Minimax Optimal Fair Classification with Bounded Demographic Disparity

Xianli Zeng[*] Guang Cheng[†] and Edgar Dobriban[‡]

March 28, 2024

Abstract

Mitigating the disparate impact of statistical machine learning methods is crucial for ensuring fairness. While extensive research aims to reduce disparity, the effect of using a *finite dataset*—as opposed to the entire population—remains unclear. This paper explores the statistical foundations of fair binary classification with two protected groups, focusing on controlling demographic disparity, defined as the difference in acceptance rates between the groups. Although fairness may come at the cost of accuracy even with infinite data, we show that using a finite sample incurs additional costs due to the need to estimate group-specific acceptance thresholds. We study the minimax optimal classification error while constraining demographic disparity to a user-specified threshold. To quantify the impact of fairness constraints, we introduce a novel measure called *fairness-aware excess risk* and derive a minimax lower bound on this measure that all classifiers must satisfy. Furthermore, we propose FairBayes-DDP+, a group-wise thresholding method with an offset that we show attains the minimax lower bound. Our lower bound proofs involve several innovations. Experiments support that FairBayes-DDP+ controls disparity at the user-specified level, while being faster and having a more favorable fairness-accuracy tradeoff than several baselines.

Contents

[*]University of Pennsylvania. zengxl19911214@gmail.com.
[†]University of California, Los Angeles. guangcheng@ucla.edu.
[‡]University of Pennsylvania. dobriban@wharton.upenn.edu.

1 Introduction

Fairness, a concept closely related to justice, has been studied for thousands of years, dating back at least to Plato's Republic (Plato, 1994; Rawls, 1971, 2001). Many laws and provisions aim to ensure fairness and protect the rights and interests of individuals, especially those of vulnerable groups. Recently, the fairness of automated decision-making systems enabled by statistical machine learning has come into question. Due to their ever-improving performance, advanced machine learning methods are increasingly being utilized in high-stakes sectors—ranging from credit lending (Ma et al., 2018) and criminal recidivism forecasting (Angwin et al., 2016) to medical diagnoses (Gupta and Mohammad, 2017)—where their decisions profoundly affect individual lives.

Concurrently, these powerful predictive models risk making discriminative decisions against certain protected groups, such as those defined by race, gender, and other characteristics (e.g., Angwin et al.,

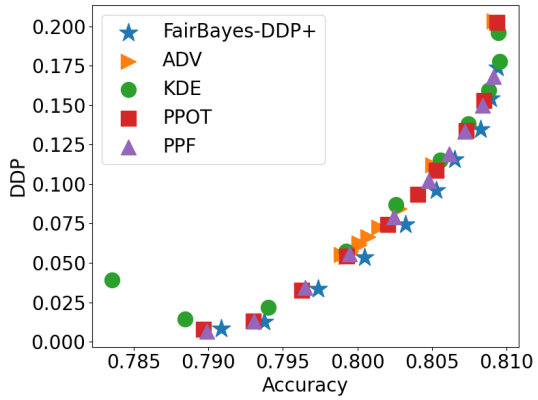


Figure 1: Our FairBayes-DDP+ method achieves a better fairness-accuracy tradeoff than other baselines on the "Adult" dataset. Here DDP is the demographic disparity, i.e., the difference in the probabilities of a positive classification among the two protected groups; see Section 7.2 for details.

Table 1: Performance of our FairBayes-DDP+ method with various pre-specified disparity levels on the "Adult" dataset; see Section 7.2. Our method controls the demographic disparity (DDP) at a user-specified value δ, while achieving a high accuracy (ACC); while having a small standard deviation (SD) over 1000 repetitions.

δ	DDP (SD)	ACC (SD)
0.00	0.008 (0.003)	0.791 (0.001)
0.02	0.013 (0.003)	0.794 (0.001)
0.04	0.033 (0.003)	0.797 (0.001)
0.06	0.054 (0.003)	0.800 (0.001)
0.08	0.074 (0.003)	0.803 (0.000)
0.10	0.096 (0.003)	0.805 (0.000)
0.12	0.115 (0.002)	0.807 (0.000)
0.14	0.135 (0.002)	0.808 (0.000)

2016; Flores et al., 2016; Corbett-Davies et al., 2023). This has motivated a growing body of work literature on the algorithmic aspects of achieving fairness (see Section 2). However, statistical considerations—such as the effect of having a finite dataset on fairness and accuracy in the entire population, and the optimal use of data—are much less studied.

To shed light on the statistical aspects of fairness, we study fair classification, where various population-level fairness criteria exist, see Section 2. Classifiers that conform to such fairness constraints and are most accurate—Bayes-optimal—in the population have been identified (Corbett-Davies et al., 2017; Menon and Williamson, 2018; Chzhen et al., 2019; Schreuder and Chzhen, 2021; Wei et al., 2021; Zeng et al., 2022, 2024). From these works, it is known that there can be fundamental trade-offs between accuracy and fairness even if the entire population is known and available to determine a classifier.

However, it is not known how much additional cost using a *finite dataset* induces. What is the best possible—minimax optimal—accuracy and fairness that we can achieve with a finite sample? For a different problem, fair regression, this has been studied by Chzhen and Schreuder (2022) and Fukuchi and Sakuma (2023); but even the definitions of fairness are unrelated.

To study fairness in classification, we consider the most commonly discussed fairness metric, demographic parity. Since the accuracy of unfair classifiers can be higher than that of fair classifiers, we introduce a novel notion of *fairness-aware excess risk* to measure performance (See Definition 4.1). This metric coincides with the conventional excess risk when the classifier is fair, and appropriately penalizes unfairness otherwise. Further, it is minimized by the most accurate—Bayes-optimal—fair classifier.

Since the properties of the data distribution affect performance, we quantify the behavior of the data near the decision boundary via the margin condition studied in non-parametric classification (Tsybakov, 2004; Audibert and Tsybakov, 2007; Lei et al., 2013). In fair binary classification with a binary protected attribute, we derive a minimax lower bound for the error when the group-wise probabilities of the positive class—or, the *regression functions*—are Hölder-smooth and the group-wise density functions of features satisfy a so-called strong density condition (Audibert and Tsybakov, 2007),

When the disparity constraint is sufficiently stringent, the group-wise acceptance thresholds need to be adjusted to satisfy the fairness constraint. Estimating these thresholds incurs an additional error, and the minimax lower bound is determined by the maximum of this and the error in estimating the regression functions for each protected group. Deriving the additional term in the lower bound requires an innovative argument, by proposing an intricate novel construction of two similar distributions with distinct decision thresholds in Le Cam's two point lower bound method (see Part II of Appendix C.2 for details).

After deriving the lower bound, we complete the minimax analysis by proposing FairBayes-DDP+, a

method that we show is minimax rate optimal. Our method is a group-wise thresholding algorithm, and improves previous estimators of fair Bayes-optimal classifiers (Menon and Williamson, 2018; Zeng et al., 2022, 2024) in two key components: (1) it identifies and adapts to possible jump discontinuities of the disparity as a function of the group-wise threshold (see Section 5.2) and (2) it introduces offsets to handle the case where the decision boundary has a positive probability (see Section 5.3). We prove that FairBayes-DDP+ is minimax optimal and asymptotically controls disparity.

We summarize our contributions as follows.

- **Minimax lower bound in binary classification with a bounded demographic disparity:** We study classification problems with a constraint on demographic disparity. We introduce the notion of fairness-aware excess risk (Definition 4.1) to measure the performance of classifiers given fairness constraints. When the data distribution satisfies appropriate versions of Hölder-smoothness condition and Tsybakov noise condition (Tsybakov, 2004), we derive the a minimax lower bound for fair classification with a bounded demographic disparity. We find that, in addition to a population-level effect, fairness may or may not have a significant effect on accuracy in a finite sample. Our analysis requires a novel construction of distributions in the lower bound, in the case where the group-wise decision thresholds need to be adjusted to satisfy the fairness constraint.

- **Minimax optimal classifier:** We introduce FairBayes-DDP+, an algorithm for binary fair classification, improving on previous methods in two key ways: (1) by adapting to possible jump discontinuities of the disparity as a function of the group-wise threshold (see Section 5.2) and (2) by introducing offsets to handle a decision boundary with a positive measure (see Section 5.3). We further prove that FairBayes-DDP+ attains minimax optimality. In experiments, we compare it with several baselines and show that it has a competitive performance. FairBayes-DDP+ controls disparity at the user-specified level, and attains a better tradeoff between fairness and accuracy in a finite sample than baselines. See Figure 1 and Table 1 for a brief example, and see Section 7 for details. Our numerical results can be reproduced with the code provided at `https://github.com/XianliZeng/FairBayes-DDP-Plus`.

2 Related Literature

There is a great deal of related work, and we can only discuss the most closely related papers.

Definitions of Fairness. Many fairness metrics have been developed. Group fairness (e.g., Calders et al., 2009; Dwork et al., 2012; Hardt et al., 2016, etc) targets parity across protected groups, while individual fairness (e.g., Joseph et al., 2016; Lahoti et al., 2019; Ruoss et al., 2020, etc) aims to provide nondiscriminatory predictions for similar individuals.

Algorithms Aiming for Fairness. There is a large literature on fair machine learning algorithms, broadly categorized into three types: pre-processing (e.g., Feldman et al., 2015; Lum and Johndrow, 2016; Johndrow and Lum, 2019; Calmon et al., 2017, etc), in-processing (e.g., Goh et al., 2016; Zafar et al., 2019; Narasimhan, 2018; Celis et al., 2019; Cotter et al., 2019; Cho et al., 2020, etc), and post-processing (e.g., Fish et al., 2016; Corbett-Davies et al., 2017; Valera et al., 2018; Menon and Williamson, 2018; Chzhen et al., 2019; Alabdulmohsin, 2020; Schreuder and Chzhen, 2021; Jang et al., 2022, etc), see Caton and Haas (2023) for a review.

Our method is a post-processing algorithm, aiming to mitigate disparities in the output of a classifier. Specifically, it is a group-wise thresholding rule (e.g., Fish et al., 2016; Corbett-Davies et al., 2017; Valera et al., 2018; Menon and Williamson, 2018; Chzhen et al., 2019; Alabdulmohsin, 2020; Schreuder and Chzhen, 2021; Jang et al., 2022, etc), estimating the probability of a positive label given the features for each protected group, and assigning thresholds to protected groups aiming for parity. Menon and Williamson (2018); Zeng et al. (2024) propose post-processing algorithms aiming to estimate the Bayes-optimal classifier, but do not study the finite-sample performance of their methods. We refine their method with an offset and show that it achieves the minimax optimal rate.

Nonparametric Classification and Minimax Optimal Rate. For a binary classification problem where the goal is to predict a label $Y \in \{0, 1\}$ based on observed d-dimensional features $x \in \mathcal{X} := \mathbb{R}^d$, a

probabilistic classifier f is a function[1] $\mathcal{X} \to [0,1]$ that specifies the probability of predicting $\widehat{Y}_f = 1$ given $X = x$, i.e., $f(x) = \mathbb{P}(\widehat{Y}_f = 1 \mid X = x)$ for all $x \in \mathcal{X}$. Classification methods include plug-in rules, which estimate the regression function $\eta : x \mapsto \mathbb{P}(Y = 1 \mid X = x)$ and makes decisions by thresholding it, and empirical risk minimizers (ERM). The convergence rates and minimax optimality of both methods have been studied (e.g., Mammen and Tsybakov, 1999; Yang, 1999, etc).

When $x \mapsto \mathbb{P}(Y = 1 \mid X = x)$ is β-Hölder-smooth, n is the sample size, and d is the dimensionality, Yang (1999) showed that the convergence rate of a plug-in classifier is $n^{-\beta/(2\beta+d)}$, the same as the convergence rate of the estimated regression function. Moreover, that work proved that the rate is minimax optimal. When the regression function is well-behaved near the decision boundary, the convergence rate is faster. By considering boundary fragments with β-smooth boundaries and noise satisfying the γ-exponent condition, Mammen and Tsybakov (1999) and Tsybakov (2004) proved that the minimax convergence rate is $n^{-\beta(\gamma+1)/[\beta(\gamma+2)+(d-1)\gamma]}$, which can be achieved by ERM rules. Audibert and Tsybakov (2007) showed that a plug-in rule with a local polynomial regression estimate is minimax optimal under the γ-exponent condition and for β-smooth regression functions, with a rate $n^{-\beta(\gamma+1)/(2\beta+d)}$.

3 Classification with a Bounded Demographic Parity

In fair binary classification problems with labels in $\mathcal{Y} = \{0,1\}$, two types of features are observed: the usual features $X \in \mathcal{X}$, and the binary protected (or, sensitive) features $A \in \mathcal{A} = \{0,1\}$[2], with respect to which we aim to be fair. For example, in a credit lending setting, X could refer to education level and income, A could indicate the race or gender of the individual, and Y could correspond to the status of repayment or defaulting on a loan. Here and below, for all $a \in \mathcal{A}$ and $y \in \mathcal{Y} = \{0,1\}$, we denote by $\mathbb{P}_X$, $\mathbb{P}_{X|A=a}$ and $\mathbb{P}_{X|A=a,Y=y}$ the marginal distribution function of X, the conditional distribution function of X given $A = a$, and the conditional distribution of X given $A = a, Y = y$, respectively.

To evaluate the fairness of a classifier, we consider demographic parity, the possibly most popular fairness metric.[3] A probabilistic classifier $f : \mathcal{X} \times \mathcal{A} \to [0,1]$ specifies the probability of predicting $\widehat{Y}_f = 1$ given $X = x$ and $A = a$, i.e., $f(x) = \mathbb{P}(\widehat{Y}_f = 1 \mid X = x, A = a)$ for $(x,a) \in \mathcal{X} \times \mathcal{A}$. The classifier f satisfies *demographic parity* if its prediction $\widehat{Y}_f$ is probabilistically independent of the protected attribute A: $\widehat{Y}_f \perp\!\!\!\perp A$, so that $\mathbb{P}_{X|A=1}\left(\widehat{Y}_f = 1\right) = \mathbb{P}_{X|A=0}\left(\widehat{Y}_f = 1\right)$. However, demographic parity may be too stringent in certain cases, and it is desired to have more flexible metrics controlling disparate impact. To measure the disparate impact of a classifier, we use the *demographic disparity* or DDP (Cho et al., 2020), i.e., the difference in the probabilities of predicting $\widehat{Y}_f = 1$ across groups:

$$\mathrm{DDP}(f) = \mathbb{P}_{X|A=1}(\widehat{Y}_f = 1) - \mathbb{P}_{X|A=0}\left(\widehat{Y}_f = 1\right). \tag{3.1}$$

We denote by $\mathcal{F}_\delta$ the set of functions satisfying the *δ-parity constraint* $|\mathrm{DDP}(f)| \leqslant \delta$, so that

$$\mathcal{F}_\delta = \{f : \mathcal{X} \times \mathcal{A} \to [0,1] \ : \ |\mathrm{DDP}(f)| \leqslant \delta\}.$$

Subject to this δ-parity constraint, we aim to minimize the misclassification error. This is achieved by *δ-fair Bayes-optimal classifiers*, defined as

$$f_\delta^\star \in \operatorname*{arg\,min}_{f \in \mathcal{F}_\delta} \mathbb{P}\left(Y \neq \widehat{Y}_f\right). \tag{3.2}$$

3.1 Fair Bayes-Optimal Classifier under Demographic Parity

The classification thresholds of fair Bayes-optimal classifiers need to be adjusted for each group, see Corbett-Davies et al. (2017); Menon and Williamson (2018); Chzhen et al. (2019); Schreuder and Chzhen (2021);

[1] All functions considered will be measurable with respect to the Borel sigma algebras on the input and output spaces; this will not be mentioned further.

[2] Conventionally, we consider $A = 0$ to represent the underprivileged group that could potentially face discrimination.

[3] In future work, we expect that our insights can seamlessly be extrapolated to other group fairness metrics, including equality of opportunity (Hardt et al., 2016) and predictive equality (Corbett-Davies et al., 2017).

Wei et al. (2021); Zeng et al. (2022, 2024) and Proposition A.1 for details. To leverage these results, we need some additional notation.

Intuitively, to minimize the error, we should output $\widehat{Y} = 1$ if the probability of $Y = 1$ given $X = x$ and $A = a$ is large. Therefore, the *group-conditional probabilities*—or, regression functions—η_a, $a \in \mathcal{A}$ defined for all $x \in \mathcal{X}$ via $\eta_a(x) := \mathbb{P}(Y = 1 \mid A = a, X = x)$, play a crucial role. All optimal classifiers f will aim to output $\widehat{Y}_f = 1$ if $\eta_a(x)$ is large.

To explain this in detail, for $a \in \mathcal{A}$, we denote $p_a := \mathbb{P}(A = a)$, and let $\mathcal{T} = [-\min(p_1, p_0), \min(p_1, p_0)]$. For $a \in \{0, 1\}$ and $t \in \mathcal{T}$, define *group-wise thresholds* via the formula $T_a(t) = 1/2 + (2a - 1)t/(2p_a)$. Let $\mathcal{F}^t$ be the class of *group-wise thresholding rules* with thresholds $T_a(t)$ for each $a \in \mathcal{A}$ as follows; where $I(\cdot)$ is the indicator function that equals unity if its argument is true, and zero otherwise:

$$\mathcal{F}^t := \left\{ f : \mathcal{X} \times \mathcal{A} \to [0, 1],\ f(x, a) = I\left(\eta_a(x) > T_a(t)\right) + \tau_a I\left(\eta_a(x) = T_a(t)\right),\ (\tau_1, \tau_0) \in [0, 1]^2 \right\}.$$

These classifiers output $\widehat{Y}_f = 1$ if $\eta_a(x)$ is large, and the thresholds $T_a(t)$ are allowed to depend on the group $a \in \mathcal{A}$. It turns out that parametrizing the thresholds of acceptance via $t \mapsto T_a(t)$ for $t \in \mathrm{T}$ suffices to obtain Bayes-optimal classifiers.

Further, define the *disparity functions* $D_- : \mathcal{T} \to \mathbb{R}$ and $D_+ : \mathcal{T} \to \mathbb{R}$ such that for all $t \in \mathcal{T}$,

$$D_-(t) = \mathbb{P}_{X|A=1}\left(\eta_1(X) > \frac{1}{2} + \frac{t}{2p_1}\right) - \mathbb{P}_{X|A=0}\left(\eta_0(X) \geqslant \frac{1}{2} - \frac{t}{2p_0}\right); \tag{3.3}$$

$$D_+(t) = \mathbb{P}_{X|A=1}\left(\eta_1(X) \geqslant \frac{1}{2} + \frac{t}{2p_1}\right) - \mathbb{P}_{X|A=0}\left(\eta_0(X) > \frac{1}{2} - \frac{t}{2p_0}\right). \tag{3.4}$$

By inspection, both functions are non-increasing, and for any $t \in \mathbb{R}$, $D_-(t) \leqslant D_+(t)$. Moreover D_- is right-continuous and D_+ is left-continuous. It is not hard to see, and it is shown in Zeng et al. (2024), that for all $t \in \mathcal{T}$, the DDP of group-wise thresholding rules ranges between $D_-(t)$ and $D_-(t)$; specifically

$$D_-(t) = \inf_{f \in \mathcal{F}^t} \mathrm{DDP}(f) \ \text{ and } \ D_+(t) = \sup_{f \in \mathcal{F}^t} \mathrm{DDP}(f).$$

In particular, $D_-(0)$ and $D_+(0)$ are, respectively, the infimum and supremum of the DDP over all unconstrained Bayes-optimal classifiers from (3.2) with $\delta = \infty$.

We will focus on the setting where the group-wise thresholds are uniquely defined, which holds if there is enough probability mass near the decision boundaries (and is ensured by our formal conditions to follow). In this case, a δ-fair Bayes-optimal classifier $f^\star_\delta$ has the following form. Define the following "inverse" of the functions D_-, D_+ on $\mathbb{R}_{\geqslant 0}$, for $\delta \geqslant 0$,

$$t^\star_\delta = \inf_{t \in \mathcal{T}} \{D_-(t) < \delta\} = \inf_{t \in \mathcal{T}} \{D_+(t) < \delta\}. \tag{3.5}$$

Since D_- is non-increasing and right-continuous, we have $D_-(t^\star_\delta) = \delta$ if D_- is continuous at $t^\star_\delta$, and $D_-(t^\star_\delta) < \delta$ while $\lim_{t \to (t^\star_\delta)^-} D_-(t) > \delta$ if D_- has a jump discontinuity at $t^\star_\delta$, A similar statement applies to D_+.

Define the *group-wise thresholds of the two groups*

$$T^\star_{\delta,1} = \frac{1}{2} + \frac{t^\star_\delta}{2p_1} \text{ and } T^\star_{\delta,0} = \frac{1}{2} - \frac{t^\star_\delta}{2p_0}. \tag{3.6}$$

Then, for some $\tau^\star_{\delta,1}, \tau^\star_{\delta,0} \in [0, 1]$—specified later in (5.11)—there is a δ-fair Bayes-optimal classifier $f^\star_\delta$ that is a group-wise thresholding rule of the form, for all x, a,

$$f^\star_\delta(x, a) = I\left(\eta_a(x) > T^\star_{\delta,a}\right) + \tau^\star_{\delta,a} I\left(\eta_a(x) = T^\star_{\delta,a}\right). \tag{3.7}$$

As discussed in Appendix A, the behavior of the fair Bayes-optimal classifiers on the decision boundary $\{x, a : \eta_a(x) = T^\star_{\delta,a}\}$ is generally not unique. For the sake of generality, for instance to deal with discrete-valued data, we will allow the decision boundary to have a positive probability mass. However, it will help to have a specific choice of the Bayes-optimal classifier to estimate. Our minimax lower bounds and rate of convergence will not depend on the specific choice of the Bayes-optimal classifier.

Moreover, we will assume without loss of generality that $D_+(0) \geqslant -D_-(0)$. This condition means that among Bayes-optimal classifiers, $\mathbb{P}_{X|A=1}(\widehat{Y}_f = 1) - \mathbb{P}_{X|A=0}(\widehat{Y}_f = 1)$ can be larger than its negative. In this sense, the group $A = 0$ is underprivileged. If this condition does not hold, we can introduce a new variable $\tilde{A}$ defined as $1 - A$, which swaps the groups characterized by $A = 1$ and $A = 0$. We will construct estimators of $D_+(0)$ and $D_-(0)$, and these may be used to decide which group is underprivileged.

Remark 1 (The impact of fairness on Bayes-optimal classifiers)**.** *Since by definition $D_-(0) \leqslant D_+(0)$, there are three possibilities for $\delta > 0$: (1) $\delta < D_-(0)$, (2) $D_-(0) \leqslant \delta < D_+(0)$, and (3) $\delta \geqslant D_+(0)$.*

1. **Fairness-impacted case:** $\delta < D_-(0)$**.** *When δ is relatively small with $\delta < D_-(0)$, no unconstrained Bayes-optimal classifier satisfies the fairness constraint $|\mathrm{DDP}(f)| \leqslant \delta$. As a result, we need to estimate the group-wise thresholds, and the fairness constraint has a significant impact on the fair Bayes-optimal classifiers. Thus, we call this case the fairness-impacted case.*

2. **Fair-boundary case:** $D_-(0) \leqslant \delta < D_+(0)$**.** *When δ is moderately large with $D_-(0) \leqslant \delta < D_+(0)$, there is at least one unconstrained classifier $f^\star \in \arg\min_{f\in\mathcal{F}} \mathbb{P}(Y \neq \widehat{Y}_f)$ satisfying $|\mathrm{DDP}(f^\star)| \leqslant \delta$, and the optimal group-wise thresholds are $1/2$ for both protected groups. In this case, we need to change the classifiers on the decision boundaries $\{x : \eta_a(x) = 1/2\}$ to satisfy fairness constraint. We thus refer to this case as the fair-boundary case. Since changes on the decision boundary do not change the accuracy, the misclassification rate of fair Bayes-optimal classifiers equals the unconstrained Bayes error.*

3. **Automatically fair case:** $\delta \geqslant D_+(0)$**.** *Finally, we call $\delta \geqslant D_+(0)$ the automatically fair case, as all unconstrained Bayes-optimal classifiers are δ-fair.*

As we can see, $t^\star_\delta$ is non-zero only in the fairness-impacted case when $\delta < D_-(0)$.

4 Minimax Lower Bound for Fair Classification

The minimax approach from statistical decision theory characterizes fundamental performance limits. An estimator is minimax rate-optimal if its convergence rate matches the minimax lower bound, i.e., the best possible rate of convergence over all estimators. In this section, we derive a minimax lower bound for the fair classification problem. This requires quantifying the performance of classifiers. We begin by introducing a proper metric for fair classification problems.

4.1 Measure of Performance

Consider first an unconstrained classification problem with a Bayes-optimal classifier $f^\star := f^\star_\infty$ defined in (3.2) with $\delta = \infty$. The performance of a classifier f is commonly measured by its excess risk over $f^\star$ (e.g., Hastie et al., 2009), defined as

$$d_R(f, f^\star) := \mathbb{P}(Y \neq \widehat{Y}_f) - \mathbb{P}(Y \neq \widehat{Y}_{f^\star}) = \sum_{a\in\mathcal{A}} p_a \int (f(x,a) - f^\star(x,a))\,(1 - 2\eta_a(x))\, d\mathbb{P}_{X|A=a}(x). \tag{4.1}$$

For fair classification problems, a first attempt may be to consider the excess risk of f over a fair Bayes-optimal classifier $f^\star_\delta$ from (3.7), i.e., $d_R(f, f^\star_\delta)$. However, in the fairness-impacted case, $d_R(f, f^\star_\delta)$ can be *negative*, as the fair Bayes-optimal classifier does not generally minimize the unconstrained risk, i.e., $f^\star_\delta \notin \arg\min_{f\in\mathcal{F}} \mathbb{P}(Y \neq \widehat{Y}_f)$. As a result, $d_R(f, f^\star_\delta)$ is not directly suitable for measuring the cost of fairness.

As an alternative, we define the following *fairness-aware excess risk* to quantify the performance of a classifier within the context of fair classification. Its functional form is analogous to (4.1), and we provide further justification below.

Definition 4.1 (Fairness-aware excess risk)**.** *Let $\delta \geqslant 0$ and $f^\star_\delta$ be a δ-fair Bayes-optimal classifier from (3.2), and consider $T^\star_{\delta,a}$ from (3.6). For any classifier $f : \mathcal{X} \times \{0,1\} \to [0,1]$, we define the fairness-aware excess risk as*

$$d_E(f, f^\star_\delta) = 2\sum_{a\in\mathcal{A}} p_a \left[\int (f(x,a) - f^\star_\delta(x,a))\left(T^\star_{\delta,a} - \eta_a(x)\right) d\mathbb{P}_{X|A=a}(x)\right]. \tag{4.2}$$

Observe first that $d_E(f, f^\star_\delta) \geqslant 0$, as from (3.7), it follows that $f(x,a) - f^\star_\delta(x,a) \leqslant 0$ when $\eta_a(x) > T^\star_{\delta,a}$ and $f(x,a) - f^\star_\delta(x,a) \geqslant 0$ when $\eta_a(x) \leqslant T^\star_{\delta,a}$. Moreover, it follows from Proposition A.1 that the choice of the δ-fair Bayes-optimal classifier $f^\star_\delta$ does not affect the value of d_E. The following result further elucidates the fairness-aware excess risk d_E, connecting it to the classical excess risk d_R.

Proposition 4.2 (Characterizing fairness-aware excess risk)**.** *For any classifier $f \in \mathcal{F}$, the fairness-aware excess risk simplifies as follows, in the cases identified in Remark 1:*

$$d_E(f, f^\star_\delta) = \begin{cases} d_R(f, f^\star_\delta), & \text{in the automatically fair and fair-boundary cases } \delta \geqslant D_-(0); \\ d_R(f, f^\star_\delta) + t^\star_\delta \left[\mathrm{DDP}(f) - \delta\right], & \text{in the fairness-impacted case } \delta < D_-(0); \end{cases}$$

Moreover, for δ-fair classifiers f with $|\mathrm{DDP}(f)| \leqslant \delta$, we have $d_R(f, f^\star_\delta) \geqslant d_E(f, f^\star_\delta)$.

We will show a *lower bound* on the minimax excess fairness-aware risk d_E over *all classifiers*, and an *upper bound* realized by an *asymptotically δ-fair classifier*. This will ensure that our method is asymptotically optimal with respect to both d_E and d_R among all δ-fair classifiers.

4.2 Conditions on the Data Distribution

In this section, we introduce conditions on the data distribution that we need in our theoretical analysis, which require some notations and definitions. For a scalar β, we denote by $\lfloor \beta \rfloor_+ := \lceil \beta \rceil - 1$ the maximal integer that is strictly less than β. For an integer $d > 0$, and a multi-index $s \in \mathbb{N}^d$, we denote $|s| = s_1 + \ldots + s_d$. Moreover, for $x \in \mathbb{R}^d$ and $s \in \mathbb{N}^d$, we denote $x^s = x_1^{s_1} \cdot \ldots \cdot x_d^{s_d}$. The first concept is the smoothness of the per-group regression functions η_a, $a \in \mathcal{A}$.

Definition 4.3 (Hölder Smoothness)**.** *Consider $\beta > 0$ and any $\lfloor \beta \rfloor_+$-times continuously differentiable real-valued function g on $\mathbb{R}^d$. For any $x \in \mathbb{R}^d$, we denote by g_x the Taylor approximation of degree $\lfloor \beta \rfloor_+$ of g at x, such that for all $x' \in \mathbb{R}^d$,*

$$g_x(x') = \sum_{s \in \mathbb{N}^d : |s| \leqslant \lfloor \beta \rfloor_+} \frac{(x' - x)^s}{s!} g^{(s)}(x).$$

For $L_\beta > 0$, the $(\beta, L_\beta, \mathbb{R}^d)$-Hölder class of functions, denoted $\Sigma(\beta, L, \mathbb{R}^d)$, is defined as the set of functions $g : \mathbb{R}^d \to \mathbb{R}$ that are $\lfloor \beta \rfloor_+$ times continuously differentiable and satisfy, for any $x, x' \in \mathbb{R}^d$, the inequality $|g(x') - g_x(x')| \leqslant L_\beta \|x' - x\|^\beta$.

The next definition is the *margin condition*, which we adapt to the fair classification problem from Tsybakov (2004); Audibert and Tsybakov (2007), Lei et al. (2013), and which controls the regularity of the regression function near the decision boundary. Let $\gamma \geqslant 0$, and let P be a distribution defined on $\mathcal{X} \times \mathcal{A} \times \mathcal{Y}$ with conditional probability functions η_a, $a \in \mathcal{A}$. For D_-, D_+ from (3.3) and (3.4) and $t^\star_\delta$ from (3.5), define the boundary probability functions $g_{\delta,-}, g_{\delta,+} : [0, \infty) \to [0, 2]$ for the positive ("+") and negative ("-") sides of $t^\star_\delta$, such that for all $\varepsilon \geqslant 0$,

$$\begin{aligned} g_{\delta,-}(\varepsilon) &= D_-(t^\star_\delta) - D_-(t^\star_\delta + \varepsilon) \\ &= \mathbb{P}_{X|A=1}\left(T^\star_{\delta,1} < \eta_1(X) \leqslant T^\star_{\delta,1} + \frac{\varepsilon}{2p_1}\right) + \mathbb{P}_{X|A=0}\left(T^\star_{\delta,0} - \frac{\varepsilon}{2p_0} \leqslant \eta_0(X) < T^\star_{\delta,0}\right); \\ g_{\delta,+}(\varepsilon) &= D_+(t^\star_\delta - \varepsilon) - D_+(t^\star_\delta) \\ &= \mathbb{P}_{X|A=1}\left(T^\star_{\delta,1} - \frac{\varepsilon}{2p_1} \leqslant \eta_1(X) < T^\star_{\delta,1}\right) + \mathbb{P}_{X|A=0}\left(T^\star_{\delta,0} < \eta_0(X) \leqslant T^\star_{\delta,0} + \frac{\varepsilon}{2p_0}\right). \end{aligned}$$

Clearly, both $g_{\delta,-}, g_{\delta,+}$ are monotone non-decreasing, while $g_{\delta,+}$ is right-continuous and $g_{\delta,-}$ is left-continuous. One can verify that all notions introduced so far can be defined not just when $\delta \in [0, \infty)$, but also when $\delta = \infty$, which corresponds to the unconstrained case.

Definition 4.4 (γ-Margin Condition, Adapted from Tsybakov (2004), Audibert and Tsybakov (2007), Lei et al. (2013))**.** *Let $\gamma \geqslant 0$, and let P be a distribution defined on $\mathcal{X} \times \mathcal{A} \times \mathcal{Y}$ with group-conditional probabilities*

$\eta_a(x,a) = \mathbb{P}(Y = 1 \mid A = a, X = x)$, *for* $x \in \mathcal{X}$ *and* $a \in \mathcal{A}$. *For* $\delta \geqslant 0$, *we say that* (η_1, η_0) *satisfies the* strong γ-margin condition *for* $\delta \in [0, \infty]$ *with respect to* P *if, first, there exist constants* $\varepsilon_0, U_\gamma > 0$ *such that,*

$$\max\{g_{\delta,-}(\varepsilon), g_{\delta,+}(\varepsilon)\} \leqslant U_\gamma \varepsilon^\gamma, \text{ for all } 0 < \varepsilon < \varepsilon_0; \tag{4.3}$$

and second, for $j \in \{+,-\}$, *if* $D_j(t^\star_\delta) = \delta$,[4] *then*

$$g_{\delta,j}(\varepsilon) \geqslant U_\gamma^{-1} \varepsilon^{1/\gamma}, \text{ for all } 0 < \varepsilon < \varepsilon_0. \tag{4.4}$$

Conditions (4.3) and (4.4) provide upper and lower bounds, respectively, on the probability mass of the regression functions near the decision boundaries. Condition (4.3) adapts the γ-exponent condition introduced by Tsybakov (2004), Audibert and Tsybakov (2007) for characterizing the convergence rate in nonparametric classification to our problem. When γ is large, the probability mass of the conditional probability function near the decision boundary decays quickly with $\varepsilon \to 0$, suggesting that the estimating the conditional probability functions η_a near the decision boundary is less challenging.

For conventional classification problems, an upper such as (4.3) bound is sufficient to characterize problem difficulty. However, for fair classification, a lower bound on the density is also necessary to characterize the estimation error of $t^\star_\delta$ when $D_-(t^\star_\delta) = \delta$ or $D_+(t^\star_\delta) = \delta$. In the fairness-impacted case from Remark 1, the difficulty of estimating $t^\star_\delta > 0$ is impacted by the behavior of D_- and D_+ near $t^\star_\delta$. This is quantified by $g_{\delta,-}(\varepsilon) = D_-(t^\star_\delta) - D_-(t^\star_\delta + \varepsilon)$ and $g_{\delta,+}(\varepsilon) = D_+(t^\star_\delta - \varepsilon) - D_+(t^\star_\delta)$ for $\varepsilon > 0$. Without Condition (4.4), D_- and D_+ could potentially be "flat", making it hard to estimate $t^\star_\delta$, as illustrated in case (3) of Figure 2. In addition, if either $D_-(t^\star_\delta) < \delta$ or $\delta < D_+(t^\star_\delta)$, we do not need a lower bound for that side of the distribution around $t^\star_\delta$, as the gap between δ and $D_j(t^\star_\delta), j \in \{+,-\}$ ensures that $t^\star_\delta$ can be estimated accurately; see case (1) of Figure 2.

Moreover, we also need to ensure that the mass of X is sufficiently "spread out", as per the following strong density condition introduced by Audibert and Tsybakov (2007). For $x \in \mathbb{R}^d$ and $r \geqslant 0$, denote by $B_{d,2}(x,r)$ the closed d-dimensional Euclidean ball centered at x with radius r.

Definition 4.5 (Strong Density Condition (Audibert and Tsybakov, 2007))**.** *Fix a compact set* $\widetilde{C} \subset \mathbb{R}^d$. *We say that a distribution of* (X, A) *with the pair of conditional distributions* $(\mathbb{P}_{X|A=1}, \mathbb{P}_{X|A=0})$ *satisfies* the strong density condition *if there exist positive constants* c_μ, r_μ, $\mu_{\min}$ *and* $\mu_{\max}$ *such that the following hold. For* $a \in \{0,1\}$, $\mathbb{P}_{X|A=a}$ *is absolutely continuous with respect to the Lebesgue measure* λ *on* $\mathbb{R}^d$, *and it is supported on a compact* (c_μ, r_μ)-regular *set* $\Omega_a \subset \widetilde{C}$, *namely*

$$\lambda\left[B_{d,2}(x,r) \cap \Omega_a\right] \geqslant c_\mu \lambda\left[B_{d,2}(x,r) \cap \widetilde{C}\right], \quad \text{for all } x \in \Omega_a \text{ and } 0 < r \leqslant r_\mu.$$

Moreover, for $a \in \{0,1\}$, *the density function* μ_a *of* $\mathbb{P}_{X|A=a}$ *with respect to the Lebesgue measure satisfies* $\mu_{\min} \leqslant \mu_a(x) \leqslant \mu_{\max}$ *for* $x \in \Omega_a$ *and* $\mu_a(x) = 0$ *otherwise.*

Letting $\delta > 0$ be the chosen disparity level, with the above definitions, our parameter space $\mathcal{P}_\Sigma(\delta, \beta, L_\beta, \gamma)$—or, $\mathcal{P}_\Sigma$ for short—is defined as:

Definition 4.6 (Parameter space)**.** *For* $\delta \geqslant 0$, $\beta, L_\beta > 0$ *and* $\gamma_\delta, \gamma_\infty \geqslant 0$, *we denote by* $\mathcal{P}_\Sigma$ *the class of all probability distributions* P *on* $\mathcal{X} \times \mathcal{A} \times \mathcal{Y}$ *satisfying the following.*

1. *The group-conditional probability functions* $\eta_a : x \mapsto \mathbb{P}(Y = 1 \mid A = a, X = x)$, *for* $x \in \mathcal{X}$ *and* $a \in \mathcal{A}$, *satisfy* $\eta_1, \eta_0 \in \Sigma(\beta, L_\beta, \mathbb{R}^d)$, *for the Hölder parameter space* Σ *from Definition 4.3.*
2. *Further,* (η_1, η_0) *satisfies the margin condition from Definition 4.4 at* δ *for* γ_δ *and at* ∞ *for* γ_∞.
3. *The pair of distributions* $(\mathbb{P}_{X|A=1}, \mathbb{P}_{X|A=0})$ *satisfies the strong density condition from Definition 4.5.*

Without loss of generality, we can assume that η_1 and η_0 share the same smoothness parameter β and satisfy the margin condition with the same parameters γ_δ, γ_∞. When $\eta_1 \in \Sigma(\beta_1, L_{\beta,1}, \mathbb{R}^d)$ satisfies the $\gamma_{\delta,1}$-margin condition and $\eta_0 \in \Sigma(\beta_0, L_{\beta,0}, \mathbb{R}^d)$ satisfies the $\gamma_{\delta,0}$-margin condition, we can set $\beta = \beta_1 \wedge \beta_0$ and $\gamma_\delta = \gamma_{\delta,1} \wedge \gamma_{\delta,0}$.

We will further assume that $\gamma_\delta \beta \leqslant d$, which in particular holds if γ_δ and β are constants; this condition is commonly used when deriving minimax lower bounds in nonparametric classification (e.g., Audibert and Tsybakov, 2007; Cai and Wei, 2021, etc.).

[4] If $D_-(t^\star_\delta) < \delta < D_+(t^\star_\delta)$, the lower bound is unnecessary.

4.3 Minimax Lower Bound

We now present our first major result: a minimax lower bound for fair classification. The estimation errors of both the regression functions and the group-wise thresholds contribute to the overall error in estimating a δ-fair Bayes-optimal classifier. Based on the discussion in Remark 1 and after Definition 4.4, there are two cases: (1) The *non-trivial fairness-impacted regime*, where $t_\delta^\star > 0$ (fairness-impacted regime from Remark 1) and $D_-(t_\delta^\star) = D_+(t_\delta^\star) = \delta$ (D_- and D_+ are continuous at $t_\delta^\star$), so that the estimation of $t_\delta^\star$ may affect the minimax lower bound; (2) The *classical regime*, which is the complement of case (1). Here, $t_\delta^\star$ does not need to be estimated, or can be estimated at a fast rate.

Theorem 4.7 (Minimax lower bound for fair classification)**.** *For a fixed $\delta \geqslant 0$, let $\beta, \gamma_\delta > 0$ be such that $\gamma_\delta \beta \leqslant d$, and consider the class of distributions $\mathcal{P}_\Sigma = \mathcal{P}_\Sigma(\delta, \beta, L_\beta, \gamma_\delta, \gamma_\infty, c_\mu, r_\mu, \mu_{\min}, \mu_{\max})$ from Definition 4.6. Let $\delta_{\text{sign}} = \text{DDP}(f_\delta^\star)$.[5] Then, there is $C > 0$ depending only on the problem hyperparameters, such that for any $n \geqslant 1$ and any classifier $\widehat{f}_{\delta,n}$ estimating the δ-fair Bayes-optimal classifier $f_\delta^\star$ from (3.7), constructed from a dataset $\mathcal{S}_n = \{(X_i, A_i, Y_i)\}_{i=1}^n$ sampled i.i.d. from some P from $\mathcal{P}_\Sigma$, we have the following.*

(1). **Non-trivial fairness-impacted regime.** *In the fairness-impacted regime from Remark 1, if in addition $D_-(t_\delta^\star) = D_+(t_\delta^\star) = \delta$, we have*

$$\inf_{\widehat{f}_{\delta,n}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}^{\otimes n} \left[d_E \left(\widehat{f}_{\delta,n}, f_\delta^\star \right) \right] \geqslant C \left[n^{-(\gamma_\delta+1)\beta/(2\beta+d)} + n^{-(\gamma_\delta+1)/(2\gamma_\delta)} \right]. \tag{4.5}$$

(2). **Classical regime.** *Otherwise, let $\gamma' = \gamma_\infty$ in the automatically fair and fair-boundary cases from Remark 1, i.e., for $D_-(0) \leqslant \delta$; and $\gamma' = \gamma_\delta$ in the fairness-impacted case from Remark 1, if further $D_-(t_\delta^\star) < \delta$ or $D_+(t_\delta^\star) > \delta$. Then, we have*

$$\inf_{\widehat{f}_{\delta,n}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}^{\otimes n} \left[d_E \left(\widehat{f}_{\delta,n}, f_\delta^\star \right) \right] \geqslant C n^{-(\gamma'+1)\beta/(2\beta+d)}. \tag{4.6}$$

We have the following observations:

1. In the fairness-impacted case, there are two sources of error: the estimation error of the regression functions near the decision boundaries, and the estimation error of the thresholds, i.e., balancing the probability of success in each group. The first is well characterized by the boundary behavior of η_a, $a \in \{0, 1\}$. For the second estimation error, when $D_-(t_\delta^\star) < \delta$ or $D_+(t_\delta^\star) > \delta$, $t_\delta^\star$ can be estimated with a rate faster than the regression functions, see Case (1) of Figure 2. When $D_-(t_\delta^\star) = D_+(t_\delta^\star) = \delta$, and D_+ or D_- is relatively "steep" near $t_\delta^\star$, one can estimate $t_\delta^\star$ at a faster rate (as shown in Case (2) of Figure 2). In contrast, when D_+ or D_- is relatively "flat" near $t_\delta^\star$, the error in estimating $t_\delta^\star$ is larger (See Case (3) of Figure 2). In addition to the population-level accuracy loss due to fairness, the constraint worsens the minimax lower bound when $\gamma_\delta > 1 + d/(2\beta)$.

2. In the automatically fair and fair-boundary cases, the lower bound (4.6) depends only on the error in estimating the regression functions near the optimal threshold $1/2$; and coincides with the lower bound from conventional non-parametric classification problems (Audibert and Tsybakov, 2007).

The proof of Theorem 4.7, presented in Section C.2, consists of two parts, considering the convergence rate of regression functions and of thresholds separately. In the first part, starting with an approach similar to Audibert and Tsybakov (2007), we construct a set of distributions indexed by the hyper-cube and meticulously verify the distributional assumptions. We then leverage Assouad's lemma to derive the lower bound from (4.6). In the second step, we consider the effect of estimating the thresholds. We depart from the existing proof ideas, introducing a novel construction which provides two very similar distributions with different optimal thresholds $t_\delta^\star$. Then, by applying Le Cam's Lemma, we establish the additional term of the lower bound for the fairness-impacted case.

[5] We have $\delta_{\text{sign}} = \delta$ in the δ-positive DDP case when $D_-(0) > \delta$ and $\delta_{\text{sign}} = -\delta$ in the δ-negative DDP case when $D_+(0) < -\delta$.

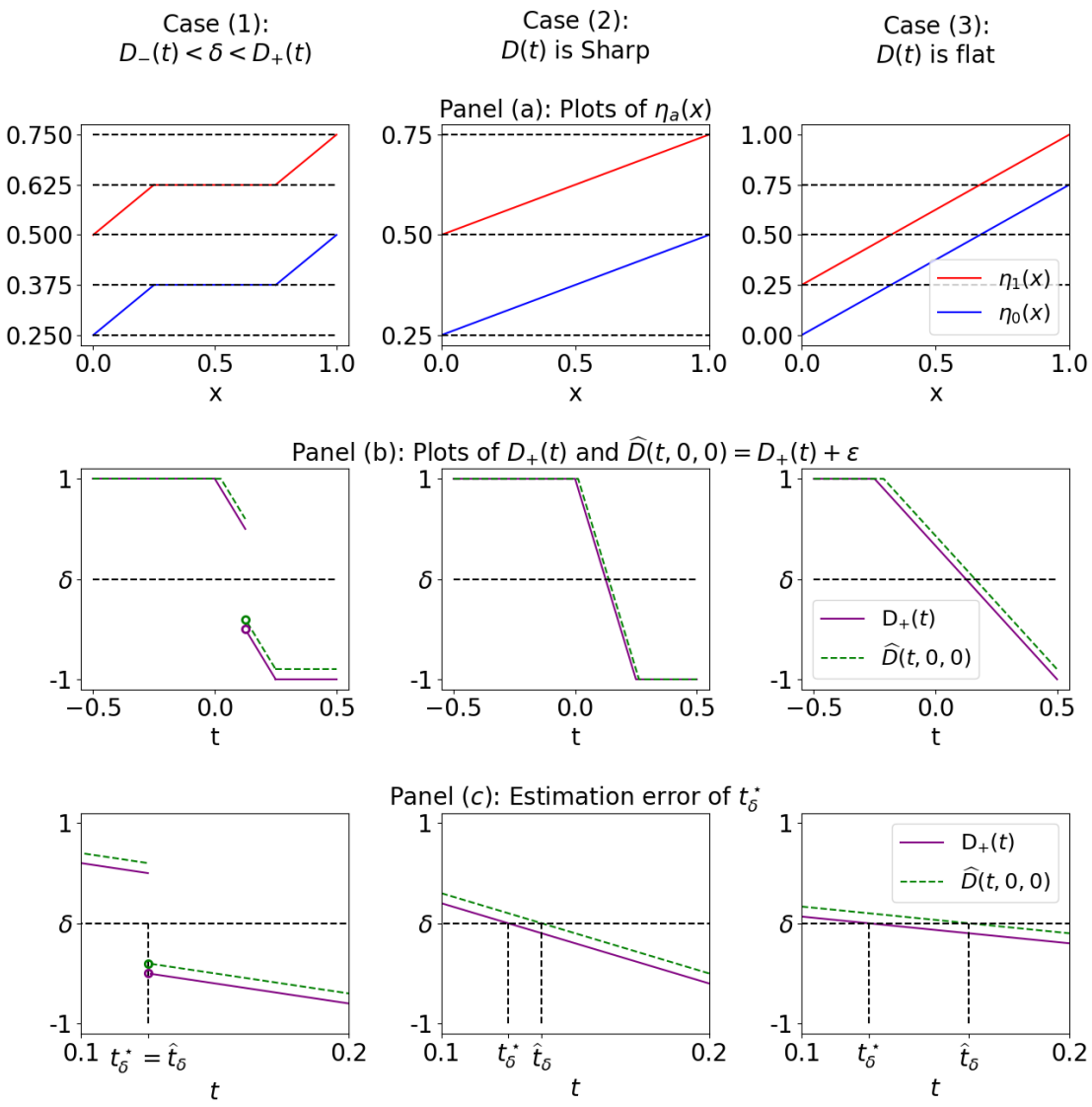


Figure 2: Estimation error of $t^\star_\delta$ in three cases, with $\mathbb{P}(A = 1) = 1/2$, $X|A = a \sim U(0,1)$ and $\delta = 0$. As we can see, when $D_-(t^\star_\delta) < \delta < D_+(t^\star_\delta)$, $t^\star_\delta$ can be estimated with a fast rate. When $\delta = D_-(t^\star_\delta)$ (or $\delta = D_+(t^\star_\delta)$), the convergence rate depends on the slope of D_- (or D_+) near $t^\star_\delta$.

5 FairBayes-DDP+: Plug-in Thresholding Rule with Offset

In this section, we complete the picture by proposing an adaptive thresholding estimator that achieves the minimax lower bounds. Together with Theorem 4.7, this establishes the minimax convergence rate in our fair classification problems. We first introduce the required estimators.

5.1 Local Polynomial Estimator of the Regression Function

In this section, we recall the definition of the local polynomial estimator of the regression function (e.g, Tsybakov, 2009; Audibert and Tsybakov, 2007, etc). For a random variable (X, Y) over $\mathbb{R}^d \times \mathbb{R}$ and n i.i.d. copies $(X_i, Y_i)_{i=1}^n$, the local polynomial estimator of the regression function $\eta : \mathbb{R}^d \to \mathbb{R}$, such that for all $x \in \mathbb{R}^d$, $\eta(x) = \mathbb{E}(Y|X = x)$, is defined as follows.

Definition 5.1. *For a bandwidth $h > 0$, $x \in \mathbb{R}^d$, for an integer $k \geqslant 0$ and a kernel $K : \mathbb{R}^d \to [0, \infty)$, denote*

by $\widehat{\theta}_x$ *a polynomial on* $\mathbb{R}^d$ *of degree* k*, whose* $k+1$ *coefficients minimize*[6] *over* $\mathbb{R}^{k+1}$

$$\sum_{i=1}^{n} \left(Y_i - \widehat{\theta}_x(X_i - x)\right)^2 K\left(\frac{X_i - x}{h}\right). \tag{5.1}$$

The local polynomial estimator $\widehat{\eta}_n^{\mathrm{LP}}(x)$ *of order* k *of the value* $\eta(x)$ *of the regression function at the point* x *is defined by* $\widehat{\eta}_n^{\mathrm{LP}}(x) = \widehat{\theta}_x(0)$ *if* $\widehat{\theta}_x$ *is a unique minimizer of* (5.1)*, and* $\widehat{\eta}_n^{\mathrm{LP}}(x) = 0$ *if* $\widehat{\theta}_x$ *is a non-unique minimizer.*

For a multi-index $s \in \mathbb{N}^d$, and $x \in \mathbb{R}^d$, we introduce the vector $U(x) = (x^s)_{|s| \leqslant k}$ and matrix $Q = (Q_{s_1,s_2})_{|s_1|,|s_2| \leqslant k}$ with

$$Q_{s_1,s_2}(x) = \sum_{i=1}^{n} (X_i - x)^{s_1+s_2} K\left(\frac{X_i - x}{h}\right).$$

Audibert and Tsybakov (2007) show that if the matrix Q is positive definite, there exists a unique polynomial on $\mathbb{R}^d$ of degree k minimizing (5.1). The corresponding local polynomial regression function estimator equals, for all $x \in \mathbb{R}^d$,

$$\eta_n^{\mathrm{LP}}(x) = \sum_{i=1}^{n} Y_i K\left(\frac{X_i - x}{h}\right) U^\top(0) Q^{-1} U(X_i - x).$$

We refer readers to Audibert and Tsybakov (2007) and Tsybakov (2009) for more details about local polynomial estimators.

5.2 Bandwidth Parameter with Possible Jump Discontinuity

We saw in Theorem 4.7 that even in the fairness-impacted case, if $D_-(t_\delta^\star) < \delta$ or $D_+(t_\delta^\star) > \delta$, the fair Bayes-optimal classifier can be estimated with the classical convergence rate for non-parametric regression. This happens if $t_\delta^\star$ is a jump discontinuity point[7] of $t \mapsto D_-(t) - \delta$ or $\mapsto D_+(t) - \delta$; by definition, D_- and D_+ share the same jump discontinuity points. Even though the disparity functions D_-, D_+ are estimated at a non-parametric rate, its jump discontinuities can be estimated with a near-parametric rate. We consider the following strategy to estimate $t_\delta^\star$, by taking possible discontinuities into account.

Let $(\Delta_n)_{n \geqslant 0}$ and $(r_n)_{n \geqslant 0}$ be two positive sequences that converge to zero slowly (e.g., at rates on the order of $(\log\log n)^{-1}$ and $(\log n)^{-1}$, respectively). First, consider the case where D_- and D_+ have a jump discontinuity at $t_\delta^\star$. If $\delta < D_+(t_\delta^\star)$, for $\Delta_n \to 0_+$, we have $D_+(t_\delta^\star) > \delta + \Delta_n > D_-(t_\delta^\star)$. This implies that $\inf_{t>0}\{D_+(t) > \delta + \Delta_n\} = t_\delta^\star$. Similarly, if $D_-(t_\delta^\star) < \delta$, we have $\inf_{t>0}\{D_-(t) > \delta - \Delta_n\} = t_\delta^\star$. Other other hand, if both D_- and D_+ are continuous at $t_\delta^\star$ with $D_-(t_\delta^\star) = D_+(t_\delta^\star) = \delta$, it holds that $\inf_{t>0}\{D_+(t) > \delta + \Delta_n\} < t_\delta^\star$ and $\inf_{t>0}\{D_-(t) > \delta + \Delta_n\} > t_\delta^\star$. Motivated by this, letting $\widehat{D}_n$ be an estimate of D_- (or D_+), we define $\widehat{t}_{\delta,\mathrm{mid}}$, $\widehat{t}_{\delta,\Delta_n,\min}$ and $\widehat{t}_{\delta,\Delta_n,\max}$ as in (5.3) and $\widehat{t}_\delta$ as in (5.4) using empirical versions of the above relations.

