m r_P^{\alpha\beta_P/\gamma - d} \rfloor r_P^d} \\ =& \frac{1 - c_w \lfloor c_{m_0} r_Q^{d\gamma\beta/\beta_P - d} \rfloor \lfloor c_m r_Q^{\alpha\beta - d\gamma\beta/\beta_P} \rfloor r_Q^d}{1 - 2^d \pi_d \lfloor c_m r_P^{\alpha\beta - d\gamma\beta/\beta_P} \rfloor r_Q^{d\gamma\beta/\beta_P}} \xrightarrow{n_Q \to \infty} 1.\end{aligned}$$

Therefore, we could set $c_w \in [\pi_d \mu^-, \pi_d \mu^+]$ to satisfy the condition $(Q_\sigma, P_\sigma) \in \mathcal{S}(\mu^+, \mu^-, c_\mu, r_\mu)$.

Verify Signal Transfer Set: We show that $\Omega^+(\gamma, C_\gamma) = \Omega$ for our construction of (Q_σ, P_σ). By the construction of the marginal distribution pair (Q_X^σ, P_X^σ) and the definition of η_σ^Q and η_σ^P in (87) and (88), for any $x \in \Omega = B_c \cup \left(\bigcup_{k=1}^m B(x_{k,l}, r_Q)\right)$,

- If $x \in B(x_{k,l}, r)$ for some $k = 1, \cdots, m$ and $l = 1, \cdots, m_0$, we have

  $$\eta_\sigma^Q(x) - \frac{1}{2} = \sigma_k (C_\beta \wedge (\frac{C_{\beta_P}}{C_\gamma})^{\frac{1}{\gamma}}) r_Q^\beta, \quad \eta_\sigma^P(x) - \frac{1}{2} = \sigma_k C_{\beta_P} r_P^{\beta_P} = \sigma_k C_{\beta_P} r_Q^{\gamma\beta}.$$

  Therefore,

  $$\operatorname{sgn}\left(\eta^Q(x) - \frac{1}{2}\right) \times (\eta^P(x) - \frac{1}{2}) \geq C_\gamma |\eta^Q(x) - \frac{1}{2}|^\gamma.$$

- If $x \in B_c$, we have $\eta_\sigma^Q(x) = \eta_\sigma^P(x) = \frac{1}{2}$. Therefore,

  $$\operatorname{sgn}\left(\eta^Q(x) - \frac{1}{2}\right) \times (\eta^P(x) - \frac{1}{2}) \geq C_\gamma |\eta^Q(x) - \frac{1}{2}|^\gamma.$$

Putting all verification steps above together, we conclude that $(Q_\sigma, P_\sigma) \in \Pi_0^{NP}$. We finish the proof of this part by applying Assouad's lemma to (Q_σ, P_σ), $\forall \sigma \in \{1, -1\}^m$.

If $\sigma, \sigma' \in \{1, -1\}^m$ differ only at one coordinate, i.e.

$$\sigma_k = -\sigma_k', \quad \sigma_l = \sigma_l' \ (\forall l \neq k),$$

we have the Hellinger distance bound as

$$\begin{aligned}H^2(Q_\sigma, Q_{\sigma'}) &= \frac{1}{2} \int \left(\sqrt{\eta_\sigma^Q(X)} - \sqrt{\eta_{\sigma'}^Q(X)}\right)^2 + \left(\sqrt{1 - \eta_\sigma^Q(X)} - \sqrt{1 - \eta_{\sigma'}^Q(X)}\right)^2 dQ_X \\ &= \sum_{l=1}^{m_0} \int_{B(x_{k,l}, r_Q)} \frac{w}{\lambda[B(x_{k,l}, r_Q)]} \left(\sqrt{\frac{1}{2} + (C_\beta \wedge (\frac{C_{\beta_P}}{C_\gamma})^{\frac{1}{\gamma}}) r_Q^\beta} - \sqrt{\frac{1}{2} - (C_\beta \wedge (\frac{C_{\beta_P}}{C_\gamma})^{\frac{1}{\gamma}}) r_Q^\beta}\right)^2 dx \\ &= \frac{1}{2} m_0 w (1 - \sqrt{1 - 2(C_\beta^2 \wedge (\frac{C_{\beta_P}}{C_\gamma})^{\frac{2}{\gamma}}) r_Q^{2\beta}}) \\ &\leq (C_\beta^2 \wedge (\frac{C_{\beta_P}}{C_\gamma})^{\frac{2}{\gamma}}) m_0 w r_Q^{2\beta}\end{aligned}$$

$$
\begin{aligned}
H^2(P_\sigma, P_{\sigma'}) &= \frac{1}{2}\int \left(\sqrt{\eta^P_\sigma(X)} - \sqrt{\eta^P_{\sigma'}(X)}\right)^2 + \left(\sqrt{1-\eta^P_\sigma(X)} - \sqrt{1-\eta^P_{\sigma'}(X)}\right)^2 dP_X \\
&= \sum_{l=1}^{m_0} \int_{B(x_{k,l}, r_Q)} \frac{w}{\lambda[B(x_{k,l}, r_Q)]} \left(\sqrt{\frac{1}{2} + C_{\beta_P} r_P^{\beta_P}} - \sqrt{\frac{1}{2} - C_{\beta_P} r_P^{\beta_P}}\right)^2 dx \\
&= \frac{1}{2} m_0 w (1 - \sqrt{1 - 2C_{\beta_P}^2 r_P^{2\beta_P}}) \\
&\leq m_0 w C_{\beta_P}^2 r_P^{2\beta_P}
\end{aligned}
$$

