Auditing ML Models for Individual Bias and Unfairness

Songkai Xue
Department of Statistics
University of Michigan

Mikhail Yurochkin
IBM Research
MIT-IBM Watson AI lab

Yuekai Sun
Department of Statistics
University of Michigan

Abstract

We consider the task of auditing ML models for individual bias/unfairness. We formalize the task in an optimization problem and develop a suite of inferential tools for the optimal value. Our tools permit us to obtain asymptotic confidence intervals and hypothesis tests that cover the target/control the Type I error rate exactly. To demonstrate the utility of our tools, we use them to reveal the gender and racial biases in Northpointe's COMPAS recidivism prediction instrument.

1 Introduction

Machine learning (ML) models are finding their way into high-stakes decision making tasks such as housing (Angwin and Parris Jr, 2016; Angwin et al., 2017) and recidivism prediction (Angwin et al., 2016). Although replacing humans with ML models eliminates human biases in the decision-making process, the models may perpetuate or even exacerbate biases in their training data. Such biases in ML systems are especially objectionable if they adversely affect minority and/or underprivileged groups of users (Barocas and Selbst, 2016). For example, in 2016 and 2017, ProPublica reported that Facebook allows advertisers to filter users by attributes protected by federal anti-discrimination law (Angwin and Parris Jr, 2016; Angwin et al., 2017). Similar reports eventually prompted state and federal level investigations into Facebook's advertising platform (Tobin, 2019a,b). Other high-profile examples of algorithmic bias/unfairness include racial bias in algorithms for estimating defendants' chances of committing another crime (Angwin et al., 2016), gender biases in resume screening systems for technical positions (Dastin, 2018), and racial bias in image search results (Allen, 2016).

In response, the data science community has proposed many formal definitions of algorithmic fairness and methods to train ML models that abide by the definitions. However, a notable gap in the literature remains: *calibrated* methods for detecting and localizing bias/unfairness in ML models. For example, in the aforementioned investigations of bias/unfairness in ML models, investigator study discrepancies between summary statistics of the output of ML models on subgroups (*e.g.* false positive rates on black and white defendants) (Angwin et al., 2016; Dastin, 2018), but they lack statistical tools to ascertain whether the discrepancies they observe are systemic or due to the inherent randomness in the data. In other words, the investigators lack tools to calibrate the statistics so that the chance of a false alarm is controlled.

In this paper, we address this issue by providing a suite of inferential tools for detecting and localizing bias/unfairness in ML models. The main benefits of the methods are

1. the methods only require *black-box* or query access to the ML model: an auditor only has to observe the output of the ML model;
2. the methods are *computationally efficient*: the main computational expense is solving a convex optimization problem;
3. the methods provide an *interpretable* pairing between inputs that localize the bias/unfairness in an ML system.

The basis of the proposed suite of inferential tools is a result on the asymptotic distribution of the optimal value of a convex optimization problem. Due to the lack of regularity in the value function of the problem, the asymptotic distribution of the optimal value is non-Gaussian. This result may be of independent interest to researchers.

Proceedings of the 23$^{\text{rd}}$International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. Copyright 2020 by the author(s).

1.1 Related work

Generally speaking, there are two kinds of mathematical definitions of algorithmic fairness: group fairness and individual fairness. Most prior work on algorithmic fairness focuses on group fairness because it is suitable for statistical analysis. Despite its prevalence, group fairness suffers from two critical issues. First, it is possible for an ML model that satisfies group fairness to be blatantly unfair from the perspective of individual users (Dwork et al., 2012). Second, there are fundamental incompatibilities between intuitive notions of group fairness (Kleinberg et al., 2016; Chouldechova, 2017).

In light of the issue with group fairness, we focus on individual fairness in this paper. At a high-level, the idea of individual fairness is a fair algorithm ought to treat similar users similarly. This idea is intuitive and has a strong legal basis. Despite its benefits, individual fairness has been dismissed as impractical because there is no consensus on which users are similar. Although this is a critical issue, it is not the focus of this paper, and we assume there is a *similarity function* that determines which users are similar and which users are dissimilar in the rest of the paper. Our tools make no restrictions on the similarity function, so auditors are free to customize the similarity function for their applications. In our computational results, we follow Yurochkin et al. (2020) by adopting a data-driven similarity function.

There is a parallel vein of work in Wasserstein distributionally robust optimization (DRO) (Blanchet and Murthy, 2019; Lee and Raginsky, 2018; Sinha et al., 2017; Blanchet et al., 2019) on obtaining confidence intervals for the population optimal value. The latest in this line of work (Blanchet et al., 2019) also obtains asymptotic distributional results on the distributionally robust optimal value. The key distinction between this line of work and our work is the robustness radius ε is fixed in our work and shrinking (usually at a $\frac{1}{n}$-rate) in the DRO literature. As we shall see, this leads to qualitatively different distributional results: the asymptotic distribution under a fixed radius is generally non-Gaussian, while the distribution under a shrinking radius is Gaussian.

2 The auditor's problem

Imagine an investigator evaluating the fairness of an ML model. The auditor wishes to detect and localize violations of *individual fairness* in the ML models. In this section, we formalize the auditor's task in a convex optimization problem. We start by recalling the definition of individual fairness by Dwork et al. (2012).

Definition 2.1. *An ML model $h : \mathcal{X} \to \mathcal{Y}$ is individually fair if there is $L > 0$ such that*

$$d_y(h(x_1), h(x_2)) \le L d_x(x_1, x_2) \text{ for all } x_1, x_2 \in \mathcal{X},$$

where d_x and d_y are metrics on the input space $\mathcal{X}$ and the output space $\mathcal{Y}$.

The fair metric d_x in Definition 2.1 encodes our intuition of which samples should be treated similarly by the ML model. We emphasize that $d_x(x_1, x_2)$ being small does NOT imply x_1 and x_2 are similar in all respects. Even if $d_x(x_1, x_2)$ is small, x_1 and x_2 may differ in certain attributes that are irrelevant to the ML task at hand, *e.g.*, protected attributes.

