FAIRNESS GUARANTEE IN ANALYSIS OF INCOMPLETE DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Missing data are prevalent and present daunting challenges in real data analysis. While there is a growing body of literature on fairness in analysis of fully observed data, there has been little work on investigating fairness in analysis of incomplete data when the goal is to develop a fair algorithm in the complete data domain where there are no missing values. In practice, a popular analytical approach for dealing with missing data is to use only the set of complete cases, i.e., observations with all features fully observed, as a representation of complete data in learning. However, depending on the missing data mechanism, the complete case domain and the complete data domain may have different data distributions and a fair algorithm in the complete case domain may show disproportionate bias towards some marginalized groups in the complete data domain. To fill this significant gap, we studying the problem of estimating fairness in the complete data domain for a model trained using observed data and evaluated in the complete case domain. We provide upper and lower bounds on the fairness estimation error and conduct numerical experiments to assess our theoretical results. Our work provides the first known results on fairness guarantee in analysis of incomplete data.

1 INTRODUCTION

Mounting evidence (Gianfrancesco et al., 2018; Buolamwini & Gebru, 2018; Kiritchenko & Mohammad, 2018; Trewin et al., 2019; Wang et al., 2019) has suggested that powerful machine learning algorithms can be unfair and lead to disproportionately unfavorable treatment towards marginalized groups. In recent years, there has been a growing body of research on addressing unfairness and bias of machine learning algorithms (Chouldechova & Roth, 2020).

Meanwhile, missing data are ubiquitous and present daunting challenges in real data analysis. Particularly, missing data, if not adequately handled, would lead to biased estimation and improper statistical inference (Little & Rubin, 2019). As such, analysis of incomplete data has been an active research area. More recently, there is also a growing recognition that missing data may have deleterious impact on algorithmic fairness. For example, in medicine, bias caused by missing values in electronic health records is identified as a significant factor contributing to unfairness of machine learning (ML) algorithms used in medicine that may exacerbate health care disparities (Rajkomar et al., 2018; Gianfrancesco et al., 2018). However, there has been little to no reported research on fairness in analysis of incomplete data.

In the presence of missing values, one popular approach for training ML models is to use only the set of complete cases, i.e., observations with all features observed, as a representation to the complete data in learning, discarding the other incomplete cases/observations. However, in practice, the ultimate goal, oftentimes, is to apply the trained models to the complete data domain where there are no missing values. Depending on the missing data mechanism, the complete case domain and the complete data domain may have different data distributions. A fair algorithm in the complete case domain may show disproportionate bias towards marginalized groups in the complete data domain. Our work addresses this significant gap when using complete cases as representations in learning.

1.1 RELATED WORKS

The growing body of literature on algorithmic fairness has been primarily focused on two types of fairness defintions, group fairness and individual fairness (Chouldechova & Roth, 2020). Group fairness emphasizes that members from different groups (e.g. gender, race etc.) should be treated similarly, while individual fairness pays more attention to treatment similarity between any of two similar individuals. In this work we investigate group fairness in analysis of incomplete data. Fairness notions are acknowledged to be case-specific, in a sense that it is not possible to achieve fairness under various definitions (Friedler et al., 2016). In binary classification problems, *demographic (or statistical) parity* (Calders & Verwer, 2010; Dwork et al., 2012) is a fairness notions that has been mostly studied. It states that the predicted outcome should be independent with sensitive attributes. However demographic parity can cause severe harm to the prediction performance when true response is dependent with sensitive attributes. As an alternative, *disparate mistreatment* (Zafar et al., 2017) states that misclassification level (regarding e.g. overall accuracy, false negative rate, false discovery rate etc.) should be similar between two sensitive groups. Similarly, Hardt et al. (2016) proposes *equalized odds*, which requires both false positive rate (FPR) and false negative rate (FNR) to be the same between two groups. In regression setting, fairness notion is usually associated with parity of loss between two groups (Agarwal et al., 2019; Oneto et al., 2019; Donini et al., 2018). In this paper we adopt accuracy parity gap as fairness notions in learning tasks including classification and regression. In fact, our results can be generalized into other fairness notions such as equal opportunity and prediction error parity with respect to mean square error.

Recently, Schumann et al. (2019) investigated fairness cross different domains and provided an upper bound of fairness on one desired domain (target domain) given that on another domain (source domain). Martínez-Plumed et al. (2019) investigated the impact of missing data on algorithmic fairness through numerical experiments. To the best of our knowledge, this is the only work on algorithmic fairness in the presence of missing data.

1.2 OUR CONTRIBUTIONS

There are a number of fundamental differences between our work and Schumann et al. (2019). Their work does not deal with missing data and associated challenges and does not consider the technique of re-weighting for domain adaptation. In addition, they provided only upper bounds on transferring fairness. The main contributions of our work are as follows: (1) the first theoretical analysis of fairness guarantee in analysis of incomplete data which provides both theoretical upper and lower bounds; (2) the first study on the effect of re-weighting in analysis of fairness across two data domains induced by missing data and the impact of the missing data mechanism; (3) evaluation of the transferred fairness (from the complete case domain to the complete data domain) through extensive numerical experiments using both synthetic and real data. Our work represents significant advances over the existing work by Martínez-Plumed et al. (2019) in studying fairness guarantee for analysis of incomplete data.

2 PROBLEM FORMULATION

2.1 PRELIMINARIES ON MISSING DATA

Suppose that the data set of interest contains n observations. If there were no missing values, each observation/case $\boldsymbol{z}_i := \{\boldsymbol{x}_i, y_i\} \in \mathcal{X} \times \mathcal{Y}$ $(i = 1, \ldots, n)$ consists of predictors $\boldsymbol{x}_i \in \mathcal{X}$ and label (response variable) $y_i \in \mathcal{Y}$. We denote the complete data matrix by $\mathbf{Z}$, whose i^{th} row is $\boldsymbol{z}_i$. Since some entries of $\mathbf{Z}$ are missing, define the indicator for observing z_{ij} or not as $r_{ij} = \mathbf{1}\{z_{ij} \text{ is observed}\}$. Denote the corresponding indicator matrix by $\mathbf{R}$. Let $\boldsymbol{z}_{(1)i}$ denote the components of $\boldsymbol{z}_i$ that are observed for observation i, and $\boldsymbol{z}_{(0)i}$ denote the components of $\boldsymbol{z}_i$ that are missing for observation i. For example, consider the case when there are two predictors and one response. If only z_{i1} is observed, then $\boldsymbol{z}_{(1)i} = z_{i1}$, $\boldsymbol{z}_{(0)i} = (z_{i2}, z_{i3})$. We then define the observed data $\mathbf{Z}_{obs}$ as the collection of the observed components from all n observations, $\left\{\boldsymbol{z}_{(1)i}, i = 1, \ldots, n\right\}$ and the missing data $\mathbf{Z}_{mis}$ as the collection of all the missing components, $\left\{\boldsymbol{z}_{(0)i}, i = 1, \ldots, n\right\}$.

There are three primary missing data mechanisms, namely, missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) (Little & Rubin, 2019). Data are said

to be *missing completely at random* (MCAR) if the distribution of $\mathbf{R}$ is independent of $\mathbf{Z}$. Data are said to be *missing at random* (MAR) if the distribution of $\mathbf{R}$ depends on $\mathbf{Z}$ only through its observed components, i.e., $\mathbf{R} \perp \mathbf{Z}_{mis}|\mathbf{Z}_{obs}$. Data are said to be *missing not at random* (MNAR) if the distribution of $\mathbf{R}$ depends on the missing components of $\mathbf{Z}$. We will investigate fairness guarantee under all three mechanisms.

In the presence of missing values, observation i is said to be a *complete case* if it is fully observed (i.e. $\boldsymbol{z}_{(1)i} = \boldsymbol{z}_i$). Let $R_i := \mathbf{1}\{\boldsymbol{z}_i \text{ is fully observed}\}$ denote the indicator of complete cases. We can then define two different data domains (distributions). The complete data domain, denoted by $\mathcal{D}_T$, is the distribution of $\boldsymbol{z}_i$ in the joint space $\mathcal{X} \times \mathcal{Y}$. The complete case domain, denoted by $\mathcal{D}_S$, is the distribution of observations that have all variables fully observed: $\boldsymbol{z}_i|R_i = 1$.

Under the MCAR mechanism, the distribution of the complete case in $\mathcal{D}_S$ is the same as the distribution of the complete data in $\mathcal{D}_T$. However under MAR or MNAR, the data distributions in $\mathcal{D}_S$ and $\mathcal{D}_T$ can be different. For example, if missingness depends on gender and other features associated with gender, then females may have a substantially higher proportion of missing values and the feature distribution in females in $\mathcal{D}_S$ may be very different from that in $\mathcal{D}_T$. As a result, an algorithm that is trained and shown to be fair in $\mathcal{D}_S$ may not be fair when evaluated in $\mathcal{D}_T$, noting that in practice we typically are more interested in the fairness guarantee in $\mathcal{D}_T$.