By the property of Hellinger distance, we have

$$
\begin{aligned}
H^2(\mathbb{P}^\sigma_{\mathcal{D}_Q} \times \mathbb{P}^\sigma_{\mathcal{D}_P}, \mathbb{P}^{\sigma'}_{\mathcal{D}_Q} \times \mathbb{P}^{\sigma'}_{\mathcal{D}_P}) \leq & n_Q H^2(Q_\sigma, Q_{\sigma'}) + n_P H^2(P_\sigma, P_{\sigma'}) \\
\leq & (C_\beta^2 \wedge (\frac{C_{\beta_P}}{C_\gamma})^{\frac{2}{\gamma}}) m_0 w n_Q r_Q^{2\beta} + m_0 w C_{\beta_P}^2 n_P r_P^{2\beta_P} \\
\leq & c_w c_{m_0} (C_\beta^2 \wedge (\frac{C_{\beta_P}}{C_\gamma})^{\frac{2}{\gamma}}) c_r^{2\beta} + c_w c_{m_0} C_{\beta_P}^2 c_r^{2\gamma\beta} \\
\leq & \frac{\sqrt{2}}{4}
\end{aligned}
$$

provided that c_r is small enough. This further indicates that

$$
TV(\mathbb{P}^\sigma_{\mathcal{D}_Q} \times \mathbb{P}^\sigma_{\mathcal{D}_P}, \mathbb{P}^{\sigma'}_{\mathcal{D}_Q} \times \mathbb{P}^{\sigma'}_{\mathcal{D}_P}) \leq \sqrt{2} H^2(\mathbb{P}^\sigma_{\mathcal{D}_Q} \times \mathbb{P}^\sigma_{\mathcal{D}_P}, \mathbb{P}^{\sigma'}_{\mathcal{D}_Q} \times \mathbb{P}^{\sigma'}_{\mathcal{D}_P}) \leq \frac{1}{2}. \tag{91}
$$

For any empirical classifier $\hat{f}$, we have

$$
\begin{aligned}
\mathcal{E}_{Q_\sigma}(\hat{f}) + \mathcal{E}_{Q_{\sigma'}}(\hat{f}) &= 2\mathbb{E}_{Q_\sigma}[|\eta^{Q_\sigma}(X) - \frac{1}{2}| \mathbf{1}\{\hat{f}(X) \neq f^*_{Q_\sigma}(X)\}] \\
&\quad + 2\mathbb{E}_{Q_{\sigma'}}[|\eta^{Q_{\sigma'}}(X) - \frac{1}{2}| \mathbf{1}\{\hat{f}(X) \neq f^*_{Q_{\sigma'}}(X)\}] \\
&\geq 2\sum_{l=1}^{m_0} \int_{B(x_{k,l}, r_Q)} \frac{w}{\lambda[B(x_{k,l}, r_Q)]} (C_\beta \wedge (\frac{C_{\beta_P}}{C_\gamma})^{\frac{1}{\gamma}}) m_0 w r_Q^\beta dx \\
&= 2(C_\beta \wedge (\frac{C_{\beta_P}}{C_\gamma})^{\frac{1}{\gamma}}) m_0 w r_Q^\beta.
\end{aligned}
$$

Combining this lower bound with (91), the Assouad's lemma shows that

$$
\begin{aligned}
\sup_{\substack{(Q,P)\in\Pi_0^{NP} \\ \Omega=\Omega_P, \gamma\beta<\beta_P}} \mathbb{E}\mathcal{E}_Q(\hat{f}) \geq \sup_{\substack{(Q_\sigma, P_\sigma) \\ \sigma\in\{1,-1\}^m}} \mathbb{E}\mathcal{E}_{Q_\sigma}(\hat{f}) &\geq \frac{1}{2}(C_\beta \wedge (\frac{C_{\beta_P}}{C_\gamma})^{\frac{1}{\gamma}}) m m_0 w r_Q^\beta \\
&\gtrsim r_Q^{\beta(1+\alpha)} \gtrsim n_P^{-\frac{\beta_P(1+\alpha)/\gamma}{2\beta_P+d}} \wedge n_Q^{-\frac{\beta_P(1+\alpha)/\gamma}{2\beta_P/\gamma+d}}.
\end{aligned}
$$

□

E.5. Proof of Lemma D.1 and D.2

Proof of Lemma D.1. For any $i = 1, \cdots, n_Q$, define the residual as $\varepsilon_i := \psi'(\beta_Q^T X_i) - Y_i \in [-1, 1]$. Define $V_{ij} := X_{ij}\varepsilon_i$, where X_{ij} is the $j-$th covariate of the i-th observation X_i. For any $t \in \mathbb{R}$, we compute the cumulant function

$$\begin{aligned}\log \mathbb{E}_{\mathcal{D}_Q}[\exp(tV_{ij})|X_i] &= \log\left(\mathbb{E}_{\mathcal{D}_Q}[\exp(t)X_{ij}Y_i]\exp(-tX_{ij}\psi'(\beta_Q^T X_i))\right)\\ &= \psi(tX_{ij} + \beta_Q^T X_i) - \psi(\beta_Q^T X_i) - \psi'(\beta_Q^T X_i)tX_{ij}.\end{aligned}$$

Hence, by the second-order Taylor series expansion, we have

$$\log \mathbb{E}_{\mathcal{D}_Q}[\exp(tV_{ij})|X_i] = \frac{t^2}{2}X_{ij}^2\psi''(\beta_Q^T X_i + \xi_i tX_{ij}) \le \frac{t^2}{8}(\frac{1}{n_Q}\sum_{i=1}^{n_Q} X_{ij}^2). \tag{92}$$

for some $\xi_i \in [0,1]$. Since $\mathbb{E}_{\mathcal{D}_Q}[X_{ij}^2] = 1$ and X_{ij}^2 is sub-exponential since $X_{ij} \sim N(0,1)$, the tail bound for the independent sum of sub-exponential random variables reads

$$\mathbb{P}_{\mathcal{D}_Q}(\frac{1}{n_Q}\sum_{i=1}^{n_Q} X_{ij}^2 \ge 2) \le 2\exp(-n_Q/4).$$

We denote the event $E := \{\max_{j=1,\cdots,d} \frac{1}{n_Q}\sum_{i=1}^{n_Q} X_{ij}^2 \le 2\}$. The union bounds give

$$\mathbb{P}_{\mathcal{D}_Q}(E^c) \le \sum_{j=1}^{d}\mathbb{P}_{\mathcal{D}_Q}(\frac{1}{n_Q}\sum_{i=1}^{n_Q} X_{ij}^2 \ge 2) \le 2d\exp(-n_Q/4) \tag{93}$$

By (92), we have that on the event E,

$$\log \mathbb{E}_{\mathcal{D}_Q}[\exp(tV_{ij})|X_i] \le \frac{t^2}{4},$$

and we obtain by the Chernoff bound that

$$\mathbb{P}_{\mathcal{D}_Q}(|\frac{1}{n_Q}\sum_{i=1}^{n_Q} V_{ij}| \ge t|E) \le 2\exp(-4n_Q t^2),$$

and the union bound is as follows:

$$\mathbb{P}_{\mathcal{D}_Q}(\max_{j=1,\cdots,d}|\frac{1}{n_Q}\sum_{i=1}^{n_Q} V_{ij}| \ge t|E) \le 2d\exp(-4n_Q t^2). \tag{94}$$

