Testing Group Fairness via Optimal Transport Projections

Nian Si[1] **Karthyek Murthy** [2] **Jose Blanchet**[1] **Viet Anh Nguyen**[1 3]

Abstract

We present a statistical testing framework to detect if a given machine learning classifier fails to satisfy a wide range of group fairness notions. The proposed test is a flexible, interpretable, and statistically rigorous tool for auditing whether exhibited biases are intrinsic to the algorithm or due to the randomness in the data. The statistical challenges, which may arise from multiple impact criteria that define group fairness and which are discontinuous on model parameters, are conveniently tackled by projecting the empirical measure onto the set of group-fair probability models using optimal transport. This statistic is efficiently computed using linear programming and its asymptotic distribution is explicitly obtained. The proposed framework can also be used to test for testing composite fairness hypotheses and fairness with multiple sensitive attributes. The optimal transport testing formulation improves interpretability by characterizing the minimal covariate perturbations that eliminate the bias observed in the audit.

1. Introduction

Algorithmic decisions are commonly conceived to have the potential of being more objective than a human's decisions, since they are generated by logical instructions and the rules of algebra. However, recent studies indicate that this may not be the case. For example, an algorithm which helps the US criminal justice system to predict recidivism rates has been shown to falsely give a higher risk for African-Americans than white Americans (Chouldechova, 2017; MultiMedia LLC, 2016). Similar biases are exhibited against female candidates in a hiring-help system developed by Amazon AI (Dastin, 2018) and an ad-targeting algorithm used by Google (Datta et al., 2015).

[1]Department of Management Science & Engineering, Stanford University [2]Engineering Systems and Design pillar, Singapore University of Technology and Design [3]VinAI Research, Vietnam. Correspondence to: Nian Si <niansi@stanford.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

A natural first explanation for the reported algorithmic biases is that the data used to train the algorithms may already be corrupted by human biases (Buolamwini & Gebru, 2018; Manrai et al., 2016). Deeper inquests have revealed insights on how common learning procedures intrinsically perpetuate the biases and potentially introduce fresh ones. The usual practice of training by minimizing empirical risk, while geared towards yielding predictions that are best when averaged over the entire population, often under-represents minority subgroups in the datasets. Moreover, even though certain sensitive attributes are forbidden by law to be used in the algorithm, the strong correlations between the sensitive attributes and other features potentially lead to biases in predictions. As reported in the studies in (Grgic-Hlaca et al., 2016; Garg et al., 2019; Barocas & Selbst, 2016; Black et al., 2020; Kleinberg et al., 2018; Lipton et al., 2018), merely masking the sensitive attributes does not address the problem.

The aforementioned biases and their impacts have sparked substantial interests in the pursuit of algorithmic fairness (Berk et al., 2018; Chouldechova & Roth, 2020; Corbett-Davies et al., 2017; Mehrabi et al., 2019). Testing whether a given machine learning algorithm is fair emerges as a question of first-order importance. In turn, designing this test for a wide range of group fairness notions (discussed in the sequel) is the main task of this paper.

Our proposed statistical hypothesis testing framework (testing framework for short) allows the auditors to systematically determine whether the biases exhibited in the audit procedure, if any, are intrinsic to the algorithm or due to the randomness in data. Moreover, our framework can be implemented as a black-box, without knowing the exact structure of the classification algorithm used.

For settings where sensitive attributes are not explicitly used as input to classification, fairness is measured on impact either at a group level or at an individual level (Barocas & Selbst, 2016). Group fairness notions seek to measure the differences in impacts across different groups and constitute the prominent means of assessing discrimination associated with group memberships. Individual fairness, on the other hand, seeks to assess if similar users are treated similarly (Dwork et al., 2012; John et al., 2020; Xue et al., 2020). Common examples of group fairness

notions include disparate impact (Zafar et al., 2017), demographic parity (statistical parity) (Calders & Verwer, 2010), equality of opportunity (Hardt et al., 2016), equalized odds (Hardt et al., 2016), etc. The specific choice is usually driven by the philosophical, sociological or legal constraints binding the application considered.

Our testing framework applies to a generic notion of group fairness which encompasses all of the above specific group fairness notions as examples. This unifying approach can also be used in contexts requiring the use of different fairness notions simultaneously and settings with multiple groups.

Since a single fairness criterion among two groups can be reduced to testing the equality in two sample conditional means, one may consider employing a Welch's t-test or a permutation test. Further, a suitable adjustment of the randomness in sample sizes, as in DiCiccio et al. (2020); Tramer et al. (2017) can be applied. Some other existing methods such as Besse et al. (2018) also only apply to one-dimensional criterion. Extensions to multiple impact criteria are not immediate, as is the equalized odds case criterion (Hardt et al., 2016), or in the presence of multiple groups. Algorithmic approaches, such as in (Saleiro et al., 2018), lack the control of the type-I (false positive error). In contrast, the framework proposed here is applicable under general multiple impact criteria and controls the type-I error exactly.

The statistical challenges, which may arise from the presence of multiple impact criteria, are conveniently handled in our framework by utilizing the machinery of optimal transport projections. This involves computing the test statistic by projecting, or in other words, optimally transporting the empirical measure to the set of probability models which satisfy the group fairness notion (or notions) considered. This gives a measure of plausibility of the classifier in satisfying the fairness criterion under the data-generating distribution and the fairness hypothesis is duly rejected if the test statistic exceeds a suitable threshold determined by the significance level. This threshold is determined from the limiting distribution of the test statistic obtained as one of the main results of this paper.

Performing statistical inference with a projection criterion is prevalent in statistics: Owen (2001) serves as a comprehensive reference for projections, or profile functions, that are computed based on likelihood ratio metrics or the Kullback-Liebler divergence. Blanchet et al. (2019) and Cisneros-Velarde et al. (2020) study statistical inference with optimal transport projections. Recently, optimal transport divergences (Villani, 2008) become an attractive tool in many recent machine learning studies, including missing data imputation, geodesic PCA, point embeddings, and repairing data with Wasserstein barycenters for training fair classifiers (Silvia et al., 2020; Gordaliza et al., 2019; Zehlike et al., 2020).

Taskesen et al. (2021) uses optimal transport projections to test a smooth relaxation of the equal opportunity criterion called probabilistic fairness; see Pleiss et al. (2017). This relaxation is required in Taskesen et al. (2021) to overcome the discontinuities in the classification boundaries which create technical complications when computing the optimal transport projections. Further, the resulting test statistic involves a non-convex optimization problem which is difficult to compute. In contrast, our work resolves the technical challenges arising from the discontinuities in classification boundaries. Moreover, our test statistic is the optimal value of a linear program, whose optimal solution offers interpretability by characterizing the minimal covariate perturbations that eliminate the bias observed in the audit. We emphasize that addressing boundary discontinuities is not simply a technical improvement. As we discuss in Section 3.3, our results show different qualitative behaviors both in the scaling and the interpretation of the optimal transport projection in terms of group fairness using optimal transport projections. In addition to enabling exact, computationally tractable, and interpretable fairness assessment for general deterministic classifiers, the technical analysis serves as a stepping stone for statistical inference in estimation tasks such as quantile regression which involve discontinuous estimating equations.

The main contributions are summarized as follows:

(1) We develop a statistical hypothesis test for assessing group fairness as per a generic notion that includes commonly used fairness criteria as special cases. The test is computationally tractable and interpretable. Besides being applicable to settings involving multiple groups, our framework is also applicable to any classifier algorithms, including but not limited to the logistic regression, SVMs, kernel methods, and nearest neighbors.

(2) We develop an extension of the statistical test for the testing problem with composite null hypothesis, addressed here as ϵ-fairness.