2.2 Fairness notions and estimand

We consider learning tasks that use features $\boldsymbol{x} \in \mathcal{X}$ to predict response $y \in \mathcal{Y}$. Each observation also has a binary sensitive attribute $A \in \{0, 1\}$. We are interested in assessing fairness of a prediction model $g : \mathcal{X} \to \mathcal{Y}$, which is learned from the observed data $\mathbf{Z}_{obs}$, in complete data domain. Let $\mathcal{E}_a(g) := \mathbb{E}_{T_a}|g(\boldsymbol{x}) - y(\boldsymbol{x})|$ denote the prediction error, where T_a represents that the expectation is taken with respect to domain $\mathcal{D}_{T_a}$, the (conditional) distribution of sensitive group $A = a$. We define a general fairness measurement to develop our theoretical results, which is applicable in various learning tasks including regression and classification.

Definition 1 (Accuracy Parity Gap)**.** *Given an arbitrary complete data domain $\mathcal{D}_T$ and predictor g, the accuracy parity gap of g is $\Delta_T(g) = |\mathcal{E}_0(g) - \mathcal{E}_1(g)|$, where subscript T indicates that the fairness estimand is defined in the complete data domain $\mathcal{D}_T$.*

This definition has close connections with various fairness notions proposed in the literature. In binary classification task where the response is binary: $y \in \{0, 1\}$, the notion *accuracy parity* has been used in Zafar et al. (2017), Friedler et al. (2016) and Zhao et al. (2019b), which requires that the prediction accuracy between two sensitive groups to be equal: $\mathcal{P}(g(\boldsymbol{x}) = y|A = 0) = \mathcal{P}(g(\boldsymbol{x}) = y|A = 1)$. Accuracy parity gap in this case is the absolute value of difference between above two quantities $|\mathcal{P}(g(\boldsymbol{x}) = y|A = 0) - \mathcal{P}(g(\boldsymbol{x}) = y|A = 1)|$. For regression tasks where the response y takes continuous value, fairness constraints on loss difference between two groups are adopted in Donini et al. (2018), Oneto et al. (2019) and Agarwal et al. (2019). Accuracy parity gap under such setting can be regarded as the difference of the mean absolute error (MAE) loss.

2.3 Fairness estimator

Since missing values can cause bias in an incomplete data set, fairness in the complete data domain is typically of primary interest. However, due to missing values, fairness of the prediction model g is assessed using only the complete cases. Fairness estimation, using the finite sum over the complete cases to approximate $\Delta_T(g)$ can be biased because of the domain difference. To mitigate such estimation bias, one useful approach is to assign weight $\omega(\boldsymbol{z}_i)$ to observation $\boldsymbol{z}_i$ and calculate the weighted sum over the complete cases to estimate $\Delta_T(g)$. Specifically, we define the weighted empirical risk (prediction error) using the complete cases $\widehat{\mathcal{E}}_a(g, \omega) := \frac{1}{\sum_{i=1}^n I(A_i=a)R_i} \sum_{i=1}^n I(A_i = a)R_i\omega(\boldsymbol{z}_i)|g(\boldsymbol{x}_i) - y_i|$, where $a \in \{0, 1\}$. Here we assume there is at least one complete case observed for each sensitive group ($\sum_{i=1}^n I(A_i = a)R_i \geq 1$). Then the proposed fairness estimator is defined as follows.

Definition 2 (Fairness estimator from complete cases)**.** *Suppose the weights assigned to complete cases are given by ω. Then the fairness estimator for predictor g in the complete data domain is*

$\widehat{\Delta}_S(g,\omega) = \left|\widehat{\mathcal{E}}_0(g,\omega) - \widehat{\mathcal{E}}_1(g,\omega)\right|$, *where subscript* S *indicates that the estimator is obtained from the data from the complete case domain* $\mathcal{D}_S$.

In practice, a popular choice of ω is the normalized inverse of the propensity score (a.k.a, the probability of being a complete case). Let $\pi(\boldsymbol{z}_i) := \mathcal{P}_T(R_i = 1|\boldsymbol{z}_i)$ denote the true propensity score (PS) model. Corollary 1 presents the results when $\omega(\boldsymbol{z}_i) = [\pi(\boldsymbol{z}_i)\mathbb{E}_S\{1/\pi(\boldsymbol{z})\}]^{-1}$, where S represents that the expectation is taken with respect to $\mathcal{D}_S$. In practice, we typically do not know the true propensity scores or the true distribution of complete cases, so we need to estimate the propensity scores and use the empirical distribution of complete cases. Various statistical and machine learning models can be used to estimate the propensity scores, such as logistic regression, random forest, support vector machines and boosting.

3 MAIN RESULTS

In this section we provide theoretical analysis for the proposed fairness estimator $\widehat{\Delta}_S(g,\omega)$ in the complete data domain $\mathcal{D}_T$. Some of the techniques in the proofs are inspired from the analysis of learning guarantee in domain adaptation in Cortes et al. (2010). Throughout, we assume the weights are normalized: $\mathbb{E}_S\omega(\boldsymbol{z}) = 1$ and are bounded away from 0 and infinity. In the regression setting, it is commonly assumed in the domain adaptation literature that y takes value inside interval $[b_1, b_2]$ for some real number b_1, b_2.

Our theoretical results can be generalized to a wide range of other fairness notions. For binary classification, all the five measurements of disparate mistreatment (accuracy, false positive rate, false negative rate, false mission rate and false discover rate) mentioned in Section 2 of Zafar et al. (2017) can be adopted to obtain our main results. Notably, *false negative rate parity* is also known as *Equal Opportunity* proposed in Hardt et al. (2016). For regression, it is straightforward to generalize our results to the fairness notation defined as the difference of L^p loss with $1 \leq p < \infty$.

3.1 AN UPPER BOUND

Let $B = \sup_{\boldsymbol{z}}\omega(\boldsymbol{z}) < +\infty$ denote the upper bound of the weights, $D_a^\omega = \mathbb{E}_{S_a}\omega(\boldsymbol{z})^2$ denote the second moment of weights, $\omega\mathcal{D}_{S_a}$ denote the distribution whose probability density function at $\boldsymbol{z}$ equals to $\omega(\boldsymbol{z})f_{S_a}(\boldsymbol{z})$ with f_{S_a} be the probability density function in $\mathcal{D}_{S_a}$. We further let n_a denote the number of complete cases observed in group $a \in \{0,1\}$. Theorem 1 below provides an upper bound of the fairness estimation error.

Theorem 1. *Assume that* g *is from a hypothesis class* $\mathcal{H}$ *with VC dimension* d *(pseudo dimension if in regression setting) and* $y \in [0,1]$. *Assume* $D_a^\omega \leq n_a/8$ *for both groups. Then, for any* $\delta > 0$, *with probability at least* $1-\delta$, *the following inequality holds:*

$$\left|\Delta_T(g) - \widehat{\Delta}_S(g,\omega)\right| \leq \sum_{a\in\{0,1\}} d_{TV}\left(\mathcal{D}_{T_a}||\omega\mathcal{D}_{S_a}\right) + \sqrt{\frac{128D_a^\omega C_d(n_a, D_a^\omega, \delta)}{n_a}} + \frac{16C_d(n_a, D_a^\omega, \delta)B}{3n_a} \tag{1}$$

where $d_{TV}\left(\mathcal{D}_{T_a}||\omega\mathcal{D}_{S_a}\right)$ *denote the total variation distance between* $\mathcal{D}_{T_a}$ *and* $\omega\mathcal{D}_{S_a}$ *and that*

$$C_d(n_a, D_a^\omega, \delta) = \begin{cases} \log\frac{(d+1)(8e)^{d+1}}{\delta} + \frac{d}{2}\log\frac{n_a}{2D_a^\omega} & \textit{in classification} \\ \log\frac{4}{\delta}\left(\frac{8e}{d}\right)^d + \frac{3d}{2}\log\frac{n_a}{\sqrt[3]{2D_a^\omega}} & \textit{in regression} \end{cases}$$

Remark 1. *If* $y \in [b_1, b_2]$, *the upper bound in Theorem 1 would be multiplied by* $b_2 - b_1$, *with requirement that* $D_a^\omega(b_2 - b_1) \leq n_a/8$.

Remark 2. D_a^ω *is always upper bounded by* B^2. *In particular, for MCAR mechanism,* $D_a^\omega = 1$.

Since the fairness estimand is defined based on prediction error, there is some similarity between the upper bound in Theorem 1 and the upper bounds for learning error in the domain adaptation literature Ben-David et al. (2010); Redko et al. (2017). The upper bound is obtained by the triangle inequality and the detailed analysis of generalization error. In the first term of the upper bound, $\omega\mathcal{D}_{S_a}$ is an approximation to the complete data domain in sensitive group a, and $d_{\text{TV}}(\mathcal{D}_{T_a}||\omega\mathcal{D}_{S_a})$ can be

viewed as the approximation error. It follows that a less accurate approximation of the complete data domain would lead to a looser upper bound on fairness estimation error. In the second and third terms, $C_d(n_a, D_a^\omega, \delta)$ is proportional to $\log n_a$. Since the third term is dominated by the second term, their sum is of order $\mathcal{O}(\log n_{\text{min}}/n_{\text{min}})^{\frac{1}{2}}$ with n_{min} representing the sample size in the minority group. It also follows that for a fixed total sample size n, the upper bound increases with sample imbalance between two groups defined by A. In addition, the missing data mechanism impacts the second moment of estimated weights $D_a^\omega \in [1, B^2]$. If a missing data mechanism leads to a larger second moment of the weights, the upper bound for the fairness estimation error would be looser. Furthermore, when the weights ω are defined using the true propensity scores, we call the resulting $\omega_0(\boldsymbol{z}_i) := (\pi(\boldsymbol{z}_i)\mathbb{E}_S(1/\pi(\boldsymbol{z})))^{-1}$ as the true weights and the first term in the upper bound in Theorem 1 would vanish in which case we have the following result.