At a high-level, we envision the auditor collects a set of audit data and evaluates the performance of the ML model on the audit data and checks for discrepancies between the performance of the model on similar samples. The presence of large discrepancies suggests the ML model violates individual fairness. This type of audit is known as a *correspondence study* in the empirical literature in social sciences; Bertrand and Mullainathan (2004)'s celebrated study of discrimination in the US labor market is a prominent example.

Mathematical preliminaries Denote the input and output space of the ML model by $\mathcal{X}$ and $\mathcal{Y}$ respectively and the sample space by $\mathcal{Z} \triangleq \mathcal{X} \times \mathcal{Y}$. We equip $\mathcal{X}$ with a metric $d_x : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$. This metric is the metric appearing in Definition 2.1; it encodes our intuition of which samples are similar and which are dissimilar. To keep things simple, we assume $\mathcal{Y}$ is a discrete set (*i.e.* the ML model is a classifier). We equip $\mathcal{Z}$ with the metric

$$d_z((x_1, y_1), (x_2, y_2)) \triangleq d_x(x_1, x_2) + \infty \times \mathbf{1}\{y_1 \neq y_2\},$$

The metric d_z encodes our intuition of which samples are similar and which are dissimilar: (x_1, y_1) and (x_2, y_2) similar if and only if (i) they share a label and (ii) x_1 and x_2 are similar according to d_x. Finally, we equip $\Delta(\mathcal{Z})$, the set of probability distributions on $\mathcal{Z}$, with the 1-Wasserstein distance. Recall the Wasserstein distance between two probability distributions P and Q on $\mathcal{Z}$ is

$$W(P, Q) = \inf_{\Pi \in \mathcal{C}(P,Q)} \int_{\mathcal{Z} \times \mathcal{Z}} c\left(z_1, z_2\right) d\Pi\left(z_1, z_2\right),$$

where $c : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_+$ is a transportation cost function and $\mathcal{C}(P, Q)$ is the set of couplings between P and Q. To encode our intuition of fairness in the Wasserstein distance, we use d_z^2 as the transportation cost function. This Wasserstein distance considers two distributions close if the mass they put on comparable segments of the sample space is similar (the placement of mass within comparable segments may differ).

Returning to the auditor's task, let h be the ML model under audit. To detect and localize disparate treatment by the ML model, the auditor picks a loss function $\ell_h : \mathcal{Z} \to \mathbb{R}_+$ to measure the performance of the model and evaluates the risk of the model $\mathbb{E}_{Z\sim P_\star}[\ell_h(Z)]$, where $P_\star$ is the data generating distribution. If there is no bias/unfairness in the ML model, then it is not possible for the auditor to increase the risk by moving (probability) mass to similar areas of the sample space. In other words, if the ML model is fair, then the value of the optimization problem

$$\begin{aligned} \max_{P\in\Delta(\mathcal{Z})} \quad & \mathbb{E}_{Z\sim P}[\ell_h(Z)] - \mathbb{E}_{Z\sim P_\star}[\ell_h(Z)] \\ \text{subject to} \quad & W(P, P_\star) \le \varepsilon, \end{aligned} \tag{2.1}$$

where $\varepsilon \ge 0$ is a transportation budget parameter and should be small. The constraint on the transportation budget compels the auditor to move mass to similar areas of the sample space.

In practice, $P_\star$ is unknown, so the auditor collects a set of audit data $\{(x_i, y_i)\}_{i=1}^n$ and solves the empirical version of (2.1):

$$\begin{aligned} \max_{P\in\Delta(\mathcal{Z})} \quad & \mathbb{E}_{Z\sim P}[\ell_h(Z)] - \mathbb{E}_{Z\sim P_n}[\ell_h(Z)] \\ & W(P, P_n) \le \varepsilon, \end{aligned} \tag{2.2}$$

where P_n is the empirical distribution of the audit data. A large optimal value is evidence that the ML model is unfair. This suggests the optimal value of this optimization problem as a test statistic. We call the optimal value of (2.2) the Fair Transport Hypothesis (FaiTH) test statistic. In summary, if the ML model is fair, then the FaiTH statistic is small.

The FaiTH statistic is robust to small changes in the similarity functions. Let $d_x, d_{x_*} : \mathcal{X}\times\mathcal{X} \to \mathbb{R}_+$ be two different similarity metrics on $\mathcal{X}$. Let $c, c_* : \mathcal{Z}\times\mathcal{Z} \to \mathbb{R}_+$ be the transportation cost functions on $\mathcal{Z}$ induced by d_x, d_{x_*}. Let $W, W_* : \Delta(\mathcal{Z})\times\Delta(\mathcal{Z}) \to \mathbb{R}_+$ be the Wasserstein distances on $\Delta(\mathcal{Z})$ induced by d_x, d_{x_*}. We start by stating the following assumptions:

(A1) the feature space $\mathcal{X}$ is bounded:

$$D \triangleq \max\{\mathsf{diam}(\mathcal{X}), \mathsf{diam}_*(\mathcal{X})\} < \infty;$$

(A2) the loss function is non-negative and bounded: $0 \le \ell_h(z) \le M$ for all $z \in \mathcal{Z}$, and L-Lipschitz with respect to d_x and d_{x_*}:

$$\begin{aligned} &\sup\nolimits_{y:(x_1,y),(x_2,y)\in\mathcal{Z}} |\ell_h(x_1,y) - \ell_h(x_2,y)| \\ &\quad \le L d_x(x_1,x_2) \wedge d_{x_*}(x_1,x_2); \end{aligned}$$

(A3) the discrepancy between the transportation cost functions is uniformly bounded:

$$\sup_{(x_1,y),(x_2,y)\in\mathcal{Z}} \left| \begin{aligned} &c((x_1,y),(x_2,y)) - \\ &c_*((x_1,y),(x_2,y)) \end{aligned} \right| \le \eta D^2.$$

The following proposition shows the robustness of the FaiTH statistic with respect to changes in the similarity functions.