We set $t = \frac{\sqrt{K+1}}{2}\sqrt{\frac{\log d}{n_Q}}$, then (94) becomes

$$\mathbb{P}_{\mathcal{D}_Q}(\max_{j=1,\cdots,d}|\frac{1}{n_Q}\sum_{i=1}^{n_Q} V_{ij}| \ge \frac{\sqrt{K+1}}{2}\sqrt{\frac{\log d}{n_Q}}|E) \le 2d^{-K}. \tag{95}$$

Putting (93) and (95) together, we have

$$\mathbb{P}_{\mathcal{D}_Q}(\max_{j=1,\cdots,d}|\frac{1}{n_Q}\sum_{i=1}^{n_Q} V_{ij}| \ge \frac{\sqrt{K+1}}{2}\sqrt{\frac{\log d}{n_Q}}) \le 2d^{-K} + 2d\exp(-n_Q/4). \tag{96}$$

Denote

$$E_Q := \max_{j=1,\cdots,d} |\frac{1}{n_Q}\sum_{i=1}^{n_Q} V_{ij}| \leq \frac{\sqrt{K+1}}{2}\sqrt{\frac{\log d}{n_Q}},$$

then we have $\mathbb{P}_{\mathcal{D}_Q}(E_Q^c) \leq 2d^{-K} + 2d\exp(-n_Q/4)$.

Define the empirical loss function as

$$l_Q(\beta) := \frac{1}{n_Q}\sum_{i=1}^{n_Q}\left\{\log(1+e^{X_i^T\beta}) - Y_iX_i^T\beta\right\}.$$

From straightforward calculation,we have

$$\nabla l_Q(\beta) = \frac{1}{n_Q}\sum_{i=1}^{n_Q}(\psi'(\beta^TX_i) - Y_i)X_i \Rightarrow \nabla l_Q(\beta_Q) = \frac{1}{n_Q}\sum_{i=1}^{n_Q} V_i,$$

$$\nabla^2 l_Q(\beta) = \frac{1}{n_Q}\sum_{i=1}^{n_Q}\psi''(\beta^TX_i)X_iX_i^T.$$

where $V_i = (\psi'(\beta_Q^TX_i) - Y_i)X_i$. We see that $l_Q(\beta)$ is convex since $\nabla^2 l_Q(\beta)$ is positive definite. Plus, we have

$$||\nabla l_Q(\beta_Q)||_\infty = \max_{j=1,\cdots,d} |\frac{1}{n_Q}\sum_{i=1}^{n_Q} V_{ij}|.$$

Define the error of the first-order Taylor series expansion of l_Q at β_Q as

$$\delta l_Q(v) = l_Q(\beta_Q + v) - l_Q(\beta_Q) - (\nabla l_Q(\beta_Q))^T v,$$

for any $v \in \mathbb{R}^d$.

Define the support of β_Q as S with the cardinality $|S| \leq s$. Denote $\{1,\cdots,d\}/S$ by S^c. Since we choose $\lambda_Q \geq \sqrt{K+1}\sqrt{\frac{\log d}{n_Q}}$, on the event E_Q we have $\lambda_Q \geq 2||\nabla l_Q(\beta_Q)||_\infty$. Proposition 5.3 of Fan et al. (2020) then implies that on the event E_Q, we have

$$||(\hat{\beta}_Q - \beta_Q)_{S^c}||_1 \leq 3||(\hat{\beta}_Q - \beta_Q)_S||_1 \Rightarrow ||\hat{\beta}_Q - \beta_Q|| \leq 4||(\hat{\beta}_Q - \beta_Q)_S||_1.$$

Suppose that n_Q is large enough such that $n_Q \geq 64\kappa_2^2 s\log d$, which is feasible as we assumed $n_Q \gg s\log d$. From Proposition 2 of the full version of Negahban et al. (2009), there exists constants $\kappa_1, \kappa_2, c_1, c_2 > 0$ such that for any $v \in \mathbb{R}^d$ satisfying that $||v||_2 \leq 1, ||v_{S^c}||_1 \leq 3||v_S||_1$, we have

$$\begin{aligned}
\delta l_Q(v) &\geq \kappa_1||v||_2^2 - \kappa_1\kappa_2\sqrt{\frac{\log d}{n_Q}}||v||_1||v||_2 \\
&\geq \kappa_1||v||_2(||v||_2 - \frac{1}{8\sqrt{s}}||v||_1) \\
&\geq \kappa_1||v||_2(||v||_2 - \frac{1}{2\sqrt{s}}||v_S||_1) \\
&\geq \kappa_1||v||_2(||v||_2 - \frac{1}{2}||v_S||_2) \\
&\geq \kappa_1||v||_2(||v||_2 - \frac{1}{2}||v||_2) \\
&= \frac{\kappa_1}{2}||v||_2^2.
\end{aligned} \tag{97}$$

with probability at least $1 - c_1 \exp(-c_2 n_Q)$ w.r.t. the distribution of $\mathcal{D}_Q$. See Appendix D.2 of the full version of Negahban et al. (2009) to get a clearer view of this statement.

Putting (96) and (97) together, we obtain that with probability at least $1 - 2d^{-K} - 2d\exp(-n_Q/4) - c_1\exp(-c_2 n_Q)$ w.r.t. the distribution of $\mathcal{D}_Q$, we have

$$\delta l_Q(v) \geq \frac{\kappa_1}{2}||v||_2^2 \quad \forall ||v||_2 \leq 1, ||v||_1 \leq 3||v_S||_1, \tag{98}$$

and

$$\lambda_Q \geq 2||\nabla l_Q(\beta_Q)||_\infty. \tag{99}$$

Applying Theorem 1 of the full version of Negahban et al. (2009) with (98) and (99), we have that with probability at least $1 - 2d^{-K} - 2d\exp(-n_Q/4) - c_1\exp(-c_2 n_Q)$,

$$||\hat{\beta}_Q - \beta_Q|| \leq \frac{4\sqrt{s}}{\kappa_1}\lambda_Q = \frac{4c_Q}{\kappa_1}\sqrt{\frac{s\log d}{n_Q}}. \tag{100}$$

since $||(\beta_Q)_{S^c}||_1 = 0$. Denote the event $\{||\hat{\beta}_Q - \beta_Q||^2 \leq \frac{4\sqrt{s}}{\kappa_1}\lambda_Q\}$ by E_0, we have

$$\mathbb{P}_{\mathcal{D}_Q}(E_0^c) \leq 2d^{-K} + 2d\exp(-n_Q/4) + c_1\exp(-c_2 n_Q).$$