Corollary 1. *If $\omega(\boldsymbol{z}_i) = \omega_0(\boldsymbol{z}_i)$, then $\widehat{\Delta}_S(g, \omega)$ is consistent for estimating $\Delta_T(g)$.*

There are several implications from Corollary 1. Under the MCAR mechanism, setting $\omega(\boldsymbol{z}_i) = 1$ would yield a consistent (unweighted) estimator. Under MAR and MNAR, since we typically do not know the true propensity scores $\pi(\boldsymbol{z}_i)$, we replace $\pi(\boldsymbol{z}_i)$ with its estimate $\hat{\pi}(\boldsymbol{z}_i)$ using a working model for $\pi(\boldsymbol{z}_i)$ which is subject to mis-specification. When the number of complete cases in both groups are sufficiently large, the first term in the upper bound would often be the dominant term. Then, $\hat{\pi}(\boldsymbol{z}_i)$ from a correctly-specified propensity score model would lead to a smaller $d_{\text{TV}}(\mathcal{D}_{T_a}||\omega\mathcal{D}_{S_a})$ and hence a tighter bound. On the other hand, an incorrectly-specified propensity score model could lead to a larger $d_{\text{TV}}(\mathcal{D}_{T_a}||\omega\mathcal{D}_{S_a})$ compared with the unweighted estimator.

3.2 A LOWER BOUND

Define $\sigma_a^2(g, \omega) := \text{Var}_{S_a}(\omega|g - y|)$ and it can be shown that $\sigma_a^2(g, \omega) \leq B^2/4$. We present the result on a lower bound for the fairness estimation error in the following theorem.

Theorem 2. *If the weight $\omega(\boldsymbol{z}_i)$ is set to be $\omega_0(\boldsymbol{z}_i)$ and that $B^2/\sigma_a^2(g, \omega) \leq \min\{n_0, n_1\}$. Then with probability at least $\frac{7}{1440}$, the following hold:*

$$12\sqrt{\frac{\sigma_0^2(g,\omega_0)}{n_0} + \frac{\sigma_1^2(g,\omega_0)}{n_1}} \geq \left|\Big(\mathcal{E}_0(g) - \mathcal{E}_1(g)\Big) - \Big(\widehat{\mathcal{E}}_0(g,\omega) - \widehat{\mathcal{E}}_1(g,\omega)\Big)\right| \geq \frac{1}{24}\sqrt{\frac{\sigma_0^2(g,\omega_0)}{n_0} + \frac{\sigma_1^2(g,\omega_0)}{n_1}} \tag{2}$$

If above holds and that $\widehat{\Delta}_S(g, \omega) \geq \frac{13}{2}\sqrt{\frac{\sigma_0^2(g,\omega_0)}{n_0} + \frac{\sigma_1^2(g,\omega_0)}{n_1}}$, we have:

$$\left|\Delta_T(g) - \widehat{\Delta}_S(g,\omega)\right| \geq \frac{1}{24}\sqrt{\frac{\sigma_0^2(g,\omega_0)}{n_0} + \frac{\sigma_1^2(g,\omega_0)}{n_1}}$$

If $\widehat{\Delta}_S(g, \omega) \leq \frac{1}{72}\sqrt{\frac{\sigma_0^2(g,\omega_0)}{n_0} + \frac{\sigma_1^2(g,\omega_0)}{n_1}}$, we have:

$$\left|\Delta_T(g) - \widehat{\Delta}_S(g,\omega)\right| \geq \frac{1}{72}\sqrt{\frac{\sigma_0^2(g,\omega_0)}{n_0} + \frac{\sigma_1^2(g,\omega_0)}{n_1}}$$

The proof involves detailed analysis of the truncation probability for the fairness difference. We also have the following corollary describing the lower bound of fairness when g is a consistent estimator for y and $|\omega(\boldsymbol{z}_i) - \omega_0(\boldsymbol{z}_i)| = \mathcal{O}_p((n_0 + n_1)^{-1/2})$.

Corollary 2. *If the weights satisfy $|\omega(\boldsymbol{z}_i) - \omega_0(\boldsymbol{z}_i)| = \mathcal{O}_p((n_0 + n_1)^{-1/2})$ and g is consistent for y (i.e. $\lim_n g(\boldsymbol{x}) = y(\boldsymbol{x})$), then for any positive δ, there exists N_0 and N_1 such that whenever $n_0 > N_0$ and $n_1 > N_1$, with probability at least $\frac{7}{10}(\frac{1}{12})^2 - \delta$*

$$\left|\Big(\mathcal{E}_0(g) - \mathcal{E}_1(g)\Big) - \Big(\widehat{\mathcal{E}}_0(g,\omega) - \widehat{\mathcal{E}}_1(g,\omega)\Big)\right| \geq \frac{1}{25}\sqrt{\frac{\sigma_0^2(g,\omega_0)}{n_0} + \frac{\sigma_1^2(g,\omega_0)}{n_1}} \tag{3}$$

Remark 3. *If y is bounded in $[b_1, b_2]$ in regression instead of the unit interval, then Theorem 2 and Corollary 2 still hold.*

Remark 4. *If the weight ω is based on the estimated $\hat{\pi}(\boldsymbol{z}_i)$ from a correctly specified propensity score model, then the optimal convergence rate for $|\omega(\boldsymbol{z}_i) - \omega_0(\boldsymbol{z}_i)|$ is $\mathcal{O}_p((n_0+n_1)^{-1/2})$. Comparing results from (1) and (3), the upper bound is of order $\mathcal{O}((\log n_{min}/n_{min})^{1/2})$ while the lower bound is of order $\mathcal{O}((n_{min})^{-1/2})$.*

Our theoretical results have additional important implications related to the missing data mechanisms. First, as long as sample size is sufficiently large as required, Theorem 1 would always hold for all mechanisms. Under the MCAR mechanism, the true propensity score is a constant and hence can be regarded as known, and the results from Theorem 2 hold for the unweighted estimator. Under the MAR mechanism, the true propensity score is generally unknown. If the correctly specified propensity score model is fitted and the estimated propensity scores converge to the true values at the rate of $\mathcal{O}_p((n_0+n_1)^{-1/2})$, the results in Corollary 2 would hold. Under the MNAR mechanism, the propensity score model depends on missing values, so it cannot be estimated without making additional modeling assumptions. If the propensity score model is mis-specified under MAR or MNAR, the results in Theorem 2 are not applicable.

4 Numerical Experiments

In this section we empirically evaluate the bias of fairness estimation in both synthetic and real data sets. Lower bound in Theorem 2 is justified and several factors that influence the fairness estimation are also investigated. Recall that our goal is to estimate fairness $\Delta_T(g)$ defined in the complete data domain, while the estimator $\hat{\Delta}_S(g,\omega)$ is obtained from the complete case domain.

4.1 Synthetic data

In our simulation studies, we consider a regression task with 10 predictors and a binary sensitive attribute $A \in \{0,1\}$. With sample size n , the predictors are generated from gaussian distribution: $x_{ij} \overset{iid}{\sim} \mathcal{N}(1-2A_i, 0.5^2)$ ($i = 1,\ldots,n$ and $j = 1,\ldots,10$). We generate the response $y_i = (\boldsymbol{x}_i^\top \beta)^2 + \epsilon_i$ where $\beta = (0.1, 0.1, 0.1, 0.1, 0.1, 1, 1, 1, 1, 1)^\top$ and $\epsilon = (\epsilon_1,\ldots,\epsilon_n)^\top \sim \mathcal{N}(0, I_n)$. Missing values are generated from the last 5 features using different propensity score models under MAR. After generating missing values, we use complete cases to fit a random forest model g for predicting y and investigate the bias of estimated fairness $\widehat{\Delta}_S(g,\omega)$ for g.

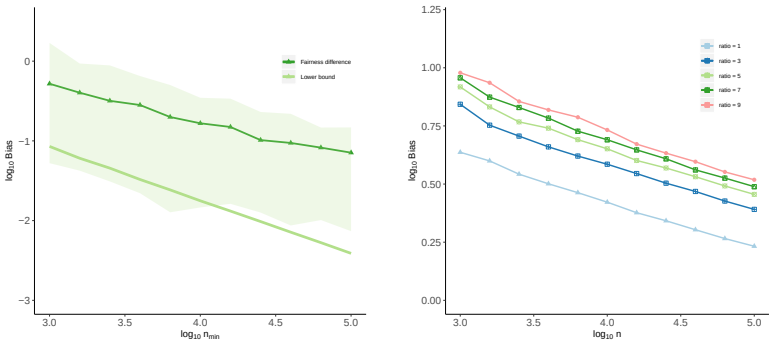
Justification of convergence rate Missing values are generated based on the following propensity score model $\text{logit}(\pi(\boldsymbol{z}_i)) = -3 + \frac{1}{5}\sum_{j=1}^{5}\sin(3x_{ij})$, where $\pi(\boldsymbol{z}_i) = \mathcal{P}(R_i = 1|\boldsymbol{z}_i)$ and $\text{logit}(p) = \log\frac{p}{1-p}$. Of note, since R_i depends on only the fully observed features, the missing data mechanism is MAR. In this experiment, we use the true weights $\omega(\boldsymbol{z}) = \omega_0(\boldsymbol{z})$ to calculate the fairness estimation bias. Figure 1-(a) shows that the lower bound from Theorem 2 is smaller than the mean of fairness estimation bias. Meanwhile, it sometimes lies inside 90% band of fairness estimation bias, implying the lower bound is not disproportionately loose. In addition, the slope of fairness difference is smaller than that of lower bound, indicating the actual convergence rate might be $o(n_{\text{min}}^{-1/2})$ under this experiment setting.