(3) The framework facilitates the exact use of fairness criterion, thus obviating the need to invoke relaxations in the absence of smoothness in the impact criteria defining group fairness. The resulting qualitative difference, in terms of the rate of convergence for resulting optimal transport projections, is previously unreported and could be of interest from the technical standpoint of analysing profile functions with discontinuous score functions.

The remainder of the paper is structured as follows. In Section 2, we introduce a generic notion of group fairness and discuss the theory of optimal transport. Section 3 details the proposed statistical test for the simple fairness null hy-

pothesis. Section 4 extends our approach to composite hypotheses. Section 5 discusses computational methods associated with the test. Numerical experiments presented in Section 6 serve to demonstrate the efficacy of the test. All technical proofs are relegated to Appendix A.

Notations. We use $\|\cdot\|_*$ to denote the dual norm of $\|\cdot\|$. We denote $(x)^- = \min\{x, 0\}$ and $(x)^+ = \max\{x, 0\}$. We use $\Rightarrow$, $\xrightarrow{p}$ and $\xrightarrow{a.s}$ to denote convergence in distribution, in probability and convergence almost surely, respectively. The support of the distribution of X is represented by $\text{supp}(X)$. We use $[n]$ to denote the set $\{1, 2, \ldots, n\}$ and $\mathbb{R}^m_+$ to denote the positive orthant $\{x \in \mathbb{R}^m_+ : x \geq 0\}$. $\delta_{(x,a,y)}$ denotes a Dirac measure on a fixed point (x, a, y).

2. Problem Setup and Preliminaries

Throughout this paper we consider the classification settings in which the deterministic classifier $\mathcal{C} : \mathcal{X} \to \mathcal{Y}$ maps the input features from $\mathcal{X} \subset \mathbb{R}^d$ to output labels in the set $\mathcal{Y} = \{0, 1\}$. Evaluation of fairness is considered with respect to a sensitive attribute A taking values in a finite set $\mathcal{A}$. For simplicity, we consider $\mathcal{A} = \{0, 1\}$ where $A = 1$ can be taken to identify the reference group. The statistical test developed in this paper and the main results are applicable more generally to settings involving a non-binary sensitive attribute (or) multiple sensitive attributes. Most notions of group fairness are stated in terms of the joint distribution $\mathbb{Q}$ of (X, A, Y), where X is the vector of input features, A is the sensitive attribute, and Y is the class label of a random sample from the population.

2.1. Notions of Group Fairness

A general reference to fairness notions can be found in Makhlouf et al. (2020, Table 14). The statistical notion of group fairness that we consider, encapsulated in Definition 1 below, is stated flexibly to include commonly used notions such as equal opportunity (Hardt et al., 2016), predictive equality (Corbett-Davies et al., 2017), equalized odds (Hardt et al., 2016), and statistical parity (Dwork et al., 2012), etc., as special cases. This flexibility is achieved by stating the definition in terms of a tuple (U, ϕ), where U is an $\mathbb{R}^s$-valued random vector completely dependent on (A, Y) and $\phi : \mathbb{R}^s \times \mathbb{R}^s \to \mathbb{R}^m$ is a function chosen to discern the differences in performance of the classifier across groups. We address (U, ϕ) as the *discerning tuple*.

Definition 1 (Generic notion of group fairness)**.** *A classifier $\mathcal{C} : \mathcal{X} \to \{0, 1\}$ is fair with respect to the discerning tuple (U, ϕ) under a probability distribution $\mathbb{Q}$ if*

$$\mathbb{E}_{\mathbb{Q}}[\mathcal{C}(X)\phi(U, \mathbb{E}_{\mathbb{Q}}[U])] = 0. \tag{1}$$

Note that $\mathcal{C}(X) = \mathbb{I}\{\mathcal{C}(X) = 1\}$, and thus equation (1) can be seen as $\mathbb{E}_{\mathbb{Q}}[\mathbb{I}\{\mathcal{C}(X) = 1\}\phi(U, \mathbb{E}_{\mathbb{Q}}[U])] = 0$. At the first glance, equation (1) seems asymmetric as it only considers the positive prediction label $\mathcal{C}(X) = 1$. However, it is easy to check that $\mathbb{E}_{\mathbb{Q}}[\phi(U, \mathbb{E}_{\mathbb{Q}}[U])] = 0$ in all the group fairness notions in Examples 1 - 6. By taking the difference, we get the symmetric guarantee that $\mathbb{E}_{\mathbb{Q}}[\mathbb{I}\{\mathcal{C}(X) = 0\}\phi(U, \mathbb{E}_{\mathbb{Q}}[U])] = 0$.

Various useful notions of fairness can be obtained by varying the choice of (U, ϕ) as illustrated in Examples 1 - 6 below. We take $\mathcal{A} = \{0, 1\}$ in Examples 1 - 4.

Example 1 (Equal opportunity (Hardt et al., 2016))**.** *A classifier $\mathcal{C} : \mathcal{X} \to \{0, 1\}$ satisfies the equal opportunity criterion relative to a distribution $\mathbb{Q}$ if*

$$\mathbb{Q}(\mathcal{C}(X) = 1 | A = 1, Y = 1) - \mathbb{Q}(\mathcal{C}(X) = 1 | A = 0, Y = 1) = 0.$$

This criterion coincides with condition (1) *with the choice $U = (\mathbb{I}_{(1,1)}(A, Y); \mathbb{I}_{(0,1)}(A, Y))$, where $\mathbb{I}_{(a,y)}(A, Y)$ denotes the indicator $\mathbb{I}(A = a, Y = y)$, and*

$$\phi : (U, \mathbb{E}_{\mathbb{Q}}[U]) \mapsto \frac{U_1}{\mathbb{E}_{\mathbb{Q}}[U_1]} - \frac{U_2}{\mathbb{E}_{\mathbb{Q}}[U_2]}. \tag{2}$$

Example 2 (Predictive equality (Corbett-Davies et al., 2017))**.** *A classifier $\mathcal{C} : \mathcal{X} \to \{0, 1\}$ satisfies the predictive equality criterion relative to a distribution $\mathbb{Q}$ if*

$$\mathbb{Q}(\mathcal{C}(X) = 1 | A = 0, Y = 0) - \mathbb{Q}(\mathcal{C}(X) = 1 | A = 1, Y = 0) = 0.$$

This criterion coincides with condition (1) *with the choice $U = [\mathbb{I}_{(1,0)}(A, Y); \mathbb{I}_{(0,0)}(A, Y)]$ and ϕ takes the same form as the function* (2)*.*

Example 3 (Equalized odds (Hardt et al., 2016))**.** *A classifier $\mathcal{C} : \mathcal{X} \to \{0, 1\}$ satisfies the equalized odds criterion relative to a distribution $\mathbb{Q}$ if it satisfies both equal opportunity and predictive equality criteria. This criterion coincides with condition* (1) *with $U = [\mathbb{I}_{(1,1)}(A, Y); \mathbb{I}_{(0,1)}(A, Y); \mathbb{I}_{(1,0)}(A, Y); \mathbb{I}_{(0,0)}(A, Y)]$ and*

$$\phi : (U, \mathbb{E}_{\mathbb{Q}}[U]) \mapsto \Big[\frac{U_1}{\mathbb{E}_{\mathbb{Q}}[U_1]} - \frac{U_2}{\mathbb{E}_{\mathbb{Q}}[U_2]}; \frac{U_3}{\mathbb{E}_{\mathbb{Q}}[U_3]} - \frac{U_4}{\mathbb{E}_{\mathbb{Q}}[U_4]}\Big].$$

Example 4 (Statistical parity (Dwork et al., 2012))**.** *A classifier $\mathcal{C} : \mathcal{X} \to \{0, 1\}$ satisfies the statistical parity criterion relative to a distribution $\mathbb{Q}$ if*

$$\mathbb{Q}(\mathcal{C}(X) = 1 | A = 1) - \mathbb{Q}(\mathcal{C}(X) = 1 | A = 0) = 0.$$

This criterion coincides with condition (1) *with the choice $U = [\mathbb{I}_1(A); \mathbb{I}_0(A)]$ and ϕ takes the same form as the function* (2)*.*

If the sensitive attribute takes multiple values or there are multiple sensitive attributes, we can still define the associated fairness notions, which correspond to different choices of (U, ϕ).