Proposition 2.2. *Under Assumptions A1–A3, the difference between the FaiTH statistics induced by d_x and d_{x_*} satisfies*

$$\left| \begin{aligned} &\max\nolimits_{P:W(P,P_n)\le\varepsilon} \mathbb{E}_{Z\sim P}[\ell_h(Z)] - \\ &\max\nolimits_{P:W_*(P,P_n)\le\varepsilon} \mathbb{E}_{Z\sim P}[\ell_h(Z)] \end{aligned} \right| \le \frac{L\eta D^2}{\sqrt{\varepsilon}}.$$

In the subsequent sections, we develop a suite of inferential tools based on the FaiTH statistic. We emphasize that

1. the auditor only needs to be able to query the output of the ML model to collect the audit data;
2. (2.2) is a linear program, so it is possible to evaluate the FaiTH statistic efficiently.

Inference for the optimal value of an optimization problem (2.2) is generally a hard task, and we focus on finite sample spaces. This simplification is common in the literature on inferential tools for optimal transport problems (Sommerfeld and Munk, 2018; Klatt et al., 2018). As we shall see, the restriction of finite spaces is sufficient for many practical problems, including evaluating the algorithmic fairness of the COMPAS recidivism prediction instrument. For a finite sample space, the auditor's problem is

$$\begin{aligned} \max_{\Pi\in\mathbb{R}_+^{|\mathcal{Z}|\times|\mathcal{Z}|}} \quad & l^\top(\Pi^\top \mathbf{1}_{|\mathcal{Z}|} - f_n) \\ & \langle C, \Pi\rangle \le \varepsilon \\ & \Pi\mathbf{1}_{|\mathcal{Z}|} = f_n, \end{aligned}$$

where $l \in \mathbb{R}_+^{|\mathcal{Z}|}$ is the vector of losses and its i-th entry is $\ell_h(z_i)$, $C \in \mathbb{R}_+^{|\mathcal{Z}|\times|\mathcal{Z}|}$ is the matrix of transportation costs and its (i,j)-th entry is $c(z_i, z_j)$, and $f_n \in \Delta_{|\mathcal{Z}|}$ is the empirical distribution of the data $\{(x_i, y_i)\}_{i=1}^n$.

3 Asymptotic distribution of the FaiTH statistic

In this section, we establish our main result on the asymptotic distribution of the FaiTH statistic. We state the main result and provide a sketch of the proof. For completeness, we also describe the key ingredients of the proof along the way.

3.1 Asymptotic distribution

The sample space of our interest is discrete: $\mathcal{Z} = \{z_1, \cdots, z_K\}$, where $K = |\mathcal{Z}|$, and the data generating distribution is $P_\star = \sum_{k=1}^K f_\star^{(k)} \delta_{z_i}$, where $f_\star = (f_\star^{(1)}, \cdots, f_\star^{(K)})^\top \in \Delta_K \triangleq \{x \in \mathbb{R}_+^K : \mathbf{1}_K^\top x = 1\}$ and δ_z is the Dirac measure at z. The auditor observes an empirical measure $P_n = \sum_{k=1}^K f_n^{(k)} \delta_{z_i}$ based on frequency summary of IID samples $Z_1, \cdots, Z_n \sim P_\star$, *i.e.*, $f_n^{(k)} = |\{i \in [n] : Z_i = z_k\}|/n$ for $k = 1, \cdots, K$, and $f_n = (f_n^{(1)}, \cdots, f_n^{(K)})^\top \in \Delta_K$. Hereafter, we do not distinguish between measures $P_\star, P_n$ and their corresponding probability vectors $f_\star, f_n$.

Consider the audit value function $\psi : \Delta_K \to \mathbb{R}_+$ defined as

$$\begin{aligned} \psi(f) \triangleq \quad & \max_{\Pi \in \mathbb{R}_+^{K\times K}} && l^\top(\Pi^\top \mathbf{1}_K - f) \\ & \text{subject to} && \langle C, \Pi \rangle \le \varepsilon \\ & && \langle D, \Pi \rangle = 0 \\ & && \Pi \mathbf{1}_K = f \end{aligned} \tag{3.1}$$

where $C \in \mathbb{R}_+^{K\times K}$ is the cost matrix, $D \in \{0,1\}^{K\times K}$ is the indicator matrix. The FaiTH statistic is the optimal value $\psi(f_n)$. The second constraint $\langle D, \Pi \rangle = 0$ explicitly encodes any restrictions on the transportation plan implicit in the transportation cost function. If $D_{i,j} = 1$, then moving mass from z_i to z_j is prohibited. This is equivalent to $c(z_i, z_j) = \infty$.

Theorem 3.1 (Asymptotic distribution of the FaiTH statistic)**.** *Let $f_\star \in \Delta_K$ and $nf_n \sim \text{Multinomial}(n; f_\star)$. Let $l = (l_1, \cdots, l_K) \in \mathbb{R}_+^K, \varepsilon \ge 0$, $C \in \mathbb{R}_+^{K\times K}$, and $D \in \{0,1\}^{K\times K}$. Define the set*

$$\begin{aligned} \Lambda = \operatorname*{arg\,max}_{\nu,\mu \ge 0, \lambda \in \mathbb{R}^K} \{ & \varepsilon\nu + f_\star^\top \lambda : \\ & \nu C + \mu D + \lambda \mathbf{1}_n^\top \preceq_{\mathbb{R}_+^{K\times K}} -\mathbf{1}_n l^\top \} \end{aligned} \tag{3.2}$$

and the multinomial covariance matrix

$$(\Sigma(p))_{i,j} = \begin{cases} p_i(1-p_i), & \text{if } 1 \le i = j \le K; \\ -p_i p_j, & \text{if } 1 \le i \ne j \le K. \end{cases}$$

The asymptotic distribution of $\psi(f_n)$ is

$$\sqrt{n}\{\psi(f_n) - \psi(f_\star)\} \xrightarrow{d} \inf\{(\lambda + l)^\top Z : (\nu, \mu, \lambda) \in \Lambda\},$$

where $Z \sim \mathcal{N}(\mathbf{0}_K, \Sigma(f_\star))$.