5.3 Plug-in Estimators with Offset

Next, we consider estimating the group-wise probability functions via plug-in estimators with an offset (Rigollet and Vert, 2009). For a density function g defined on $\mathcal{X}$, the λ-level set of g is $\Lambda_g(\lambda) = \{x \in \mathcal{X},\ g(x) > \lambda\}$. If $\widehat{g}$ is a consistent estimator of g, a first thought is to estimate $\Lambda_g(\lambda)$ by the plug-in estimator $\widehat{\Lambda}_g(\lambda) = \{x \in \mathcal{X},\ \widehat{g}(x) > \lambda\}$. However, this can be inconsistent if the boundary set $\{x \in \mathcal{X},\ g(x) = \lambda\}$ has positive probability. Alternatively, Rigollet and Vert (2009) proposed plug-in density estimators with offset $(\ell_n)_{n \geqslant 1}$:

$$\widetilde{\Lambda}_{g,\ell_n}(\lambda) = \widehat{\Lambda}_g(\lambda + \ell_n) = \{x \in \mathcal{X},\ \widehat{g}(x) > \lambda + \ell_n\},$$

where $\ell_n \geqslant 0$ tends to zero as n tends to infinity; see Appendix D for further discussion. Similarly, $\{x \in \mathcal{X},\ g(x) < \lambda\}$ and $\{x \in \mathcal{X},\ g(x) = \lambda\}$ can be consistently estimated by $\{x \in \mathcal{X},\ \widehat{g}(x) \leqslant \lambda - \ell_n\}$ and $\{x \in \mathcal{X},\ \lambda - \ell_n < \widehat{g}(x) \leqslant \lambda + \ell_n\}$, respectively; and we will adapt such ideas to our problem.

[6] A minimizer always exists, but may not be unique.

[7] For a function $D : \mathbb{R} \to \mathbb{R}$, we say that t_0 is a jump discontinuity point of D if both the left limit $\lim_{t \to t_0^-} D(t)$ and right limit $\lim_{t \to t_0^+} D(t)$ of D at t_0 are finite, and $\lim_{t \to t_0^+} D(t) \neq \lim_{t \to t_0^-} D(t)$.

Algorithm 1 FairBayes-DDP+: Thresholding Rule with Offset for Fair Classification under Demographic Parity

Input: Disparity level δ. Dataset $\mathcal{S}_n = S_{n,1} \cup S_{n,0}$ with $\mathcal{S}_n = \{(x_i, a_i, y_i)\}_{i=1}^n$ and, for $a \in \{0,1\}$, $S_{n,a} = \{(x_{a,j}, a, y_{a,j})\}_{j=1}^{n_a}$.

Step 1: Estimate η_a and η_0 by local polynomial estimators:

Let $U(x) = (x^s)_{|s|\leqslant\lfloor\beta\rfloor_+}$, $Q_a = (Q_{a,s_1,s_2})_{|s_1|,|s_2|\leqslant\lfloor\beta\rfloor_+}$ with $Q_{a,s_1,s_2} = \sum_{j=1}^{n_a} (x_{a,j} - x)^{s_1+s_2} K\left(\frac{x_{a,j}-x}{h_{n,a}}\right)$, and $\bar{B}_a = (\bar{B}_{a,s_1,s_2})_{|s_1|,|s_2|\leqslant\lfloor\beta\rfloor_+}$ with $\bar{B}_{a,s_1,s_2} = n_a^{-1} h_{n,a}^{-(d+s_1+s_2)} Q_{a,s_1,s_2}$.

Define

$$\eta_{n,a}^{\mathrm{LP}}(x) = \sum_{j=1}^{n_a} Y_{a,j} K\left(\frac{x_{a,j}-x}{h_{n,a}}\right) U^\top(0) Q^{-1} U(X_i - x).$$

Let

$$\widehat{\eta}_a(x) = \begin{cases} 0; & \lambda_{\min}(\bar{B}_a) \leqslant (\log n_a)^{-1} \text{ or } \eta_{n,a}^{\mathrm{LP}}(x) < 0; \\ 1; & \lambda_{\min}(\bar{B}_a) > (\log n_a)^{-1} \text{ and } \eta_{n,a}^{\mathrm{LP}}(x) > 1; \\ \eta_{n,a}^{\mathrm{LP}}(x), & \text{otherwise.} \end{cases}$$

Step 2: Estimate the optimal thresholds:

$$\widehat{D}_n(t, \ell_{n,1}, \ell_{n,0}) = \frac{1}{n_1}\sum_{j=1}^{n_1} I\left(\widehat{\eta}_1(x_{1,j}) > \frac{1}{2} + \frac{nt}{2n_1} + \ell_{n,1}\right) - \frac{1}{n_0}\sum_{j=1}^{n_0} I\left(\widehat{\eta}_0(x_{0,j}) > \frac{1}{2} - \frac{nt}{2n_0} + \ell_{n,0}\right). \quad (5.2)$$

Set

$$\begin{cases} \widehat{t}_{\delta,\mathrm{mid}} = \inf_{t\geqslant 0}\left\{\widehat{D}_n(t,0,0) < \delta\right\}; \\ \widehat{t}_{\delta,\Delta_n,\min} = \inf_{t\geqslant 0}\left\{\widehat{D}_n(t,0,0) < \delta + \Delta_n\right\}; \\ \widehat{t}_{\delta,\Delta_n,\max} = \inf_{t\geqslant 0}\left\{\widehat{D}_n(t,0,0) < \delta - \Delta_n\right\}, \end{cases} \quad (5.3)$$

$$\widehat{t}_\delta = \begin{cases} \widehat{t}_{\delta,\Delta_n,\min}, & \widehat{t}_{\delta,\mathrm{mid}} - \widehat{t}_{\delta,\Delta_n,\min} \leqslant r_n; \\ \widehat{t}_{\delta,\Delta_n,\max}, & \widehat{t}_{\delta,\mathrm{mid}} - \widehat{t}_{\delta,\Delta_n,\min} > r_n \text{ and } \widehat{t}_{\delta,\Delta_n,\max} - \widehat{t}_{\delta,\mathrm{mid}} \leqslant r_n; \\ \widehat{t}_{\delta,\mathrm{mid}}, & \widehat{t}_{\delta,\mathrm{mid}} - \widehat{t}_{\delta,\Delta_n,\min} > r_n \text{ and } \widehat{t}_{\delta,\Delta_n,\max} - \widehat{t}_{\delta,\mathrm{mid}} > r_n. \end{cases} \quad (5.4)$$

and

$$\widehat{\tau}_{\delta,1} = \rho\left(\frac{\widehat{\pi}_{n,0,+} - \widehat{\pi}_{n,1,+} + \widehat{\delta}}{\widehat{\pi}_{n,1,=}}\right) \quad \text{and} \quad \widehat{\tau}_{\delta,0} = \rho\left(\frac{\widehat{\pi}_{n,1,+} - \widehat{\pi}_{n,0,+} - \widehat{\delta}}{\widehat{\pi}_{n,0,=}}\right). \quad (5.5)$$

with

$$\widehat{\pi}_{n,a,+} = \frac{1}{n_a}\sum_{j=1}^{n_a} I\left(\widehat{\eta}_a(x_{a,j}) > \widehat{T}_{\delta,a} + \ell_{n,a}\right), \quad \widehat{\pi}_{n,a,=} = \frac{1}{n_a}\sum_{j=1}^{n_a} I\left(\widehat{T}_{\delta,a} - \ell_{n,a} < \widehat{\eta}_a(x_{a,j}) \leqslant \widehat{T}_{\delta,a} + \ell_{n,a}\right), \quad (5.6)$$

and

$$\widehat{\delta} = \delta \cdot I\left(\widehat{D}_n(0, \ell_{n,1}, -\ell_{n,0}) > \delta\right). \quad (5.7)$$

Output: $\widehat{f}_{\delta,n}$ from (5.13).

5.4 FairBayes-DDP+: Plug-in Estimator with Offset for Fair Classification

In this section, we introduce our FairBayes-DDP+ method, a classifier for minimax optimal classification under demographic parity. This method is based on a two-stage plug-in estimator with offsets, where the first stage estimates the regression functions for each group and the second stage estimates the thresholding adjustment parameter $t_\delta^\star$. As elaborated in Appendix A, the behavior of the fair Bayes-optimal classifier on the decision boundary is generally not unique. For identifiability, we estimate a specific Bayes-optimal

classifier, chosen such that the probability of $\widehat{Y}=1$ is minimized on the decision boundaries. Consider the truncation function $\rho:\mathbb{R}\cup\{\infty\}\to[0,1]$ defined as

$$\rho(x)=\begin{cases}0, & x\leqslant 0\\ 1, & x\geqslant 1;\\ x, & \text{otherwise.}\end{cases} \tag{5.8}$$

and the adjustment of δ in the fairness-impacted case given by

$$\widetilde{\delta}=\delta I\left(D_-(0)>\delta\right)=\begin{cases}0, & D_-(0)\leqslant\delta;\\ \delta, & D_-(0)>\delta.\end{cases} \tag{5.9}$$

We set, for $a\in\{0,1\}$,

$$\pi^\star_{a,+}=\mathbb{P}_{X|A=a}(\eta_a(X)>T^\star_{\delta,a}) \qquad \text{and} \qquad \pi^\star_{a,=}=\mathbb{P}_{X|A=a}(\eta_a(X)=T^\star_{\delta,a}), \tag{5.10}$$

and interpreting $x/0=0$ for all $x\in\mathbb{R}$ here and in what follows, the randomization probabilities

$$\tau^\star_{\delta,1}=\rho\left(\frac{\pi^\star_{0,+}-\pi^\star_{1,+}+\widetilde{\delta}}{\pi^\star_{1,=}}\right) \qquad \text{and} \qquad \tau^\star_{\delta,0}=\rho\left(\frac{\pi^\star_{1,+}-\pi^\star_{0,+}-\widetilde{\delta}}{\pi^\star_{0,=}}\right). \tag{5.11}$$

These choices lead to a specific fair Bayes-optimal classifier with the following properties.

Proposition 5.2 (Properties of a specific fair Bayes-optimal classifier)**.** *Consider the fair Bayes-optimal classifier $f^\star_\delta$ from (3.7), with the thresholds $T^\star_{\delta,a}$ from (3.6) and the randomization probabilities $\tau^\star_{\delta,a}$ from (5.11).*

- *In the fairness-impacted case where $D_-(0)>\delta$, we have that* $\mathrm{DDP}(f^\star_\delta)=\delta$;
- *In the automatically fair and fair-boundary cases where $D_-(0)\leqslant\delta$, we have that*

$$\mathrm{DDP}(f^\star_\delta)=0 \text{ if } D_-(0)\leqslant 0, \quad \text{and} \quad \mathrm{DDP}(f^\star_\delta)=D_-(0) \text{ if } 0<D_-(0)\leqslant\delta.$$

Now, suppose we have a dataset $\mathcal{S}_n=\{(x_i,a_i,y_i)\}_{i=1}^n$. We separate the data according to the protected information: for $a\in\{0,1\}$, we let $\mathcal{S}_{n,a}=\{(x_i,a_i,y_i)\in\mathcal{S}_n,a_i=a\}$ with $n_a:=|\mathcal{S}_{n,a}|$. The j-th element of $\mathcal{S}_{n,a}$ is denoted as $(x_{a,j},a,y_{a,j})$, for $j\in[n_a]$.

Step 1: First, we estimate η_a via local polynomial estimation using $\mathcal{S}_n$. Specifically, consider a kernel $K:\mathbb{R}^d\to\mathbb{R}$ satisfying

$$\begin{aligned}&(1)\ \text{there is } c>0 \text{ with } K(x)\geqslant cI(\|x\|\leqslant c), \ \text{ for all } x\in\mathbb{R}^d; && (2)\ \textstyle\int_{\mathbb{R}^d}K(t)dt=1;\\ &(3)\ \textstyle\int_{\mathbb{R}^d}(1+\|t\|^{4\beta})K^2(t)dt<\infty; && (4)\ \textstyle\sup_{t\in\mathbb{R}^d}(1+\|t\|^{2\beta})K(t)<\infty.\end{aligned} \tag{5.12}$$

One can take, for example, K as the Gaussian kernel. Let $h_{n,a}>0$, and consider the matrix $\bar{B}_a=(\bar{B}_{a,s_1,s_2})_{|s_1|,|s_2|\leqslant\lfloor\beta\rfloor_+}$, where

$$\bar{B}_{a,s_1,s_2}(x)=\frac{1}{n_a h_{n,a}^d}\sum_{j=1}^{n_a}\left(\frac{x_{a,j}-x}{h_{n,a}}\right)^{s_1+s_2}K\left(\frac{x_{a,j}-x}{h_{n,a}}\right).$$

Define the regression function estimator $\widehat{\eta}_a$ as follows. If the smallest eigenvalue of the matrix $\bar{B}_a$ is greater than or equal to $(\log n_a)^{-1}$, for all x, we set $\widehat{\eta}_a(x)$ equal to the projection of $\widehat{\eta}_a^{\mathrm{LP}}(x)$ on the interval $[0,1]$, where $\widehat{\eta}_a^{\mathrm{LP}}$ is the $LP(\lfloor\beta\rfloor_+)$ estimator of η_a with bandwidth $h_{n,a}>0$ and kernel K satisfying (5.12). If the smallest eigenvalue of $\bar{B}_a$ is less than $(\log n_a)^{-1}$, we set $\widehat{\eta}_a(x)=0$.

Step 2: In the second step, we start by estimating the acceptance threshold for each protected group, via solving a one-dimensional empirical fairness constraint and then determining the prediction on the decision boundaries when those have strictly positive estimated probability. We observe that the thresholds of the fair Bayes-optimal classifier from (3.7) balance the probability measures of the level sets $\{x\in\mathcal{X},\eta_1(x)>T^\star_{\delta,1}\}$

and $\{x \in \mathcal{X}, \eta_0(x) > T^\star_{\delta,0}\}$. As a result, we can incorporate plug-in estimation with an offset for level set estimation. Specifically, for $a \in \{0,1\}$, some $\ell_{n,a} > 0$, and any $\zeta \in \mathbb{R}$, we estimate $\mathbb{P}_{X|A=a}(\eta_a(X) > \zeta)$ and $\mathbb{P}_{X|A=a}(\eta_a(X) \geqslant \zeta)$ by $n_a^{-1}\sum_{j=1}^{n_a} I\left(\widehat{\eta}_a\left(x_{a,j}\right) > \zeta + \ell_{n,a}\right)$ and $n_a^{-1}\sum_{j=1}^{n_a} I\left(\widehat{\eta}_a\left(x_{a,j}\right) > \zeta - \ell_{n,a}\right)$, respectively. With this, the probability of the decision boundary can be consistently estimated. Based on (3.7), we consider the group-wise thresholding rule defined for all x, a by

$$f^t_\ell(x,a) = I\left(\widehat{\eta}_a(x) > \frac{1}{2} + \frac{nt}{2(2a-1)n_a} + \ell_{n,a}\right) + \tau_a I\left(\left|\widehat{\eta}_a(x) - \frac{1}{2} + \frac{nt}{2(2a-1)n_a}\right| \leqslant \ell_{n,a}\right),$$

where η_a and p_a are estimated by plug-in estimators.

Next, our goal is to construct estimates $\widehat{t}_\delta$ and $\widehat{\tau}_{\delta,a}$ such that the proposed classifier approximately satisfies the fairness constraint. With $\widehat{D}_n(t, \ell_{n,1}, \ell_{n,0})$ from (5.2), we consider plug-in estimators of $D_-(t)$ and $D_+(t)$ given by $\widehat{D}_n(t, \ell_{n,1}, -\ell_{n,0})$ and $\widehat{D}_n(t, -\ell_{n,1}, \ell_{n,0})$ with $\ell_{n,1}, \ell_{n,0} > 0$, respectively. We estimate $t^\star_\delta$ using the approach introduced in Section 5.2. Let $\widehat{t}_\delta$ defined as in (5.4) and let, for $a \in \{0,1\}$, $\widehat{T}_{\delta,a} = 1/2 + (2a-1)n\widehat{t}_\delta/n_a$. We estimate $\tau^\star_{\delta,1}$ and $\tau^\star_{\delta,1}$ specified in (5.11) by $\widehat{\tau}_{\delta,a}$ from (5.5), using the plug-in estimates with offsets $(\widehat{\pi}_{n,a,+}, \widehat{\pi}_{n,a,=})$ of $(\pi^\star_{a,+}, \pi^\star_{a,+})$ from (5.6), respectively. Also, $\widetilde{\delta}$ from (5.9) is estimated by $\widehat{\delta}$ from (5.7). Our final estimate of the δ-fair Bayes-optimal classifier is

$$\widehat{f}_{\delta,n}(x,a) = I\left(\widehat{\eta}_a(x) > \widehat{T}_{\delta,a} + \ell_{n,a}\right) + \widehat{\tau}_{\delta,a} I\left(|\widehat{\eta}_a(x) - \widehat{T}_{\delta,a}| \leqslant \ell_{n,a}\right). \tag{5.13}$$

Our plug-in method is directly motivated by the fair Bayes-optimal classifier from Theorem A.1. The offsets $\ell_{n,a}, a \in \{0,1\}$ are carefully designed to handle estimation on the boundaries.

Remark 2. *In Step* 1 *of our method, we use the local polynomial estimators only for theoretical purposes, as they lead to an upper bound matching the minimax lower bound. However, as we show in our experiments, in practice we can use other methods, such as support vector machines or deep neural networks, to estimate the regression function for improved performance.*

6 Asymptotic Analysis of FairBayes-DDP+

In this section, we study the statistical properties of FairBayes-DDP+. We first derive the convergence rate of our plug-in method, establishing its minimax optimality. We then show that the constraint $|\mathrm{DDP}(f)| \leqslant \delta$ is satisfied up to a vanishing error term.

6.1 Convergence Rate and Minimax Optimality

In this section, we establish the convergence rate of FairBayes-DDP+. The rate depends on the pointwise convergence of $\widehat{\eta}_a$ to η_a, $a \in \{0,1\}$. To quantify this rate, the following definition describes a notion of pointwise convergence of a sequence of estimators of the conditional probability functions η_a, $a \in \{0,1\}$.

Definition 6.1 (Pointwise convergence)**.** *Let $\mathcal{P}$ be a class of distributions for (X, A, Y) and fix $U_\eta > 0$. Let $(\phi_{n,1})_{n\geqslant 1}$ and $(\phi_{n,0})_{n\geqslant 1}$ be two positive, monotonically non-increasing sequences. We say that the estimator sequence $(\widehat{\eta}_{n,1}, \widehat{\eta}_{n,0})_{n\geqslant 1}$, where $\widehat{\eta}_{n,1}, \widehat{\eta}_{n,0}$ is constructed using a sample of size n, converges pointwise at rate $(\phi_{n,1}, \phi_{n,0})_{n\geqslant 1}$ uniformly over $\mathcal{P}$ if there are positive constants $c_{1,\eta}$, $c_{2,\eta}$ and L_η, as well as a set $\Omega \subset \mathcal{X}$, such that $\mathbb{P}(\Omega) = 1$ and, for $a \in \{0,1\}$,*

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}^{\otimes n}\left(\sup_{x\in\Omega} |\widehat{\eta}_{n,a}(x) - \eta_a(x)| > \varepsilon\right) \leqslant c_{1,\eta} \exp\left(-c_{2,\eta}\left(\varepsilon/\phi_{n,a}\right)^2\right), \quad L_\eta \phi_{n,a} < \varepsilon < U_\eta. \tag{6.1}$$

We will usually drop the subscript n and write $\widehat{\eta}_a = \widehat{\eta}_{n,a}$. In the rest of this paper, we let, for $\varepsilon > 0$, for $i \in [4]$ and $\iota \in \{t, T, t_1, t_2, r, D, \pi\}$, and for quantities $c_{i,\iota} > 0$,

$$\psi_{n,1,\iota}(\varepsilon) = c_{1,\iota} \exp\left(-c_{2,\iota}\left(\varepsilon/[\phi_{n,1} \vee \phi_{n,0}]\right)^2\right) \text{ and } \psi_{n,2,\iota}(\varepsilon) = c_{3,\iota} \exp\left(-c_{4,\iota} n\varepsilon^2\right). \tag{6.2}$$

With $t^\star_\delta$ from (3.5), D_- and D_+ from (3.3) and (3.4), we denote by $\widetilde{I}^\star(\delta)$ the indicator function of the non-trivial fairness-impacted regime introduced in Theorem 4.7, i.e.,

$$\widetilde{I}^\star(\delta) = I(\{t^\star_\delta > 0\} \cap \{D_-(t^\star_\delta) = D_+(t^\star_\delta) = \delta\}). \tag{6.3}$$

Moreover, for $\varepsilon > 0$ and $r_n > 0$, we write

$$\omega(\varepsilon, r_n) = \widetilde{I}^\star(\delta) \cdot \varepsilon + (1 - \widetilde{I}^\star(\delta)) \cdot r_n, \tag{6.4}$$

which selects ε in the non-trivial fairness-impacted regime, and r_n otherwise. We derive a a general and abstract convergence rate for $\widehat{t}_\delta$ below, assuming the convergence of $\widehat{\eta}_a$. Later we will apply this result to our concrete setting. For two scalars a, b, we denote their maximum by $\max\{a, b\}$ or $a \vee b$, and their minimum by $\min\{a, b\}$ or $a \wedge b$.

Theorem 6.2 (Error bound for estimating $t^\star_\delta$)**.** *Let $\mathcal{P}$ be a class of densities on $\mathcal{X} \times \mathcal{A} \times \mathcal{Y}$, let $\delta \geqslant 0$, and let $(\phi_{n,1}, \phi_{n,0})_{n \geqslant 1}$ be two positive, monotonically non-increasing sequences such that, for constants $c_\mu > 0$, $\tilde{\mu}_a \geqslant 1/2$ for $a \in \{0, 1\}$, we have $\phi_{n,a} \geqslant c_\mu n^{-\tilde{\mu}_a}$. Suppose that $(\widehat{\eta}_1, \widehat{\eta}_0)$ are $(\phi_{n,1}, \phi_{n,0})_{n \geqslant 1}$-pointwise convergent to (η_1, η_0) as per Definition 6.1, uniformly over $\mathcal{P}$. Then, with D_- and D_+ from (3.3) and (3.4), there are constants $c_{i,t}$, $i \in [4]$, L_t, U_t and $U_{\Delta,t}$ such that, if $L_t(\phi_{n,1} \vee \phi_{n,0}) < r_n < U_t$, $2(g_{\delta,-}(4r_n) \vee g_{\delta,+}(4r_n)) < \Delta_n < U_\Delta$ and $L_t(\phi_{n,1} \vee \phi_{n,0}) < \varepsilon < r_n$ hold, then with $\psi_{n,j,t}$, $j \in \{1, 2\}$ from (6.2) and ω from (6.4), we have for any $\varepsilon > 0$ that*

$$\mathbb{P}^{\otimes n}\left(|\widehat{t}_\delta - t^\star_\delta| > \varepsilon\right) \leqslant \psi_{n,1,t}(\varepsilon) + \sum_{j \in \{-,+\}} I(\delta = D_j(t^\star_\delta)) \psi_{n,2,t}\left(g_{\delta,j}\left(\omega(\varepsilon, r_n)\right)\right). \tag{6.5}$$

The bound in (6.5) consists of two parts. The first is determined by the convergence rates of $\widehat{\eta}_1$ and $\widehat{\eta}_0$, as $\widehat{t}_\delta$ is based on them. The second depends on the behavior of the conditional probability functions near the decision boundary, for the same reason as explained after Theorem 4.7. When $D_-(t^\star_\delta) < \delta$ or $\delta < D_-(t^\star_\delta)$, the second term in (6.5) disappears, and the convergence rate of $\widehat{t}_\delta$ depends only on the convergence of $\widehat{\eta}_1$ and $\widehat{\eta}_0$.

By definition, we have $\widehat{T}_{\delta,a} = 1/2 + (2a - 1)\widehat{t}_\delta n/(2n_a)$ and $T^\star_{\delta,a} = 1/2 + (2a - 1)t^\star_\delta/(2p_a)$. Moreover, n/n_a is a root-n-consistent estimate of $1/p_a$ when $p_a > 0$. Building on these observations, we can show the following corollary, still in an abstract setting:

Corollary 6.3 (Error bound for optimal thresholds)**.** *Under the condition of Theorem 6.2, there are constants $c_{i,T}$, $i \in [4]$, U_T, L_T and U_Δ such that, if $L_T(\phi_{n,1} \vee \phi_{n,0}) < r_n < U_T$, $2(g_{\delta,-}(4r_n) \vee g_{\delta,+}(4r_n)) < \Delta_n < U_{\Delta,T}$ and $L_T(\phi_{n,1} \vee \phi_{n,0}) < \varepsilon < r_n$ hold, then with $\psi_{n,j,T}$, $j \in \{1, 2\}$, from (6.2) and ω from (6.4), we have for any $\varepsilon > 0$ that*

$$\mathbb{P}^{\otimes n}\left(|\widehat{T}_{\delta,a} - T^\star_{\delta,a}| > \varepsilon\right) \leqslant \psi_{n,1,T}(\varepsilon) + \sum_{j \in \{-,+\}} I(\delta = D_j(t^\star_\delta)) \psi_{n,2,T}\left(g_{\delta,j}\left(\omega(\varepsilon, r_n)\right)\right). \tag{6.6}$$

Assuming the convergence rates of the estimated regression functions in (6.1), and with the results on the thresholds from (6.5), we can show that FairBayes-DDP+ is asymptotically fair and accurate, again first in an abstract setting.

Theorem 6.4 (Fairness-aware excess risk upper bound; abstract version)**.** *For any $\delta \geqslant 0$, suppose that the conditions of Theorem 6.2 hold. Suppose further that the regression functions (η_1, η_0) satisfy the γ_δ-margin condition. Then, the plug-in estimate $\widehat{f}_{\delta,n}$ with offsets $\ell_{n,a}$ from (5.13), $a \in \{0, 1\}$, satisfies, with $\widetilde{I}^\star(\delta)$ from (6.3),*

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}^{\otimes n}\left[d_E\left(\widehat{f}_{\delta,n}, f^\star_\delta\right)\right] \leqslant C\left((\phi_{n,1} \vee \phi_{n,0} \vee \ell_{n,1} \vee \ell_{n,0}) + \widetilde{I}^\star(\delta) n^{-1/(2\gamma_\delta)}\right)^{\gamma_\delta + 1}. \tag{6.7}$$

The convergence of $d_E(\widehat{f}_{\delta,n}, f^\star_\delta)$ remains unaffected by the offsets when $\ell_{n,1} \vee \ell_{n,0} \leqslant C(\phi_{n,1} \vee \phi_{n,0})$. Indeed, the boundary effects are negligible when considering d_E, as the expression $(f(x,a) - f^\star_\delta(x,a))(T^\star_{\delta,a} - \eta_a(x)) \equiv 0$ holds for any f on the boundary sets. However, the offsets are key to ensuring the asymptotic fairness of

our method, as demonstrated in the next section. Additionally, they also impact the accuracy through Proposition 4.2.

We can make this result concrete by leveraging the point-wise convergence of the local polynomial estimator from Audibert and Tsybakov (2007), for the appropriate choice of $h_{n,a}$. With this we now show that FairBayes-DDP+ achieves the minimax lower bound derived in Theorem 4.7.

Corollary 6.5 (Fairness-aware excess risk upper bound)**.** *Consider the class of densities $\mathbb{P}_\Sigma$ defined in Definition 4.6. For any $\delta > 0$, consider the FairBayes-DDP+ classifier $\widehat{f}_{\delta,n}$ from (5.13), where for $a \in \{0,1\}$, η_a is estimated by the local polynomial estimators of (η_1, η_0) with kernel K satisfying (5.12) and $h_{n,a} \asymp n_a^{-1/(2\beta+d)}$, $\Delta_n \asymp (\log\log n)^{-1}$, $r_n \asymp (\log n)^{-1}$, and the offsets satisfy $\ell_{n,1}, \ell_{n,0} \asymp n^{-\beta/(2\beta+d)}$. Then, we have, with $\widetilde{I}^\star(\delta)$ from (6.3),*

$$\sup_{\mathbb{P}\in\mathcal{P}_\Sigma} \mathbb{E}^{\otimes n}\left[d_E\left(\widehat{f}_{\delta,n}, f_\delta^\star\right)\right] \leqslant C\left[n^{-(\gamma_\delta+1)\beta/(2\beta+d)} + \widetilde{I}^\star(\delta) n^{-(\gamma_\delta+1)/(2\gamma_\delta)}\right].$$

The convergence rate stated in equation (4.5) matches the minimax lower bound specified in Theorem 4.7. Specifically,

(1) In the non-trivial fairness-impacted regime, i.e., $t_\delta^\star > 0$ and $D_-(t_\delta^\star) = D_+(t_\delta^\star) = \delta$, we have $\widetilde{I}^\star(\delta) = 1$ and
$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}^{\otimes n}\left[d_E\left(\widehat{f}_{\delta,n}, f_\delta^\star\right)\right] \leqslant C\left[n^{-(\gamma_\delta+1)\beta/(2\beta+d)} + n^{-(\gamma_\delta+1)/(2\gamma_\delta)}\right].$$

(2) In the classical regime, i.e., (2.1) $t_\delta^\star > 0$ or (2.2) $D_-(t_\delta^\star) < \delta$ or (2.3) $D_+(t_\delta^\star) > \delta$, we have $\widetilde{I}^\star(\delta) = 0$. Thus, with $\gamma' = \gamma_\infty$ for the automatically fair and fair-boundary cases ($t_\delta^\star = 0$), and $\gamma' = \gamma_\delta$ for the fairness impacted case ($t_\delta^\star > 0$), we have $\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}^{\otimes n}\left[d_E\left(\widehat{f}_{\delta,n}, f_\delta^\star\right)\right] \leqslant C n^{-(\gamma'+1)\beta/(2\beta+d)}$.

This implies that our FairBayes-DDP+ classifier is minimax optimal.

6.2 Asymptotic Fairness

For any $\delta \geqslant 0$, at the population level, our fairness constraint enforces that $|\mathrm{DDP}(f)| \leqslant \delta$. However, based on a finite sample, in general one may slightly violate the constraint. When $\delta = 0$, Fukuchi and Sakuma (2023) defined a learning algorithm with output $\widehat{f}_{\delta,n}$ to be (α,ξ)-consistently fair—for $\alpha > 0$ and $\xi > 0$—for an unfairness measure $U : \mathcal{F} \to \mathbb{R}$, if there are constants $n_0 \geqslant 0$ and $C > 0$ independent of n such that $\mathbb{P}(U(\widehat{f}_{\delta,n}) > Cn^{-\alpha}) \leqslant \xi$ for all $n \geqslant n_0$, over the randomness from the training data. We adapt this definition to fair classification.

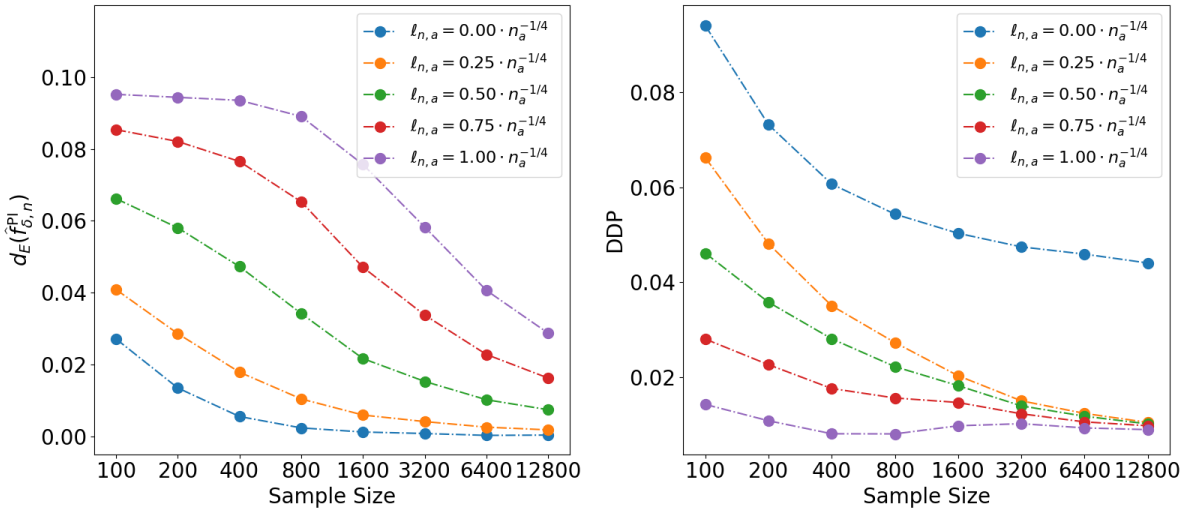
Definition 6.6. *A sequence of classifiers $\widehat{f}_{\delta,n}$ depending on a sample of size n is (δ,α,ξ)-consistently fair under demographic parity if there are constants $n_0 \geqslant 0$ and $C > 0$ independent of n, such that $\mathbb{P}(|\mathrm{DDP}(\widehat{f}_{\delta,n})| > \delta + Cn^{-\alpha}) \leqslant \xi$ for all $n \geqslant n_0$, over the randomness from the training data.*

The following theorem demonstrates that the FairBayes-DDP+ algorithm is consistently fair.

Theorem 6.7. *For any $\delta \geqslant 0$ and $\xi > 0$, there is $C_\xi > 0$ such that, with $\Delta_n \asymp (\log\log n)^{-1}$, $r_n \asymp (\log n)^{-1}$ and offsets satisfying $C_\xi(\phi_{n,1} \vee \phi_{n,0}) < \ell_{n,1} \wedge \ell_{n,0} < r_n$, the FairBayes-DDP+ estimate of the δ-fair Bayes-optimal classifier is $(\delta, 1/2, \xi)$-consistently fair. In particular, there exist constants $c_{D,i}$, $i \in [6]$ and L_ε such that for $L_\varepsilon(\ell_{n,1} \vee \ell_{n,0})^\gamma < \varepsilon \leqslant \sqrt{8(p_1 \wedge p_0)}$, we have*

$$\mathbb{P}^{\otimes n}\left(\left|\mathrm{DDP}(\widehat{f}_{\delta,n})\right| > \delta + \varepsilon\right) \leqslant \psi_{n,1,D}(\ell_{n,1} \wedge \ell_{n,0}) + \sum_{j\in\{-,+\}} I(\delta = D_j(t_\delta^\star))\psi_{n,2,D}\left(g_{\delta,j}(\omega(\ell_{n,1} \wedge \ell_{n,0}, r_n))\right) + c_{5,D}\exp(-c_{6,D} n\varepsilon^2). \tag{6.8}$$

Theorem 6.7 demonstrates that if $\ell_{n,1} \wedge \ell_{n,0} > C_\xi(\phi_{n,1} \vee \phi_{n,0})$, then the disparity level of $\widehat{f}_{\delta,n}$ will be no more than the pre-specified level δ, up to a small term of order $n^{-1/2}$. This lower bound for offsets is necessary to ensure that the boundary sets and their probability measures are consistently estimated. Smaller offsets could lead to inconsistent estimators of $(\tau_{\delta,1}^\star, \tau_{\delta,0}^\star)$, which would increase the risk of violating the fairness constraint. Moreover, the level of offsets also determines the tradeoff between fairness and accuracy. Larger offsets lead to slower convergence rates for the measure d_E, but also to a smaller probability of disparity.



(a) Estimated fairness-aware excess risk

(b) Estimated demographic disparity

Figure 3: Estimated fairness-aware excess risk and DDP of our FairBayes-DDP+ classifier $\widehat{f}_{\delta,n}$ in the setting from Section 7.1, for various sample sizes.

Table 2: Estimated fairness-aware excess risk and DDP of our FairBayes-DDP+ classifier $\widehat{f}_{\delta,n}$ in the setting from Section 7.1, with $\ell_{n,a} = 0.25 \cdot n_a^{-1/(2\beta+d)}$ and various sample sizes.

Sample size	100	200	400	800	1600	3200	6400	12800
$d_E(\widehat{f}_{\delta,n})$	0.041	0.029	0.018	0.010	0.006	0.004	0.003	0.002
(SD)	(0.024)	(0.019)	(0.012)	(0.008)	(0.005)	(0.005)	(0.005)	(0.005)
DDP	0.066	0.048	0.035	0.027	0.020	0.015	0.012	0.010
(SD)	(0.051)	(0.036)	(0.027)	(0.022)	(0.016)	(0.012)	(0.010)	(0.008)

7 Simulation Studies

7.1 Simulation Studies

In this section, we conduct simulation studies to illustrate the numerical performance of our method. We consider a data-generating progress with standard components, similar to e.g., Cai and Wei (2021):

(1) Protected attribute: The protected attribute A follows the Bernoulli distribution with parameter $1/2$.

(2) Common feature: The common features $X = (X_1, X_2)$ are two-dimensional. For $a = \{0, 1\}$, the conditional distribution of (X_1, X_2) given the protected feature $A = a$ follows the uniform distribution on $[-1, 1]^2$.

(3) Regression functions: The conditional probability of $Y = 1$ given $(X_1, X_2, A) = (x_1, x_2, a)$ is

$$\eta_a(x_1, x_2) = \frac{1 + (2a-1)s_1}{2} + \frac{s_2 \cdot \mathrm{sign}(x_1)}{2} \left(|x_1|(1 - |x_2|)\right)^\beta,$$

for all $(x_1, x_2) \in [-1, 1]^2$ and $a \in \mathcal{A}$. Here s_1, s_2 and β are hyperparameters that determine the group-wise thresholds, the margin condition, and the smoothness of the regression function. we set $s_2 > s_1 > 0$ so that $s_1 + s_2 \leqslant 1$.

It is clear that, for $a = \{0, 1\}$, $\eta_a \in [0, 1]$ is β-smooth, and it also satisfies the γ_δ-margin condition with $\gamma_\delta = 1/\beta$ when $\delta = 0$ and $\gamma_\delta = 1$ otherwise. As shown in Section G, the δ-fair Bayes-optimal classifier takes

Table 3: Estimated fairness-aware excess risk and DDP of our FairBayes-DDP+ classifier $\widehat{f}_{\delta,n}$ in the setting from Section 7.1, for various pre-specified disparity levels.

δ	0.00	0.05	0.10	0.15	0.20	0.25	0.30
$d_E(\widehat{f}_{\delta,n})$	0.002	0.002	0.002	0.003	0.003	0.004	0.005
(SD)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
DDP	0.010	0.050	0.100	0.150	0.200	0.250	0.300
(SD)	(0.008)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.014)

values

$$f_\delta^\star(x_1, x_2, a) = I\left(\eta_a(x_1, x_2) > \frac{1}{2} + (2a-1)t_\delta^\star\right)$$

for x_1, x_2, a, and this choice is unique almost surely with respect to the distribution of the data. Here $t_\delta^\star$ satisfies $(s_1 - s_2)/2 \leqslant t_\delta^\star \leqslant s_1/2$ and solves the equation

$$\left(\frac{s_1 - 2t_\delta^\star}{s_2}\right)^{\frac{1}{\beta}} \left(1 - \frac{1}{\beta}\ln\left(\frac{s_1 - 2t_\delta^\star}{s_2}\right)\right) = \delta \wedge \left[\left(\frac{s_1}{s_2}\right)^{\frac{1}{\beta}} - \frac{1}{\beta}\left(\frac{s_1}{s_2}\right)^{\frac{1}{\beta}} \ln\left(\frac{s_1}{s_2}\right)\right].$$

Moreover, with $q_\delta^\star = ((s_1 - 2t_\delta^\star)/s_2)^{1/\beta}$, the misclassification rate of $f_\delta^\star$ is given by

$$R(f_\delta^\star) = \frac{1}{2} - \frac{s_1 q_\delta^\star}{2}\left(1 - \ln\left(q_\delta^\star\right)\right) - \frac{s_2}{2(\beta+1)}\left(\frac{1 - (q_\delta^\star)^{\beta+1}}{\beta+1} + (q_\delta^\star)^{\beta+1}\ln\left(q_\delta^\star\right)\right).$$

In our experiments, we set $s_1 = 0.2$, $s_2 = 0.8$, $\beta = 1$ and generate samples of size $n_{\text{train}} = 2^j \cdot 50$, $j \in [6]$ from the source distribution. For each sample size, we estimate the regression functions by local polynomial estimators with a Gaussian kernel. Additionally, we vary the bandwidth $h_{n,a}$ from $0.5 \cdot n_a^{-1/4}$ to $5 \cdot n_a^{-1/4}$, where n_a is the sample size associated with group $A = a$, and select the bandwidth that yields the best performance on a validation set of size $n_{\text{val}} = 1000$. For estimating the thresholds, we let $\Delta_n = 0.1 \cdot (\log\log n)^{-1}$, $r_n = 0.1 \cdot (\log n)^{-1}$ and consider offsets with levels $\ell_{n,a} = \{0, 0.25, 0.5, 0.75, 1\} \cdot n_a^{-1/4}$ to evaluate the effect of offsets.

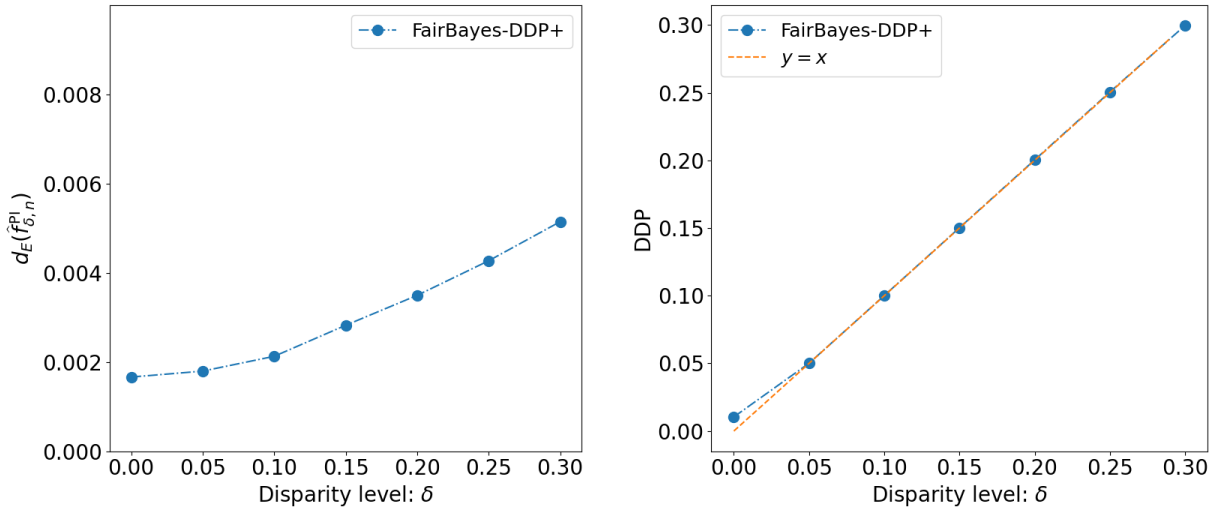
For the resulting FairBayesDDP+ classifier $\widehat{f}_{\delta,n}$, we estimate the fairness-aware excess risk $d_E(\widehat{f}_{\delta,n})$ and disparity $\text{DDP}(\widehat{f}_{\delta,n})$ on a test set with size $n_{\text{test}} = 1000$. We repeat the experiments 1000 times. The results with $\delta = 0$ are summarized in Figure 3 and Table 2. As we can see, both the fairness-aware excess risk and disparity converge to zero as the sample size increases, lending support to the asymptotic consistency and fairness of our method. For a given sample size, larger offsets lead to a a slower convergence of the fairness-aware excess risk d_E and a faster convergence of the DDP, which is consistent with our theoretical results from Theorem 6.4 and Theorem 6.7.

Next, we set $n_{\text{train}} = 12800$, $\ell_{n,a} = 0.25 \cdot n_a^{-1/4}$ and consider different pre-specified levels of δ. Again, we set $\beta = 1$, and the bandwidth is chosen on a grid from $0.5 \cdot n_a^{-1/4}$ to $5 \cdot n_a^{-1/4}$ to optimize performance on a validation set. Figure 4 and Table 3 present the fairness-aware excess risks and disparity levels of our estimator under various pre-specified levels of disparity, based on 1000 simulations. As we can see, FairBayes-DDP+ effectively controls the disparity and achieves a vanishing fairness-aware excess risk.

7.2 Empirical Data Analysis

To further support our theory and our proposed method, we conduct experiments on the benchmark "Adult" dataset (Becker and Kohavi, 1996), and compare our method with strong baseline methods.

Data Description. The "Adult" dataset is a commonly considered dataset in fair statistical learning. It contains data on a sample of individuals. The target variable y measures if the income of an individual is more than \$50,000. Age, marriage status, education level, and other related variables are included in x, and the protected attribute a refers to gender. To support our asymptotic theory from Section 6, we select three continuous features—"age", "year of education", and "working hours per week"—as predictors; these



(a) Estimated and population values of the fairness-aware excess risk.

(b) Estimated and population values of the demographic disparity.

Figure 4: Estimated and population values of the fairness-aware excess risk and DDP of our FairBayes-DDP+ classifier $\widehat{f}_{\delta,n}$ in the setting from Section 7.1, for various desired disparity levels.

features have the largest empirical marginal correlation with the label. We adopt a standard data processing approach as in e.g., Cho et al. (2020). In addition, we split the usual training set into a training part (70%) and a validation part (30%) for model selection.

Baselines. We consider several strong baselines proposed recently for fair classification: (1) Adversarial training (ADV, Zhang et al. (2018)), (2) KDE-based constrained optimization (KDE, Cho et al. (2020)), (3) Post-processing through optimal transport (PPOT, Xian et al. (2023)), and (4) Post-processing through flipping (PPF, Chen et al. (2023)).

Training details. For our Fair Bayes-DDP+ method, we estimate the regression functions over three features using local polynomial estimators. We use the Gaussian kernel and set the smoothness hyperparameter as $\beta = 3$; which influences the choices below. We select the bandwidth $h_{n,a}$ with the best performance on the validation set, ranging from $0.5\times n_a^{-1/(2\beta+d)}$ to $5\times n_a^{-1/(2\beta+d)}$. To estimate the group-wise thresholds, we let $\Delta_n = 0.1\cdot(\log\log n)^{-1}$ and $r_n = 0.1\cdot(\log n)^{-1}$. The offsets are set as $\ell_{n,a} = 0.1\cdot n_a^{-\beta/(2\beta+d)}$, for all a.

For other methods, we follow the training settings from Cho et al. (2020). A three-layer fully connected neural network with 32 hidden neurons is trained with the Adam optimizer with the default hyperparameters $(\beta_1, \beta_2) = (0.9, 0.999)$. The batch size, training epochs, and learning rate are set to be 512, 200 and 0.1, respectively. For adversarial training (Zhang et al., 2018), we further use a two-layer fully connected neural network with 16 hidden neurons as the discriminator. In all cases, we train the model on the training set and perform early stopping based on the validation set. All experiments use PyTorch; we repeat them 50 times.[8]

Simulation Results. We first evaluate the FairBayes-DDP+ algorithm with various pre-determined levels of disparity. We present the simulation results in Table 1. We observe that FairBayes-DDP+ controls the disparity level at the pre-determined values, as desired.

We then compare FairBayes-DDP+ with baseline methods in Table 4. We observe that FairBayes-DDP+, PPOT, and PPF demonstrate comparable performance in terms of both accuracy and disparity control. This similarity arises because they are all post-processing methods that aim to estimate the fair Bayes-optimal classifier and are able to control the disparity directly. In contrast, KDE and ADV are in-processing methods where the disparity is controlled implicitly by tuning hyperparameters controlling the training process. Consequently, they exhibit inferior performance in disparity control compared to the

[8]The randomness of the experiments comes from the stochasticity of the batch selection in the optimization algorithm.

Table 4: Classification accuracy and DDP on the "Adult" dataset

Methods	Parameters	ACC	DDP
FairBayes-DDP+ (Proposed)	$\delta = 0$	0.791 (0.001)	0.008 (0.003)
ADV(Zhang et al., 2018)	$\alpha = 5$	0.799 (0.004)	0.055 (0.017)
KDE(Cho et al., 2020)	$\lambda = 0.95$	0.784 (0.002)	0.039 (0.008)
PPOT (Xian et al., 2023)	$\delta = 0$	0.790 (0.001)	0.008 (0.004)
PPF (Chen et al., 2023)	$\delta = 0$	0.790 (0.001)	0.007 (0.003)

post-processing methods.

To further support FairBayes-DDP+, we compare its fairness-accuracy tradeoff with that of other baseline methods. For FairBayes-DDP+, PPOT and PPF, the level of unfairness is directly controlled, ranging from zero to the empirical DDP of the unconstrained classifier. In KDE-based constrained optimization, fairness and accuracy are balanced through a tuning parameter that controls the ratio between the loss and the fairness regularization term. We let this tuning parameter λ vary from 0.05 to 0.95 to explore a wide range of the tradeoff. In adversarial training, the tradeoff is controlled by changing the hyperparameter α that handles the gradient of the discriminator. We vary this parameter from zero to five. We empirically find that in this range, the performance is representative and suffices for comparison. More details about the effects of λ and α can be found in Cho et al. (2020) and Zhang et al. (2018), respectively.

Figure 1 presents the empirical fairness-accuracy tradeoff, where each point represents a particular tuning parameter. Our FairBayes-DDP+ algorithm demonstrates the best tradeoff, followed by PPOT and PPF. For a given disparity level, FairBayes-DDP+ achieves the highest accuracy. The KDE method performs satisfactorily in the high disparity regime; however, it may lose accuracy in the low disparity regime. This loss in accuracy could be attributed to its use of a Huber surrogate loss to handle the non-differentiability of the absolute value function at zero. Here, adversarial training does not reduce the DDP to near zero, possibly due to the instability of minimax training.

8 Summary and Discussion

In this paper, we develop minimax optimal classifiers having a bounded demographic disparity. Under appropriate smoothness and margin conditions, we show that there can be an additional term in the minimax lower bound, caused by the error in estimating the per-class thresholds. We also propose the FairBayes-DDP+ method for fair classification, prove its minimax optimality, and illustrate it in simulations and empirical data analysis. In this work, our theory rests on the low-dimensional optimality of local polynomial methods, however, empirically the plug-in method works well in higher-dimensional settings by leveraging neural nets (Zeng et al. (2024)). Formalizing this rigorously remains an intriguing direction for future work.

Acknowledgements

This work was partially supported by ARO W911NF-20-1-0080, ARO W911NF-23-1-0296, NSF 2031895, NSF DMS 2046874, ONR N00014-21-1-2843, ONR N00014-18-2759, NSF – SCALE MoDL (2134209), a JP Morgan Faculty Award and the Sloan Foundation.

References

I. Alabdulmohsin. Fair classification via unconstrained optimization. *arXiv preprint arXiv:2005.14621*, 2020.

J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica, May 2016. URL `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`.

P. Assouad. Deux remarques sur l'estimation. *Comptes rendus des séances de l'Académie des sciences. Série 1, Mathématique*, 296(23):1021–1024, 1983.

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

B. Becker and R. Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

T. T. Cai and H. Wei. Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *The Annals of Statistics*, 49(1):100–128, 2021.

T. Calders, F. Kamiran, and M. Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18. IEEE, 2009.

F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K. R. Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017.

S. Caton and C. Haas. Fairness in machine learning: A survey. *ACM Computing Surveys*, 2023.

L. E. Celis, L. Huang, V. Keswani, and N. K. Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 319–328. Association for Computing Machinery, 2019.

W. Chen, Y. Klochkov, and Y. Liu. Post-hoc bias scoring is optimal for fair classification. *arXiv preprint arXiv:2310.05725*, 2023.

J. Cho, G. Hwang, and C. Suh. A fair classifier using kernel density estimation. In *Advances in Neural Information Processing Systems*, pages 15088–15099. Curran Associates, Inc., 2020.

E. Chzhen and N. Schreuder. A minimax framework for quantifying risk-fairness trade-off in regression. *The Annals of Statistics*, 50(4):2416–2442, 2022.

E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil. Leveraging labeled and unlabeled data for consistent fair binary classification. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019.

S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806. Association for Computing Machinery, 2017.

S. Corbett-Davies, J. D. Gaebler, H. Nilforoshan, R. Shroff, and S. Goel. The measure and mismeasure of fairness. *Journal of Machine Learning Research*, 24(312):1–117, 2023.

A. Cotter, H. Jiang, M. Gupta, S. Wang, T. Narayan, S. You, and K. Sridharan. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *Journal of Machine Learning Research*, 20 (172):1–59, 2019.

L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*. Springer New York, NY, 1996.

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, page 214–226. Association for Computing Machinery, 2012.

M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. Association for Computing Machinery, 2015.

B. Fish, J. Kun, and Á. D. Lelkes. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 144–152. SIAM, 2016.

A. W. Flores, K. Bechtel, and C. T. Lowenkamp. False positives, false negatives, and false analyses: A rejoinder to Machine Bias: "There's software used across the country to predict future criminals. And it's biased against blacks". *Federal Probation*, 80(2):38–46, 2016.

K. Fukuchi and J. Sakuma. Demographic parity constrained minimax optimal regression under linear model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

G. Goh, A. Cotter, M. Gupta, and M. P. Friedlander. Satisfying real-world goals with dataset constraints. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2016.

M. Gupta and Q. Mohammad. Advances in AI and ML are reshaping healthcare. TechCrunch, Mar. 2017. URL `https://techcrunch.com/2017/03/16/advances-in-ai-and-ml-are-reshaping-healthcare/`.

M. Hardt, E. Price, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2016.

T. Hastie, R. Tibshirani, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer New York, NY, 2nd edition, 2009.

T. Jang, P. Shi, and X. Wang. Group-aware threshold adaptation for fair classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6988–6995, 2022.

J. E. Johndrow and K. Lum. An algorithm for removing sensitive information: Application to race-independent recidivism prediction. *The Annals of Applied Statistics*, 13(1):189–220, 2019.

M. Joseph, M. Kearns, J. H. Morgenstern, and A. Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2016.

P. Lahoti, K. P. Gummadi, and G. Weikum. iFair: Learning individually fair data representations for algorithmic decision making. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 1334–1345. IEEE, 2019.

J. Lei, J. Robins, and L. Wasserman. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287, 2013.

K. Lum and J. Johndrow. A statistical framework for fair predictive algorithms. *arXiv preprint arXiv:1610.08077*, 2016.

X. Ma, J. Sha, D. Wang, Y. Yu, Q. Yang, and X. Niu. Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. *Electronic Commerce Research and Applications*, 31:24–39, 2018.