Example 5 (Equal opportunity with a non-binary sensitive attribute)**.** *Let $\mathcal{A} = \{0, 1, 2, \ldots, k\}$. A classifier $\mathcal{C} : \mathcal{X} \to \{0, 1\}$ satisfies the equal opportunity criterion relative to a probability measure $\mathbb{Q}$ if*

$$\mathbb{Q}\left(\mathcal{C}(X) = 1 | A = t, Y = 1\right) - \mathbb{Q}\left(\mathcal{C}(X) = 1 | A = 0, Y = 1\right) = 0 \quad \forall t \in \mathcal{A}\backslash\{0\}.$$

This criterion coincides with condition (1) *with the choice $U = (\mathbb{I}_{(0,1)}(A, Y); \mathbb{I}_{(1,1)}(A, Y); \ldots \mathbb{I}_{(k,1)}(A, Y))$ and $\phi = (\phi_1, \phi_2, \ldots, \phi_t)$ with*

$$\phi_t : (U, \mathbb{E}_{\mathbb{Q}}[U]) \mapsto \frac{U_t}{\mathbb{E}_{\mathbb{Q}}[U_t]} - \frac{U_1}{\mathbb{E}_{\mathbb{Q}}[U_1]} \quad \forall t \in \mathcal{A}\backslash\{0\}.$$

Example 6 (Equal opportunity with multiple sensitive attributes)**.** *Suppose we have K sensitive attributes, $A_1, A_2, \ldots, A_K$, all taking values in a superset $\mathcal{A}$. A classifier $\mathcal{C} : \mathcal{X} \to \{0, 1\}$ satisfies the equal opportunity criterion relative to a probability measure $\mathbb{Q}$ if*

$$\mathbb{Q}\left(\mathcal{C}(X) = 1 | A_t = 1, Y = 1\right) - \mathbb{Q}\left(\mathcal{C}(X) = 1 | A_t = 0, Y = 1\right) = 0 \quad \forall t \in [K].$$

This criterion coincides with condition (1) *with the choice*

$$U_t = \mathbb{I}_{(1,1)}(A_t, Y) \text{ and } U_{t+K} = \mathbb{I}_{(0,1)}(A_t, Y) \quad \forall t \in [K]$$

and $\phi = (\phi_1, \phi_2, \ldots, \phi_t)$ with

$$\phi_t : (U, \mathbb{E}_{\mathbb{Q}}[U]) \mapsto \frac{U_t}{\mathbb{E}_{\mathbb{Q}}[U_t]} - \frac{U_{t+K}}{\mathbb{E}_{\mathbb{Q}}[U_{t+K}]} \quad \forall t \in [K].$$

2.2. Optimal Transport and the Wasserstein Distance

We next introduce the notion of optimal transport costs, of which Wasserstein distances is a special case. Let $P(\mathcal{Z})$ denote the set of all probability distributions on $\mathcal{Z} \triangleq \mathcal{X} \times \mathcal{A} \times \mathcal{Y}$.

Definition 2 (Optimal transport costs, Wasserstein distances)**.** *Given a lower semicontinuous function $c : \mathcal{Z} \times \mathcal{Z} \to [0, \infty]$, the optimal transportation cost $W_c(\mathbb{Q}_1, \mathbb{Q}_2)$ between any two distributions $\mathbb{Q}_1, \mathbb{Q}_2 \in P(\mathcal{Z})$ is given by,*

$$W_c(\mathbb{Q}_1, \mathbb{Q}_2) = \min_{\pi \in \Pi(\mathbb{Q}_1, \mathbb{Q}_2)} \mathbb{E}_\pi \left[c\left(Z, Z'\right)\right],$$

where $\Pi(\mathbb{Q}_1, \mathbb{Q}_2)$ is the set of all joint distributions of (Z, Z') such that the law of $Z = (X, A, Y)$ is $\mathbb{Q}_1$ and that of $Z' = (X', A', Y')$ is $\mathbb{Q}_2$.

If $c(\cdot, \cdot)$ is a metric on $\mathcal{Z}$, then $W_c(\cdot)$ is the type-1 Wasserstein distance; see Villani (2008, Chapter 6). The quantity $W_c(\mathbb{Q}_1, \mathbb{Q}_2)$ can be interpreted as the least transportation cost incurred in transporting mass from $\mathbb{Q}_1$ to $\mathbb{Q}_2$, when the cost of transporting unit mass from location $z \in \mathcal{Z}$ to location $z' \in \mathcal{Z}$ is given by $c(z, z')$.

Throughout the paper, we assume that the function c is decomposable as

$$c\left((x, a, y), (x', a', y')\right) = \bar{c}(x, x') + \infty \cdot |a - a'| + \infty \cdot |y - y'|,$$

for some $\bar{c} : \mathcal{X} \times \mathcal{X} \to [0, \infty]$ satisfying (i) $\bar{c}(x, x') = 0$ if and only if $x = x'$ and (ii) $\bar{c}(x, x') = \bar{c}(x', x)$ for all $x, x' \in \mathcal{X}$. In the above expression, we interpret $\infty \times 0 = 0$. Examples of $\bar{c}(\cdot, \cdot)$ that are useful in our context include

$$\bar{c}(x, x') = \|x - x'\|, \tag{3a}$$

and also

$$\bar{c}(x, x') = k(x, x) - 2k(x, x') + k(x', x'), \tag{3b}$$

where $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a suitable reproducing kernel. Another useful example of $\bar{c}(\cdot)$ is specified in terms of the discrete metric suitable for use in the presence of discrete categorical features: Suppose that the feature vector $X = (X_D, X_C)$, with X_D denoting the set of discrete features taking values in a countable set $\mathbb{D} \subset \mathbb{R}^{d_1}$ and X_C denoting the set of continuous features taking values in $\mathbb{R}^{d_2}$. We have $d_1 + d_2 = d$. In this instance, it is feasible to restrict the transportation to elements in $\mathbb{D} \times \mathbb{R}^{d_2}$ by considering

$$\begin{aligned}\bar{c}\left((x_D, x_C), (x'_D, x'_C)\right) &= \|x_C - x'_C\| + \delta \mathbb{I}\{\{x_D, x'_D\} \subset \mathbb{D}, x_D \neq x'_D\} \\ &\quad + \infty \cdot \mathbb{I}\{\{x_D, x'_D\} \nsubseteq \mathbb{D}, x_D \neq x'_D\},\end{aligned} \tag{4}$$

for some $\delta > 0$. Further, we allow the cost function to be dependent on the sensitive attribute. Following the same line, Hui et al. (2021) recently demonstrates a test power gain by tuning properly a sensitive-attribute-dependent transportation cost function.