Effect of sample imbalance We fix the total sample size and alter the sample imbalance between two groups. We generate missing values using MAR mechanism $\text{logit}(\pi(\boldsymbol{z}_i)) = -1 + \frac{1}{5}\sum_{j=1}^{5} x_{ij}$. To estimate the fairness, we adopt logistic regression for propensity score estimation, which is correctly-specified. The resulting fairness estimation bias is shown in Figure 1-(b), in which we set sample imbalance ratio, which is the number of complete data in group $A_i = 0$ divided by that in group $A_i = 1$, from 1 to 9. For a fixed n, increasing sample imbalance will lead to larger bias $|\Delta_T(g) - \hat{\Delta}_S(g,\omega)|$. This result supports that sample imbalance could harm the fairness estimation.

Effect of different weights We compare the performance of fairness estimator among 7 estimators with different weights: $\omega(\boldsymbol{z}_i) = 1$ (i.e., unweighted estimator), the true weights $\omega(\boldsymbol{z}_i) = \omega_0(\boldsymbol{z}_i)$ and 5 weights estimated via $\omega(\boldsymbol{z}_i) = \widehat{\omega}_0(\boldsymbol{z}_i) := \sum_{i=1}^{n} R_i / \left(\hat{\pi}(\boldsymbol{z}_i)\sum_{i=1}^{n}\{R_i/\hat{\pi}(\boldsymbol{z}_i)\}\right)$ where $\hat{\pi}(\boldsymbol{z}_i)$ is obtained from logistic regression(both correctly and incorrectly specified), random forest (RF)

estimator, support vector machine (SVM) as well as extreme gradient boosting (XGB) estimators. Missing values are generated from MAR mechanism $\text{logit}(\pi(\boldsymbol{z}_i)) = 3 - \frac{1}{5}\sum_{j=1}^{5} x_{ij}^3$. We alter the complete data's sample size so that both group has the same number of complete cases (in expectation). In particualr we use $\{x_{ij}^3\}$ to fit the first logistic regression model, which is correctly specified. In the second model, we use $\{|x_{ij}|\}$, which leads to an incorrectly specified model. From Figure 1-(c), using true inverse probability weights or correctly specified model (logistic regression) lead to smaller gap between fairness in complete data and complete case domains. At the same time, incorrectly-specified propensity score model and non-parametric propensity score models could result in larger $|\Delta_T(g) - \hat{\Delta}_S(g, \omega)|$. In this case, unweighted estimator has the largest fairness estimation bias.

Effect of sensitive groups' domains We add an additional parameter M in the distribution of designs: x_{ij} i.i.d. drawn from $\mathcal{N}(1 - 2MA_i, 0.5^2)$. Then we generate missing values using MAR mechanism $\text{logit}(\pi(\boldsymbol{z}_i)) = 2 - 4A_i$. We use a correctly-specified logistic regression model to estimate propensity score model. Fairness estimation bias with different M are shown in Figure 1-(d). From the figure, with increasing M, the distance between two sensitive group's domain increases. At the same time the fairness estimation bias also increases. This implies that it can be harder to guarantee fairness in complete data domain when two sensitive groups' distributions are more different.



(a) Justification of bounds

(b) Effect of sample imbalance

(c) Effect of PS estimator

(d) Effect of sensitive group domains

Figure 1: Experiments on synthetic data. In (a), 90% confidence band and mean value of fairness estimation bias are calculated over 200 repeats. Each curve in (b) (c) (d) plots the mean fairness estimation bias over 50 repeats.

4.2 Real datasets

In analysis of each real dataset, we artificially generate missing values under the three missing data mechanisms, train a random forest prediction model g for the outcome of interest using the set of complete cases, and then evaluate the fairness estimation error of g. For the estimated fairness $\widehat{\Delta}_S(g, \omega)$, We compare multiple options of ω, a) $\omega(\boldsymbol{z}_i) = 1$ (i.e., un-

		Specification of ω					
		Unweighted	True	Logistic	RF	SVM	XGB
COMPAS	MCAR	**1.27 ± 0.81**	1.30 ± 1.01	1.32 ± 1.04	1.33 ± 0.92	2.02 ± 1.81	1.41 ± 1.08
($\times 10^{-2}$)	MAR	3.12 ± 2.05	**1.91 ± 1.61**	2.03 ± 1.69	2.16 ± 1.71	8.05 ± 5.36	3.44 ± 2.40
	MNAR	3.92 ± 1.90	**2.78 ± 1.73**	3.82 ± 1.90	3.84 ± 2.11	3.81 ± 2.03	3.51 ± 2.07
ADNI	MCAR	3.00 ± 2.63	2.96 ± 2.47	**2.45 ± 2.09**	2.70 ± 2.27	3.09 ± 2.55	3.06 ± 2.46
($\times 10^{-3}$)	MAR	3.24 ± 2.71	**2.41 ± 1.67**	2.73 ± 2.37	2.71 ± 1.72	3.00 ± 2.63	3.22 ± 2.68
	MNAR	3.13 ± 2.48	**2.13 ± 1.82**	2.61 ± 2.02	2.35 ± 1.88	3.16 ± 2.61	3.07 ± 1.55

Table 1: Bias in fairness estimation $|\Delta_T(g) - \widehat{\Delta}_S(g,\omega)|$ with different options for ω and missing data mechanisms in analysis of the COMPAS and ADNI datasets. Mean ± SD over 50 repeats.

weighted), b) the true inverse probability weights $\omega(\boldsymbol{z}_i) = \omega_0(\boldsymbol{z}_i)$, and c) $\omega(\boldsymbol{z}_i) = \widehat{\omega}_0(\boldsymbol{z}_i) = \sum_{i=1}^n R_i / \left(\hat{\pi}(\boldsymbol{z}_i) \sum_{i=1}^n \{R_i/\hat{\pi}(\boldsymbol{z}_i)\}\right)$ where $\hat{\pi}(\boldsymbol{z}_i)$ is obtained from logistic regression, RF, SVM and XGB. We compute the fairness estimation bias $|\Delta_T(g) - \hat{\Delta}_S(g,\omega)|$, where $\Delta_T(g)$ is approximated using the complete data in the original real dataset before missing values are generated. This procedure is repeated 50 times for each dataset. Of note, in all real data analyses, the logistic regression model is the correctly specified model for $\pi(\boldsymbol{z}_i)$.

COMPAS recidivism dataset COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) (Northpointe, 2010) is a risk assessment instrument developed by Northpointe Inc. A set of questions is completed by defendants to be used in prediction of "Risk of Recidivism." The dataset analyzed in this work contains records of defendants from Broward County from 2013 and 2014. Prior work has demonstrated bias of COMPAS towards certain groups of defendants (e.g. race, gender and age) (Angwin et al., 2016). In our analysis, gender is treated as the sensitive attribute and nine features are used to predict two-year recidivism (defined by arrest within 2 years) (Rudin et al., 2018). We generate missing values for the last feature and the outcome variable under three missing mechanisms: MCAR, $\text{logit}(\pi(\boldsymbol{z}_i)) = 0.5$; MAR, $\text{logit}(\pi(\boldsymbol{z}_i)) = 3 + \sum_{j=1}^8 x_{ij}$; MNAR, $\text{logit}(\pi(\boldsymbol{z}_i)) = 2y + 2\prod_{j=6}^9 x_{ij}$. As shown in Table 1, all options of ω lead to comparable results under MCAR, noting that all of them are valid under MCAR. Under MAR and MNAR, however, the estimator using the true weights is most accurate and followed by logistic regression, noting that logistic regression is the correctly specified model for $\pi(\boldsymbol{z}_i)$ in this data analysis.

ADNI gene expression data We analyze a dataset from Alzheimer's Disease Neuroimaging Initiative which contains gene expression and clinical data for 649 patients. In our analysis, we use the top 1000 transcriptomic features with highest positive correlation with gender, the sensitive feature. The outcome variable is the VBM right hippocampal volume, ranging between $[0.4, 0.6]$. Missing values are generated for the last 900 features under the three missing data mechanisms: MCAR, $\text{logit}(\pi(\boldsymbol{z}_i)) = 0.5$; MAR, $\text{logit}(\pi(\boldsymbol{z}_i)) = 2 - \frac{1}{25}\sum_{j=1}^{50} x_{ij}$; MNAR, $\text{logit}(\pi(\boldsymbol{z}_i)) = 2 - \frac{1}{25}\sum_{j=101}^{150} x_{ij}$. As shown in Table 1, the main findings are consistent with those from the analysis of COMPAS dataset.

