

Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved

Jiahao Chen
cjiahao@gmail.com

Nathan Kallus
Cornell Tech
New York, New York, USA
kallus@cornell.edu

Xiaojie Mao*
Cornell Tech
New York, New York, USA
xm77@cornell.edu

Geoffry Svacha
svacha@gmail.com

Madeleine Udell
Cornell University
Ithaca, New York, USA
udell@cornell.edu

ABSTRACT

Assessing the fairness of a decision making system with respect to a protected class, such as gender or race, is challenging when class membership labels are unavailable. Probabilistic models for predicting the protected class based on observable proxies, such as surname and geolocation for race, are sometimes used to impute these missing labels for compliance assessments. Empirically, these methods are observed to exaggerate disparities, but the reason why is unknown. In this paper, we decompose the biases in estimating outcome disparity via threshold-based imputation into multiple interpretable bias sources, allowing us to explain when over- or underestimation occurs. We also propose an alternative weighted estimator that uses soft classification, and show that its bias arises simply from the conditional covariance of the outcome with the true class membership. Finally, we illustrate our results with numerical simulations and a public dataset of mortgage applications, using geolocation as a proxy for race. We confirm that the bias of threshold-based imputation is generally upward, but its magnitude varies strongly with the threshold chosen. Our new weighted estimator tends to have a negative bias that is much simpler to analyze and reason about.

CCS CONCEPTS

- Social and professional topics → Race and ethnicity; Geographic characteristics;
- Applied computing → IT governance; Law.

KEYWORDS

fair lending, disparate impact, protected class, racial discrimination, race imputation, probabilistic proxy model, Bayesian Improved Surname Geocoding

*Corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

FAT* '19, January 29–31, 2019, Atlanta, GA, USA

© 2019 Association for Computing Machinery.

ACM ISBN 978-1-4503-6125-5/19/01...\$15.00

<https://doi.org/10.1145/3287560.3287594>

ACM Reference Format:

Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. 2019. Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved. In *FAT* '19: Conference on Fairness, Accountability, and Transparency (FAT* '19), January 29–31, 2019, Atlanta, GA, USA*. ACM, New York, NY, USA, 13 pages. <https://doi.org/10.1145/3287560.3287594>

1 INTRODUCTION

Models for high stakes decision making have ethical and legal needs to demonstrate a lack of discrimination with respect to protected classes [2, 26]. Examples of such decisions include employment and compensation [14, 20], university admissions [4], and sentence and bail setting [3, 7, 16]. Another example relevant to the financial services industry is credit decisioning [6], which is a classification problem where these ethical concerns are enshrined in concrete regulatory compliance requirements. Credit decisions must be shown to comply with a myriad of federal and state fair lending laws, some of which are summarized in [6]¹. Some of these laws define protected classes, such as race and gender, where discrimination on the basis of a customer’s membership in these classes is prohibited. Table 1 summarizes the protected classes defined by the Fair Housing Act (FHA) [28] and Equal Credit Opportunity Act (ECOA) [29].

Table 1: Protected classes defined under fair lending laws.

Law	FHA[28]	ECOA[29]
age		X
color	X	X
disability	X	
exercised rights under CCPA		X
familial status (household composition)	X	
gender identity	X	
marital status (single or married)		X
national origin	X	X
race	X	X
recipient of public assistance		X
religion	X	X
sex	X	X

¹In this paper, we restrict our discussion and citation of applicable laws to those of the United States of America.

1.1 Assessing fairness with unknown protected class membership

When demonstrating that credit decisions comply with these fair lending laws, we sometimes run into situations where fairness and bias assessments must be done on populations without knowing their memberships in protected classes, because it is illegal or operationally difficult to do so. For example, credit card and auto loan companies must demonstrate that the way they extend credit is not racially discriminatory, yet are not allowed to ask applicants what race they are when they apply for credit.² Similarly, health plans can only solicit race and ethnicity information for new members but cannot obtain the same information for existing members [18]. Given the lack of secure protocols that permit disparity evaluation with encrypted protected classes [23], disparate impact assessments for these situations have to impute the mostly (or entirely) missing labels corresponding to the protected class, usually by relying on observed proxy variables that can predict class memberships. The imputed protected classes are then used by regulators in assessing disparate impact (but they are not allowed to be used in decision making). Generally, any model that imputes the missing protected attribute value based on other, observed variables is known as a *proxy model*, and such a model that is based on predicting conditional class membership probabilities is known as a *probabilistic proxy model*.

For example, for assessing adverse action with regard to race in credit decisions, regulators like the Bureau of Consumer Financial Protection (BCFP)³ have been known in the past to use a probabilistic proxy model to impute the customers' unknown race labels [12]. They used a naïve Bayes classifier, the Bayesian Improved Surname Geocoding (BISG) method, to predict the probability of race membership given the customer's surname and address of residence [19]. Specifically, assuming that surname and location are statistically independent given race, BISG uses Bayes's rule to compute race membership probabilities from the conditional distributions of surname given race and of location given race as inferred by census data. This methodology [12] notably supported a \$98 million fine against a major auto loan lender [11]. This case generated some controversy [8–10, 13, 24], in part due to empirical findings that the amount of disparate impact estimated by BISG appears to overestimate true disparities [1, 31]. However, the cause for this overestimation phenomenon is unknown, as is whether overestimation is to be expected always, or whether or not underestimation of disparate impact is also possible. This observation forms the motivation for our current work, which is broadly applicable to any fairness assessment where an unobserved protected class must be imputed using a proxy model. The aim is not to criticize the use of proxy models in general, but rather to provide a more informed analysis of the statistical biases inherent in any assessment where membership in protected classes must be imputed.

Main results. This paper investigates the bias in estimating demographic disparity (Definition 2.2) when a proxy model is employed

²Lenders may ask applicants to self-identify in a voluntary basis, with the understanding that the answer will not affect the outcome of the application and that the information is collected for compliance assessment only [15, 12 CFR §1002.5(b)].

³Formerly the Consumer Financial Protection Bureau (CFPB). Citations and references reflect the name at time of publication.

to impute a protected class. We present the first theoretical results describing when the use of proxy models can lead to biased estimates of outcome disparity, which explains the overestimates observed in the past, and also offers insights on the practical use of proxy models. More specifically, our key contributions are:

- (1) We derive the (statistical) bias for the commonly used thresholded estimator, where a label is assigned only if the proxy model predicts a label with probability exceeding a predefined threshold [1, 12, 31] (Definition 2.4, Theorem 3.3). We decompose its bias into multiple sources, which gives a set of interpretable conditions under which the thresholded estimator can over- or underestimate the outcome disparity.
- (2) We present a new weighted estimator for demographic disparity (Definition 2.5) that uses soft classification based on proxy model outputs as opposed to hard imputation. We derive its bias (Theorem 3.1) and find that the weighted estimator has only one bias source.
- (3) We validate our results on a public mortgage data set, using geolocation as the sole variable in a proxy model for race. We identify the specific source of bias that can account for the overestimation of the thresholded estimator, which can explain the overestimation bias of using proxy methods observed in previous literature.
- (4) We discover that the estimation bias is sensitive to the threshold used in class imputation based on the proxy model. This shows the intrinsic limitation of the thresholded estimator.

2 EVALUATING THE FAIRNESS OF A BINARY DECISION

We have three main variables of interest:

Binary decision Y with $Y = 1$ representing a favorable outcome, such as the approval of loan application or college admission offer, and $Y = 0$ representing an unfavorable outcome.

Protected class A such as gender or race; often, we will write $A = a$ for the *advantaged group* and $A = b$ for the *disadvantaged group*.

Proxy variable Z a set of covariates taking values $z \in \mathcal{Z}$ used to predict A in a probabilistic proxy model.

We present only the binary case $A \in \{a, b\}$ for simplicity. Unless otherwise stated, for multiclass A , our results generalize straightforwardly to the pairwise outcome disparity between any advantaged group and any disadvantaged group. (Additional details are given in Appendices B and C.)

Definition 2.1. The *mean group outcome* for the group $A = u$ is

$$\mu(u) = \mathbb{E}(Y | A = u),$$

where $u \in \{a, b\}$ and \mathbb{E} is the usual expectation with respect to population distribution.

Definition 2.2. The *demographic disparity*, or Calders-Verwer gap [5, 22], δ , is the difference in mean group outcomes between the advantaged and disadvantaged groups:

$$\delta = \mu(a) - \mu(b) \tag{1}$$

A positive demographic disparity δ means that a higher proportion of the advantaged group $A = a$ receive a favorable outcome

than the disadvantaged group b , i.e. the disadvantaged group experiences an outcome disparity. Demographic disparity is simple to understand and is widely used [25, 30, 32], despite its flaws [17, 21].

When the labels for the protected class A are known, the demographic disparity can be simply and reliably estimated by the difference of within-group sample means. Otherwise, a probabilistic proxy model may be used to estimate the probabilities $\mathbb{P}(A = a | Z)$ and $\mathbb{P}(A = b | Z)$ of membership to the different groups within A . BISG [19] is an example of a probabilistic proxy model where Z is taken to be surname and geolocation and conditional probabilities are estimated using the naïve Bayes methodology. The goal of this paper is to quantify the bias in estimating demographic disparity using a proxy model. We neglect the estimation bias intrinsic to estimating the proxy model itself, so that $\{\mathbb{P}(A = u | Z) : u \in \{a, b\}\}$ describes the true population distribution of A conditioned on Z and hence the intrinsic uncertainty of predicting A from Z .