By the theorem setting, we have $2d^{-K} \lesssim \frac{s\log d}{n_Q} \wedge \frac{s\log d}{n_P}$, so it remains to show that the term $2d\exp(-n_Q/4) + c_1\exp(-c_2 n_Q)$ is asymptotically less than $\frac{s\log d}{n_Q} \wedge \frac{s\log d}{n_P}$. Given the condition $\log n_Q \gg d$, we could set n_Q to be large enough such that

$$2d\exp(-n_Q/4) = \exp(-n_Q/4 + \log d) \leq \exp(-n_Q/8).$$

Hence, it suffices to show that for any $c > 0$, we have $\exp(-cn_Q) \ll \frac{s\log d}{n_Q} \wedge \frac{s\log d}{n_P}$. Note that

$$n_Q \gg \log\frac{n_P}{s\log d} \Rightarrow \exp(cn_Q - \frac{n_P}{s\log d}) \gg 1 \Rightarrow \exp(-cn_Q) \ll \frac{s\log d}{n_P}.$$

Since it is trivial that $\exp(-cn_Q) \ll \frac{s\log d}{n_Q}$, we have $\exp(-cn_Q) \ll \frac{s\log d}{n_Q} \wedge \frac{s\log d}{n_P}$, which finishes the proof. □

Proof of Lemma D.2. For any $i = 1, \cdots, n_P$, define the residual as

$$\varepsilon_i^P := \psi'(\beta_P^T X_i^P) - Y_i^P \in [-1, 1].$$

Define $V_{ij}^P := X_{ij}^P \varepsilon_i^P$, where X_{ij}^P is the $j-$th covariate of the i-th observation X_i^P. Following the completely identical procedures as given in Lemma D.1, we have that

$$\mathbb{P}_{\mathcal{D}_P}(\max_{j=1,\cdots,d}|\frac{1}{n_P}\sum_{i=1}^{n_P} V_{ij}^P| \geq \frac{\sqrt{K+1}}{2}\sqrt{\frac{\log d}{n_P}}) \leq 2d^{-K} + 2d\exp(-n_P/4). \tag{101}$$

Denote

$$E_P := \max_{j=1,\cdots,d}|\frac{1}{n_Q}\sum_{i=1}^{n_P} V_{ij}| \leq \frac{\sqrt{K+1}}{2}\sqrt{\frac{\log d}{n_P}},$$

then we have $\mathbb{P}_{\mathcal{D}_P}(E_P^c) \leq 2d^{-K} + 2d\exp(-n_P/4)$. Similarly, define the empirical loss function as

$$l_P(\beta) := \frac{1}{n_P}\sum_{i=1}^{n_P}\left\{\log(1+e^{(X_i^P)^T\beta}) - Y_i^P(X_i^P)^T\beta\right\}.$$

Similarly, we have

$$||\nabla l_P(\beta_Q)||_\infty = \max_{j=1,\cdots,d}|\frac{1}{n_P}\sum_{i=1}^{n_P}V_{ij}^P|,$$

and

$$\nabla^2 l_P(\beta) = \frac{1}{n_P}\sum_{i=1}^{n_P}\psi''(\beta^T X_i^P)X_i^P(X_i^P)^T.$$

Define

$$h = \frac{||\beta_P||}{||\beta_Q||}\beta_Q,$$

which could be seen as the "rotated" β_P in order to have the sparsity pattern with the same norm. Due to the angle constraint with the parameter Δ between β_Q and β_P, we have

$$||\beta_P - h|| \leq \Delta||\beta_P|| \leq U\Delta.$$

We expand the first-order Taylor series at h, which is different from the classical analysis in Lemma D.1. Define the error of the first-order Taylor series expansion of l_P at h as

$$\delta l_P(v) = l_P(h+v) - l_P(h) - (\nabla l_P(h))^T v,$$

for any $v \in \mathbb{R}^d$. Define the support of β_Q as S. Since we choose $\lambda_P \geq \sqrt{K+1}\sqrt{\frac{\log d}{n_P}}$, on the event E_P we have $\lambda_P \geq 2||\nabla l_P(\beta_P)||_\infty$.

Let $F(v) = l_P(h+v) - l(v) + \lambda_P(||h+v||_1 - ||h||_1)$ and $\hat{v} = \hat{\beta}_P - h$. Then $F(\hat{v}) \leq 0$. It could be observed that on the event

$$E_P \cap \{||\frac{1}{n_P}\sum_{i=1}^{n_P}X_i^P(X_i^P)^T||_2 \leq 2\},$$

we have

$$\begin{aligned}
l_P(h+v) - l(v) &\geq -|\nabla l_P(h)^T v| \\
&\geq -|\nabla l_P(\beta_P)^T v| - |(\nabla l_P(\beta_P) - \nabla l_P(h))^T v| \\
&\geq -||\nabla l_P(\beta_P)||_\infty ||v||_1 - ||\nabla l_P(\beta_P) - \nabla l_P(h)||_2 ||v||_2 \\
&\geq -\frac{\lambda_P}{2}||v||_1 - \max\{||\nabla^2 l_P||_2\} \cdot ||\beta_P - h||_2 \cdot ||v||_2 \\
&\geq -\frac{\lambda_P}{2}||v||_1 - \frac{1}{4}||\frac{1}{n_P}\sum_{i=1}^{n_P}X_i^P(X_i^P)^T||_2 \cdot ||\beta_P - h||_2 \cdot ||v||_2 \\
&\geq -\frac{\lambda_P}{2}||v||_1 - \frac{U\Delta}{2}||v||_2
\end{aligned} \tag{102}$$

The first inequality is due to convexity of l_p. The third inequality is because that

$$u^T v \leq ||u||_\infty ||v||_1, \ u^T v \leq ||u||_2 ||v||_2$$

for any vector pair u, v. The fourth inequality is due to the first-order Taylor expansion for the derivative and the matrix norm inequality

$$||Au||_2 \leq ||A||_2||u||_2$$

for any matrix A and vector u.