The set Λ in Theorem 3.1 is the set of optimal points of the dual problem of $\psi(f_\star)$, which coincides with the set of Lagrange multipliers of $\psi(f_\star)$ satisfying the optimality conditions. It is generally a convex set. However, if Λ is a singleton, then the asymptotic distribution is Gaussian. This is the generic case, as the inequality constraint in the auditor's problem is generally active. The dual optimum is only non-unique when the inequality constraint is redundant. The left panel of Figure 1 shows a histogram of the values of $\sqrt{n}\{\psi(f_n) - \psi(f_\star)\}$ and its asymptotic distribution.

3.2 Directionally differentiable statistical functionals and delta method

A standard tool for deriving the asymptotic distribution of a statistical functional is the delta method.

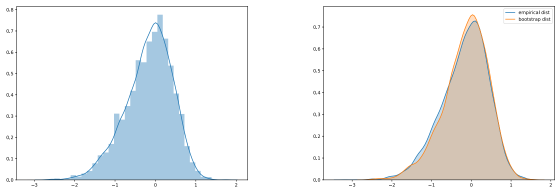


Figure 1: Asymptotic approximation (left panel) and bootstrap approximation (right panel) to the sampling distribution of the FaiTH statistic.

However, the delta method requires the statistical functional to be differentiable (van der Vaart, 1998). Although the audit value function is not differentiable, it is convex and directionally differentiable. As we shall see, this allows us to appeal to a version of the delta method for directionally differentiable functions.

Definition 3.2 (Hadamard directional derivatives)**.** $\mathbb{D}$ *and $\mathbb{E}$ are Banach spaces. A map $\phi : \mathbb{D}_\phi \subseteq \mathbb{D} \to \mathbb{E}$ is called Hadamard directionally differentiable at $\theta_0 \in \mathbb{D}$ tangentially to $\mathbb{D}_0 \subseteq \mathbb{D}$ if there is a map $\phi'_{\theta_0} : \mathbb{D} \to \mathbb{E}$ such that*

$$\lim_{h' \to h, t \to 0^+} \tfrac{1}{t}(\phi(\theta_0 + th') - \phi(\theta_0)) = \phi'_{\theta_0}(h)$$

for any $h \in \mathbb{D}_0$

The audit value function is closely related to the *optimal value function* of the auditor's problem. The optimal value function describes the sensitivity of the optimal value of an optimization problem to perturbations of the problem parameters. Under suitable conditions, the optimal value function is directionally differentiable.

There is a more general version of the delta method for directionally differentiable statistical functionals (Shapiro, 1991; Dümbgen, 1993; Römisch, 2014). Although this version is common in the stochastic optimization literature, it rarely appears in the statistics literature.

Theorem 3.3 (Delta method)**.** *Suppose the following assumptions hold:*

1. *$\mathbb{D}$ and $\mathbb{E}$ are Banach spaces;*
2. *$\phi : \mathbb{D}_\phi \subseteq \mathbb{D} \to \mathbb{E}$ is Hadamard directionally differentiable at θ_0 tangentially to $\mathbb{D}_0$;*
3. *$\theta_0 \in \mathbb{D}_\phi$ and $\hat{\theta}_n : \{X_i\}_{i=1}^n \to \mathbb{D}_\phi$ satisfies $r_n\{\hat{\theta}_n - \theta_0\} \xrightarrow{d} \mathbb{G}_0$ in $\mathbb{D}$ for some $r_n \uparrow \infty$;*
4. *$\mathbb{G}_0$ is tight and its support is included in $\mathbb{D}_0$.*

Then, we have

$$r_n\{\phi(\hat{\theta}_n) - \phi(\theta_0)\} \xrightarrow{d} \phi'_{\theta_0}(\mathbb{G}_0) \text{ in } \mathbb{E}.$$

3.3 Proof sketch of Theorem 3.1

Since $\psi(f)$ can be viewed as the optimal value function of a class of maximization problems parameterized by

f, we can show $\psi(f)$ is Hadamard directionally differentiable at $f_\star$, and give an exact derivative formula by using Proposition 4.27 in Bonnans and Shapiro (2000).

Theorem 3.4. *Under the same assumptions of Theorem 3.1, $\psi(f)$ is Hadamard directionally differentiable at $f_\star$. Furthemore, the derivative is given by*

$$\begin{aligned}\psi'_{f_\star}(h) &= \lim_{\substack{h'\to h\\ t\to 0^+}} \frac{\psi(f_\star + th') - \psi(f_\star)}{t} \\ &= \inf\{(\lambda + l)^\top h : (\nu, \mu, \lambda) \in \Lambda\},\end{aligned}$$

where the convex set Λ is defined by (3.2).

With Theorem 3.4, we can directly show the asymptotic distribution result by applying delta method for Hadamard directionally differentiable functionals.

4 Testing whether an ML model is fair

Theorem 3.1, while insightful, is not immediately useful for inference because the asymptotic distribution depends on the unknown f_*. In this section, we show that a bootstrap approximation to the asymptotic distribution is valid, so it is possible to perform inference with the bootstrap. Due to the non-differentiability of the audit value function (3.1), Efron's non-parametric boostrap (Efron, 1979) is generally invalid. Instead, we consider m-out-of-n bootstrap (Dümbgen, 1993) and a numerical bootstrap (Hong and Li, 2018, 2020).