3. Test For Simple Null Hypothesis via Optimal Transport

Recall that $P(\mathcal{Z})$ denotes the set of all probability distributions on $\mathcal{Z} \triangleq \mathcal{X} \times \mathcal{A} \times \mathcal{Y}$. Let

$$\mathcal{F} = \{\mathbb{Q} \in P(\mathcal{Z}) : \mathbb{E}_{\mathbb{Q}}[\mathcal{C}(X)\phi(U, \mathbb{E}_{\mathbb{Q}}[U])] = 0\}$$

be the collection of distributions under which the classifier $\mathcal{C}(\cdot)$ is fair, as deemed by Definition 1. Given N independent samples $\{x_i, a_i, y_i\}_{i=1}^N$ from a distribution $\mathbb{P}$ of (X, A, Y), we are interested in the statistical test with the hypotheses

$$\mathcal{H}_0 : \mathbb{P} \in \mathcal{F} \quad \text{against} \quad \mathcal{H}_1 : \mathbb{P} \notin \mathcal{F}.$$

With the null hypothesis $\mathcal{H}_0$ being that the classifier $\mathcal{C}(\cdot)$ is fair, our statistical test will detect the failure of $\mathcal{C}(\cdot)$ in meeting the fairness criterion (in Definition 1) under the data generating distribution. To develop a suitable test statistic, let $\hat{\mathbb{P}}^N = N^{-1}\sum_{i=1}^N \delta_{(x_i,a_i,y_i)}$ denote the empirical measure of the samples obtained from a distribution $\mathbb{P} \in P(\mathcal{Z})$. We define the projection of $\hat{\mathbb{P}}^N$ onto $\mathcal{F}$ as

$$\begin{aligned}\mathcal{P}(\hat{\mathbb{P}}^N) &\triangleq \inf_{\mathbb{Q}\in\mathcal{F}} W_c(\mathbb{Q}, \hat{\mathbb{P}}^N) \\ &= \left\{ \begin{array}{ll} \inf & W_c(\mathbb{Q}, \hat{\mathbb{P}}^N) \\ \text{s.t.} & \mathbb{E}_{\mathbb{Q}}[\mathcal{C}(X)\phi(U, \mathbb{E}_{\mathbb{Q}}[U])] = 0. \end{array} \right. \end{aligned} \tag{5}$$

We adopt the statistical hypothesis framework: for a prespecified significance level α,

$$\text{reject } \mathcal{H}_0 \text{ if } s_{_N} > \eta_{1-\alpha},$$

where $s_{_N}$ is a test statistic that depends on the projection distance $\mathcal{P}(\hat{\mathbb{P}}^N)$, and $\eta_{1-\alpha}$ is the $(1-\alpha)\times 100\%$ quantile of a limiting distribution.

3.1. Linear Programming Formulation for Projection

Our aim here is to reformulate the *in*finite dimensional projection formulation (5) as a finite dimensional linear program. For this purpose, let us define $d : \mathcal{X} \to [0, \infty]$ as

$$d(x) \triangleq \inf\left\{\bar{c}(x, x') : x' \in \mathcal{X},\ \mathcal{C}(x') = 1 - \mathcal{C}(x)\right\}, \tag{6}$$

which gives a measure of distance to the region with classifier label different from that at x, and $d(x) = 0$ means that x on the decision boundary. The value $d(x)$ is readily computed for commonly used classifiers such as linear classifiers (as shown in the proof of Lemma 1 below in the supplement) and kernelized classifiers. In the case of a classifier defined in terms of kernels, say as in,

$$\mathcal{C}(x) = \mathbb{I}\left(\sum_{i=1}^n \alpha_i k(x_i, x) + b \geq 0\right),$$

one may use the transportation cost (3b) and $d(x)$ admits a closed-form expression

$$d(x) = \left(\sum_{i=1}^n \alpha_i k(x_i, x) + b\right)^2 / (\alpha^\top \mathbb{K}\alpha),$$

where $\mathbb{K}$ is an $n \times n$ matrix with entries $\mathbb{K}_{i,j} = k(x_i, x_j)$.

Proposition 1 (Primal reformulation)**.** *The projection distance $\mathcal{P}(\hat{\mathbb{P}}^N)$ is equal to the optimal value of a linear program. More specifically, we have*

$$\mathcal{P}(\hat{\mathbb{P}}^N) = \left\{ \begin{array}{ll} \min_p & \dfrac{1}{N}\displaystyle\sum_{i\in[N]} p_i d(x_i) \\ \text{s.t.} & p \in [0,1]^N, \\ & \displaystyle\sum_{i\in[N]} [1 - 2\mathcal{C}(x_i)]\phi(u_i, \mathbb{E}_{\hat{\mathbb{P}}^N}[U])p_i \\ & \qquad = -\displaystyle\sum_{i\in[N]} \mathcal{C}(x_i)\phi(u_i, \mathbb{E}_{\hat{\mathbb{P}}^N}[U]). \end{array} \right. \tag{7}$$

Naturally, one may study the above linear program by considering its dual formulation. Define the following function

$$\mathcal{D}(\hat{\mathbb{P}}^N) \triangleq \max_{\gamma\in\mathbb{R}^m} \left\{ \begin{array}{l} \frac{1}{N}\sum_{i\in[N]} \gamma^\top \phi(u_i, \mathbb{E}_{\hat{\mathbb{P}}^N}[U])\mathcal{C}(x_i) + \\ \left(d(x_i) + [1 - 2\mathcal{C}(x_i)]\,\gamma^\top \phi(u_i, \mathbb{E}_{\hat{\mathbb{P}}^N}[U])\right)^- \end{array} \right\},$$

where recall the notation that $(x)^- = \min\{x, 0\}$. Strong duality of linear programming asserts that $\mathcal{P}(\hat{\mathbb{P}}^N)$ and $\mathcal{D}(\hat{\mathbb{P}}^N)$ are dual to each other.

Proposition 2 (Strong duality)**.** *Strong duality holds, i.e., $\mathcal{P}(\hat{\mathbb{P}}^N) = \mathcal{D}(\hat{\mathbb{P}}^N)$.*

3.2. Asymptotic Behavior of the Projection Distance

The goal of this subsection is to study the limiting behavior of the projection distance $\mathcal{P}(\hat{\mathbb{P}}^N)$ as the sample size N increases. Proposition 2 implies that it is sufficient to examine the asymptotic behavior of $\mathcal{D}(\hat{\mathbb{P}}^N)$. To present the regularity assumptions under which the limiting behavior can be unravelled, we set $\mu = \mathbb{E}_{\mathbb{P}}[U]$ and define Φ as

$$\Phi : \mathcal{X} \to \mathbb{R}, \quad \Phi(X) = (2\mathcal{C}(X) - 1)d(X).$$

Assumption 1 (Continuous density and derivatives)**.** *There exists $\eta > 0$ such that the below conditions are satisfied:*

a) *The probability distribution of $\Phi(X)$ has a positive continuous density $f(\cdot)$ in the interval $(-v, v)$, i.e., $\mathbb{P}(\Phi(X) \in [-v, u)) = \int_{-v}^{u} f(\nu)\mathrm{d}\nu$ for $u \in (-v, v)$.*

b) *For every $u \in$ supp(U), the function $\phi(u, z)$ has a continuous derivative (Jacobian matrix) $\phi_z(u, z)$ in the neighborhood z satisfying $\|z - \mu\|_2 < v$. In addition, $\Sigma_1 \triangleq \mathbb{E}_{\mathbb{P}}[\phi(U, \mu)\phi(U, \mu)^\top \,|\, d(X) = 0] \succ 0$.*

Assumption 2 (Continuous conditional probability)**.** *For the case where supp(U) is a finite set, the conditional probability $\mathbb{P}(U = u \,|\, \Phi(X) = t)$ is continuous around $t = 0$ for every $u \in$ supp(U).*

We are now ready to state the main result of this section concerning the asymptotic behavior of the projection distance.