On the other hand, as $h_S = h$ by definition of h, we have

$$\begin{aligned} ||h+v||_1 - ||h||_1 &= ||h + v_S + v_{S^c}||_1 - ||h||_1 \\ &\geq ||h + v_{S^c}||_1 - ||v_S||_1 - ||h||_1 \\ &= ||h||_1 + ||v_{S^c}||_1 - ||v_S||_1 - ||h||_1 \\ &= ||v_{S^c}||_1 - ||v_S||_1. \end{aligned} \tag{103}$$

Combining (102), (103), and $F(\hat{v}) \leq 0$, we have

$$U\Delta||\hat{v}||_2 \geq \lambda_P(||\hat{v}_{S^c}||_1 - 3||\hat{v}_S||_1) \Longrightarrow ||\hat{v}||_1 \leq 4||\hat{v}_S||_1 + \frac{U\Delta}{\lambda_P}||\hat{v}||_2. \tag{104}$$

We apply Proposition 2 of the full version of Negahban et al. (2009) again. Suppose that n_P is large enough such that $n_P \geq 64\kappa_2^2 s \log d$, which is feasible as we assumed $n_P \gg s\log d$. From Proposition 2 of the full version of Negahban et al. (2009), there exists constants $\kappa_1, \kappa_2, c_1, c_2 > 0$ such that

$$\begin{aligned} \delta l_P(\hat{v}) &\geq \kappa_1||\hat{v}||_2^2 - \kappa_1\kappa_2\sqrt{\frac{\log d}{n_P}}||\hat{v}||_1||\hat{v}||_2 \\ &\geq \kappa_1||\hat{v}||_2(||\hat{v}||_2 - \kappa_2\sqrt{\frac{\log d}{n_P}}||\hat{v}||_1) \\ &\geq \kappa_1||\hat{v}||_2(||\hat{v}||_2 - 4\kappa_2\sqrt{\frac{\log d}{n_P}}||\hat{v}_S||_1 - \kappa_2\sqrt{\frac{\log d}{n_P}}\frac{U\Delta}{\lambda_P}||\hat{v}||_2) \\ &\geq \kappa_1||\hat{v}||_2(||\hat{v}||_2 - \frac{1}{2\sqrt{s}}||\hat{v}_S||_1 - \frac{\kappa_2 U\Delta}{\sqrt{K+1}}||\hat{v}||_2) \\ &\geq \kappa_1||\hat{v}||_2(||\hat{v}||_2 - \frac{1}{2}||\hat{v}_S||_2 - \frac{1}{4}||\hat{v}||_2) \\ &\geq \kappa_1||\hat{v}||_2(||\hat{v}||_2 - \frac{1}{2}||\hat{v}||_2 - \frac{1}{4}||\hat{v}||_2) \\ &= \frac{\kappa_1}{4}||\hat{v}||_2^2. \end{aligned} \tag{105}$$

with probability at least $1 - c_1\exp(-c_2 n_Q)$ w.r.t. the distribution of $\mathcal{D}_Q$. The fourth inequality is due to $n_P \geq 64\kappa_2^2 s\log d$, and $\lambda_P \geq \sqrt{K+1}\sqrt{\frac{\log d}{n_P}}$. Recall that in the lemma we assumed that Δ is smaller than a constant, and in (105) we specifically assume that $\frac{\kappa_2 U\Delta}{\sqrt{K+1}} \leq \frac{1}{4}$, i.e. $\Delta \leq \frac{\sqrt{K+1}}{4\kappa_2 U}$.

From the proof of Theorem 1 of Bickel and Levina (2008), by applying Lemma A.3 of Bickel and Levina (2008) and the union bound, we have

$$\mathbb{P}_{\mathcal{D}_P}(||\frac{1}{n_P}\sum_{i=1}^{n_P} X_i^P(X_i^P)^T - I_d|| \geq t) \leq 3d\exp(c_3 n t^2)$$

for some constant $c_3 > 0$ and any $t > 0$. Hence,

$$\mathbb{P}_{\mathcal{D}_P}(||\frac{1}{n_P}\sum_{i=1}^{n_P} X_i^P(X_i^P)^T||\ 2) \leq \mathbb{P}_{\mathcal{D}_P}(||\frac{1}{n_P}\sum_{i=1}^{n_P} X_i^P(X_i^P)^T - I_d|| \geq 1) \leq 3d\exp(c_3 n). \tag{106}$$

Putting (101), (105), and (106) together, we obtain that with probability at least $1 - 2d^{-K} - 2d\exp(-n_P/4) - c_1\exp(-c_2 n_P) - 3d\exp(c_3 n)$ w.r.t. the distribution of $\mathcal{D}_P$, we have

$$\delta l_P(\hat{v}) \geq \frac{\kappa_1}{4}||\hat{v}||_2^2 \tag{107}$$

and

$$\lambda_P \geq 2||\nabla l_P(\beta_Q)||_\infty. \tag{108}$$

With a similar analysis as the one in Lemma D.1 based on the theorem setting, we have $2d^{-K} + 2d\exp(-n_P/4) + c_1\exp(-c_2 n_P) + 3d\exp(c_3 n) \leq C_P\left(\frac{s\log d}{n_Q} \wedge \frac{s\log d}{n_P}\right)$. Therefore, it remains to show that $||\hat{\beta}_P - \beta_P||_2 \lesssim \sqrt{s}\lambda_P + \Delta$ when (107) and (108) hold.

Applying (102) and (103) again, we have

$$\begin{aligned}
0 \geq F(\hat{v}) &\geq \nabla l_P(h)^T\hat{v} + \frac{\kappa_1}{4}||\hat{v}||_2^2 + \lambda_P(||h+\hat{v}||_1 - ||h||_1) \\
&\geq \frac{\kappa_1}{4}||\hat{v}||_2^2 - \frac{\lambda_P}{2}||v||_1 - \frac{U\Delta}{2}||v||_2 + \lambda_P(||v_{S^c}||_1 - ||v_S||_1) \\
&= \frac{\kappa_1}{4}||\hat{v}||_2^2 - \frac{\lambda_P}{2}(||v_{S^c}||_1 + ||v_S||_1) - \frac{U\Delta}{2}||v||_2 + \lambda_P(||v_{S^c}||_1 - ||v_S||_1) \\
&\geq \frac{\kappa_1}{4}||\hat{v}||_2^2 - \frac{3\lambda_P}{2}||v_S||_1 - \frac{U\Delta}{2}||v||_2 \\
&\geq \frac{\kappa_1}{4}||\hat{v}||_2^2 - \frac{3\sqrt{s}\lambda_P}{2}||v_S||_2 - \frac{U\Delta}{2}||v||_2 \\
&\geq \frac{\kappa_1}{4}||\hat{v}||_2^2 - \left(\frac{3\sqrt{s}\lambda_P}{2} + \frac{U\Delta}{2}\right)||v||_2.
\end{aligned}$$

given (107) and (108), which implies

$$||\hat{v}||_2 \leq \frac{4}{\kappa_1}(\frac{3\sqrt{s}\lambda_P}{2} + \frac{U\Delta}{2}) \lesssim \sqrt{s}\lambda_P + \Delta.$$