2.1 Thresholded estimator

In order to introduce the estimators for demographic disparity, assume that we have N independent and identically distributed (iid) samples $(Y_i, Z_i)_{i=1}^N$. The true membership A_i of the i th sample is unknown, but we have access to the probabilistic proxy estimates $\{\mathbb{P}(A_i = u | Z_i) : u \in \{a, b\}, i \in \{1, \dots, N\}\}$ by simply applying our proxy model to the observed proxy variable Z_i .

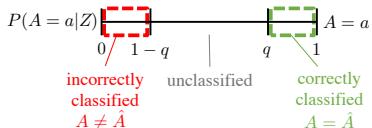
The ordinary approach is to predict a single value for class membership, \hat{A}_i [1, 31]:

Definition 2.3. Let $q \in [\frac{1}{2}, 1]$. Then the *thresholded estimated membership* \hat{A}_i for the i th unit is

$$\hat{A}_i = \begin{cases} a, & \mathbb{P}(A_i = a | Z_i) > q, \\ b, & \mathbb{P}(A_i = b | Z_i) > q, \\ \text{NA, otherwise,} \end{cases}$$

where NA stand for an unclassified observation that is excluded from the subsequent outcome disparity evaluation.

Considering a unit with $A = a$, this estimation rule can be summarized pictorially as follows



where dashed boxes represent correct ($\hat{A} = A$, green) and incorrect classifications ($\hat{A} \neq A$, red), and the middle is unclassified.

We can use these predicted labels to estimate the mean group outcomes and demographic disparity by mimicking the simple estimator for group sample means difference, but imputing \hat{A} for the unknown A . This leads to the following estimator:

Definition 2.4. Let $(Y_i, Z_i)_{i=1}^N$ be N iid samples, $\{\hat{A}_i\}_{i=1}^N$ be the estimated labels according to the thresholding rule in Definition 2.3, and $\mathbb{I}(S)$ be the indicator function for some set S . Then, the *thresholded estimators* for mean group outcomes and demographic

Table 2: Lending policy of Example 2.6, which always accepts high income applicants and never low income applicants. Each number in the table cells represents the number of applicants who truly belong to race a or b . The percentages in parentheses represent the average loan approval rate in that quadrant.

Neighborhood	True race	
	a	b
high income	70 (100%) ⁱ	30 (100%) ⁱ
low income	30 (0%) ⁱⁱ	70 (0%)

ⁱ Misclassified as race a in high-income neighborhood.

ⁱⁱ Misclassified as race b in low-income neighborhood.

disparity are

$$\hat{\mu}_q(u) = \frac{\sum_{i=1}^N \mathbb{I}(\hat{A}_i = u) Y_i}{\sum_{i=1}^N \mathbb{I}(\hat{A}_i = u)}, u \in \{a, b\},$$

$$\hat{\delta}_q = \hat{\mu}_q(a) - \hat{\mu}_q(b). \quad (2)$$

2.2 Weighted estimator

The form of the thresholded estimators above shows clearly the use of a hard classification rule. Under a probabilistic proxy model, this hard classification rule inevitably misclassifies some individuals, given the intrinsic uncertainty in classifying the protected class. Moreover, the threshold rule results in a group of unclassified individuals who are removed from the outcome disparity evaluation. To avoid these problems, we propose a new estimator that accounts for the soft classification generated by the proxy.

Definition 2.5. Let $(Y_i, Z_i)_{i=1}^N$ be defined as above. Then, the *weighted estimators* for mean group outcomes and demographic disparity are

$$\hat{\mu}_W(u) = \frac{\sum_{i=1}^N \mathbb{P}(A_i = a | Z_i) Y_i}{\sum_{i=1}^N \mathbb{P}(A_i = a | Z_i)}, u \in \{a, b\}$$

$$\hat{\delta}_W = \hat{\mu}_W(a) - \hat{\mu}_W(b) \quad (3)$$

2.3 Toy examples

In this part, we use two hypothetical toy examples to demonstrate the intuition regarding when the thresholded estimator (2) can overestimate the demographic disparity. In both examples, we consider evaluating the disparity of loan approval with respect to two races $A = a$ and $A = b$. However, the true race A is unknown, so the geolocation Z is used to estimate race. Suppose that there are only two neighborhoods where all people live in one neighborhood have high income and all people live in the other neighborhood have low income. We further assume that the high-income neighborhood is primarily occupied by the advantaged group a and the low-income neighborhood is primarily occupied by the disadvantaged group b . Therefore, the thresholding rule with $q = 0.5$ classifies all people living in the high-income neighborhood as the advantaged group, i.e., $\hat{A} = a$, and all people living in the low-income neighborhood as the disadvantaged group, i.e., $\hat{A} = b$.

Table 3: Lending policy of Example 2.7, with affirmative action that favors the disadvantaged group over the advantaged group at any given income level.

Neighborhood	True race	
	<i>a</i>	<i>b</i>
high income	70 (70%)	30 (80%) ⁱ
low income	30 (20%) ⁱⁱ	70 (30%)

ⁱ Misclassified as race *a* in high-income neighborhood.

ⁱⁱ Misclassified as race *b* in low-income neighborhood.

Example 2.6. Consider an extreme lending policy: all people with high income get their loans approved, while all people with low income are rejected, no matter what their races are. See Table 2 for the illustration. Simple calculations show that the true demographic disparity $\delta = 40\%$ ⁴ while the thresholded estimator $\hat{\delta}_{0.5} = 100\%$ ⁵. Thus the thresholded estimator overestimates the demographic disparity. The reason for the overestimation is that the race proxy is correlated with the loan approval outcome because of the dependence between geolocation and socioeconomic status: people who live in the neighborhood primarily occupied by the advantaged group are also more likely to get loan approval because they have relatively high socioeconomic status, e.g. high income in this example. These people are classified as advantaged group because their neighborhoods are associated with high probability of belonging to the advantaged group. As a result, the thresholding rule misclassifies people who are from the disadvantaged group but get loan approval as the advantaged group. In this way, the thresholding rule leads to underestimates of the loan acceptance rate of the disadvantaged group. Analogously, the thresholding rule leads to overestimates of the loan acceptance rate of the advantaged group because it misclassifies people from the advantaged group but not likely to get loan approval as the disadvantaged group. Consequently, the misclassification of the protected attribute is dependent with the outcome, which ultimately leads to the overestimation of demographic disparity. The dependence between the misclassification and the outcome results from the *inter-geolocation* outcome variation: the socioeconomic status, race proxy probabilities (i.e., race ratios), and loan acceptance rates vary across different neighborhoods such that the favorable outcome is positively correlated with the probability of belonging to the advantaged group.

Example 2.7. Now consider a hypothetical lending policy with affirmative action that approves the disadvantaged group with higher rate than the advantaged group with the same income level. But overall people with high income are still more likely to be accepted. See Table 3 for a concrete example. Simple calculations show that the true demographic disparity $\delta = 10\%$ ⁶, which means that the disadvantaged group overall has lower chance to get loan approval due to their population concentration in the low-income neighborhood. However, the thresholded estimator gives $\hat{\delta}_{0.5} = 46\%$ ⁷

⁴ $\delta = \frac{1}{100}(70 \times 1 + 30 \times 0) - \frac{1}{100}(30 \times 1 + 70 \times 0) = 0.4$.

⁵ $\hat{\delta}_{0.5} = \frac{1}{100}(70 \times 1 + 30 \times 1) - \frac{1}{100}(30 \times 0 + 70 \times 0) = 1$.

⁶ $\delta = \frac{1}{100}(70 \times 0.7 + 30 \times 0.2) - \frac{1}{100}(30 \times 0.8 + 70 \times 0.3) = 0.1$.

⁷ $\hat{\delta}_{0.5} = \frac{1}{100}(70 \times 0.7 + 30 \times 0.8) - \frac{1}{100}(30 \times 0.2 + 70 \times 0.3) = 0.46$.

that overestimates the demographic disparity. The reason for the overestimation is that the disadvantaged group have higher approval rate than their neighbors from the advantaged group. As a result, misclassifying part of the disadvantaged group as the advantaged group raises the average approval rate for the advantaged group. Similarly, misclassifying part of the advantaged group as the disadvantaged group lowers down the average approval rate for the disadvantaged group. Consequently, the misclassification of the protected attribute is also dependent with the outcome and ultimately leads to overestimation of demographic disparity. In contrast to Example 2.6, the dependence between the misclassification and the outcome results from the intra-geolocation outcome variation: people from different protected groups living in the same locations have different chance of getting the favorable outcome.

In both examples, the protected attribute misclassification made by the thresholding rule is not uniformly at random. Instead, the misclassification shows systematic pattern with respect to the outcome either due to the inter-geolocation outcome variation (Example 2.6) or intra-geolocation outcome variation (Example 2.7). In Section 3, we will formalize the main intuition in these two examples and show that the interplay of the two bias sources captured by these two examples determines the overestimation or underestimation of the thresholded estimator. In Section 4, we will further identify the specific bias source that accounts for the estimation bias of the thresholded estimator in a mortgage data set.

3 BIAS IN THRESHOLDED AND WEIGHTED ESTIMATORS

In this section, we derive the asymptotic biases for the thresholded estimator (2) and weighted estimator (3) for demographic disparity. We also provide some interpretable sufficient conditions under which these two estimators overestimate or underestimate the demographic disparity.