Theorem 1 (Limit theorem for $\mathcal{D}(\hat{\mathbb{P}}^N)$)**.** *Suppose that $\{X_1, U_1\}, ..., \{X_n, U_n\}$ are independently obtained from the distribution $\mathbb{P}$ and that Assumptions 1 and 2 are satisfied. Then under the null hypothesis $\mathcal{H}_0$,*

$$N \times \mathcal{D}(\hat{\mathbb{P}}^N) \Rightarrow \max_{\gamma \in \mathbb{R}^m} \left\{ \gamma^\top V - \frac{1}{2}\gamma^\top S \gamma \right\} = \frac{1}{2} V^\top S^{-1} V,$$

where $S = f(0)\Sigma_1$, $V \sim \mathcal{N}(0, \Sigma)$, Σ is the covariance matrix of $\phi(U,\mu)\mathcal{C}(X) + \mathbb{E}_{\mathbb{P}}\left[\phi_z(U,\mu)\mathcal{C}(X)\right] U$, and $\Rightarrow$ denotes the convergence in distribution.

The finite cardinality of the outcome space of U in Assumption 2 is not restrictive: Theorem 1 still holds under *in*finite cardinality under an equivalent assumption. Details can be found in Appendix A. For the fairness notions in Examples 1 - 4, we report in Corollary 1 below the specific closed-form limit distributions obtained from Theorem 1.

Corollary 1. *Suppose that $\phi(U,\mu) = \frac{U_1}{\mu_1} - \frac{U_2}{\mu_2}$ with U_1, U_2 satisfying $U_1 U_2 = 0$ (with probability 1), as in Examples 1, 2, and 4. Then we have the limiting distribution,*

$$\begin{aligned} &V^\top S^{-1} V/2 \\ =&\frac{\sigma^2 \chi^2(1)}{2f(0)\left(\mu_2^2 \mathbb{E}_{\mathbb{P}}[U_1 | d(X) = 0] + \mu_1^2 \mathbb{E}_{\mathbb{P}}[U_2 | d(X) = 0]\right)}, \end{aligned}$$

where $\chi^2(1)$ is a chi-squared distribution with one degree of freedom and

$$\begin{aligned} \sigma^2 =& \operatorname{var}\left\{\mathcal{C}(X)\left(\mu_2 U_1 - \mu_1 U_2\right)\right. \\ &\left. + U_2 \mathbb{E}_{\mathbb{P}}\left[U_1 \mathcal{C}(X)\right] - U_1 \mathbb{E}_{\mathbb{P}}\left[U_2 \mathcal{C}(X)\right]\right\}. \end{aligned}$$

For Example 3, we have

$$\begin{aligned} &V^\top S^{-1} V/2 \\ =&\frac{\sigma_1^2 \chi_1(1)^2}{2f(0)\left(\mu_2^2 \mathbb{E}_{\mathbb{P}}[U_1 | d(X) = 0] + \mu_1^2 \mathbb{E}_{\mathbb{P}}[U_2 | d(X) = 0]\right)} \\ +&\frac{\sigma_2^2 \chi_2(1)^2}{2f(0)\left(\mu_4^2 \mathbb{E}_{\mathbb{P}}[U_3 | d(X) = 0] + \mu_3^2 \mathbb{E}_{\mathbb{P}}[U_4 | d(X) = 0]\right)}. \end{aligned}$$

where $\chi_1^2(1)$ and $\chi_2^2(1)$ are two independent chi-squared distributions with one degree of freedom and

$$\begin{aligned} \sigma_1^2 =& \operatorname{var}\left\{\mathcal{C}(X)\left(\mu_2 U_1 - \mu_1 U_2\right)\right. \\ &\left. + U_2 \mathbb{E}_{\mathbb{P}}\left[U_1 \mathcal{C}(X)\right] - U_1 \mathbb{E}_{\mathbb{P}}\left[U_2 \mathcal{C}(X)\right]\right\}, \\ \sigma_2^2 =& \operatorname{var}\left\{\mathcal{C}(X)\left(\mu_4 U_3 - \mu_3 U_4\right)\right. \\ &\left. + U_4 \mathbb{E}_{\mathbb{P}}\left[U_3 \mathcal{C}(X)\right] - U_3 \mathbb{E}_{\mathbb{P}}\left[U_4 \mathcal{C}(X)\right]\right\}. \end{aligned}$$

Assumption 1*a*) is satisfied for a broad class of classification models interesting in practice. Lemma 1 below identifies that Assumption 1*a*) is satisfied even if there are some discrete features.

Lemma 1. *Suppose that $X = (X_D, X_C)$, where X_D takes values in a finite subset $\mathbb{D} \subset \mathbb{R}^{d_1}, d = d_1 + d_2$ and X_C is an $\mathbb{R}^{d_2}$-valued random vector whose conditional distribution $X_C \,|\, X_D = x$ has a positive density in $\mathbb{R}^{d_2}$ for every $x \in \mathbb{D}$. Let the cost function $\bar{c}(\cdot)$ be given by (3a) or (4). Further, if the classifier is written as $\mathcal{C}(X) = \mathbb{I}\{\ell(\theta^\top X) \geq \tau\}$ where $\ell(\cdot)$ is a continuous and increasing function with $\lim_{x \to +\infty} \ell(x) > \tau > \lim_{x \to -\infty} \ell(x)$ and $\theta = (\theta_D, \theta_C)$ with $\theta_C \neq 0$, we have that Assumption 1.*a*) *is satisfied.*

With $\mathcal{P}(\hat{\mathbb{P}}^N) = \mathcal{D}(\hat{\mathbb{P}}^N)$ as in Proposition 2, Theorem 1 reveals that one can use $s_{_N} \triangleq N \times \mathcal{P}(\hat{\mathbb{P}}^N)$ as a test statistic to reject $\mathcal{H}_0$. In particular, for a prespecified significance level $\alpha \in (0,1)$, let $\eta_{1-\alpha}$ denote the $(1-\alpha) \times 100\%$ quantile of the generalized chi-squared distribution given by the law of $V^\top S^{-1} V/2$. Specific computation and estimation procedures required to compute the statistic $s_{_N}$ and the quantile $\eta_{1-\alpha}$ are discussed in Section 5.

3.3. The Structure of the Wasserstein Projection

In this subsection, we characterize the projection measure $\mathbb{Q}$ and provide a heuristic justification for the convergence rate of Theorem 1. Note that the proof of Proposition 1 also leads to an ε-optimizer sequence for (5).

Proposition 3 (ε-optimizer)**.** *Suppose $d(x_i) < +\infty$ for every $i \in [N]$. Let x_i^ε be an ε-optimizer obtained by solving (6) with $x = x_i$, for $i \in [N]$. Let $\{p_i^\star\}_{i=1}^N$ be an optimal solution of the problem* (7)*. Then, the measure*

$$\mathbb{Q}^\varepsilon \triangleq \frac{1}{N} \sum_{i=1}^N (1 - p_i^\star) \delta_{(x_i, a_i, y_i)} + p_i^\star \delta_{(x_i^\varepsilon, a_i, y_i)}$$

is an ε-optimizer of the problem (5)*.*

Proposition 3 indicates in the optimal transportation plan, the transporter moves mass from x_i to x_i^ε when $p_i^\star \neq 0$. Since the optimal solution of a linear programming problem occurs at corner points, most of $p_i^\star$ should be either zero or one. Further, the proof of Theorem 1 shows that the number of non-zero values in $\{p_i^\star\}_{i=1}^N$ is of the order $O_p(N^{1/2})$ and for each non-zero $p_i^\star$, the moving distance $d(x_i)$ is of the order $O_p(N^{-1/2})$ under the null hypothesis. Therefore, $\mathcal{D}(\hat{\mathbb{P}}^N)$ is of the order $O_p(N^{-1})$. This statistical phenomenon is due to the discontinuity of the estimating function $\mathcal{C}(X)\phi(U, \mathbb{E}_{\mathbb{Q}}[U])$ in X, where the transporter is able to move a small amount of probability mass, but the move results in a significant change of the value for the estimating function around the discontinuity region. The $O_p(N^{-1})$ convergence rate is in contrast to the rate in Blanchet et al. (2019), Taskesen et al. (2021) and Cisneros-Velarde et al. (2020), where the estimating function is assumed to be continuous. For the continuous estimating function, it is optimal to move every point $O_p(N^{-1/2})$ distances, which results in a $O_p(N^{-1/2})$ convergence rate.