5 Discussions

In this paper, we provide both an upper and lower bounds on fairness estimation error in the complete data space and evaluate the theoretical properties in extensive numerical experiments. Our work provides the first known results on fairness guarantee when learning with incomplete data as representations. We expect the work to offer insight into this area of research. Indeed this area offers fertile ground for future research. For example, many methods have been developed for handling missing data including imputation. It is of potential interest to investigate the impact of these methods on algorithmic fairness and refine these methods for the purpose of developing fair ML algorithms using the imputed data.

REFERENCES

Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. arXiv preprint arXiv:1905.12843, 2019.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchne. Machine bias. ProPublica. Retrieved from https://www.propublica.org/article/ how-we-analyzed-the-compas-recidivism-algorithm, 2016.

Martin Anthony and Peter L Bartlett. Neural network learning: Theoretical foundations. cambridge university press, 2009.

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

Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. Data decisions and theoretical implications when adversarially learning fair representations. arXiv preprint arXiv:1707.00075, 2017.

Victor Bouvier, Philippe Very, Clément Chastagnol, Myriam Tami, and Céline Hudelot. Robust domain adaptation: Representations, weights and inductive bias. arXiv preprint arXiv:2006.13629, 2020.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency, pp. 77–91, 2018.

Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. Data Mining and Knowledge Discovery, 21(2):277–292, 2010.

Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. arXiv preprint arXiv:1810.08810, 2018.

Alexandra Chouldechova and Aaron Roth. A snapshot of the frontiers of fairness in machine learning. Communications of the ACM, 63(5):82–89, 2020.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 797–806, 2017.

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

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

Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In Advances in Neural Information Processing Systems, pp. 2791–2801, 2018.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In Proceedings of the 3rd innovations in theoretical computer science conference, pp. 214–226, 2012.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, pp. 259–268, 2015.

Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im) possibility of fairness. arXiv preprint arXiv:1609.07236, 2016.

Milena A Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. Potential biases in machine learning algorithms using electronic health record data. JAMA internal medicine, 178(11):1544–1547, 2018.

Paula Gordaliza, Eustasio Del Barrio, Gamboa Fabrice, and Jean-Michel Loubes. Obtaining fairness using optimal transport theory. In International Conference on Machine Learning, pp. 2357–2365, 2019.

Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. The case for process fairness in learning: Feature selection for fair decision making. In NIPS Symposium on Machine Learning and the Law, volume 1, pp. 2, 2016.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In Advances in neural information processing systems, pp. 3315–3323, 2016.

David Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension. J. Comb. Theory, Ser. A, 69(2):217–232, 1995.

Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In 2011 IEEE 11th International Conference on Data Mining Workshops, pp. 643–650. IEEE, 2011.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 35–50. Springer, 2012.

Svetlana Kiritchenko and Saif M Mohammad. Examining gender and race bias in two hundred sentiment analysis systems. arXiv preprint arXiv:1805.04508, 2018.

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. Algorithmic fairness. In Aea papers and proceedings, volume 108, pp. 22–27, 2018.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In Advances in neural information processing systems, pp. 4066–4076, 2017.

Roderick JA Little and Donald B Rubin. Statistical analysis with missing data, volume 793. John Wiley & Sons, 2019.

Fernando Martínez-Plumed, Cèsar Ferri, David Nieves, and José Hernández-Orallo. Fairness and missing values. arXiv preprint arXiv:1905.12728, 2019.

Northpointe. Compas risk & need assessment system: Selected questions posed by inquiring agencies. 2010.

Luca Oneto, Michele Donini, and Massimiliano Pontil. General fair empirical risk minimization. arXiv preprint arXiv:1901.10080, 2019.

Dana Pessach and Erez Shmueli. Algorithmic fairness. arXiv preprint arXiv:2001.09784, 2020.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In Advances in Neural Information Processing Systems, pp. 5680–5689, 2017.

Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. Ensuring fairness in machine learning to advance health equity. Annals of internal medicine, 169(12):866–872, 2018.

Ievgen Redko, Amaury Habrard, and Marc Sebban. Theoretical analysis of domain adaptation with optimal transport. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 737–753. Springer, 2017.

Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. arXiv preprint arXiv:1703.03717, 2017.

Cynthia Rudin, Caroline Wang, and Beau Coker. The age of secrecy and unfairness in recidivism prediction. arXiv preprint arXiv:1811.00731, 2018.

Candice Schumann, Xuezhi Wang, Alex Beutel, Jilin Chen, Hai Qian, and Ed H Chi. Transfer of machine learning fairness across domains. arXiv preprint arXiv:1906.09688, 2019.

Shari Trewin, Sara Basson, Michael Muller, Stacy Branham, Jutta Treviranus, Daniel Gruen, Daniel Hebert, Natalia Lyckowski, and Erich Manser. Considerations for ai fairness for people with disabilities. AI Matters, 5(3):40–63, 2019.

Sahil Verma and Julia Rubin. Fairness definitions explained. In 2018 IEEE/ACM International Workshop on Software Fairness (FairWare), pp. 1–7. IEEE, 2018.

Christina Wadsworth, Francesca Vera, and Chris Piech. Achieving fairness through adversarial learning: an application to recidivism prediction. arXiv preprint arXiv:1807.00199, 2018.

Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In Proceedings of the IEEE International Conference on Computer Vision, pp. 5310–5319, 2019.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In Proceedings of the 26th international conference on world wide web, pp. 1171–1180, 2017.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In International Conference on Machine Learning, pp. 325–333, 2013.

Han Zhao and Geoff Gordon. Inherent tradeoffs in learning fair representations. In Advances in neural information processing systems, pp. 15675–15685, 2019.

Han Zhao, Remi Tachet des Combes, Kun Zhang, and Geoffrey J Gordon. On learning invariant representation for domain adaptation. arXiv preprint arXiv:1901.09453, 2019a.

Han Zhao, Amanda Coston, Tameem Adel, and Geoffrey J Gordon. Conditional learning of fair representations. arXiv preprint arXiv:1910.07162, 2019b.

A PROOF OF THEOREM 1

We begin by stating and proving the following theorem, which will play an important role in the proof of Theorem 1. Suppose we have N observations $\{\boldsymbol{z}_i\}_{i=1}^N$ drawn from an arbitrary domain $\mathcal{D}_S$ and let

$$\widehat{\mathcal{E}}(g,\omega) := \frac{1}{\sum_{i=1}^N R_i} \sum_{i=1}^N R_i \omega(\boldsymbol{z}_i) |g(\boldsymbol{x}_i) - y_i|$$

Furthermore we define $D^\omega = \mathbb{E}_S \omega(\boldsymbol{z})^2$, in this section we use $\mathbb{E}_S$ to represent that the expectation is taken with respect to this arbitrary domain $\mathcal{D}_S$. Then we have the following result:

Theorem 3. *Let $\mathcal{H}$ be a hypothesis set with VC-dimension (pseudo dimension in regression setting) d. Let g be an arbitrary prediction model from hypothesis set $\mathcal{H}$. If $D^\omega \leq N/8$, then, for any $\delta > 0$, with probability at least $1-\delta$, the following bound holds:*

$$\left| \mathbb{E}_S \omega(\boldsymbol{z}) \left| g(\boldsymbol{x}) - y(\boldsymbol{x}) \right| - \widehat{\mathcal{E}}(g,\omega) \right| \leq \sqrt{\frac{128 D^\omega C_d(N, D^\omega, \delta)}{N}} + \frac{16 C_d(N, D^\omega, \delta) B}{3N} \tag{4}$$

$$C_d(N, D^\omega, \delta) = \begin{cases} \log \frac{(d+1)(8e)^{d+1}}{\delta} + \frac{d}{2} \log \frac{N}{2D^\omega} & \textit{in classification} \\ \log \frac{2}{\delta} \left(\frac{8e}{d}\right)^d + \frac{3d}{2} \log \frac{N}{\sqrt[3]{2D^\omega}} & \textit{in regression} \end{cases}$$

Proof. The proof follows a standard approach in deriving generalization error bound related to VC-dimension (or pseudo-dimension). Recall that an observation $\boldsymbol{z} = \{\boldsymbol{x}, y\}$, we begin by letting $f_g(\boldsymbol{z}) := \omega(\boldsymbol{z}) |g(\boldsymbol{x}) - y(\boldsymbol{x})|$ and $\sigma^2 = \mathbb{E}\omega^2$. Then since $g \in \mathcal{H}$, we let $\mathcal{F}$ denote the set of f_g. In the rest of the proof, we simply ignore subscription g and let $f(\boldsymbol{x}) := \omega(\boldsymbol{z}) |g(\boldsymbol{x}) - y(\boldsymbol{x})|$. This is possible since the analysis holds for arbitrary $g \in \mathcal{H}$, i.e. $f \in \mathcal{F}$. We simplify the notation by defining

$$\widehat{\mathcal{E}}(g,\omega) = \mathbb{P}_N f(\boldsymbol{z})$$

and

$$\mathbb{E}_S \omega(\boldsymbol{z}) \left| g(\boldsymbol{x}) - y(\boldsymbol{x}) \right| = \mathbb{P} f(\boldsymbol{z})$$