3.1 Weighted estimator

THEOREM 3.1. *Let A be a binary protected class with values a and b . The bias of the weighted estimator $\hat{\delta}_W$ in Definition 2.5 for demographic disparity δ in (1) is*

$$\hat{\delta}_W - \delta = [\hat{\mu}_W(a) - \mu(a)] - [\hat{\mu}_W(b) - \mu(b)],$$

where as $N \rightarrow \infty$, the biases in the weighted estimators for the mean group outcomes $\hat{\mu}_W(u)$, for $u \in \{a, b\}$, converge almost surely to

$$\hat{\mu}_W(u) - \mu(u) \xrightarrow{a.s.} -\frac{\mathbb{E}[\text{Cov}(\mathbb{I}(A=u), Y | Z)]}{\mathbb{P}(A=u)}. \quad (4)$$

We omit the *a.s.* (almost sure) notation later for brevity.

COROLLARY 3.2. *If Y is independent of A conditionally on Z , then the weighted estimator for demographic disparity, $\hat{\delta}_W$, is asymptotically unbiased.*

If Y is independent of A conditionally on Z , then the advantaged group and disadvantaged group with the same Z values are equally treated in terms of the average outcome. Nevertheless, this does not contradict the existence of overall disparity against the disadvantaged group in terms of the unconditioned average outcome, as an example of Simpson's paradox [27]. The conditional independence

assumption required by Corollary 3.2 is trivially satisfied if Y is the output of some function $f(X)$, and Z includes the input features X , since Y is now determined entirely by Z . Such a situation arises naturally when Y is the output of machine learning algorithms. In the Appendix C.4, we show a semi-synthetic example based on the mortgage dataset where the weighted estimator is asymptotically unbiased when the proxy model is well constructed so that the conditional independence assumption is satisfied.

However, this conditional independence assumption may not hold in practice and the weighted estimator may be biased. Consider the example of loan application with race proxy based on geolocation. If within most locations, the affirmative action described in Example 2.7 is present, i.e., the disadvantaged group is more likely to get a loan than their neighbors from the advantaged group, then Y covaries negatively with $\mathbb{I}(A = a)$ and positively with $\mathbb{I}(A = b)$ conditionally on $Z = z$, for most values of $z \in \mathcal{Z}$, which implies that the weighted estimator overestimates the demographic disparity. On the contrary, if within most locations, the advantaged group is more likely to get loan than their neighbors belonging to the disadvantaged group, then Y covaries in the exact opposite way, implying that the weighted estimator underestimates the demographic disparity. Overall, the estimation bias of the weighted estimator depends on the intra-geolocation variation of the loan application outcome.

3.2 Thresholded estimator

Before stating the asymptotic bias for the estimator (2), we define the following terms for $u \in \{a, b\}$:

$$\begin{aligned}\Delta_1(u) &= \mathbb{E}[Y | \mathbb{P}(A = u | Z) > q, A = u^c] \\ &\quad - \mathbb{E}[Y | \mathbb{P}(A = u | Z) > q, A = u], \\ \Delta_2(u) &= \mathbb{E}[Y | \mathbb{P}(A = u | Z) \leq q, A = u] \\ &\quad - \mathbb{E}[Y | \mathbb{P}(A = u | Z) > q, A = u],\end{aligned}$$

where q is the threshold used for estimating race, and u^c is the class opposite to u , i.e., $u^c = b$ if $u = a$ and $u^c = a$ if $u = b$. $\Delta_1(u)$ measures the outcome mean discrepancy for two different protected groups *within* the same proxy probability range, and $\Delta_2(u)$ measures the outcome mean discrepancy *across* different proxy probability ranges for the same protected group. Consider the example of loan application with race proxy based on geolocation. In this example, $\Delta_1(u)$ measures the loan approval disparity between two race groups who live in locations that are primarily occupied by one of these two races (in terms of the threshold q). In contrast $\Delta_2(u)$ measures the loan approval rate disparity between people belonging to a race who live in locations that are primarily occupied by this race group and people belonging to this race who live in locations that are less occupied by this race. Therefore, $\Delta_1(u)$ roughly characterizes the *intra-geolocation* variation of loan outcome and $\Delta_2(u)$ roughly characterizes the *inter-geolocation* variation of loan outcome.

THEOREM 3.3. *Let A be a binary protected class with values a and b . The bias for the thresholded estimator $\hat{\delta}_q$ in Definition 2.4 is:*

$$\hat{\delta}_q - \delta = [\hat{\mu}_q(a) - \mu(a)] - [\hat{\mu}_q(b) - \mu(b)],$$

and as $N \rightarrow \infty$, for $u \in \{a, b\}$ and $u^c \in \{a, b\}$ as the class opposite to u ,

$$\begin{aligned}\hat{\mu}_q(u) - \mu(u) &\rightarrow \Delta_1(u)C_1(u) - \Delta_2(u)C_2(u) \\ &\quad + (\Delta_1(u) - \Delta_2(u))C_3(u),\end{aligned}$$

where

$$C_1(u) = \mathbb{P}(\hat{A} = u | A = u)\mathbb{P}(A = u^c | \hat{A} = u),$$

$$C_2(u) = \mathbb{P}(A = u | \hat{A} = u)\mathbb{P}(\hat{A} \neq u | A = u), \text{ and}$$

$$C_3(u) = \mathbb{P}(\hat{A} \neq u | A = u)\mathbb{P}(A = u^c | \hat{A} = u).$$

Here $\hat{A} \neq u$ means that $\hat{A} = u^c$ or $\hat{A} = NA$.

The generalization to multiclass A is straightforward and in Appendix C.1 we show that the bias formula in Theorem 3.3 for binary protected class capture the main effects of the bias for multiclass A .

Theorem 3.3 shows that the occurrence of overestimation or underestimation depends on complex interplay of the Δ terms. In the following corollary, we provide the simplest set of sufficient conditions for the overestimation and underestimation to demonstrate the main intuition.

COROLLARY 3.4. *Suppose*

$$\Delta_1(a) \geq 0, \tag{5i}$$

$$-\Delta_1(b) \geq 0, \tag{5ii}$$

$$-\Delta_2(a) \geq 0, \text{ and} \tag{5iii}$$

$$\Delta_2(b) \geq 0, \tag{5iv}$$

and at least one inequality holds strictly. Then in the limit $N \rightarrow \infty$, $\hat{\mu}_q(a) > \mu(a)$, $\hat{\mu}_q(b) < \mu(b)$, and $\hat{\delta}_q > \delta$, i.e. the thresholded estimator overestimates the demographic disparity.

Conversely, let (i') $\Delta_1(a) \leq 0$, (ii') $-\Delta_1(b) \leq 0$, (iii') $-\Delta_2(a) \leq 0$, and (iv') $\Delta_2(b) \leq 0$, and at least one inequality holds strictly. Then in the limit $N \rightarrow \infty$, $\hat{\mu}_q(a) < \mu(a)$, $\hat{\mu}_q(b) > \mu(b)$, and $\hat{\delta}_q < \delta$, i.e. the thresholded estimator underestimates the demographic disparity.

The conditions (5i)–(5iv) are demonstrated in Figure 1.

Conditions (5i)–(5ii) exactly formalize the intuition captured by Example 2.7. In our loan application example, using a geolocation-based race proxy, these two conditions capture the intra-geolocation outcome variation: on average higher proportion of disadvantaged group b receive the favorable outcome than the advantaged group a among all the geolocations that are primarily occupied by one race (in terms of the threshold q).

The conditions (5iii)–(5iv) exactly formalize the intuition captured by Example 2.6. They characterize the dependence between the decision outcome and the proxy probability conditionally on the protected attribute. Specifically, condition (5iii) holds when the decision outcome is positively correlated with the probability of belonging to the advantaged group, while condition (5iv) holds when the decision outcome is negatively correlated with the probability of belonging to the disadvantaged group conditionally on the true protected attribute. In the example of loan application with race proxy based on geolocation, geolocation is correlated with both race ratio and socioeconomic status (e.g., income, FICO score, etc.): locations that are primarily occupied by the advantaged racial groups tend to be associated with relatively higher socioeconomic status while locations that are primarily occupied by the disadvantaged

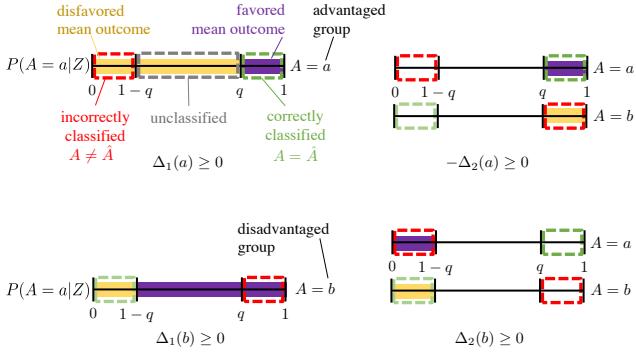


Figure 1: Illustration of conditions (5i)–(5iv) in Corollary 3.4, illustrating the complex interplay between membership in a protected class A and outcome Y that give rise to overestimated demographic disparity using the thresholded estimator of Definition 2.4. The horizontal axes show $P(A = a|Z)$, the probability that a probabilistic proxy model predicts membership in the advantaged group a . Dashed boxes represent correct ($\hat{A} = A$, green) and incorrect classifications ($\hat{A} \neq A$, red), from the thresholded estimated membership rule of Definition 2.3, with the interval $(1 - q, q)$ always unclassified. Solid bars represent favored (purple) and disfavored mean outcomes (yellow).

racial groups tend to be associated with relatively lower socioeconomic status. As a result, people who live in locations dominant by the advantaged racial group are assigned with high probability of belonging to the advantaged group, and they are more likely to get approved in loan application because of relatively higher socioeconomic status. Conversely, the loan approval is negatively correlated with the probability of belonging to disadvantaged groups. This example demonstrates that (5iii)–(5iv) are likely to hold when the predictors Z are also strongly correlated with the decision outcome, e.g., the geolocation is correlated with the loan approval because of the socioeconomic status disparities across different geolocations.