4.1 Boostrapping the asymptotic distribution of the FaiTH statistic

We start by describing the failure of Efron's non-parametric bootstrap. Let f_n^* be the empirical distribution of n independent samples from f_n. The non-parametric bootstrap approximates the distribution of the FaiTH statistic with the distribution of $\sqrt{n}(\psi(f_n^*) - \psi(f_n))$. This distribution is known as the bootstrap distribution, and the non-parametric bootstrap is consistent if the bootstrap distribution converges weakly to the asymptotic distribution:

$$\sup_{g\in \mathrm{BL}_1(\mathbb{R})} \left| \begin{array}{c} \mathbb{E}^*\left[g\left(\sqrt{n}\left\{\psi(f_n^*) - \psi(f_n)\right\}\right)|f_n\right] \\ -\mathbb{E}\left[g\left(\sqrt{n}\left\{\psi(f_n) - \psi(f_\star)\right\}\right)\right] \end{array} \right| \xrightarrow{p} 0,$$

where $\mathrm{BL}_1(\mathbb{R})$ is 1-Lipschitz subset of the $\|\cdot\|_\infty$ ball. Unfortunately, if ψ is only directionally differentiable (but not differentiable), then the non-parametric bootstrap may fail (Bickel et al., 2012; Andrews, 2000). In fact, it is known that if $\sqrt{n}(f_n - f_*)$ has a Gaussian asymptotic distribution, then the non-parametric bootstrap is consistent if and only if ψ is (Hadamard) differentiable (Fang and Santos, 2019). Unfortunately, as saw in Section 3, the FaiTH statistic is a generally non-differentiable function of the empirical distribution.

Before discussing alternatives to the non-parametric bootstrap, we observe that the audit value function is differentiable at f_* whenever Λ is a singleton. In such problems, $\sqrt{n}(f_n - f_*)$ has a Gaussian asymptotic distribution, so the non-parametric bootstrap is consistent. One practical heuristic to check for failure of the non-parametric bootstrap is checking whether the bootstrap distribution is Gaussian: non-Gaussianity suggests failure of the non-parametric bootstrap.

Fortunately, there are several alternatives to the non-parametric bootstrap that remain consistent for non-differentiable statistical functionals. We refer to these methods as non-standard bootstrap methods. Three promiment methods are the m-out-of-n bootstrap (Dümbgen, 1993; Shao, 1994; Bickel and Sakov, 2008), subsampling (Politis et al., 1999), and the numerical bootstrap (Hong and Li, 2018, 2020). In our computational results, we rely on the m-out-of-n bootstrap and the numerical bootstrap. We provide detailed descriptions of both methods in Section B of the Supplementary Materials.

Theorem 4.1 (Consistency of m-out-of-n bootstrap)**.** *Let $mf^*_{n,m} \sim$ Multinomial$(m; f_n)$. As long as $m = m(n) \to \infty$ and $m/n \to 0$, we have*

$$\sup_{g\in \mathrm{BL}_1(\mathbb{R})} \left| \begin{array}{c} \mathbb{E}^*\left[g\left(\sqrt{m}\left\{\psi(f^*_{n,m}) - \psi(f_n)\right\}\right)|f_n\right] \\ -\mathbb{E}\left[g\left(\sqrt{n}\left\{\psi(f_n) - \psi(f_\star)\right\}\right)\right] \end{array} \right| \xrightarrow{p} 0.$$

Theorem 4.2 (Consistency of numerical derivative method)**.** *Let $z_n^* \sim \mathcal{N}(\mathbf{0}_K, \Sigma(f_n); \mathbb{T})$, a Gaussian distribution truncated in $\mathbb{T}$, where $\mathbb{T} = \mathbb{T}(f_n, \epsilon) = \{x \in \mathbb{R}^K : f_n + \epsilon x \in \mathbb{R}^K_+\}$. As long as $\epsilon = \epsilon(n) \to 0$ and $\sqrt{n}\epsilon \to \infty$, we have*

$$\sup_{g\in \mathrm{BL}_1(\mathbb{R})} \left| \begin{array}{c} \mathbb{E}^*\left[g\left(\epsilon^{-1}\left\{\psi(f_n + \epsilon z_n^*) - \psi(f_n)\right\}\right)|f_n\right] \\ -\mathbb{E}\left[g\left(\sqrt{n}\left\{\psi(f_n) - \psi(f_\star)\right\}\right)\right] \end{array} \right| \xrightarrow{p} 0.$$

4.2 Inference for the audit value

The preceding bootstrap methods complete our suite of inferential tools for the audit value. In this subsection, we demonstrate the utility of the tools by forming confidence intervals and testing restrictions on the audit value.

One of the most basic inferential tasks is forming a confidence interval of the audit value. Such confidence intervals may be used to give an *asymptotically exact certificate* of individual fairness for ML models. Let c_q^* be the q-th quantile of the bootstrap distribution:

$$c_q^* = \inf\{c \in \mathbb{R} : \mathbb{P}(\sqrt{m}\{\psi(f^*_{n,m}) - \psi(f_n)\} \le c) \ge q\},$$

where $0 \le q \le 1$. In practice, c_q^* is estimated by q-th quantile of output $\mathcal{S}$ of Algorithm 1 in the Supplementary Materials. Since the approximation error can be made arbitrarily small by increasing number of bootstrap iterations B, we ignore this error in our results.