We further let $\mathbb{D}_N$ denote the training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$. We also consider a set of "ghost" sample $\mathbb{D}'_N = \{(\boldsymbol{x}'_i, y'_i)\}_{i=1}^N$ with size N. In reality we do not have access to them but we will make use of them to prove the theorem. We let f_N be the maximizer of $|\mathbb{P} f - \mathbb{P}_N f|$ in $\mathcal{F}$. Similar to $\mathbb{P}_N$, we can define $\mathbb{P}'_N$ for the ghost samples. Firstly notice that

$$I\left(|\mathbb{P} f_N - \mathbb{P}_N f_N| > t\right) I\left(|\mathbb{P} f_N - \mathbb{P}'_N f_N| < t/2\right) \leq I\left(|\mathbb{P}'_N f_N - \mathbb{P}_N f_N| > t/2\right)$$

Taking expecation with respect to the ghost sample yields

$$I\left(|\mathbb{P} f_N - \mathbb{P}_N f_N| > t\right) P_{D'_N}\left(|\mathbb{P} f_N - \mathbb{P}'_N f_N| < t/2\right) \leq P_{\mathbb{D}'_N}\left(|\mathbb{P}'_N f_N - \mathbb{P}_N f_N| > t/2\right)$$

Chebyshev's inequality gives

$$P_{D'_N}\left(|\mathbb{P} f_N - \mathbb{P}'_N f_N| > t/2\right) \leq \frac{4\mathrm{Var}\left[f_N\right]}{Nt^2} = \frac{4\sigma^2}{Nt^2} \leq \frac{4D^\omega}{Nt^2}$$

where the last inequality is given by the fact that $D^\omega \geq \mathbb{E}\omega^2 = \sigma^2$ (see detailed argument in Cortes et al. (2010)). This in turn gives

$$I\left(|\mathbb{P} f_N - \mathbb{P}_N f_N| > t\right) \left(1 - \frac{4D^\omega}{Nt^2}\right) \leq P_{D'_N}\left(|\mathbb{P}'_N f_N - \mathbb{P}_N f_N| > t/2\right)$$

when $t \geq \sqrt{\frac{8D^\omega}{N}}$, we have

$$P\left(\sup_{f \in \mathcal{F}} |\mathbb{P} f - \mathbb{P}_N f| \geq t\right) \leq 2P\left(\sup_{f \in \mathcal{F}} |\mathbb{P}'_N f - \mathbb{P}_N f| \geq t/2\right) \tag{5}$$

We further define $\mathcal{F}_{|N} = \{(f(\boldsymbol{x}_1, y_1), \ldots, f(\boldsymbol{x}_N, y_N)) \mid f \in \mathcal{H}\}$. Then

$$P\left(\sup_{f \in \mathcal{F}} |\mathbb{P}'_N f - \mathbb{P}_N f| \geq t/2\right) = P\left(\sup_{f \in \mathcal{F}_{|2N}} |\mathbb{P}'_N f - \mathbb{P}_N f| \geq t/2\right)$$

Let $\mathcal{N}_1(t/8, \mathcal{F}, 2N)$ denote the uniform covering number defined as

$$\mathcal{N}_1(t/8, \mathcal{F}, 2N) = \max_{\mathbb{D}_N, \mathbb{D}'_N} \mathcal{N}\left(t/8, \mathcal{F}_{|2N}, d_1\right) \tag{6}$$

where $\mathcal{N}\left(t/8, \mathcal{F}_{|2N}, d_1\right)$ is the $t/8$-covering number of set $\mathcal{F}_{|2N}$ with respect to L^1 distance. Now define $\mathcal{G} \subseteq \mathcal{F}_{|2N}$ as a $t/8$-cover of $\mathcal{F}_{|2N}$ with $|\mathcal{G}| \leq \mathcal{N}_1(t/8, \mathcal{F}, 2N)$. Then

$$\begin{aligned} P\left(\sup_{f \in \mathcal{F}_{|2N}} |\mathbb{P}'_N f - \mathbb{P}_N f| \geq t/2\right) &\leq P\left(\sup_{g \in \mathcal{G}} |\mathbb{P}'_N g - \mathbb{P}_N g| \geq t/4\right) \\ &\leq \mathcal{N}_1(t/8, \mathcal{F}, 2N) \sup_{g \in \mathcal{G}} P\left(|\mathbb{P}'_N g - \mathbb{P}_N g| \geq t/4\right) \\ &\leq 2\mathcal{N}_1(t/8, \mathcal{F}, 2N) \sup_{g \in \mathcal{G}} P\left(|\mathbb{P}_N g - \mathbb{P} g| \geq t/8\right) \end{aligned} \tag{7}$$

Then according to classical bound of covering number (Theorem 1 in Haussler (1995)):

$$\mathcal{N}\left(t/8, \mathcal{F}_{|2N}, d_1\right) < e(d+1)\left(\frac{16e}{t}\right)^d$$

The part remained is analysis of the probability $\sup_{g \in \mathcal{G}} P\left(|\mathbb{P}_N g - \mathbb{P} g| \geq t/8\right)$. By Bernstein's inequality

$$\sup_{g \in \mathcal{G}} P\left(|\mathbb{P}_N g - \mathbb{P} g| \geq t/8\right) \leq \exp\left(\frac{-Nt^2/128}{D^\omega + tB/24}\right) \tag{8}$$

Combining results above yields

$$\begin{aligned} P\left(\sup_{f \in \mathcal{F}} |\mathbb{P} f - \mathbb{P}_N f| \geq t\right) &< 4e(d+1)\left(\frac{16e}{t}\right)^d \exp\left(\frac{-Nt^2/128}{D^\omega + tB/24}\right) \\ &\leq 4e(d+1)\left(16e\sqrt{\frac{N}{8D^\omega}}\right)^d \exp\left(\frac{-Nt^2/128}{D^\omega + tB/24}\right) \end{aligned}$$

Let $\delta = 4e(d+1)\left(16e\sqrt{\frac{N}{8D^\omega}}\right)^d \exp\left(\frac{-Nt^2/128}{D^\omega + tB/24}\right)$. Simplify the equation gives

$$Nt^2 - \frac{16C_d(N, D^\omega, \delta)B}{3}t - 128D^\omega C_d(N, D^\omega, \delta) = 0$$

where

$$C_d(N, D^\omega, \delta) = \log\frac{(d+1)(8e)^{d+1}}{2\delta} + \frac{d}{2}\log\frac{N}{2D^\omega}$$

This equation has non-negative solution

$$\begin{aligned} t_\delta &= \frac{\frac{16}{3}C_d(N, D^\omega, \delta)B + \sqrt{(\frac{16}{3}C_d(N, D^\omega, \delta)B)^2 + 512ND^\omega C_d(N, D^\omega, \delta)}}{2N} \\ &\leq \frac{16C_d(N, D^\omega, \delta)B}{3N} + \sqrt{\frac{128D^\omega C_d(N, D^\omega, \delta)}{N}} \end{aligned}$$

Thus we have

$$1 - \delta \leq P\left(\sup_{f \in \mathcal{F}} |\mathbb{P} f - \mathbb{P}_N f| \leq t_\delta\right) \leq P\left(\sup_{f \in \mathcal{F}} |\mathbb{P} f - \mathbb{P}_N f| \leq \frac{16C_d(N, D^\omega, \delta)B}{3N} + \sqrt{\frac{128D^\omega C_d(N, D^\omega, \delta)}{N}}\right)$$

As an alternative result, we can make use of the Bennett's inequality instead of Bernstein's inequality, which results in

$$P\left(\sup_{f \in \mathcal{F}} |\mathbb{P} f - \mathbb{P}_N f| \leq \sqrt{\frac{2BC_d(N, D^\omega, \delta)}{\log(1 + B/D^\omega)N}}\right) \geq 1 - \delta$$

Now we consider the regression case, in which we assume $\mathcal{H}$ has pseudo dimension d. Notice that all the results above hold before equation (7). Now by Theorem 12.2 in Anthony & Bartlett (2009), we have that

$$\mathcal{N}\left(t/8, \mathcal{F}_{|2N}, d_1\right) \leq \mathcal{N}\left(t/8, \mathcal{F}_{|2N}, d_\infty\right) \leq \left(\frac{16Ne}{td}\right)^d$$

when $N \geq d/2$. Combining with (8) yields

$$\begin{aligned} P\left(\sup_{f\in\mathcal{F}} |\mathbb{P}f - \mathbb{P}_N f| \geq t\right) &< 2\left(\frac{16Ne}{td}\right)^d \exp\left(\frac{-Nt^2/128}{D^\omega + tB/24}\right) \\ &\leq 2\left(\frac{16Ne}{d}\sqrt{\frac{N}{8D^\omega}}\right)^d \exp\left(\frac{-Nt^2/128}{D^\omega + tB/24}\right) \end{aligned}$$

Let $\delta = 2\left(\frac{16Ne}{d}\sqrt{\frac{N}{8D^\omega}}\right)^d \exp\left(\frac{-Nt^2/128}{D^\omega+tB/24}\right)$ and define

$$C_d(N, D^\omega, \delta) = \log\frac{2}{\delta}\left(\frac{8e}{d}\right)^d + \frac{3d}{2}\log\frac{N}{\sqrt[3]{2D^\omega}}$$

Following exactly the same procedure as above, we can have

$$1-\delta \leq P\left(\sup_{f\in\mathcal{F}} |\mathbb{P}f - \mathbb{P}_N f| \leq t_\delta\right) \leq P\left(\sup_{f\in\mathcal{F}} |\mathbb{P}f - \mathbb{P}_N f| \leq \frac{16C_d(N, D^\omega, \delta)B}{3N} + \sqrt{\frac{128D^\omega C_d(N, D^\omega, \delta)}{N}}\right)$$