The following corollary shows that (5i)–(5iv) have effects with different magnitudes on the overestimation bias:

COROLLARY 3.5. Let A be a binary protected class with values a and b . For $u \in \{a, b\}$ and u^c as the class opposite to u , the quantities $C_1(u), C_2(u), C_3(u)$ in Theorem 3.3 are related in the following way:

(i) $C_2(u) > C_3(u)$ if and only if

$$\mathbb{P}(A = u | \hat{A} = u) > \mathbb{P}(A = u^c | \hat{A} = u). \quad (6)$$

(ii) $C_2(u) > C_1(u)$ if and only if

$$\mathbb{P}(A = u) > \mathbb{P}(\hat{A} = u) \quad (7)$$

Thus if the conditions (6) and (7) both hold, then $C_2(u) > C_1(u)$ and $C_2(u) > C_3(u)$.

Condition (6) holds as long as the proxy model is reasonably predictive of the true protected class. Condition (7) usually holds for the thresholded estimated membership (Definition 2.3) when a high threshold q is used: a fairly high fraction of observations are unclassified under the thresholded estimation rule with high threshold q ,

so the fraction of observations classified into one protected class is usually lower than the fraction of observations actually belonging to that protected class. Therefore, when a high threshold q is used, usually conditions (5iii)–(5iv) contribute more to the overestimation bias than (5i)–(5ii). As a result, even if (5i)–(5ii) are violated, as long as (5iii)–(5iv) hold, the thresholded estimator is still very likely to overestimate demographic disparity. In contrast, when a low threshold is used, the difference between $\mathbb{P}(A = u) > \mathbb{P}(\hat{A} = u)$ is usually smaller, and thus $C_2(u) - C_1(u)$ is also smaller. In this case, (5i)–(5ii) and (5iii)–(5iv) may have comparable contribution to the estimation bias. In Section 4.2, we will verify these facts by using different thresholds on the mortgage dataset.

4 NUMERICAL RESULTS

4.1 Analysis of bias terms in synthetic data

In this part, we simulate an example where geolocation is used to construct proxy for race. We use this example to demonstrate different sources of overestimation and underestimation bias of the weighted estimator and the thresholded estimator.

Experimental setup. We consider race as the unknown protected attribute A with $A = a$ being the advantaged racial group and $A = b$ being the disadvantaged racial group. Suppose there are three different neighborhoods ($Z = z_1, z_2, z_3$) where the proportions of people belonging to the advantaged group are 0.2, 0.5, 0.8 respectively. These proportions are in turn the race proxy probabilities. For example, all people who live in the neighborhood z_1 are assigned with $(\mathbb{P}(A = a | Z = z_1) = 0.2, \mathbb{P}(A = b | Z = z_1) = 0.8)$ as their race proxy. If we apply the thresholded estimation rule with $q = 0.75$, the estimated races for people living in neighborhoods tracts are:

	$Z = z_1$	$Z = z_2$	$Z = z_3$
\hat{A}	b	NA	a

We assume that geolocation is strongly correlated with people's income X : $\mathbb{E}(X | Z = z_1) < \mathbb{E}(X | Z = z_2) = 2 < \mathbb{E}(X | Z = z_3)$. Therefore, z_1 can be considered as the low-income tract where the proportion of people belonging to the disadvantaged group is high ($\mathbb{P}(A = b | Z = z_3) = 0.8$), z_2 can be considered as middle-income tract where the proportions of people belonging to the advantaged group and the disadvantaged group are equal ($\mathbb{P}(A = a | Z = z_2) = \mathbb{P}(A = b | Z = z_2) = 0.5$), and z_3 can be considered as high-income tract where the proportion of people belonging to the advantaged group is also high ($\mathbb{P}(A = a | Z = z_1) = 0.8$). The decision outcome Y is loan approval, which we assume solely depends on the income X : we simulate Y according to $\mathbb{P}(Y = 1 | X) = 1/(1 + \exp(-\lambda(X - 2)))$ with $\lambda > 0$. Thus people with higher income are more likely to get approved ($Y = 1$). Here λ controls the extent to which loan approval Y depends on income X . Since income X solely depends on geolocation Z , λ indirectly controls the dependence of the outcome Y on geolocation Z and thus on the corresponding race proxy probabilities ($\mathbb{P}(A = a | Z), \mathbb{P}(A = b | Z)$).

In each experiment, we set the total population size of neighborhoods z_1, z_2, z_3 to be 3000, 4000, and 5000 respectively. Each experiment is repeated 30 times and the average estimated demographic disparity from the thresholded estimator, the weighted estimator, and using the true race is shown in Figure 2.

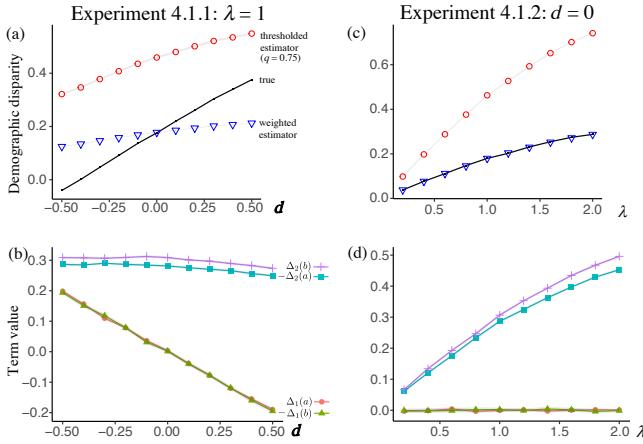


Figure 2: Top row: performance of the thresholded ($\hat{\delta}_{q=0.75}$, red circles) and weighted ($\hat{\delta}_W$, blue triangles) estimators of demographic disparity in Experiments 4.1.1 (a) and 4.1.2 (c), relative to the true demographic disparity (black line). Bottom row: values of the terms $\Delta_1(a)$, $-\Delta_1(b)$, $\Delta_2(a)$, and $-\Delta_2(b)$ in Corollary 3.4 in Experiments 4.1.1 (b) and 4.1.2 (d), showing their contributions to the bias. Shown in light gray is the confidence interval within two standard deviations, averaged over 30 simulations, which in all cases is comparable to the width of the plotted lines.

Experiment 4.1.1. The income is normally distributed according to $N(\mu_X, 0.25)$ with mean value μ_X depending on both geolocation and race:

μ_X	$Z = z_1$	$Z = z_2$	$Z = z_3$
$A = a$	$1 + d$	$2 + d$	$3 + d$
$A = b$	$1 - d$	$2 - d$	$3 - d$

where d controls the discrepancy of income X between the two race groups within the same geolocation. We fix $\lambda = 1$ and vary d from -0.5 to 0.5 , which indirectly varies the magnitude of intra-geolocation decision outcome variation between the two races. By construction, d should affect only the $\Delta_1(a), \Delta_1(b)$ in (5i)–(5ii).

Experiment 4.1.2. The income is normally distributed according to $N(\mu_X, 0.25)$ with the mean value μ_X depending on only geolocation:

	$Z = z_1$	$Z = z_2$	$Z = z_3$
μ_X	1	2	3

In other words, people with different races within the same geolocation have the same income distribution and thus the same decision outcome distribution. Therefore, there is no *intra-geolocation* decision outcome variation between the two races. As a result, when we vary λ from 0.2 to 2 , only *inter-geolocation* outcome variation changes, which affects $\Delta_2(a), \Delta_2(b)$ in conditions (5iii)–(5iv).

Figure 2(b) shows that in Experiment 4.1.1, only $\Delta_1(a)$ and $\Delta_1(b)$ vary strongly with d , so they are largely responsible for the bias observed in Figure 2(a). When $d < 0$, the disadvantaged group have comparatively higher incomes and are thus more likely to be

approved in loan application. This implies that conditions (5i)–(5ii) hold and therefore result in overestimation bias for the thresholded estimator. Furthermore, Y covaries negatively with $\mathbb{I}(A = a)$ and positively with $\mathbb{I}(A = b)$ conditionally on Z , also resulting in an overestimation bias for the weighted estimator. When $d > 0$, however, conditions (5i)–(5ii) are violated. The terms $\Delta_1(a)$ and $\Delta_1(b)$ start to counteract overestimation bias, thus the overestimation bias decreases with d . Furthermore, Y now covaries positively with $\mathbb{I}(A = a)$ and negatively with $\mathbb{I}(A = b)$, also resulting in an underestimation bias for the weighted estimator.