The two-sided equal-tailed confidence interval for the audit value $\psi(f_\star)$ with asymptotic coverage probability $1-\alpha$ is

$$\mathrm{CI}_{\text{two-sided}} = \left[\psi(f_n) - \frac{c^*_{1-\alpha/2}}{\sqrt{n}}, \psi(f_n) - \frac{c^*_{\alpha/2}}{\sqrt{n}}\right]. \quad (4.1)$$

Theorem 4.3 (Asymptotic coverage of two-sided confidence interval)**.** *For any $f_\star \in \Delta_K$, we have*

$$\liminf_{n\to\infty} \mathbb{P}\left(\psi(f_\star) \in \mathrm{CI}_{\text{two-sided}}\right) \geq 1-\alpha.$$

Compared to other certificates of individual fairness (*e.g.*, the certificate in Yurochkin et al. (2020)), our certificate is asymptotically exact. This is a consequence of the asymptotic exactness of the coverage of the confidence interval (4.1).

Another basic inferential task is testing restrictions on the audit value. In light of the (asymptotic) validity two-sided confidence region (4.1), it is possible to test simple restrictions of the form $\psi(f_*) = \delta$, for some $\delta > 0$, by checking whether δ falls in the $(1-\alpha)$-level confidence region. By the duality between confidence intervals and hypothesis tests, this test has asymptotic Type I error rate at most α. In the rest of this subsection, we consider the task of testing a compound hypothesis of the form $\psi(f_*) < \delta$.

Definition 4.4. (δ–fairness). *For a constant $\delta \geq 0$, an ML system is called δ–fair if $\psi(f_\star) \leq \delta$.*

In order to test whether or not an ML system is δ–fair, the auditor considers hypothesis testing problem

$$H_0 : \psi(f_\star) \leq \delta \quad \text{versus} \quad H_1 : \psi(f_\star) > \delta. \quad (4.2)$$

The one-sided confidence interval for the audit value $\psi(f_\star)$ with asymptotic coverage probability $1-\alpha$ is

$$\mathrm{CI}_{\text{one-sided}} = \left[\psi(f_n) - \frac{c^*_{1-\alpha}}{\sqrt{n}}, \infty\right).$$

We reject the null hypothesis H_0 if the one-sided confidence interval does not cover δ, *i.e.*,

$$\delta \notin \left[\psi(f_n) - \frac{c^*_{1-\alpha}}{\sqrt{n}}, \infty\right).$$

Theorem 4.5 (Asymptotic validity of test)**.** *For any $\delta \geq 0$, we have*

$$\limsup_{n\to\infty} \sup_{f_\star \in \Delta_K : \psi(f_\star) \leq \delta} \mathbb{P}_{f_\star}\left(\delta \notin \mathrm{CI}_{\text{one-sided}}\right) \leq \alpha.$$

If $\psi(f_\star) > \delta$, then $\lim_{n\to\infty} \mathbb{P}\left(\delta \notin \mathrm{CI}_{\text{one-sided}}\right) = 1$.

The choice of threshold δ is application dependent, and there is no generic recipe to pick δ. It reflects the auditor's tolerance on fairness level of an ML system. For example, in recidivism prediction, a reasonable threshold may be the rate of miscarriage of justice. In other words, the auditor expects the performance of the recidivism prediction instrument to deteriorate by no more than the inherent error rate in the criminal justice system. We demonstrate the suitability of this choice in our computational results.

5 Computational results

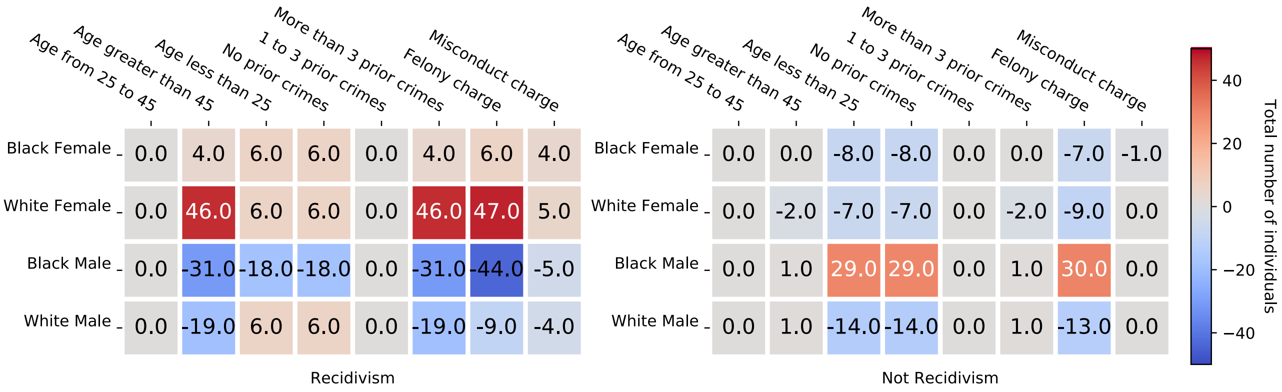
We shall verify correctness of our methodology using widely studied COMPAS dataset (Angwin et al., 2016). Originally it was shown that COMPAS score used for providing recommendation to the judge if a person will recommit or not is biased against certain groups of individuals. In Angwin et al. (2016), it was shown that COMPAS score is strongly biased against men and minorities.

To apply our methodology it remains to choose metric and loss function for the auditor's problem. We make choices to facilitate simplicity and interpretability of the analysis. For the metric we consider any two observations which only differ in race or gender to have distance zero between each other and infinity otherwise. For the loss we shall consider 0-1 loss, then FaiTH value can be understood as missclassification rates induced by the solution of the auditor's problem (2.2) and threshold δ corresponds to the amount of classification errors that the auditor believes it is justified for the problem. Here we choose $\delta = 0.0365$, which is the midpoint of the results reported by various studies on the number of innocent prisoners in the United States (Wikipedia).