□

Now we are ready to prove Theorem 1:

Proof. From Theorem 3, notice that for both groups $a \in \{0, 1\}$:

$$\left|\mathcal{E}_a(g) - \mathbb{E}_{S_a}\omega(\boldsymbol{x})(g(\boldsymbol{x}) - y(\boldsymbol{x}))\right| = \left|\mathbb{E}_{S_a}(\omega_0(\boldsymbol{x}) - \omega(\boldsymbol{x}))\,|g(\boldsymbol{x}) - y(\boldsymbol{x})|\,\right| \tag{9}$$

By triangle inequality, combining (9) and (4) yields that with at least probability $1 - 2\delta$:

$$\begin{aligned} &\left|\left|\mathcal{E}_0(g) - \mathcal{E}_1(g)\right| - \left|\widehat{\mathcal{E}}_0(g,\omega) - \widehat{\mathcal{E}}_1(g,\omega)\right|\right| \\ \leq &\sum_{a\in\{0,1\}} \left|\mathbb{E}_{S_a}(\omega_0(\boldsymbol{x}) - \omega(\boldsymbol{x}))\,|g(\boldsymbol{x}) - y(\boldsymbol{x})|\,\right| + \left|\mathbb{E}_{S_a}\omega(\boldsymbol{x})\,|g(\boldsymbol{x}) - y(\boldsymbol{x})| - \widehat{\mathcal{E}}_a(g,\omega)\right| \\ \leq &\sum_{a\in\{0,1\}} \left|\mathbb{E}_{S_a}(\omega_0(\boldsymbol{x}) - \omega(\boldsymbol{x}))\,|g(\boldsymbol{x}) - y(\boldsymbol{x})|\,\right| + \sqrt{\frac{128D_a^\omega C_d(n_a, D_a^\omega, \delta)}{n_a}} + \frac{16C_d(n_a, D_a^\omega, \delta)B}{3n_a} \end{aligned} \tag{10}$$

Notice that by definition of total variation distance:

$$d_{\text{TV}}(\mathcal{D}_{T_a}||\widehat{\omega}\mathcal{D}_{S_a}) \geq \left|\mathbb{E}_{S_a}(\omega_0(\boldsymbol{x}) - \omega(\boldsymbol{x}))\,|g(\boldsymbol{x}) - y(\boldsymbol{x})|\,\right|$$

Finally substituting δ to $\delta/2$ yields the result.

□

B PROOF OF THEOREM 2

Proof. Since we have $\omega = \omega_0$, we only use ω in the proof for the sake of simplicity. The proof is inspired by the technique used in Theorem 9 in Cortes et al. (2010). Assume $(\boldsymbol{x}, y) = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n_0+n_1}$ is the complete cases, in which n_0 of the data belongs to sensitive group $A = 0$. Let

$$\phi_g(\boldsymbol{x}, y) = \widehat{\mathcal{E}}_0(g,\omega) - \widehat{\mathcal{E}}_1(g,\omega) = \frac{1}{n_0}\sum_{i=1}^{n_0}\omega(\boldsymbol{x}_{0,i}, y_{0,i})|g(\boldsymbol{x}_{0,i}) - y_{0,i}| - \frac{1}{n_1}\sum_{i=1}^{n_1}\omega(\boldsymbol{x}_{1,i}, y_{1,i})|g(\boldsymbol{x}_{1,i}) - y_{1,i}|$$

with $g \in \mathcal{H}$. Obviously when true weights are adopted, we have $\mathbb{E}\phi_g(\boldsymbol{x}, y) = \mathcal{E}_0(g) - \mathcal{E}_1(g)$. Without loss of generality, we assume $n_0 < n_1$. We let $\sigma_i^2 := \mathrm{Var}_{S_i}(\omega|g - y|)$. Furthermore let $\sigma^2 = \sigma_0^2 + (n_0/n_1)\sigma_1^2$. Now consider $U := \frac{\mathbb{E}\phi_g(\boldsymbol{x},y) - \phi_g(\boldsymbol{x},y)}{\sigma}$. Notice that

$$\mathbb{E}Z^2 = \frac{\frac{1}{n_0}\sigma_0^2 + \frac{1}{n_1}\sigma_1^2}{\sigma^2} = \frac{1}{n_0} \tag{11}$$

Meanwhile we can split the expectation into

$$\mathbb{E}U^2\mathbf{1}_{|U|\in[0,1/(k\sqrt{n_0}))} + \mathbb{E}U^2\mathbf{1}_{|U|\in[1/(k\sqrt{n_0}),u/\sqrt{n_0})} + \mathbb{E}U^2\mathbf{1}_{|U|\in[u/\sqrt{n_0},+\infty)}$$

which is upper bounded by

$$\frac{1}{k^2 n_0} + \frac{u^2}{n_0}\mathcal{P}(u/\sqrt{n_0} > |U| > 1/(k\sqrt{n_0})) + \mathbb{E}U^2\mathbf{1}_{|U|\in[u/\sqrt{n_0},+\infty)}$$

Combined with (11) yields

$$\mathcal{P}(u/\sqrt{n_0} > |U| > 1/(k\sqrt{n_0})) \geq \frac{k^2-1}{k^2u^2} - \frac{n_0}{u^2}\mathbb{E}U^2\mathbf{1}_{|U|\in[u/\sqrt{n_0},+\infty)} \tag{12}$$

Now notice that $n_0\mathbb{E}U^2\mathbf{1}_{|U|\in[u/\sqrt{n_0},+\infty)}$ can be written as

$$\begin{aligned} n_0\mathbb{E}U^2\mathbf{1}_{|U|\in[u/\sqrt{n_0},+\infty)} &= \int_0^{+\infty} \mathcal{P}\left[n_0|U|^2 1_{|U|>\frac{u}{\sqrt{n_0}}} > t\right] dt \\ &= \int_0^{u^2} \mathcal{P}\left[|U| > \frac{u}{\sqrt{n_0}}\right] dt + \int_{u^2}^{+\infty} \mathcal{P}\left[|U| > \sqrt{\frac{t}{n_0}}\right] dt \\ &= u^2\mathcal{P}\left[|U| > \frac{u}{\sqrt{n_0}}\right] + \int_{u^2}^{+\infty} \mathcal{P}\left[|U| > \sqrt{\frac{t}{n_0}}\right] dt \end{aligned} \tag{13}$$

The probability in the last line can be upper bounded by

$$\begin{aligned} &\mathcal{P}\left[|U| > \sqrt{\frac{t}{n_0}}\right] \leq \mathcal{P}\left[|(\frac{1}{n_0}\sum_{i=1}^{n_0}\omega(\boldsymbol{x}_{0,i}, y_{0,i})|g(\boldsymbol{x}_{0,i}) - y_{0,i}| - \mathbb{E}_{S_0}\omega(\boldsymbol{x}, y)|g(\boldsymbol{x}) - y|)| > \frac{\sigma}{2}\sqrt{\frac{t}{n_0}}\right] \\ &+ \mathcal{P}\left[|(\frac{1}{n_1}\sum_{i=1}^{n_1}\omega(\boldsymbol{x}_{1,i}, y_{1,i})|g(\boldsymbol{x}_{1,i}) - y_{1,i}| - \mathbb{E}_{S_1}\omega(\boldsymbol{x}, y)|g(\boldsymbol{x}) - y|)| > \frac{\sigma}{2}\sqrt{\frac{t}{n_0}}\right] \\ &\leq \exp\left(-\frac{\sigma^2 t}{8\sigma_0^2 + 4/3B\sigma\sqrt{\frac{t}{n_0}}}\right) + \exp\left(-\frac{(n_1/n_0)\sigma^2 t}{8\sigma_1^2 + 4/3B\sigma\sqrt{\frac{t}{n_0}}}\right) \end{aligned}$$

where the second inequality is given by Bernstein's inequality. We now state that

$$\frac{(n_1/n_0)\sigma^2 t}{8\sigma_1^2 + 4/3B\sigma\sqrt{\frac{t}{n_0}}} \geq \frac{t}{8 + 4/3\sqrt{t}}$$

To see this, consider two cases $\sigma > \sigma_1$ and $\sigma \leq \sigma_1$. If $\sigma > \sigma_1$ then

$$\frac{(n_1/n_0)\sigma^2 t}{8\sigma_1^2 + 4/3B\sigma\sqrt{\frac{t}{n_0}}} \geq \frac{\sigma^2 t}{8\sigma^2 + 4/3B\sigma\sqrt{\frac{t}{n_0}}} \geq \frac{t}{8 + 4/3\sqrt{t}}$$

where the second inequality is given by the assumption $B^2/\sigma^2 \leq n_0$. If $\sigma \leq \sigma_1$

$$\frac{(n_1/n_0)\sigma^2 t}{8\sigma_1^2 + 4/3B\sigma\sqrt{\frac{t}{n_0}}} = \frac{t((n_1/n_0)\sigma_0^2 + \sigma_1^2)}{8\sigma_1^2 + 4/3B\sigma\sqrt{\frac{t}{n_0}}} \geq \frac{t\sigma_1^2}{8\sigma_1^2 + 4/3B\sigma_1^2\sqrt{\frac{t}{n_0}}} \geq \frac{t}{8 + 4/3\sqrt{t}}$$