Figure 2(d) shows that in Experiment 4.1.2, only the terms $\Delta_2(a)$ and $\Delta_2(b)$ vary strongly with λ , so these terms are responsible for the variation observed in the thresholded estimator $\hat{\delta}_q$ in Figure 2(c). As λ increases, both $-\Delta_2(a)$ and $\Delta_2(b) > 0$ increase, along with the overestimation bias of $\hat{\delta}_q$. In contrast, the weighted estimator $\hat{\delta}_W$ is unbiased because the income distribution does not depend on race, and so Y is independent with $\mathbb{I}(A = a)$ and $\mathbb{I}(A = b)$ conditionally on Z .

While presented only for the particular choices of $\lambda = 1$ for Experiment 4.1.1 and $d = 0$ for Experiment 4.1.2, the observation of which terms vary strongly with d and λ hold true for quite a few different choices.

4.2 Estimation biases in the HMDA mortgage data set with geolocation proxy for race

In this section, we use the public HMDA (Home Mortgage Disclosure Act) data set⁸ to demonstrate demographic disparity estimation bias when using a probabilistic proxy for race. This data set contains mortgage loan application records in the U.S. for which the geolocation (state, county, and census tract), self-reported race/ethnicity, and loan origination outcome were reported, and has been used in the literature to evaluate the BISG race proxy [1, 12, 31]. We use the loan data for the years 2011–2012, consistent with [12]. We denote $Y = 1$ if a loan application was approved or originated, and $Y = 0$ if it was denied. The final sample contains around 17 million observations with non-missing geolocation, race, and loan origination outcome information.

We consider a race proxy based only on geolocation, as the public data set is anonymized and omits surnames. This proxy is derived from the racial and ethnic composition of the U.S. population that is over 18 years of age, using the census tracts of the 2010 decennial census⁹. For example, the proportion of Hispanic, white, black, and API (Asian or Pacific Islander) in census tract 020100, Autauga County, Alabama are 0.02, 0.86, 0.10, 0.01 respectively, according to the 2010 census. This quadruple is assigned to all applicants from this census tract as their race proxy. This probabilistic proxy, while different from BISG, is nevertheless sufficient to demonstrate the general result regarding disparity estimation bias in Section 3. We present results for Hispanic, white, and black subpopulations in this section, and defer the result for API to the Appendix.

Estimation bias of thresholded estimator and weighted estimator. In Figure 3, we show the estimation bias of the thresholded estimator with different thresholds and the weighted estimator. Clearly the

⁸Data link: <https://www.consumerfinance.gov/data-research/hmda/explore>.

⁹We use the census tract level geolocation-only proxy constructed by the CFPB, as described in <https://github.com/cfpb/proxy-methodology>.

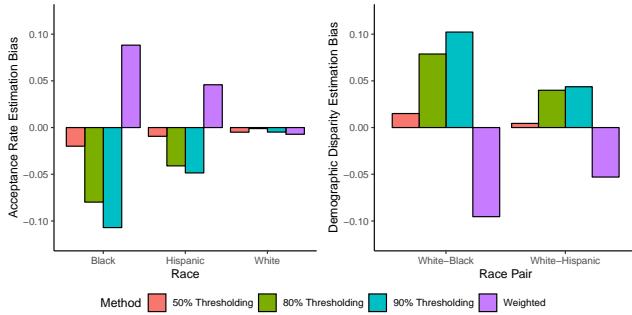


Figure 3: Left: estimation biases of loan acceptance rates μ for different races in the HMDA data set of Section 4.2, using the thresholded estimator for mean group outcome $\hat{\mu}_q$ ($q = 0.5, 0.8, 0.9$) from Definition 2.4, as well as the weighted estimator $\hat{\mu}_W$ from Definition 2.5, relative to the true mean group outcome μ calculated using the actual race labels. **Right:** estimation biases of demographic disparity δ between pairs of races, using the thresholded estimator $\hat{\delta}_q$ from Definition 2.4 and weighted estimator $\hat{\delta}_W$ from Definition 2.5, relative to the true demographic disparity δ calculated using the actual race labels.

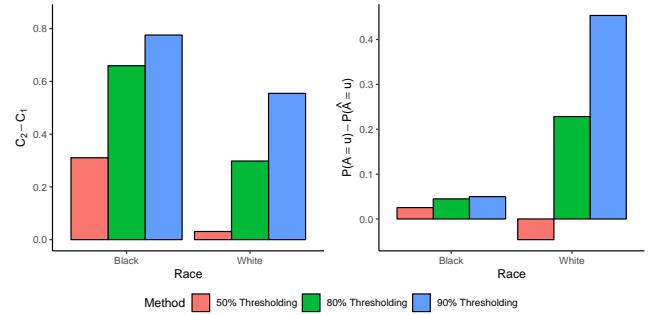


Figure 5: Left: the difference between the terms $C_2(u)$ and $C_1(u)$ in Theorem 3.3 for races $u = \text{black}$ or white and thresholds $q = 0.5, 0.8$ and 0.9 for estimating the race labels according to the thresholded rule of Definition 2.3. As q increases, $C_2(u)$ is increasingly larger than $C_1(u)$, and thus the $\Delta_2(u)$ terms contribute more to the estimation bias than the $\Delta_1(u)$ terms. **Right:** difference between the probabilities of the true and estimated race labels, $P(A=u) - P(\hat{A}=u)$. As q increases from 0.5 to 0.9 , the proportion of unclassified samples increases from 0.065 to 0.63 , which causes $P(\hat{A}=u)$ to decrease drastically. As a result, $C_2(u) - C_1(u)$ increases due to Corollary 3.5.

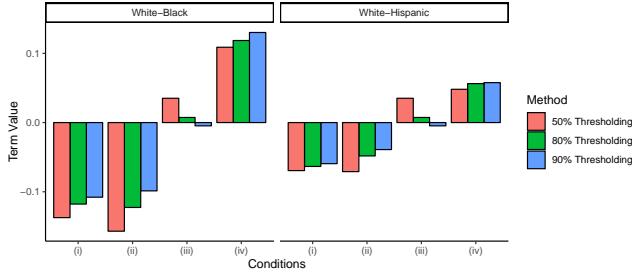


Figure 4: The values of the terms $\Delta_1(a)$, $-\Delta_1(b)$, $\Delta_2(a)$, and $-\Delta_2(b)$ **in Corollary 3.4 for the mortgage data set of Section 4.2**, for thresholds $q = 0.5, 0.8$, and 0.9 . Negative values demonstrate partial violation of the conditions of Corollary 3.4.

thresholded estimator underestimates the loan acceptance rate of black and Hispanic groups but it estimates the loan acceptance rate for White group accurately. As a result, the thresholded estimator overestimates the demographic disparity. Moreover, the overestimation bias tends to decrease as the threshold q decreases. In contrast, the weighted estimator displays the opposite performance and underestimates the demographic disparity.

Bias source of thresholded estimator. Figure 4 shows the conditions (5i)–(5iv). Only (5iv) strictly holds, meaning that people from the disadvantaged group (Hispanic or black) living in census tracts where their race ratio exceeds the corresponding classification threshold have lower average loan acceptance rate than people from the disadvantaged group living in census tracts where their race ratio is not as high as the classification threshold. This captures the *inter-geolocation* outcome variation: the loan approval is negatively

correlated with the disadvantaged group prevalence across the geolocations. In Example 2.6, we give a reason why loan acceptance is correlated with the race proxy probabilities: geolocation is correlated with both race proportions (i.e., race proxy probabilities) and socioeconomic status (e.g., income, FICO score, etc.) that affects loan approval. In Appendix C.3, we validate that such correlations are indeed obvious in the mortgage data set.

In contrast to conditions (5iv), conditions (5i)–(5ii) are strictly violated, and condition (5iii) is slightly violated. However, the thresholded estimators still have overestimation bias because (5iii)–(5iv) have dominant effects, as shown in Figure 5. Moreover, as the threshold q increases, conditions (5iii)–(5iv) start to dominate, increasing the overestimation bias. Nevertheless, the apparent reduction of overestimation bias when using lower threshold does not mean that using lower threshold is better in practice. Instead, the reduction of overestimation bias is the consequence of delicate counterbalance between the two opposing bias sources: the violation of conditions (5i)–(5ii) contributes to underestimation bias and the conditions (5iii)–(5iv) contribute to overestimation bias. Thus, the smaller bias from using a lower threshold is not a robust finding. For example, for the experiments in Section 4, changing the threshold from 0.75 to 0.5 does not affect either the race imputation result or the demographic disparity estimation bias. This reflects the intrinsic complexity of the thresholded estimator.

In summary, according to Corollary 3.5, the fact that the thresholded estimation rule for race excludes many unclassified examples makes the overestimation predominantly be determined by the inter-geolocation outcome variation captured by conditions (5iii)–(5iv), especially when the threshold q is very high. Moreover, strong racial segregation and socioeconomic status disparities across different geolocations make condition (5iii) or (5iv) (if not both) very

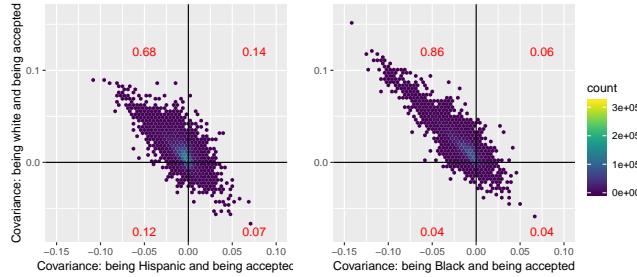


Figure 6: The population distribution across different combinations of within-census-tract-covariances between the loan acceptance and the membership to advantaged group or the membership to disadvantaged group. The red numbers in each quadrant represents the proportion of people who live in census tracts with the corresponding covariance combination.

likely to hold. As a result, the thresholded estimator tends to overestimate the demographic disparity, especially when high threshold is used. Since BISG also uses geolocation for race estimation, the reported overestimation bias of using thresholded estimator with BISG should be at least largely (if not all) due to the same reason. Moreover, the estimation bias of thresholded can be very sensitive to the threshold value because the threshold value influences the interplay of the bias sources in (5i)–(5iv). This reflects the intrinsic limitation of the thresholded estimator.