Therefore, let us emphasize again the key qualitative difference between our contributions and those of Taskesen et al. (2021). A statistical noise gives the empirical appearance of unfairness in two ways: (A) small statistical fluctuations around all data points; (B) a small sub-population with large outcome fluctuations around the decision boundary. Taskesen et al. (2021) studied scenario (A) and our paper studies scenario (B).

4. Test For Composite Null Hypothesis via Optimal Transport

In settings where the notion of exact group fairness becomes restrictive or unattainable, it becomes attractive to test whether the deviation from fairness, if any, from a given group's viewpoint is not more than a prespecified small extent. The question of verifying ϵ-fairness, from an one-sided perspective, can be similarly formulated as follows. Following Section 3, we define

$$\mathcal{F}_\epsilon = \{\mathbb{Q} : \mathbb{E}_{\mathbb{Q}}[\mathcal{C}(X)\phi(U, \mathbb{E}_{\mathbb{Q}}[U])] \leq \epsilon\},$$

where $\epsilon \in \mathbb{R}^m_+$ is a tolerance level prespecified by the fairness auditor. We are interested in the statistical test with the hypotheses

$$\mathcal{H}_0 : \mathbb{P} \in \mathcal{F}_\epsilon \quad \text{against} \quad \mathcal{H}_1 : \mathbb{P} \notin \mathcal{F}_\epsilon. \tag{8}$$

A suitable statistical test for this formulation will serve the purpose of detecting failure of $\mathcal{C}(\cdot)$ in meeting the one-sided fairness condition within an ϵ tolerance. Similar to Problem 5, we define the projection of $\hat{\mathbb{P}}^N$ onto $\mathcal{F}_\epsilon$ as

$$\mathcal{P}_\epsilon(\hat{\mathbb{P}}^N) \triangleq \left\{ \begin{array}{ll} \inf & W_c(\mathbb{Q}, \hat{\mathbb{P}}^N) \\ \text{s.t.} & \mathbb{E}_{\mathbb{Q}}[\mathcal{C}(X)\phi(U, \mathbb{E}_{\mathbb{Q}}[U])] \leq \epsilon. \end{array} \right.$$

4.1. Linear Programming Formulation for Projection

Proposition 4 (Primal reformulation)**.** *The projection distance $\mathcal{P}_\epsilon(\hat{\mathbb{P}}^N)$ is equal to the optimal value of a linear program. More specifically, we have*

$$\mathcal{P}_\epsilon(\hat{\mathbb{P}}^N) \tag{9}$$
$$= \left\{ \begin{array}{ll} \min_p & \frac{1}{N} \sum_{i\in[N]} p_i d(x_i) \\ \text{s.t.} & p \in [0,1]^N \\ & \sum_{i\in[N]} (1 - 2\mathcal{C}(x_i))\phi(u_i, \mathbb{E}_{\hat{\mathbb{P}}^N}[U])p_i \\ & \qquad + \sum_{i\in[N]} \mathcal{C}(x_i)\phi(u_i, \mathbb{E}_{\hat{\mathbb{P}}^N}[U]) \leq N\epsilon. \end{array} \right.$$

As in Section 3, $\mathcal{P}_\epsilon(\hat{\mathbb{P}}^N)$ is amenable to be studied via the respective dual function,

$$\mathcal{D}_\epsilon(\hat{\mathbb{P}}^N) \triangleq$$
$$\max_{\gamma\in\mathbb{R}^m_+} -\gamma^\top\epsilon + \frac{1}{N}\sum_{i\in[N]} \Big\{\gamma^\top\phi(u_i, \mathbb{E}_{\hat{\mathbb{P}}^N}[U])\mathcal{C}(x_i) + \big(d(x_i) + (1 - 2\mathcal{C}(x_i))\,\gamma^\top\phi(u_i, \mathbb{E}_{\hat{\mathbb{P}}^N}[U])\big)^- \Big\}.$$

Proposition 5 (Strong duality)**.** *Strong duality holds, i.e., $\mathcal{P}_\epsilon(\hat{\mathbb{P}}^N) = \mathcal{D}_\epsilon(\hat{\mathbb{P}}^N)$.*

4.2. Asymptotic Behavior of the Projection Distance

We next study the limit of $\mathcal{D}_\epsilon(\hat{\mathbb{P}}^N)$ as the sample size N tends to infinity. In order to state the theorem, let us introduce notation for asymptotic stochastic ordering. We say that a sequence of random elements $\{A_n\}_{n\geq 1}$ satisfies $A_n \lesssim_D B$ if for every continuous and bounded non-decreasing function g,

$$\limsup_{n\to\infty} \mathbb{E}\left[g(A_n)\right] \leq \mathbb{E}\left[g(B)\right].$$

Theorem 2. *Suppose that $\{X_1, U_1\}, ..., \{X_n, U_n\}$ are independently obtained from the distribution $\mathbb{P}$ and that Assumptions 1 and 2 are satisfied. Then under the null hypothesis $\mathcal{H}_0$,*

$$N \times \mathcal{D}_\epsilon(\hat{\mathbb{P}}^N) \lesssim_D \max_{\gamma\in\mathbb{R}^m_+} \left\{\gamma^\top V - \frac{1}{2}\gamma^\top S\gamma\right\}, \tag{10}$$

where $S = f(0)\Sigma_1$, $V \sim \mathcal{N}(0, \Sigma)$, and Σ is the covariance of $\phi(U,\mu)\mathcal{C}(X) + \mathbb{E}_{\mathbb{P}}\left[\phi_z(U,\mu)\mathcal{C}(X)\right]U$. In particular, if $\phi(U,\mu)$ is one-dimensional ($m = 1$), we have

$$\max_{\gamma\in\mathbb{R}^m_+} \left\{\gamma^\top V - \frac{1}{2}\gamma^\top S\gamma\right\} = \frac{1}{2}S^{-1}V^2\mathbb{I}\{V \geq 0\}.$$

With $\mathcal{P}_\epsilon(\hat{\mathbb{P}}^N) = \mathcal{D}_\epsilon(\hat{\mathbb{P}}^N)$ as in Proposition 5, Theorem 2 reveals that one can use $s_{_N}(\epsilon) \triangleq N \times \mathcal{P}_\epsilon(\hat{\mathbb{P}}^N)$ as a test statistic to reject $\mathcal{H}_0$ and $\eta_{1-\alpha}$, defined by the $(1-\alpha)\times 100\%$ quantile of the right hand side bounding variable in (10), as a threshold. We then follow the same hypothesis testing procedure defined in Section 3. Since Theorem 2 only provides a stochastic upper-bound, we actually use a *conservative* quantile and the type I error is less than or equal to the desired significance level α asymptotically.

5. Computation and Estimation Procedures

5.1. Computations of the Test Statistic

Based on Proposition 1, we propose a sorting-based algorithm for computing $\mathcal{P}(\hat{\mathbb{P}}^N)$ for one-dimensional $\phi(\cdot)$ (that is, $m = 1$). The steps involve transporting points which are close to the decision boundary and have significant contributions towards improving fairness if prediction labels

are flipped. The exact steps are described in Algorithm 1. Note that the algorithm requires only information on $\{\mathcal{C}(x_i), d(x_i) : i \in [N]\}$, instead of the whole functional structure of the classifier $\mathcal{C}$.