Similarly $\frac{\sigma^2 t}{8\sigma_0^2 + 4/3B\sigma\sqrt{\frac{t}{n_0}}} \geq \frac{t}{8+4/3\sqrt{t}}$. Notice that $\sqrt{t} \geq 1/k$, take $k = 24$, we have

$$\mathcal{P}\left[|U| > \sqrt{\frac{t}{n_0}}\right] \leq 2\exp\left(-\frac{t}{8 + 4/3\sqrt{t}}\right) \leq 2\exp\left(-\frac{3\sqrt{t}}{5}\right)$$

Plug into (13) gives that

$$n_0\mathbb{E}U^2\mathbf{1}_{|U|\in[u/\sqrt{n_0},+\infty)} \leq 2u^2\exp\left(-\frac{3u}{5}\right)+\int_{u^2}^{+\infty}2\exp\left(-\frac{3\sqrt{t}}{5}\right)dt$$
$$=2\left(u^2+\frac{10u}{3}+2\left(\frac{5}{3}\right)^2\right)\exp\left(-\frac{3u}{5}\right)$$

when $u=12$, above is smaller than 0.29. In this case, (12) yields

$$\mathcal{P}(|U|>1/(24\sqrt{n_0}))>\mathcal{P}(u/\sqrt{n_0}>|U|>1/(24\sqrt{n_0}))=\frac{575}{576u^2}-\frac{0.29}{u^2}=\frac{7}{10u^2}=\frac{7}{10}\left(\frac{1}{12}\right)^2 \tag{14}$$

Finally we have the observation

$$\mathcal{P}(|U|>1/(24\sqrt{n_0}))=\mathcal{P}\left[\left|\Big(\mathcal{E}_0(g)-\mathcal{E}_1(g)\Big)-\Big(\widehat{\mathcal{E}}_0(g,\omega)-\widehat{\mathcal{E}}_1(g,\omega)\Big)\right|>\frac{1}{24}\sqrt{\frac{\sigma_0^2}{n_0}+\frac{\sigma_1^2}{n_1}}\right]$$

which completes the proof of first argument. To see the second argument, it can be proven that when

$$\widehat{\Delta}_S(g,\omega)\geq\frac{13}{2}\sqrt{\frac{\sigma_0^2(g,\omega)}{n_0}+\frac{\sigma_1^2(g,\omega)}{n_1}}>\frac{u+1/24}{2}\sqrt{\frac{\sigma_0^2(g,\omega)}{n_0}+\frac{\sigma_1^2(g,\omega)}{n_1}}$$

the interval $\left[\widehat{\Delta}_S(g,\omega)-u/\sqrt{n_0},\widehat{\Delta}_S(g,\omega)+u/\sqrt{n_0}\right]$ will never intersect with the interval $\left[-\widehat{\Delta}_S(g,\omega)-1/(24\sqrt{n_0}),-\widehat{\Delta}_S(g,\omega)+1/(24\sqrt{n_0})\right]$. Hence we have

$$\mathcal{P}\left[\left|\Delta_T(g)-\widehat{\Delta}_S(g,\omega)\right|>\frac{1}{24}\sqrt{\frac{\sigma_0^2(g,\omega)}{n_0}+\frac{\sigma_1^2(g,\omega)}{n_1}}\right]>\mathcal{P}(u/\sqrt{n_0}>|U|>1/(24\sqrt{n_0}))$$

The third argument can be proved in a similar way: when $\widehat{\Delta}_S(g,\omega)\leq\frac{1}{72}\sqrt{\frac{\sigma_0^2(g,\omega)}{n_0}+\frac{\sigma_1^2(g,\omega)}{n_1}}$, we have that

$$\left|\Delta_T(g)-\widehat{\Delta}_S(g,\omega)\right|\geq\left|\Big(\mathcal{E}_0(g)-\mathcal{E}_1(g)\Big)-\Big(\widehat{\mathcal{E}}_0(g,\omega)-\widehat{\mathcal{E}}_1(g,\omega)\Big)\right|-2\widehat{\Delta}_S(g,\omega)$$
$$\geq\frac{1}{72}\sqrt{\frac{\sigma_0^2(g,\omega)}{n_0}+\frac{\sigma_1^2(g,\omega)}{n_1}}$$

□

C PROOF OF COROLLARY 2

Proof. When the propensity score model is not accurate, by Theorem 2 we have

$$\left|\Big(\mathcal{E}_0(g)-\mathcal{E}_1(g)\Big)-\Big(\widehat{\mathcal{E}}_0(g,\omega)-\widehat{\mathcal{E}}_1(g,\omega)\Big)\right|\geq\frac{1}{24}\sqrt{\frac{\sigma_0^2(g,\omega_0)}{n_0}+\frac{\sigma_1^2(g,\omega_0)}{n_1}} \tag{15}$$

with probability at least $\frac{7}{10}(\frac{1}{12})^2$. Now recall the convergence assumption of estimated propensity score model: $|\omega-\omega_0|=\mathcal{O}_p((n_0+n_1)^{-1/2})$. We have that $D_a^{\omega}-D_a^{\omega_0}=\mathbb{E}_{S_a}\omega^2-\mathbb{E}_{S_a}\omega_0^2=\mathcal{O}_p((n_0+n_1)^{-1/2})$ for both groups $a\in\{0,1\}$. This yields

$$\sqrt{\frac{\sigma_0^2(g,\omega)}{n_0}+\frac{\sigma_1^2(g,\omega)}{n_1}}=\sqrt{\frac{\sigma_0^2(g,\omega_0)}{n_0}+\frac{\sigma_1^2(g,\omega_0)}{n_1}}+o_p(n_0^{-1/2}+n_1^{-1/2}) \tag{16}$$

Meanwhile, since g is consistent, we have

$$\left|\mathbb{E}_{S_a}\omega(\boldsymbol{x})(g(\boldsymbol{x})-y(\boldsymbol{x}))-\mathcal{E}_a(g)\right|=\left|\mathbb{E}_{S_a}(\omega-\omega_0)|g(\boldsymbol{x})-y(\boldsymbol{x})|\right|=o_p((n_0+n_1)^{-1/2}) \tag{17}$$

for both group.

$$
\begin{aligned}
&\left|\Big(\mathcal{E}_0(g)-\mathcal{E}_1(g)\Big)-\Big(\widehat{\mathcal{E}}_0(g,\omega)-\widehat{\mathcal{E}}_1(g,\omega)\Big)\right| \\
\geq&\left|\Big(\mathbb{E}_{S_0}\omega(\boldsymbol{x})(g(\boldsymbol{x})-y(\boldsymbol{x}))-\mathbb{E}_{S_1}\omega(\boldsymbol{x})(g(\boldsymbol{x})-y(\boldsymbol{x}))\Big)-\Big(\widehat{\mathcal{E}}_0(g,\omega)-\widehat{\mathcal{E}}_1(g,\omega)\Big)\right| \\
-&\left|\Big(\mathbb{E}_{S_0}\omega(\boldsymbol{x})(g(\boldsymbol{x})-y(\boldsymbol{x}))-\mathbb{E}_{S_1}\omega(\boldsymbol{x})(g(\boldsymbol{x})-y(\boldsymbol{x}))\Big)-\Big(\mathcal{E}_0(g)-\mathcal{E}_1(g)\Big)\right| \\
\geq&\frac{1}{24}\sqrt{\frac{\sigma_0^2(g,\omega)}{n_0}+\frac{\sigma_1^2(g,\omega)}{n_1}}-\left|\Big(\mathbb{E}_{S_0}\omega(\boldsymbol{x})(g(\boldsymbol{x})-y(\boldsymbol{x}))-\mathbb{E}_{S_1}\omega(\boldsymbol{x})(g(\boldsymbol{x})-y(\boldsymbol{x}))\Big)-\Big(\mathcal{E}_0(g)-\mathcal{E}_1(g)\Big)\right|
\end{aligned}
$$

where the last inequality holds with probability at least $\frac{7}{1440}$. Combining equation (16) and (17), we know that for arbitrary δ, there exists N_0 and N_1 such that whenever $n_0 > N_0$ and $n_1 > N_1$, with probability at least $1-\delta$,

$$
\begin{aligned}
&\left|\Big(\mathbb{E}_{S_0}\omega(\boldsymbol{x})(g(\boldsymbol{x})-y(\boldsymbol{x}))-\mathbb{E}_{S_1}\omega(\boldsymbol{x})(g(\boldsymbol{x})-y(\boldsymbol{x}))\Big)-(\mathcal{E}_0(g,\omega)-\mathcal{E}_1(g,\omega))\right| \\
\leq&\left(\frac{1}{24}-\frac{1}{25}\right)\sqrt{\frac{\sigma_0^2(g,\omega_0)}{n_0}+\frac{\sigma_1^2(g,\omega_0)}{n_1}}
\end{aligned}
$$

Thus with probability at least $\frac{7}{1440}-\delta$,

$$
\left|\Big(\mathcal{E}_0(g)-\mathcal{E}_1(g)\Big)-\Big(\widehat{\mathcal{E}}_0(g,\omega)-\widehat{\mathcal{E}}_1(g,\omega)\Big)\right|\geq\frac{1}{25}\sqrt{\frac{\sigma_0^2(g,\omega_0)}{n_0}+\frac{\sigma_1^2(g,\omega_0)}{n_1}}
$$

which yields the desired argument.

□