Bias source of the weighted estimator. Figure 6 shows the distribution of population who live in census tracts with different combinations of $\text{Cov}(\mathbb{I}(A = a), Y | Z)$ and $\text{Cov}(\mathbb{I}(A = b), Y | Z)$ with a representing white and b representing black or Hispanic. Theorem 3.1 implies that the upper left quadrants account for the underestimation bias of weighted estimator for demographic disparity, as these quadrants contain all census tracts where the loan acceptance is positively correlated with membership to the advantaged group ($\text{Cov}(\mathbb{I}(A = a), Y | Z) > 0$), while negatively correlated with the membership to the disadvantaged group ($\text{Cov}(\mathbb{I}(A = b), Y | Z) < 0$). Figure 6 shows that the majority of people live in census tracts falling in the upper left quadrants. In other words, most people live in census tracts where white people have a higher average loan approval rate than their black or Hispanic neighbors. This explains the underestimation bias of the weighted estimator since the estimation bias of the weighted estimator is solely determined by the intra-geolocation outcome variation according to Theorem 3.1.

Summary. The empirical results show that our theoretical analysis provides convincing explanations for the observed estimation bias of the thresholded estimator and the weighted estimator. We show that the bias sources of the weighted estimator and the thresholded estimator are very different: the bias of the weighted estimator is solely determined by the intra-geolocation variation, and the bias of the thresholded estimator is determined by both inter-geolocation variation and intra-geolocation variation. When high threshold is used, the inter-geolocation variation dominates.

In the mortgage dataset, we show that the racial segregation, socioeconomic status, outcome disparity pattern with respect to geolocation make the thresholded estimator tend to overestimate the demographic disparity, and the weighted estimator tend to underestimate the demographic disparity. This explains the overestimation bias of the thresholded estimator with BISG reported by previous literature. Moreover, our results show that the estimation bias of thresholded estimator is sensitive to the threshold because of complex interplay of different bias sources. As a result, the observed estimation bias pattern in one setting may hardly generalize to other settings. In contrast, the weighted estimator only has one bias source, and is thus easier to reason about.

5 CONCLUSIONS

This paper presents the first theoretical analysis of bias in outcome disparity assessments using a probabilistic proxy model for the unobserved protected class. In Theorem 3.3, we derived the bias of a thresholded estimator (Definition 2.4) that has been described in the literature. We also gave sufficient conditions in Corollary 3.4 to understand when this methodology is biased, and to what extent. Our theoretical analysis is valid whenever a proxy model is used with a thresholded estimator to impute protected class membership, and is thus consistent with previous studies that had observed an overestimation bias of the thresholded estimator based on BISG. When applied to the public HMDA mortgage dataset with geolocation as the sole proxy for race membership (Section 4.2), we found that the estimation bias of the thresholded estimator depends on the complex interaction of multiple different biases, producing a strong sensitivity to the precise value of the threshold used. Further studies will be needed to demonstrate the robustness of this numerical finding, particularly when using other proxy models and other measures of fairness. Nevertheless, our work signals caution for choosing the *ad hoc* thresholds which are used in practice, as without the ground truth labels, we are unable to determine the optimal choice of threshold.

To alleviate the theoretical challenges of the thresholded estimator, we also proposed a weighted estimator (Definition 2.5), which propagated the uncertainty resulting from the probabilistic proxy onto the final estimand. We found that the estimation bias of this weighted estimator has only one bias source, arising from the loan approval discrepancy between difference races within the same geolocation, which led to an overall underestimate. As the behavior of this estimator’s bias is simpler to reason about, we believe that the weighted estimator may be a useful new method to incorporate into outcome disparity evaluations when proxy models are used, especially if the sign of estimation bias can be determined with external knowledge.

REFERENCES

- [1] Arthur P Baines and Marsha J Courchane. 2014-11. Fair Lending: Implications for the Indirect Auto Finance Market. <https://www.craij.com/sites/default/files/publications/Fair-Lending-Implications-for-the-Indirect-Auto-Finance-Market.pdf>
- [2] Solon Barocas and Andrew Selbst. 2016. Big Data’s Disparate Impact. *California Law Review* 104, 1 (2016), 671–729. <https://doi.org/10.15779/Z38BG31>
- [3] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research* (2018), 1–42. <https://doi.org/10.1177/0049124118782533> arXiv:1703.09207

- [4] P. J. Bickel, E. A. Hammel, and J. W. O'Connell. 1975. Sex Bias in Graduate Admissions: Data from Berkeley. *Science* 187, 4175 (1975), 398–404. <https://doi.org/10.1126/science.187.4175.398>
- [5] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21, 2 (2010), 277–292. <https://doi.org/10.1007/s10618-010-0190-x>
- [6] Jiahao Chen. 2018. Fair Lending Needs Explainable Models for Responsible Recommendation. In *Proceedings of the 2nd FATREC Workshop on Responsible Recommendation (FATREC'18)*. ACM. arXiv:1808.04684
- [7] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5, 2 (2017). <https://doi.org/10.1089/big.2016.0047> arXiv:1703.0056
- [8] Committee on Financial Services. 2015-11. Unsafe at Any Bureaucracy: CFPB Junk Science and Indirect Auto Lending. https://financialservices.house.gov/uploadedfiles/11-24-15_cfpb_indirect_auto_staff_report.pdf
- [9] Committee on Financial Services. 2016-01. Unsafe at Any Bureaucracy, Part II: How the Bureau of Consumer Financial Protection Removed Anti-fraud Safeguards to Achieve Political Goals. https://financialservices.house.gov/uploadedfiles/cfpb_indirect_auto_part_ii.pdf
- [10] Committee on Financial Services. 2017-01. Unsafe at Any Bureaucracy, Part III: The CFPB's Vitiated Legal Case Against Auto-Lenders. https://financialservices.house.gov/uploadedfiles/1-18-17_cfpb_indirect_auto_staff_report_iii.pdf
- [11] Consumer Financial Protection Bureau. 2013-12. CFPB and DOJ Order Ally to Pay \$80 Million to Consumers Harmed by Discriminatory Auto Loan Pricing. <https://www.consumerfinance.gov/about-us/newsroom/cfpb-and-doj-order-ally-to-pay-80-million-to-consumers-harmed-by-discriminatory-auto-loan-pricing/>
- [12] Consumer Financial Protection Bureau. 2014. Using publicly available information to proxy for unidentified race and ethnicity: a methodology and assessment. <https://www.consumerfinance.gov/data-research/research-reports/using-publicly-available-information-to-proxy-for-unidentified-race-and-ethnicity/>
- [13] Consumer Financial Protection Bureau. 2018-05. Statement of the Bureau of Consumer Financial Protection on enactment of S.J. Res. 57. <https://www.consumerfinance.gov/about-us/newsroom/statement-bureau-consumer-financial-protection-enactment-sj-res-57/>
- [14] Delores A Conway and Harry V Roberts. 1983. Reverse Regression, Fairness, and Employment Discrimination. *Journal of Business & Economic Statistics* 1, 1 (1983), 75–85. <https://doi.org/10.1080/07350015.1983.10509326>
- [15] Division of Consumer and Community Affairs. 2011-07. 12 CFR Supplement I to Part 202 - Official Staff Interpretations. https://www.law.cornell.edu/cfr/text/12/appendix-Supplement_I_to_part_202
- [16] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances* 4, 1 (2018), eaao5580. <https://doi.org/10.1126/sciadv.aao5580>
- [17] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ACM, 214–226. <https://doi.org/10.1145/2090236.2090255>
- [18] Marc N. Elliott, Allen Fremont, Peter A Morrison, Philip Pantoja, and Nicole Lurie. 2008. A new method for estimating race/ethnicity and associated disparities where administrative records lack self-reported race/ethnicity. *Health Services Research* 43, 5p1 (2008), 1722–1736. <https://doi.org/10.1111/j.1475-6773.2008.00854.x>
- [19] Marc N. Elliott, Peter A. Morrison, Allen Fremont, Daniel F. McCaffrey, Philip Pantoja, and Nicole Lurie. 2009-04. Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Services and Outcomes Research Methodology* 9, 2 (2009-04), 69–83. <https://doi.org/10.1007/s10742-009-0047-1>
- [20] William H. Greene. 1984. Reverse regression: The algebra of discrimination. *Journal of Business and Economic Statistics* 2, 2 (1984), 117–120. <https://doi.org/10.1080/07350015.1984.10509378>
- [21] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, 3315–3323. arXiv:1610.02413
- [22] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-Aware Classifier with Prejudice Remover Regularizer. In *Machine Learning and Knowledge Discovery in Databases*, Peter A Flach, Tijl De Bie, and Nello Cristianini (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 35–50. https://doi.org/10.1007/978-3-642-33486-3_3
- [23] Niki Kilbertus, Adrià Gascón, Matt J Kusner, Michael Veale, Krishna P Gummadi, and Adrian Weller. 2018. Blind Justice: Fairness with Encrypted Sensitive Attributes, In Proceedings of the 35th International Conference on Machine Learning. *Proceedings of Machine Learning Research* 80, 2630–2639. arXiv:1806.03281
- [24] James Rufus Koren. 2016-08. Feds use Rand formula to spot discrimination. The GOP calls it junk science. *Los Angeles Times* (2016-08). <http://www.latimes.com/business/la-fi-rand-elliott-20160824-snap-story.html>
- [25] Zachary C Lipton, Alexandra Chouldechova, and Julian McAuley. 2017. Does mitigating ML's impact disparity require treatment disparity? (2017). arXiv:1711.07076
- [26] Cecilia Munoz, Megan Smith, and DJ Patil. 2016. *Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights*. Technical Report May. https://www.whitehouse.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf
- [27] Edward Hugh Simpson. 1951. The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society Series B* 13, 2 (1951), 238–241. <http://www.jstor.org/stable/2984065>
- [28] US Congress. 1968. 42 U.S.C. §3601 ff.: Fair Housing Act. <https://www.justice.gov/crt/fair-housing-act-2>
- [29] US Congress. 1974-10. 15 U.S.C. §1691 ff.: Equal Credit Opportunity Act. https://www.ecfr.gov/cgi-bin/text-idx?tpl=/ecfrbrowse/Title12/12ecfr202_main_.02.tpl
- [30] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification, In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. *Proceedings of Machine Learning Research* 54, 962–970. arXiv:1507.05259
- [31] Yan Zhang. 2016-01. Assessing Fair Lending Risks Using Race/Ethnicity Proxies. *Management Science* 64, 1 (2016-01), 178–197. <https://doi.org/10.1287/mnsc.2016.2579>
- [32] Indré Žliobaitė. 2015. On the relation between accuracy and fairness in binary classification. (2015). arXiv:1505.05723