E. Mammen and A. B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.

A. K. Menon and R. C. Williamson. The cost of fairness in binary classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 107–118. PMLR, 2018.

H. Narasimhan. Learning with complex loss functions and constraints. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, pages 1646–1654. PMLR, 2018.

Plato. *The Republic*. MIT, 1994. URL `http://classics.mit.edu/Plato/republic.html`.

J. Rawls. *Justice as fairness: A restatement*. Harvard University Press, 2001.

J. Rawls. *A theory of justice*. Harvard University Press, 1971.

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

A. Ruoss, M. Balunovic, M. Fischer, and M. Vechev. Learning certified individually fair representations. In *Advances in Neural Information Processing Systems*, pages 7584–7596. Curran Associates, Inc., 2020.

N. Schreuder and E. Chzhen. Classification with abstention but without disparities. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 1227–1236. PMLR, 2021.

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

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer New York, NY, 2009.

I. Valera, A. Singla, and M. Gomez Rodriguez. Enhancing the accuracy and fairness of human decision making. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2018.

D. Wei, K. N. Ramamurthy, and F. Calmon. Optimized score transformation for consistent fair classification. *Journal of Machine Learning Research*, 22(258):1–78, 2021.

R. Xian, L. Yin, and H. Zhao. Fair and optimal classification via post-processing. In *Proceedings of the 40th International Conference on Machine Learning*, pages 37977–38012. PMLR, 2023.

Y. Yang. Minimax nonparametric classification .i. rates of convergence. *IEEE Transactions on Information Theory*, 45(7):2271–2284, 1999.

M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75):1–42, 2019.

X. Zeng, E. Dobriban, and G. Cheng. Fair Bayes-optimal classifiers under predictive parity. In *Advances in Neural Information Processing Systems*, pages 27692–27705. Curran Associates, Inc., 2022.

X. Zeng, G. Cheng, and E. Dobriban. Bayes-optimal fair classification with linear disparity constraints via pre-, in-, and post-processing. *arXiv preprint arXiv: 2402.02817*, 2024.

B. H. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340. Association for Computing Machinery, 2018.

Appendix

Additional Notation and Definitions

In this appendix, we use some additional notation. For a real-valued function f defined on $[a, b)$ for some $a < b$, we denote by $\lim_{x\to a^+} f(x)$ the limit from the right of f at a, if it exists. Similarly, if f is defined on $(b, a]$ for $b < a$, we denote by $\lim_{x\to a^-} f(x)$ the limit from the left of f at a, if it exists. For an interval $[a, b]$, and scalars $c \in \mathbb{R}$, $d > 0$, we denote $c + d[a, b] = [c + da, c + db]$. For an integer $p \geqslant 1$, we let e_j, $j \in [p]$ be the j-th standard basis vector, with $e_{jj} = 1$ and $e_{jk} = 0$ for $k \neq j$. For an integer $p \geqslant 1$ and $x = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d$, denote by $B_{d,p}(x, r)$ the d-dimensional ℓ_p ball with center x and radius $r \geqslant 0$, i.e., $B_{d,p}(x, r) = \{y = (y_1, \ldots, y_d)^\top : \sum_{j=1}^d |y_j - x_j|^p \leqslant r^p\}$. Moreover, let $V_{d,p}$ be the volume of $B_{d,p}(0, 1)$. For $q > 0$ and $z = (z_1, \ldots, z_d)^\top \in [0, 1]^d$, we define $\mathcal{C}_{z,q} = \{x = (x_1, \ldots, x_d)^\top : |x_i - z_i| \leqslant 4q^{-1}, i = 1, 2, \ldots, d\}$

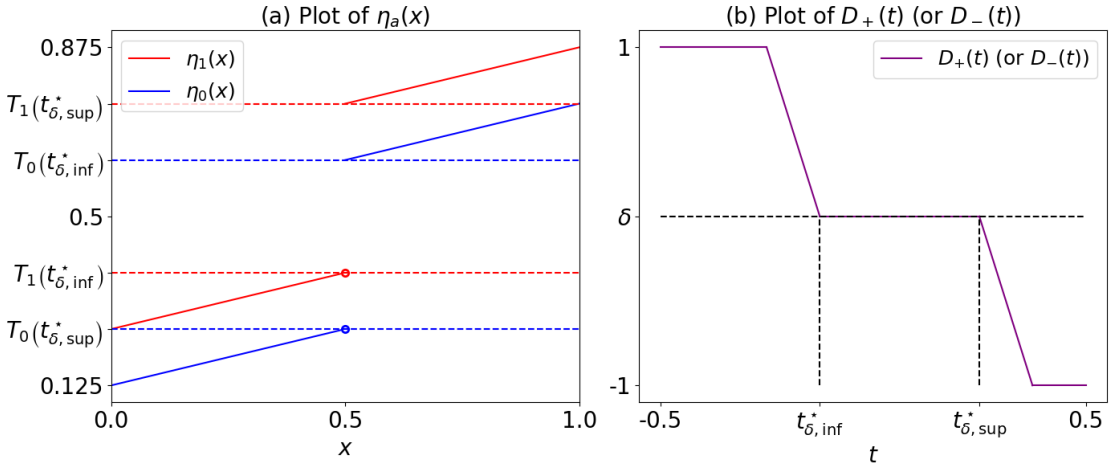


Figure 5: When both $D_+(t)$ and $D_-(t)$ are flat at δ (or $-\delta$), there exists an interval $[t^\star_{\delta,\text{inf}}, t^\star_{\delta,\text{sup}}]$ within which the conditions $D_+(t) = D_-(t) = \delta$ (or $-\delta$) always hold. In other words, the corresponding classifiers satisfy the hard constraint. Here, we set $\mathbb{P}(A = 1) = 1/2$, $X|A = a \sim U(0, 1)$ and $\delta = 0$.

as the cube of side length $8/q$ centered at z, and $\mathcal{D}_{z,q} = B_{d,2}(z, 2q^{-1}) \setminus B_{d,2}(z, q^{-1})$ as a hyperspherical shell. The interior of a set S is denoted by $\text{int}S$. For a classifier f, we also define $R(f) := \mathbb{P}\left(Y \neq \widehat{Y}_f\right)$.

Without a fairness constraint, a *Bayes-optimal classifier*, which minimizes the misclassification rate, is defined as $f^\star \in \text{argmin}_f \left[\mathbb{P}(Y \neq \widehat{Y}_f)\right]$. Bayes-optimal classifiers are the "best possible" method when fairness is not a concern. Denoting the indicator function by $I(\cdot)$, a classical result (see e.g., Devroye et al., 1996, etc) is that all Bayes-optimal classifiers $f^\star : \mathcal{X} \times \{0, 1\} \to [0, 1]$ have the form

$$f^\star(x, a) = I\left(\eta_a(x) > 1/2\right) + \tau_a I\left(\eta_a(x) = 1/2\right), \tag{.1}$$

for all $(x, a) \in \mathcal{X} \times \{0, 1\}$, where and $\tau_0, \tau_1 \in [0, 1]$ are any two constants.

A Fair Bayes-optimal Classifier with a Nonzero Disparity

Here we give the general form of Bayes-optimal classifiers for the case where the group-wise decision thresholds are not unique, following Zeng et al. (2024). For any $\delta > 0$, define the following quantities, which can be viewed as "inverses" of the functions D_-, D_+ in the various cases:

$$t^\star_{\delta,\text{inf}} = \begin{cases} \inf\{t : D_-(t) \leqslant \delta\} = \inf\{t : D_+(t) \leqslant \delta\}, & \text{in the fairness-impacted case } D_-(0) > \delta; \\ 0, & \text{otherwise,} \end{cases} \tag{A.1}$$

$$t^\star_{\delta,\text{sup}} = \begin{cases} \inf\{t : D_-(t) < \delta\} = \inf\{t : D_+(t) < \delta\}, & \text{in the fairness-impacted case } D_-(0) > \delta; \\ 0, & \text{otherwise.} \end{cases} \tag{A.2}$$

We need both $t^\star_{\delta,\text{inf}}, t^\star_{\delta,\text{sup}}$ to account for the case where D_-, D_+ are "flat" at δ or $-\delta$, so that they can take any value over a nonempty interval, see Figure 5 for an illustration.

Proposition A.1 (Fair Bayes-optimal classifiers)**.** *For any $\delta \geqslant 0$, all δ-fair Bayes-optimal classifiers $f^\star_\delta$ have the following form: for any $T_{\delta,1} \in [1/2 + t^\star_{\delta,\text{inf}}/(2p_1), 1/2 + t^\star_{\delta,\text{sup}}/(2p_1)]$ and $T_{\delta,0} \in [1/2 - t^\star_{\delta,\text{sup}}/(2p_0), 1/2 - t^\star_{\delta,\text{inf}}/(2p_0)]$, there are $(\tau_{\delta,1}, \tau_{\delta,0}) \in [0, 1]^2$, such that for all x, a,*

$$f^\star_\delta(x, a) = I\left(\eta_a(x) > T_{\delta,a}\right) + \tau_{\delta,a} I\left(\eta_a(x) = T_{\delta,a}\right). \tag{A.3}$$

Further, $(T_{\delta,1}, T_{\delta,0})$ and $(\tau_{\delta,1}, \tau_{\delta,0})$ are determined by the following constraints:

(1). When $D_-(0) \leqslant \delta$,

$$\left|\mathbb{P}_{X|A=1}\left(\widehat{Y}_{f^\star_\delta} = 1\right) - \mathbb{P}_{X|A=0}\left(\widehat{Y}_{f^\star_\delta} = 1\right)\right| \leqslant \delta. \tag{A.4}$$

(2). When $D_-(0) > \delta$,

$$\mathbb{P}_{X|A=1}\left(\widehat{Y}_{f_\delta^\star}=1\right)-\mathbb{P}_{X|A=0}\left(\widehat{Y}_{f_\delta^\star}=1\right)=\delta. \tag{A.5}$$

The form of $(\tau_{\delta,1},\tau_{\delta,0})$ *is provided below. For scalars* a,b,c,d *with* $b\leqslant c$ *and* $d>0$, *let* $\rho((a+[b,c])/d)$ *represent* $[\rho((a+b/d)),\rho((a+c)/d)]$ *with* ρ *from* (5.8), *let* $\widetilde{\delta}$ *and* $\widetilde{\delta}$ *be defined in* (5.9), *and let, for* $a\in\{0,1\}$, $\pi^\star_{a,+}$ *and* $\pi^\star_{a,=}$ *be defined in* (5.10). *We have the following four cases:*

- *Case (1). When* $\pi^\star_{1,=}=\pi^\star_{0,=}=0$, $(\tau_{\delta,1},\tau_{\delta,0})\in[0,1]^2$ *can be arbitrary.*
- *Case (2). When* $\pi^\star_{1,=}>\pi^\star_{0,=}=0$, $\tau_{\delta,0}\in[0,1]$ *can be arbitrary, and we can take*

$$\tau_{\delta,1}\in\rho\left(\frac{\pi^\star_{0,+}-\pi^\star_{1,+}+\left[\widetilde{\delta},\widetilde{\delta}\right]}{\pi^\star_{1,=}}\right).$$

- *Case (3). Similarly, when* $\pi^\star_{0,=}>\pi^\star_{1,=}=0$, $\tau_{\delta,1}\in[0,1]$ *can be arbitrary, and we can take*

$$\tau_{\delta,0}\in\rho\left(\frac{\pi^\star_{1,+}-\pi^\star_{0,+}+\left[-\widetilde{\delta},-\widetilde{\delta}\right]}{\mathbb{P}_{X|A=0}\left(\eta_0(X)=T_{\delta,0}\right)}\right).$$

- *Case (4). Finally, when* $\pi^\star_{1,+}>0$ *and* $\pi^\star_{0,+}>0$, *we can take*

$$\tau_{\delta,1}\in\rho\left(\frac{\pi^\star_{0,+}-\pi^\star_{1,+}+\left[\widetilde{\delta},\widetilde{\delta}+\pi^\star_{0,=}\right]}{\pi^\star_{1,=}}\right),\quad\text{and}\quad\tau_{\delta,0}\in\rho\left(\frac{\pi^\star_{1,+}+\tau_{\delta,1}\pi^\star_{1,=}-\pi^\star_{0,+}+\left[-\widetilde{\delta},-\widetilde{\delta}\right]}{\pi^\star_{0,=}}\right).$$

Remark 3. *When* $\eta_1(X)$ *and* $\eta_0(X)$ *have density functions on* $[0,1]$, *we have for* $a\in\{0,1\}$, $\mathbb{P}_{X|A=a}(\eta_a(X)=1/2+(2a-1)t^\star_\delta/(2p_a))=0$ *and the optimal classifier is deterministic. With* $t^\star_\delta\in[t^\star_{\delta,\inf},t^\star_{\delta,\sup}]$, *and for all* x,a, *it takes values*

$$f^\star_\delta(x,a)=I\left(\eta_a(x)>\frac{1}{2}+\frac{(2a-1)t^\star_\delta}{2p_a}\right). \tag{A.6}$$

Proofs

In Sections B to E, we present the proofs of our theoretical results from the main text. We first introduce several technical lemmas that are essential for proving our theoretical results (Section B), and defer their proofs to Section F.

B Additional Lemmas

Lemma B.1. *For any* $z=(z_1,\ldots,z_d)^\top\in\mathbb{R}^d$, $R>0$ *and* $x=(x_1,\ldots,x_d)^\top\in B_{d,1}(z,R)$, *we have, for* $0\leqslant r\leqslant 2R/(d+2)$,

$$\lambda[B_{d,1}(z,R)\cap B_{d,2}(x,r)]\geqslant\frac{V_{d,1}}{V_{d,2}2^d}\cdot\lambda[B_{d,2}(x,r)].$$

Lemma B.2. *For any* $z=(z_1,\ldots,z_d)^\top\in\mathbb{R}^d$, $R>0$ *and* $x=(x_1,\ldots,x_d)^\top\in B_{d,2}(z,R)$, *we have, for* $0\leqslant r\leqslant R$,

$$\lambda[B_{d,2}(z,R)\cap B_{d,2}(x,r)]\geqslant\frac{3^{\frac{d+1}{2}}V_{d-1,2}}{2^{d+1}(d+1)V_{d,2}}\cdot\lambda[B_{d,2}(x,r)].$$

Lemma B.3. *Let* z *be a point such that* $\mathcal{C}_{z,q}\subset[0,1]^d$. *Then, for any* $r>0$, *there is* $0<C_r\leqslant 1$ *also depending on* d, *such that, for any* $x=(x_1,\ldots,x_d)^\top\in[0,1]^d\setminus\mathcal{D}_{z,q}$,

$$\lambda[B_{d,2}(x,r)\cap(\mathcal{C}_{z,q}\setminus\mathcal{D}_{z,q})]\geqslant C_r\lambda[B_{d,2}(x,r)\cap\mathcal{C}_{z,q}].$$

Lemma B.4. *For any classifier $f : \mathcal{X} \times \mathcal{A} \to [0,1]$, we have*

$$R(f) = \sum_{a\in\{0,1\}} p_a \int \left[(1-2\eta_a(x))\, f(x,a) + \eta_a(x)\right] d\mathbb{P}_{X|A=a}(x),$$

and

$$\mathrm{DDP}(f) = \sum_{a\in\{0,1\}} \int (2a-1) f(x,a) d\mathbb{P}_{X|A=a}(x). \tag{B.1}$$

Lemma B.5. *Let $\mathcal{S}_n = \mathcal{S}_{n,1} \cup \mathcal{S}_{n,0}$ with $\mathcal{S}_n = \{(x_i, a_i, y_i)\}_{i=1}^n$ and, for $a \in \{0,1\}$, $S_{n,a} = \{(x_{a,j}, a, y_{a,j})\}_{j=1}^{n_a}$ being an i.i.d. sample. We have, for $a \in \{0,1\}$,*

$$\mathbb{P}^{\otimes n}\left(\left|\frac{n_a}{n} - p_a\right| \geqslant \varepsilon\right) \leqslant 2\exp\left(-2n\varepsilon^2\right). \tag{B.2}$$

Moreover, if $p_1 \wedge p_0 > 0$, we have, for $a \in \{0,1\}$ and $\varepsilon \leqslant 1/p_a$,

$$\mathbb{P}^{\otimes n}\left(\left|\frac{n}{n_a} - \frac{1}{p_a}\right| \geqslant \varepsilon\right) \leqslant 2\exp\left(-\frac{np_a^4\varepsilon^2}{2}\right). \tag{B.3}$$

Lemma B.6. *For $a \in \{0,1\}$, we have that if $\varepsilon \leqslant \sqrt{p_a/2}$,*

$$\mathbb{P}^{\otimes n}\left(\sup_{T\in\mathbb{R}}\left|\frac{1}{n_a}\sum_{j=1}^{n_a} I\left(\eta_a\left(x_{a,j}\right) > T\right) - \mathbb{P}_{X|A=a}\left(\eta_a(X) > T\right)\right| > \varepsilon\right) \leqslant 4\exp\left(-np_a\varepsilon^2\right).$$

Lemma B.7. *Let D_- and D_+ be defined as in (3.3) and (3.4), respectively. For $t \in \mathbb{R}$, we denote*

$$\begin{aligned} D_{n,-}(t) &= \tfrac{1}{n_1}\textstyle\sum_{j=1}^{n_1} I\left(\eta_1\left(x_{1,j}\right) > \frac{1}{2} + \frac{t}{2p_1}\right) - \frac{1}{n_0}\sum_{j=1}^{n_0} I\left(\eta_0\left(x_{0,j}\right) \geqslant \frac{1}{2} - \frac{t}{2p_0}\right), \\ D_{n,+}(t) &= \tfrac{1}{n_1}\textstyle\sum_{j=1}^{n_1} I\left(\eta_1\left(x_{1,j}\right) \geqslant \frac{1}{2} + \frac{t}{2p_1}\right) - \frac{1}{n_0}\sum_{j=1}^{n_0} I\left(\eta_0\left(x_{0,j}\right) > \frac{1}{2} - \frac{t}{2p_0}\right). \end{aligned}$$

Then, for $\varepsilon \leqslant \sqrt{(p_1 \wedge p_0)/2}$,

$$\max\left\{\mathbb{P}^{\otimes n}\left(|D_{n,+}(t) - D_+(t)| > \varepsilon\right), \mathbb{P}^{\otimes n}\left(|D_{n,-}(t) - D_-(t)| > \varepsilon\right)\right\} \leqslant 8\exp\left(-\frac{n(p_1\wedge p_0)\varepsilon^2}{4}\right). \tag{B.4}$$

Lemma B.8. *For a set $\{\iota_0, \iota_1, ..., \iota_K\} \subset \mathbb{R}$ and $K > 0$, and for $\iota \in \{\iota_0, \iota_1, ..., \iota_K\}$, let $\psi_{n,1,\iota}$ and $\psi_{n,2,\iota}$ be defined for all $\varepsilon > 0$ as in (6.2), with $c_{i,\iota} > 0, i \in [4]$:*

$$\psi_{n,1,\iota}(\varepsilon) = c_{1,\iota}\exp\left(-c_{2,\iota}\left(\varepsilon/[\phi_{n,1} \vee \phi_{n,0}]\right)^2\right) \text{ and } \psi_{n,2,\iota}(\varepsilon) = c_{3,\iota}\exp\left(-c_{4,\iota}n\varepsilon^2\right).$$

Under the margin condition (4.4), for any $C_k > 0, k \in [K]$, there exists $U_\varepsilon > 0$ such that, for $j \in \{-,+\}$ and $\varepsilon < U_\varepsilon$, with $c_{1,\iota_0} = \sum_{k=1}^K c_{1,\iota_k}$, $c_{2,\iota_0} = \min_{k\in[K]}\left(c_{2,\iota_k}C_k^2\right)$, $c_{3,\iota_0} = \sum_{k=1}^K c_{3,\iota_k}$ and $c_{4,\iota_0} = \left(\min_{k\in[K]} c_{4,\iota_k}\cdot U_\gamma^{-2}C_k^\gamma\right) \wedge 1$, we have for all $\varepsilon > 0$ that

$$\sum_{k=1}^K \psi_{n,1,\iota_k}(C_k\varepsilon) \leqslant \psi_{n,1,\iota_0}(\varepsilon), \tag{B.5}$$

and

$$I\left(\delta = D_j(t_\delta^\star)\right)\left(\sum_{k=1}^K \psi_{n,2,\iota_k}\left(g_{\delta,j}(\omega(C_k\varepsilon, r_n))\right)\right) \leqslant I\left(\delta = D_j(t_\delta^\star)\right)\psi_{n,2,\iota_0}\left(g_{\delta,j}(\omega(\varepsilon, r_n))\right). \tag{B.6}$$

Lemma B.9. *Under the conditions of Theorem 6.2, there are constants L_{t_1}, U_{t_1} and c_{i,t_1}, $i \in [4]$ determining functions ψ_{n,i,t_1} in (6.2), such that, for $\Delta_n \geqslant 0$ and $L_{t_1}(\phi_{n,1} \vee \phi_{n,0}) < \varepsilon < U_{t_1}$, with $t_\delta^\star$ from (3.5),*

$$\mathbb{P}^{\otimes n}\left(\widehat{D}_n\left(t_\delta^\star + \varepsilon, 0, 0\right) \geqslant \delta + \Delta_n\right) \leqslant \psi_{n,1,t_1}(\varepsilon) + I\left(\delta = D_-(t_\delta^\star)\right)\psi_{n,2,t_1}\left(\Delta_n + g_{\delta,-}\left(\frac{\varepsilon}{2}\right)\right), \tag{B.7}$$

and

$$\mathbb{P}^{\otimes n}\left(\widehat{D}_n\left(t_\delta^\star - \varepsilon, 0, 0\right) \leqslant \delta - \Delta_n\right) \leqslant \psi_{n,1,t_1}(\varepsilon) + I\left(\delta = D_+(t_\delta^\star)\right)\psi_{n,2,t_1}\left(\Delta_n + g_{\delta,+}\left(\frac{\varepsilon}{2}\right)\right). \tag{B.8}$$

Lemma B.10. *Under the conditions of Theorem 6.2, there are constants L_{t_2}, U_{t_2} and c_{i,t_2}, $i \in [4]$ determining functions ψ_{n,i,t_2} in (6.2), such that, for $L_{t_2}(\phi_{n,1} \vee \phi_{n,0}) < \varepsilon < U_{t_2}$,*

(1) if $D_-(t^\star_\delta) = \delta$, for $g_{\delta,-}(2\varepsilon) < \Delta_n \leqslant g_{\delta,-}(2\varepsilon) + \sqrt{(p_1 \wedge p_0)/2}$,

$$\mathbb{P}^{\otimes n}\left(\widehat{D}_n\left(t^\star_\delta + \varepsilon, 0, 0\right) \leqslant \delta - \Delta_n\right) \leqslant \psi_{n,1,t_2}(\varepsilon) + \psi_{n,2,t_2}\left(\Delta_n - g_{\delta,-}\left(2\varepsilon\right)\right); \tag{B.9}$$

(2) if $D_-(t^\star_\delta) < \delta$, for $\Delta_n < (\delta - D_-(t^\star_\delta))/2$,

$$\mathbb{P}^{\otimes n}\left(\widehat{D}_n\left(t^\star_\delta + \varepsilon, 0, 0\right) \geqslant \delta - \Delta_n\right) \leqslant \psi_{n,1,t_2}(\varepsilon); \tag{B.10}$$

(3) if $D_+(t^\star_\delta) = \delta$, for $g_{\delta,+}(2\varepsilon) < \Delta_n \leqslant g_{\delta,+}(2\varepsilon) + \sqrt{(p_1 \wedge p_0)/2}$,

$$\mathbb{P}^{\otimes n}\left(\widehat{D}_n\left(t^\star_\delta - \varepsilon, 0, 0\right) \geqslant \delta + \Delta_n\right) \leqslant \psi_{n,1,t_2}(\varepsilon) + \psi_{n,2,t_2}\left(\Delta_n - g_{\delta,+}\left(2\varepsilon\right)\right); \tag{B.11}$$

(4) if $D_+(t^\star_\delta) < \delta$, for $\Delta_n < (D_+(t^\star_\delta) - \delta)/2$,

$$\mathbb{P}^{\otimes n}\left(\widehat{D}_n\left(t^\star_\delta - \varepsilon, 0, 0\right) \leqslant \delta + \Delta_n\right) \leqslant \psi_{n,1,t_2}(\varepsilon). \tag{B.12}$$

Lemma B.11. *Under the conditions of Theorem 6.2, there are constants L_r, U_r, $U_{\Delta,r}$ and $c_{i,r}$, $i \in [4]$ determining functions $\psi_{n,i,r}$ in (6.2) , such that, for $L_r(\phi_{n,1} \vee \phi_{n,0}) < r_n < U_r$, and $2(g_{\delta,-}(4r_n) \vee g_{\delta,+}(4r_n)) < \Delta_n < U_{\Delta,r}$,*

(1) if $t^\star_\delta = 0$,

$$\mathbb{P}^{\otimes n}\left(\widehat{t}_{\delta,\mathrm{mid}} - \widehat{t}_{\delta,\Delta_n,\min} > r_n\right) \leqslant \psi_{n,1,r}(r_n) + I(\delta = D_-(t^\star_\delta))\psi_{n,2,r}\left(g_{\delta,-}\left(\frac{r_n}{2}\right)\right); \tag{B.13}$$

(2) if $t^\star_\delta > 0$ and $D_+(t^\star_\delta) > \delta$,

$$\mathbb{P}^{\otimes n}\left(\widehat{t}_{\delta,\mathrm{mid}} - \widehat{t}_{\delta,\Delta_n,\min} > r_n\right) \leqslant \psi_{n,1,r}\left(\frac{r_n}{2}\right) + I(\delta = D_-(t^\star_\delta))\psi_{n,2,r}\left(g_{\delta,-}\left(\frac{r_n}{4}\right)\right); \tag{B.14}$$

(3) if $t^\star_\delta > 0$ and $D_+(t^\star_\delta) = \delta > D_-(t^\star_\delta)$,

$$\mathbb{P}^{\otimes n}\left(\widehat{t}_{\delta,\mathrm{mid}} - \widehat{t}_{\delta,\Delta_n,\min} \leqslant r_n\right) \leqslant \psi_{n,1,r}(r_n) + \psi_{n,2,r}\left(g_{\delta,+}\left(\frac{r_n}{2}\right)\right), \tag{B.15}$$

and

$$\mathbb{P}^{\otimes n}\left(\widehat{t}_{\delta,\Delta_n,\max} - \widehat{t}_{\delta,\mathrm{mid}} > r_n\right) \leqslant \psi_{n,1,r}\left(\frac{r_n}{2}\right) + \psi_{n,2,r}\left(g_{\delta,+}\left(\frac{r_n}{4}\right)\right); \tag{B.16}$$

(4) if $t^\star_\delta > 0$ and $D_-(t^\star_\delta) = D_+(t^\star_\delta) = \delta$,

$$\mathbb{P}^{\otimes n}\left(\widehat{t}_{\delta,\mathrm{mid}} - \widehat{t}_{\delta,\Delta_n,\min} \leqslant r_n\right) \leqslant \psi_{n,1,r}(r_n) + \psi_{n,2,r}\left(g_{\delta,+}\left(\frac{r_n}{2}\right)\right), \tag{B.17}$$

and

$$\mathbb{P}^{\otimes n}\left(\widehat{t}_{\delta,\mathrm{mid}} - \widehat{t}_{\delta,\Delta_n,\min} \leqslant r_n\right) \leqslant \psi_{n,1,r}(r_n) + \psi_{n,2,r}\left(g_{\delta,-}\left(\frac{r_n}{2}\right)\right). \tag{B.18}$$

Lemma B.12. *Under the conditions of Theorem 6.2, let η_1 and η_0 satisfy the γ-exponent condition in the upper bound from Definition 4.4 at level $T^\star_1$ with respect to $\mathbb{P}_{X|A=1}$ and at level $T^\star_0$ with respect to $\mathbb{P}_{X|A=0}$, respectively. Then, for $a \in \{0,1\}$, the plug-in estimator with offset $\ell_{n,1}, \ell_{n,0} > 0$, $r_n \asymp (\log\log n)^{-1}$ and $\Delta_n \asymp (\log n)^{-1}$ satisfies, for some positive constant C,*

$$\begin{aligned}\mathbb{E}^{\otimes n}\int I\{\eta_a(x) > T^\star_{\delta,a},\, \widehat{\eta}_a(x) \leqslant \widehat{T}_{\delta,a} + \ell_{n,a}\}\left|\eta_a(x) - T^\star_{\delta,a}\right| d\mathbb{P}_{X|A=a}(x) \\ \leqslant C\left(\left(\phi_{n,1} \vee \phi_{n,0} \vee \ell_{n,a}\right)^{\gamma+1} + I\left(0 < D_-\left(t^\star_\delta\right) = \delta = D_+\left(t^\star_\delta\right)\right) \cdot n^{-\frac{\gamma+1}{2\gamma}}\right).\end{aligned}$$

An analogous bound holds for $\mathbb{E}^{\otimes n}\int I\{\eta_a(x) < T^\star_{\delta,a},\, \widehat{\eta}_a(x) \geqslant \widehat{T}_{\delta,a} - \ell_{n,a}\}\left|\eta_a(x) - T^\star_{\delta,a}\right| d\mathbb{P}_{X|A=a}(x)$.

Lemma B.13. *Under the condition of Theorem 6.2, there are constants U_δ, L_δ, $U_{\Delta,\delta}$ $c_{i,\delta}$, $i \in [4]$ determining functions $\psi_{n,i,\delta}$ in (6.2) such that, for $L_\delta(\phi_{n,1} \vee \phi_{n,0}) < r_n < U_\delta$ and $0 < \Delta_n < U_{\Delta,\delta}$, with $\widetilde{\delta}$ from (5.9) and $\widehat{\delta}$ from (5.7),*

$$\mathbb{P}^{\otimes n}\left(\widehat{\delta} \neq \widetilde{\delta}\right) \leqslant \psi_{n,1,\delta}(r_n) + \psi_{n,2,\delta}(\Delta_n). \tag{B.19}$$

In the following lemmas, we denote, for $a \in \{0,1\}$,

$$\widehat{\pi}_{a,+} = \mathbb{P}_{X|A=a}\left(\widehat{\eta}_a(X) > \widehat{T}_{\delta,a} + \ell_{n,a}\right); \qquad \widehat{\pi}_{a,=} = \mathbb{P}_{X|A=a}\left(|\widehat{\eta}_a(X) - \widehat{T}_{\delta,a}| \leqslant \ell_{n,a}\right). \tag{B.20}$$

Lemma B.14. *There is L_{π_1}, U_{π_1}, U_{Δ,π_1} and $c_{i,\pi}$, $i \in [4]$ determining functions ψ_{n,i,π_1}, $i \in \{1,2\}$ in (6.2), such that, for $L_{\pi_1}(\phi_{n,1} \vee \phi_{n,0}) < r_n < U_{\pi_1}$, $2\left(g_{\delta,-}(4r_n) \vee g_{\delta,+}(4r_n)\right) < \Delta_n < U_{\Delta,\pi_1}$, $\varepsilon > 4U_\gamma(4(p_1\ell_{n,1} \vee p_0\ell_{n,0}))^\gamma$ and $2L_{\pi_1}(\phi_{n,1} \vee \phi_{n,0}) < \ell_{n,1}, \ell_{n,0} < 2r_n$, with $\omega(\varepsilon, r_n)$ from (6.4), we have*

$$\begin{aligned}
&\max\left\{\mathbb{P}^{\otimes n}(\widehat{\pi}_{a,+} > \pi^\star_{a,+} + \varepsilon), \mathbb{P}^{\otimes n}(\widehat{\pi}_{a,+} < \pi^\star_{a,+} - \varepsilon), \mathbb{P}^{\otimes n}(\widehat{\pi}_{a,=} > \pi^\star_{a,=} + \varepsilon), \mathbb{P}^{\otimes n}(\widehat{\pi}_{a,=} < \pi^\star_{a,=} - \varepsilon)\right\} \\
&\quad \leqslant \psi_{n,1,\pi}\left(\ell_{n,a}\right) + \sum_{j=\{-,+\}} I(\delta = D_j(t^\star_\delta))\psi_{n,2,\pi}\left(g_{\delta,j}(\omega(\ell_{n,a}, r_n))\right).
\end{aligned} \tag{B.21}$$

Lemma B.15. *With the same L_{π_1}, U_{π_1}, U_{Δ,π_1} and $c_{i,\pi}$, $i \in [4]$ as in Lemma B.14, we have, for $L_T(\phi_{n,1} \vee \phi_{n,0}) < r_n < U_T$, $2\left(g_{\delta,-}(4r_n) \vee g_{\delta,+}(4r_n)\right) < \Delta_n < U_\Delta$, $4U_\gamma(4(p_1\ell_{n,1} \vee p_0\ell_{n,0}))^\gamma < \varepsilon \leqslant \sqrt{(p_1 \wedge p_0)/2}$ and $2(L_\eta \vee L_T)(\phi_{n,1} \vee \phi_{n,0}) < \ell_{n,1}, \ell_{n,0} < 2r_n$, with $\omega(\varepsilon, r_n)$ from (6.4), $\widehat{\pi}_{n,a,+}, \widehat{\pi}_{n,a,=}$ from (5.6) and $\pi^\star_{a,+}, \pi^\star_{a,=}$ from (5.10),*

$$\begin{aligned}
&\max\left\{\mathbb{P}^{\otimes n}(\widehat{\pi}_{n,a,+} > \pi^\star_{a,+} + \varepsilon), \mathbb{P}^{\otimes n}(\widehat{\pi}_{n,a,+} < \pi^\star_{a,+} - \varepsilon), \mathbb{P}^{\otimes n}(\widehat{\pi}_{n,a,=} > \pi^\star_{a,=} + \varepsilon), \mathbb{P}^{\otimes n}(\widehat{\pi}_{n,a,=} < \pi^\star_{a,=} - \varepsilon)\right\} \\
&\quad \leqslant \psi_{n,1,\pi}\left(\ell_{n,a}\right) + \sum_{j=\{-,+\}} I(\delta = D_j(t^\star_\delta))\psi_{n,2,\pi}\left(g_{\delta,j}(\omega(\ell_{n,a}, r_n))\right) + 4\exp\left(-\frac{np_a\varepsilon^2}{4}\right).
\end{aligned} \tag{B.22}$$

Lemma B.16. *Let $b > 0$. For $0 < \varepsilon < b/2$, we have, with ρ from (5.8),*

$$\rho\left(\frac{a+2\varepsilon}{b-\varepsilon}\right) - \rho\left(\frac{a}{b}\right) \leqslant \frac{6\varepsilon}{b} \quad \text{and} \quad \rho\left(\frac{a-2\varepsilon}{b+\varepsilon}\right) - \rho\left(\frac{a}{b}\right) \geqslant \frac{-6\varepsilon}{b}. \tag{B.23}$$

Lemma B.17. *There exist constants L_π, U_π, $L_{\varepsilon,\pi}$, $U_{\Delta,\pi}$, $c_{5,\pi}$, $c_{6,\pi}$ and with the same $c_{i,\pi}$, $i \in [4]$ as in Lemma B.14 such that, for $a \in \{0,1\}$, $L_\pi(\phi_{n,1} \vee \phi_{n,0}) < r_n < U_\pi$, $2\left(g_{\delta,-}(4r_n) \vee g_{\delta,+}(4r_n)\right) < \Delta_n < U_{\Delta,\pi}$, $L_{\varepsilon,\pi}(\ell_{n,1} \vee \ell_{n,0})^\gamma < \varepsilon \leqslant \sqrt{(p_1 \wedge p_0)/2}$ and $(L_\eta \vee L_T)(\phi_{n,1} \vee \phi_{n,0}) < \ell_{n,a}/2 < r_n$, with $\widehat{\pi}_{a,+}, \widehat{\pi}_{a,=}$ from (B.20), $\widehat{\tau}_{\delta,a}$ from (5.5), $\pi^\star_{a,+}, \pi^\star_{a,=}$ from (5.10) and $\tau^\star_{\delta,a}$ from (5.11),*

$$\begin{aligned}
&\max\left\{\mathbb{P}^{\otimes n}\left(\widehat{\pi}_{a,=}\widehat{\tau}_{\delta,a} > \pi^\star_{a,=}\tau^\star_{\delta,a} + \varepsilon\right), \mathbb{P}^{\otimes n}\left(\widehat{\pi}_{a,=}\widehat{\tau}_{\delta,a} < \pi^\star_{a,=}\tau^\star_{\delta,a} - \varepsilon\right)\right\} \\
&\qquad \leqslant 4\psi_{n,1,\pi}\left(\ell_{n,a}\right) + 4\sum_{j=\{-,+\}} I(\delta = D_j(t^\star_\delta))\psi_{n,2,\pi}\left(g_{\delta,j}(\omega(\ell_{n,a}, r_n))\right) + c_{5,\pi}\exp\left(-c_{6,\pi}n\varepsilon^2\right).
\end{aligned} \tag{B.24}$$

The following proposition from Audibert and Tsybakov (2007) demonstrates the point-wise convergence of the local polynomial estimator.

Proposition B.18. *Let $\mathcal{P}$ be a class of probability distributions for (X, Y), such that the regression function $\eta(x) = \mathbb{P}(Y = 1 \mid X = x)$ belongs to the Hölder class $\Sigma(\beta, L_\beta, \mathbb{R}^d)$ and the marginal law of X satisfies the strong density condition. Let $(X_i, Y_i)_{i=1}^n$ an i.i.d. sample from $\mathbb{P}$, and $\widehat{\eta}$ be the local polynomial estimator with kernel K satisfying (5.12) and $h_n \asymp n^{-1/(2\beta+d)}$. Then there exist constants $C_1, C_2 > 0$ such that for any $\delta > 0$, $n \geqslant 1$ we have*

$$\sup_{P \in \mathcal{P}} \mathbb{P}^{\otimes n}\left(|\widehat{\eta}(x) - \eta(x)| > \varepsilon\right) \leqslant C_1\exp\left(-C_2 n^{\frac{2\beta}{2\beta+d}}\varepsilon^2\right),$$

for almost all x with respect to $\mathbb{P}_X$. The constants C_1, C_2 depend only on β, d, L, c_0, r_0, $\mu_{\min}$, $\mu_{\max}$, and on the kernel K.

Proposition B.18 shows that the local polynomial estimators $(\widehat{\eta}_1, \widehat{\eta}_0)$ are $(\phi_{n,1}, \phi_{n,0})_{n \geqslant 1}$-pointwise convergent with $\phi_{n,1} = \phi_{n,0} = n^{-\beta/(2\beta+d)}$.

C Proofs of Results in Section 4

C.1 Proof of Proposition 4.2

By (3.5) and (A.5), we have $t^\star_\delta = 0$ when $\delta \geqslant D_-(0)$ and

$$t^\star_\delta \left[\mathbb{P}_{X|A=1}\left(\widehat{Y}_{f^\star_\delta} = 1\right) - \mathbb{P}_{X|A=0}\left(\widehat{Y}_{f^\star_\delta} = 1\right)\right] = \begin{cases} 0, & D_-(0) \leqslant \delta; \\ t^\star_\delta \delta, & D_-(0) > \delta. \end{cases}$$

Using Lemma B.4, it follows that

$$\begin{aligned}
d_E\left(f, f^\star_\delta\right) &= 2\sum_{a\in\mathcal{A}} p_a \int_{\mathcal{X}} \left(\frac{1}{2} + \frac{(2a-1)t^\star_\delta}{2p_a} - \eta_a(x)\right)\left(f(x,a) - f^\star_\delta(x,a)\right) d\mathbb{P}_{X|A=a}(x) \\
&= d_R(f, f^\star_\delta) + 2\sum_{a\in\mathcal{A}} p_a \left(\frac{(2a-1)t^\star_\delta}{2p_a}\right)\left[\int \left(f(x,a) - f^\star_\delta(x,a)\right) d\mathbb{P}_{X|A=a}(x)\right] \\
&= d_R(f, f^\star_\delta) + \sum_{a\in\mathcal{A}} (2a-1)t^\star_\delta \left[\mathbb{P}_{X|A=a}\left(\widehat{Y}_f = 1\right) - \mathbb{P}_{X|A=a}\left(\widehat{Y}_{f^\star_\delta} = 1\right)\right] \\
&= d_R(f, f^\star_\delta) + t^\star_\delta \left[\mathbb{P}_{X|A=1}\left(\widehat{Y}_f = 1\right) - \mathbb{P}_{X|A=1}\left(\widehat{Y}_{f^\star_\delta} = 1\right) - \mathbb{P}_{X|A=0}\left(\widehat{Y}_f = 1\right) + \mathbb{P}_{X|A=0}\left(\widehat{Y}_{f^\star_\delta} = 1\right)\right] \\
&= \begin{cases} d_R(f, f^\star_\delta), & D_-(0) \leqslant \delta; \\ d_R(f, f^\star_\delta) + t^\star_\delta\left[\mathrm{DDP}(f) - \delta\right], & D_-(0) > \delta. \end{cases}
\end{aligned}$$

Now note that $t^\star_\delta > 0$ when $D_-(0) > \delta$ and $t^\star_\delta < 0$ when $\delta < -D_+(0)$. This implies that if $|\mathrm{DDP}(f)| \leqslant \delta$, then $d_R\left(f, f^\star_\delta\right) \geqslant d_E(f, f^\star_\delta)$.

C.2 Proof of Theorem 4.7

In the automatically fair and fair-boundary cases when $\delta \geqslant \max\{|D_-(0)|, |D_+(0)|\}$, all unconstrained Bayes-optimal classifiers are δ-fair Bayes-optimal classifiers. In this scenario, the fair classification problem is simply a standard unconstrained classification problem, and the minimax lower bound is the same as the lower bound (4.6) from Audibert and Tsybakov (2007).

Next, we consider the fairness-impacted case. In what follows, we assume $\delta = 0$ and write $\gamma := \gamma_0$ without loss of generality. We will generally omit mentioning δ further in this proof. In addition to the usual lower bound for classification problems, in the fairness-impacted case, the minimax lower bound may contain a second term due to the estimation of thresholds. Accordingly, the proof of the theorem also contains two parts.

In the first part, we start from the strategy of Audibert and Tsybakov (2007); with some modifications, either in order to streamline the proof, or as required by the fairness constraint. For $\vec{\sigma} \in \{0,1\}^m$, we construct a family of distributions $\mathbb{P}_{\vec{\sigma}}$ on $[0,1]^d \times \{0,1\} \times \{0,1\}$ such that $D_{\vec{\sigma},r}(t^\star_{\vec{\sigma}}) < 0 < D_{\vec{\sigma},l}(t^\star_{\vec{\sigma}})$, and apply Assouad's lemma adapted to the fair classification problem. In the second part, we construct two distributions, $\mathbb{P}_1$ and $\mathbb{P}_{-1}$, on $\mathbb{R}^d \times \{0,1\} \times \{0,1\}$ such that $D_{\pm1,r}(t^\star_{\pm1}) = 0 = D_{\pm1,l}(t^\star_{\pm1})$, and then apply Le Cam's lemma to show that the second term appears in the lower bound for the fairness-impacted case.

Part I: For an integer $q \geqslant 1$ divisible by eight, we consider the following regular grid in the unit cube:

$$\mathcal{G}_q = \left\{\left(\frac{8k_1+4}{q}, \frac{8k_2+4}{q}, \ldots, \frac{8k_d+4}{q}\right) : k_i \in \{0,1,...,q/8-1\}, i \in [d]\right\}. \tag{C.1}$$

Observe that the cardinality of $\mathcal{G}_q$ is $M = 8^{-d}q^d$, and denote by $x_1, x_2, ..., x_M$ the points in $\mathcal{G}_q$. Let $m \leqslant M$ be a positive integer to be specified later. Writing $B_0 = [0,1]^d \setminus \cup_{j=1}^m B_{d,2}(x_j, 2q^{-1})$, we have that $B_0, B_{d,2}(x_1, 2q^{-1}), \ldots, B_{d,2}(x_m, 2q^{-1})$ forms a partition of $[0,1]^d$; note in particular any two distinct points x_i, x_j are at distance at least $8/q$, and so the balls do not intersect. We next define a collection $\mathcal{H} = \{\mathbb{P}_{\vec{\sigma}} : \vec{\sigma} \in \{0,1\}^m\}$ of probability distributions $\mathbb{P}_{\vec{\sigma}}$ on $\mathcal{Z} = \mathbb{R}^d \times \{0,1\} \times \{0,1\}$, indexed by the vertices of the hypercube, by specifying the marginal distributions of X and A, and the conditional distribution $\mathbb{P}_{Y|X,A=a}$.

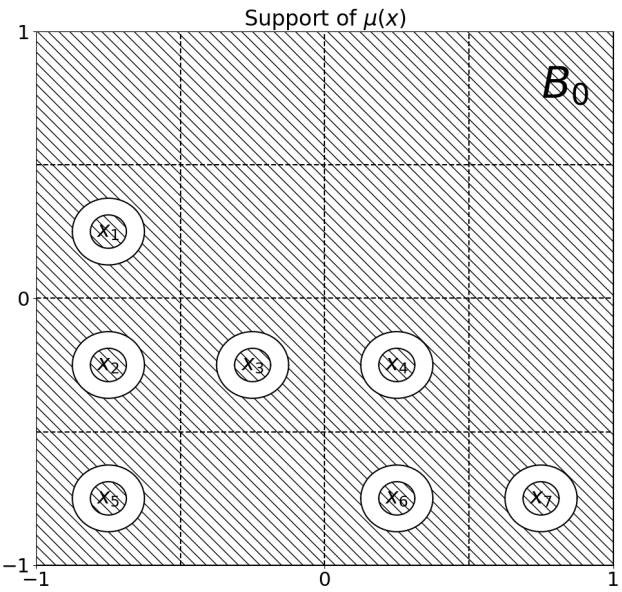


Figure 6: The shaded areas illustrate the support of μ in a two-dimensional setting.

- *Construction of marginal distributions of X and A*:

  We construct A and X to be independently distributed with the marginal distributions of X and A not depending on $\vec{\sigma}$. For any $\mathbb{P}_{\vec{\sigma}} \in \mathcal{H}$, we set $\mathbb{P}_{\vec{\sigma}}(A = 1) = 1/2$. For a certain w with $0 < w < m^{-1}$, to be chosen later, X has a density μ with respect to the Lebesgue measure λ on $\mathbb{R}^d$, defined in the following way:

$$\mu(x) = \begin{cases} w/\lambda[B_{d,2}(x_j, q^{-1})], & x \in B_{d,2}(x_j, q^{-1}), \quad j \in [m]; \\ (1 - mw)/\lambda[B_0], & x \in B_0; \\ 0, & \text{otherwise.} \end{cases} \tag{C.2}$$

  Figure 6 provides an illustration of the support of the function $\mu(x)$. Note that μ has a constant density over the ball $B_{d,2}(x_j, q^{-1})$ [9], for all j; as well as on B_0. Clearly, μ is a probability density function on $[0,1]^d$.

- *Construction of conditional distribution of Y given X and A*:

  Let u be an infinitely differentiable and non-increasing function on $[0, \infty)$, with bounded derivatives of all orders, such that $u(x) = 1$ when $x \leqslant 1$, $u(x) \in (0,1)$ when $x \in (1,2)$, and $u(x) = 0$ when $x \geqslant 2$. For $0 < c_\beta \leqslant q^\beta/2$, let $\phi : \mathbb{R}^d \to [0, \infty)$ be the function defined, for $x \in \mathbb{R}^d$, as

$$\phi(x) = c_\beta q^{-\beta} u(q\|x\|), \tag{C.3}$$

  and note that by the choice of c_β and the properties of u, $\phi(x) \leqslant 1/2$ for all x. For any $\mathbb{P}_{\vec{\sigma}} \in \mathcal{H}$, denote $\eta_{\vec{\sigma},a}(x,a) = \mathbb{P}_{\vec{\sigma}}(Y = 1 \mid X = x, A = a)$ for all x, a. We set

$$\eta_{\vec{\sigma},1}(x) = \begin{cases} (1 + \sigma_j \phi(x - x_j))/2, & x \in B_{d,2}(x_j, 2q^{-1}),\ j \in [m]; \\ 1/2, & \text{otherwise;} \end{cases} \tag{C.4}$$

  and

$$\eta_{\vec{\sigma},0}(x) = 1/2, \ x \in [0,1]^d. \tag{C.5}$$

[9] Here, we use q^{-1} rather than $2q^{-1}$.

Next, we consider the fair Bayes-optimal classifiers under $\mathbb{P}_{\vec{\sigma}} \in \mathcal{H}$. Define $D_{\vec{\sigma},-}$ to be D_- from (3.3) for the distribution $\mathbb{P}_{\vec{\sigma}}$ and define $D_{\vec{\sigma},+}$, $t^\star_{\vec{\sigma}}$, $T^\star_{\vec{\sigma},a}$, $\tau^\star_{\vec{\sigma},a}$, $g_{\vec{\sigma},a}$ similarly. By the definition of the distribution of X from (C.2), by (C.4), (C.5), and due to our choice $w < m^{-1}$,

$$\begin{aligned} -1 \leqslant D_{\vec{\sigma},-}(0) &= \mathbb{P}_{\vec{\sigma},X|A=1}\left(\eta_{\vec{\sigma},1}(X) > \frac{1}{2}\right) - \mathbb{P}_{\vec{\sigma},X|A=0}\left(\eta_{\vec{\sigma},0}(X) \geqslant \frac{1}{2}\right) \\ &= \sum_{j=1}^m I(\sigma_j = 1) \int_{B_{d,2}(x_j, q^{-1})} \mu(x)dx - 1 = w\sum_{j=1}^m I(\sigma_j = 1) - 1 < 0. \end{aligned}$$

Similarly,

$$D_{\vec{\sigma},+}(0) = \mathbb{P}_{\vec{\sigma},X|A=1}\left(\eta_{\vec{\sigma},1}(X) \geqslant \frac{1}{2}\right) - \mathbb{P}_{\vec{\sigma},X|A=0}\left(\eta_{\vec{\sigma},0}(X) > \frac{1}{2}\right) = 1 > 0.$$

This implies $D_+(0) \geqslant -D_-(0)$. In addition, one can verify that for any $t > 0$, one has $D_{\vec{\sigma},+}(t) < 0$. Thus, we have $t^\star_{\vec{\sigma}} = \sup_t \{D_{\vec{\sigma},+}(t) > 0\} = 0$ and $D_{\vec{\sigma},-}(t^\star_{\vec{\sigma}}) < 0 < D_{\vec{\sigma},+}(t^\star_{\vec{\sigma}})$. Further, from (3.6), $T^\star_{\vec{\sigma},a} = 1/2$ for $a \in \{0,1\}$. Moreover, due to (A.3), and since $\delta = 0$ and $D_{\vec{\sigma},l}(t) < 0$, (A.5) becomes

$$\begin{aligned} &\mathbb{P}_{\vec{\sigma},X|A=1}\left(\eta_{\vec{\sigma},1}(X) > \frac{1}{2}\right) + \tau^\star_{\vec{\sigma},1}\mathbb{P}_{\vec{\sigma},X|A=1}\left(\eta_{\vec{\sigma},1}(X) = \frac{1}{2}\right) \\ &- \mathbb{P}_{\vec{\sigma},X|A=0}\left(\eta_{\vec{\sigma},0}(X) > \frac{1}{2}\right) - \tau^\star_{\vec{\sigma},0}\mathbb{P}_{\vec{\sigma},X|A=0}\left(\eta_{\vec{\sigma},0}(X) = \frac{1}{2}\right) = 0. \end{aligned}$$

Since by (C.2), (C.4), and (C.5), $\mathbb{P}_{\vec{\sigma},X|A=1}(\eta_{\vec{\sigma},1}(X) > 1/2) = w\sum_{j=1}^m I(\sigma_j = 1)$, while $\mathbb{P}_{\vec{\sigma},X|A=0}(\eta_{\vec{\sigma},0}(X) > 1/2) = 0$, and $\mathbb{P}_{\vec{\sigma},X|A=0}(\eta_{\vec{\sigma},0}(X) = 1/2) = 1$, this is equivalent to

$$w\sum_{j=1}^m I(\sigma_j = 1) + \tau^\star_{\vec{\sigma},1}\mathbb{P}_{\vec{\sigma},X|A=1}(\eta_{\vec{\sigma},1}(X) = 1/2) = \tau^\star_{\vec{\sigma},0}.$$

Hence, to ensure that a classifier is Bayes-optimal, due to (A.5) in Theorem A.1, it suffices to take $\tau^\star_{\vec{\sigma},1} = 0$ and $\tau^\star_{\vec{\sigma},0} = w\sum_{j=1}^m I(\sigma_j = 1) \in [0,1]$. Thus, based on (A.3), using that $T^\star_{\vec{\sigma},a} = 1/2$, a fair Bayes-optimal classifier is given by

$$\begin{cases} f^\star_{\vec{\sigma}}(x,1) = I\left(x \in \bigcup_{j\in[m]:\sigma_j=1} B_{d,2}(x_j, 2q^{-1})\right); \\ f^\star_{\vec{\sigma}}(x,0) = w\sum_{j=1}^m I(\sigma_j = 1). \end{cases} \tag{C.6}$$

Now, we verify the distributional conditions:

- *Smoothness Condition from Definition 4.3*: For any $b \in \mathbb{N}^d$ such that $|b| \leqslant \lfloor\beta\rfloor_+$, the partial derivative $D^b\eta_{\vec{\sigma},1}$ at $x \in B_{d,2}(x_j, 2q^{-1})$ exists and $D^b\eta_{\vec{\sigma},1}(x) = c_\beta q^{|b|-\beta}\phi_j/2 \cdot D^b u(q\|x - x_j\|)$. Since $\|\cdot\|$ is infinitely differentiable on $\mathbb{R}^d/\{0\}$ and $u(\cdot)$ is infinitely differentiable on $[0,\infty)$ with $u(t) = 1$ for $0 < t \leqslant 1$, we have that $u(\|\cdot\|)$ is infinitely differentiable on $\mathbb{R}^d$ with bounded derivatives of all order. Thus, there is a constant M such that $|D^b u(q\|x - x_j\|)| < M$. Therefore, for any $j \in \{1,\ldots,m\}$ and any $x, x' \in B_{d,2}(x_j, 2q^{-1})$, we have, when c_β is small enough, that $|\eta_{\vec{\sigma},1}(x') - \eta_{\vec{\sigma},1}(x)| \leqslant L_\beta\|x' - x\|^\beta$. This implies that $\eta_{\vec{\sigma},a}$ belongs to the Hölder class $\Sigma(\beta, L_\beta, \mathbb{R}^d)$.