The proof is finished by observing that

$$||\hat{\beta}_P - \beta_P||_2 \leq ||\hat{v}||_2 + ||h - \beta_P||_2 \leq ||\hat{v}||_2 + U\Delta \lesssim \sqrt{s}\lambda_P + \Delta.$$

□

E.6. Proofs of Auxiliary Results

Lemma E.1. If the distribution pair (Q, P) belongs to Π_{BA}^{NP} defined in Section 4 with corresponding parameters, then (Q, P) satisfies Assumption 2 with

$$\varepsilon(z;\gamma, C_\gamma/2) = \left(C_\alpha z^{1+\alpha}\right) \wedge \left(2^{\frac{1+\alpha}{\gamma}} C_\alpha C_\gamma^{-\frac{1+\alpha}{\gamma}} \Delta^{\frac{1+\alpha}{\gamma}}\right).$$

Proof. Suppose $(Q,P)\in\Pi_{BA}^{NP}$. If $x\in\Omega$ satisfies that $|\eta^Q(x)-\frac{1}{2}|\geq(\frac{2\Delta}{C_\gamma})^{\frac{1}{\gamma}}$, then by the definition of Π_{BA}^{NP} we have

$$\Delta\leq\frac{C_\gamma}{2}|\eta^Q(x)-\frac{1}{2}|^\gamma,$$
$$\Longrightarrow s(x)\geq C_\gamma|\eta^Q(x)-\frac{1}{2}|^\gamma-\Delta\geq\frac{C_\gamma}{2}|\eta^Q(x)-\frac{1}{2}|^\gamma.$$

which indicates that $x\in\Omega^+(\gamma,C_\gamma/2)$ and $\Omega^-(\gamma,C_\gamma/2)\subset\{x\in\Omega:|\eta^Q(x)-\frac{1}{2}|<(\frac{2\Delta}{C_\gamma})^{\frac{1}{\gamma}}\}$. Therefore,

$$\begin{aligned}&Q(X\in\Omega^-(\gamma,C_\gamma/2),0<|\eta^Q(X)-\frac{1}{2}|\leq z)\\ \leq&Q(0<|\eta^Q(X)-\frac{1}{2}|\leq(\frac{2\Delta}{C_\gamma})^{\frac{1}{\gamma}}\wedge z)\\ \leq&C_\alpha((\frac{2\Delta}{C_\gamma})^{\frac{\alpha}{\gamma}}\wedge z^\alpha),\end{aligned}$$

Moreover, since

$$\sup_{x\in\Omega^-(\gamma,C_\gamma),|\eta^Q(x)-\frac{1}{2}|\leq z}|\eta^Q(x)-\frac{1}{2}|\leq(\frac{2\Delta}{C_\gamma})^{\frac{1}{\gamma}}\wedge z,$$

we have

$$\begin{aligned}&\mathbb{E}_{(X,Y)\sim Q}\left[|\eta^Q(X)-\frac{1}{2}|\mathbf{1}\{s(X)\leq C_\gamma|\eta^Q(X)-\frac{1}{2}|^\gamma\leq C_\gamma z^\gamma\}\right]\\ \leq&\left(\sup_{x\in\Omega^-(\gamma,C_\gamma),|\eta^Q(x)-\frac{1}{2}|\leq z}|\eta^Q(x)-\frac{1}{2}|\right)\cdot Q(X\in\Omega^-(\gamma,C_\gamma/2),0<|\eta^Q(X)-\frac{1}{2}|\leq z)\\ \leq&C_\alpha\left((\frac{2\Delta}{C_\gamma})^{\frac{\alpha}{\gamma}}\wedge z^\alpha\right)\cdot\left((\frac{2\Delta}{C_\gamma})^{\frac{1}{\gamma}}\wedge z\right)\end{aligned}$$

Therefore, Assumption 2 holds with

$$\begin{aligned}\varepsilon(z;\gamma,C_\gamma/2)&=C_\alpha\left((\frac{2\Delta}{C_\gamma})^{\frac{\alpha}{\gamma}}\wedge z^\alpha\right)\cdot\left((\frac{2\Delta}{C_\gamma})^{\frac{1}{\gamma}}\wedge z\right)\\&=\left(C_\alpha z^{1+\alpha}\right)\wedge\left(2^{\frac{1+\alpha}{\gamma}}C_\alpha C_\gamma^{-\frac{1+\alpha}{\gamma}}\Delta^{\frac{1+\alpha}{\gamma}}\right).\end{aligned}$$

□

Lemma E.2. Suppose there exists a continuous function $p(\cdot;\gamma,C_\gamma):[0,1]\to[0,1]$ such that for any $\eta\in[0,1]$,

$$Q(\text{sgn}\left(\eta-\frac{1}{2}\right)\times(\eta^P-\frac{1}{2})\geq C_\gamma|\eta-\frac{1}{2}|^\gamma|\eta^Q=\eta)\leq p(\eta;\gamma,C_\gamma).$$

Then Assumption 2 holds with

$$\varepsilon(z;\gamma,C_\gamma)=\int_{\frac{1}{2}-z}^{\frac{1}{2}+z}|\eta-\frac{1}{2}|p(\eta;\gamma,C_\gamma)dF_\eta^Q(\eta)\leq C_\alpha z^{1+\alpha}\times\sup_{\eta\in[\frac{1}{2}-z,\frac{1}{2}+z]}p(\eta;\gamma,C_\gamma)$$

where F_η^Q is defined as the cumulative distribution function of η^Q w.r.t. Q_X.

Proof. The first equality holds by observing the transform using Funibi's Theorem:

$$\int_{\Omega^-(\gamma,C_\gamma)} |\eta^Q(X) - \frac{1}{2}|\mathbf{1}\{0 < |\eta^Q(X) - \frac{1}{2}| \le z\} dQ_X$$
$$= \int_{\frac{1}{2}-z}^{\frac{1}{2}+z} |\eta - \frac{1}{2}| Q_{\eta^P}(\text{sgn}\left(\eta - \frac{1}{2}\right) \times (\eta^P - \frac{1}{2}) \ge C_\gamma |\eta - \frac{1}{2}|^\gamma | \eta^Q = \eta) dF_\eta^Q(\eta).$$

where Q_{η^P} is the marginal distribution of η^P with respect to Q. The second equality holds by the provided lemma condition. □

We present a lemma that establishes a high probability uniform bound on the distance between any point and its K-nearest neighbors. This result improves upon Lemma 9.1 in Cai and Wei (2021) by providing a tighter bound that leverages the Hoeffding's inequality, and the rest of our proof is similar. The proof follows a similar approach, so we only provide a more concise presentation here.