5.1 Audit guidelines and interpretation

In this subsection we give practical guidelines for an auditor wishing to assess performance of an ML system. We will investigate performance of a vanilla logistic regression (LR) classifier trained on COMPAS dataset to predict if a person will re-offend. We use 70% of the COMPAS dataset to train the classifier and the remaining 30% to audit it using black-box access to the trained model. To determine if an ML system is individually fair we compute the FaiTH value and report lower and upper bounds of the 95% two-sided confidence interval ($\mathrm{CI}^{(2)}_{\text{lower}}$ and $\mathrm{CI}^{(2)}_{\text{upper}}$) and lower bound of the 95% one-sided confidence interval ($\mathrm{CI}^{(1)}_{\text{lower}}$) using methodology described in the preceding sections. We fail to reject the hypothesis that a classifier is individually fair if a pre-specified value of δ is contained in the confidence interval.

We repeat the experiment 50 times and summarize the results in Table 1. Common group-fairness metrics are reported and FaiTH is applied to test previously proposed fair classification techniques motivated by the notion of group fairness. Before discussing the relation to group fairness, we complete the audit analysis of the logistic regression. Both one- and two-sided confidence intervals lower bounds are equal to $0.05 > \delta$ on average, meaning that auditor should reject the individual fairness hypothesis of the logistic regression classifier.



Recidivism

	Age from 25 to 45	Age greater than 45	Age less than 25	No prior crimes	1 to 3 prior crimes	More than 3 prior crimes	Felony charge	Misconduct charge
Black Female	0.0	4.0	6.0	6.0	0.0	4.0	6.0	4.0
White Female	0.0	46.0	6.0	6.0	0.0	46.0	47.0	5.0
Black Male	0.0	-31.0	-18.0	-18.0	0.0	-31.0	-44.0	-5.0
White Male	0.0	-19.0	6.0	6.0	0.0	-19.0	-9.0	-4.0

Not Recidivism

	Age from 25 to 45	Age greater than 45	Age less than 25	No prior crimes	1 to 3 prior crimes	More than 3 prior crimes	Felony charge	Misconduct charge
Black Female	0.0	0.0	-8.0	-8.0	0.0	0.0	-7.0	-1.0
White Female	0.0	-2.0	-7.0	-7.0	0.0	-2.0	-9.0	0.0
Black Male	0.0	1.0	29.0	29.0	0.0	1.0	30.0	0.0
White Male	0.0	1.0	-14.0	-14.0	0.0	1.0	-13.0	0.0

Figure 2: Transport map of vanilla logistic regression on audit dataset. (Number in each grid shows the change in total number of individuals after transport.)

Table 1: Numerical comparisons of multiple fairness methods.

	FaiTH	$\text{CI}^{(2)}_{\text{lower}}$	$\text{CI}^{(2)}_{\text{upper}}$	$\text{CI}^{(1)}_{\text{lower}}$	Accuracy	AOD	EOD	SPD
LR	$.06 \pm .02$	$.05 \pm .02$	$.07 \pm .03$	$.05 \pm .02$	$\mathbf{.67} \pm .01$	$-.23 \pm .04$	$-.19 \pm .04$	$-.26 \pm .03$
ADB	$.18 \pm .06$	$.16 \pm .05$	$.20 \pm .06$	$.16 \pm .05$	$.65 \pm .01$	$-.05 \pm .13$	$-.01 \pm .12$	$-.08 \pm .13$
RWT	$.15 \pm .02$	$.13 \pm .02$	$.17 \pm .02$	$.14 \pm .02$	$.66 \pm .01$	$\mathbf{-.02} \pm .04$	$\mathbf{.01} \pm .04$	$\mathbf{-.06} \pm .04$
LFR	$.07 \pm .05$	$.06 \pm .04$	$.08 \pm .05$	$.06 \pm .05$	$.66 \pm .01$	$-.09 \pm .09$	$-.06 \pm .07$	$-.13 \pm .08$
RLR	$\mathbf{.02} \pm .02$	$\mathbf{.01} \pm .02$	$\mathbf{.02} \pm .02$	$\mathbf{.01} \pm .02$	$.66 \pm .01$	$-.19 \pm .03$	$-.15 \pm .03$	$-.22 \pm .03$

In this situation auditor may utilize the adversarial distribution computed to evaluate the FaiTH statistics in (2.2) to investigate the patterns of individual fairness violation. We present such analysis in Figure 2. On the left heat map we show the change in distribution of the features of individuals labeled as recidivists in the audit data (counts of the distribution maximizing (2.2) minus counts of the audit dataset distribution). We can interpret the figure column-wise: there are 31 black males and 19 white males older than 45 that were correctly classified as recidivists, but would be misclassified as non-reoffenders if they were to be white females (or black females for the 4 of them); similar argument holds for recidivists with more than 3 prior crimes and/or a felony charge. In summary, we see that white females are treated by the classifier as a privileged group. The right figure shows analogous heat map for individuals labeled as non-reoffenders in the audit data. Among others we see that young white males and females, and black females correctly classified to not commit recidivism would be classified as recidivists if they were to be black males. Previous study of the COMPAS dataset reports white females as the privileged group and black males as unprivileged (ProPublica), aligning with our findings. We can also make an additional observation based on our analysis: people in the age group of 25 to 45 and/or those with 1 to 3 prior crimes were treated individually fair by the classifier. Auditor may utilize such findings to provide recommendations to the ML system provider if the system fails to pass the FaiTH test without disclosing the audit data.

Relation to group fairness We proceed to evaluate the individual fairness hypothesis for several group fairness approaches proposed in the literature. We consider three algorithms available in the IBM AIF360 toolkit (Bellamy et al., 2018). Two preprocessing techniques: Reweighting (RWT) (Kamiran and Calders, 2012) that modifies data weights in the training loss, and Learning Fair Representation (LFR) (Zemel et al., 2013) that finds transformed feature space obfuscating information about protected attributes. And an in-processing technique: Adversarial Debiasing (ADB) (Zhang et al., 2018) that learns a group-fair predictor by reducing the ability of a corresponding adversary to predict protected attributes. We also report common group fairness metrics (for all prefered value is close to 0): average odds difference (AOD), equal opportunity difference (EOD) and statistical parity difference (SPD). Results are summarized in Table 1: all of these methods succeed in reducing the group biases, however they tend to exacerbate individual fairness violations as can be seen from the FaiTH value. For example, Reweighting method appears to mitigate most of the group biases, but investigating corresponding logistic regression fit we find that it assigns large coefficient to the race variable. In other words, decision of the corresponding classifier is majorly affected by the race, which is not permissible from the perspective of individual fairness and an alarm is raised by FaiTH.