Algorithm 1 Computing $\mathcal{P}(\hat{\mathbb{P}}^N)$ for one-dimensional $\phi(\cdot)$

1: **Input:** Data $\{d(x_i), \mathcal{C}(x_i), \phi(u_i, \mathbb{E}_{\hat{\mathbb{P}}^N}[U])\}_{i=1}^N$.
2: **Output:** the optimal value $\mathcal{P}(\hat{\mathbb{P}}^N)$.
3: Let $s \leftarrow -\sum_{i\in[N]} \mathcal{C}(x_i)\phi(u_i, \mathbb{E}_{\hat{\mathbb{P}}^N}[U])$;
4: For $i \in [N]$, compute

$$t_i \leftarrow d(x_i)^{-1}(1 - 2\mathcal{C}(x_i))\phi(u_i, \mathbb{E}_{\hat{\mathbb{P}}^N}[U])\mathrm{sgn}(s);$$

5: Sort $t_1, \ldots, t_N$ in descending order, where $t_{(i)}$ denotes the i-th largest one and let $d_{(i)}$ be the corresponding distance;
6: Initialize $V = 0$ and let $s \leftarrow |s|$;
7: **for** $i \leftarrow 1$ to T **do**
8: **if** $t_{(i)}d_{(i)} < s$ **then**
9: $s \leftarrow s - t_{(i)}d_{(i)}$ and $V \leftarrow V + d_{(i)}$;
10: **else**
11: $V \leftarrow V + t_{(i)}^{-1}s$ and break;
12: **end if**
13: **end for**
14: Output $\mathcal{P}(\hat{\mathbb{P}}^N) \leftarrow V/N$.

With the output $\mathcal{P}(\hat{\mathbb{P}}^N)$ returned by Algorithm 1, computation of the test statistic $s_{_N} = N \times \mathcal{P}(\hat{\mathbb{P}}^N)$ is immediate. To obtain $\mathcal{P}_\epsilon(\hat{\mathbb{P}}^N)$ similarly, one may modify Line 3 in Algorithm 1 as in

$$s \leftarrow -\left(\sum_{i\in[N]} \mathcal{C}(x_i)\phi(u_i, \mathbb{E}_{\hat{\mathbb{P}}^N}[U]) - N\epsilon\right)^+.$$

With $\mathrm{sgn}(0) = 0$ assigning $t_i = 0$ for all $i \in [N]$ in Step 4, we take $0/0 = 0$ in Line 11 in Algorithm 1 in order to obtain $\mathcal{P}_\epsilon(\hat{\mathbb{P}}^N)$. It is easy to see that the time complexity of Algorithm 1 is the same of the time complexity of the sorting algorithm, which is generally $O(N \log N)$. For instances where $m > 1$, one may solve either problem (7) or problem (9) with a standard linear program solver to obtain the respective values $\mathcal{P}(\hat{\mathbb{P}}^N)$ or $\mathcal{P}_\epsilon(\hat{\mathbb{P}}^N)$, which is also solvable in polynomial time. Therefore, our hypothesis test is more computationally efficient than the test proposed in Taskesen et al. (2021), which requires solving a non-convex optimization problem.

5.2. Computations of the Quantile of the Limiting Distributions

We use the conditional density estimator and the Nadaraya-Watson estimator (Tsybakov, 2008, Section 1) to estimate $f(0)$ and Σ_1, i.e.,

$$\hat{f}(0) = \frac{1}{Nh}\sum_{i\in[N]} K\left(\frac{\Phi(x_i)}{h}\right), \text{ and}$$

$$\hat{\Sigma}_1 = \frac{\sum_{i\in[N]} K\left(\frac{\Phi(x_i)}{h}\right)\phi\left(u_i, \mathbb{E}_{\hat{\mathbb{P}}^N}[U]\right)\phi\left(u_i, \mathbb{E}_{\hat{\mathbb{P}}^N}[U]\right)^\top}{\sum_{i\in[N]} K\left(\frac{\Phi(x_i)}{h}\right)},$$

where $h > 0$ is the bandwidth parameter, and $K(\cdot)$ is a kernel function that is symmetric and integrates to one. By combining the above two estimates, an empirical estimate for S, denoted by $\hat{S}$ is computed via,

$$\frac{1}{Nh}\sum_{i\in[N]} K\left(\frac{\Phi(x_i)}{h}\right)\phi\left(u_i, \mathbb{E}_{\hat{\mathbb{P}}^N}[U]\right)\phi\left(u_i, \mathbb{E}_{\hat{\mathbb{P}}^N}[U]\right)^\top.$$

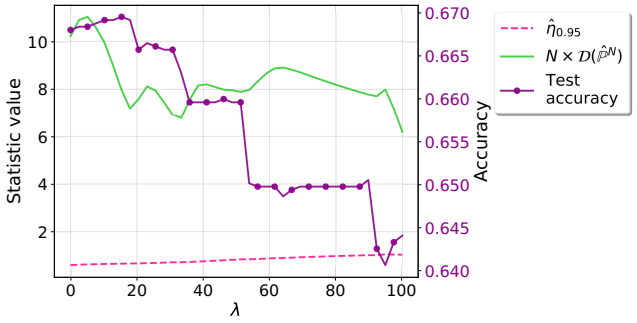
Under some mild conditions, by choosing $h = O(N^{-1/5})$, we have $\|S - \hat{S}\| = O(N^{-2/5})$; see, for example, Härdle (1990, Theorem 4.2.1) and Tsybakov (2008, Proposition 1.7). By combining the empirical covariance estimator for Σ, we obtain a quantile estimate $\hat{\eta}_{1-\alpha}$.

6. Numerical Experiments

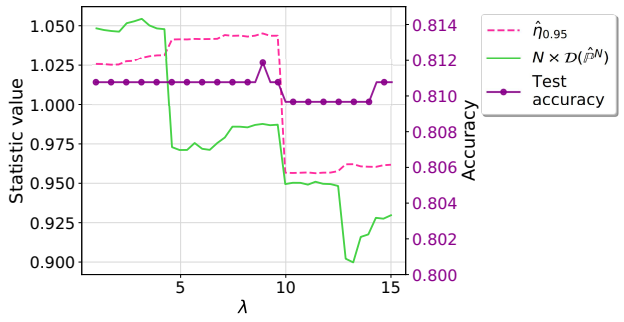
Our experiments use the following three datasets: Arrhythmia (Dua & Graff, 2017), COMPAS (MultiMedia LLC, 2016) and Drug (Fehrman et al., 2017). The details of the datasets are provided in Appendix B.2.

In the first experiment, we test the fairness of the Tikhonov-regularized logistic and SVM classifiers by the equal opportunity criterion. We randomly split 70%-30% of the data as a train-test set. Figure 1 reports the test statistics, fairness rejection threshold, and the accuracy of the classifier. Figure 1(a) shows the result of regularized logistics classifier in COMPAS dataset, while Figure 1(b) shows the result of SVM classifier in the Drug dataset. We observe that a strong regularization only reduces the test statistics very mildly, and the Wasserstein projection tests suggest we reject the fair null hypothesis even when the regularization power is sufficiently large, which presents a different phenomenon from the probabilistic fairness test results shown in Taskesen et al. (2021). We here provide a heuristic explanation for this difference. Consider a logistic classifier $\mathcal{C}(x) = \mathbb{I}\{1/(1 + \exp(-\theta^\top x)) \geq 0.5\}$. Since regularization usually induces shrinkage, the regularized classifier could be approximated by $\mathcal{C}_\varepsilon(x) = \mathbb{I}\{1/(1 + \exp(-\varepsilon\theta^\top x)) \geq 0.5\}$, and large regularization power corresponds to small ε. Note that $\mathcal{C}_\varepsilon(x) = \mathcal{C}(x)$ no matter how small $\varepsilon > 0$ is. However, for the probabilistic notion, $\mathcal{C}_\varepsilon$ will output approximately equal probabilities for both labels, which tends to be probabilistic fair when ε is very

small. The experiment thus demonstrates probabilistic fairness does not imply the exact fairness in general.