Lemma E.3 (K-NN Distance Bound)**.** There exists a constant $c_D > 0$ such that with probability at least $1 - c_D \frac{n_Q}{k_Q} \exp(-2k_Q)$ w.r.t. the distribution of $X_{1:n_Q}$, for all $x \in \Omega \subset [0,1]^d$,

$$||X_{(k_Q)}(x) - x|| \le c_D (\frac{k_Q}{n_Q})^{\frac{1}{d}}. \tag{109}$$

Plus, with probability at least $1 - c_D \frac{n_P}{k_P} \exp(-2k_P)$ w.r.t. the distribution of $X^P_{1:n_P}$, for all $x \in \Omega^P \subset [0,1]^d$,

$$||X^P_{(k_P)}(x) - x|| \le c_D (\frac{k_P}{n_P})^{\frac{1}{d}}. \tag{110}$$

Proof. It suffices to prove (109) since the proof of (110) can be obtained by symmetrically replacing the relevant quantities.

Let $(Q, P) \in \Pi^{NP}$ and take any $x \in \Omega$ and $r < r_\mu$. Since $\frac{dQ_X}{d\lambda} \ge \mu^-$, we have

$$Q(X \in B(x,r)) \ge \mu^- \lambda(B(x,r) \cap \Omega) \ge c_\mu \mu^- \pi_d r^d, \tag{111}$$

where $\pi_d = \lambda(B(0,1))$ is the volume of a d-dimensional unit sphere.

If $k_Q \ge (\frac{1}{4} \wedge \frac{c_\mu \mu^- \pi_d r_\mu^d}{2})$, then $c_D(\frac{k_Q}{n_Q})^{\frac{1}{d}} \ge c_D(\frac{1}{4})^{\frac{1}{d}}$, and we can set c_D to be large enough to satisfy (109) using the trivial bound $||X_{(k_Q)}(x) - x|| \le d^{\frac{1}{2}}$. Therefore, we only need to consider the case when $k_Q \le (\frac{1}{4} \wedge \frac{c_\mu \mu^- \pi_d r_\mu^d}{2})$.

Set $r_0 = (\frac{2k_Q}{c_\mu \mu^- \pi_d n_Q})^{\frac{1}{d}}$. From $k_Q \le \frac{c_\mu \mu^- \pi_d r_\mu^d}{2} n_Q$ we have $r_0 < r_\mu$. Therefore, (111) tells

$$Q(X \in B(x, r_0)) \ge \frac{2k_Q}{n_Q}.$$

Let $S(x) = \sum_{i=1}^{n_Q} \mathbf{1}\{X_i \in B(x, r_0)\}$. Since $X_1, \cdots, X_{n_Q}$ are independent, S follows a binomial distribution with parameters n_Q and $\frac{2k_Q}{n_Q}$. By Hoeffding's inequality,

$$\mathbb{P}_{X_{1:n_Q}}(S(x) < k_Q) = \mathbb{P}_{X_{1:n_Q}}(S - \mathbb{E}_{\mathcal{D}_Q}[S] < -k_Q) \le \exp(-\frac{2k_Q^2}{k^Q}) = \exp(-2k_Q). \tag{112}$$

Suppose that M balls with radius r_0 centered at $x_1, \cdots, x_M$ satisfy that

$$[0,1]^d \subset \bigcup_{m=1}^{M} B(x_m, r_0).$$

It is feasible to find such M balls with $M \leq Cr_0^{-d}$ for some $C > 0$ large enough. By the union bound,

$$\mathbb{P}_{X_{1:n_Q}}(\min_{1\leq m\leq M}\{S(x_m)\} < k_Q) \leq M\exp(-2k_Q) \leq Cr_0^{-d}\exp(-2k_Q). \tag{113}$$

For any $x \in \Omega$, there exists some $1 \leq m' \leq M$ such that $x_{m'} \in B(x, r_0)$. Note that $S(x_{m'}) \geq k_Q$ implies that

$$||X_{(k_Q)}(x) - x|| \leq 2r_0.$$

Therefore, we have

$$\mathbb{P}_{X_{1:n_Q}}(\forall x \in \Omega, ||X_{(k_Q)}(x) - x|| \leq 2r_0) \geq \mathbb{P}_{X_{1:n_Q}}(\min_{1\leq m\leq M}\{S(x_m)\} \geq k_Q) \geq 1 - Cr_0^{-d}\exp(-2k_Q),$$

i.e. with probability at least $1 - \frac{2C}{c_\mu \mu^- \pi_d}\frac{n_Q}{k_Q}\exp(-2k_Q)$ w.r.t. the distribution of $X_{1:n_Q}$, we have

$$||X_{(k_Q)}(x) - x|| \leq 2(\frac{2}{c_\mu \mu^- \pi_d})^{\frac{1}{d}}(\frac{k_Q}{n_Q})^{\frac{1}{d}}.$$

This completes the proof by setting c_D large enough. □

Lemma E.4. For any empirical classifier $\hat{f}$ and $\alpha, \beta \in \mathbb{R}^d$ such that $||\alpha||$ and $||\beta||$ are bounded between c and C for some constant $C > c > 0$, and $\langle\alpha, \beta\rangle \in [0, \pi/2]$. Then we have

$$\mathcal{E}_{Q_\alpha}(\hat{f}) + \mathcal{E}_{Q_\beta}(\hat{f}) \geq \frac{c\sigma(C)(1-\sigma(C))}{20\pi}\langle\alpha, \beta\rangle^2.$$

Proof. Without loss of generality and due to the symmetry property of of $N(0, I_d)$, we could rotate α and β at the same time so that

$$\alpha = (||\alpha||, 0, 0, \cdots, 0), \quad \beta = (||\beta||\cos\langle\alpha, \beta\rangle, ||\beta||\sin\langle\alpha, \beta\rangle, 0, 0, \cdots, 0).$$

Therefore, we assume that only the first coordinate of α and the first and second coordinates of β can be non-zero.