Appendix A PROOFS FOR SECTION 3

PROOF OF THEOREM 3.1. By the strong law of large numbers, we have the almost sure convergence for the estimator

$$\begin{aligned}\hat{\mu}_W(a) &= \frac{\frac{1}{N} \sum_{i=1}^N \mathbb{P}(A_i = a | Z_i) Y_i}{\frac{1}{N} \sum_{i=1}^N \mathbb{P}(A_i = a | Z_i)} \\ &\xrightarrow{\text{a.s.}} \frac{\mathbb{E}[\mathbb{P}(A = a | Z)Y]}{\mathbb{E}[\mathbb{P}(A = a | Z)]} \\ &= \frac{\mathbb{E}[\mathbb{E}(\mathbb{I}(A = a) | Z)\mathbb{E}(Y | Z)]}{\mathbb{P}(A = a)}.\end{aligned}$$

For the true mean group outcome, introduce the trivial conditional expectation with respect to Z :

$$\mu(a) = \frac{\mathbb{E}[\mathbb{I}(A = a)Y]}{\mathbb{P}(A = a)} = \frac{\mathbb{E}[\mathbb{E}(\mathbb{I}(A = a)Y | Z)]}{\mathbb{P}(A = a)}$$

Thus, we can regroup the terms in the difference to recognize the conditional covariance

$$\begin{aligned}\hat{\mu}_W(a) - \mu(a) &\xrightarrow{\text{a.s.}} - \frac{\mathbb{E}[\mathbb{E}(\mathbb{I}(A = a)Y | Z) - \mathbb{E}(\mathbb{I}(A = a) | Z)\mathbb{E}(Y | Z)]}{\mathbb{P}(A = a)} \\ &= - \frac{\mathbb{E}[\text{Cov}(\mathbb{I}(A = a), Y | Z)]}{\mathbb{P}(A = a)}.\end{aligned}$$

The analogous results hold for $A = b$ everywhere. \square

PROOF OF COROLLARY 3.2. When Y is independent of A conditionally on Z , it immediately follows that

$$\text{Cov}(\mathbb{I}(A = a), Y | Z) = \text{Cov}(\mathbb{I}(A = b), Y | Z) = 0,$$

and thus

$$\begin{aligned}\hat{\mu}_W(a) - \mu(a) &\xrightarrow{\text{a.s.}} 0 \\ \hat{\mu}_W(b) - \mu(b) &\xrightarrow{\text{a.s.}} 0 \\ \therefore \hat{\delta}_W - \delta &\xrightarrow{\text{a.s.}} 0\end{aligned}$$

\square

PROOF OF THEOREM 3.3. Define the following events for $u_1, u_2 \in \{a, b\}$:

$$\begin{aligned}\mathcal{E}_q^+(u_1, u_2) &= \{\mathbb{P}(A = u_1 | Z) > q, A = u_2\} \\ \mathcal{E}_q^-(u_1, u_2) &= \{\mathbb{P}(A = u_1 | Z) \leq q, A = u_2\}.\end{aligned}$$

Then

$$\begin{aligned}\{A = a\} &= \mathcal{E}_q^+(a, a) \cup \mathcal{E}_q^-(a, a), \\ \{\hat{A} = a\} &= \mathcal{E}_q^+(a, a) \cup \mathcal{E}_q^+(a, b),\end{aligned}$$

where \hat{A} is the estimated protected attribute according to the thresholding rule (Definition 2.3).

It follows that

$$\begin{aligned}\mu(a) &= \frac{\mathbb{E}[\mathbb{I}(A = a)Y]}{\mathbb{P}(A = a)} \\ &= \frac{\mathbb{E}[\mathbb{I}(\mathcal{E}_q^+(a, a))Y] + \mathbb{E}[\mathbb{I}(\mathcal{E}_q^-(a, a))Y]}{\mathbb{P}(\mathcal{E}_q^+(a, a)) + \mathbb{P}(\mathcal{E}_q^-(a, a))} \\ \hat{\mu}_W(a) &\xrightarrow{\text{a.s.}} \frac{\mathbb{E}[\mathbb{I}(\hat{A} = a)Y]}{\mathbb{P}(\hat{A} = a)} \\ &= \frac{\mathbb{E}[\mathbb{I}(\mathcal{E}_q^+(a, a))Y] + \mathbb{E}[\mathbb{I}(\mathcal{E}_q^+(a, b))Y]}{\mathbb{P}(\mathcal{E}_q^+(a, a)) + \mathbb{P}(\mathcal{E}_q^+(a, b))}.\end{aligned}$$

Then simple algebra shows that

$$\begin{aligned}\hat{\mu}_W(a) - \mu(a) &\xrightarrow{\text{a.s.}} \Delta_1(a)C_1(a) - \Delta_2(a)C_2(a) \\ &\quad + (\Delta_1(a) - \Delta_2(a))C_3(a),\end{aligned}$$

where

$$\begin{aligned}\Delta_1(a) &= \mathbb{E}[Y | \mathcal{E}_q^+(a, b)] - \mathbb{E}[Y | \mathcal{E}_q^+(a, a)], \\ \Delta_2(a) &= \mathbb{E}[Y | \mathcal{E}_q^-(a, a)] - \mathbb{E}[Y | \mathcal{E}_q^+(a, a)],\end{aligned}$$

and

$$\begin{aligned}C_1(a) &= \mathbb{P}(\hat{A} = a | A = a)\mathbb{P}(A = b | \hat{A} = a), \\ C_2(a) &= \mathbb{P}(A = a | \hat{A} = a)\mathbb{P}(\hat{A} \neq a | A = a), \\ C_3(a) &= \mathbb{P}(\hat{A} \neq a | A = a)\mathbb{P}(A = b | \hat{A} = a).\end{aligned}$$

Similarly, we can prove that

$$\begin{aligned}\hat{\mu}_W(b) - \mu(b) &\xrightarrow{\text{a.s.}} \Delta_1(b)C_1(b) - \Delta_2(b)C_2(b) \\ &\quad + (\Delta_1(b) - \Delta_2(b))C_3(b),\end{aligned}$$

where $\Delta_1(b), \Delta_2(b), C_1(b), C_2(b), C_3(b)$ are defined analogously.

In summary, as $N \rightarrow \infty$, for $u \in \{a, b\}$ and u^c as the class opposite to u ,

$$\begin{aligned}\hat{\mu}_q(u) - \mu(u) &\rightarrow \Delta_1(u)C_1(u) - \Delta_2(u)C_2(u) \\ &\quad + (\Delta_1(u) - \Delta_2(u))C_3(u),\end{aligned}$$

where $\Delta_1(u)$ and $\Delta_2(u)$ are defined in Section 3.2, and

$$\begin{aligned}C_1(u) &= \mathbb{P}(\hat{A} = u | A = u)\mathbb{P}(A = u^c | \hat{A} = u), \\ C_2(u) &= \mathbb{P}(A = u | \hat{A} = u)\mathbb{P}(\hat{A} \neq u | A = u), \\ C_3(u) &= \mathbb{P}(\hat{A} \neq u | A = u)\mathbb{P}(A = u^c | \hat{A} = u).\end{aligned}$$

\square

PROOF OF COROLLARY 3.4. The conclusions follow immediately from Theorem 3.3. \square

PROOF OF COROLLARY 3.5. First, (i) is obvious according to the formulas of $C_2(u), C_3(u)$ in Theorem 3.3.