- *Margin Condition from Definition 4.4*: Since $D_{\vec{\sigma},r}(t^\star_{\vec{\sigma}}) < 0 < D_{\vec{\sigma},l}(t^\star_{\vec{\sigma}})$, we only need to verify condition (4.3). We have from (C.4) that, for $\vec{\sigma} \in \{0,1\}^m$,

$$g_{\vec{\sigma},+}(\varepsilon) = \mathbb{P}_{\vec{\sigma},X|A=1}\left(\frac{1}{2} - \varepsilon \leqslant \eta_{\vec{\sigma},1}(X) < \frac{1}{2}\right) + \mathbb{P}_{\vec{\sigma},X|A=0}\left(\frac{1}{2} < \eta_{\vec{\sigma},0}(X) \leqslant \frac{1}{2} + \varepsilon\right) = 0$$

and

$$g_{\vec{\sigma},-}(\varepsilon) = \mathbb{P}_{\vec{\sigma},X|A=1}\left(\frac{1}{2} < \eta_{\vec{\sigma},1}(X) \leqslant \frac{1}{2} + \varepsilon\right) + \mathbb{P}_{\vec{\sigma},X|A=0}\left(\frac{1}{2} - \varepsilon \leqslant \eta_{\vec{\sigma},0}(X) < \frac{1}{2}\right)$$

$$= \sum_{j=1}^{m} I\left(\sigma_j = 1\right) \cdot \mathbb{P}_{\vec{\sigma}, X|A=1}\left(0 < \phi(X - x_j) < 2\varepsilon\right) \leqslant \sum_{j=1}^{m} \int_{B_{d,2}(x_j, 2q^{-1})} I(0 < \phi(x - x_j) < 2\varepsilon)\mu(x)dx$$
$$= \sum_{j=1}^{m} \int_{B_{d,2}(x_j, q^{-1})} I(0 < \phi(x - x_j) < 2\varepsilon)\frac{w}{\lambda[B_{d,2}(x_j, q^{-1})]}dx = mw \cdot I\left(\varepsilon > c_\beta q^{-\beta}/2\right),$$

where in the last step we have used that due to (C.3), we have $\phi(x - x_j) \leqslant c_\beta q^{-\beta}$ for all x. Therefore, the γ-margin condition is satisfied if $mw = \Theta(q^{-\gamma\beta})$.

- *Strong density condition from Definition 4.5*: Let $m = \lfloor q^{d-\gamma\beta} \rfloor$, $w = c_w q^{-d}$ for some $c_w > 0$ and take $q \to \infty$ as $n \to \infty$. The condition $\gamma\beta \leqslant d$ ensures that $m \geqslant 1$. Next, note that since by definition X is independent of A, we have $\mu_a = \mu$ for all a. Now, let $\Omega_\mu = B_0 \cup_{j=1}^{m} B_{d,2}(x_j, q^{-1})$ be the support of μ. Recall that $\mathcal{C}_{z,q} = \{y : |y_i - z_i| \leqslant 4q^{-1}, i = 1, 2, \ldots, d\}$ and $\mathcal{D}_{z,q} = B_{d,2}(z, 2q^{-1}) \setminus B_{d,2}(z, q^{-1})$. Recalling $\mathcal{G}_q$ from (C.1), we have that

$$[0,1]^d = \cup_{z \in \mathcal{G}_q} \mathcal{C}_{z,q} \quad \text{and} \quad \Omega_\mu = [0,1]^d \setminus (\cup_{i=1}^{m} \mathcal{D}_{x_i,q}).$$

For $x \in \Omega_\mu$, we have, for all i that $x \in [0,1]^d \setminus \mathcal{C}_{x_i,q}$. Then, by Lemma B.3, for any $r > 0$, there exists $0 < C_r \leqslant 1$ such that,

$$\lambda[B_{d,2}(x, r) \cap \Omega_\mu] = \lambda\left[B_{d,2}(x, r) \cap [0,1]^d\right] - \sum_{i=1}^{m} \lambda[B_{d,2}(x, r) \cap \mathcal{D}_{x_i,q}]$$
$$= \sum_{z \in \mathcal{G}_q} \lambda[B_{d,2}(x, r) \cap \mathcal{C}_{z,q}] - \sum_{i=1}^{m} \lambda[B_{d,2}(x, r) \cap \mathcal{D}_{x_i,q}]$$
$$= \sum_{z \in \mathcal{G}_q \setminus \left(\cup_{i=1}^{m} \{x_i\}\right)} \lambda[B_{d,2}(x, r) \cap \mathcal{C}_{z,q}] + \sum_{i=1}^{m} \lambda[B_{d,2}(x, r) \cap \mathcal{C}_{x_i,q}] - \sum_{i=1}^{m} \lambda[B_{d,2}(x, r) \cap \mathcal{D}_{x_i,q}]$$
$$= \sum_{z \in \mathcal{G}_q \setminus \left(\cup_{i=1}^{m} \{x_i\}\right)} \lambda[B_{d,2}(x, r) \cap \mathcal{C}_{z,q}] + \sum_{i=1}^{m} \lambda[B_{d,2}(x, r) \cap (\mathcal{C}_{x_i,q} \setminus \mathcal{D}_{x_i,q})]$$
$$\geqslant \sum_{z \in \mathcal{G}_q \setminus \left(\cup_{i=1}^{m} \{x_i\}\right)} \lambda[B_{d,2}(x, r) \cap \mathcal{C}_{z,q}] + C_r \sum_{i=1}^{m} \lambda[B_{d,2}(x, r) \cap \mathcal{C}_{x_i,q}]$$
$$\geqslant C_r \sum_{z \in \mathcal{G}_q} \lambda[B_{d,2}(x, r) \cap \mathcal{C}_{z,q}] = C_r \lambda[B_{d,2}(x, r) \cap [0,1]^d].$$

On the other hand, if $x \in \cup_{j=1}^{m} B_{d,2}(x_j, q^{-1})$, due to (C.2), we have due to the choice of w that $\mu(x) = c_w q^{-d}/\lambda[B_{d,2}(x_j, q^{-1})] = c_w V_{d,2}^{-1}$. On the other hand, if $x \in B_0$, due to (C.2), we have $\mu(x) = (1 - mc_w q^{-d})/(1 - m2^d q^{-d} V_{d,2})$. Thus, the conditional distribution of X given $A = a$ satisfies the strong density condition with $c_\mu = C_r$, $r_\mu = 1$, and $\mu_{\min} \leqslant \mu_{\max}$ if $\mu_{\min} \leqslant (1 - mc_w q^{-d})/(1 - m2^d q^{-d} V_{d,2}) \leqslant \mu_{\max}$ and $\mu_{\min} \leqslant c_w/V_{d,2} \leqslant \mu_{\max}$.

Finally, we derive the first term of the minimax lower bound. For $r \in \{0, 1\}$ and $\vec{\sigma} \in \{-1, 1\}^m$, denote $\vec{\sigma}_{j,r} = (\sigma_1, \ldots, \sigma_{j-1}, r, \sigma_{j+1}, \ldots, \sigma_m)$, Clearly, $\mathbb{P}_{\vec{\sigma}_{j,1}}^{\otimes n}$ is absolutely continuous with respect to $\mathbb{P}_{\vec{\sigma}_{j,0}}^{\otimes n}$. Moreover, recall that the total variation distance between $\mathbb{P}_{\vec{\sigma}_{j,1}}^{\otimes n}$ and $\mathbb{P}_{\vec{\sigma}_{j,0}}^{\otimes n}$ can be expressed as

$$\mathrm{TV}\left(\mathbb{P}_{\vec{\sigma}_{j,1}}^{\otimes n}, \mathbb{P}_{\vec{\sigma}_{j,0}}^{\otimes n}\right) = 1 - \int \left(\frac{\mathbb{P}_{\vec{\sigma}_{j,1}}^{\otimes n}(z)}{\mathbb{P}_{\vec{\sigma}_{j,0}}^{\otimes n}(z)} \wedge 1\right) d\mathbb{P}_{\vec{\sigma}_{j,0}}^{\otimes n}(z).$$

Now, we provide an upper bound on the Kullback–Leibler divergence between $\mathbb{P}_{\vec{\sigma}_{j,1}}^{\otimes n}$ and $\mathbb{P}_{\vec{\sigma}_{j,0}}^{\otimes n}$, taking $j = 1$ without loss of generality. As $\mathbb{P}_{\vec{\sigma}_{1,1}}(A = a) = \mathbb{P}_{\vec{\sigma}_{1,0}}(A = a) = 1/2$ for $a \in \{0, 1\}$, recalling $\eta_{\vec{\sigma},a}$ from

(C.4) and (C.5), we have

$$\begin{aligned}
&\mathrm{KL}\left(\mathbb{P}_{\vec{\sigma}_{1,1}}, \mathbb{P}_{\vec{\sigma}_{1,0}}\right) = \sum_{a\in\mathcal{A}} \mathbb{P}_{\vec{\sigma}_{1,1}}(A=a) \int \eta_{\vec{\sigma}_{1,1},a}(x) \log \frac{\eta_{\vec{\sigma}_{1,1},a}(x)}{\eta_{\vec{\sigma}_{1,0},a}(x)} \mu(x)dx \\
&= \frac{1}{2}\left[\int \eta_{\vec{\sigma}_{1,1},1}(x) \log \frac{\eta_{\vec{\sigma}_{1,1},1}(x)}{\eta_{\vec{\sigma}_{1,0},1}(x)} \mu(x)dx + \int (1-\eta_{\vec{\sigma}_{1,1},1}(x)) \log \frac{1-\eta_{\vec{\sigma}_{1,1},1}(x)}{1-\eta_{\vec{\sigma}_{1,0},1}(x)} \mu(x)dx\right] \\
&= \frac{1}{2}\int_{B_{d,2}(x_1,2q^{-1})} \left[\frac{1+\phi(x-x_1)}{2}\log\left(1+\phi(x-x_1)\right) + \frac{1-\phi(x-x_1)}{2}\log\left(1-\phi(x-x_1)\right)\right]\mu(x)dx \\
&\leqslant \frac{1}{2}\int_{B_{d,2}(x_1,2q^{-1})} \left[\frac{1+\phi(x-x_1)}{2}\phi(x-x_1) - \frac{1-\phi(x-x_1)}{2}\phi(x-x_1)\right]\mu(x)dx \\
&\leqslant \frac{1}{2}\int_{B_{d,2}(x_1,2q^{-1})} \left(\phi^2(x-x_1)\frac{w}{\lambda\left[B_{d,2}(x_1,q^{-1})\right]}\right)dx \leqslant \frac{w}{2}c_\beta^2 q^{-2\beta}.
\end{aligned}$$

Here, the first inequality holds since, for $0 < x < 1$, $\log(1+x) < x$ and $\log(1-x) < -x$; and the last inequality holds due to (C.3). By Pinsker's inequality (Tsybakov, 2009), we therefore have

$$\mathrm{TV}\left(\mathbb{P}^{\otimes n}_{\vec{\sigma}_{1,1}}, \mathbb{P}^{\otimes n}_{\vec{\sigma}_{1,-1}}\right) \leqslant \frac{1}{2}\sqrt{\mathrm{KL}\left(\mathbb{P}^{\otimes n}_{\vec{\sigma}_{1,1}}, \mathbb{P}^{\otimes n}_{\vec{\sigma}_{1,-1}}\right)} = \frac{1}{2}\sqrt{n\mathrm{KL}\left(\mathbb{P}_{\vec{\sigma}_{1,1}}, \mathbb{P}_{\vec{\sigma}_{1,-1}}\right)} \leqslant \sqrt{\frac{1}{8}c_\beta\sqrt{nw}q^{-\beta}}. \tag{C.7}$$

Recalling that $w = c_w q^{-d}$ and taking $q = n^{1/(2\beta+d)}$ with $c_w = 2c_\beta^{-2}$, we have $\mathrm{TV}\left(\mathbb{P}^{\otimes n}_{\vec{\sigma}_{1,1}}, \mathbb{P}^{\otimes n}_{\vec{\sigma}_{1,-1}}\right) \leqslant 1/2$.

To complete the proof for the first term in minimax lower bound, we apply Assouad's lemma (Assouad, 1983; Tsybakov, 2009) to the class $\mathcal{H}$. Let ν denote the distribution of a Bernoulli variable with parameter $1/2$, so that for $\sigma \sim \nu$, $\nu(\sigma=1) = \nu(\sigma=0) = \frac{1}{2}$. For data-dependent sets $\widehat{G}_{a,1}$ and $\widehat{G}_{a,\tau}$ for $a \in \{0,1\}$, let $\widehat{f}_{\delta,n}$ be the classifier with, for all x, a,

$$\widehat{f}_{\delta,n}(x,a) = I(x \in \widehat{G}_{a,1}) + \widehat{\tau} I(x \in \widehat{G}_{a,\tau}).$$

We use $\mathbb{E}_{\vec{\sigma}}$ to denote expectation under the distribution $\mathbb{P}_{\vec{\sigma}}$. Then, by the definition of d_E from (4.2), using that $T^\star_{\vec{\sigma},a} = 1/2$, and by (C.6), (C.4)

$$\begin{aligned}
&\sup_{\mathbb{P}\in\mathcal{P}_\Sigma} \mathbb{E}^{\otimes n} d_E(\widehat{f}_{\delta,n}, f^\star) \geqslant \sup_{\vec{\sigma}\in\{0,1\}^m} \mathbb{E}^{\otimes n}_{\vec{\sigma}} d_E(\widehat{f}_{\delta,n}, f^\star_{\vec{\sigma}}) \\
&= \sup_{\vec{\sigma}\in\{0,1\}^m} \mathbb{E}^{\otimes n}_{\vec{\sigma}} \left\{\int \left(\widehat{f}_{\delta,n}(x,1) - f^\star_{\vec{\sigma}}(x,1)\right)\left(\frac{1}{2} - \eta_{\vec{\sigma},1}(x)\right)\mu(x)dx\right\} \\
&= \sup_{\vec{\sigma}\in\{0,1\}^m} \mathbb{E}^{\otimes n}_{\vec{\sigma}} \left\{\sum_{j=1}^m \int_{B_{d,2}(x_j,2q^{-1})} \left|\widehat{f}_{\delta,n}(x,1) - \sigma_j\right| \frac{\phi(x-x_j)}{2}\mu(x)dx\right\} \\
&\geqslant \frac{1}{2}\mathbb{E}^{\otimes m}_\nu \left\{\mathbb{E}^{\otimes n}_{\vec{\sigma}} \left\{\sum_{j=1}^m \int_{B_{d,2}(x_j,2q^{-1})} \left|\widehat{f}_{\delta,n}(x,1) - \sigma_j\right| \phi(x-x_j)\mu(x)dx\right\}\right\}.
\end{aligned}$$

In the last line, we have written $\mathbb{E}^{\otimes m}_\nu$ for the expectation over $\vec{\sigma} = (\sigma_1, \ldots, \sigma_m)$ with i.i.d. $\sigma_i \sim \nu$ for all $i \in [m]$. Recalling the definition $\vec{\sigma}_{j,0}$, the last term equals

$$\begin{aligned}
&\frac{1}{2}\mathbb{E}^{\otimes m}_\nu \left\{\sum_{j=1}^m \mathbb{E}^{\otimes n}_{\vec{\sigma}_{j,0}} \left\{\frac{\mathbb{P}^{\otimes n}_{\vec{\sigma}}}{\mathbb{P}^{\otimes n}_{\vec{\sigma}_{j,0}}} \int_{B_{d,2}(x_j,2q^{-1})} \left|\widehat{f}_{\delta,n}(x,1) - \sigma_j\right| \phi(x-x_j)\mu(x)dx\right\}\right\} \\
&= \frac{1}{2}\mathbb{E}^{\otimes\{m-1\}}_\nu \left\{\sum_{j=1}^m \mathbb{E}_{\sigma_j\sim\nu}\mathbb{E}^{\otimes n}_{\vec{\sigma}_{j,0}} \left\{\frac{\mathbb{P}^{\otimes n}_{\vec{\sigma}}}{\mathbb{P}^{\otimes n}_{\vec{\sigma}_{j,0}}} \int_{B_{d,2}(x_j,2q^{-1})} \left|\widehat{f}_{\delta,n}(x,1) - \sigma_j\right| \phi(x-x_j)\mu(x)dx\right\}\right\} \\
&\geqslant \frac{1}{2}\mathbb{E}^{\otimes\{m-1\}}_\nu \left\{\sum_{j=1}^m \mathbb{E}^{\otimes n}_{\vec{\sigma}_{j,0}} \left\{\left(\frac{\mathbb{P}^{\otimes n}_{\vec{\sigma}_{j,1}}}{\mathbb{P}^{\otimes n}_{\vec{\sigma}_{j,0}}} \wedge 1\right) \mathbb{E}_{\sigma_j\sim\nu} \int_{B_{d,2}(x_j,2q^{-1})} \left|\widehat{f}_{\delta,n}(x,1) - \sigma_j\right| \phi(x-x_j)\mu(x)dx\right\}\right\}.
\end{aligned}$$

Since for all $z \in [0,1]$, $\mathbb{E}_{\sigma_j \sim \nu}|z - \sigma_j| = (|z-1| + |z|)/2 = 1/2$, this can be further written as

$$\frac{1}{4}\mathbb{E}_{\nu}^{\otimes\{m-1\}}\left\{\sum_{j=1}^{m}\mathbb{E}_{\vec{\sigma}_{j,0}}^{\otimes n}\left\{\left(1 - \mathrm{TV}\left(\mathbb{P}_{\vec{\sigma}_{j,1}}^{\otimes n}, \mathbb{P}_{\vec{\sigma}_{j,-1}}^{\otimes n}\right)\right)\int_{B_{d,2}(x_j, 2q^{-1})}\phi(x - x_j)\mu(x)dx\right\}\right\}.$$

Using (C.7) with $m = \lfloor q^{d-\gamma\beta}\rfloor$, $w = c_w q^{-d}$ and $q = n^{1/(2\beta+d)}$, as well as by

$$\int_{B_{d,2}(x_1, 2q^{-1})}\phi(x - x_1)\mu(x)dx = \int_{B_{d,2}(x_1, q^{-1})}\phi(x - x_j)\frac{w}{\lambda\left[B_{d,2}(x_1, q^{-1})\right]}dx = wc_\beta q^{-\beta},$$

this is lower bounded by $\frac{mwc_\beta q^{-\beta}}{8} \geqslant C' n^{-\frac{\beta(\gamma+1)}{2\beta+d}}$, as desired. This finishes the argument of the first part.

Part 2. In this part, we apply Le Cam's method to prove that the second term on the right hand side of (4.5) appears in the lower bound; see Figure 7 for an illustration of the construction. Recall that $V_{d,p}$ is the volume of a d-dimensional unit ℓ_p ball and let $v_d = V_{d,1}^{-1/d}$. We will construct two distributions $\mathbb{P}_1$ and $\mathbb{P}_{-1}$ on $\mathcal{X} \times \{0,1\}$ with $\mathcal{X} = [-3,3] \times [-1,1]^{d-1}$. For $0 < s \leqslant 1$ specified later, let $\mathcal{B}_1 = \{x = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d : |x_1 + (1+s)v_d| + |x_2| + \ldots |x_d| \leqslant v_d\}$, $\mathcal{B}_2 = \{x = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d : |x_1| + |x_2| + \ldots |x_d| \leqslant sv_d\}$, and $\mathcal{B}_3 = \{x = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d : |x_1 - (1+s)v_d| + |x_2| + \ldots |x_d| \leqslant v_d\}$. By construction, we have $\lambda[\mathcal{B}_1] = \lambda[\mathcal{B}_3] = 1$ and $\lambda[\mathcal{B}_2] = s^d$. We will the use subscripts $+1$ and -1 to denote quantities corresponding to $\mathbb{P}_1$ and $\mathbb{P}_{-1}$, respectively.

- *Construction of marginal distributions of X and A*: We set $\mathbb{P}_1(A=1) = \mathbb{P}_{-1}(A=1) = 0.5$ and for $j \in \{-1,1\}$ and $a \in \{0,1\}$, denote by $\mu_{j,a}$ the conditional density function of X given $A = a$ under $\mathbb{P}_j$, defined as, for $x = (x_1, ..., x_d)^\top$,

$$\begin{cases} \mu_{1,1}(x) = \mu_{-1,0}(x) = \begin{cases} 1/(2+2s^d), & x_1 \in \mathcal{B}_1 \cup \mathcal{B}_2; \\ 1/2, & x_1 \in \mathcal{B}_3; \\ 0, & \text{otherwise}; \end{cases} \\ \mu_{1,0}(x) = \mu_{-1,1}(x) = \begin{cases} 1/2, & x_1 \in \mathcal{B}_1; \\ 1/(2+2s^d), & x_1 \in \mathcal{B}_2 \cup \mathcal{B}_3; \\ 0, & \text{otherwise}. \end{cases} \end{cases}$$

- *Construction of conditional distribution of Y given X and A*: For $j \in \{-1,1\}$ and all x, a, consider the regression functions $\eta_{j,a}(x) = \mathbb{P}_j(Y = 1 \mid A = a, X = x)$, defined as

$$\eta_{j,a}(x) = \eta_a(x) = \begin{cases} 1/2 + (2a-1)/4 - c_\beta(sv_d)^{d/\gamma} - c_\beta(-x_1 - sv_d)^{d/\gamma}, & -(2+s)v_d \leqslant x_1 < -sv_d; \\ 1/2 + (2a-1)/4 - c_\beta(sv_d)^{d/\gamma} + c_\beta(x_1 + sv_d)^{d/\gamma}, & -sv_d \leqslant x_1 < 0; \\ 1/2 + (2a-1)/4 + c_\beta(sv_d)^{d/\gamma} - c_\beta(-x_1 + sv_d)^{d/\gamma}, & 0 \leqslant x_1 < sv_d; \\ 1/2 + (2a-1)/4 + c_\beta(sv_d)^{d/\gamma} + c_\beta(x_1 - sv_d)^{d/\gamma}, & sv_d \leqslant x_1 \leqslant (2+s)v_d. \end{cases} \tag{C.8}$$

  Here $c_\beta \in (0,1)$ is chosen small enough that $\eta_{j,a} \in \Sigma\left(\beta, L_\beta, \mathbb{R}^d\right)$ for all j, a, which can be done since $\gamma\beta \leqslant d$.

Next, we consider the fair Bayes-optimal classifiers under $\mathbb{P}_1$ and $\mathbb{P}_{-1}$. Define, for $j \in \{-1,1\}$, $D_{j,r}$ to be D_- from (3.3) for the distribution $\mathbb{P}_j$ and define $D_{j,\ell}$, $t_j^\star$, $T_{j,a}^\star$, $\tau_{j,a}^\star$, $g_{j,a}$ similarly. It can be readily verified that

$$\begin{aligned} &\mathbb{P}_{1,X|A=1}(\eta_{1,1}(X) > 3/4 + c_\beta(sv_d)^{d/\gamma}) = \mathbb{P}_{1,X|A=0}(\eta_{1,0}(X) \geqslant 1/4 - c_\beta(sv_d)^{d/\gamma}) \\ &= \mathbb{P}_{-1,X|A=1}(\eta_{-1,1}(X) > 3/4 - c_\beta(sv_d)^{d/\gamma}) = \mathbb{P}_{-1,X|A=0}(\eta_{-1,0}(X) \geqslant 1/4 + c_\beta(sv_d)^{d/\gamma}) = \frac{1}{2}. \end{aligned} \tag{C.9}$$

For instance, $\eta_{1,1}(x) > 3/4 + c_\beta(sv_d)^{d/\gamma}$ holds if and only if $sv_d \leqslant x_1 \leqslant (2+s)v_d$. Noting that $\mu_{j,a}(x) \equiv 0$ when $x \notin \cup_{j=1}^3 \mathcal{B}_j$, we have

$$\mathbb{P}_{1,X|A=1}(sv_d \leqslant X_1 \leqslant (2+s)v_d) = \int_{(\cup_{j=1}^3 \mathcal{B}_j) \cap \{sv_d \leqslant x_1 \leqslant (2+s)v_d\}} \mu_{1,1}(x)dx = \int_{\mathcal{B}_3} \mu_{1,1}(x)dx = \frac{1}{2}\lambda(\mathcal{B}_3) = 1/2.$$

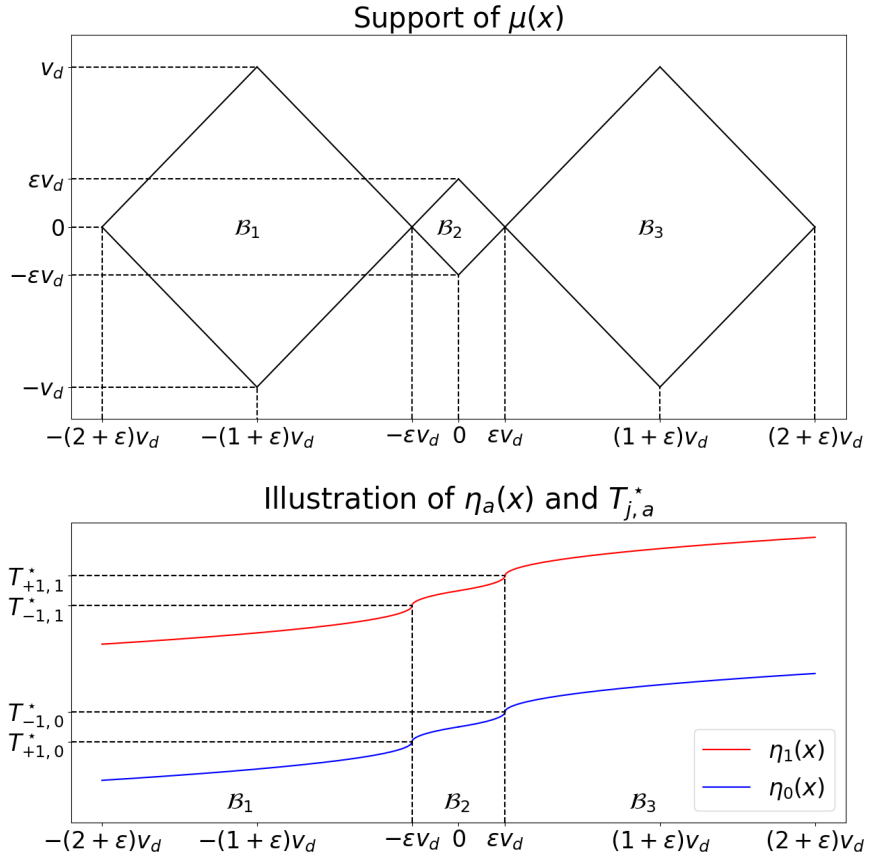


Figure 7: Illustration of constructions in the lower bound based on a two-dimensional setting. In the upper panel, we plot the support of $\mu(x)$, i.e., $\mathcal{B}_1$, $\mathcal{B}_2$ and $\mathcal{B}_3$. In the lower panel, we plot the regression functions for two groups (Here, we only plot $\eta_a(x_1)$ as $\eta_a(x)$ only depends on the first coordinate of x), and the fair thresholds under $\mathbb{P}_{+1}$ and $\mathbb{P}_{-1}$.

Recalling (3.3), using that $p_0 = p_1 = 1/2$, (C.9) further implies that for $j \in \{-1, 1\}$,

$$\begin{aligned}
&D_{j,r}\left(\frac{1}{4} + j \cdot c_\beta (sv_d)^{d/\gamma}\right) \\
&= \mathbb{P}_{j,X|A=1}\left(\eta_{j,1}(X) > \frac{3}{4} + j \cdot c_\beta (sv_d)^{d/\gamma}\right) - \mathbb{P}_{j,X|A=0}\left(\eta_{j,0}(X) \geqslant \frac{1}{4} - j \cdot c_\beta (sv_d)^{d/\gamma}\right) = \frac{1}{2} - \frac{1}{2} = 0.
\end{aligned}$$

As both $\eta_1(X)$ and $\eta_0(X)$ are continuous random variable on $\cup_{j=1}^3 \mathcal{B}_j$, We thus deduce that $t_j^\star = 1/4 + j \cdot c_\beta (sv_d)^{d/\gamma}$. In fact, for any $\varepsilon > 0$, we have

$$\begin{aligned}
&D_{j,r}(t_j^\star - \varepsilon) = D_{j,r}(t_j^\star - \varepsilon) - D_{j,r}(t_j^\star) + D_{j,r}(t_j^\star) \\
&= \mathbb{P}_{j,X|A=1}\left(\frac{1}{2} + t^\star - \varepsilon < \eta_{j,1}(X) \leqslant \frac{1}{2} + t^\star\right) + \mathbb{P}_{j,X|A=0}\left(\frac{1}{2} - t^\star \leqslant \eta_{j,0}(X) < \frac{1}{2} - t^\star + \varepsilon\right) > 0,
\end{aligned}$$

and

$$\begin{aligned}
&D_{j,r}(t_j^\star + \varepsilon) = D_{j,r}(t_j^\star + \varepsilon) - D_{j,r}(t_j^\star) + D_{j,r}(t_j^\star) \\
&= -\mathbb{P}_{j,X|A=1}\left(\frac{1}{2} + t^\star < \eta_{j,1}(X) \leqslant \frac{1}{2} + t^\star + \varepsilon\right) - \mathbb{P}_{j,X|A=0}\left(\frac{1}{2} - t^\star - \varepsilon \leqslant \eta_{j,0}(X) < \frac{1}{2} - t^\star\right) < 0.
\end{aligned}$$

Hence based on (A.6), if we set $\tau^\star_{j,1} = \tau^\star_{j,0} = 0$, a fair Bayes-optimal classifier is given by

$$f^\star_{\pm 1}(x,a) = I\left(\eta_{\pm 1,a}(x) > 1/2 + (2a-1)\left(1/4 \pm c_\beta (s v_d)^{d/\gamma}\right)\right). \tag{C.10}$$

Now, we verify the distributional conditions:

- *Smoothness Condition from Definition 4.3*: Since $\gamma\beta \leqslant d$, for any $b \in \mathbb{N}^d$ such that $|b| \leqslant \lfloor\beta\rfloor_+$, we have that $|b| \leqslant d/\gamma$. Thus, for $a \in \{0,1\}$, the partial derivative $D^b \eta_{j,a}$ exists and is bounded by $\Pi_{j=1}^{|b|}(d/\gamma - j + 1)(|x_1| + s v_d)^{d/\gamma - |b|}$. As $|x_1|$ and s are bounded, when $c_\beta > 0$ is small enough, for $j \in \{-1,1\}$ and $a \in \{0,1\}$, $\eta_{j,a}$ belongs to the Hölder class $\Sigma(\beta, L_\beta, \mathbb{R}^d)$.
- *Margin Condition from Definition 4.4*: In this case, we have $D_{j,-}(t^\star_j) = 0 = D_{j,+}(t^\star_j)$ for $j = \{-1,1\}$. To verify the margin condition, we need to provide both lower and upper bounds for

$$g_{\delta,j}(t^\star_j, \varepsilon) = \sum_a \mathbb{P}_{j,X|A=a}\left(0 < |\eta_{j,a} - T^\star_{j,a}| < \varepsilon\right) \quad \text{for } j = \{-1,1\}, \varepsilon \in (0, \varepsilon_0],$$

with some $\varepsilon_0 > 0$. We set $\varepsilon_0 = c_\beta (v_d)^{d/\gamma}$. By construction, we have for $j \in \{-1,1\}$ and all $x \in \mathcal{X}$ that $\mu_{j,1}(x) = \mu_{-j,0}(x)$ and $\eta_{-j,1}(x) = \eta_{-j,0}(x) - 1/2$. Moreover, $T^\star_{j,1} - T^\star_{-j,0} = 1/2 + 1/4 + j \cdot c_\beta (s v_d)^{d/\gamma} - (1/2 - (1/4 - j \cdot c_\beta (s v_d)^{d/\gamma})) = 1/2$. Thus,

$$\begin{aligned}
&\mathbb{P}_{j,X|A=1}\left(T^\star_{j,1} < \eta_{j,1}(X) < T^\star_{j,1} + \varepsilon\right) = \int_{\mathcal{X}} I\left(T^\star_{j,1} < \eta_{j,1}(x) < T^\star_{j,1} + \varepsilon\right)\mu_{j,1}(x)dx \\
&= \int_{\mathcal{X}} I\left(T^\star_{j,1} - \frac{1}{2} < \eta_{j,1}(x) - \frac{1}{2} < T^\star_{j,1} + \varepsilon - \frac{1}{2}\right)\mu_{j,1}(x)dx \\
&= \int_{\mathcal{X}} I\left(T^\star_{-j,0} < \eta_{-j,0}(x) < T^\star_{-j,0} + \varepsilon\right)\mu_{-j,0}(x)dx \\
&= \mathbb{P}_{-j,X|A=0}\left(T^\star_{-j,0} < \eta_{-j,0}(X) < T^\star_{-j,0} + \varepsilon\right).
\end{aligned} \tag{C.11}$$

Again, by construction, we have for $a \in \{0,1\}$ and $x \in \mathcal{X}$ that $\mu_{1,a}(x) = \mu_{1,1-a}(-x)$ and $\eta_{1,1-a}(-x) = 1 - \eta_{1,a}(x)$. Moreover, $T^\star_{1,a} = 1/2 + (2a-1)t^\star_1 = 1 - (1/2 + (1-2a)t^\star_1) = 1 - T^\star_{1,1-a}$. Thus,

$$\begin{aligned}
&\mathbb{P}_{1,X|A=a}\left(T^\star_{1,a} < \eta_{1,a}(X) < T^\star_{1,a} + \varepsilon\right) = \int_{\mathcal{X}} I\left(1 - T^\star_{1,a} - \varepsilon < 1 - \eta_{1,a}(x) < 1 - T^\star_{1,a}\right)\mu_{1,a}(x)dx \\
&= \int_{\mathcal{X}} I\left(T^\star_{1,1-a} - \varepsilon < \eta_{1,1-a}(-x) < T^\star_{1,1-a}\right)\mu_{1,1-a}(-x)dx \\
&= \int_{\mathcal{X}} I\left(T^\star_{1,1-a} - \varepsilon < \eta_{1,1-a}(x) < T^\star_{1,1-a}\right)\mu_{1,1-a}(x)dx \\
&= \mathbb{P}_{1,X|A=1-a}\left(T^\star_{1,1-a} - \varepsilon < \eta_{1,1-a}(X) < T^\star_{1,1-a}\right).
\end{aligned} \tag{C.12}$$

Thus, for $j \in \{-1,1\}$,

$$\begin{aligned}
g_{\delta,j}(t^\star_j, \varepsilon) &= \sum_a \mathbb{P}_{j,X|A=a}\left(0 < |\eta_{j,a} - T^\star_{j,a}| < \varepsilon\right) \\
&= \sum_a \left(\mathbb{P}_{j,X|A=a}\left(T^\star_{j,a} - \varepsilon < \eta_{j,a} < T^\star_{j,a}\right) + \mathbb{P}_{j,X|A=a}\left(T^\star_{j,a} < \eta_{j,a} < T^\star_{j,a} + \varepsilon\right)\right) \\
&= 2\left(\mathbb{P}_{1,X|A=1}\left(T^\star_{1,1} - \varepsilon < \eta_{1,1} < T^\star_{1,1}\right) + \mathbb{P}_{1,X|A=1}\left(T^\star_{1,1} < \eta_{1,1} < T^\star_{1,1} + \varepsilon\right)\right),
\end{aligned}$$

where the last equality follows (C.11) and (C.12). Next, we provide upper bounds for $\mathbb{P}_{1,X|A=1}\left(T^\star_{1,1} - \varepsilon < \eta_{1,1} < T^\star_{1,1}\right)$ and $\mathbb{P}_{1,X|A=1}\left(T^\star_{1,1} < \eta_{1,1} < T^\star_{1,1} + \varepsilon\right)$ when $0 < \varepsilon \leqslant c_\beta (v_d)^{d/\gamma}$. We first observe the following two facts:

  - Fact 1. By construction, we have that

$$\begin{aligned}
&\mathcal{B}_1 \subset \{x : -(2+s)v_d \leqslant x_1 \leqslant -v_d\}, \quad \mathcal{B}_2 \subset \{x : -s v_d \leqslant x_1 \leqslant s v_d\}, \\
&\text{and } \mathcal{B}_3 \subset \{x : v_d \leqslant x_1 \leqslant (2+s)v_d\}.
\end{aligned} \tag{C.13}$$

– Fact 2. Recalling that for any $d \geqslant 1$, $V_{d,1}$ is the volume of the unit ℓ_1 ball, we have

$$\begin{aligned} V_{d,1} &= \int_{\sum_{j=1}^d |x_j| \leqslant 1} dx = \int_{-1}^1 \left(\int_{\sum_{j=2}^d |x_j| \leqslant 1-|x_1|} dx_2 \dots dx_d \right) dx_1 \\ &= 2\int_0^1 \left(\int_{\sum_{j=2}^d |x_j| \leqslant 1-x_1} dx_2 \dots dx_d \right) dx_1 = 2\int_0^1 \left(\int_{\sum_{j=2}^d |x_j| \leqslant x_1} dx_2 \dots dx_d \right) dx_1 \\ &= 2V_{d-1,1} \int_0^1 x_1^{d-1} dx_1 = \frac{2V_{d-1,1}}{d}. \end{aligned} \tag{C.14}$$

This further implies $V_{d,1} = 2^d/(d!)$ and $v_d = (d!)^{1/d}/2$.

For $\mathbb{P}_{1,X|A=1}\left(T^\star_{1,1} < \eta_{1,1} < T^\star_{1,1} + \varepsilon\right)$, since $\eta_{1,1}(x)$ only depends on x_1, is continuous, and is strictly increasing as a function of x_1, its inverse $\eta_{1,1}^{-1}$ as a function of x_1 exists with $\eta_{1,1}^{-1}(T^\star_{1,1}) = sv_d$. Moreover, for $0 < \varepsilon \leqslant \varepsilon_0$, $\eta_{1,1}^{-1}(T^\star_{1,1} + \varepsilon) = sv_d + (\varepsilon/c_\beta)^{\gamma/d}$. Thus, $T^\star_{1,1} < \eta_{1,1}(x) < T^\star_{1,1} + \varepsilon$ is equivalent to $sv_d < x_1 < sv_d + (\varepsilon/c_\beta)^{\gamma/d}$. Further, by (C.13), we have $\mathcal{B}_j \cap \{sv_d < x_1 < sv_d + (\varepsilon/c_\beta)^{\gamma/d}\} = \emptyset$ for $j = 1, 2$. Thus, for $0 < \varepsilon \leqslant \varepsilon_0$,

$$\begin{aligned} &\mathbb{P}_{1,X|A=1}\left(T^\star_{1,1} < \eta_{1,1}(X) < T^\star_{1,1} + \varepsilon\right) = \mathbb{P}_{1,X|A=1}\left(sv_d < X_1 < sv_d + \left(\frac{\varepsilon}{c_\beta}\right)^{\gamma/d}\right) \\ &= \int I\left(sv_d < x_1 < sv_d + \left(\frac{\varepsilon}{c_\beta}\right)^{\gamma/d}\right) \mu_{1,1}(x)dx = \sum_{j=1}^3 \int_{\mathcal{B}_j} I\left(sv_d < x_1 < sv_d + \left(\frac{\varepsilon}{c_\beta}\right)^{\gamma/d}\right) \mu_{1,1}(x)dx \\ &= \frac{1}{2}\int_{\mathcal{B}_3} I\left(sv_d < x_1 < sv_d + \left(\frac{\varepsilon}{c_\beta}\right)^{\gamma/d}\right) dx. \end{aligned} \tag{C.15}$$

Since $0 < \varepsilon \leqslant c_\beta (v_d)^{d/\gamma}$, and recalling (C.14), this further equals

$$\begin{aligned} &\frac{1}{2}\int_{sv_d}^{sv_d + \left(\frac{\varepsilon}{c_\beta}\right)^{\gamma/d}} \left(\int_{\sum_{j=2}^d |x_j| \leqslant v_d - |x_1 - (1+s)v_d|} dx_2 \dots dx_d \right) dx_1 \\ &= \frac{1}{2}\int_{sv_d}^{sv_d + \left(\frac{\varepsilon}{c_\beta}\right)^{\gamma/d}} \left(\int_{\sum_{j=2}^d |x_j| \leqslant x_1 - sv_d} dx_2 \dots dx_d \right) dx_1 = \frac{1}{2}\int_0^{\left(\frac{\varepsilon}{c_\beta}\right)^{\gamma/d}} \left(\int_{\sum_{j=2}^d |x_j| \leqslant x_1} dx_2 \dots dx_d \right) dx_1 \\ &= \frac{1}{2}\int_0^{\left(\frac{\varepsilon}{c_\beta}\right)^{\gamma/d}} V_{d-1,1} x_1^{d-1} dx_1 = \frac{V_{d-1,1}}{2d}\left(\frac{\varepsilon}{c_\beta}\right)^{\gamma} = \frac{V_{d,1}}{4}\left(\frac{\varepsilon}{c_\beta}\right)^{\gamma}. \end{aligned} \tag{C.16}$$

To bound $\mathbb{P}_{1,X|A=1}\left(T^\star_{1,1} - \varepsilon < \eta_{1,1} < T^\star_{1,1}\right)$, we first study the inverse $\eta_{1,1}^{-1}$ of $\eta_{1,1}$, viewed as a function of x_1. We have $\eta_{1,1}^{-1}(T^\star_{1,1}) = sv_d$ and

$$\eta_{1,1}^{-1}(T^\star_{1,1} - \varepsilon) = \begin{cases} sv_d - \left(\frac{\varepsilon}{c_\beta}\right)^{\gamma/d}, & 0 < \varepsilon \leqslant c_\beta (sv_d)^{d/\gamma}; \\ \left(2\left(sv_d\right)^{d/\gamma} - \frac{\varepsilon}{c_\beta}\right)^{\gamma/d} - sv_d, & c_\beta (sv_d)^{d/\gamma} < \varepsilon \leqslant 2c_\beta (sv_d)^{d/\gamma}; \\ -\left(\frac{\varepsilon}{c_\beta} - 2\left(sv_d\right)^{d/\gamma}\right)^{\gamma/d} - sv_d, & 2c_\beta (sv_d)^{d/\gamma} < \varepsilon \leqslant 2c_\beta (sv_d)^{d/\gamma} + c_\beta (2v_d)^{d/\gamma}. \end{cases}$$

Thus, for $0 < \varepsilon \leqslant 2c_\beta (sv_d)^{d/\gamma} + c_\beta (2v_d)^{d/\gamma}$, $T^\star_{1,1} - \varepsilon < \eta_{1,1}(x) < T^\star_{1,1}$ is equivalent to

$$\begin{cases} sv_d - \left(\frac{\varepsilon}{c_\beta}\right)^{\gamma/d} < x_1 < sv_d, & \text{when } 0 < \varepsilon \leqslant c_\beta (sv_d)^{d/\gamma}; \\ \left(2\left(sv_d\right)^{d/\gamma} - \frac{\varepsilon}{c_\beta}\right)^{\gamma/d} - sv_d < x_1 < sv_d, & \text{when } c_\beta (sv_d)^{d/\gamma} < \varepsilon \leqslant 2c_\beta (sv_d)^{d/\gamma}; \\ -\left(\frac{\varepsilon}{c_\beta} - 2\left(sv_d\right)^{d/\gamma}\right)^{\gamma/d} - sv_d < x_1 < sv_d, & \text{when } 2c_\beta (sv_d)^{d/\gamma} < \varepsilon \leqslant 2c_\beta (sv_d)^{d/\gamma} + c_\beta (2v_d)^{d/\gamma}. \end{cases}$$

Now we consider several cases.

– (1) $0 < \varepsilon \leqslant c_\beta (sv_d)^{d/\gamma}$. We can write

$$\begin{aligned}
&\mathbb{P}_{1,X|A=1}\left(T^\star_{1,1} - \varepsilon < \eta_{1,1}(X) < T^\star_{1,1}\right) = \mathbb{P}_{1,X|A=1}\left(sv_d - \left(\frac{\varepsilon}{c_\beta}\right)^{\gamma/d} < X_1 < sv_d\right) \\
&= \int I\left(sv_d - \left(\frac{\varepsilon}{c_\beta}\right)^{\gamma/d} < x_1 < sv_d\right)\mu_{1,1}(x)dx \\
&= \sum_{j=1}^{3}\int_{\mathcal{B}_j} I\left(sv_d - \left(\frac{\varepsilon}{c_\beta}\right)^{\gamma/d} < x_1 < sv_d\right)\mu_{1,1}(x)dx.
\end{aligned}$$

In this case, we have $sv_d - (\varepsilon/c_\beta)^{\gamma/d} \geqslant 0$. Using (C.13), we have $\mathcal{B}_j \cap \{x : sv_d - (\varepsilon/c_\beta)^{\gamma/d} < x_1 < sv_d\} = \emptyset$ for $j = 1, 3$. Thus, this further equals

$$\begin{aligned}
&\int_{\mathcal{B}_2} I\left(sv_d - \left(\frac{\varepsilon}{c_\beta}\right)^{\gamma/d} < x_1 < sv_d\right)\frac{1}{2+2s^d}dx \\
&= \frac{1}{2+2s^d}\int_{sv_d - \left(\frac{\varepsilon}{c_\beta}\right)^{\gamma/d}}^{sv_d}\left(\int_{\sum_{j=2}^d |x_j| \leqslant sv_d - |x_1|} dx_2 \ldots dx_d\right)dx_1 \\
&= \frac{1}{2+2s^d}\int_0^{\left(\frac{\varepsilon}{c_\beta}\right)^{\gamma/d}}\left(\int_{\sum_{j=2}^d |x_j| \leqslant x_1} dx_2 \ldots dx_d\right)dx_1 \\
&= \frac{V_{d-1,1}}{2+2s^d}\int_0^{\left(\frac{\varepsilon}{c_\beta}\right)^{\gamma/d}} x_1^{d-1}dx_1 = \frac{V_{d,1}}{4+4s^d}\left(\frac{\varepsilon}{c_\beta}\right)^{\gamma}.
\end{aligned}$$

As $1 \leqslant 1 + s^d \leqslant 2$ when $0 \leqslant s \leqslant 1$, we thus have, for $0 < \varepsilon \leqslant c_\beta (sv_d)^{d/\gamma}$,

$$\frac{V_{d,1}}{8}\left(\frac{\varepsilon}{c_\beta}\right)^{\gamma} \leqslant \mathbb{P}_{1,X|A=1}\left(T^\star_{1,1} - \varepsilon < \eta_{1,1}(X) < T^\star_{1,1}\right) \leqslant \frac{V_{d,1}}{4}\left(\frac{\varepsilon}{c_\beta}\right)^{\gamma}.$$

– (2) $c_\beta (sv_d)^{d/\gamma} < \varepsilon \leqslant 2c_\beta (sv_d)^{d/\gamma}$. In this case, we have $-sv_d \leqslant \left(2\left(sv_d\right)^{d/\gamma} - \frac{\varepsilon}{c_\beta}\right)^{\gamma/d} - sv_d < 0$. Again, by (C.13), we have $\mathcal{B}_j \cap \{x : sv_d - (\varepsilon/c_\beta)^{\gamma/d} < x_1 < sv_d\} = \emptyset$ for $j = 1, 3$. Thus,

$$\begin{aligned}
&\mathbb{P}_{1,X|A=1}\left(T^\star_{1,1} - \varepsilon < \eta_{1,1}(X) < T^\star_{1,1}\right) = \mathbb{P}_{1,X|A=1}\left(\left(2\left(sv_d\right)^{d/\gamma} - \frac{\varepsilon}{c_\beta}\right)^{\gamma/d} - sv_d < X_1 < sv_d\right) \\
&= \int I\left(\left(2\left(sv_d\right)^{d/\gamma} - \frac{\varepsilon}{c_\beta}\right)^{\gamma/d} - sv_d < x_1 < sv_d\right)\mu_{1,1}(x)dx \\
&= \sum_{j=1}^{3}\int_{\mathcal{B}_j} I\left(\left(2\left(sv_d\right)^{d/\gamma} - \frac{\varepsilon}{c_\beta}\right)^{\gamma/d} - sv_d < x_1 < sv_d\right)\mu_{1,1}(x)dx \\
&= \int_{\mathcal{B}_2} I\left(\left(2\left(sv_d\right)^{d/\gamma} - \frac{\varepsilon}{c_\beta}\right)^{\gamma/d} - sv_d < x_1 < sv_d\right)\frac{1}{2+2s^d}dx. \qquad \text{(C.17)}
\end{aligned}$$

On one hand, since $I\left(\left(2\left(sv_d\right)^{d/\gamma} - \varepsilon/c_\beta\right)^{\gamma/d} - sv_d < x_1 < sv_d\right) \leqslant 1$, this is upper bounded by

$$\int_{\mathcal{B}_2}\frac{1}{2+2s^d}dx = \frac{1}{2+2s^d}\lambda(\mathcal{B}_2) = \frac{V_{d,1}}{2+2s^d}(sv_d)^d < \frac{V_{d,1}}{2}\left(\frac{\varepsilon}{c_\beta}\right)^{\gamma},$$

where the last inequality holds since $1 + s^d > 1$ and $(\varepsilon/c_\beta)^\gamma > (sv_d)^d$ when $\varepsilon > c_\beta (sv_d)^{d/\gamma}$.

On the other hand, since $I\left(\left(2\left(sv_d\right)^{d/\gamma}-\varepsilon/c_\beta\right)^{\gamma/d}-sv_d<x_1<sv_d\right)\geqslant I\left(0<x_1<sv_d\right)$ when $\left(2\left(sv_d\right)^{d/\gamma}-\varepsilon/c_\beta\right)^{\gamma/d}-sv_d<0$, the (C.17) is lower bounded by

$$\begin{aligned}&\int_{\mathcal{B}_2} I\left(0<x_1<sv_d\right)\frac{1}{2+2s^d}dx=\frac{1}{2}\int_{\mathcal{B}_2}\frac{1}{2+2s^d}dx=\frac{1}{4+4s^d}\lambda(\mathcal{B}_2)\\&=\frac{V_{d,1}}{4+4s^d}(sv_d)^d\geqslant\frac{V_{d,1}}{2^{3+\gamma}}\left(\frac{\varepsilon}{c_\beta}\right)^\gamma,\end{aligned}$$

where the last inequality holds since $1+s^d<2$ and $(\varepsilon/c_\beta)^\gamma\leqslant 2^\gamma(sv_d)^d$ when $\varepsilon\leqslant 2c_\beta(sv_d)^{d/\gamma}$. As a result, we have, for $c_\beta(sv_d)^{d/\gamma}<\varepsilon\leqslant 2c_\beta(sv_d)^{d/\gamma}$,

$$\frac{V_{d,1}}{2^{3+\gamma}}\left(\frac{\varepsilon}{c_\beta}\right)^\gamma\leqslant\mathbb{P}_{1,X|A=1}\left(T^\star_{1,1}-\varepsilon<\eta_{1,1}(X)<T^\star_{1,1}\right)\leqslant\frac{V_{d,1}}{2}\left(\frac{\varepsilon}{c_\beta}\right)^\gamma.$$

– (3) $2c_\beta(sv_d)^{d/\gamma}<\varepsilon\leqslant 2c_\beta(sv_d)^{d/\gamma}+c_\beta(2v_d)^{d/\gamma}$. In this case, we have $-\left(\varepsilon/c_\beta-2\left(sv_d\right)^{d/\gamma}\right)^{\gamma/d}-sv_d<-sv_d$. By (C.13), $\mathcal{B}_2\cap\left\{x:-\left(\varepsilon/c_\beta-2\left(sv_d\right)^{d/\gamma}\right)^{\gamma/d}-sv_d<x_1<sv_d\right\}=\mathrm{int}\mathcal{B}_2$, where int denotes the interior of a set, and $\mathcal{B}_3\cap\left\{x:-\left(\varepsilon/c_\beta-2\left(sv_d\right)^{d/\gamma}\right)^{\gamma/d}-sv_d<x_1<sv_d\right\}=\emptyset$. Thus,

$$\begin{aligned}&\mathbb{P}_{1,X|A=1}\left(T^\star_{1,1}-\varepsilon<\eta_{1,1}(X)<T^\star_{1,1}\right)=\mathbb{P}_{1,X|A=1}\left(-\left(\varepsilon/c_\beta-2\left(sv_d\right)^{d/\gamma}\right)^{\gamma/d}-sv_d<X_1<sv_d\right)\\&=\int I\left(-\left(\varepsilon/c_\beta-2\left(sv_d\right)^{d/\gamma}\right)^{\gamma/d}-sv_d<x_1<sv_d\right)\mu_{1,1}(x)dx\\&=\sum_{j=1}^3\int_{\mathcal{B}_j} I\left(-\left(\varepsilon/c_\beta-2\left(sv_d\right)^{d/\gamma}\right)^{\gamma/d}-sv_d<x_1<sv_d\right)\mu_{1,1}(x)dx\\&=\int_{\mathcal{B}_1} I\left(-\left(\varepsilon/c_\beta-2\left(sv_d\right)^{d/\gamma}\right)^{\gamma/d}-sv_d<x_1<sv_d\right)\frac{1}{2+2s^d}dx+\int_{\mathrm{int}\mathcal{B}_2}\frac{1}{2+2s^d}dx.\end{aligned}$$

The first term further equals

$$\begin{aligned}&\frac{1}{2+2s^d}\int_{-\left(\varepsilon/c_\beta-2(sv_d)^{d/\gamma}\right)^{\gamma/d}-sv_d}^{-sv_d}\left(\int_{\sum_{j=2}^d|x_j|\leqslant v_d-|x_1+(1+s)v_d|}dx_2\ldots dx_d\right)dx_1\\&=\frac{1}{2+2s^d}\int_{-\left(\varepsilon/c_\beta-2(sv_d)^{d/\gamma}\right)^{\gamma/d}-sv_d}^{-sv_d}\left(\int_{\sum_{j=2}^d|x_j|\leqslant -x_1-sv_d}dx_2\ldots dx_d\right)dx_1\\&=\frac{1}{2+2s^d}\int_0^{\left(\varepsilon/c_\beta-2(sv_d)^{d/\gamma}\right)^{\gamma/d}}\left(\int_{\sum_{j=2}^d|x_j|\leqslant x_1}dx_2\ldots dx_d\right)dx_1\\&=\frac{V_{d,1}}{4+4s^d}\left(\varepsilon/c_\beta-2\left(sv_d\right)^{d/\gamma}\right)^\gamma.\end{aligned}$$

We have on one hand,

$$\frac{s^d}{2+2s^d}+\frac{V_{d,1}}{4+4s^d}\left(\frac{\varepsilon}{c_\beta}-2\left(sv_d\right)^{d/\gamma}\right)^\gamma<\frac{V_{d,1}}{2+2s^d}(sv_d)^d+\frac{V_{d,1}}{4+4s^d}\left(\frac{\varepsilon}{c_\beta}\right)^\gamma\leqslant\frac{3V_{d,1}}{4}\left(\frac{\varepsilon}{c_\beta}\right)^\gamma,$$

where the last inequality holds since $1+s^d>1$ and $(\varepsilon/c_\beta)^\gamma>(sv_d)^d$ when $\varepsilon>2c_\beta(sv_d)^{d/\gamma}>c_\beta(sv_d)^{d/\gamma}$. On the other hand, we have, when $2c_\beta(sv_d)^{d/\gamma}<\varepsilon\leqslant 4c_\beta(sv_d)^{d/\gamma}$,

$$\frac{s^d}{2+2s^d}+\frac{V_{d,1}}{4+4s^d}\left(\frac{\varepsilon}{c_\beta}-2\left(sv_d\right)^{d/\gamma}\right)^\gamma\geqslant\frac{s^d}{2+2s^d}=\frac{V_{d,1}}{2+2s^d}(sv_d)^d\geqslant\frac{V_{d,1}}{2^{2+2\gamma}}\left(\frac{\varepsilon}{c_\beta}\right)^\gamma,$$

and when $4c_\beta(sv_d)^{d/\gamma} < \varepsilon$,

$$\frac{s^d}{2+2s^d} + \frac{V_{d,1}}{4+4s^d}\left(\frac{\varepsilon}{c_\beta} - 2\left(sv_d\right)^{d/\gamma}\right)^\gamma \geqslant \frac{V_{d,1}}{4+4s^d}\left(\frac{\varepsilon}{2c_\beta}\right)^\gamma \geqslant \frac{V_{d,1}}{2^{3+\gamma}}\left(\frac{\varepsilon}{c_\beta}\right)^\gamma,$$

As a result, we have, for $2c_\beta(sv_d)^{d/\gamma} < \varepsilon \leqslant 2c_\beta(sv_d)^{d/\gamma} + c_\beta(2v_d)^{d/\gamma}$,

$$\frac{V_{d,1}}{2^{3+2\gamma}}\left(\frac{\varepsilon}{c_\beta}\right)^\gamma \leqslant \mathbb{P}_{1,X|A=1}\left(T^\star_{1,1} - \varepsilon < \eta_{1,1}(X) < T^\star_{1,1}\right) \leqslant \frac{3V_{d,1}}{4}\left(\frac{\varepsilon}{c_\beta}\right)^\gamma.$$

In particular, we have, for $0 < \varepsilon \leqslant c_\beta v_d^{d/\gamma}$,

$$\frac{V_{d,1}}{2^{3+2\gamma}}\left(\frac{\varepsilon}{c_\beta}\right)^\gamma \leqslant \mathbb{P}_{1,X|A=1}\left(T^\star_{1,1} - \varepsilon < \eta_{1,1}(X) < T^\star_{1,1}\right) \leqslant \frac{3V_{d,1}}{4}\left(\frac{\varepsilon}{c_\beta}\right)^\gamma. \tag{C.18}$$

Combining (C.15), (C.16) and (C.18), we get, for $j \in \{-1,+1\}$ and $0 < \varepsilon \leqslant c_\beta v_d^{d/\gamma}$,

$$V_{d,1}(2^{-1} + 2^{-(2+2\gamma)})\left(\frac{\varepsilon}{c_\beta}\right)^\gamma \leqslant g_{\delta,j}(t^\star_j, \varepsilon) = 2\mathbb{P}_{1,X|A=1}\left(0 < |\eta_{1,1} - T^\star_{1,1}| < \varepsilon\right) \leqslant 2V_{d,1}\left(\frac{\varepsilon}{c_\beta}\right)^\gamma.$$

Thus, for $0 < x \leqslant 1/2$, $g^{-1}_{\delta,j}(t^\star_j, x) \leqslant c_\beta\left(2x/V_{d,1}\right)^{1/\gamma}$, and the γ-margin condition from Definition 4.4 is satisfied.