(a) Regularized logistics classifier in COMPAS



(b) SVM classifier in Drug dataset

Figure 1. Test statistics and accuracy of regularized classifiers on test data with a rejection threshold. The green line is the test statistics; the pink dashed line is the rejection threshold at the significance level $\alpha = 0.05$; the purple line is the test accuracy; λ denotes the regularization parameter, where larger λ means stronger regularization power.

In the second experiment, we compare a fair algorithm proposed in Donini et al. (2018) with a naive SVM classifier (parametrized by the ridge regularization λ) in three datasets: Arrhythmia, COMPAS and Drug. We randomly split 70%-30% of the data as a train-test set and we replicate this procedure 1,000 times. We will test the fairness in terms of the equal opportunity and equalized odds criteria, and for the equal opportunity criteria, we will further show results using Welch's test (Welch's test is not applicable for multi-dimensional equalized odds criteria). Tables 1 and 2 show a rejection percentage of the naive SVM and the method in Donini et al. (2018) at the significance level $\alpha = 0.05$ in those 1,000 replications using our test according to the equal opportunity and equalized odds criteria, respectively. Table 3 shows the test results using Welch's test according to the equal opportunity criterion. Our test results demonstrate that the method in Donini et al. (2018) has a significantly lower rejection rate, which means it is substantially more fair than the naive method.

Table 1. Rejection percentage of the Naive SVM and the method in Donini et al. (2018) at the significance level $\alpha = 0.05$ according to the equal opportunity criterion using our test.

	Arrhythmia	COMPAS	Drug
Naive SVM	68.4%	100%	30.1%
Donini et al. (2018)	11.6%	16.6%	21.6%

Table 2. Rejection percentage of the Naive SVM and the method in Donini et al. (2018) at the significance level $\alpha = 0.05$ according to the equalized odds criterion using our test.

	Arrhythmia	COMPAS	Drug
Naive SVM	75.1%	100%	30.5%
Donini et al. (2018)	13.7%	21.7%	17.2%

Table 3. Rejection percentage of the Naive SVM and the method in Donini et al. (2018) at the significance level $\alpha = 0.05$ according to the equal opportunity criterion using Welch's test.

	Arrhythmia	COMPAS	Drug
Naive SVM	76.1%	100%	35.5%
Donini et al. (2018)	14.0%	16.1%	23.0%

More experiments are conducted in Appendix B.1 to empirically validate the convergence result in Theorem 1 and our proposed hypothesis test method.

Acknowledgement

Material in this paper is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-20-1-0397. Additional support is gratefully acknowledged from NSF grants 1915967, 1820942, and 1838576 and Singapore Ministry of Education's AcRF grant MOE2019-T2-2-163.

References

Barocas, S. and Selbst, A. D. Big data's disparate impact. *California Law Review*, 104:671–732, 2016.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, pp. 0049124118782533, 2018.

Besse, P., del Barrio, E., Gordaliza, P., and Loubes, J.-M. Confidence intervals for testing disparate impact in fair learning. *arXiv preprint arXiv:1807.06362*, 2018.

Black, E., Yeom, S., and Fredrikson, M. Fliptest: fairness testing via optimal transport. In *Proceedings of*

the 2020 Conference on Fairness, Accountability, and Transparency, pp. 111–121, 2020.

Blanchet, J., Kang, Y., and Murthy, K. Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pp. 77–91, 2018.

Calders, T. and Verwer, S. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017.

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

Cisneros-Velarde, P., Petersen, A., and Oh, S.-Y. Distributionally robust formulation and model selection for the graphical lasso. In *International Conference on Artificial Intelligence and Statistics*, pp. 756–765. PMLR, 2020.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. 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.

Dastin, J. Amazon scraps secret AI recruiting tool that showed bias against women. *San Fransico, CA: Reuters. Retrieved on October*, 9:2018, 2018.

Datta, A., Tschantz, M. C., and Datta, A. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *Proceedings on Privacy Enhancing Technologies*, 2015(1):92–112, 2015.

DiCiccio, C., Vasudevan, S., Basu, K., Kenthapadi, K., and Agarwal, D. Evaluating fairness using permutation tests. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1467–1477, 2020.

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J. S., and Pontil, M. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, pp. 2791–2801, 2018.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

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

Fehrman, E., Muhammad, A. K., Mirkes, E. M., Egan, V., and Gorban, A. N. The five factor model of personality and evaluation of drug consumption risk. In *Data Science*, pp. 231–242. Springer, 2017.

Garg, S., Perot, V., Limtiaco, N., Taly, A., Chi, E. H., and Beutel, A. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 219–226, 2019.

Gordaliza, P., Barrio, E. D., Fabrice, G., and Loubes, J.-M. Obtaining fairness using optimal transport theory. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 2357–2365, 2019.

Grgic-Hlaca, N., Zafar, M. B., Gummadi, K. P., and Weller, A. 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.

Härdle, W. *Applied Nonparametric Regression*. Cambridge University Press, 1990.

Hardt, M., Price, E., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29*, pp. 3315–3323, 2016.

Hui, Y., Xie, J., Blanchet, J., and Glynn, P. Empirical optimal transport projections with non-symmetric costs. *preprint*, 2021.

John, P. G., Vijaykeerthy, D., and Saha, D. Verifying individual fairness in machine learning models. In *Conference on Uncertainty in Artificial Intelligence*, pp. 749–758. PMLR, 2020.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Rambachan, A. Algorithmic fairness. In *AEA Papers and Proceedings*, volume 108, pp. 22–27, 2018.

Lipton, Z., McAuley, J., and Chouldechova, A. Does mitigating ML's impact disparity require treatment disparity? In *Advances in Neural Information Processing Systems*, pp. 8125–8135, 2018.

Makhlouf, K., Zhioua, S., and Palamidessi, C. On the applicability of ML fairness notions. *arXiv preprint arXiv:2006.16745*, 2020.

Manrai, A. K., Funke, B. H., Rehm, H. L., Olesen, M. S., Maron, B. A., Szolovits, P., Margulies, D. M., Loscalzo, J., and Kohane, I. S. Genetic misdiagnoses and the potential for health disparities. *New England Journal of Medicine*, 375(7):655–665, 2016.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

MultiMedia LLC. *Machine Bias*, 2016. Available at https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Owen, A. B. *Empirical Likelihood*. CRC Press, 2001.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pp. 5680–5689, 2017.

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., Rodolfa, K. T., and Ghani, R. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*, 2018.

Silvia, C., Ray, J., Tom, S., Aldo, P., Heinrich, J., and John, A. A general approach to fairness with optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 3633–3640, 2020.

Taskesen, B., Blanchet, J., Kuhn, D., and Nguyen, V. A. A statistical test of probabilistic fairness. *Accepted to ACM Conference on Fairness, Accountability, and Transparency*, 2021.

Tramer, F., Atlidakis, V., Geambasu, R., Hsu, D., Hubaux, J.-P., Humbert, M., Juels, A., and Lin, H. Fairtest: Discovering unwarranted associations in data-driven applications. In *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 401–416. IEEE, 2017.

Tsybakov, A. B. *Introduction to Nonparametric Estimation*. Springer, 2008.

Villani, C. *Optimal Transport: Old and New*, volume 338. Springer, 2008.

Xue, S., Yurochkin, M., and Sun, Y. Auditing ML models for individual bias and unfairness. In *International Conference on Artificial Intelligence and Statistics*, pp. 4552–4562. PMLR, 2020.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. 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.

Zehlike, M., Hacker, P., and Wiedemann, E. Matching code and law: achieving algorithmic fairness with optimal transport. *Data Mining and Knowledge Discovery*, 34(1):163–200, 2020.