Second, note that

$$\begin{aligned}C_1(u) &= \mathbb{P}(\hat{A} = u | A = u)(1 - \mathbb{P}(A = u | \hat{A} = u)), \\ C_2(u) &= \mathbb{P}(A = u | \hat{A} = u)(1 - \mathbb{P}(\hat{A} = u | A = u)).\end{aligned}$$

Thus

$$\begin{aligned} C_2(u) - C_1(u) &= \mathbb{P}(A = u \mid \hat{A} = u) - \mathbb{P}(\hat{A} = u \mid A = u) \\ &= \mathbb{P}(A = u, \hat{A} = u) \left[\frac{1}{\mathbb{P}(\hat{A} = u)} - \frac{1}{\mathbb{P}(A = u)} \right] \\ &= \frac{\mathbb{P}(A = u, \hat{A} = u)}{\mathbb{P}(\hat{A} = u)\mathbb{P}(A = u)} (\mathbb{P}(A = u) - \mathbb{P}(\hat{A} = u)). \end{aligned}$$

Therefore, $\mathbb{P}(\hat{A} = u) < \mathbb{P}(A = u)$ if and only if

$$C_2(u) > C_1(u).$$

□

Appendix B MULTILEVEL UNKNOWN PROTECTED CLASS

In this section, we suppose that the protected class A can have more than two values and the value space of A is denoted as \mathcal{A} . Denote $\mathcal{A}' = \mathcal{A} \setminus \{a, b\}$, i.e., \mathcal{A}' is the set of all values of A other than a and b .

COROLLARY B.1. *The quantities $C_1(u), C_2(u), C_3(u)$ for $u \in \{a, b\}$ in Theorem 3.3 are related in the following way:*

- (i) *If $\mathbb{P}(A = u \mid \hat{A} = u) > \mathbb{P}(A = u^c \mid \hat{A} = u)$, then $C_2(u) > C_3(u)$*
- (ii) *If $\mathbb{P}(A = u) + \mathbb{P}(A \in \mathcal{A}', \hat{A} = u) > \mathbb{P}(\hat{A} = u)$, then $C_2(u) > C_1(u)$.*

Thus if the conditions in (i) and (ii) both hold, then $C_2(u) > C_1(u)$ and $C_2(u) > C_3(u)$.

This means in the multiclass case, $C_2(u) > C_1(u)$ can hold more easily: it can hold even if $\mathbb{P}(A = u) < \mathbb{P}(\hat{A} = u)$, which explains why in figure 5 $C_2(u) > C_1(u)$ even though $\mathbb{P}(A = u) < \mathbb{P}(\hat{A} = u)$ where u stands for White.

Appendix C ADDITIONAL RESULTS ON THE MORTGAGE DATASET

C.1 Bias According to Theorem 3.3

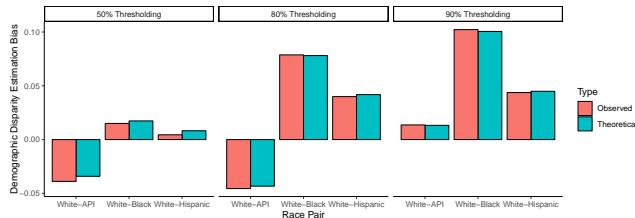


Figure 7: The observed bias is the actual estimation bias of the thresholded estimators with different thresholds, as shown in the right subfigure of Figure 3. The theoretical bias is computed according to Theorem 3.3. This figure shows that the theoretical bias formula in Theorem 3.3 for binary protected class approximates the observed bias for multiclass protected class very well. Therefore, Theorem 3.3 indeed captures the main bias sources of thresholded estimators for both binary and multiclass protected class.

C.2 Results for API

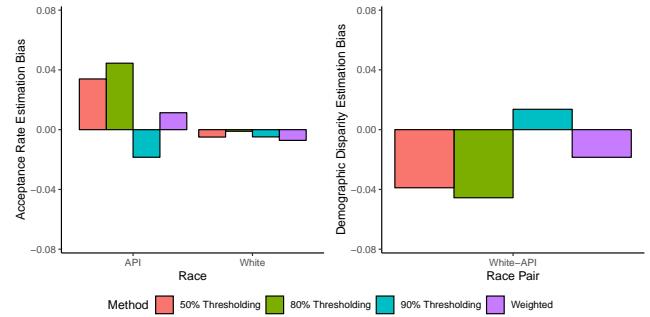


Figure 8: Left figure: the acceptance rate estimation bias for the API race. Right figure: demographic disparity estimation bias with respect to White and API.

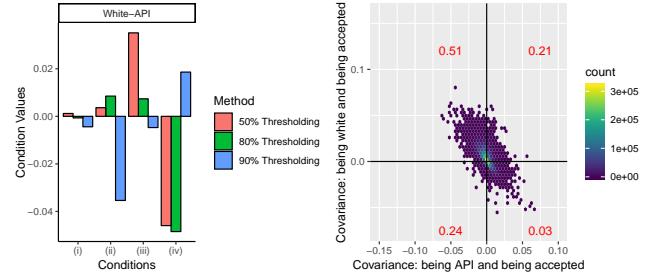


Figure 9: Left figure: the left hand side quantities in conditions 5i–5iv in Corollary 3.4 when using thresholded estimator to estimate the demographic disparity with respect to White and API. Right figure: the population distribution across census tracts with different combinations of covariance terms in Theorem 3.1 with White as the advantaged group and API as the disadvantaged group.

In Figure 8, we can observe that the estimation bias is quite different when estimating the demographic disparity for White and API. The thresholded estimator could overestimate or underestimate the demographic disparity while the weighted estimator only very slightly underestimate the demographic disparity. This is not totally surprising because the API population have very different geolocation distribution and socioeconomic status distribution than Hispanic and Black. API population tend to mix with other races in terms of living locations and the average socioeconomic status disparity between API and White is much smaller than the average socioeconomic status disparity between other minority groups and White. Figure 9 shows that the proportion of people who live in census tracts where API are accepted at lower average rate while White are accepted at higher average rate is much smaller than the proportion of people who live in census tracts where Black or Hispanic are accepted at lower average rate while White are accepted at higher average rate. This explains why the underestimation bias of the weighted estimator is very small.

C.3 Correlation between the race probability and socioeconomic status

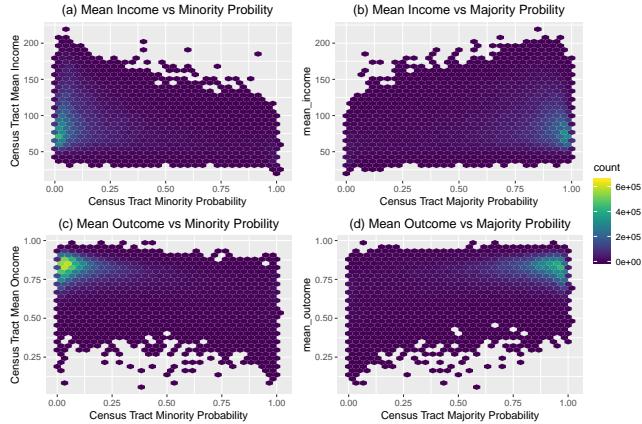


Figure 10: Figure (a) and (b) show the population distribution across census tracts with different mean yearly income versus different probability of belonging to the majority group and different probability of belonging to the minority groups. Figure (c) and (d) show the population distribution across census tracts with different average loan acceptance rate versus different probability of belonging to the majority group and different probability of belonging to the minority groups. The majority group here refers to White and the minority groups here refer to Black or Hispanic.

Figure 10 shows the correlation between the race proxy probabilities and yearly income or average loan acceptance rate. We can clearly observe that census tracts with high White probability overall have more mass on higher income and higher loan acceptance rate while census tracts with high Hispanic or Black probability have more mass on lower income and lower loan acceptance rate. This figure validates the fact that the geolocation encodes socioeconomic status disparities such that the race proxy probabilities are correlated with socioeconomic status variables (e.g., income in this example) and the loan approval outcome. These correlations account for the conditions (5iii)–(5iv) in the Corollary 3.4.

C.4 An example of unbiased weighted estimator

To provide an example where the weighted estimator can be unbiased, we construct a semi-synthetic dataset based on the mortgage dataset. We aggregate the mortgage applicants' yearly income into deciles and use this discrete income as the predictor for race. We simulate the loan approval outcome Y with $\mathbb{P}(Y = 1 \mid Z = z) = 1/(1 + \exp(-(z - 5.5)))$, i.e., the loan approval outcome only depends on the discrete income Z . By construction, the loan approval outcome Y is independent with race A conditionally on Z , thus according to Theorem 3.1 the weighted estimator should be unbiased. This theoretical result is verified in Figure 11. This example shows that the weighted estimator with well constructed proxy model can estimate the demographic disparity perfectly.

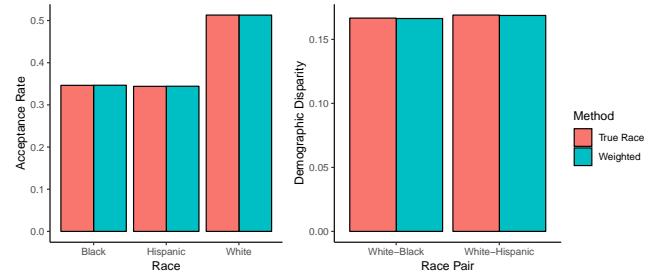


Figure 11: The loan acceptance rate and demographic disparity estimated by using the true race and by the weighted estimator for the semi-synthetic data where income is used to estimate race and it also determines the loan outcome. Weighted estimator is unbiased in this case because race is independent with the outcome conditionally on income by construction.