- *Strong density condition from Definition 4.5*: Denote $C_d = [-3,3]^d$ and by $\Omega_\mu = \cup_{j=1}^3 \mathcal{B}_j$ the support of $\mu_{\pm 1,a}$. By construction, we have, for $a \in \{0,1\}$ and $j \in \{-1,1\}$, since $s \leqslant 1$,

$$\frac{1}{4} \leqslant \frac{1}{2+2s^d} \leqslant \mu_{j,a}(x) \leqslant \frac{1}{2}.$$

Thus, we can take $\mu_{\min} = 1/4$ and $\mu_{\max} = 1/2$. We then show that Ω_μ is a regular set by considering the following two cases: (1) $x \in \mathcal{B}_1 \cup \mathcal{B}_3$ and (2) $x \in \mathcal{B}_2$. Let $\Xi_d = V_{d,1}/(V_{d,2}2^d)$.

  - (1) When $x \in \mathcal{B}_1 \cup \mathcal{B}_3$: Without loss of generality, we assume $x \in \mathcal{B}_1$. By Lemma B.1 where $\mathcal{B}_1 = B_{d,1}(z,R)$ with $z = ((1+s)v_d, \ldots, 0)^\top$ and $R = v_d$, we have, when $r \leqslant 2v_d/(d+2)$,

$$\lambda[B_{d,2}(x,r) \cap \Omega_\mu] \geqslant \lambda[B_{d,2}(x,r) \cap \mathcal{B}_1] \geqslant \Xi_d \lambda[B_{d,2}(x,r)] \geqslant \Xi_d \lambda[B_{d,2}(x,r) \cap C_d].$$

  - (2) When $x \in \mathcal{B}_2$: Without loss of generality, we assume $x_1 > 0$.
    - When $r \leqslant 2sv_d/(d+2)$, by Lemma B.1 where $\mathcal{B}_2 = B_{d,1}(z,R)$ with $z = 0$ and $R = sv_d$, we have

$$\lambda[B_{d,2}(x,r) \cap \Omega_\mu] \geqslant \lambda[B_{d,2}(x,r) \cap \mathcal{B}_2] \geqslant \Xi_d \lambda[B_{d,2}(x,r)] \geqslant \Xi_d \lambda[B_{d,2}(x,r) \cap C_d].$$

    - When $2sv_d/(d+2) < r \leqslant 2\sqrt{2}sv_d$, we have, by Lemma B.1 where $\mathcal{B}_2 = B_{d,1}(z,R)$ with $z = 0$ and $R = sv_d$,

$$\begin{aligned}
&\lambda[B_{d,2}(x,r) \cap \Omega_\mu] \geqslant \lambda[B_{d,2}(x, 2sv_d/(d+2)) \cap \mathcal{B}_2] \\
&\geqslant \Xi_d \lambda[B_{d,2}(x, 2sv_d/(d+2))] = \frac{\Xi_d}{2^{d/2}(d+2)^d}\lambda[B_{d,2}(x, 2\sqrt{2}sv_d)] \\
&\geqslant \frac{\Xi_d}{2^{d/2}(d+2)^d}\lambda[B_{d,2}(x,r)] \geqslant \frac{\Xi_d}{2^{d/2}(d+2)^d}\lambda[B_{d,2}(x,r) \cap C_d].
\end{aligned}$$

    - When $\sqrt{2}sv_d < r/2 < v_d/(d+2)$, we denote $z = (sv_d, 0, \ldots, 0)^\top \in \mathcal{B}_3$. As $x_1 > 0$, we have $z \in B_{2,d}(x, \sqrt{2}sv_d)$. Noting that when $r > 2\sqrt{2}sv_d$, $r - \sqrt{2}sv_d > r/2$, this further implies $B_{2,d}(z, r/2) \subset B_{2,d}(z, r - \sqrt{2}sv_d) \subset B_{2,d}(x,r)$. Now, since $z \in \mathcal{B}_1$, by Lemma B.1 again, as $\mathcal{B}_3 = B_{d,1}(z,R)$ with $z = (-(1+s)v_d, \ldots, 0)^\top$ and $R = v_d$, we have

$$\lambda[B_{2,d}(x,r) \cap \Omega_\mu] \geqslant \lambda[B_{2,d}(z, r/2) \cap \mathcal{B}_3] \geqslant \Xi_d \lambda[\mathcal{B}_1] \geqslant \Xi_d \lambda[\mathcal{B}_1 \cap C_d].$$

The strong density condition is thus satisfied.

Using (C.8), and in particular that $\eta_{1,1} = \eta_{-1,1}$ or $\mathbb{P}_{1,Y|X,A=a}(y) = \mathbb{P}_{-1,Y|X,A=a}(y)$, the Kullback–Leibler divergence between $\mathbb{P}_1$ and $\mathbb{P}_{-1}$ can be expressed as

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_1, \mathbb{P}_{-1}) &= \int_{\mathcal{X}\times\mathcal{A}\times\mathcal{Y}} \log \frac{d\mathbb{P}_1(x,a,y)}{d\mathbb{P}_{-1}(x,a,y)} d\mathbb{P}_1(x,a,y) \\
&= \int_{\mathcal{X}\times\mathcal{A}\times\mathcal{Y}} \log \frac{d\mathbb{P}_{1,A}(a) d\mathbb{P}_{1,X|A=a}(x) d\mathbb{P}_{1,Y|X,A=a}(y)}{d\mathbb{P}_{-1,A}(a) d\mathbb{P}_{-1,X|A=a}(x) d\mathbb{P}_{-1,Y|X,A=a}(y)} d\mathbb{P}_1(x,a,y) \\
&= \sum_{a\in\{0,1\}} \mathbb{P}_1(A=a) \int_{\mathcal{X}} \log \frac{d\mathbb{P}_{1,X|A=a}(x)}{d\mathbb{P}_{-1,X|A=a}(x)} d\mathbb{P}_{1,X|A=a}(x).
\end{aligned}$$

Using the definition of $\mu_{j,a}$, and in particular that for all a, $\mu_{1,a} \neq \mu_{-1,a}$ only for $x \in \mathcal{B}_1 \cup \mathcal{B}_3$, as well as that $\lambda[\mathcal{B}_1] = \lambda[\mathcal{B}_3] = 1$, this further equals

$$\begin{aligned}
&\sum_{a\in\{0,1\}} \frac{1}{2} \int \log \frac{\mu_{1,a}(x)}{\mu_{-1,a}(x)} \mu_{1,a}(x) dx \\
&= \frac{1}{2} \int_{\mathcal{B}_1\cup\mathcal{B}_3} \left[\frac{1}{2} \log(1+s^d) - \frac{1}{2+2s^d} \log(1+s^d) \right] dx = \frac{s^d}{2+2s^d} \log(1+s^d) \leqslant C s^{2d}.
\end{aligned}$$

By Pinsker's inequality, we have

$$\mathrm{TV}\left(\mathbb{P}_1^{\otimes n}, \mathbb{P}_{-1}^{\otimes n}\right) \leqslant \frac{1}{2} \sqrt{\mathrm{KL}\left(\mathbb{P}_1^{\otimes n}, \mathbb{P}_{-1}^{\otimes n}\right)} = \frac{1}{2} \sqrt{n \mathrm{KL}\left(\mathbb{P}_1, \mathbb{P}_{-1}\right)} \leqslant C' \sqrt{n} s^d. \tag{C.19}$$

Recall that ν denotes the distribution of a Rademacher variable. We have, by (4.2),

$$\begin{aligned}
&\sup_{\mathbb{P}\in\mathcal{P}_\Sigma} \mathbb{E}_{\mathbb{P}} d_E(\widehat{f}_{\delta,n}, f^\star) \geqslant \sup_{j\in\{-1,1\}} \mathbb{E}_{\mathbb{P}_j}^{\otimes n} d_E\left(\widehat{f}_{\delta,n}, f_j^\star\right) \\
&= \sup_{j\in\{-1,1\}} \mathbb{E}_{\mathbb{P}_j}^{\otimes n} \sum_{a\in\{0,1\}} \left\{ \int_{\cup_{j=1}^3 \mathcal{B}_j} \left(\widehat{f}_{\delta,n}(x,a) - f_j^\star(x,a)\right) \left(T_{j,a}^\star - \eta_{j,a}(x)\right) \mu_{j,a}(x) dx \right\} \\
&\geqslant \sup_{j\in\{-1,1\}} \mathbb{E}_{\mathbb{P}_j}^{\otimes n} \sum_{a\in\{0,1\}} \left\{ \int_{\mathcal{B}_2} \left(\widehat{f}_{\delta,n}(x,a) - f_j^\star(x,a)\right) \left(T_{j,a}^\star - \eta_{j,a}(x)\right) \mu_{j,a}(x) dx \right\} \\
&\geqslant \mathbb{E}_\nu \mathbb{E}_{\mathbb{P}_\nu}^{\otimes n} \sum_{a\in\{0,1\}} \left\{ \int_{\mathcal{B}_2} \left(\widehat{f}_{\delta,n}(x,a) - f_\nu^\star(x,a)\right) \left(T_{\nu,a}^\star - \eta_{\nu,a}(x)\right) \mu_{\nu,a}(x) dx \right\}.
\end{aligned} \tag{C.20}$$

Here, the second to last inequality holds since $\left(\widehat{f}_{\delta,n}(x,a) - f_j^\star(x,a)\right) \left(T_{j,a}^\star - \eta_{j,a}(x)\right) \geqslant 0$, which follows the fact that $\widehat{f}_{\delta,n}(x,a) \in [0,1]$, while $f_j^\star(x,a) = 1$ when $\eta_{j,a}(x) > T_{j,a}^\star$ and $f_j^\star(x,a) = 0$ when $\eta_{j,a}(x) < T_{j,a}^\star$.

Now, define the distribution $\mathbb{P}_0 = (\mathbb{P}_1 + \mathbb{P}_{-1})/2$. We have that $\mathbb{P}_1^{\otimes n}$ and $\mathbb{P}_{-1}^{\otimes n}$ are absolutely continuous with respect to $\mathbb{P}_0^{\otimes n}$. Moreover, by (C.13),

$$\eta_{j,a}(x) = \begin{cases}
1/2 + (2a-1)/4 - c_\beta (s v_d)^{d/\gamma} - c_\beta (-x_1 - s v_d)^{d/\gamma}, & x \in \mathcal{B}_1; \\
1/2 + (2a-1)/4 - c_\beta (s v_d)^{d/\gamma} + c_\beta (x_1 + s v_d)^{d/\gamma}, & x \in \mathcal{B}_2 \cap \{x_1 < 0\}; \\
1/2 + (2a-1)/4 + c_\beta (s v_d)^{d/\gamma} - c_\beta (-x_1 + s v_d)^{d/\gamma}, & x \in \mathcal{B}_2 \cap \{x_1 \geqslant 0\}; \\
1/2 + (2a-1)/4 + c_\beta (s v_d)^{d/\gamma} + c_\beta (x_1 - s v_d)^{d/\gamma}, & x \in \mathcal{B}_3.
\end{cases}$$

In particular, we have, for $x \in \mathcal{B}_2$,

$$\eta_{j,a}(x) = 1/2 + (2a-1)/4 + \mathrm{sign}(x_1) c_\beta (s v_d)^{d/\gamma} - \mathrm{sign}(x_1) c_\beta (-|x_1| + s v_d)^{d/\gamma}. \tag{C.21}$$

and

$$1/2 + (2a-1)/4 - c_\beta (s v_d)^{d/\gamma} \leqslant \eta_{j,a}(x) \leqslant 1/2 + (2a-1)/4 + c_\beta (s v_d)^{d/\gamma}.$$

By (C.10), for $x \in \mathrm{int}(\mathcal{B}_2)$, we thus have

$$f^\star_{+1}(x,1) = f^\star_{-1}(x,0) = 0 \quad \text{and} \quad f^\star_{-1}(x,1) = f^\star_{+1}(x,0) = 1.$$

In other words, we have $f^\star_j(x,a) = I(j+1 \neq 2a)$ on $\mathrm{int}(\mathcal{B}_2)$. Moreover, since the boundary of $\mathcal{B}_2$ has zero measure, we can set $f^\star_j(x,a)$ arbitrarily on this boundary. Thus, (C.20) can be written and bounded as

$$\begin{aligned}
&\mathbb{E}_\nu \sum_{a\in\{0,1\}} \mathbb{E}^{\otimes n}_{\mathbb{P}_0}\left(\frac{d\mathbb{P}^{\otimes n}_\nu}{d\mathbb{P}^{\otimes n}_0}\right)\left\{\int_{\mathcal{B}_2}\left(\widehat{f}_{\delta,n}(x,a) - I(\nu+1\neq 2a)\right)\left(T^\star_{\nu,a} - \eta_{\nu,a}(x)\right)\mu_{\nu,a}(x)dx\right\} \\
&\geqslant \sum_{a\in\{0,1\}} \mathbb{E}^{\otimes n}_{\mathbb{P}_0}\left(\frac{d\mathbb{P}^{\otimes n}_1}{d\mathbb{P}^{\otimes n}_0}\wedge\frac{d\mathbb{P}^{\otimes n}_{-1}}{d\mathbb{P}^{\otimes n}_0}\right)\mathbb{E}_\nu\left\{\int_{\mathcal{B}_2}\left(\widehat{f}_{\delta,n}(x,a) - I(\nu+1\neq 2a)\right)\left(T^\star_{\nu,a} - \eta_{\nu,a}(x)\right)\mu_{\nu,a}(x)dx\right\} \\
&= \mathbb{E}_{\mathbb{P}_0}(1 - \mathrm{TV}\left(\mathbb{P}^{\otimes n}_1, \mathbb{P}^{\otimes n}_{-1}\right) \sum_{a\in\{0,1\}} \mathbb{E}_\nu\left\{\int_{B_2}\left(\widehat{f}_{\delta,n}(x,a) - I(\nu+1\neq 2a)\right)\left(T^\star_{\nu,a} - \eta_{\nu,a}(x)\right)\mu_{\nu,a}(x)dx\right\}.
\end{aligned} \tag{C.22}$$

Then, if we take $s \asymp n^{-1/(2d)}$ with $\mathrm{TV}\left(\mathbb{P}^{\otimes n}_1, \mathbb{P}^{\otimes n}_{-1}\right) \leqslant C'\sqrt{n}s^d < 1/2$, by (C.19), (C.22) is further lower bounded by

$$\begin{aligned}
&\frac{1}{2}\sum_{a\in\{0,1\}} \mathbb{E}_\nu\left\{\int_{B_2}\left(\widehat{f}_{\delta,n}(x,a) - I(\nu+1\neq 2a)\right)\left(T^\star_{\nu,a} - \eta_{\nu,a}(x)\right)\mu_{\nu,a}(x)dx\right\} \\
&= \frac{1}{4}\int_{B_2}\left(\widehat{f}_{\delta,n}(x,1) - I(1+1\neq 2)\right)\left(T^\star_{1,1} - \eta_{1,1}(x)\right)\mu_{1,1}(x)dx \\
&+ \frac{1}{4}\int_{B_2}\left(\widehat{f}_{\delta,n}(x,1) - I(-1+1\neq 2)\right)\left(T^\star_{-1,1} - \eta_{-1,1}(x)\right)\mu_{-1,1}(x)dx \\
&+ \frac{1}{4}\int_{B_2}\left(\widehat{f}_{\delta,n}(x,0) - I(1+1\neq 0)\right)\left(T^\star_{1,0} - \eta_{1,0}(x)\right)\mu_{1,0}(x)dx \\
&+ \frac{1}{4}\int_{B_2}\left(\widehat{f}_{\delta,n}(x,0) - I(-1+1\neq 0)\right)\left(T^\star_{-1,0} - \eta_{-1,0}(x)\right)\mu_{-1,0}(x)dx.
\end{aligned}$$

By (C.21) and the fact that $T^\star_{j,a} = 1/2 + (2a-1)(1/4 + j\cdot c_\beta(sv_d)^{d/\gamma})$,

$$\eta_{j,a} - T^\star_{j,a} = -c_\beta(2a-1)j(sv_d)^{d/\gamma} + c_\beta \mathrm{sign}(x_1)\left((sv_d)^{d/\gamma} - (sv_d - |x_1|)^{d/\gamma}\right).$$

Moreover, we have $\mu_{j,a}(x) = 1/(2+2s^d)$ for $x \in \mathcal{B}_2$. Thus, denoting $W_d = (sv_d)^{d/\gamma}$ and $h(x) = \mathrm{sign}(x_1)\left[(sv_d)^{d/\gamma} - (sv_d - |x|)^{d/\gamma}\right]$, this further equals

$$\begin{aligned}
&\frac{c_\beta}{8+8s^d}\left\{\int_{B_2}\widehat{f}_{\delta,n}(x,1)\left(W_d - h(x)\right)dx\right\} + \frac{c_\beta}{8+8s^d}\left\{\int_{B_2}\left(\widehat{f}_{\delta,n}(x,1) - 1\right)\left(-W_d - h(x)\right)dx\right\} \\
&+ \frac{c_\beta}{8+8s^d}\left\{\int_{\mathcal{B}_2}\left(\widehat{f}_{\delta,n}(x,0) - 1\right)\left(-W_d - h(x)\right)dx\right\} + \frac{c_\beta}{8+8s^d}\left\{\int_{\mathcal{B}_2}\widehat{f}_{\delta,n}(x,0)\left(W_d - h(x)\right)dx\right\} \\
&= \frac{c_\beta}{8+8s^d}\left\{\int_{B_2}\left(\widehat{f}_{\delta,n}(x,1)W_d - \widehat{f}_{\delta,n}(x,1)h(x) - \widehat{f}_{\delta,n}(x,1)W_d - \widehat{f}_{\delta,n}(x,1)h(x) + W_d + h(x)\right)dx\right\} \\
&+ \frac{c_\beta}{8+8s^d}\left\{\int_{\mathcal{B}_2}\left(-\widehat{f}_{\delta,n}(x,0)W_d - \widehat{f}_{\delta,n}(x,0)h(x) + W_d + h(x) + \widehat{f}_{\delta,n}(x,0)W_d - \widehat{f}_{\delta,n}(x,0)h(x)\right)dx\right\}.
\end{aligned}$$

Cancelling terms, this equals

$$\frac{c_\beta}{8+8s^d}\left\{\int_{B_2}\left(-2\widehat{f}_{\delta,n}(x,1)h(x) + W_d + h(x)\right)dx\right\}$$

$$+\frac{c_\beta}{8+8s^d}\left\{\int_{\mathcal{B}_2}\left(-2\widehat{f}_{\delta,n}(x,0)h(x)+W_d+h(x)\right)dx\right\}$$
$$=\frac{c_\beta}{4+4s^d}\left\{\int_{B_2}\left(-\left(\widehat{f}_{\delta,n}(x,1)+\widehat{f}_{\delta,n}(x,0)\right)h(x)+W_d+h(x)\right)dx\right\}. \tag{C.23}$$

Note that $0\leqslant\widehat{f}_{\delta,n}(x,a)\leqslant 1$ for $a\in\{0,1\}$. By definition of W_d and $h(x)$, we have, for $a\in\{0,1\}$,

$$W_d+h(x)-2\widehat{f}_{\delta,n}(x,a)h(x)=W_d+(1-2\widehat{f}_{\delta,n}(x,a))h(x)\geqslant W_d-|h(x)|$$
$$=(sv_d)^{d/\gamma}-\left((sv_d)^{d/\gamma}-(sv_d-|x_1|)^{d/\gamma}\right)=(sv_d-|x_1|)^{d/\gamma}.$$

Then, (C.23) is lower bounded by

$$\frac{c_\beta}{4+4s^d}\int_{\mathcal{B}_2}(sv_d-|x_1|)^{d/\gamma}dx=\frac{c_\beta}{2+2s^d}\int_{\mathcal{B}_2\cap\{x_1>0\}}(sv_d-x_1)^{d/\gamma}dx$$
$$=\frac{c_\beta}{2+2s^d}\int_0^{sv_d}\left(\int_{\sum_{j=2}^d|x_j|\leqslant sv_d-x_1}(sv_d-x_1)^{d/\gamma}dx_2\ldots dx_d\right)dx_1$$
$$=\frac{c_\beta}{2+2s^d}\int_0^{sv_d}(sv_d-x_1)^{d/\gamma}\left(\int_{\sum_{j=2}^d|x_j|\leqslant sv_d-x_1}dx_2\ldots dx_d\right)dx_1$$
$$=\frac{c_\beta V_{d-1,1}}{2+2s^d}\int_0^{sv_d}(sv_d-x_1)^{d/\gamma}(sv_d-x_1)^{d-1}dx_1$$
$$=\frac{c_\beta\gamma V_{d-1,1}}{(2+2s^d)(1+\gamma)d}(sv_d)^{d/\gamma+d}\geqslant C_{\beta,\gamma,d}s^{\frac{d}{\gamma}+d}\geqslant C'n^{-\frac{1+\gamma}{2\gamma}}.$$

In the last inequality, we have used that $s\asymp n^{-1/(2d)}$. This finishes the proof.

D Proofs of Theorems in Section 5

The following counterexample from Rigollet and Vert (2009) demonstrates the advantage of offsets. Assume that $\mathcal{X}\subset\mathbb{R}$, that the density g is such that $g(x)=1/2$ for $x\in[0,1]$ and that $g(x)<1/2$ elsewhere. Assume $\widehat{g}$ is an consistent estimator of g such that $\|\widehat{g}-g\|_\infty\leqslant\varepsilon$ for some small $\varepsilon>0$. If $\widehat{g}(x)=g(x)+\varepsilon$ for $x\in[0,1]$, we have $\Lambda_f(1/2)=\emptyset$ and $\widehat{\Lambda}_f(1/2)\supset[0,1]$. Thus, the standard plug-in estimate fails to estimate $\Lambda_f(1/2)$ consistently even as ε tends to 0. However, $\widetilde{\Lambda}_{g,\ell_n}(1/2)$ with a positive offset $\ell_n>\varepsilon$ can become consistent.

D.1 Proof of Proposition 5.2

- In the fairness-impacted case, we have $D_-(0)>0$ and $\widetilde{\delta}=\delta$. Recalling (3.3), (3.4), (3.5) and (5.10), we have, by the left continuity of D_+ and the right continuity of D_-,

$$D_+(t^\star_\delta)=\mathbb{P}_{X|A=1}\left(\eta_1(X)\geqslant\frac{1}{2}+\frac{t^\star_\delta}{2p_1}\right)-\mathbb{P}_{X|A=0}\left(\eta_0(X)>\frac{1}{2}-\frac{t^\star_\delta}{2p_0}\right)$$
$$=\mathbb{P}_{X|A=1}\left(\eta_1(X)>T^\star_{\delta,1}\right)+\mathbb{P}_{X|A=1}\left(\eta_1(X)=T^\star_{\delta,1}\right)-\mathbb{P}_{X|A=0}\left(\eta_0(X)>T^\star_{\delta,0}\right)$$
$$=\pi^\star_{1,+}+\pi^\star_{1,=}-\pi^\star_{0,+}\geqslant\delta,$$

and

$$D_-(t^\star_\delta)=\mathbb{P}_{X|A=1}\left(\eta_1(X)>\frac{1}{2}+\frac{t^\star_\delta}{2p_1}\right)-\mathbb{P}_{X|A=0}\left(\eta_0(X)\geqslant\frac{1}{2}-\frac{t^\star_\delta}{2p_0}\right)$$
$$=\mathbb{P}_{X|A=1}\left(\eta_1(X)>T^\star_{\delta,1}\right)-\mathbb{P}_{X|A=0}\left(\eta_0(X)>T^\star_{\delta,0}\right)-\mathbb{P}_{X|A=0}\left(\eta_0(X)=T^\star_{\delta,0}\right)$$
$$=\pi^\star_{1,+}-\pi^\star_{0,+}-\pi^\star_{0,=}\leqslant\delta.$$

It follows that

$$\delta - \pi^\star_{1,+} + \pi^\star_{0,+} = \delta - D_+(t^\star_\delta) + \pi^\star_{1,=} \leqslant \pi^\star_{1,=},$$

and thus, and interpreting $x/0 = 0$ for all $x \in \mathbb{R}$ in what follows, $\frac{\pi^\star_{0,+} - \pi^\star_{1,+} + \delta}{\pi^\star_{1,=}} \leqslant 1$. Similarly, $\pi^\star_{1,+} - \pi^\star_{0,+} - \delta = D_-(t^\star_\delta) - \delta + \pi^\star_{0,=} \leqslant \pi^\star_{0,=}$ and $\frac{\pi^\star_{1,+} - \pi^\star_{0,+} - \delta}{\pi^\star_{0,=}} \leqslant 1$. As a result, with $x \mapsto \sigma(x) = \max(x, 0)$,

$$\begin{aligned}
\mathrm{DDP}(f^\star_\delta) &= \sum_{a\in\{0,1\}} \left[(2a-1)\left(\mathbb{P}_{X|A=a}\left(\eta_a(X) > T^\star_{\delta,a}\right) + \tau^\star_{\delta,a}\mathbb{P}_{X|A=a}\left(\eta_a(X) = T^\star_{\delta,a}\right)\right)\right] \\
&= \pi^\star_{1,+} + \tau^\star_{\delta,1}\cdot\pi^\star_{1,=} - \pi^\star_{0,+} - \tau^\star_{\delta,0}\cdot\pi^\star_{0,=} \\
&= \pi^\star_{1,+} + \pi^\star_{1,=}\rho\left(\frac{\pi^\star_{0,+} - \pi^\star_{1,+} + \delta}{\pi^\star_{1,=}}\right) - \pi^\star_{0,+} - \pi^\star_{0,=}\rho\left(\frac{\pi^\star_{1,+} - \pi^\star_{0,+} - \delta}{\pi^\star_{0,=}}\right) \\
&= \pi^\star_{1,+} + \sigma\left(\pi^\star_{0,+} - \pi^\star_{1,+} + \delta\right) - \pi^\star_{0,+} - \sigma\left(\pi^\star_{1,+} - \pi^\star_{0,+} - \delta\right) = \delta,
\end{aligned}$$

were the last equation follows by checking the cases $\pi^\star_{0,+} - \pi^\star_{1,+} + \delta \geqslant 0$ and $\pi^\star_{0,+} - \pi^\star_{1,+} + \delta < 0$.

- In the automatically-fair and fair-boundary cases, we have $D_-(0) \leqslant 0$ and $\widetilde{\delta} = 0$.
  - If $D_-(0) \leqslant 0$, we have,

    $$-\pi^\star_{1,=} \leqslant D_+(0) - \pi^\star_{1,=} = \pi^\star_{1,+} - \pi^\star_{0,+} = D_-(0) + \pi^\star_{0,=} \leqslant \pi^\star_{0,=}.$$

    It holds that $(\pi^\star_{0,+} - \pi^\star_{1,+})/\pi^\star_{1,=} \leqslant 1$ and $(\pi^\star_{1,+} - \pi^\star_{0,+})/\pi^\star_{0,=} \leqslant 1$. As a result, with $x \mapsto \sigma(x) = \max(x, 0)$,

    $$\begin{aligned}
    \mathrm{DDP}(f^\star_\delta) &= \pi^\star_{1,+} + \pi^\star_{1,=}\rho\left(\frac{\pi^\star_{0,+} - \pi^\star_{1,+}}{\pi^\star_{1,=}}\right) - \pi^\star_{0,+} - \pi^\star_{0,=}\rho\left(\frac{\pi^\star_{1,+} - \pi^\star_{0,+}}{\pi^\star_{0,=}}\right) \\
    &= \pi^\star_{1,+} + \sigma\left(\pi^\star_{0,+} - \pi^\star_{1,+}\right) - \pi^\star_{0,+} - \pi^\star_{0,=}\sigma\left(\pi^\star_{1,+} - \pi^\star_{0,+}\right) = 0,
    \end{aligned}$$

    were the last equation follows by checking the cases $\pi^\star_{0,+} - \pi^\star_{1,+} \geqslant 0$ and $\pi^\star_{0,+} - \pi^\star_{1,+} < 0$.
  - If $D_-(0) > 0$, we have,

    $$\pi^\star_{1,+} - \pi^\star_{0,+} = D_-(0) + \pi^\star_{0,=} > \pi^\star_{0,=}.$$

    It holds that $(\pi^\star_{0,+} - \pi^\star_{1,+})/\pi^\star_{1,=} \leqslant 0$ and $(\pi^\star_{1,+} - \pi^\star_{0,+})/\pi^\star_{0,=} > 1$. As a result, with $x \mapsto \sigma(x) = \max(x, 0)$,

    $$\begin{aligned}
    \mathrm{DDP}(f^\star_\delta) &= \pi^\star_{1,+} + \pi^\star_{1,=}\rho\left(\frac{\pi^\star_{0,+} - \pi^\star_{1,+}}{\pi^\star_{1,=}}\right) - \pi^\star_{0,+} - \pi^\star_{0,=}\rho\left(\frac{\pi^\star_{1,+} - \pi^\star_{0,+}}{\pi^\star_{0,=}}\right) \\
    &= \pi^\star_{1,+} + -\pi^\star_{0,+} - \pi^\star_{0,=} = D_-(0).
    \end{aligned}$$

E Proofs of Theorems in Section 6

E.1 Proof of Theorem 6.2

Recall the definition of $\widehat{t}_{\delta,\mathrm{mid}}$, $\widehat{t}_{\delta,\Delta_n,\mathrm{min}}$ and $\widehat{t}_{\delta,\Delta_n,\mathrm{max}}$ from (5.3). By construction, we have for any $\varepsilon > 0$ that

$$\begin{aligned}
&\left\{\widehat{t}_{\delta,\mathrm{mid}} > t^\star_\delta + \varepsilon\right\} \subset \left\{\widehat{D}_n(t^\star_\delta + \varepsilon, 0, 0) \geqslant \delta\right\}; && \left\{\widehat{t}_{\delta,\mathrm{mid}} < t^\star_\delta + \varepsilon\right\} \subset \left\{\widehat{D}_n(t^\star_\delta + \varepsilon, 0, 0) \leqslant \delta\right\}; \\
&\left\{\widehat{t}_{\delta,\Delta_n,\mathrm{min}} > t^\star_\delta + \varepsilon\right\} \subset \left\{\widehat{D}_n(t^\star_\delta + \varepsilon, 0, 0) \geqslant \delta + \Delta_n\right\}; && \left\{\widehat{t}_{\delta,\Delta_n,\mathrm{min}} < t^\star_\delta + \varepsilon\right\} \subset \left\{\widehat{D}_n(t^\star_\delta + \varepsilon, 0, 0) \leqslant \delta + \Delta_n\right\}; \\
&\left\{\widehat{t}_{\delta,\Delta_n,\mathrm{max}} > t^\star_\delta + \varepsilon\right\} \subset \left\{\widehat{D}_n(t^\star_\delta + \varepsilon, 0, 0) \geqslant \delta - \Delta_n\right\}; && \left\{\widehat{t}_{\delta,\Delta_n,\mathrm{max}} < t^\star_\delta + \varepsilon\right\} \subset \left\{\widehat{D}_n(t^\star_\delta + \varepsilon, 0, 0) \leqslant \delta - \Delta_n\right\}.
\end{aligned}$$

Let L_{t_1}, U_{t_1} and c_{i,t_1} be defined in Lemma B.9; let L_{t_2}, U_{t_2} and c_{i,t_2} be defined in Lemma B.10; and let L_r, U_r, $U_{\Delta,r}$, and $c_{i,r}$ be defined in Lemma B.11. During this proof, we take $L_t = L_{t_1} \vee L_{t,2} \vee L_r$ and $U_t = U_{t_1} \wedge U_{t,2} \wedge U_r$. We further denote

$$\mathcal{E}_t = \psi_{n,1,t}(\varepsilon) + \sum_{j=\{-1,1\}} I(\delta = D_j(t_\delta^\star))\psi_{n,2,t}\left(g_{\delta,j}\left(\omega(\varepsilon, r_n)\right)\right).$$

Next, we consider the following four cases: (1) $t_\delta^\star = 0$, (2) $t_\delta^\star > 0$ and $D_+(t_\delta^\star) > \delta$, (3) $t_\delta^\star > 0$ and $D_+(t_\delta^\star) = \delta > D_-(t_\delta^\star)$ and (4) $t_\delta^\star > 0$ and $D_+(t_\delta^\star) = \delta = D_-(t_\delta^\star)$.

Case (1) $t_\delta^\star = 0$.

In this case, we have $\widetilde{I}^\star(\delta) = 0$ and $\omega(\varepsilon, r_n) = r_n$. By construction, we have $\widehat{t}_\delta = \widehat{t}_{\delta,\Delta_n,\min}$ when $\widehat{t}_{\delta,\mathrm{mid}} - \widehat{t}_{\delta,\Delta_n,\min} \leqslant r_n$. By (B.7) of Lemma B.9, (B.13) of Lemma B.11 and Lemma B.8, when $L_t(\phi_{n,1} \vee \phi_{n,0}) < r_n < U_t$, $2(g_{\delta,-}(4r_n) \vee g_{\delta,+}(4r_n)) < \Delta_n < U_{\Delta,t} := U_{\Delta,r}$ and $L_t(\phi_{n,1} \vee \phi_{n,0}) < \varepsilon < r_n$, we have, with $c_{i,t} > 0, i \in [4]$ that

$$\begin{aligned}
&\mathbb{P}^{\otimes n}\left(\widehat{t}_\delta > t_\delta^\star + \varepsilon\right) = \mathbb{P}^{\otimes n}\left(\widehat{t}_\delta > \varepsilon\right) = \mathbb{P}^{\otimes n}\left(\widehat{t}_\delta > \varepsilon, \widehat{t}_\delta = \widehat{t}_{\delta,\Delta_n,\min}\right) + \mathbb{P}^{\otimes n}\left(\widehat{t}_\delta > \varepsilon, \widehat{t}_\delta \neq \widehat{t}_{\delta,\Delta_n,\min}\right) \\
&\leqslant \mathbb{P}^{\otimes n}\left(\widehat{t}_{\delta,\Delta_n,\min} > \varepsilon\right) + \mathbb{P}^{\otimes n}\left(\widehat{t}_\delta \neq \widehat{t}_{\delta,\Delta_n,\min}\right) \\
&\leqslant \mathbb{P}^{\otimes n}\left(\widehat{D}_n(t_\delta^\star + \varepsilon, 0, 0) \geqslant \delta + \Delta_n\right) + \mathbb{P}^{\otimes n}\left(\widehat{t}_{\delta,\mathrm{mid}} - \widehat{t}_{\delta,\Delta_n,\min} > r_n\right) \\
&\leqslant \psi_{n,1,t_1}(\varepsilon) + I(\delta = D_-(t_\delta^\star))\psi_{n,2,t_1}\left(\Delta_n + g_{\delta,-}\left(\frac{\varepsilon}{2}\right)\right) + \psi_{n,1,r}(r_n) + I(\delta = D_-(t_\delta^\star))\psi_{n,2,r}\left(g_{\delta,-}\left(\frac{r_n}{2}\right)\right) \\
&\leqslant \psi_{n,1,t_1}(\varepsilon) + I(\delta = D_-(t_\delta^\star))\psi_{n,2,t_1}\left(g_{\delta,-}\left(\frac{r_n}{4}\right)\right) + \psi_{n,1,r}(r_n) + I(\delta = D_-(t_\delta^\star))\psi_{n,2,r}\left(g_{\delta,-}\left(\frac{r_n}{4}\right)\right) \\
&\leqslant \psi_{n,1,t}(\varepsilon) + I(\delta = D_-(t_\delta^\star))\psi_{n,2,t}\left(g_{\delta,-}\left(r_n\right)\right) \leqslant \mathcal{E}_t.
\end{aligned}$$

Here, the third last inequality holds since $\Delta_n + g_{\delta,-}(\varepsilon/2) > \Delta_n > 4g_{\delta,-}(4r_n) > g_{\delta,-}(r_n/4)$.

Moreover, by definition, $\mathbb{P}^{\otimes n}\left(\widehat{t}_\delta < t_\delta^\star - \varepsilon\right) \leqslant \mathbb{P}^{\otimes n}\left(\widehat{t}_\delta < 0\right) = 0 \leqslant \mathcal{E}_t$.

Case (2) $t_\delta^\star > 0$ and $\delta < D_+(t_\delta^\star)$.

In this case, we have $\widetilde{I}^\star(\delta) = 0$ and $\omega(\varepsilon, r_n) = r_n$. By construction, we have $\widehat{t}_\delta = \widehat{t}_{\delta,\Delta_n,\min}$ when $\widehat{t}_{\delta,\mathrm{mid}} - \widehat{t}_{\delta,\Delta_n,\min} \leqslant r_n$. By (B.7) of Lemma B.9, (B.14) of Lemma B.11 and Lemma B.8, when $L_t(\phi_{n,1} \vee \phi_{n,0}) < r_n < U_t$, $2(g_{\delta,-}(4r_n) \vee g_{\delta,+}(4r_n)) < \Delta_n < U_{\Delta,t} := U_{\Delta,r}$ and $L_t(\phi_{n,1} \vee \phi_{n,0}) < \varepsilon < r_n$, we have, with $c_{i,t} > 0, i \in [4]$ that

$$\begin{aligned}
&\mathbb{P}^{\otimes n}\left(\widehat{t}_\delta > t_\delta^\star + \varepsilon\right) = \mathbb{P}^{\otimes n}\left(\widehat{t}_\delta > t_\delta^\star + \varepsilon, \widehat{t}_\delta = \widehat{t}_{\delta,\Delta_n,\min}\right) + \mathbb{P}^{\otimes n}\left(\widehat{t}_\delta > t_\delta^\star + \varepsilon, \widehat{t}_\delta \neq \widehat{t}_{\delta,\Delta_n,\min}\right) \\
&\leqslant \mathbb{P}^{\otimes n}\left(\widehat{t}_{\delta,\Delta_n,\min} > t_\delta^\star + \varepsilon\right) + \mathbb{P}^{\otimes n}\left(\widehat{t}_\delta \neq \widehat{t}_{\delta,\Delta_n,\min}\right) \\
&\leqslant \mathbb{P}^{\otimes n}\left(\widehat{D}_n(t_\delta^\star + \varepsilon, 0, 0) \geqslant \delta + \Delta_n\right) + \mathbb{P}^{\otimes n}\left(\widehat{t}_{\delta,\mathrm{mid}} - \widehat{t}_{\delta,\Delta_n,\min} > r_n\right) \\
&\leqslant \psi_{n,1,t_1}(\varepsilon) + I(\delta = D_-(t_\delta^\star))\psi_{n,2,t_1}\left(\Delta_n + g_{\delta,-}\left(\frac{\varepsilon}{2}\right)\right) + \psi_{n,1,r}\left(\frac{r_n}{2}\right) + I(\delta = D_-(t_\delta^\star))\psi_{n,2,r}\left(g_{\delta,-}\left(\frac{r_n}{4}\right)\right) \\
&\leqslant \psi_{n,1,t_1}(\varepsilon) + I(\delta = D_-(t_\delta^\star))\psi_{n,2,t_1}\left(g_{\delta,-}\left(\frac{r_n}{4}\right)\right) + \psi_{n,1,r}\left(\varepsilon\right) + I(\delta = D_-(t_\delta^\star))\psi_{n,2,r}\left(g_{\delta,-}\left(\frac{r_n}{4}\right)\right) \\
&\leqslant \psi_{n,1,t}(\varepsilon) + I(\delta = D_-(t_\delta^\star))\psi_{n,2,t}\left(g_{\delta,-}\left(r_n\right)\right) \leqslant \mathcal{E}_t.
\end{aligned}$$

Here, the third last inequality holds since $\Delta_n - g_{\delta,-}(\varepsilon/2) > \Delta_n - g_{\delta,-}(4r_n) > g_{\delta,-}(4r_n) \geqslant g_{\delta,-}(r_n/4)$.

On the other hand, by (B.12) of Lemma B.10, (B.16) of Lemma B.11 and Lemma B.8, when $L_t(\phi_{n,1} \vee \phi_{n,0}) < r_n < U_t$, $2(g_{\delta,-}(4r_n) \vee g_{\delta,+}(4r_n)) < \Delta_n < U_{\Delta,t} := U_{\Delta,r} \wedge ((D_+(t_\delta^\star) - \delta)/2)$ and $L_t(\phi_{n,1} \vee \phi_{n,0}) < \varepsilon < r_n$, we have, with $c_{i,t} > 0, i \in [4]$ that

$$\begin{aligned}
&\mathbb{P}^{\otimes n}\left(\widehat{t}_\delta < t_\delta^\star - \varepsilon\right) = \mathbb{P}^{\otimes n}\left(\widehat{t}_\delta < t_\delta^\star - \varepsilon, \widehat{t}_\delta = \widehat{t}_{\delta,\Delta_n,\min}\right) + \mathbb{P}^{\otimes n}\left(\widehat{t}_\delta < t_\delta^\star - \varepsilon, \widehat{t}_\delta \neq \widehat{t}_{\delta,\Delta_n,\min}\right) \\
&\leqslant \mathbb{P}^{\otimes n}\left(\widehat{t}_{\delta,\Delta_n,\min} < t_\delta^\star - \varepsilon\right) + \mathbb{P}^{\otimes n}\left(\widehat{t}_\delta \neq \widehat{t}_{\delta,\Delta_n,\min}\right) \\
&\leqslant \mathbb{P}^{\otimes n}\left(\widehat{D}_n(t_\delta^\star - \varepsilon, 0, 0) \leqslant \delta + \Delta_n\right) + \mathbb{P}^{\otimes n}\left(\widehat{t}_{\delta,\mathrm{mid}} - \widehat{t}_{\delta,\Delta_n,\min} > r_n\right)
\end{aligned}$$

$$\leqslant \psi_{n,1,t_2}(\varepsilon) + \psi_{n,1,r}\left(\frac{r_n}{2}\right) + I(\delta = D_-(t_\delta^\star))\psi_{n,2,r}\left(g_{\delta,-}\left(\frac{r_n}{4}\right)\right)$$
$$\leqslant \psi_{n,1,t_1}(\varepsilon) + \psi_{n,1,r}(\varepsilon) + I(\delta = D_-(t_\delta^\star))\psi_{n,2,r}\left(g_{\delta,-}\left(\frac{r_n}{4}\right)\right)$$
$$\leqslant \psi_{n,1,t}(\varepsilon) + I(\delta = D_-(t_\delta^\star))\psi_{n,2,t}\left(g_{\delta,-}(r_n)\right) \leqslant \mathcal{E}_t.$$

Case (3) $t_\delta^\star > 0$ and $D_-(t_\delta^\star) < \delta = D_+(t_\delta^\star)$.

Again, we have $\widetilde{I}^\star(\delta) = 0$ and $\omega(\varepsilon, r_n) = r_n$. By construction, we have $\widehat{t}_\delta = \widehat{t}_{\delta,\Delta_n,\max}$ when $\widehat{t}_{\delta,\text{mid}} - \widehat{t}_{\delta,\Delta_n,\min} > r_n$ and $\widehat{t}_{\delta,\Delta_n,\max} - \widehat{t}_{\delta,\text{mid}} < r_n$. By (B.10) of Lemma B.10, (B.15), (B.16) of Lemma B.11 and Lemma B.8, when $L_t(\phi_{n,1} \vee \phi_{n,0}) < r_n < U_t$, $2(g_{\delta,-}(4r_n) \vee g_{\delta,+}(4r_n)) < \Delta_n < U_{\Delta,t} := U_{\Delta,r} \wedge ((\delta - D_-(t_\delta^\star)/2)$ and $L_t(\phi_{n,1} \vee \phi_{n,0}) < \varepsilon < r_n$, we have, with $c_{i,t} > 0, i \in [4]$ that

$$\begin{aligned}
&\mathbb{P}^{\otimes n}\left(\widehat{t}_\delta > t_\delta^\star + \varepsilon\right) = \mathbb{P}^{\otimes n}\left(\widehat{t}_\delta > t_\delta^\star + \varepsilon, \widehat{t}_\delta = \widehat{t}_{\delta,\Delta_n,\max}\right) + \mathbb{P}^{\otimes n}\left(\widehat{t}_\delta > t_\delta^\star + \varepsilon, \widehat{t}_\delta \neq \widehat{t}_{\delta,\Delta_n,\max}\right)\\
&\leqslant \mathbb{P}^{\otimes n}\left(\widehat{t}_{\delta,\Delta_n,\max} > t_\delta^\star + \varepsilon\right) + \mathbb{P}^{\otimes n}\left(\widehat{t}_\delta \neq \widehat{t}_{\delta,\Delta_n,\max}\right)\\
&\leqslant \mathbb{P}^{\otimes n}\left(\widehat{D}_n(t_\delta^\star + \varepsilon, 0, 0) \geqslant \delta - \Delta_n\right) + \mathbb{P}^{\otimes n}\left(\widehat{t}_{\delta,\text{mid}} - \widehat{t}_{\delta,\Delta_n,\min} \leqslant r_n\right) + \mathbb{P}^{\otimes n}\left(\widehat{t}_{\delta,\Delta_n,\max} - \widehat{t}_{\delta,\text{mid}} > r_n\right)\\
&\leqslant \psi_{n,1,t_2}(\varepsilon) + \psi_{n,1,r}(r_n) + \psi_{n,2,r}\left(g_{\delta,+}\left(\frac{r_n}{2}\right)\right) + \psi_{n,1,r}\left(\frac{r_n}{2}\right) + \psi_{n,2,r}\left(g_{\delta,+}\left(\frac{r_n}{4}\right)\right)\\
&\leqslant \psi_{n,1,t_2}(\varepsilon) + 2\psi_{n,1,r}\left(\frac{r_n}{2}\right) + 2\psi_{n,2,r}\left(g_{\delta,+}\left(\frac{r_n}{4}\right)\right) \leqslant \psi_{n,1,t}(\varepsilon) + \psi_{n,2,t}\left(g_{\delta,+}(r_n)\right) \leqslant \mathcal{E}_t.
\end{aligned}$$

Here, the third last inequality holds since $\Delta_n - g_{\delta,-}(2\varepsilon) > \Delta_n - g_{\delta,-}(4r_n) > g_{\delta,+}(4r_n) \geqslant g_{\delta,+}(r_n/4)$.

On the other hand, by (B.8) of Lemma B.9, (B.15), (B.16) of Lemma B.11 and Lemma B.8, when $L_t(\phi_{n,1} \vee \phi_{n,0}) < r_n < U_t$, $2(g_{\delta,-}(4r_n) \vee g_{\delta,+}(4r_n)) < \Delta_n < U_{\Delta,t} := U_{\Delta,r}$ and $L_t(\phi_{n,1} \vee \phi_{n,0}) < \varepsilon < r_n$, we have, with $c_{i,t} > 0, i \in [4]$ that

$$\begin{aligned}
&\mathbb{P}^{\otimes n}\left(\widehat{t}_\delta < t_\delta^\star - \varepsilon\right) = \mathbb{P}^{\otimes n}\left(\widehat{t}_\delta < t_\delta^\star - \varepsilon, \widehat{t}_\delta = \widehat{t}_{\delta,\Delta_n,\max}\right) + \mathbb{P}^{\otimes n}\left(\widehat{t}_\delta < t_\delta^\star - \varepsilon, \widehat{t}_\delta \neq \widehat{t}_{\delta,\Delta_n,\max}\right)\\
&\leqslant \mathbb{P}^{\otimes n}\left(\widehat{t}_{\delta,\Delta_n,\max} < t_\delta^\star - \varepsilon\right) + \mathbb{P}^{\otimes n}\left(\widehat{t}_\delta \neq \widehat{t}_{\delta,\Delta_n,\max}\right)\\
&\leqslant \mathbb{P}^{\otimes n}\left(\widehat{D}_n(t_\delta^\star - \varepsilon, 0, 0) \leqslant \delta - \Delta_n\right) + \mathbb{P}^{\otimes n}\left(\widehat{t}_{\delta,\text{mid}} - \widehat{t}_{\delta,\Delta_n,\min} \leqslant r_n\right) + \mathbb{P}^{\otimes n}\left(\widehat{t}_{\delta,\Delta_n,\max} - \widehat{t}_{\delta,\text{mid}} > r_n\right)\\
&\leqslant \psi_{n,1,t_1}(\varepsilon) + \psi_{n,2,t_1}\left(\Delta_n + g_{\delta,+}\left(\frac{\varepsilon}{2}\right)\right) + \psi_{n,1,r}(r_n) + \psi_{n,2,r}\left(g_{\delta,+}\left(\frac{r_n}{2}\right)\right) + \psi_{n,1,r}\left(\frac{r_n}{2}\right)\\
&+ \psi_{n,2,r}\left(g_{\delta,+}\left(\frac{r_n}{4}\right)\right)\\
&\leqslant \psi_{n,1,t_1}(\varepsilon) + \psi_{n,2,t_1}\left(g_{\delta,+}\left(\frac{r_n}{4}\right)\right) + 2\psi_{n,1,r}\left(\frac{r_n}{2}\right) + 2\psi_{n,2,r}\left(g_{\delta,+}\left(\frac{r_n}{4}\right)\right) \leqslant \psi_{n,1,t}(\varepsilon) + \psi_{n,2,t}\left(g_{\delta,-}(r_n)\right) \leqslant \mathcal{E}_t.
\end{aligned}$$

Case (4) $t_\delta^\star > 0$ and $D_-(t_\delta^\star) = \delta = D_+(t_\delta^\star)$.