For any $\lambda \in [0, 1]$, we define $w_\lambda := \lambda\alpha + (1-\lambda)\beta$. Define $\lambda_0, \lambda_1 \in [0, 1]$ as the value satisfying that

$$\langle\alpha, w_{\lambda_0}\rangle = \langle w_{\lambda_1}, \beta\rangle = \langle\alpha, \beta\rangle/4.$$

For simplicity, we abbreviate w_{λ_0} as w_0 and w_{λ_1} as w_1. Define the area

$$D := \{x \in \mathbb{R}^d : (w_0^T x)(w_1^T x) < 0, \frac{1}{4} \leq x_1^2 + x_2^2 \leq 1\}.$$

It is easy to see that $D \subset \{x \in \mathbb{R}^d : (\alpha^T x)(\beta^T x) < 0, \frac{1}{2} \leq x_1^2 + x_2^2 \leq 1\}$. Moreover, we see that if $x \in D$, we have

$$|\alpha^T x| = ||\alpha||x_1 \geq ||\alpha||\sin(\frac{\langle\alpha, \beta\rangle}{4})(x^1 + x^2)^{\frac{1}{2}},$$

$$|\beta^T x| \geq ||\beta|| \sin(\frac{\langle \alpha, \beta \rangle}{4})(x^1 + x^2)^{\frac{1}{2}}$$

since the angle between $(x_1, x_2, 0 \cdots, 0)$ and normal planes of α and β are both in $[\frac{\langle \alpha,\beta \rangle}{4}, \frac{3\langle \alpha,\beta \rangle}{4}]$. Therefore, since $\sigma'(t) = \sigma(t)(1 - \sigma(t))$ decreases with growing $|t|$, and $|\alpha^T x| \leq ||\alpha|| \leq C$ on $x \in D$, we have that if $x \in D$.

$$\begin{aligned}
|\eta^{Q_\alpha}(x) - \frac{1}{2}| &\geq |\sigma(\sin(\frac{\langle \alpha, \beta \rangle}{4})||\alpha||(x_1^2 + x_2^2)^{\frac{1}{2}}) - \frac{1}{2}| \\
&\geq \sigma(C)(1 - \sigma(C)) \sin(\frac{\langle \alpha, \beta \rangle}{4})||\alpha||(x_1^2 + x_2^2)^{\frac{1}{2}} \\
&\geq \frac{1}{2} c\sigma(C)(1 - \sigma(C)) \sin(\frac{\langle \alpha, \beta \rangle}{4}).
\end{aligned}$$

Similarly,

$$|\eta^{Q_\beta}(x) - \frac{1}{2}| \geq \frac{1}{2} c\sigma(C)(1 - \sigma(C)) \sin(\frac{\langle \alpha, \beta \rangle}{4}).$$

Hence, we have

$$\begin{aligned}
\mathcal{E}_{Q_\alpha}(\hat{f}) + \mathcal{E}_{Q_\beta}(\hat{f}) &= 2\mathbb{E}_{Q_\alpha}[|\eta^{Q_\alpha}(X) - \frac{1}{2}|\mathbf{1}\{(\hat{f} - \frac{1}{2})(\alpha^T x) < 0\}] \\
&+ 2\mathbb{E}_{Q_\beta}[|\eta^{Q_\beta}(X) - \frac{1}{2}|\mathbf{1}\{(\hat{f} - \frac{1}{2})(\beta^T x) < 0\}] \\
&\geq 2\mathbb{E}_{Q_\alpha}[|\eta^{Q_\alpha}(X) - \frac{1}{2}|\mathbf{1}\{(\hat{f} - \frac{1}{2})(\alpha^T x) < 0, x \in D\}] \\
&+ 2\mathbb{E}_{Q_\beta}[|\eta^{Q_\beta}(X) - \frac{1}{2}|\mathbf{1}\{(\hat{f} - \frac{1}{2})(\beta^T x) < 0, x \in D\}] \\
&\geq c\sigma(C)(1 - \sigma(C)) \sin(\frac{\langle \alpha, \beta \rangle}{4}) \times \\
&\mathbb{E}_{X \sim N(0, I_d)}[\left(\mathbf{1}\{(\hat{f} - \frac{1}{2})(\alpha^T x) < 0\} + \mathbf{1}\{(\hat{f} - \frac{1}{2})(\beta^T x) < 0\}\right) \mathbf{1}\{x \in D\}].
\end{aligned}$$

The change of the expectation operator in the last inequality is due to the fact that $|\eta^{Q_\alpha}(X) - \frac{1}{2}|\mathbf{1}\{(\hat{f} - \frac{1}{2})(\alpha^T x) < 0\}$ and $|\eta^{Q_\beta}(X) - \frac{1}{2}|\mathbf{1}\{(\hat{f} - \frac{1}{2})(\beta^T x) < 0\}$ do not depend on the distribution of the response value.

If $x \in D$, then $(\alpha^T x)(\beta^T x) < 0$, which implies

$$\mathbf{1}\{(\hat{f} - \frac{1}{2})(\alpha^T x) < 0\} + \mathbf{1}\{(\hat{f} - \frac{1}{2})(\beta^T x) < 0\} = 1.$$

Therefore, we could further bound $\mathcal{E}_{Q_\alpha}(\hat{f}) + \mathcal{E}_{Q_\beta}(\hat{f})$ by

$$\begin{aligned}
\mathcal{E}_{Q_\alpha}(\hat{f}) + \mathcal{E}_{Q_\beta}(\hat{f}) &\geq c\sigma(C)(1 - \sigma(C)) \sin(\frac{\langle \alpha, \beta \rangle}{4}) Q_X(D) \\
&\geq c\sigma(C)(1 - \sigma(C)) \sin(\frac{\langle \alpha, \beta \rangle}{4}) \langle \alpha, \beta \rangle Q(X_1^2 + X_2^2 \in [\frac{1}{4}, 1]) \\
&\geq \frac{c\sigma(C)(1 - \sigma(C))}{20\pi} \langle \alpha, \beta \rangle^2.
\end{aligned}$$

where $\Psi(\cdot)$ is the cumulative distribution function of the univariate standard normal distribution. Note that we utilized the numerical results that

$$\sin(\frac{\langle \alpha, \beta \rangle}{4}) \geq \frac{1}{2\pi} \langle \alpha, \beta \rangle,$$

and the cumulative distribution function of chi-squared distribution with 2 degrees of freedom

$$Q(X_1^2 + X_2^2 \in [\frac{1}{4}, 1]) = e^{-\frac{1}{2}} - e^{-\frac{1}{8}} \geq \frac{1}{10}.$$

□