In this case, we have $\widetilde{I}^\star(\delta) = 1$ and $\omega(\varepsilon, r_n) = \varepsilon$. By construction, we have $\widehat{t}_\delta = \widehat{t}_{\delta,\text{mid}}$ when $\min(\widehat{t}_{\delta,\text{mid}} - \widehat{t}_{\delta,\Delta_n,\min}, \widehat{t}_{\delta,\Delta_n,\max} - \widehat{t}_{\delta,\text{mid}}) > r_n$. By (B.7), (B.8) of Lemma B.9, (B.17), (B.18) of Lemma B.11 and Lemma B.8, when $L_t(\phi_{n,1} \vee \phi_{n,0}) < r_n < U_t$, $2(g_{\delta,-}(4r_n) \vee g_{\delta,+}(4r_n)) < \Delta_n < U_{\Delta,t} := U_{\Delta,r}$ and $L_t(\phi_{n,1} \vee \phi_{n,0}) < \varepsilon < r_n$, we have, with $c_{i,t} > 0, i \in [4]$ that,

$$\begin{aligned}
&\mathbb{P}^{\otimes n}\left(\widehat{t}_\delta > t_\delta^\star + \varepsilon\right) = \mathbb{P}^{\otimes n}\left(\widehat{t}_\delta > t_\delta^\star + \varepsilon, \widehat{t}_\delta = \widehat{t}_{\delta,\text{mid}}\right) + \mathbb{P}^{\otimes n}\left(\widehat{t}_\delta > t_\delta^\star + \varepsilon, \widehat{t}_\delta \neq \widehat{t}_{\delta,\text{mid}}\right)\\
&\leqslant \mathbb{P}^{\otimes n}\left(\widehat{t}_{\delta,\text{mid}} > t_\delta^\star + \varepsilon\right) + \mathbb{P}^{\otimes n}\left(\widehat{t}_\delta \neq \widehat{t}_{\delta,\text{mid}}\right)\\
&\leqslant \mathbb{P}^{\otimes n}\left(\widehat{D}_n(t_\delta^\star + \varepsilon, 0, 0) \geqslant \delta\right) + \mathbb{P}^{\otimes n}\left(\widehat{t}_{\delta,\text{mid}} - \widehat{t}_{\delta,\Delta_n,\min} \leqslant r_n\right) + \mathbb{P}^{\otimes n}\left(\widehat{t}_{\delta,\Delta_n,\max} - \widehat{t}_{\delta,\text{mid}} \leqslant r_n\right)\\
&\leqslant \psi_{n,1,t_1}(\varepsilon) + \psi_{n,2,t_1}\left(g_{\delta,-}\left(\frac{\varepsilon}{2}\right)\right) + 2\psi_{n,1,r}(r_n) + \psi_{n,2,r}\left(g_{\delta,+}\left(\frac{r_n}{2}\right)\right) + \psi_{n,2,r}\left(g_{\delta,-}\left(\frac{r_n}{2}\right)\right)\\
&\leqslant \psi_{n,1,t}(\varepsilon) + \psi_{n,2,t}\left(g_{\delta,-}(\varepsilon)\right) + \psi_{n,2,t}\left(g_{\delta,+}(\varepsilon)\right) \leqslant \mathcal{E}_t.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\mathbb{P}^{\otimes n}\left(\widehat{t}_\delta < t_\delta^\star - \varepsilon\right) = \mathbb{P}^{\otimes n}\left(\widehat{t}_\delta < t_\delta^\star - \varepsilon, \widehat{t}_\delta = \widehat{t}_{\delta,\text{mid}}\right) + \mathbb{P}^{\otimes n}\left(\widehat{t}_\delta < t_\delta^\star - \varepsilon, \widehat{t}_\delta \neq \widehat{t}_{\delta,\text{mid}}\right)\\
&\leqslant \mathbb{P}^{\otimes n}\left(\widehat{t}_{\delta,\text{mid}} < t_\delta^\star - \varepsilon\right) + \mathbb{P}^{\otimes n}\left(\widehat{t}_\delta \neq \widehat{t}_{\delta,\text{mid}}\right)
\end{aligned}$$

$$\leqslant \mathbb{P}^{\otimes n}\left(\widehat{D}_n(t^\star_\delta-\varepsilon,0,0)\leqslant\delta\right)+\mathbb{P}^{\otimes n}\left(\widehat{t}_{\delta,\mathrm{mid}}-\widehat{t}_{\delta,\Delta_n,\min}\leqslant r_n\right)+\mathbb{P}^{\otimes n}\left(\widehat{t}_{\delta,\Delta_n,\max}-\widehat{t}_{\delta,\mathrm{mid}}\leqslant r_n\right)$$
$$\leqslant \psi_{n,1,t_1}(\varepsilon)+\psi_{n,2,t_1}\left(g_{\delta,+}\left(\frac{\varepsilon}{2}\right)\right)+2\psi_{n,1,r}(r_n)+\psi_{n,2,r}\left(g_{\delta,+}\left(\frac{r_n}{2}\right)\right)+\psi_{n,2,r}\left(g_{\delta,-}\left(\frac{r_n}{2}\right)\right)$$
$$\leqslant \psi_{n,1,t}(\varepsilon)+\psi_{n,2,t}\left(g_{\delta,-}\left(\varepsilon\right)\right)+\psi_{n,2,t}\left(g_{\delta,+}\left(\varepsilon\right)\right)\leqslant\mathcal{E}_t.$$

E.2 Proof of Theorem 6.4

By definition, we have

$$\widehat{f}_{\delta,n}(x,a)=\begin{cases}1, & \left\{(x,a):\widehat{\eta}_a(x)>\widehat{T}_{\delta,a}+\ell_{n,a}\right\},\\ \widehat{\tau}_{\delta,a}, & \left\{(x,a):\left|\widehat{\eta}_a(x)-\widehat{T}_{\delta,a}\right|<\ell_{n,a}\right\},\\ 0, & \left\{(x,a):\widehat{\eta}_a(x)<\widehat{T}_{\delta,a}-\ell_{n,a}\right\};\end{cases}\quad\text{and } f^\star_\delta(x,a)=\begin{cases}1, & \left\{(x,a):\eta_a(x)>T^\star_{\delta,a}\right\},\\ \tau^\star_{\delta,a}, & \left\{(x,a):\eta_a(x)=T^\star_{\delta,a}\right\},\\ 0, & \left\{(x,a):\eta_a(x)<T^\star_{\delta,a}\right\}.\end{cases}$$

It follows that

$$\widehat{f}_{\delta,n}(x,a)-f^\star_\delta(x,a)=\begin{cases}0, & \left\{(x,a):\widehat{\eta}_a(x)>\widehat{T}_{\delta,a}+\ell_{n,a},\eta_a(x)>T^\star_{\delta,a}\right\},\\ 1-\tau^\star_{\delta,a}, & \left\{(x,a):\widehat{\eta}_a(x)>\widehat{T}_{\delta,a}+\ell_{n,a},\eta_a(x)=T^\star_{\delta,a}\right\},\\ 1, & \left\{\widehat{\eta}_a(x)>\widehat{T}_{\delta,a}+\ell_{n,a},(x,a):\eta_a(x)<T^\star_{\delta,a}\right\},\\ \widehat{\tau}_{\delta,a}-1, & \left\{(x,a):\left|\widehat{\eta}_a(x)-\widehat{T}_{\delta,a}\right|\leqslant\ell_{n,a},\eta_a(x)>T^\star_{\delta,a}\right\},\\ \widehat{\tau}_{\delta,a}-\tau^\star_{\delta,a}, & \left\{(x,a):\left|\widehat{\eta}_a(x)-\widehat{T}_{\delta,a}\right|\leqslant\ell_{n,a},\eta_a(x)=T^\star_{\delta,a}\right\},\\ \widehat{\tau}_{\delta,a}, & \left\{\left|\widehat{\eta}_a(x)-\widehat{T}_{\delta,a}\right|\leqslant\ell_{n,a},(x,a):\eta_a(x)<T^\star_{\delta,a}\right\},\\ -1, & \left\{(x,a):\widehat{\eta}_a(x)\leqslant\widehat{T}_{\delta,a}-\ell_{n,a},\eta_a(x)>T^\star_{\delta,a}\right\},\\ -\tau^\star_{\delta,a}, & \left\{(x,a):\widehat{\eta}_a(x)<\widehat{T}_{\delta,a}-\ell_{n,a},\eta_a(x)=T^\star_{\delta,a}\right\},\\ 0, & \left\{\widehat{\eta}_a(x)<\widehat{T}_{\delta,a}-\ell_{n,a},(x,a):\eta_a(x)<T^\star_{\delta,a}\right\}.\end{cases}$$

Since $T^\star_{\delta,a}-\eta_a(x)=0$ on $\left\{(x,a):\eta_a(x)=T^\star_{\delta,a}\right\}$, we have

$$\begin{aligned}&(\widehat{f}_{\delta,n}(x,a)-f^\star_\delta(x,a))(T^\star_{\delta,a}-f^\star_\delta(x,a))\\ &=\begin{cases}T^\star_{\delta,a}-\eta_a(x), & \{(x,a):\widehat{\eta}_a(x)>\widehat{T}_{\delta,a}+\ell_{n,a},\eta_a(x)<T^\star_{\delta,a}\};\\ (\widehat{\tau}_{\delta,a}-1)(T^\star_{\delta,a}-\eta_a(x)), & \{(x,a):|\widehat{\eta}_a(x)-\widehat{T}_{\delta,a}|\leqslant\ell_{n,a},\eta_a(x)>T^\star_{\delta,a}\};\\ \widehat{\tau}_{\delta,a}(T^\star_{\delta,a}-\eta_a(x)), & \{(x,a):|\widehat{\eta}_a(x)-\widehat{T}_{\delta,a}|\leqslant\ell_{n,a},\eta_a(x)<T^\star_{\delta,a}\};\\ -(T^\star_{\delta,a}-\eta_a(x)), & \{(x,a):\widehat{\eta}_a(x)<\widehat{T}_{\delta,a}-\ell_{n,a},\eta_a(x)>T^\star_{\delta,a}\};\\ 0, & \text{otherwise},\end{cases}\\ &\leqslant|\eta_a(x)-T^\star_{\delta,a}|\left[I(\eta_a(x)>T^\star_{\delta,a},\widehat{\eta}_a(x)\leqslant\widehat{T}_{\delta,a}+\ell_{n,a})+I(\eta_a(x)<T^\star_{\delta,a},\widehat{\eta}_a(x)\geqslant\widehat{T}_{\delta,a}-\ell_{n,a})\right].\end{aligned}$$

Then, by (4.2) and Lemma B.12 for $a\in\{0,1\}$, we conclude that

$$\begin{aligned}&\mathbb{E}^{\otimes n}\left[d_E\left(\widehat{f}_{\delta,n},f^\star\right)\right]=2\sum_{a\in\{0,1\}}p_a\mathbb{E}^{\otimes n}\left[\int(\widehat{f}_{\delta,n}(x,a)-f^\star(x,a))(T^\star_{\delta,a}-\eta_a(x))d\mathbb{P}_{X|A=a}(x)\right]\\ &=2\sum_{a\in\{0,1\}}p_a\mathbb{E}^{\otimes n}\int I\{\eta_a(x)>T^\star_{\delta,a},\,\widehat{\eta}_a(x)\leqslant\widehat{T}_{\delta,a}+\ell_{n,a}\}\left|\eta_a(x)-T^\star_{\delta,a}\right|d\mathbb{P}_{X|A=a}(x)\\ &+2\sum_{a\in\{0,1\}}p_a\mathbb{E}^{\otimes n}\int I\{\eta_a(x)<T^\star_{\delta,a},\,\widehat{\eta}_a(x)\geqslant\widehat{T}_{\delta,a}-\ell_{n,a}\}\left|\eta_a(x)-T^\star_{\delta,a}\right|d\mathbb{P}_{X|A=a}(x)\\ &\leqslant C\left((\phi_{n,1}\vee\phi_{n,0}\vee\ell_{n,1}\vee\ell_{n,0})+\widetilde{I}^\star(\delta)n^{-1/(2\gamma_\delta)}\right)^{\gamma_\delta+1}.\end{aligned}$$

E.3 Proof of Theorem 6.7

Recalling (3.5) and (5.10), we have by definition

$$\begin{aligned}
D_+(t^\star_\delta) &= \mathbb{P}_{X|A=1}\left(\eta_1(X) \geqslant \frac{1}{2} + \frac{t^\star_\delta}{2p_1}\right) - \mathbb{P}_{X|A=0}\left(\eta_0(X) > \frac{1}{2} - \frac{t^\star_\delta}{2p_0}\right) \\
&= \mathbb{P}_{X|A=1}\left(\eta_1(X) > T^\star_{\delta,1}\right) + \mathbb{P}_{X|A=1}\left(\eta_1(X) = T^\star_{\delta,1}\right) - \mathbb{P}_{X|A=0}\left(\eta_0(X) > T^\star_{\delta,0}\right) \\
&= \pi^\star_{1,+} + \pi^\star_{1,=} - \pi^\star_{0,+} \geqslant \delta,
\end{aligned}$$

and

$$\begin{aligned}
D_-(t^\star_\delta) &= \mathbb{P}_{X|A=1}\left(\eta_1(X) > \frac{1}{2} + \frac{t^\star_\delta}{2p_1}\right) - \mathbb{P}_{X|A=0}\left(\eta_0(X) \geqslant \frac{1}{2} - \frac{t^\star_\delta}{2p_0}\right) \\
&= \mathbb{P}_{X|A=1}\left(\eta_1(X) > T^\star_{\delta,1}\right) - \mathbb{P}_{X|A=0}\left(\eta_0(X) > T^\star_{\delta,0}\right) - \mathbb{P}_{X|A=0}\left(\eta_0(X) = T^\star_{\delta,0}\right) \\
&= \pi^\star_{1,+} - \pi^\star_{0,+} - \pi^\star_{0,=} \leqslant \delta.
\end{aligned}$$

Recall that for all x, a,

$$\widehat{f}_{\delta,n}(x,a) = I\left(\widehat{\eta}_a(X) > \widehat{T}_{\delta,a} + \ell_{n,a}\right) + \widehat{\tau}_{\delta,a} I\left(\left|\widehat{\eta}_a(x) - \widehat{T}_{\delta,a}\right| \leqslant \ell_{n,a}\right).$$

The disparity level of our method $\widehat{f}_{\delta,n}$ can be expressed as

$$\begin{aligned}
\mathrm{DDP}(\widehat{f}_{\delta,n}) &= \mathbb{P}_{X|A=1}\left(\widehat{Y}_{\widehat{f}_{\delta,n}} = 1\right) - \mathbb{P}_{X|A=0}\left(\widehat{Y}_{\widehat{f}_{\delta,n}} = 1\right) \\
&= \mathbb{P}_{X|A=1}\left(\widehat{\eta}_1(X) > \widehat{T}_{\delta,1} + \ell_{n,1}\right) + \widehat{\tau}_{\delta,1}\mathbb{P}_{X|A=1}\left(\left|\widehat{\eta}_1(x) - \widehat{T}_{\delta,1}\right| \leqslant \ell_{n,1}\right) \\
&\quad - \mathbb{P}_{X|A=0}\left(\widehat{\eta}_0(X) > \widehat{T}_{\delta,0} + \ell_{n,0}\right) + \widehat{\tau}_{\delta,0}\mathbb{P}_{X|A=0}\left(\left|\widehat{\eta}_0(x) - \widehat{T}_{\delta,0}\right| \leqslant \ell_{n,0}\right) \\
&= \widehat{\pi}_{1,+} + \widehat{\pi}_{1,=}\widehat{\tau}_{\delta,1} - \widehat{\pi}_{0,+} - \widehat{\pi}_{0,=}\widehat{\tau}_{\delta,0}.
\end{aligned}$$

From Proposition 5.2, we have $|\pi^\star_{1,+} + \pi^\star_{1,=}\tau^\star_{\delta,1} - \pi^\star_{0,+} - \pi^\star_{0,=}\tau^\star_{\delta,0}| \leqslant \delta$. Then, by Lemmas B.14 and B.17, we have, for $(L_\pi \vee L_{\pi_1})(\phi_{n,1} \vee \phi_{n,0}) < r_n < (U_\pi \wedge U_{\pi_1})$, $2\left(g_{\delta,-}(4r_n) \vee g_{\delta,+}(4r_n)\right) < \Delta_n < (U_{\Delta,\pi} \wedge U_{\Delta,\pi_1})$, $L_{\varepsilon,\pi} < \varepsilon/4 \leqslant \sqrt{(p_1 \wedge p_0)/2}$ and $(L_\eta \vee L_T)(\phi_{n,1} \vee \phi_{n,0}) < \ell_{n,a}/2 < r_n$,

$$\begin{aligned}
&\mathbb{P}^{\otimes n}\left(\mathrm{DDP}(\widehat{f}_{\delta,n}) > \delta + \varepsilon\right) = \mathbb{P}^{\otimes n}\left(\mathrm{DDP}(\widehat{f}_{\delta,n}) > \pi^\star_{1,+} + \pi^\star_{1,=}\tau^\star_{\delta,1} - \pi^\star_{0,+} - \pi^\star_{0,=}\tau^\star_{\delta,0} + \varepsilon\right) \\
&\leqslant \mathbb{P}^{\otimes n}\left(\widehat{\pi}_{1,+} - \pi^\star_{1,+} > \frac{\varepsilon}{4}\right) + \mathbb{P}^{\otimes n}\left(\widehat{\tau}_{\delta,1}\widehat{\pi}_{1,=} - \pi^\star_{1,=}\tau^\star_{\delta,1} > \frac{\varepsilon}{4}\right) \\
&\qquad + \mathbb{P}^{\otimes n}\left(\widehat{\pi}_{0,+} - \pi^\star_{0,+} < -\frac{\varepsilon}{4}\right) + \mathbb{P}^{\otimes n}\left(\widehat{\pi}_{0,=}\widehat{\tau}_{\delta,0} - \pi^\star_{0,=}\tau^\star_{\delta,0} < -\frac{\varepsilon}{4}\right) \\
&\leqslant 10\psi_{n,1,\pi}\left(\ell_{n,1} \wedge \ell_{n,0}\right) + 10\sum_{j\in\{-,+\}} I\left(\delta = D_j(t^\star_\delta)\right)\psi_{n,2,\pi}\left(g\left(\xi\left(\ell_{n,1} \wedge \ell_{n,0}, r_n\right)\right)\right) + 2c_{5,\pi}\exp\left(-\frac{c_{6,\pi}n\varepsilon^2}{16}\right).
\end{aligned}$$

For $c_{1,D} = 20c_{1,\pi}$, $c_{2,D} = c_{2,\pi}/16$, $c_{3,D} = 20c_{3,\pi}$, $c_{4,D} = c_{4,\pi}/(4^{2\gamma})$, $c_{5,D} = c_{5,\pi}$, and $c_{6,D} = c_{6,\pi}/16$, this is upper bounded by

$$\frac{1}{2}\psi_{n,1,D}\left(\ell_{n,1} \wedge \ell_{n,0}\right) + \frac{1}{2}\psi_{n,2,D}\left(g\left(\xi\left(\ell_{n,1} \wedge \ell_{n,0}, r_n\right)\right)\right) + \frac{1}{2}c_{5,D}\exp\left(-c_{6,D}n\varepsilon^2\right).$$

Similarly, the same upper bound holds for $\mathbb{P}^{\otimes n}\left(\mathrm{DDP}(\widehat{f}_{\delta,n}) < -\delta - \varepsilon\right)$. As a result, there exist constants $c_{D,i}$, $i \in [6]$ such that (6.8) holds.

F Proofs of Lemmas

F.1 Proof of Lemma B.1

Without loss of generality, we assume $z = 0$ by translation, and that $R = 1$ by scaling. The vertices of the polyhedron $B_{d,1}(0,1)$ are $\{\sigma \cdot e_j : j \in [d], \sigma \in \{-1,1\}\}$. Without loss of generality, we assume that the vertex

of $B_{d,1}(0,1)$ closest to x is e_1. Then, we have $\sum_{j=1}^d |x_j| \leqslant 1$ and $x_1 > \max_{j=2,\ldots,d} |x_j|$. When $r \leqslant 1/(d+2)$, we will show that

$$B_{d,1}\left(x - \frac{re_1}{2}, \frac{r}{2}\right) \subset B_{d,2}(x,r) \cap B_{d,1}(0,1). \tag{F.1}$$

First, as $B_{d,1}(x,r) \subset B_{d,2}(x,r)$, we have $B_{d,1}(x - e_1 r/2, r/2) \subset B_{d,1}(x,r) \subset B_{d,2}(x,r)$. Next, let $y = (y_1, \ldots, y_d)^\top \in B_{d,1}(x - e_1 r/2, r/2)$. We have

$$\left|y_1 - x_1 + \frac{r}{2}\right| + \sum_{j=2}^d |y_j - x_j| \leqslant \frac{r}{2}.$$

We consider the following two cases: (1) $\sum_{j=1}^d |x_j| \leqslant 1 - r$ and (2) $1 - r < \sum_{j=1}^d |x_j| \leqslant 1$.

(1) When $\sum_{j=1}^d |x_j| \leqslant 1 - r$, we have

$$\begin{aligned}\sum_{j=1}^d |y_j| &\leqslant \left|y_1 - x_1 + \frac{r}{2}\right| + \sum_{j=2}^d |y_j - x_j| + \left|x_1 - \frac{r}{2}\right| + \sum_{j=2}^d |x_j| \\ &\leqslant \frac{r}{2} + \left|x_1 - \frac{r}{2}\right| + \sum_{j=2}^d |x_j| \leqslant \frac{r}{2} + \frac{r}{2} + \sum_{j=1}^d |x_j| \leqslant 1.\end{aligned}$$

(2) When $1 - r < \sum_{j=1}^d |x_j| \leqslant 1$, we have $x_1 > 1 - r - \sum_{j=2}^d |x_j| \geqslant 1 - r - (d-1)x_1$. It follows that $x_1 > (1-r)/d \geqslant r/2$ as $r \leqslant 2/(d+2)$. Thus, starting with the same argument as above,

$$\sum_{j=1}^d |y_j| \leqslant \frac{r}{2} + \left|x_1 - \frac{r}{2}\right| + \sum_{j=2}^d |x_j| = \frac{r}{2} + x_1 - \frac{r}{2} + \sum_{j=1}^d |x_j| \leqslant 1.$$

This implies (F.1), and we then have

$$\lambda\left[B_{d,1}(0,1) \cap B_{d,2}(x,r)\right] \geqslant \lambda\left[B_{d,1}\left(x - \frac{re_1}{2}, \frac{r}{2}\right)\right] = \frac{V_{d,1} r^d}{2^d} = \frac{V_{d,1}}{V_{d,2} 2^d} \cdot V_{d,2} r^d = \frac{V_{d,1}}{V_{d,2} 2^d} \lambda[B_{d,2}(x,r)].$$

F.2 Proof of Lemma B.2

As in the proof of Lemma B.1, without loss of generality, we can assume $z = 0$ and $R = 1$. Further, since the ℓ_2-ball is rotation-invariant, we can assume without loss of generality that $x = x_1 e_1$ with $x_1 \in [0,1]$. For $r \leqslant 1$, and for $y = (y_1, \ldots, y_d)^\top \in B_{d,2}(x - e_1, 1) \cap B_{d,2}(x,r)$, we have

$$(y_1 - x_1 + 1)^2 + \sum_{j=2}^d y_j^2 \leqslant 1 \quad \text{and} \quad (y_1 - x_1)^2 + \sum_{j=2}^d y_j^2 \leqslant r^2.$$

It follows that

$$-\sqrt{1 - \sum_{j=2}^d y_j^2} \leqslant x_1 - \sqrt{r^2 - \sum_{j=2}^d y_j^2} \leqslant y_1 \leqslant \sqrt{1 - \sum_{j=2}^d y_j^2} + x_1 - 1 \leqslant \sqrt{1 - \sum_{j=2}^d y_j^2}.$$

It follows that $y \in B_{d,2}(0,1)$. This shows that $B_{d,2}(x - e_1, 1) \cap B_{d,2}(x,r) \subset B_{d,2}(0,1) \cap B_{d,2}(x,r)$. Moreover,

$$\{y : \|y - x + e_1\|_2^2 = 1\} \cap \{y : \|y - x\|_2^2 = r^2\} = \left\{y : y_1 = x_1 - \frac{r^2}{2}, \sum_{j=2}^d y_j^2 = r^2 - \frac{r^4}{4}\right\}.$$

It follows that

$$\lambda[B_{d,2}(0,1) \cap B_{d,2}(x,r)] \geqslant \lambda[B_{d,2}(x - e_1, 1) \cap B_{d,2}(x,r)]$$

$$= \int I\left(x_1 - r \leqslant y_1 \leqslant x_1 - r^2/2, \sum_{j=2}^{d} y_j^2 \leqslant r^2 - (y_1 - x_1)^2\right) dy$$
$$+ \int I\left(x_1 - r^2/2 \leqslant y_1 \leqslant x_1, \sum_{j=2}^{d} y_j^2 \leqslant 1 - (y_1 - x_1 + 1)^2\right) dy$$
$$= \int_{x_1 - r}^{x_1 - \frac{r^2}{2}} \left(r^2 - (y_1 - x_1)^2\right)^{\frac{d-1}{2}} V_{d-1,2} dy_1 + \int_{x_1 - \frac{r^2}{2}}^{x_1} \left(1 - (y_1 - x_1 + 1)^2\right)^{\frac{d-1}{2}} V_{d-1,2} dy_1.$$

By substituting $t = \sqrt{r^2 - (y_1 - x_1)^2}$ for the first integral, and using $r \leqslant 1$, this is lower bounded by

$$V_{d-1,2} \int_0^{\sqrt{r^2 - \frac{r^4}{4}}} \frac{t^d}{\sqrt{r^2 - t^2}} dt \geqslant V_{d-1,2} \int_0^{\frac{\sqrt{3}r}{2}} \frac{t^d}{r} dt$$
$$\geqslant r^d V_{d,2} \frac{3^{\frac{d+1}{2}} V_{d-1,2}}{2^{d+1}(d+1)V_{d,2}} = \frac{3^{\frac{d+1}{2}} V_{d-1,2}}{2^{d+1}(d+1)V_{d,2}} \cdot \lambda[B_{d,2}(x,r)].$$

F.3 Proof of Lemma B.3

We consider two cases: (1) $x \in [0,1]^d \setminus \mathcal{C}_{z,q}$ and (2) $x \in \mathcal{C}_{z,q} \setminus \mathcal{D}_{z,q}$.

- Case (1): $x \in [0,1]^d \setminus \mathcal{C}_{z,q}$.

  Since the result holds for any $C_r \leqslant 1$ when $B_{d,2}(x,r) \cap \mathcal{D}_{z,r} = \emptyset$, we only need to consider the case that $B_{d,2}(x,r) \cap \mathcal{D}_{z,r} \neq \emptyset$. In this case, we have $\|x - z\| < 2q^{-1} + r$. Letting $z_x = z + 3q^{-1}(x - z)/\|x - z\|$, we can verify that $B_{d,2}\left(z_x, q^{-1}\right) \subset B_{d,2}(x,r) \cap \left(\mathcal{C}_{z,q} \setminus \mathrm{int}\mathcal{D}_{z,q}\right)$. In fact, for any point $y \in B_{d,2}\left(z_x, q^{-1}\right)$, we have

  $$\|y - x\| \leqslant \|y - z_x\| + \|z_x - x\| \leqslant q^{-1} + \|x - z\| - 3q^{-1} = \|x - z\| - 2q^{-1} < r.$$

  Moreover,

  $$\|y - z\| \leqslant \|y - z_x\| + \|z_x - z\| = q^{-1} + 3q^{-1} = 4q^{-1},$$

  and $\|y - z\| \geqslant \|z_x - z\| - \|y - z_x\| = 3q^{-1} - q^{-1} = 2q^{-1}$. It follows that

  $$y \in B_{d,2}(x,r) \cap \left(B_{d,2}(z, 4q^{-1}) \setminus \mathrm{int}B_{d,2}(z, 2q^{-1})\right) \subset B_{d,2}(x,r) \cap \left(\mathcal{C}_{z,q} \setminus \mathrm{int}\mathcal{D}_{z,q}\right).$$

  Since $\lambda[\mathcal{C}_{z,q}] = 8^d q^{-d}$, and the boundary of $\mathcal{D}_{z,q}$ has zero Lebesgue measure, we thus have

  $$\lambda[B_{d,2}(x,r) \cap (\mathcal{C}_{z,q} \setminus \mathcal{D}_{z,q})] \geqslant \lambda[B_{d,2}(z_x, q^{-1})] = V_{d,2} q^{-d} = 8^{-d} V_{d,2} \lambda[\mathcal{C}_{z,q}]$$
  $$\geqslant 8^{-d} V_{d,2} \lambda[B_{d,2}(x,r) \cap \mathcal{C}_{z,q}].$$

- (2) Case (2): $x \in \mathcal{C}_{z,q} \setminus \mathcal{D}_{z,q}$. In this case, we consider two sub-cases: (2.1) $x \in B_{d,2}(z, q^{-1})$; and (2.2) $x \in \mathcal{C}_{z,q} \setminus B_{d,2}(z, 2q^{-1})$.

  - Case (2.1): $x \in B_{d,2}(z, q^{-1})$.

    * When $r \leqslant q^{-1}$, we have $B_{d,2}(x,r) \subset B_{d,2}(z, 2q^{-1}) \subset \mathcal{C}_{z,q}$ since, for any point $y \in B_{d,2}(x,r)$, $\|y - z\| \leqslant \|y - x\| + \|x - z\| \leqslant r + q^{-1} \leqslant 2q^{-1}$. Thus, by Lemma B.2 with $R = q^{-1}$, denoting $\Psi_d = 3^{\frac{d+1}{2}} V_{d-1,2}/(2^{d+1}(d+1)V_{d,2})$, we have

      $$\lambda[B_{d,2}(x,r) \cap (\mathcal{C}_{z,q} \setminus \mathcal{D}_{z,q})] = \lambda[B_{d,2}(x,r) \cap B_{d,2}(z, q^{-1})]$$
      $$\geqslant \Psi_d \lambda[B_{d,2}(x,r)] = \Psi_d \lambda[B_{d,2}(x,r) \cap \mathcal{C}_{z,q}].$$

    * When, $q^{-1} < r < (4\sqrt{d} + 1)q^{-1}$, we have

      $$\lambda[B_{d,2}(x,r) \cap (\mathcal{C}_{z,q} \setminus \mathcal{D}_{z,q})] \geqslant \lambda[B_{d,2}(x, q^{-1}) \cap (\mathcal{C}_{z,q} \setminus \mathcal{D}_{z,q})]$$

$$= \lambda[B_{d,2}(x,q^{-1}) \cap B_{d,2}(z,q^{-1})] \geqslant \Psi_d \lambda[B_{d,2}(x,q^{-1})] = \frac{\Psi_d}{(4\sqrt{d}+1)^d}\lambda[B_{d,2}(x,(4\sqrt{d}+1)q^{-1})]$$
$$\geqslant \frac{\Psi_d}{(4\sqrt{d}+1)^d}\lambda[B_{d,2}(x,(4\sqrt{d}+1)q^{-1}) \cap \mathcal{C}_{z,q}] \geqslant \frac{\Psi_d}{(4\sqrt{d}+1)^d}\lambda[B_{d,2}(x,r) \cap \mathcal{C}_{z,q}].$$

* When $r \geqslant (4\sqrt{d}+1)q^{-1}$, we have $\mathcal{C}_{z,q} \subset B_{d,2}(x,r)$ since for any point $y \in \mathcal{C}_{z,q}$, $\|y-x\| \leqslant \|y-z\| + \|z-x\| \leqslant 4\sqrt{d}q^{-1} + q^{-1} \leqslant (4\sqrt{d}+1)q^{-1}$. We thus have,

$$\lambda[B_{d,2}(x,r) \cap (\mathcal{C}_{z,q} \setminus \mathcal{D}_{z,q})] = \lambda[\mathcal{C}_{z,q} \setminus \mathcal{D}_{z,q}] = (8^d - (2^d-1)V_{d,2})q^{-d}$$
$$= \frac{8^d - (2^d-1)V_{d,2}}{8^d}\lambda[\mathcal{C}_{z,q}] = \frac{8^d - (2^d-1)V_{d,2}}{8^d}\lambda[B_{d,2}(x,r) \cap \mathcal{C}_{z,q}]. \tag{F.2}$$

– Case (2.2): $x \in \mathcal{C}_{z,q} \setminus B_{d,2}(z, 2q^{-1})$.

* When $r \leqslant q^{-1}$, the result holds when $\|x-z\| > 3q^{-1}$ since in this case, $B_{d,2}(x,r) \cap \mathcal{D}_{z,q} = \emptyset$. When $\|x-z\| \leqslant 3q^{-1}$, we consider the set $\mathcal{M} = \{\tilde{y} = 2x - y : y \in B_{d,2}(x,r) \cap \mathcal{D}_{z,q}\}$. Since $\mathcal{M}$ is a translation and reflection of $B_{d,2}(x,r) \cap \mathcal{D}_{z,q}$, we clearly have $\lambda[\mathcal{M}] = \lambda[B_{d,2}(x,r) \cap \mathcal{D}_{z,q}]$. Moreover, we can verify that $\mathcal{M} \subset B_{d,2}(x,r) \cap \left(B_{d,2}(z,4q^{-1}) \setminus B_{d,2}(z,3q^{-1})\right) \subset B_{d,2}(x,r) \cap (\mathcal{C}_{z,q} \setminus \mathcal{D}_{z,q})$. In fact, for any $\tilde{y} \in \mathcal{M}$ with $y = 2x - \tilde{y} \in B_{d,2}(x,r) \cap (\mathcal{C}_{z,q} \setminus \mathcal{D}_{z,q})$, we have $\|\tilde{y} - x\| = \|x - y\| \leqslant r \leqslant q^{-1}$ and

$$\|\tilde{y} - z\| \leqslant \|\tilde{y} - x\| + \|x - z\| \leqslant q^{-1} + 3q^{-1} = 4q^{-1}.$$

Moreover, since $x = (y+\tilde{y})/2$, it holds that $\|z-x\| \leqslant \max(\|z-\tilde{y}\|, \|z-y\|)$. As $\|z-y\| \leqslant 2q^{-1} < 3q^{-1} \leqslant \|z-x\|$, we have $\|\tilde{y}-z\| \geqslant 3q^{-1}$. As a result,

$$\tilde{y} \in B_{d,2}(x,r) \cap \left(B_{d,2}(z,4q^{-1}) \setminus B_{d,2}(z,3q^{-1})\right).$$

Therefore,

$$\lambda[B_{d,2}(x,r) \cap (\mathcal{C}_{z,q} \setminus \mathcal{D}_{z,q})] \geqslant \lambda[\mathcal{M}] = \lambda[B_{d,2}(x,r) \cap \mathcal{D}_{z,q}],$$

which, using $\mathcal{C}_{z,q} = \left(\mathcal{C}_{z,q} \setminus \mathcal{D}_{z,q}\right) \cup \mathcal{D}_{z,q}$, implies that

$$\lambda[B_{d,2}(x,r) \cap (\mathcal{C}_{z,q} \setminus \mathcal{D}_{z,q})] \geqslant \frac{1}{2}\lambda[B_{d,2}(x,r) \cap \mathcal{C}_{z,q}]. \tag{F.3}$$

* When $q^{-1} < r < 8\sqrt{d}q^{-1}$. We consider the set $\mathcal{N} = \{\tilde{y} = x + (rq)^{-1}(y-x) : y \in B_{d,2}(x,r) \cap \mathcal{C}_{z,q}\}$. Since $\mathcal{N}$ is a scaling of $B_{d,2}(x,r) \cap \mathcal{C}_{z,q}$ with scaling coefficient $(rq)^{-1}$, we clearly have $\lambda[\mathcal{N}] = (rq)^{-d}\lambda[B_{d,2}(x,r) \cap \mathcal{C}_{z,q}]$. Moreover, we can verify that $\mathcal{N} \subset B_{d,2}(x,q^{-1}) \cap \mathcal{C}_{z,q}$. In fact, for any $\tilde{y} \in \mathcal{N}$ with $y = x + rq(y-x) \in B_{d,2}(x,r) \cap \mathcal{C}_{z,q}$, we have $\|\tilde{y} - x\| = \|(rq)^{-1}(y-x)\| \leqslant q^{-1}$, and

$$\begin{aligned}
|\tilde{y}_i - z_i| &= |\tilde{y}_i - x_i + x_i - z_i| \leqslant |\tilde{y}_i - x_i| + |x_i - z_i| I\left((\tilde{y}_i - x_i)\cdot(x_i - z_i) \geqslant 0\right) \\
&= (rq)^{-1}|y_i - x_i| + |x_i - z_i| I\left((y_i - x_i)\cdot(x_i - z_i) \geqslant 0\right) \\
&= |y_i - x_i| + |x_i - z_i| I\left((y_i - x_i)\cdot(x_i - z_i) \geqslant 0\right) \\
&= \max\left(|y_i - x_i|, |y_i - z_i|\right) \leqslant 4q^{-1},
\end{aligned}$$

which uses that $(rq)^{-1} \leqslant 1$ when $r > q^{-1}$ and $I\left((\tilde{y}_i - x_i)\cdot(x_i - z_i) \geqslant 0\right) = I\left((y_i - x_i)\cdot(x_i - z_i) \geqslant 0\right)$. As a result, $\tilde{y} \in B_{d,2}(x,q^{-1}) \cap \mathcal{C}_{z,q}$.
Now, using (F.3), we have, for $q^{-1} < r < 8\sqrt{d}q^{-1}$,

$$\lambda[B_{d,2}(x,r) \cap (\mathcal{C}_{z,q} \setminus \mathcal{D}_{z,q})] \geqslant \lambda[B_{d,2}(x,q^{-1}) \cap (\mathcal{C}_{z,q} \setminus \mathcal{D}_{z,q})]$$
$$= \frac{1}{2}\lambda[B_{d,2}(x,q^{-1}) \cap \mathcal{C}_{z,q}] \geqslant \frac{1}{2}\lambda[\mathcal{N}] = \frac{1}{2r^d q^d}\lambda[B_{d,2}(x,r) \cap \mathcal{C}_{z,q}] \geqslant \frac{1}{2^{3d+1}d^{\frac{d}{2}}}\lambda[B_{d,2}(x,r) \cap \mathcal{C}_{z,q}].$$

* When $r \geqslant 8\sqrt{d}q^{-1}$, we have $\mathcal{C}_{z,q} \subset B_{d,2}(x,r)$, since for any point $y \in \mathcal{C}_{z,q}$, $\|y-x\| \leqslant \|y-z\| + \|z-x\| \leqslant 4\sqrt{d}q^{-1} + 4\sqrt{d}q^{-1} \leqslant 8\sqrt{d}q^{-1}$. Thus, the conclusion follows from the same reasoning as (F.2).

F.4 Proof of Lemma B.4

By definition, $\widehat{Y}$ is conditionally independent of Y given X and A. Thus,

$$\mathbb{P}(\widehat{Y}_f = 1, Y = 0 \mid X = x, A = a) = f(x,a)(1-\eta_a(x)), \text{ and}$$
$$\mathbb{P}(\widehat{Y}_f = 1, Y = 0 \mid X = x, A = a) = \eta_a(x)(1-f(x,a)).$$

This implies that

$$\begin{aligned}
R(f) = \mathbb{P}(Y \neq \widehat{Y}_f) &= \sum_{a\in\mathcal{A}} p_a \mathbb{P}(Y \neq \widehat{Y}_f | A = a) \\
&= \sum_{a\in\{0,1\}} p_a \int_{\mathcal{X}} \left(\mathbb{P}(\widehat{Y}_f = 1, Y = 0 \mid X = x, A = a) + \mathbb{P}(\widehat{Y}_f = 1, Y = 0 \mid X = x, A = a)\right) d\mathbb{P}_{X|A=a}(x) \\
&= \sum_{a\in\{0,1\}} p_a \int_{\mathcal{X}} \left(f(x,a)(1-\eta_a(x)) + \eta_a(x)(1-f(x,a))\right) d\mathbb{P}_{X|A=a}(x) \\
&= \sum_{a\in\{0,1\}} p_a \int_{\mathcal{X}} \left((1-2\eta_a(x))f(x,a) + \eta_a(x)\right) d\mathbb{P}_{X|A=a}(x).
\end{aligned}$$

For the second result,

$$\begin{aligned}
\mathrm{DDP}(f) &= \mathbb{P}\left(\widehat{Y}_f = 1 \mid A = 1\right) - \mathbb{P}\left(\widehat{Y}_f = 1 \mid A = 0\right) \\
&= \sum_{a\in\{0,1\}} (2a-1)\mathbb{P}\left(\widehat{Y}_f = 1 \mid A = a\right) = \sum_{a\in\{0,1\}} \int (2a-1) f(x,a) d\mathbb{P}_{X|A=a}(x).
\end{aligned}$$

This finishes the proof.

F.5 Proof of Lemma B.5

For $a \in \{0,1\}$ and $(x_j, a_j, y_j) \in \mathcal{S}_a$, denote $I_{a,j} = I(a_j = a)$. We have that $I_{a,j}$ are i.i.d. copies of $I(A = a)$ with $n_a = \sum_{j=1}^n I_{a,j}$, $I_{a,j} \in \{0,1\}$, and $\mathbb{E}(I_{a,j}) = p_a$. Then, by Hoeffding's inequality, $\mathbb{P}^{\otimes n}\left(\left|\frac{n_a}{n} - p_a\right| \geqslant \varepsilon\right) \leqslant 2\exp\left(-2n\varepsilon^2\right)$. Next, when $|n_a/n - p_a| \leqslant \delta \leqslant p_a/2$, we have

$$\left|\frac{n}{n_a} - \frac{1}{p_a}\right| = \frac{n}{n_a p_a}\left|\frac{n_a}{n} - p_a\right| \leqslant \frac{2\delta}{p_a^2}.$$

Thus, by taking $\varepsilon = \frac{2\delta}{p_a^2} \leqslant 1/p_a$,

$$\mathbb{P}^{\otimes n}\left(\left|\frac{n}{n_a} - \frac{1}{p_a}\right| \geqslant \varepsilon\right) \leqslant \mathbb{P}^{\otimes n}\left(\left|\frac{n_a}{n} - p_a\right| \geqslant \frac{p_a^2\varepsilon}{2}\right) \leqslant 2\exp\left(-\frac{np_a^4\varepsilon^2}{2}\right).$$

F.6 Proof of Lemma B.6

When $\varepsilon \leqslant \sqrt{p_a/2}$ and $|n_a/n - p_a| \leqslant \varepsilon\sqrt{p_a/2}$, we have, $n_a \geqslant n\left(p_a - \varepsilon\sqrt{p_a/2}\right) \geqslant np_a/2$. By the Dvoretzky–Kiefer–Wolfowitz inequality and (B.2) of Lemma B.5, when $\varepsilon \leqslant \sqrt{p_a/2}$,

$$\begin{aligned}
&\mathbb{P}^{\otimes n}\left(\sup_{T\in\mathbb{R}}\left|\frac{1}{n_a}\sum_{j=1}^{n_a} I\left(\eta_a\left(x_{a,j}\right) > T\right) - \mathbb{P}_{X|A=a}\left(\eta_a(X) > T\right)\right| > \varepsilon\right) \\
&= \mathbb{P}^{\otimes n}\left(\sup_{T\in\mathbb{R}}\left|\frac{1}{n_a}\sum_{j=1}^{n_a} I\left(\eta_a\left(x_{a,j}\right) > T\right) - \mathbb{P}_{X|A=a}\left(\eta_a(X) > T\right)\right| > \varepsilon, \left|\frac{n_a}{n} - p_a\right| \leqslant \sqrt{\frac{p_a}{2}}\varepsilon\right)
\end{aligned}$$

$$
\begin{aligned}
&+\mathbb{P}^{\otimes n}\left(\sup_{T\in\mathbb{R}}\left|\frac{1}{n_a}\sum_{j=1}^{n_a} I\left(\eta_a\left(x_{a,j}\right)>T\right)-\mathbb{P}_{X|A=a}\left(\eta_a(X)>T\right)\right|>\varepsilon, \left|\frac{n_a}{n}-p_a\right|>\sqrt{\frac{p_a}{2}}\varepsilon\right)\\
&\leqslant 2\exp\left(-2n\left(p_a-\sqrt{\frac{p_a}{2}}\varepsilon\right)\varepsilon^2\right)+\mathbb{P}^{\otimes n}\left(\left|\frac{n_a}{n}-p_a\right|\geqslant\sqrt{\frac{p_a}{2}}\varepsilon\right)\\
&\leqslant 2\exp\left(-np_a\varepsilon^2\right)+2\exp\left(-np_a\varepsilon^2\right)=4\exp\left(-np_a\varepsilon^2\right).
\end{aligned}
$$

F.7 Proof of Lemma B.7

Here, we only study the first term on the left hand side of (B.4); the same argument applies to the second one. For $t\in\mathbb{R}$ and all $j\in[n_1]$, let $I_{1,j}(t)=I(\eta_1(x_{1,j})>1/2+t/(2p_1))$, and for all $j\in[n_0]$, let $I_{0,j}(t)=I(\eta_0(x_{0,j})\geqslant 1/2-t/(2p_0))$. We have, for $a\in\{0,1\}$ and $j\in[n_a]$,

$$
\begin{aligned}
&\mathbb{E}_{X|A=1}\left(I_{1,j}(t)\right)=\mathbb{P}_{X|A=1}\left(\eta_1(X)>\tfrac{1}{2}+\tfrac{t}{2p_1}\right); \quad \mathbb{E}_{X|A=0}\left(I_{0,j}(t)\right)=\mathbb{P}_{X|A=0}\left(\eta_0(X)\geqslant\tfrac{1}{2}-\tfrac{t}{2p_0}\right);\\
&D_{n,-}(t)=\tfrac{1}{n_1}\sum_{j=1}^{n_1} I_{1,j}(t)-\tfrac{1}{n_0}\sum_{j=1}^{n_0} I_{0,j}(t); \qquad D_{-}(t)=\tfrac{1}{n_1}\sum_{j=1}^{n_1}\mathbb{E}[I_{1,j}(t)]-\tfrac{1}{n_0}\sum_{j=1}^{n_0}\mathbb{E}[I_{0,j}(t)].
\end{aligned}
$$

By Lemma B.6, we have, for $\varepsilon\leqslant\sqrt{(p_1\wedge p_0)/2}$ and $a\in\{0,1\}$

$$
\mathbb{P}^{\otimes n}\left(\sup_{T\in\mathbb{R}}\left|\frac{1}{n_a}\sum_{j=1}^{n_a}\left(I_{a,j}(t)-\mathbb{E}_{X|A=a}[I_{a,j}(t)]\right)\right|>\varepsilon\right)\leqslant 4\exp\left(-np_a\varepsilon^2\right).
$$

It follows that, for $\varepsilon\leqslant\sqrt{(p_1\wedge p_0)/2}$,

$$
\begin{aligned}
&\mathbb{P}^{\otimes n}\left(|D_{n,r}(t)-D_{-}(t)|>\varepsilon\right)\\
&\leqslant\sum_{a=0}^{1}\mathbb{P}^{\otimes n}\left(\left|\frac{1}{n_a}\sum_{j=1}^{n_a}\left(I_{a,j}(t)-\mathbb{E}_{X|A=a}[I_{a,j}(t)]\right)\right|>\frac{\varepsilon}{2}\right)\leqslant 8\exp\left(-\frac{n(p_1\wedge p_0)\varepsilon^2}{4}\right).
\end{aligned}
$$

F.8 Proof of Lemma B.8

Proof. For (B.5), by construction, we have, with $c_{1,\iota_0}=\sum_{k=1}^{K}c_{1,\iota_k}$ and $c_{2,\iota_0}=\min_{k\in[K]}\left(c_{2,\iota_k}C_k^2\right)$,

$$
\begin{aligned}
&\sum_{k=1}^{K}\psi_{n,1,\iota_k}(C_k\varepsilon)=\sum_{k=1}^{K}\left(c_{1,\iota_k}\exp\left(-c_{2,\iota_k}\left(\frac{C_k\varepsilon}{\phi_{n,1}\vee\phi_{n,0}}\right)^2\right)\right)\\
&\leqslant\left(\sum_{k=1}^{K}c_{1,\iota_k}\right)\exp\left(\left(-\min_{k\in[K]}\left(c_{2,\iota_k}C_k^2\right)\right)\left(\frac{\varepsilon}{\phi_{n,1}\vee\phi_{n,0}}\right)^2\right)=\psi_{n,1,\iota_0}(\varepsilon).
\end{aligned}
$$

For (B.6), under the margin condition (4.4), we have, for $j\in\{-,+\}$, $k\in[K]$ and $\left(\max_{k\in[K]}c_k\vee 1\right)\varepsilon<\varepsilon_0$, if $\delta=D_j\left(t_\delta^\star\right)$, that

$$
g_{\delta,j}(C_k)>U_\gamma^{-1}(C_k\varepsilon)^\gamma>U_\gamma^{-2}C_k^\gamma(U_\gamma\varepsilon^\gamma)>U_\gamma^{-2}C_k^\gamma g_{\delta,j}(\varepsilon).
$$

Thus, with $\widetilde{I}^\star(\delta)$ from (6.3),

$$
\begin{aligned}
&g_{\delta,j}(\omega(C_k\varepsilon,r_n))=\widetilde{I}^\star(\delta)\cdot g_{\delta,j}(C_k\varepsilon)+(1-\widetilde{I}^\star(\delta))\cdot g_{\delta,j}(r_n)>\widetilde{I}^\star(\delta)\cdot U_\gamma^{-2}C_k^\gamma\cdot g_{\delta,j}(\varepsilon)+(1-\widetilde{I}^\star(\delta))\cdot g_{\delta,j}(r_n)\\
&>(U_\gamma^{-2}C_k^\gamma\wedge 1)\left(\widetilde{I}^\star(\delta)\cdot g_{\delta,j}(\varepsilon)+(1-\widetilde{I}^\star(\delta))\cdot g_{\delta,j}(r_n)\right)=(U_\gamma^{-2}C_k^\gamma\wedge 1)g_{\delta,j}(\omega(\varepsilon,r_n)).
\end{aligned}
$$

It then follows that, with $c_{3,\iota_0}=\sum_{k=1}^{K}c_{3,\iota_k}$ and $c_{4,\iota_0}=\left(\min_{k\in[K]}c_{4,\iota_k}\cdot U_\gamma^{-2}C_k^\gamma\right)\wedge 1$,

$$
I\left(\delta=D_j(t_\delta^\star)\right)\left(\sum_{k=1}^{K}\psi_{n,2,\iota_k}\left(g_{\delta,j}\left(\omega(C_1\varepsilon,r_n)\right)\right)\right)=I\left(\delta=D_j(t_\delta^\star)\right)\left(\sum_{k=1}^{K}c_{3,\iota_k}\exp\left(-c_{4,\iota_k}ng_{\delta,j}^2\left(\omega(C_1\varepsilon,r_n)\right)\right)\right)
$$

$$\leqslant I\left(\delta = D_j(t_\delta^\star)\right) \sum_{k=1}^{K} c_{3,\iota_k} \exp\left(-\left(\left(c_{4,\iota_k} \cdot U_\gamma^{-2} C_1^\gamma\right) \wedge 1\right) n g_{\delta,j}^2\left(\omega(\varepsilon, r_n)\right)\right)$$

$$\leqslant I\left(\delta = D_j(t_\delta^\star)\right) \left(\left(\sum_{k=1}^{K} c_{3,\iota_k}\right) \exp\left(-\left(\left(\min_{k\in[K]} c_{4,\iota_k} \cdot U_\gamma^{-2} C_k^\gamma\right) \wedge 1\right) n g_{\delta,j}^2\left(\omega(\varepsilon, r_n)\right)\right)\right)$$

$$= I\left(\delta = D_j(t_\delta^\star)\right) \psi_{n,2,\iota_0}\left(g_{\delta,j}(\omega(\varepsilon, r_n))\right).$$

□

F.9 Proof of Lemma B.9

We only prove (B.7), since (B.8) can be verified in the same vein. We define the following events:

$$E_1 = \left\{\widehat{D}_n\left(t_\delta^\star + \varepsilon, 0, 0\right) \geqslant \delta + \Delta_n\right\}; \qquad E_2 = \left\{\max_j \left|\widehat{\eta}_1\left(x_{1,j}\right) - \eta_1\left(x_{1,j}\right)\right| \leqslant \tfrac{\varepsilon}{8p_1}\right\};$$
$$E_3 = \left\{\max_j \left|\widehat{\eta}_0\left(x_{0,j}\right) - \eta_0\left(x_{0,j}\right)\right| \leqslant \tfrac{\varepsilon}{8p_0}\right\}; \qquad E_4 = \left\{\left|\tfrac{n}{n_1} - \tfrac{1}{p_1}\right| \leqslant \tfrac{\varepsilon}{5p_1}\right\};$$
$$E_5 = \left\{\left|\tfrac{n}{n_0} - \tfrac{1}{p_0}\right| \leqslant \tfrac{\varepsilon}{5p_0}\right\}; \qquad E_6 = \left\{D_{n,-}\left(t_\delta^\star + \tfrac{\varepsilon}{2}\right) < \delta + \Delta_n\right\}.$$

When E_2 and E_4 hold with $\varepsilon \leqslant 1/4$, we have, with $t_\delta^\star$ and $T_{\delta,a}^\star$ from (3.5),

$$\frac{1}{n_1}\sum_{j=1}^{n_1} I\left(\widehat{\eta}_1\left(x_{1,j}\right) > 1/2 + \frac{n\left(t_\delta^\star + \varepsilon\right)}{2n_1}\right) = \frac{1}{n_1}\sum_{j=1}^{n_1} I\left(\widehat{\eta}_1\left(x_{1,j}\right) > T_{\delta,1}^\star + \frac{\varepsilon}{2p_1} + \frac{t_\delta^\star + \varepsilon}{2}\left(\frac{n}{n_1} - \frac{1}{p_1}\right)\right)$$
$$\leqslant \frac{1}{n_1}\sum_{j=1}^{n_1} I\left(\eta_1\left(x_{1,j}\right) > T_{\delta,1}^\star + \frac{\varepsilon}{2p_1} - \frac{(1+\varepsilon)\varepsilon}{10p_1} - \left|\widehat{\eta}_1\left(x_{1,j}\right) - \eta_1\left(x_{1,j}\right)\right|\right)$$
$$\leqslant \frac{1}{n_1}\sum_{j=1}^{n_1} I\left(\eta_1\left(x_{1,j}\right) > T_{\delta,1}^\star + \frac{\varepsilon}{4p_1}\right).$$

Similarly, when $E_{a,3}$ and $E_{a,5}$ hold with $\varepsilon \leqslant 1/4$,

$$\frac{1}{n_0}\sum_{j=1}^{n_0} I\left(\widehat{\eta}_0\left(x_{0,j}\right) > 1/2 - \frac{n\left(t_\delta^\star + \varepsilon\right)}{2n_0}\right) \geqslant \frac{1}{n_0}\sum_{j=1}^{n_0} I\left(\eta_0\left(x_{0,j}\right) \geqslant T_{\delta,0}^\star - \frac{\varepsilon}{4p_0}\right).$$

This implies that, when $E_2, \ldots, E_6$ hold with $\varepsilon \leqslant 1/4$, we have

$$\widehat{D}_n\left(t_\delta^\star + \varepsilon, 0, 0\right) = \frac{1}{n_1}\sum_{j=1}^{n_1} I\left(\widehat{\eta}_1\left(x_{1,j}\right) > 1/2 + \frac{n\left(t_\delta^\star + \varepsilon\right)}{2n_1}\right) - \sum_{j=1}^{n_0} I\left(\widehat{\eta}_0\left(x_{0,j}\right) > 1/2 - \frac{n\left(t_\delta^\star + \varepsilon\right)}{2n_0}\right)$$
$$\leqslant \frac{1}{n_1}\sum_{j=1}^{n_1} I\left(\eta_1\left(x_{1,j}\right) > T_{\delta,1}^\star + \frac{\varepsilon}{4p_1}\right) - \frac{1}{n_0}\sum_{j=1}^{n_0} I\left(\eta_0\left(x_{0,j}\right) \geqslant T_{\delta,0}^\star - \frac{\varepsilon}{4p_0}\right) = D_{n,-}\left(t_\delta^\star + \frac{\varepsilon}{2}, 0, 0\right) < \delta + \Delta_n.$$

It follows that $\mathbb{P}^{\otimes n}(\cap_{j=1}^6 E_j) = 0$ and

$$\mathbb{P}^{\otimes n}\left(\widehat{D}_n\left(t_\delta^\star + \varepsilon, 0, 0\right) > \delta\right) \leqslant \mathbb{P}^{\otimes n}\left(\cap_{j=1}^6 E_j\right) + \sum_{j=2}^{6} \mathbb{P}^{\otimes n}\left(E_j^c\right) = \sum_{j=2}^{6} \mathbb{P}^{\otimes n}\left(E_j^c\right).$$

Next, we bound $\mathbb{P}^{\otimes n}(E_2^c), \ldots, \mathbb{P}^{\otimes n}(E_6^c)$ in order. First, as $(\widehat{\eta}_1, \widehat{\eta}_0)$ are $(\phi_{n,1}, \phi_{n,0})_{n\geqslant 1}$-pointwise convergent, by (6.1), we have, for $8L_\eta(p_1\phi_{n,1} \vee p_0\phi_{n,0}) < \varepsilon < 8U_\eta(p_1 \wedge p_0)$, i.e., when $L_\eta\phi_{n,a} < \varepsilon/(8p_a) < U_\eta$ for $a \in \{0, 1\}$, that

$$\mathbb{P}^{\otimes n}(E_2^c) \leqslant c_{1,\eta} \exp\left(-\frac{c_{2,\eta}}{64p_1^2}\left(\frac{\varepsilon}{\phi_{n,1}}\right)^2\right) \text{ and } \mathbb{P}^{\otimes n}(E_3^c) \leqslant c_{1,\eta} \exp\left(-\frac{c_{2,\eta}}{64p_0^2}\left(\frac{\varepsilon}{\phi_{n,0}}\right)^2\right).$$

Second, using (B.3) of Lemma B.5 and that $\phi_{n,a} \geqslant c_\mu n^{-1/2}$, for $0 < \varepsilon \leqslant 5$, so that $\varepsilon/(5p_a) \leqslant 1/p_a$, we have

$$\mathbb{P}^{\otimes n}\left(E_4^c\right) \leqslant 2\exp\left(-\frac{np_1^2\varepsilon^2}{50}\right) \leqslant 2\exp\left(-\frac{c_\mu^2 p_1^2}{50}\left(\frac{\varepsilon}{\phi_{n,1}}\right)^2\right), \text{ and } \mathbb{P}^{\otimes n}\left(E_5^c\right) \leqslant 2\exp\left(-\frac{c_\mu^2 p_0^2}{50}\left(\frac{\varepsilon}{\phi_{n,0}}\right)^2\right).$$

Finally, to bound $\mathbb{P}^{\otimes n}(E_6^c)$, denote $c_\delta = \delta - D_-(t_\delta^\star)$ and note that

$$\begin{aligned}
&\mathbb{P}^{\otimes n}\left(E_6^c\right) = \mathbb{P}^{\otimes n}\left(D_{n,-}\left(t_\delta^\star + \frac{\varepsilon}{2}\right) \geqslant \delta + \Delta_n\right) = \mathbb{P}^{\otimes n}\left(\left(D_{n,-} - D_-\right)\left(t_\delta^\star + \frac{\varepsilon}{2}\right) \geqslant \delta + \Delta_n - D_-\left(t_\delta^\star + \frac{\varepsilon}{2}\right)\right) \\
&\leqslant 8\exp\left(-\frac{n(p_1 \wedge p_0)\left(\delta + \Delta_n - D_-\left(t_\delta^\star + \frac{\varepsilon}{2}\right)\right)^2}{4}\right) = 8\exp\left(-\frac{n(p_1 \wedge p_0)\left(c_\delta + \Delta_n + g_{\delta,-}\left(\frac{\varepsilon}{2}\right)\right)^2}{4}\right) \\
&\leqslant 8I\left(\delta = D_-(t_\delta^\star)\right)\exp\left(-\frac{n(p_1 \wedge p_0)\left(\Delta_n + g_{\delta,-}\left(\frac{\varepsilon}{2}\right)\right)^2}{4}\right) + 8\exp\left(-\frac{n(p_1 \wedge p_0)c_\delta^2}{4}\right).
\end{aligned}$$

For $\varepsilon \leqslant 5c_\delta/\sqrt{2(p_1 \vee p_0)}$, this can be upper bounded by

$$\begin{aligned}
&8I\left(\delta = D_-(t_\delta^\star)\right)\exp\left(-\frac{n(p_1 \wedge p_0)\left(\Delta_n + g_{\delta,-}\left(\frac{\varepsilon}{2}\right)\right)^2}{4}\right) + 8\exp\left(-\frac{n(p_1 \wedge p_0)^2\varepsilon^2}{50}\right) \\
&\leqslant 8I\left(\delta = D_-(t_\delta^\star)\right)\exp\left(-\frac{n(p_1 \wedge p_0)\left(\Delta_n + g_{\delta,-}\left(\frac{\varepsilon}{2}\right)\right)^2}{4}\right) + 8\exp\left(-\frac{c_\mu^2(p_1 \wedge p_0)^2}{50}\left(\frac{\varepsilon}{\phi_{n,1} \vee \phi_{n,0}}\right)^2\right).
\end{aligned}$$

In conclusion, by taking $L_{t_1} = 8L_\eta(p_1 \vee p_0)$ and $U_{t_1} = (8p_1U_\eta) \wedge (8p_0U_\eta) \wedge (5c_\delta/\sqrt{2(p_1 \vee p_0)}) \wedge (1/4)$, we have that, for $L_{t_1}(\phi_{n,1} \vee \phi_{n,0}) < \varepsilon < U_{t_1}$, (B.7) holds with $\psi_{n,1,t_1}(\varepsilon) = c_{1,t_1}\exp(-c_{2,t_1}(\varepsilon/[\phi_{n,1} \vee \phi_{n,0}])^2)$ and $\psi_{n,2,t_1}(\varepsilon) = c_{3,t_1}\exp\left(-c_{4,t_1}n\varepsilon^2\right)$, where $c_{1,t_1} = 2c_{1,\eta} + 12$, $c_{2,t_1} = (c_{2,\eta} \wedge c_\mu^2(p_1 \wedge p_0)^2)/(64(p_1 \vee p_0)^2 \vee 50)$, $c_{3,t_1} = 8$, and $c_{4,t_1} = (p_1 \wedge p_0)/4$.

F.10 Proof of Lemma B.10

Here, we only prove (B.9) and (B.10), since (B.11) and (B.12) can be verified similarly. To prove (B.9), we define the following events:

$$\begin{aligned}
&\overline{E}_1 = \left\{\widehat{D}_n\left(t_\delta^\star + \varepsilon, 0, 0\right) \leqslant \delta - \Delta_n\right\}; && \overline{E}_2 = \left\{\max_j\left|\widehat{\eta}_1\left(x_{1,j}\right) - \eta_1\left(x_{1,j}\right)\right| \leqslant \tfrac{\varepsilon}{4p_1}\right\}; \\
&\overline{E}_3 = \left\{\max_j\left|\widehat{\eta}_0\left(x_{0,j}\right) - \eta_0\left(x_{0,j}\right)\right| \leqslant \tfrac{\varepsilon}{4p_0}\right\}; && \overline{E}_4 = \left\{\left|\tfrac{n}{n_1} - \tfrac{1}{p_1}\right| \leqslant \tfrac{\varepsilon}{4p_1}\right\}; \\
&\overline{E}_5 = \left\{\left|\tfrac{n}{n_0} - \tfrac{1}{p_0}\right| \leqslant \tfrac{\varepsilon}{4p_0}\right\}; && \overline{E}_6 = \left\{D_{n,l}\left(t_\delta^\star + 2\varepsilon\right) > \delta - \Delta_n\right\}.
\end{aligned}$$

Following the arguments used for proving (B.7), we have the following facts:

(1) With $\varepsilon \leqslant 1$, $\mathbb{P}^{\otimes n}\left(\widehat{D}_n\left(t_\delta^\star + \varepsilon, 0, 0\right) \leqslant \delta - \Delta_n\right) \leqslant \sum_{j=2}^6 \mathbb{P}^{\otimes n}\left(\overline{E}_j^c\right)$.

(2) For $4L_\eta(p_1\phi_{n,1} \vee p_0\phi_{n,0}) < \varepsilon < 4U_\eta(p_1 \wedge p_0)$, i.e., when $L_\eta\phi_{n,a} < \varepsilon/(4p_a) < U_\eta$ for $a \in \{0,1\}$,

$$\mathbb{P}^{\otimes n}(\overline{E}_2^c) \leqslant c_{1,\eta}\exp\left(-\frac{c_{2,\eta}}{16p_1^2}\left(\frac{\varepsilon}{\phi_{n,1}}\right)^2\right) \text{ and } \mathbb{P}^{\otimes n}(\overline{E}_3^c) \leqslant c_{1,\eta}\exp\left(-\frac{c_{2,\eta}}{16p_0^2}\left(\frac{\varepsilon}{\phi_{n,0}}\right)^2\right).$$

(3) For $0 < \varepsilon \leqslant 4$, so that $\varepsilon/(4p_a) \leqslant 1/p_a$,

$$\mathbb{P}^{\otimes n}\left(\overline{E}_4^c\right) \leqslant 2\exp\left(-\frac{c_\mu^2 p_1^2}{32}\left(\frac{\varepsilon}{\phi_{n,1}}\right)^2\right) \text{ and } \mathbb{P}^{\otimes n}\left(\overline{E}_5^c\right) \leqslant 2\exp\left(-\frac{c_\mu^2 p_0^2}{32}\left(\frac{\varepsilon}{\phi_{n,0}}\right)^2\right).$$

Now, to bound $\mathbb{P}^{\otimes n}(\overline{E}_6^c)$, since $D_-(t_\delta^\star) = \delta$, we have

$$\delta - \Delta_n - D_-\left(t_\delta^\star + 2\varepsilon\right) = D_-(t_\delta^\star) - D_-\left(t_\delta^\star + 2\varepsilon\right) - \Delta_n = g_{\delta,-}\left(2\varepsilon\right) - \Delta_n.$$

Since $\Delta_n \asymp (\log\log n)^{-1}$ and $\varepsilon < r_n \asymp (\log n)^{-1}$, we have from Condition 4.3 that when n is large enough,

$$(\log\log n)^{-1} \asymp \Delta_n - U_\gamma 2^\gamma \varepsilon^\gamma < \Delta_n - g_{\delta,-}(2\varepsilon) < \sqrt{(p_1 \wedge p_0)/2}.$$

Then, by Lemma B.7,

$$\begin{aligned}\mathbb{P}^{\otimes n}\left(\overline{E}_6^c\right) &= \mathbb{P}^{\otimes n}\left(D_{n,-}\left(t_\delta^\star + 2\varepsilon\right) \leqslant \delta - \Delta_n\right) = \mathbb{P}^{\otimes n}\left(\left(D_{n,-} - D_-\right)\left(t_\delta^\star + 2\varepsilon\right) \leqslant \delta - \Delta_n - D_-\left(t_\delta^\star + 2\varepsilon\right)\right) \\ &= \mathbb{P}^{\otimes n}\left(\left(D_{n,-} - D_-\right)\left(t_\delta^\star + 2\varepsilon\right) \leqslant -\left(\Delta_n - g_{\delta,-}(2\varepsilon)\right)\right) \leqslant 8\exp\left(-\frac{n(p_1 \wedge p_0)\left(\Delta_n - g_{\delta,-}(2\varepsilon)\right)^2}{4}\right).\end{aligned}$$

In conclusion, by taking $L_{t_2} = 4L_\eta(p_1 \vee p_0)$ and $U_{t_2} = (4p_1U_\eta) \wedge (4p_0U_\eta) \wedge 1$, we have that, for $L_{t_2}(\phi_{n,1} \vee \phi_{n,0}) < \varepsilon < U_{t_2}$, (B.9) holds with $\psi_{n,1,t_2}(\varepsilon) = c_{1,t_2}\exp(-c_{2,t_2}(\varepsilon/[\phi_{n,1} \vee \phi_{n,0}])^2)$ and $\psi_{n,2,t_2}(\varepsilon) = c_{3,t_2}\exp\left(-c_{4,t_2}n\varepsilon^2\right)$, where $c_{1,t_2} = 2c_{1,\eta} + 4$, $c_{2,t_2} = (c_{2,\eta} \wedge c_\mu^2(p_1 \wedge p_0)^2)/(16(p_1 \vee p_0)^2 \vee 32)$, $c_{3,t_2} = 8$, and $c_{4,t_2} = (p_1 \wedge p_0)/4$.

Next, to prove (B.10), we define the following events:

$$\begin{aligned}\widetilde{E}_1 &= \left\{\widehat{D}_n\left(t_\delta^\star + \varepsilon, 0, 0\right) \geqslant \delta - \Delta_n\right\}; & \widetilde{E}_2 &= \left\{\max_j \left|\widehat{\eta}_1\left(x_{1,j}\right) - \eta_1\left(x_{1,j}\right)\right| \leqslant \tfrac{\varepsilon}{4p_1}\right\}; \\ \widetilde{E}_3 &= \left\{\max_j \left|\widehat{\eta}_0\left(x_{0,j}\right) - \eta_0\left(x_{0,j}\right)\right| \leqslant \tfrac{\varepsilon}{4p_0}\right\}; & \widetilde{E}_4 &= \left\{\left|\tfrac{n}{n_1} - \tfrac{1}{p_1}\right| \leqslant \tfrac{\varepsilon}{4p_1}\right\}; \\ \widetilde{E}_5 &= \left\{\left|\tfrac{n}{n_0} - \tfrac{1}{p_0}\right| \leqslant \tfrac{\varepsilon}{4p_0}\right\}; & \widetilde{E}_6 &= \left\{D_{n,l}\left(t_\delta^\star\right) < \delta - \Delta_n\right\}.\end{aligned}$$

Again, by the arguments used in proving (B.7), we obtain the following facts:

(1) With $\varepsilon \leqslant 1$, $\mathbb{P}^{\otimes n}\left(\widehat{D}_n\left(t_\delta^\star + \varepsilon, 0, 0\right) \geqslant \delta - \Delta_n\right) \leqslant \sum_{j=2}^6 \mathbb{P}^{\otimes n}\left(\widetilde{E}_j^c\right)$.

(2) For $4L_\eta(p_1\phi_{n,1} \vee p_0\phi_{n,0}) < \varepsilon < 4U_\eta(p_1 \wedge p_0)$, i.e., when $L_\eta\phi_{n,a} < \varepsilon/(4p_a) < U_\eta$ for $a \in \{0,1\}$,

$$\mathbb{P}^{\otimes n}(\widetilde{E}_2^c) \leqslant c_{1,\eta}\exp\left(-\frac{c_{2,\eta}}{16p_1^2}\left(\frac{\varepsilon}{\phi_{n,1}}\right)^2\right) \text{ and } \mathbb{P}^{\otimes n}(\widetilde{E}_3^c) \leqslant c_{1,\eta}\exp\left(-\frac{c_{2,\eta}}{16p_0^2}\left(\frac{\varepsilon}{\phi_{n,0}}\right)^2\right).$$

(3) For $0 < \varepsilon \leqslant 4$, so that $\varepsilon/(4p_a) \leqslant 1/p_a$,

$$\mathbb{P}^{\otimes n}\left(\widetilde{E}_4^c\right) \leqslant 2\exp\left(-\frac{c_\mu^2 p_1^2}{32}\left(\frac{\varepsilon}{\phi_{n,1}}\right)^2\right) \text{ and } \mathbb{P}^{\otimes n}\left(\widetilde{E}_5^c\right) \leqslant 2\exp\left(-\frac{c_\mu^2 p_0^2}{32}\left(\frac{\varepsilon}{\phi_{n,0}}\right)^2\right).$$

Now, to bound $\mathbb{P}^{\otimes n}(\widetilde{E}_6^c)$, let $c_\delta = \delta - D_-(t_\delta^\star)$. For $\Delta_n < c_\delta/2$ and $\varepsilon \leqslant c_\delta\sqrt{2/(p_1 \vee p_0)}$, by Lemma B.7,

$$\begin{aligned}\mathbb{P}^{\otimes n}\left(\widetilde{E}_6^c\right) &= \mathbb{P}^{\otimes n}\left(D_{n,-}\left(t_\delta^\star\right) \geqslant \delta - \Delta_n\right) = \mathbb{P}^{\otimes n}\left(\left(D_{n,-} - D_-\right)\left(t_\delta^\star\right) \geqslant \delta - \Delta_n - D_-\left(t_\delta^\star\right)\right) \\ &= \mathbb{P}^{\otimes n}\left(\left(D_{n,-} - D_-\right)\left(t_\delta^\star\right) \geqslant \frac{c_\delta}{2}\right) \leqslant 8\exp\left(\frac{n(p_1 \wedge p_0)c_\delta^2}{16}\right) \\ &\leqslant 8\exp\left(-\frac{n(p_1 \wedge p_0)^2\varepsilon^2}{32}\right) \leqslant 8\exp\left(-\frac{c_\mu^2(p_1 \wedge p_0)^2}{32}\left(\frac{\varepsilon}{\phi_{n,1} \vee \phi_{n,0}}\right)^2\right).\end{aligned}$$

In conclusion, by taking $L_{t_2} = 4L_\eta(p_1 \vee p_0)$ and $U_{t_2} = (4p_1U_\eta) \wedge (4p_0U_\eta) \wedge c_\delta\sqrt{2/(p_1 \vee p_0)} \wedge 1$, we have that, for $L_{t_2}(\phi_{n,1} \vee \phi_{n,0}) < \varepsilon < U_{t_2}$, (B.10) holds with $\psi_{n,1,t_2}(\varepsilon) = c_{1,t_2}\exp(-c_{2,t_2}(\varepsilon/[\phi_{n,1} \vee \phi_{n,0}])^2)$, where $c_{1,t_2} = 2c_{1,\eta} + 12$ and $c_{2,t_2} = (c_{2,\eta} \wedge c_\mu^2(p_1 \wedge p_0)^2)/(16(p_1 \vee p_0)^2 \vee 32)$.

F.11 Proof of Lemma B.11

Proof. Here, we only prove (B.13), (B.14) and (B.15) as the other three claims can be verified similarly: (B.16) is analogous to (B.14), (B.17) and (B.18) are analogous to (B.15).

For (B.13), recalling the definition of $\widehat{t}_{\delta,\Delta_n,\min}$ and $\widehat{t}_{\delta,\mathrm{mid}}$ in (5.3), we have $0 \leqslant \widehat{t}_{\delta,\Delta_n,\min} \leqslant \widehat{t}_{\delta,\mathrm{mid}}$ and $\{\widehat{t}_{\delta,\mathrm{mid}} > r_n\} \subset \{\widehat{D}_n(r_n,0,0) \geqslant \delta\}$. Now, $t^\star_\delta = 0$ if and only if $D_-(0) \leqslant \delta$. Further, take $L_r = L_{t_1}$ and $U_r = U_{t,1}$. By Lemma B.9, when $L_r < r_n < U_r$, we have, with $c_{i,r} = c_{i,t_1}, i \in [4]$, that

$$\begin{aligned}
&\mathbb{P}^{\otimes n}\left(\widehat{t}_{\delta,\mathrm{mid}} - \widehat{t}_{\delta,\Delta_n,\min} > r_n\right) \leqslant \mathbb{P}^{\otimes n}\left(\widehat{t}_{\delta,\mathrm{mid}} > r_n\right) \\
&\leqslant \mathbb{P}^{\otimes n}\left(\widehat{D}_n(r_n,0,0) \geqslant \delta\right) \leqslant \psi_{n,1,r}(r_n) + I(\delta = D_-(t^\star_\delta))\psi_{n,2,r}\left(g_{\delta,-}\left(\frac{r_n}{2}\right)\right).
\end{aligned}$$

For (B.14), by the definition of $\widehat{t}_{\delta,\Delta_n,\min}$ and $\widehat{t}_{\delta,\mathrm{mid}}$, we have $0 \leqslant \widehat{t}_{\delta,\Delta_n,\min} \leqslant \widehat{t}_{\delta,\mathrm{mid}}$, as well as

$$\left\{\widehat{t}_{\delta,\Delta_n,\min} < t^\star_\delta - \frac{r_n}{2}\right\} \subset \left\{\widehat{D}_n\left(t^\star_\delta - \frac{r_n}{2}, 0, 0\right) \leqslant \delta + \Delta_n\right\}$$

and $\left\{\widehat{t}_{\delta,\mathrm{mid}} > t^\star_\delta + \frac{r_n}{2}\right\} \subset \left\{\widehat{D}_n\left(t^\star_\delta + \frac{r_n}{2}, 0, 0\right) \geqslant \delta\right\}$.

Take $L_r = 2(L_{t_1} \vee L_{t_2})$, $U_r = 2(U_{t,1} \wedge U_{t,2})$ and $U_{\Delta_1} = (D_+(t^\star_\delta) - \delta)/2$. By Lemmas B.9 and B.10, when $L_b(\phi_{n,1} \vee \phi_{n,0}) < r_n < U_r$ and $g_{\delta,+}(4r_n) < \Delta_n < U_{\Delta,r}$, we have, with $c_{1,r} = c_{1,t_1} + c_{1,t_2}$, $c_{2,r} = c_{2,t_1} \wedge c_{2,t_2}$, $c_{3,r} = c_{3,t_2}$ and $c_{4,r} = c_{4,t_1}$ that

$$\begin{aligned}
&\mathbb{P}^{\otimes n}\left(\widehat{t}_{\delta,\mathrm{mid}} - \widehat{t}_{\delta,\Delta_n,\min} > r_n\right) \leqslant \mathbb{P}^{\otimes n}\left(\widehat{t}_{\delta,\Delta_n,\min} < t^\star_\delta - \frac{r_n}{2}\right) + \mathbb{P}^{\otimes n}\left(\widehat{t}_{\delta,\mathrm{mid}} > t^\star_\delta + \frac{r_n}{2}\right) \\
&\leqslant \mathbb{P}^{\otimes n}\left(\widehat{D}_n\left(t^\star_\delta - \frac{r_n}{2}, 0, 0\right) \leqslant \delta + \Delta_n\right) + \mathbb{P}^{\otimes n}\left(\widehat{D}_n\left(t^\star_\delta + \frac{r_n}{2}, 0, 0\right) \geqslant \delta\right) \\
&\leqslant \psi_{n,1,t_2}\left(\frac{r_n}{2}\right) + \psi_{n,1,t_1}\left(\frac{r_n}{2}\right) + I(\delta = D_-(t^\star_\delta))\psi_{n,2,t_1}\left(g_{\delta,-}\left(\frac{r_n}{4}\right)\right) \\
&\leqslant \psi_{n,1,r}\left(\frac{r_n}{2}\right) + I(\delta = D_-(t^\star_\delta))\psi_{n,2,r}\left(g_{\delta,-}\left(\frac{r_n}{4}\right)\right).
\end{aligned}$$

For (B.15), again by the definition of $\widehat{t}_{\delta,\Delta_n,\min}$ and $\widehat{t}_{\delta,\mathrm{mid}}$, we have $0 \leqslant \widehat{t}_{\delta,\Delta_n,\min} \leqslant \widehat{t}_{\delta,\mathrm{mid}}$, as well as

$$\left\{\widehat{t}_{\delta,\Delta_n,\min} > t^\star_\delta - 2r_n\right\} \subset \left\{\widehat{D}_n\left(t^\star_\delta - 2r_n, 0, 0\right) \geqslant \delta + \Delta_n\right\}$$

and $\left\{\widehat{t}_{\delta,\mathrm{mid}} < t^\star_\delta - r_n\right\} \subset \left\{\widehat{D}_n\left(t^\star_\delta - r_n, 0, 0\right) \leqslant \delta\right\}$.

Take $L_r = L_{t_1} \vee (L_{t_2}/2)$, $U_r = (U_{t,1} \wedge U_{t,2})/2$ and $U_{\Delta,r} = \sqrt{(p_1 \wedge p_0)/2} \wedge ((\delta - D_-(t^\star_\delta))/2)$. By Lemmas B.9 and B.10, when $L_r(\phi_{n,1} \vee \phi_{n,0}) < r_n < U_r$ and $g_{\delta,+}(4r_n) < \Delta_n < U_{\Delta_1}$, we have, with $c_{1,r} = c_{1,t_1} + c_{1,t_2}$, $c_{2,r} = c_{2,t_1} \wedge c_{2,t_2}$, $c_{3,r} = c_{3,t_1} + c_{3,t_2}$ and $c_{4,r} = c_{4,t_1} \wedge c_{4,t_2}$ that

$$\begin{aligned}
&\mathbb{P}^{\otimes n}\left(\widehat{t}_{\delta,\mathrm{mid}} - \widehat{t}_{\delta,\Delta_n,\min} < r_n\right) \leqslant \mathbb{P}^{\otimes n}\left(\widehat{t}_{\delta,\mathrm{mid}} > t^\star_\delta - 2r_n\right) + \mathbb{P}^{\otimes n}\left(\widehat{t}_{\delta,\mathrm{mid}} < t^\star_\delta - r_n\right) \\
&\leqslant \mathbb{P}^{\otimes n}\left(\widehat{D}_n\left(t^\star_\delta - 2r_n, 0, 0\right) \geqslant \delta + \Delta_n\right) + \mathbb{P}^{\otimes n}\left(\widehat{D}_n\left(t^\star_\delta - r_n, 0, 0\right) \leqslant \delta\right) \\
&\leqslant \psi_{n,1,t_2}(2r_n) + \psi_{n,2,t_2}\left(\Delta_n - g_{\delta,+}\left(4r_n\right)\right) + \psi_{n,1,t_1}(r_n) + \psi_{n,2,t_1}\left(g_{\delta,+}\left(\frac{r_n}{2}\right)\right) \\
&\leqslant \psi_{n,1,r}(r_n) + \psi_{n,2,r}\left(\left(\Delta_n - g_{\delta,+}\left(4r_n\right)\right) \wedge g_{\delta,+}\left(\frac{r_n}{2}\right)\right) \leqslant \psi_{n,1,r}(r_n) + \psi_{n,2,r}\left(g_{\delta,+}\left(\frac{r_n}{2}\right)\right).
\end{aligned}$$

The last inequality holds since, when $\Delta_n > 2g_{\delta,+}(4r_n)$, $\Delta_n - g_{\delta,+}(4r_n) > g_{\delta,+}(4r_n) > g_{\delta,+}(r_n/2)$. □

F.12 Proof of Lemma B.12

Here, we only prove the first claim with $a = 1$, since the other results can be derived with the same argument. During the proof, we use C, C_1, C_2, etc., to represent constants that may vary from line to line. Let $0 < C_J < 1/4$ be a constant determined later. For $J_n = \lfloor C_J \cdot \log_2 n \rfloor$ and $1 \leqslant j \leqslant J_n$, with L_η, U_η and

$(\phi_{n,1},\phi_{n,0})_{n\geqslant 1}$ from Definition 6.1, L_T and U_T from Corollary 6.3, D_- from (3.3), D_+ from (3.3), $t^\star_\delta$ from (3.5), δ from (5.9), and γ from the γ-exponent condition in the upper bound from Definition 4.4, we denote

$$a_{n,j} = 2^j(L_\eta \vee L_T)(\phi_{n,1} \vee \phi_{n,0}), \tag{F.4}$$

$$b_{n,j} = \begin{cases} 2^{2+j/\gamma} L_\gamma^{-1/\gamma} n^{-1/2\gamma}, & t^\star_\delta > 0 \text{ and } D_-(t^\star_\delta) = D_+(t^\star_\delta) = \delta; \\ 0, & \text{otherwise}, \end{cases} \tag{F.5}$$

$$d_{n,j} = a_{n,j} + b_{n,j} + \ell_{n,1}. \tag{F.6}$$

Here, we set C_J such that $d_{n,J_n} < \Delta_n \asymp (\log n)^{-1}$. From Definition 6.1, it follows that

$$\mathbb{P}^{\otimes n}\left(|\widehat{\eta}_1 - \eta_1^\star| > \frac{1}{2} a_{n,j-1}\right) \leqslant c_{1,\eta} \exp\left(-c_{2,\eta}(L_\eta \vee L_T)^2 2^{2j-2}\right) \leqslant C_1 \exp\left(-C_2 2^{2j}\right),$$

where in the second inequality we have defined the constants C_1, C_2 appropriately. We first handle the case where $t^\star_\delta > 0$ and $D_-(t^\star_\delta) = D_+(t^\star_\delta) = \delta$, and then consider the complement of this case.

Scenario 1: $t^\star_\delta > 0$ and $D_-(t^\star_\delta) = D_+(t^\star_\delta) = \delta$.

In this case, we have

$$d_{n,j} = a_{n,j} + b_{n,j} + \ell_{n,1} \leqslant C\left(2^j(\phi_{n,1} \vee \phi_{n,0}) + 2^{j/\gamma} n^{-1/2\gamma} + \ell_{n,1}\right).$$

Moreover, by the margin condition (4.4) from Definition 4.4,, we have

$$g_{\delta,-}\left(\frac{b_{n,j}}{2}\right) \wedge g_{\delta,+}\left(\frac{b_{n,j}}{2}\right) \geqslant L_\gamma 2^{-2\gamma} b_{n,j}^\gamma = 2^j n^{-\frac{1}{2}}.$$

By Corollary 6.3, and using appropriate parts of $1/2 \cdot a_{n,j-1} + b_{n,j-1}$ in the upper bound,

$$\begin{aligned}
&\mathbb{P}^{\otimes n}\left(\widehat{T}_{\delta,1} - T^\star_{\delta,1} > 1/2 \cdot a_{n,j-1} + b_{n,j-1}\right) \\
&\leqslant c_{1,T} \exp\left(-c_{2,T}\left(\frac{a_{n,j-1}}{2(\phi_{n,1} \vee \phi_{n,0})}\right)^2\right) + c_{3,T} \exp\left(-c_{4,T} n \left(g_{\delta,-}\left(\frac{b_{n,j}}{2}\right) \wedge g_{\delta,+}\left(\frac{b_{n,j}}{2}\right)\right)^2\right) \\
&\leqslant c_{1,T} \exp\left(-c_{2,T} 2^{2(j-2)}(L_\eta \vee L_T)^2\right) + c_{3,T} \exp\left(-c_{4,T} 2^{2j}\right) \leqslant C_1 \exp\left(-C_2 2^{2j}\right).
\end{aligned}$$

Thus, we have

$$\mathbb{P}^{\otimes n}\left(\widehat{T}_{\delta,1} - T^\star_{\delta,1} > a_{n,j-1}/2 + b_{n,j-1}\right) \leqslant C_1 \exp\left(-C_2 2^{2j}\right). \tag{F.7}$$

Scenario 2: Other cases.

In this case, we have $d_{n,j} = a_{n,j} + \ell_{n,1} \leqslant C\left(2^j(\phi_{n,1} \vee \phi_{n,0}) + \ell_{n,1}\right)$. By the margin condition (4.4) from Definition 4.4, we have, for $j \in \{-,+\}$ that $g_{\delta,j}(r_n/4) > L_\gamma 2^{-2\gamma} r_n^\gamma$. It follows that,

$$I(\delta = D_-(t^\star_\delta))\psi_{n,2,T}\left(g_{\delta,-}\left(\frac{r_n}{4}\right)\right) + I(\delta = D_+(t^\star_\delta))\psi_{n,2,T}\left(g_{\delta,+}\left(\frac{r_n}{4}\right)\right) \leqslant 2\psi_{n,2,T}\left(L_\gamma 2^{-2\gamma} r_n^\gamma\right).$$

When $r_n \asymp (\log\log n)^{-1}$, we have $n r_n^{2\gamma} \geqslant C n^{\frac{1}{2}} \geqslant C 2^{2J_n}$. Then, by Corollary 6.3, we have

$$\begin{aligned}
&\mathbb{P}^{\otimes n}\left(\widehat{T}_{\delta,1} - T^\star_{\delta,1} > a_{n,j-1}/2 + b_{n,j-1}\right) = \mathbb{P}^{\otimes n}\left(\widehat{T}_{\delta,1} - T^\star_{\delta,1} > a_{n,j-1}/2\right) \\
&\leqslant \psi_{n,1,T}(a_{n,j-1}/2) + 2\psi_{n,2,T}(L_\gamma 2^{-2\gamma} r_n^\gamma) \\
&= c_{1,T} \exp\left(-c_{2,T} 2^{2(j-2)}(L_\eta \vee L_T)^2\right) + 2c_{3,T} \exp\left(-c_{3,T} n L_\gamma^2 2^{-4\gamma} r_n^{2\gamma}\right) \leqslant C_1 \exp\left(-C_2 2^{2j}\right).
\end{aligned}$$

Now, we consider the disjoint sets

$$\mathcal{C}_0 = \left\{\eta_1 > T^\star_{\delta,1} + d_{n,J_n},\, \widehat{\eta}_1 \leqslant \widehat{T}_{\delta,1} + \ell_{n,1}\right\}; \quad \mathcal{C}_1 = \left\{T^\star_{\delta,1} < \eta_1 \leqslant T^\star_{\delta,1} + d_{n,1},\, \widehat{\eta}_1 \leqslant \widehat{T}_{\delta,1} + \ell_{n,1}\right\};$$
$$\mathcal{C}_j = \left\{T^\star_{\delta,1} + d_{n,j-1} < \eta_1 \leqslant T^\star_{\delta,1} + d_{n,j},\, \widehat{\eta}_1 \leqslant \widehat{T}_{\delta,1} + \ell_{n,1}\right\}, \quad j \geqslant 2.$$

Clearly, $\left\{x : \eta_1(x) > T^\star_{\delta,1},\, \widehat{\eta}_1(x) \leqslant \widehat{T}_{\delta,1} + \ell_{n,1}\right\} \subset \cup_{j=0}^{J_n} \mathcal{C}_j$. We bound the quantity of interest over the sets $\mathcal{C}_j$ individually.

On $\mathcal{C}_1$: Since $\mathcal{C}_1 \subset \{0 < |\eta_1 - T^\star_{\delta,1}| \leqslant d_{n,1}\}$, using the margin condition from Definition 4.4, we have

$$\mathbb{E}^{\otimes n}\left[\int_{\mathcal{C}_1} \left|\eta_1(x) - T^\star_{\delta,1}\right| d\mathbb{P}_{X|A=1}(x)\right] \leqslant d_{n,1}\mathbb{P}_{X|A=1}\left(T^\star_{\delta,1} < \eta_1(X) \leqslant T^\star_{\delta,1} + d_{n,1}\right) \leqslant U_\gamma d_{n,1}^{1+\gamma}.$$

On $\mathcal{C}_j$ for $2 \leqslant j \leqslant J_n$: For $j \geqslant 2$, we have due to the definitions from (F.4), $\mathcal{C}_j \subset \mathcal{C}_{j,1} \cup \mathcal{C}_{j,2}$ with

$$\mathcal{C}_{j,1} = \left\{|\widehat{\eta}_1 - \eta_1| > \frac{1}{2}a_{n,j-1}\right\} \cap \left\{T^\star_{\delta,1} < \eta_1(x) \leqslant T^\star_{\delta,1} + d_{n,j}\right\},$$

and

$$\mathcal{C}_{j,2} = \left\{\widehat{T}_{\delta,1} - T^\star_{\delta,1} > \frac{1}{2}a_{n,j-1} + b_{n,j-1}\right\} \cap \left\{T^\star_{\delta,1} < \eta_1(x) \leqslant T^\star_{\delta,1} + d_{n,j}\right\}.$$

Using Fubini's theorem, the margin condition from Definition 4.4, and (F.7),

$$\begin{aligned}
&\mathbb{E}^{\otimes n}\left[\int_{\mathcal{C}_j} \left|\eta_1(x) - T^\star_{\delta,1}\right| d\mathbb{P}_{X|A=1}(x)\right] \\
&\leqslant \mathbb{E}^{\otimes n}\left[\int_{\mathcal{C}_{j,1}} \left|\eta_1(x) - T^\star_{\delta,1}\right| d\mathbb{P}_{X|A=1}(x)\right] + \mathbb{E}^{\otimes n}\left[\int_{\mathcal{C}_{j,2}} \left|\eta_1(x) - T^\star_{\delta,1}\right| d\mathbb{P}_{X|A=1}(x)\right] \\
&= d_{n,j}\int \left[\mathbb{P}^{\otimes n}\left(|\widehat{\eta}_1(x) - \eta_1(x)| > \frac{1}{2}a_{n,j-1}\right) + \mathbb{P}^{\otimes n}\left(\widehat{T}_{\delta,1} - T^\star_{\delta,1} > \frac{1}{2}a_{n,j-1} + b_{n,j-1}\right)\right] \\
&\qquad\qquad \cdot I\left(T^\star_{\delta,1} < \eta_1(x) \leqslant T^\star_{\delta,1} + d_{n,j}\right) d\mathbb{P}_{X|A=1}(x) \\
&\leqslant d_{n,j}C_1 \exp\left(-C_2 2^{2j}\right)\mathbb{P}_{X|A=1}(T^\star_{\delta,1} < \widehat{T}_{\delta,1} \leqslant T^\star_{\delta,1} + d_{n,j}) \leqslant C_1 U_\gamma d_{n,j}^{\gamma+1} \exp\left(-C_2 2^{2j}\right).
\end{aligned}$$

On $\mathcal{C}_0$: Finally,

$$\mathcal{C}_0 \subset \{|\widehat{\eta}_1 - \eta_1| > a_{n,J_n}/2\} \cup \left\{\widehat{T}_{\delta,1} - T^\star_{\delta,1} > a_{n,J_n}/2 + b_{n,J_n}\right\}.$$

Using that for all x, $|\eta_1(x) - T^\star_{\delta,1}| \leqslant 1$, by Fubini's theorem, and by the margin condition from Definition 4.4, and (F.7), we obtain that

$$\begin{aligned}
&\mathbb{E}^{\otimes n}\left(\int_{\mathcal{C}_0} \left|\eta_1(x) - T^\star_{\delta,1}\right| d\mathbb{P}_{X|A=1}\right) \leqslant \mathbb{E}^{\otimes n}\mathbb{P}_{X|A=1}\left(\mathcal{C}_0\right) \\
&\leqslant \mathbb{P}^{\otimes n}\left(|\widehat{\eta}_1 - \eta_1| > \frac{1}{2}a_{n,J_n}\right) + \mathbb{P}^{\otimes n}\left(\widehat{T}_{\delta,1} - T^\star_{\delta,1} > \frac{1}{2}a_{n,J_n} + b_{n,J_n}\right) \\
&\leqslant 2C_1 \exp\left(-C_2 2^{2J_n}\right) = 2C_1 \exp\left(-C_2 n^{2C_J}\right) \leqslant C\left(\phi_{n,1} \vee \phi_{n,0} \vee \ell_{n,1} \vee n^{-\frac{1}{2\gamma}}\right)^\gamma.
\end{aligned}$$

To conclude, we have

$$\begin{aligned}
&\mathbb{E}^{\otimes n}\int_{\eta_1(x) > T^\star_{\delta,1}, \widehat{\eta}_1(x) \leqslant \widehat{T}_{\delta,1} + \ell_{n,1}} |\eta_1(x) - T^\star_1| d\mathbb{P}_{X|A=1}(x) \leqslant \sum_{j=0}^{J_n} \mathbb{E}^{\otimes n}\int_{\mathcal{C}_j} |\eta_1(x) - T^\star_1| d\mathbb{P}_{X|A=1}(x) \\
&\leqslant U_\gamma d_{n,1}^{1+\gamma} + \sum_{j=2}^{J_n} C_1 U_\gamma d_{n,j}^{\gamma+1} \exp\left(-C_2 2^{2j}\right)
\end{aligned}$$

$$\leqslant \begin{cases} C\sum_{j=1}^{J_n}\exp\left(-C_2 2^{2j}\right)\left(2^j(\phi_{n,1}\vee\phi_{n,0})+2^{j/\gamma}n^{-\frac{1}{2\gamma}}+\ell_{n,1}\right)^{\gamma+1}, & 0<D_-(t_\delta^\star)=\delta=D_+(t_\delta^\star); \\ C\sum_{j=1}^{J_n}\exp\left(-C_2 2^{2j}\right)\left(2^j(\phi_{n,1}\vee\phi_{n,0})+\ell_{n,1}\right)^{\gamma+1}, & \text{otherwise}, \end{cases}$$
$$\leqslant C\left((\phi_{n,1}\vee\phi_{n,0}\vee\ell_{n,1})^{\gamma+1}+I\left(0<D_-(t_\delta^\star)=\delta=D_+(t_\delta^\star)\right)\cdot n^{-\frac{\gamma+1}{2\gamma}}\right).$$

This finishes the proof.

F.13 Proof of Lemma B.13

By the definition of $\widetilde{\delta}$ from and $\widehat{\delta}$, we have

$$\mathbb{P}^{\otimes n}\left(\widehat{\delta}\neq\widetilde{\delta}\right)=\begin{cases}\mathbb{P}^{\otimes n}\left(\widehat{D}_n(r_n,0,0)\geqslant\delta+\Delta_n\right), & D_-(0)\leqslant\delta;\\ \mathbb{P}^{\otimes n}\left(\widehat{D}_n(r_n,0,0)<\delta+\Delta_n\right), & D_-(0)>\delta.\end{cases}$$

We consider the two cases in order:

- Case (1): $D_-(0)\leqslant\delta$.

  In this case, we have $t_\delta^\star=0$. By (B.7) of Lemma B.9, with L_{t_1}, U_{t_1} and $c_{i,t_1}, i\in[4]$ from Lemma B.9, we have, for $L_{t_1}(\phi_{n,1}\vee\phi_{n,0})<r_n<U_{t_1}$ and $\Delta_n>0$,

  $$\mathbb{P}^{\otimes n}\left(\widehat{D}_n(r_n,0,0)\geqslant\delta+\Delta_n\right)\leqslant\psi_{n,1,t_1}(r_n)+\psi_{n,2,t_1}(\Delta_n).$$

  Thus, (B.19) holds with $c_{i,\delta}=c_{i,t_1}, i\in[4]$.

- Case (2): $D_-(0)>\delta$.

  In this case, we define the following events:

  $$\widehat{E}_1=\left\{\widehat{D}_n(r_n,0,0)<\delta+\Delta_n\right\};\qquad \widehat{E}_2=\left\{\max_j|\widehat{\eta}_1(x_{1,j})-\eta_1(x_{1,j})|\leqslant\tfrac{r_n}{2p_1}\right\};$$
  $$\widehat{E}_3=\left\{\max_j|\widehat{\eta}_0(x_{0,j})-\eta_0(x_{0,j})|\leqslant\tfrac{r_n}{2p_0}\right\};\qquad \widehat{E}_4=\left\{D_{n,-}(2r_n)\geqslant\delta+\Delta_n\right\}.$$

  Follow the same arguments for proving (B.7), we have the following facts:

  (1) with $\Delta_n>0$,

  $$\mathbb{P}^{\otimes n}\left(\widehat{D}_n(r_n,0,0)<\delta+\Delta_n\right)\leqslant\mathbb{P}^{\otimes n}\left(\widehat{D}_n(r_n,0,0)<D_-(0)-\Delta_n\right)\leqslant\sum_{j=2}^{4}\mathbb{P}^{\otimes n}\left(\widehat{E}_j^c\right).$$

  (2) for $2L_\eta(p_1\phi_{n,1}\vee p_0\phi_{n,0})<r_n<2U_\eta(p_1\wedge p_0)$, i.e., when $L_\eta\phi_{n,a}<\varepsilon/(2p_a)<U_\eta$ for $a\in\{0,1\}$,

  $$\mathbb{P}^{\otimes n}(\widehat{E}_4^c)\leqslant c_{1,\eta}\exp\left(-\frac{c_{2,\eta}}{4p_1^2}\left(\frac{\varepsilon}{\phi_{n,1}}\right)^2\right)\text{ and }\mathbb{P}^{\otimes n}(\widehat{E}_3^c)\leqslant c_{1,\eta}\exp\left(-\frac{c_{2,\eta}}{4p_0^2}\left(\frac{\varepsilon}{\phi_{n,0}}\right)^2\right).$$

  Now, to bound $\mathbb{P}^{\otimes n}(\widehat{E}_4^c)$. We have $t_\delta^\star>0$ when $D_-(0)>\delta$. For $r_n<t_\delta^\star/4$,

  $$\delta+\Delta_n-D_-(2r_n)\leqslant D_+(t_\delta^\star)+\Delta_n-D_-(t_\delta^\star/2)\leqslant\Delta_n.$$

  Then, for $\Delta_n\leqslant\sqrt{(p_1\wedge p_0)/2}$, by Lemma B.7,

  $$\mathbb{P}^{\otimes n}\left(\widehat{E}_4^c\right)=\mathbb{P}^{\otimes n}\left(D_{n,-}(2r_n)\geqslant\delta+\Delta_n\right)=\mathbb{P}^{\otimes n}\left((D_{n,-}-D_-)(2r_n)\geqslant\delta+\Delta_n-D_-(2r_n)\right)$$
  $$=\mathbb{P}^{\otimes n}\left((D_{n,-}-D_-)(2r_n)\geqslant\Delta_n\right)\leqslant 8\exp\left(\frac{n(p_1\wedge p_0)\Delta_n^2}{4}\right).$$

  In conclusion, by taking $L_\delta=2L_\eta(p_1\vee p_0)$, $U_\delta=(t_\delta^\star/4)\wedge(2U_\gamma(p_1\wedge p_0))$ and $U_{\Delta,\delta}=\sqrt{(p_1\wedge p_0)/2}$, we have that, for $L_\delta(\phi_{n,1}\vee\phi_{n,0})<r_n<U_\delta$ and $0<\Delta_n<U_{\Delta,\delta}$, (B.19) holds with with $c_{1,\delta}=2c_{1,\eta}$, $c_{2,\delta}=c_{2,\eta}/(4(p_1\vee p_0)^2)$, $c_{3,\delta}=8$ and $c_{4,\delta}=(p_1\wedge p_0)/4$.

F.14 Proof of Lemma B.14

During the proof, we denote $\mathcal{E}_{\pi,a} = \psi_{n,1,\pi}(\ell_{n,a}) + \sum_{j\in\{-,+\}} I(\delta = D_j(t^\star_\delta))\psi_{n,2,\pi}(\xi(\ell_{n,a}, r_n))$. For the first two terms of (B.21), note that

$$\begin{aligned}
&I\left(\eta_a(x) > T^\star_{\delta,1} + \ell_{n,a} + |\widehat{\eta}_a(x) - \eta_a(x)| + |\widehat{T}_{\delta,a} - T^\star_{\delta,a}|\right) \\
&\leqslant I\left(\widehat{\eta}_a(x) > \widehat{T}_{\delta,a} + \ell_{n,a}\right) = I\left(\eta_a(x) > T^\star_{\delta,a} + \ell_{n,a} + \eta_a(x) - \widehat{\eta}_a(x) + \widehat{T}_{\delta,a} - T^\star_{\delta,a}\right) \qquad \text{(F.8)} \\
&\leqslant I\left(\eta_a(x) > T^\star_{\delta,a} + \ell_{n,a} - |\widehat{\eta}_a(x) - \eta_a(x)| - |\widehat{T}_{\delta,a} - T^\star_{\delta,a}|\right).
\end{aligned}$$

Using the margin condition from Definition 4.4, it follows that, on the event that $\sup_{x\in\Omega}|\widehat{\eta}_a(x) - \eta_a(x)| \leqslant \ell_{n,a}/2$ and $|\widehat{T}_{\delta,a} - T^\star_{\delta,a}| \leqslant \ell_{n,a}/2$,

$$\mathbb{P}_{X|A=a}(\eta_a(X) > T^\star_{\delta,a}) - U_\gamma(4p_a\ell_{n,a})^\gamma \leqslant \mathbb{P}_{X|A=a}(\widehat{\eta}_a(X) > \widehat{T}_{\delta,a} + \ell_{n,a}) \leqslant \mathbb{P}_{X|A=a}(\eta_a(X) > T^\star_{\delta,a}). \qquad \text{(F.9)}$$

Thus, if we take $L_{\pi_1} = L_t$, $U_{\pi_1} = U_T$ and $U_{\Delta,\pi_1} = U_{\Delta,T}$, by Corollary 6.3 and as $(\widehat{\eta}_1, \widehat{\eta}_0)$ are $(\phi_{n,1}, \phi_{n,0})$-pointwise convergent, there exist constants $c_{i,\pi} > 0.i \in [4]$ such that, for $L_{\pi_1}(\phi_{n,1} \vee \phi_{n,0}) < r_n < U_{\pi_1}$, $2(g_{\delta,-}(4r_n) \vee g_{\delta,+}(4r_n)) < \Delta_n < U_{\Delta,\pi_1}$, $\varepsilon > U_\gamma(4p_a\ell_{n,a})^\gamma$ and $L_{\pi_1}(\phi_{n,1} \vee \phi_{n,0}) < \ell_{n,a}/2 < r_n$, with the functions $\psi_{n,j,\iota}$ from (6.2),

$$\begin{aligned}
&\mathbb{P}^{\otimes n}(\widehat{\pi}_{a,+} > \pi^\star_{a,+} + \varepsilon) \leqslant \mathbb{P}^{\otimes n}(\widehat{\pi}_{a,+} > \pi^\star_{a,+}) \\
&\leqslant \mathbb{P}^{\otimes n}\left(\sup_{x\in\Omega}|\widehat{\eta}_a(x) - \eta_a(x)| > \frac{\ell_{n,a}}{2}\right) + \mathbb{P}^{\otimes n}\left(|\widehat{T}_{\delta,a} - T^\star_{\delta,a}| > \frac{\ell_{n,a}}{2}\right) \\
&\leqslant \psi_{n,1,\eta}\left(\frac{\ell_{n,a}}{2}\right) + \psi_{n,1,T}\left(\frac{\ell_{n,a}}{2}\right) + \sum_{j\in\{-,+\}} I(\delta = D_j(t^\star_\delta))\psi_{n,2,T}\left(\xi\left(\frac{\ell_{n,a}}{2}, r_n\right)\right) \\
&\leqslant \psi_{n,1,\pi}(\ell_{n,a}) + \sum_{j\in\{-,+\}} I(\delta = D_j(t^\star_\delta))\psi_{n,2,\pi}(\xi(\ell_{n,a}, r_n)) = \mathcal{E}_{\pi,a}.
\end{aligned}$$

Similarly,

$$\mathbb{P}^{\otimes n}(\widehat{\pi}_{a,+} < \pi^\star_{a,+} - \varepsilon) \leqslant \mathbb{P}^{\otimes n}(\widehat{\pi}_{a,+} < \pi^\star_{a,+} - U_\gamma(4p_a\ell_{n,a})^\gamma) \leqslant \mathcal{E}_{\pi,a}.$$

For the last two terms of (B.21), on the event that $\sup_{x\in\Omega}|\widehat{\eta}_1(x) - \eta_1(x)| \leqslant \ell_{n,1}/2$ and $|\widehat{T}_{\delta,1} - T^\star_{\delta,1}| < \ell_{n,1}/2$,

$$I(\eta_1(x) = T^\star_{\delta,1}) \leqslant I(|\widehat{\eta}_1(x) - \widehat{T}_{\delta,1}| \leqslant \ell_{n,1}) \leqslant I(|\eta_1(x) - T^\star_{\delta,1}| \leqslant \ell_{n,1} + |\widehat{\eta}_1(x) - \eta_1(x)| + |\widehat{T}_{\delta,1} - T^\star_{\delta,1}|).$$

Again, by the margin condition from Definition 4.4, on the event that $\sup_{x\in\Omega}|\widehat{\eta}_1(x) - \eta_1(x)| \leqslant \ell_{n,1}/2$ and $|\widehat{T}_{\delta,1} - T^\star_{\delta,1}| < \ell_{n,1}/2$,

$$\mathbb{P}_{X|A=1}(\eta_1(x) = T^\star_{\delta,1}) + 2U_\gamma(4p_1\ell_{n,1})^\gamma \geqslant \mathbb{P}_{X|A=1}(|\widehat{\eta}_1(x) - \widehat{T}_{\delta,1}| \leqslant \ell_{n,1}) \geqslant \mathbb{P}_{X|A=1}(\eta_1(x) = T^\star_{\delta,1}).$$

Thus, by Corollary 6.3 and the fact that $(\widehat{\eta}_1, \widehat{\eta}_0)$ are $(\phi_{n,1}, \phi_{n,0})$-pointwise convergent, we have, for $L_{\pi_1}(\phi_{n,1} \vee \phi_{n,0}) < r_n < U_{\pi_1}$, $2(g_{\delta,-}(4r_n) \vee g_{\delta,+}(4r_n)) < \Delta_n < U_{\Delta,\pi_1}$, $\varepsilon > U_\gamma(4p_a\ell_{n,a})^\gamma$ and $(L_\eta \vee L_T)(\phi_{n,1} \vee \phi_{n,0}) < \ell_{n,a}/2 < r_n$ that

$$\mathbb{P}^{\otimes n}(\widehat{\pi}_{1,=} > \pi^\star_{1,=} + \varepsilon) \leqslant \mathbb{P}^{\otimes n}(\widehat{\pi}_{1,=} > \pi^\star_{1,=} + 2U_\gamma(4p_1\ell_{n,1})^\gamma) \leqslant \mathcal{E}_{\pi,a}$$

and $\mathbb{P}^{\otimes n}(\widehat{\pi}_{1,=} < \pi^\star_{1,=} - \varepsilon) \leqslant \mathbb{P}^{\otimes n}(\widehat{\pi}_{1,=} < \pi^\star_{1,=}) \leqslant \mathcal{E}_{\pi,a}$, finishing the proof.

F.15 Proof of Lemma B.15

During the proof, we denote,

$$\mathcal{E}'_{\pi,a} = \psi_{n,1,\pi}(\ell_{n,a}) + \sum_{j\in\{,+} I(\delta = D_j(t^\star_\delta)) \cdot \psi_{n,2,\pi}(g_{\delta,j}(\ell_{n,a})) + 4\exp\left(-\frac{np_a\varepsilon^2}{4}\right)$$

For $a \in \{0,1\}$, define the following events:

$$E_{a,1} = \left\{ \max_{j\in[n_a]} |\widehat{\eta}_a(x_{a,j}) - \eta_a(x_{a,j})| \leqslant \ell_{n,a}/2 \right\}, \qquad E_{a,2} = \left\{ |\widehat{T}_{\delta,a} - T^\star_{\delta,a}| \leqslant \ell_{n,a}/2 \right\},$$

$$E_{a,3} = \left\{ \frac{1}{n_a}\sum_{j=1}^{n_a} I\left(\eta_a(x_{a,j}) > T^\star_{\delta,a}\right) - \mathbb{P}_{X|A=a}\left(\eta_a(X) > T^\star_{\delta,a}\right) \leqslant \varepsilon \right\},$$

$$E_{a,4} = \left\{ \frac{1}{n_a}\sum_{j=1}^{n_a} I(\eta_a(x_{a,j}) \geqslant T^\star_{\delta,a} + 2\ell_{n,a}) - \mathbb{P}_{X|A=a}(\eta_a(X) > T^\star_{\delta,a} + 2\ell_{n,a}) \geqslant -\frac{\varepsilon}{2} \right\},$$

$$E_{a,5} = \left\{ \frac{1}{n_a}\sum_{j=1}^{n_a} I(|\eta_a(x_{a,j}) - T^\star_{\delta,a}| \leqslant 2\ell_{n,a}) - \mathbb{P}_{X|A=a}(|\eta_a(X) - T^\star_{\delta,a}| \leqslant 2\ell_{n,a}) \leqslant \frac{\varepsilon}{2} \right\},$$

$$E_{a,6} = \left\{ \frac{1}{n_a}\sum_{j=1}^{n_a} I(\eta_a(x_{a,j}) = T^\star_{\delta,a}) - \mathbb{P}_{X|A=a}(\eta_a(X) = T^\star_{\delta,a}) \geqslant -\varepsilon \right\}.$$

We recall the functions $\psi_{n,j,\iota}$ from (6.2), for $j \in \{1,2\}$, defined by constants $c_{i,\iota}$ with $\iota \in \{\eta, T, \pi\}$ and $i \in [4]$. For the first two terms in (B.22), we have from (F.8) with $x = x_{a,j}$ that, when $E_{a,i}$, $i \in [3]$ hold,

$$\begin{aligned}\widehat{\pi}_{n,a,+} &= \frac{1}{n_a}\sum_{j=1}^{n_a} I\left(\widehat{\eta}_a(x_{a,j}) > \widehat{T}_{\delta,a} + \ell_{n,a}\right) \leqslant \frac{1}{n_a}\sum_{j=1}^{n_a} I\left(\eta_a(x_{a,j}) > T^\star_{\delta,a}\right) \\ &\leqslant \mathbb{P}_{X|A=1}\left(\eta_a(X) > T^\star_{\delta,a}\right) + \varepsilon = \pi^\star_{a,+} + \varepsilon.\end{aligned}$$

Moreover, when $E_{a,1}$, $E_{a,2}$ and $E_{a,4}$ hold with $\varepsilon \geqslant 4U_\gamma(4p_a\ell_{n,a})^\gamma$, by the margin condition from Definition 4.4 and the definition of $\pi^\star_{a,+}$ from (B.20),

$$\begin{aligned}\widehat{\pi}_{n,a,+} &= \frac{1}{n_a}\sum_{j=1}^{n_a} I\left(\widehat{\eta}_a(x_{a,j}) > \widehat{T}_{\delta,a} + \ell_{n,a}\right) \geqslant \frac{1}{n_a}\sum_{j=1}^{n_a} I\left(\eta_a(x_{a,j}) > T^\star_{\delta,a} + 2\ell_{n,a}\right) \\ &\geqslant \mathbb{P}_{X|A=1}\left(\eta_a(X) > T^\star_{\delta,a} + 2\ell_{n,a}\right) - \frac{\varepsilon}{2} \geqslant \pi^\star_{a,+} - \frac{\varepsilon}{2} - \mathbb{P}_{X|A=a}\left(T^\star_{\delta,a} < \eta_a(X) \leqslant T^\star_{\delta,a} + 2\ell_{n,a}\right) \\ &\geqslant \pi^\star_{a,+} - \frac{\varepsilon}{2} - U_\gamma(4p_a\ell_{n,a})^\gamma \geqslant \pi^\star_{a,+} - \varepsilon.\end{aligned}$$

Thus, by Corollary 6.3, Lemma B.6, Lemma B.8 and the fact that $(\widehat{\eta}_1, \widehat{\eta}_0)$ are $(\phi_{n,1}, \phi_{n,0})$-pointwise convergent, we have, for $L_{\pi_1}(\phi_{n,1} \vee \phi_{n,0}) < r_n < U_{\pi_1}$, $2\left(g_{\delta,-}(4r_n) \vee g_{\delta,+}(4r_n)\right) < \Delta_n < U_{\Delta,\pi_1}$, $4U_\gamma(4p_a\ell_{n,a})^\gamma \leqslant \varepsilon \leqslant \sqrt{p_a/2}$ and $L_{\pi_1}(\phi_{n,1} \vee \phi_{n,0}) \leqslant \ell_{n,a}/2 \leqslant r_n$, with $c_{i,\pi} > 0, i \in [4]$, that

$$\begin{aligned}&\mathbb{P}^{\otimes n}(\widehat{\pi}_{n,a,+} > \pi^\star_{a,+} + \varepsilon) \leqslant \sum_{j=1}^{3} \mathbb{P}^{\otimes n}(E^c_{a,j}) \\ &\leqslant \psi_{n,1,\eta}\left(\frac{\ell_{n,a}}{2}\right) + \psi_{n,1,T}\left(\frac{\ell_{n,a}}{2}\right) + \sum_{j\in\{-,+\}} I(\delta = D_j(t^\star_\delta)\cdot\psi_{n,2,T}\left(g_{\delta,j}\xi\left(\frac{\ell_{n,a}}{2}, r_n\right)\right) + 4\exp\left(-np_a\varepsilon^2\right) \\ &\leqslant \psi_{n,1,\pi}\left(\ell_{n,a}\right) + \sum_{j\in\{-,+\}} I(\delta = D_j(t^\star_\delta)\psi_{n,2,\pi}\left(g_{\delta,j}\left(\xi\left(\ell_{n,a}, r_n\right)\right)\right) + 4\exp\left(-np_a\varepsilon^2\right) = \mathcal{E}'_{\pi,a}.\end{aligned}$$

Similarly,

$$\mathbb{P}^{\otimes n}\left(\widehat{\pi}_{n,a,+} < \pi^\star_{a,+} - \varepsilon\right) \leqslant \sum_{j=1,2,4} \mathbb{P}^{\otimes n}(E^c_{a,j}) \leqslant \mathcal{E}'_{\pi,a}.$$

For the last two terms in (B.22), note that

$$I\left(|\eta_a(x) - T^\star_{\delta,a}| \leqslant \ell_{n,a} - |\widehat{\eta}_a(x) - \eta_a(x)| - |\widehat{T}_{\delta,a} - T^\star_{\delta,a}|\right) \leqslant I\left(|\widehat{\eta}_a(x) - \widehat{T}_{\delta,a}| \leqslant \ell_{n,a}\right)$$

$$\leqslant I\left(|\eta_a(x) - T^\star_{\delta,a}| \leqslant \ell_{n,a} + |\widehat{\eta}_a(x) - \eta_a(x)| + |\widehat{T}_{\delta,a} - T^\star_{\delta,a}|\right).$$

When $E_{a,1}$, $E_{a,2}$ and $E_{a,5}$ hold with $\varepsilon > 4U_\gamma(4p_a\ell_{n,a})^\gamma$, by the margin condition from Definition 4.4,

$$\begin{aligned}
\widehat{\pi}_{n,a,=} &= \frac{1}{n_a}\sum_{j=1}^{n_a} I\left(|\widehat{\eta}_a(x_{a,j}) - \widehat{T}_{\delta,a}| \leqslant \ell_{n,a}\right) \leqslant \frac{1}{n_a}\sum_{j=1}^{n_a} I\left(|\eta_a(x_{a,j}) - T^\star_{\delta,a}| \leqslant 2\ell_{n,a}\right) \\
&\leqslant \mathbb{P}_{X|A=1}\left(|\eta_a(X) - T^\star_{\delta,a}| \leqslant 2\ell_{n,a}\right) + \frac{\varepsilon}{2} = \pi^\star_{a,=} + \frac{\varepsilon}{2} + \mathbb{P}_{X|A=1}\left(0 < |\eta_a(X) - T^\star_{\delta,a}| \leqslant 2\ell_{n,a}\right) \\
&\leqslant \pi^\star_{a,=} + \frac{\varepsilon}{2} + 2U_\gamma(4p_a\ell_{n,a})^\gamma \leqslant \pi^\star_{a,=} + \varepsilon.
\end{aligned}$$

and when $E_{a,1}$, $E_{a,2}$ and $E_{a,6}$ hold,

$$\begin{aligned}
\widehat{\pi}_{n,a,=} &= \frac{1}{n_a}\sum_{j=1}^{n_a} I\left(|\widehat{\eta}_a(x_{a,j}) - \widehat{T}_{\delta,a}| > \ell_{n,a}\right) \geqslant \frac{1}{n_a}\sum_{j=1}^{n_a} I\left(\eta_a(x_{a,j}) = T^\star_{\delta,a}\right) \\
&\geqslant \mathbb{P}_{X|A=1}\left(\eta_a(X) = T^\star_{\delta,a}\right) - \varepsilon = \pi^\star_{a,=} - \varepsilon.
\end{aligned}$$

Thus, we have, for $L_{\pi_1}(\phi_{n,1} \vee \phi_{n,0}) < r_n < U_{\pi_1}$, $2\left(g_{\delta,-}(4r_n) \vee g_{\delta,+}(4r_n)\right) < \Delta_n < U_{\Delta,\pi_1}$, $4U_\gamma(4p_a\ell_{n,a})^\gamma \leqslant \varepsilon \leqslant \sqrt{p_a/2}$ and $(L_\eta \vee L_T)(\phi_{n,1} \vee \phi_{n,0}) \leqslant \ell_{n,a}/2 \leqslant r_n$, with $c_{i,\pi} > 0, i \in [4]$ that

$$\mathbb{P}^{\otimes n}\left(\widehat{\pi}_{n,a,=} > \pi^\star_{a,=} + \varepsilon\right) \leqslant \mathbb{P}^{\otimes n}(E^c_{a,1}) + \mathbb{P}^{\otimes n}(E^c_{a,2}) + \mathbb{P}^{\otimes n}(E^c_{a,5}) \leqslant \mathcal{E}'_{\pi,a}.$$

Similarly,

$$\mathbb{P}^{\otimes n}\left(\widehat{\pi}_{n,a,=} < \pi^\star_{a,=} - \varepsilon\right) \leqslant \mathbb{P}^{\otimes n}(E^c_{a,1}) + \mathbb{P}^{\otimes n}(E^c_{a,2}) + \mathbb{P}^{\otimes n}(E^c_{a,6}) \leqslant \mathcal{E}'_{\pi,a}.$$

F.16 Proof of Lemma B.16

Recalling the function ρ from (5.8), we consider the following three cases in order: (1) $a \geqslant b > 0$; (2) $a \leqslant 0 < b$. and (3) $0 < a < b$.

Case 1: If $a \geqslant b > 0$, we have $\rho(a/b) = 1$. Moreover, with $0 < \varepsilon < b/2$, $(a+2\varepsilon)/(b-\varepsilon) > 1$ and $(a-2\varepsilon)/(b+\varepsilon) - 1 = (a-b-3\varepsilon)/(b+\varepsilon) > -3\varepsilon/b > -6\varepsilon/b$. It follows that, with $0 < \varepsilon < b/2$:

$$\rho\left(\frac{a+2\varepsilon}{b-\varepsilon}\right) - \rho\left(\frac{a}{b}\right) = 0 < \frac{6\varepsilon}{b} \quad \text{and} \quad \rho\left(\frac{a-2\varepsilon}{b+\varepsilon}\right) - \rho\left(\frac{a}{b}\right) > -\frac{6\varepsilon}{b}.$$

Case 2: If $a \leqslant 0 < b$, we have $\rho(a/b) = 0$. Moreover, with $0 < \varepsilon < b/2$, $(a+2\varepsilon)/(b-\varepsilon) m 4\varepsilon/b < \frac{6\varepsilon}{b}$ and $(a-2\varepsilon)/(b+\varepsilon) < 0$. It follows that, with $0 < \varepsilon < b/2$:

$$\rho\left(\frac{a+2\varepsilon}{b-\varepsilon}\right) - \rho\left(\frac{a}{b}\right) \leqslant \frac{4\varepsilon}{b} \quad \text{and} \quad \rho\left(\frac{a-2\varepsilon}{b+\varepsilon}\right) - \rho\left(\frac{a}{b}\right) = 0 \geqslant -\frac{6\varepsilon}{b}.$$

Case 3: If $0 < a < b$, we have $\rho(a/b) = a/b$. Moreover, with $0 < \varepsilon < b/2$,

$$\frac{a+2\varepsilon}{b-\varepsilon} - \frac{a}{b} = \frac{(a+2b)\varepsilon}{b(b-\varepsilon)} \leqslant \frac{3b\varepsilon}{b(b-\varepsilon)} \leqslant \frac{6\varepsilon}{b}$$

and

$$\frac{a-2\varepsilon}{b+\varepsilon} - \frac{a}{b} = \frac{(-a-2b)\varepsilon}{b(b+\varepsilon)} \geqslant \frac{-3b\varepsilon}{b(b+\varepsilon)} \geqslant -\frac{3\varepsilon}{b} \geqslant -\frac{6\varepsilon}{b}.$$

Hence (B.23) follows.

F.17 Proof of Lemma B.17

We consider the following two cases for $\pi^\star_{a,=}$ from (5.10): (1) $\pi^\star_{a,=}=0$ and (2) $\pi^\star_{a,=}>0$.

Case (1): When $\pi^\star_{a,=}=0$, since by (5.5), $0\leqslant\widehat{\tau}_{\delta,a}\leqslant 1$ and since $\widehat{\pi}_{a,=}\geqslant 0$ for $\widehat{\pi}_{a,=}$ from (B.20), for any $\varepsilon>0$, we have

$$\mathbb{P}^{\otimes n}\left(\widehat{\pi}_{a,=}\widehat{\tau}_{\delta,a}<\pi^\star_{a,=}\tau^\star_{\delta,a}-\varepsilon\right)\leqslant\mathbb{P}^{\otimes n}\left(\widehat{\pi}_{a,=}<0\right)=0.$$

Moreover, if we take $L_\pi=L_{\pi_1}$, $U_\pi=U_{\pi_1}$ $L_{\varepsilon,\pi}=2^{2+2\gamma}U_\gamma(p_1\vee p_0)^\gamma$ and $U_{\Delta,\pi}=U_{\Delta,\pi_1}$, by (B.21) from Lemma B.14, for $L_{\pi_1}(\phi_{n,1}\vee\phi_{n,0})<r_n<U_{\pi_1}$, $2\left(g_{\delta,-}(4r_n)\vee g_{\delta,+}(4r_n)\right)<\Delta_n<U_\Delta,\pi_1$, $L_{\varepsilon,\pi}(\ell_{n,1}\vee\ell_{n,0})^\gamma\leqslant\varepsilon\leqslant\sqrt{p_a/2}$ and $(L_\eta\vee L_T)(\phi_{n,1}\vee\phi_{n,0})\leqslant\ell_{n,a}/2\leqslant r_n$, we have

$$\begin{aligned}&\mathbb{P}^{\otimes n}\left(\widehat{\pi}_{a,=}\widehat{\tau}_{\delta,a}>\pi^\star_{a,=}\tau^\star_{\delta,a}+\varepsilon\right)=\mathbb{P}^{\otimes n}\left(\widehat{\pi}_{a,=}\widehat{\tau}_{\delta,a}>\varepsilon\right)\leqslant\mathbb{P}^{\otimes n}\left(\widehat{\pi}_{a,=}>\pi^\star_{a,=}+\varepsilon\right)\\&\leqslant\psi_{n,1,\pi}\left(\ell_{n,a}\right)+\sum_{j\in\{-,+\}}I(\delta=D_j(t^\star_\delta))\psi_{n,2,\pi}\left(g_{\delta,j}\omega(\ell_{n,a},r_n)\right).\end{aligned}$$

Thus, (B.24) holds with $c_{5,\pi}=c_{6,\pi}=0$.

Case (2): When $\pi^\star_{a,=}>0$, we distinguish the following two cases:

(A). When $\widehat{\delta}=\widetilde{\delta}$, $\widehat{\pi}_{n,a,+}\geqslant\pi^\star_{a,+}-\pi^\star_{a,=}\varepsilon/12$, $\widehat{\pi}_{n,1-a,+}\leqslant\pi^\star_{1-a,+}+\pi^\star_{a,=}\varepsilon/12$, $\widehat{\pi}_{n,a,=}\geqslant\pi^\star_{a,=}-\pi^\star_{a,=}\varepsilon/12$ and $\widehat{\pi}_{a,=}\leqslant\pi^\star_{a,=}+\varepsilon/2$, recalling $\widehat{\delta}_a$ from Algorithm 1, and using Lemma B.16, we have

$$\begin{aligned}&\widehat{\pi}_{a,=}\widehat{\tau}_{\delta,a}=\widehat{\pi}_{a,=}\rho\left(\frac{\widehat{\pi}_{n,1-a,+}-\widehat{\pi}_{n,a,+}+\widehat{\delta}}{\widehat{\pi}_{n,a,=}}\right)\leqslant\left(\pi^\star_{a,=}+\frac{\varepsilon}{2}\right)\rho\left(\frac{\pi^\star_{1-a,+}-\pi^\star_{a,+}+\widetilde{\delta}+\varepsilon\pi^\star_{a,=}/6}{\pi^\star_{a,=}-\varepsilon\pi^\star_{a,=}/12}\right)\\&\leqslant\left(\pi^\star_{a,=}+\frac{\varepsilon}{2}\right)\left(\rho\left(\frac{\pi^\star_{1-a,+}-\pi^\star_{a,+}+\delta}{\pi^\star_{a,=}}\right)+\frac{\varepsilon}{2}\right)\leqslant\pi^\star_{a,=}\tau^\star_{\delta,a}+\varepsilon;\end{aligned}$$

Thus, if we take $L_\pi=L_{\pi_1}\vee L_\delta$, $U_\pi=U_{\pi_1}\wedge U_\delta$, $L_{\varepsilon,\pi}=48U_\gamma\left(((4p_1)^\gamma/\pi^\star_{1,=})\vee((4p_0)^\gamma/\pi^\star_{0,=})\right)$ and $U_{\Delta,\pi}=U_{\Delta,\pi_1}\wedge U_{\Delta,\delta}$, by Lemmas B.13, B.14 and B.15, we have, with $L_\pi(\phi_{n,1}\vee\phi_{n,0})<r_n<U_\pi$, $\Delta_n<U_{\Delta,\pi}$, $2(L_{\pi_1}\vee L_\delta)(\phi_{n,1}\vee\phi_{n,0})<\ell_{n,1},\ell_{n,0}<2r_n$, and $L_{\varepsilon,\pi}(\ell_{n,1}\vee\ell_{n,0})^\gamma<\varepsilon\leqslant\sqrt{p_a/2}$, that

$$\begin{aligned}&\mathbb{P}^{\otimes n}\left(\widehat{\pi}_{a,=}\rho\left(\frac{\widehat{\pi}_{n,1-a,+}-\widehat{\pi}_{n,a,+}+\widehat{\delta}_a}{\widehat{\pi}_{n,a,=}}\right)>\pi^\star_{a,=}\rho\left(\frac{\pi^\star_{1-a,+}-\pi^\star_{a,+}+\delta}{\pi^\star_{a,=}}\right)+\varepsilon\right)\\&\leqslant\mathbb{P}^{\otimes n}\left(\widehat{\pi}_{n,a,+}<\pi^\star_{a,+}-\frac{\pi^\star_{a,=}\varepsilon}{12}\right)+\mathbb{P}^{\otimes n}\left(\widehat{\pi}_{n,1-a,+}>\pi^\star_{1-a,+}+\frac{\pi^\star_{a,=}\varepsilon}{12}\right)\\&+\mathbb{P}^{\otimes n}\left(\widehat{\pi}_{n,a,=}<\pi^\star_{a,=}-\frac{\pi^\star_{a,=}\varepsilon}{12}\right)+\mathbb{P}^{\otimes n}\left(\widehat{\pi}_{a,=}>\pi^\star_{a,=}+\frac{\varepsilon}{2}\right)\\&\leqslant 4\psi_{n,1,\pi}\left(\ell_{n,a}\right)+4\sum_{j\in\{-,+\}}I(\delta=D_j(t^\star_\delta))\psi_{n,2,\pi}\left(g_{\delta,j}\omega(\ell_{n,a},r_n)\right)+12\exp\left(-\frac{np_a(\pi^\star_{a,=})^2\varepsilon^2}{576}\right).\end{aligned}$$

Choosing $c_{5,\pi}=12$ and $c_{6,\pi}=p_a\pi^{\star 2}_{a,=}/576$ proves the first claimed inequality.

(B). When $\widehat{\pi}_{n,a,+}\leqslant\pi^\star_{a,+}+\pi^\star_{a,=}\varepsilon/12$, $\widehat{\pi}_{n,1-a,+}\geqslant\pi^\star_{1-a,+}-\pi^\star_{a,=}\varepsilon/12$, $\widehat{\pi}_{n,a,=}\leqslant\pi^\star_{a,=}+\pi^\star_{a,=}\varepsilon/12$ and $\widehat{\pi}_{a,=}\geqslant\pi^\star_{a,=}-\varepsilon/2$, we have

$$\begin{aligned}&\widehat{\pi}_{a,=}\rho\left(\frac{\widehat{\pi}_{n,1-a,+}-\widehat{\pi}_{n,a,+}+\widehat{\delta}_a}{\widehat{\pi}_{n,a,=}}\right)\geqslant\left(\pi^\star_{a,=}-\frac{\varepsilon}{2}\right)\rho\left(\frac{\pi^\star_{1-a,+}-\pi^\star_{a,+}-\delta-\varepsilon\pi^\star_{a,=}/6}{\pi^\star_{a,=}+\varepsilon\pi^\star_{a,=}/12}\right)\\&\geqslant\left(\pi^\star_{a,=}-\frac{\varepsilon}{2}\right)\left(\rho\left(\frac{\pi^\star_{1-a,+}-\pi^\star_{a,+}-\delta}{\pi^\star_{a,=}}\right)-\frac{\varepsilon}{2}\right)\geqslant\pi^\star_{a,=}\rho\left(\frac{\pi^\star_{1-a,+}-\pi^\star_{a,+}-\delta}{\pi^\star_{a,=}}\right)-\varepsilon.\end{aligned}$$

Thus, similarly to case (A),

$$\mathbb{P}^{\otimes n}\left(\widehat{\pi}_{a,=}\rho\left(\frac{\widehat{\pi}_{n,1-a,+}-\widehat{\pi}_{n,a,+}+\widehat{\delta}_a}{\widehat{\pi}_{n,a,=}}\right)<\pi^\star_{a,=}\rho\left(\frac{\pi^\star_{1-a,+}-\pi^\star_{a,+}-\delta}{\pi^\star_{a,=}}\right)-\varepsilon\right)$$

$$\leqslant \mathbb{P}^{\otimes n}\left(\widehat{\pi}_{n,a,+} > \pi^\star_{a,+} + \frac{\pi^\star_{a,=}\varepsilon}{12}\right) + \mathbb{P}^{\otimes n}\left(\widehat{\pi}_{n,1-a,+} < \pi^\star_{1-a,+} - \frac{\pi^\star_{a,=}\varepsilon}{12}\right)$$
$$+ \mathbb{P}^{\otimes n}\left(\widehat{\pi}_{n,a,=} > \pi^\star_{a,=} + \frac{\pi^\star_{a,=}\varepsilon}{12}\right) + \mathbb{P}^{\otimes n}\left(\widehat{\pi}_{a,=} < \pi^\star_{a,=} - \frac{\varepsilon}{2}\right),$$

and the second claimed inequality follows as in case (A).

G Bayes-optimal Classifier for Data Distribution from Section 7.1

In this section, we derive the δ-fair Bayes optimal classifier and its misclassification rate for the data distribution proposed in Section 7.1. Let $p_a = P(A = 1)$ be the probability of A being 1, let μ_a the conditional density function of (X_1, X_2) given $A = a$, and let $\eta_a(X_1, X_2)$ be the probability of $Y = 1$ given $A = a$ and X_1, X_2. According to the construction, for $a \in \{0, 1\}$ and $(x_1, x_2) \in [-1, 1]^2$,

$$(1)\ p_a = \frac{1}{2},\ (2)\ \mu_a(x_1, x_2) \equiv \frac{1}{4} \text{ and } (3)\ \eta_a(x_1, x_2) = \frac{1 + (2a-1)s_1}{2} + \frac{s_2 \cdot \operatorname{sign}(x_1)}{2}\left(|x_1|(1-|x_2|)\right)^\beta. \quad \text{(G.1)}$$

In the following, we define the half cubes $\mathbb{B}^2_+$ and $\mathbb{B}^2_-$ as

$$\mathbb{B}^2_+ = \left\{x = (x_1, x_2) \in [-1, 1]^2, x_1 \geqslant 0\right\} \text{ and } \mathbb{B}^2_- = \left\{x = (x_1, x_2) \in [-1, 1]^2, x_1 < 0\right\},$$

respectively. As $(|x_1|(1-|x_2|))^\beta \in [0, 1]$ on $[-1, 1]^2$, we have

$$\begin{cases} \eta_a(x_1, x_2) \in \left[\frac{1+(2a-1)s_1}{2}, \frac{1+(2a-1)s_1+s_2}{2}\right] \text{ on } \mathbb{B}^2_+; \\ \eta_a(x_1, x_2) \in \left[\frac{1+(2a-1)s_1-s_2}{2}, \frac{1+(2a-1)s_1}{2}\right] \text{ on } \mathbb{B}^2_-. \end{cases} \quad \text{(G.2)}$$

In order to derive $D_-(t)$, we first calculate $\mathbb{P}_{X_1,X_2|A=1}(\eta_1(X_1, X_2) > 1/2 + t)$ and $\mathbb{P}_{X_1,X_2|A=0}(\eta_0(X_1, X_2) > 1/2 - t)$. For $\mathbb{P}_{X_1,X_2|A=1}(\eta_1(X_1, X_2) > 1/2 + t)$, we consider two cases: (1) $(s_1 - s_2)/2 \leqslant t \leqslant s_1/2$ and (2) $s_1/2 < t \leqslant (s_1 + s_2)/2$.

- Case (1): $(s_1 - s_2)/2 \leqslant t \leqslant s_1/2$.

  In this case, we have $(1 + s_1 - s_2)/2 \leqslant 1/2 + t \leqslant (1 + s_1)/2$. By (G.2),

$$\left\{\eta_1(x_1, x_2) > \frac{1}{2} + t\right\} \cap \mathbb{B}^2_+ = \mathbb{B}^2_+, \quad \text{(G.3)}$$

  and

$$\begin{aligned}
&\left\{\eta_1(x_1, x_2) > \frac{1}{2} + t\right\} \cap \mathbb{B}^2_- = \left\{\operatorname{sign}(x_1)\left(|x_1|\left(1-|x_2|\right)\right)^\beta > \frac{2t - s_1}{s_2}\right\} \cap \mathbb{B}^2_- \\
&= \left\{-\left(-x_1\left(1-|x_2|\right)\right)^\beta > \frac{2t-s_1}{s_2}\right\} \cap \mathbb{B}^2_- = \left\{-x_1\left(1-|x_2|\right) < \left(\frac{s_1-2t}{s_2}\right)^{\frac{1}{\beta}}\right\} \cap \mathbb{B}^2_- \\
&= \left\{(1-|x_2|) < -\frac{1}{x_1}\left(\frac{s_1-2t}{s_2}\right)^{\frac{1}{\beta}}\right\} \cap \mathbb{B}^2_- = \left\{-1 \leqslant x_1 < 0, 1 + \frac{1}{x_1}\left(\frac{s_1-2t}{s_2}\right)^{\frac{1}{\beta}} < |x_2| \leqslant 1\right\} \\
&= \left\{-1 \leqslant x_1 \leqslant -\left(\frac{s_1-2t}{s_2}\right)^{\frac{1}{\beta}}, 1 + \frac{1}{x_1}\left(\frac{s_1-2t}{s_2}\right)^{\frac{1}{\beta}} < |x_2| \leqslant 1\right\} \\
&\cup \left\{-\left(\frac{s_1-2t}{s_2}\right)^{\frac{1}{\beta}} < x_1 < 0, |x_2| \leqslant 1\right\}. \quad \text{(G.4)}
\end{aligned}$$

Here, the last equality holds since the event $\left\{|x_2| > 1 + ((s_1 - 2t)/s_2)^{(1/\beta)}/x_1\right\}$ is implied by $-((s_1 - 2t)/s_2)^{(1/\beta)} < x_1 < 0$. Then, by (G.1),

$$\begin{aligned}
&\mathbb{P}_{X_1,X_2|A=1}\left(\eta_1(X_1,X_2) > \frac{1}{2} + t\right) = \int_{\{\eta_1(x_1,x_2) > \frac{1}{2}+t\}} \mu_1(x_1,x_2)dx_1dx_2 \\
&= \frac{1}{4}\left(\int_{\{\eta_1(x_1,x_2) > \frac{1}{2}+t\}\cap\mathbb{B}_+^2} dx_1dx_2 + \int_{\{\eta_1(x_1,x_2) > \frac{1}{2}+t\}\cap\mathbb{B}_-^2} dx_1dx_2\right) \\
&= \frac{1}{2} + \frac{1}{4}\int_{-\left(\frac{s_1-2t}{s_2}\right)^{\frac{1}{\beta}}}^{0}\left(\int_{-1}^{1} dx_2\right)dx_1 + \frac{1}{4}\int_{-1}^{-\left(\frac{s_1-2t}{s_2}\right)^{\frac{1}{\beta}}}\left(\int_{1+\frac{1}{x_1}\left(\frac{2t-s_1}{s_2}\right)^{\frac{1}{\beta}} < |x_2| \leqslant 1} dx_2\right)dx_1 \\
&= \frac{1}{2} + \frac{1}{2}\int_{-\left(\frac{s_1-2t}{s_2}\right)^{\frac{1}{\beta}}}^{0} dx_1 + \frac{1}{2}\int_{-1}^{-\left(\frac{s_1-2t}{s_2}\right)^{\frac{1}{\beta}}}\left(-\frac{1}{x_1}\left(\frac{s_1-2t}{s_2}\right)^{\frac{1}{\beta}}\right)dx_1 \\
&= \frac{1}{2} + \frac{1}{2}\left(\frac{s_1-2t}{s_2}\right)^{\frac{1}{\beta}} - \frac{1}{2}\left(\frac{s_1-2t}{s_2}\right)^{\frac{1}{\beta}}\left(\ln(-x_1)\Big|_{-1}^{-\left(\frac{s_1-2t}{s_2}\right)^{\frac{1}{\beta}}}\right) \\
&= \frac{1}{2} + \frac{1}{2}\left(\frac{s_1-2t}{s_2}\right)^{\frac{1}{\beta}}\left(1 - \frac{1}{\beta}\ln\left(\frac{s_1-2t}{s_2}\right)\right).
\end{aligned}$$

- Case (2): $s_1/2 \leqslant t \leqslant (s_1+s_2)/2$.
  In this case, we have $(1+s_1)/2 < 1/2 + t < (1+s_1+s_2)/2$. Again, by (G.2),

$$\left\{\eta_1(x_1,x_2) > \frac{1}{2} + t\right\} \cap \mathbb{B}_-^2 = \emptyset, \tag{G.5}$$

and

$$\begin{aligned}
&\left\{\eta_1(x_1,x_2) > \frac{1}{2} + t\right\} \cap \mathbb{B}_+^2 = \left\{\operatorname{sign}(x_1)\left(|x_1|\left(1 - |x_2|\right)\right)^{\beta} > \frac{2t-s_1}{s_2}\right\} \cap \mathbb{B}_+^2 \\
&= \left\{x_1\left(1-|x_2|\right) > \left(\frac{2t-s_1}{s_2}\right)^{\frac{1}{\beta}}\right\} \cap \mathbb{B}_+^2 = \left\{\left(1-|x_2|\right) > \frac{1}{x_1}\left(\frac{2t-s_1}{s_2}\right)^{\frac{1}{\beta}}\right\} \cap \mathbb{B}_+^2 \\
&= \left\{0 \leqslant x_1 \leqslant 1, |x_2| < 1 - \frac{1}{x_1}\left(\frac{2t-s_1}{s_2}\right)^{\frac{1}{\beta}}\right\} \\
&= \left\{\left(\frac{2t-s_1}{s_2}\right)^{\frac{1}{\beta}} \leqslant x_1 \leqslant 1, |x_2| < 1 - \frac{1}{x_1}\left(\frac{2t-s_1}{s_2}\right)^{\frac{1}{\beta}}\right\}.
\end{aligned} \tag{G.6}$$

Here, the last equality holds since the set $\left\{|x_2| < 1 - ((2t-s_1)/s_2)^{(1/\beta)}/x_1\right\}$ is empty set when $0 \leqslant x_1 < (2t-s_1)/s_2)^{(1/\beta)}$. Thus,

$$\begin{aligned}
&\mathbb{P}_{X_1,X_2|A=1}\left(\eta_1(X_1,X_2) > \frac{1}{2} + t\right) = \int_{\{\eta_1(x_1,x_2) > t\}} \mu_1(x_1,x_2)dx_1dx_2 \\
&= \frac{1}{4}\left(\int_{\{\eta_1(x_1,x_2) > t\}\cap\mathbb{B}_+^2} dx_1dx_2 + \int_{\{\eta_1(x_1,x_2) > t\}\cap\mathbb{B}_-^2} dx_1dx_2\right) \\
&= \frac{1}{4}\int_{\{\eta_1(x_1,x_2) > t\}\cap\mathbb{B}_+^2} dx_1dx_2 = \frac{1}{4}\int_{\left(\frac{2t-s_1}{s_2}\right)^{\frac{1}{\beta}}}^{1}\left(\int_{\frac{1}{x_1}\left(\frac{2t-s_1}{s_2}\right)^{\frac{1}{\beta}}-1}^{1-\frac{1}{x_1}\left(\frac{2t-s_1}{s_2}\right)^{\frac{1}{\beta}}} dx_2\right)dx_1 \\
&= \frac{1}{2}\int_{\left(\frac{2t-s_1}{s_2}\right)^{\frac{1}{\beta}}}^{1}\left(1 - \frac{1}{x_1}\left(\frac{2t-s_1}{s_2}\right)^{\frac{1}{\beta}}\right)dx_1
\end{aligned}$$

$$= \frac{1}{2}\left(1 - \left(\frac{2t-s_1}{s_2}\right)^{\frac{1}{\beta}}\right) - \frac{1}{2}\left(\frac{2t-s_1}{s_2}\right)^{\frac{1}{\beta}} \left(\ln x \Big|_{\left(\frac{2t-s_1}{s_2}\right)^{\frac{1}{\beta}}}^{1}\right)$$
$$= \frac{1}{2} - \frac{1}{2}\left(\frac{2t-s_1}{s_2}\right)^{\frac{1}{\beta}} \left(1 - \frac{1}{\beta}\ln\left(\frac{2t-s_1}{s_2}\right)\right).$$

In summary, we have,

$$\mathbb{P}_{X_1,X_2|A=1}\left(\eta_1(X) > \frac{1}{2}+t\right) = \begin{cases} 1, & t \leqslant \frac{s_1-s_2}{2}; \\ \frac{1}{2} + \frac{1}{2}\left(\frac{s_1-2t}{s_2}\right)^{\frac{1}{\beta}}\left(1 - \frac{1}{\beta}\ln\left(\frac{s_1-2t}{s_2}\right)\right), & \frac{s_1-s_2}{2} < t \leqslant \frac{s_1}{2}; \\ \frac{1}{2} - \frac{1}{2}\left(\frac{2t-s_1}{s_2}\right)^{\frac{1}{\beta}}\left(1 - \frac{1}{\beta}\ln\left(\frac{2t-s_1}{s_2}\right)\right), & \frac{s_1}{2} \leqslant t \leqslant \frac{s_1+s_2}{2}; \\ 0, & t > \frac{s_1+s_2}{2}. \end{cases} \tag{G.7}$$

To find $\mathbb{P}_{X_1,X_2|A=0}\left(\eta_0(X_1,X_2) \geqslant \frac{1}{2}-t\right)$, we note that $\eta_0(x_1,x_2) = \eta_1(x_1,x_2) - s_1$ and $\mu_0(x_1,x_2) \equiv \mu_1(x_1,x_2)$. Thus,

$$\begin{aligned} &\mathbb{P}_{X_1,X_2|A=0}\left(\eta_0(X_1,X_2) \geqslant \frac{1}{2}-t\right) = \int_{\{\eta_0(x_1,x_2)>\frac{1}{2}-t\}} \mu_0(x_1,x_2)dx_1dx_2 \\ &= \int_{\{\eta_1(x_1,x_2)>\frac{1}{2}+s_1-t\}} \mu_1(x_1,x_2)dx_1dx_2 = \mathbb{P}_{X_1,X_2|A=1}\left(\eta_1(X) > \frac{1}{2}+s_1-t\right) \\ &= \begin{cases} 1, & s_1 - t \leqslant \frac{s_1-s_2}{2}; \\ \frac{1}{2} + \frac{1}{2}\left(\frac{2t-s_1}{s_2}\right)^{\frac{1}{\beta}}\left(1 - \frac{1}{\beta}\ln\left(\frac{2t-s_1}{s_2}\right)\right), & \frac{s_1-s_2}{2} < s_1 - t \leqslant \frac{s_1}{2}; \\ \frac{1}{2} - \frac{1}{2}\left(\frac{s_1-2t}{s_2}\right)^{\frac{1}{\beta}}\left(1 - \frac{1}{\beta}\ln\left(\frac{s_1-2t}{s_2}\right)\right), & \frac{s_1}{2} < s_1 - t \leqslant \frac{s_1+s_2}{2}; \\ 0, & s_1 - t > \frac{s_1+s_2}{2}, \end{cases} \\ &= \begin{cases} 0, & t \leqslant \frac{s_1-s_2}{2}; \\ \frac{1}{2} - \frac{1}{2}\left(\frac{s_1-2t}{s_2}\right)^{\frac{1}{\beta}}\left(1 - \frac{1}{\beta}\ln\left(\frac{s_1-2t}{s_2}\right)\right), & \frac{s_1-s_2}{2} < t \leqslant \frac{s_1}{2}; \\ \frac{1}{2} + \frac{1}{2}\left(\frac{2t-s_1}{s_2}\right)^{\frac{1}{\beta}}\left(1 - \frac{1}{\beta}\ln\left(\frac{2t-s_1}{s_2}\right)\right), & \frac{s_1}{2} < t \leqslant \frac{s_1+s_2}{2}; \\ 1, & t > \frac{s_1+s_2}{2}. \end{cases} \end{aligned} \tag{G.8}$$

Recalling (3.3), by (G.7) and (G.8),

$$\begin{aligned} D_-(t) &= \mathbb{P}_{X_1,X_2|A=1}\left(\eta_1(X_1,X_2) > \frac{1}{2}+t\right) - \mathbb{P}_{X_1,X_2|A=0}\left(\eta_0(X_1,X_2) > \frac{1}{2}-t\right) \\ &= \begin{cases} 1, & t \leqslant \frac{s_1-s_2}{2}; \\ \left(\frac{s_1-2t}{s_2}\right)^{\frac{1}{\beta}}\left(1 - \frac{1}{\beta}\ln\left(\frac{s_1-2t}{s_2}\right)\right), & \frac{s_1-s_2}{2} < t \leqslant \frac{s_1}{2}; \\ -\left(\frac{2t-s_1}{s_2}\right)^{\frac{1}{\beta}}\left(1 - \frac{1}{\beta}\ln\left(\frac{2t-s_1}{s_2}\right)\right), & \frac{s_1}{2} < t \leqslant \frac{s_1+s_2}{2}; \\ -1, & t > \frac{s_1+s_2}{2}. \end{cases} \end{aligned}$$

Since $s_1 < s_2$, we have $D_-(0) = \left(\frac{s_1}{s_2}\right)^{\frac{1}{\beta}}\left(1 - \frac{1}{\beta}\ln\left(\frac{s_1}{s_2}\right)\right) > 0$. Thus, based on (A.6), the δ-fair Bayes-optimal classifier is given, for all x, a, by

$$f_\delta^\star(x,a) = I\left(\eta_a(x) > \frac{1}{2} + (2a-1)t_\delta^\star\right) + \tau_{\delta,a}^\star I\left(\eta_a(x) = \frac{1}{2} + (2a-1)t_\delta^\star\right),$$

with $t_\delta^\star = I(D_-(0) > \delta) \cdot \sup_t\{D_-(t) > \delta\}$ and $(\tau_{\delta,1}^\star, \tau_{\delta,0}^\star) \in [0,1]^2$ being arbitrary. Specifically, $t_\delta^\star$ satisfies $0 \leqslant t_\delta^\star < s_1/2$ and

$$\left(\frac{s_1 - 2t_\delta^\star}{s_2}\right)^{\frac{1}{\beta}} \left(1 - \frac{1}{\beta}\ln\left(\frac{s_1 - 2t}{s_2}\right)\right) = \delta \wedge \left[\left(\frac{s_1}{s_2}\right)^{\frac{1}{\beta}} \left(1 - \frac{1}{\beta}\ln\left(\frac{s_1}{s_2}\right)\right)\right].$$

For the misclassification rate of $f_\delta^\star$, we have

$$\begin{aligned}
\mathbb{P}(Y \neq \widehat{Y}_f) &= \sum_{a\in\{0,1\}} p_a \int \left[(1 - 2\eta_a(x_1,x_2))\, f_\delta^\star(x_1,x_2,a) + \eta_a(x_1,x_2)\right] d\mathbb{P}_{X_1,X_2|A=a}(x) \\
&= \frac{1}{8}\sum_{a\in\{0,1\}} \int_{-1}^1\int_{-1}^1 \left[(1 - 2\eta_a(x_1,x_2))\, f_\delta^\star(x_1,x_2,a) + \eta_a(x_1,x_2)\right] dx_1dx_2 \\
&= \frac{1}{8}\int_{-1}^1\int_{-1}^1 \left[f_\delta^\star(x_1,x_2,1) + f_\delta^\star(x_1,x_2,1)\right] dx_1dx_2 + \frac{1}{8}\int_0^1\int_0^1 \left[\eta_1(x_1,x_2) + \eta_0(x_1,x_2)\right] dx_1dx_2 \\
&\quad - \frac{1}{4}\int_{\{\eta_1(x_1,x_2) > \frac{1}{2} + t_\delta^\star\}} \eta_1(x_1,x_2)dx_1dx_2 - \frac{1}{4}\int_{\{\eta_0(x_1,x_2) > \frac{1}{2} - t_\delta^\star\}} \eta_0(x_1,x_2)dx_1dx_2 \\
&=: \text{(I)} + \text{(II)} + \text{(III)} + \text{(IV)}.
\end{aligned}$$

We now calculate the above four terms in turn.

For (I), denote $q_\delta^\star = ((s_1 - 2t_\delta^\star)/s_2)^{(1/\beta)}$. Since $0 \leqslant t_\delta^\star \leqslant s_1/2$, by (G.7) and (G.8),

$$\begin{aligned}
\text{(I)} &= \frac{1}{2}\left[\mathbb{P}_{X_1,X_2|A=1}\left(\eta_1(X_1,X_2) > \frac{1}{2} + t_\delta^\star\right) + \mathbb{P}_{X_1,X_2|A=0}\left(\eta_0(X_1,X_2) > \frac{1}{2} - t_\delta^\star\right)\right] \\
&= \frac{1}{4} + \frac{1}{4}q_\delta^\star\left(1 - \ln\left(q_\delta^\star\right)\right) + \frac{1}{4} - \frac{1}{4}q_\delta^\star\left(1 - \ln\left(q_\delta^\star\right)\right) = \frac{1}{2}.
\end{aligned}$$

Further,

$$\begin{aligned}
\text{(II)} &= \frac{1}{8}\int_{-1}^1\int_{-1}^1 \left[1 + s_2 \cdot \operatorname{sign}(x_1)\left(|x_1|(1 - |x|_2)\right)^\beta\right] dx_1dx_2 \\
&= \frac{1}{2} + s_2\left(\int_{-1}^0 (-1)^{\beta+1}x_1^\beta dx_1 + \int_0^1 x_1^\beta dx_1\right)\cdot\int_{-1}^1 (1 - |x_2|)^\beta dx_2 \\
&= \frac{1}{2} + s_2\cdot\left(-\int_0^1 x_1^\beta dx_1 + \int_0^1 x_1^\beta dx_1\right)\cdot\int_{-1}^1 (1 - |x_2|)^\beta dx_2 = \frac{1}{2}.
\end{aligned}$$

For (III), by (G.3) and (G.4), we have that

$$\begin{aligned}
\text{(III)} &= -\frac{1}{4}\int_{\{\eta_1(x_1,x_2) > \frac{1}{2} + t_\delta^\star\}} \eta_1(x_1,x_2)dx_1dx_2 \\
&= -\frac{1}{4}\left(\int_{\{\eta_1(x_1,x_2) > \frac{1}{2} + t_\delta^\star\}\cap\mathbb{B}_+^2} \eta_1(x_1,x_2)dx_1dx_2 + \int_{\{\eta_1(x_1,x_2) > \frac{1}{2} + t_\delta^\star\}\cap\mathbb{B}_-^2} \eta_1(x_1,x_2)dx_1dx_2\right) \\
&= -\frac{1}{4}\int_0^1\int_{-1}^1 \eta_1(x_1,x_2)dx_1dx_2 - \frac{1}{4}\int_{-\left(\frac{s_1 - 2t_\delta^\star}{s_2}\right)^{\frac{1}{\beta}}}^0 \left(\int_{-1}^1 \eta_1(x_1,x_2)dx_2\right) dx_1 \\
&\quad - \frac{1}{4}\int_{-1}^{-q_\delta^\star}\left(\int_{1 + \frac{q_\delta^\star}{x_1} < |x_2| \leqslant 1} \eta_1(x_1,x_2)dx_2\right) dx_1 \\
&= -\frac{1}{4}\int_0^1\int_{-1}^1 \frac{1 + s_1 + s_2\left(|x_1|(1 - |x_2|)\right)^\beta}{2} dx_1dx_2 \\
&\quad - \frac{1}{4}\int_{-q_\delta^\star}^0\left(\int_{-1}^1 \frac{1 + s_1 - s_2\left(|x_1|(1 - |x_2|)\right)^\beta}{2} dx_2\right) dx_1
\end{aligned}$$

$$-\frac{1}{4}\int_{-1}^{-q_\delta^\star}\left(\int_{1+\frac{q_\delta^\star}{x_1}<|x_2|\leqslant 1}\frac{1+s_1-s_2\left(|x_1|(1-|x_2|)\right)^\beta}{2}dx_2\right)dx_1.$$

This further equals

$$\begin{aligned}
&-\frac{1+s_1}{4}-\frac{s_2}{8}\int_0^1 x_1^\beta dx_1\cdot\int_{-1}^1(1-|x_2|)^\beta dx_1dx_2-\frac{1+s_1}{4}q_\delta^\star\\
&+\frac{s_2}{8}\int_{-q_\delta^\star}^0(-x_1)^\beta dx_1\cdot\int_{-1}^1(1-|x_2|)^\beta dx_2dx_1-\frac{1+s_1}{4}q_\delta^\star\int_{-1}^{-q_\delta^\star}-\frac{1}{x_1}dx_1\\
&+\frac{s_2}{8}\int_{-1}^{-q_\delta^\star}(-x_1)^\beta\left(\int_{1+\frac{q_\delta^\star}{x_1}<|x_2|\leqslant 1}(1-|x_2|)^\beta dx_2\right)dx_1\\
&=-\frac{1+s_1}{4}-\frac{s_2}{4(\beta+1)^2}-\frac{1+s_1}{4}q_\delta^\star+\frac{s_2}{4(\beta+1)^2}\left(q_\delta^\star\right)^{\beta+1}+\frac{1+s_1}{4}q_\delta^\star\ln\left(q_\delta^\star\right)\\
&+\frac{s_2}{4(\beta+1)}\int_{-1}^{-q_\delta^\star}-\frac{\left(q_\delta^\star\right)^{\beta+1}}{x_1}dx_1\\
&=-\frac{1+s_1}{4}\left(1+q_\delta^\star-q_\delta^\star\ln\left(q_\delta^\star\right)\right)-\frac{s_2}{4(\beta+1)}\left(\frac{1-\left(q_\delta^\star\right)^{\beta+1}}{\beta+1}+\left(q_\delta^\star\right)^{\beta+1}\ln\left(q_\delta^\star\right)\right).
\end{aligned}$$

For (IV), by (G.5), (G.6), $s_1/2\leqslant s_1-t^\star\leqslant s_1$ and as $\eta_0(x_1,x_2)=\eta_1(x_1,x_2)-s_1$, we have that

$$\begin{aligned}
\text{(IV)}&=-\frac{1}{4}\int_{\{\eta_0(x_1,x_2)>\frac{1}{2}-t_\delta^\star\}}\eta_0(x_1,x_2)dx_1dx_2=-\frac{1}{4}\int_{\{\eta_1(x_1,x_2)>\frac{1}{2}+s_1-t_\delta^\star\}}\left(\eta_0(x_1,x_2)\right)dx_1dx_2\\
&=-\frac{1}{4}\int_{\{\eta_1(x_1,x_2)>\frac{1}{2}+s_1-t_\delta^\star\}\cap\mathbb{B}_+^2}\eta_0(x_1,x_2)dx_1dx_2-\frac{1}{4}\int_{\{\eta_1(x_1,x_2)>\frac{1}{2}+s_1-t_\delta^\star\}\cap\mathbb{B}_-^2}\eta_0(x_1,x_2)dx_1dx_2\\
&=-\frac{1}{4}\int_{\left(\frac{s_1-2t_\delta^\star}{s_2}\right)^{\frac{1}{\beta}}}^1\left(\int_{\frac{1}{x_1}\left(\frac{s_1-2t_\delta^\star}{s_2}\right)^{\frac{1}{\beta}}-1}^{1-\frac{1}{x_1}q_\delta^\star}\eta_0(x_1,x_2)dx_2\right)dx_1\\
&=-\frac{1}{4}\int_{q_\delta^\star}^1\left(\int_{\frac{1}{x_1}q_\delta^\star-1}^{1-\frac{1}{x_1}q_\delta^\star}\frac{1-s_1+s_2\left(|x_1|(1-|x_2|)\right)^\beta}{2}dx_2\right)dx_1\\
&=-\frac{1-s_1}{4}\int_{q_\delta^\star}^1\left(1-\frac{1}{x_1}q_\delta^\star\right)dx_1-\frac{s_2}{8}\int_{q_\delta^\star}^1 x_1^\beta\left(\int_{\frac{1}{x_1}q_\delta^\star-1}^{1-\frac{1}{x_1}q_\delta^\star}(1-|x_2|)^\beta dx_2\right)dx_1\\
&=-\frac{1-s_1}{4}\left(1-q_\delta^\star\right)-\frac{1-s_1}{4}q_\delta^\star\ln\left(q_\delta^\star\right)-\frac{s_2}{4(\beta+1)}\int_{q_\delta^\star}^1\left(x_1^\beta-\frac{1}{x_1}\left(q_\delta^\star\right)^{\beta+1}\right)dx_1\\
&=-\frac{1-s_1}{4}\left(1-q_\delta^\star+q_\delta^\star\ln\left(q_\delta^\star\right)\right)-\frac{s_2}{4(\beta+1)}\left(\frac{1-\left(q_\delta^\star\right)^{\beta+1}}{\beta+1}+\left(q_\delta^\star\right)^{\beta+1}\ln\left(q_\delta^\star\right)\right).
\end{aligned}$$

Summing up, we have

$$\begin{aligned}
R(f_\delta^\star)&=\text{(I)}+\text{(II)}+\text{(III)}+\text{(IV)}\\
&=\frac{1}{2}+\frac{1}{2}-\frac{1+s_1}{4}\left(1+q_\delta^\star-q_\delta^\star\ln\left(q_\delta^\star\right)\right)-\frac{1-s_1}{4}\left(1-q_\delta^\star+q_\delta^\star\ln\left(q_\delta^\star\right)\right)\\
&\quad-\frac{s_2}{2(\beta+1)}\left(\frac{1-\left(q_\delta^\star\right)^{\beta+1}}{\beta+1}+\left(q_\delta^\star\right)^{\beta+1}\ln\left(q_\delta^\star\right)\right)\\
&=\frac{1}{2}-\frac{s_1q_\delta^\star}{2}\left(1-\ln\left(q_\delta^\star\right)\right)-\frac{s_2}{2(\beta+1)}\left(\frac{1-\left(q_\delta^\star\right)^{\beta+1}}{\beta+1}+\left(q_\delta^\star\right)^{\beta+1}\ln\left(q_\delta^\star\right)\right).
\end{aligned}$$