5.2 Model selection under FaiTH constraint

In this subsection, we propose a generic model selection strategy under δ-fairness constraint, and present the strategy by logistic regression with ℓ_1 penalty.

The idea of strategy is to select candidates of models which pass the fairness hypothesis testing (4.2). To be precise, we filter all models through comparison between the fairness threshold δ and the CI lower bound of audit value evaluated on validation dataset. Then among these candidates, we select the model which has the lowest validation error.

The dataset is splited into training, validation, and audit dataset. We fit ℓ_1-regularized logistic regression (RLR) by minimizing $\mathcal{L}(Z,\beta)+\lambda\|\beta\|_1$, where β is vector of regression coefficients, Z is the training set, $\mathcal{L}$ is the logistic loss, and $\lambda > 0$ is a tuning parameter.

Figure 3 demonstrates trade-off between accuracy and fairness. Strong penalty (*i.e.*, small value of $\frac{1}{\lambda}$) results in tiny FaiTH statistic but huge validation error, and on the contrary, weak penalty (*i.e.*, large value of $\frac{1}{\lambda}$) leads to undesirable fairness level but satisfactory accuracy. The broken orange line shows lower bounds of 95% confidence interval (one-sided) of validation audit value for each λ. Note that a tuning parameter λ passes the δ–fairness test if and only if its corresponding CI lower bound is smaller than δ, so the range of that orange broken line lies under green dotted line determines all candidates of δ–fair tuning parameters. Choosing the tuning parameter which has lowest validation error among these candidates outputs the selected $\frac{1}{\lambda} = 0.0145$. We note that gender is not selected so that prediction without using gender can effectively ensure model's individual fairness and keep comparable prediction accuracy at the same time.

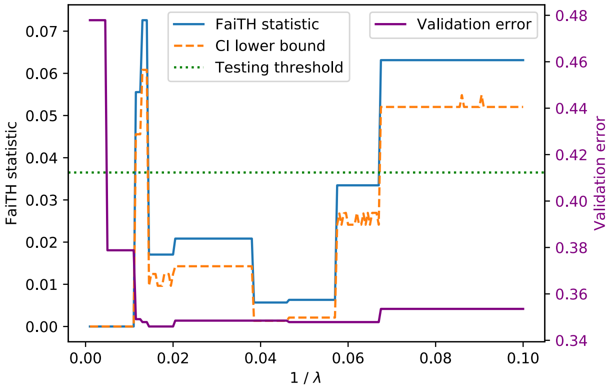


Figure 3: Performance of logistic regression with ℓ_1 penalty on validation dataset.

Solution pathes of regression coefficients are depicted in Figure 4. The vertical dotted line $\frac{1}{\lambda} = 0.0145$ shows the selected model. Whether or not an individual has prior crimes is of the greatest significance for predicting recidivism since the corresponding coefficient pops out firstly. The other five selected variables are "more than 3 prior crimes", race, "age greater than 45", "misconduct charge", and "age less than 25" in sequence.

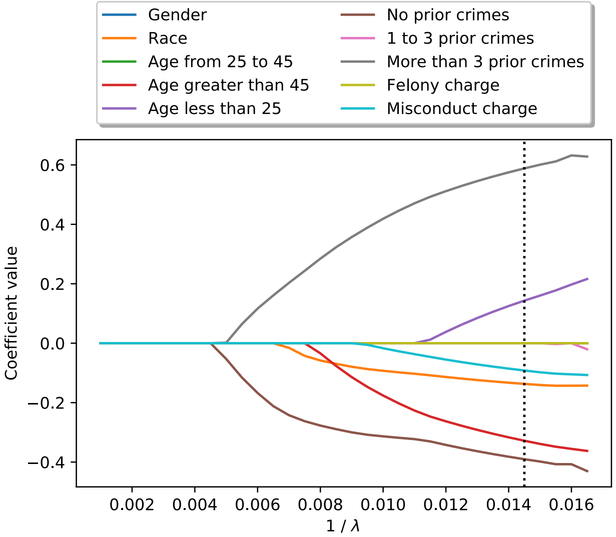


Figure 4: Solution pathes of logistic regression with ℓ_1 penalty.

We run our model selection strategy for 50 times and make comparison with other methods in Table 1. RLR continues to have low FaiTH value when we computed on the audit dataset and is the only method for which we fail to reject the individual fairness hypothesis. RLR also has better group fairness scores than the baseline, however not as good as those of other group fairness approaches. We note that RLR is a simple model selection based approach that is plausible due to the development of our FaiTH methodology. Combining FaiTH with prior ideas used for group fairness may layout a pass for training ML systems with strong guarantees for both individual and group fairness.

6 Summary and discussion

In this paper, we developed a suite of inferential tools for detecting and localizing individual bias/unfairness in the ML model. Our tools only require black-box access to the ML model and are computationally efficient. Further, they allow auditors to control the false alarm rate and provide asymptotically exact certificates of fairness. We demonstrated the utility of our tools by using them to reveal the gender and racial biases in Northpointe's COMPAS recidivism prediction instrument.

Acknowledgements

This work was supported by the National Science Foundation under grants DMS-1830247 and DMS-1916271.

References

Antoine Allen. The 'three black teenagers' search shows it is society, not Google, that is racist — Antoine Allen. *The Guardian*, June 2016. ISSN 0261-3077.

Donald W. K. Andrews. Inconsistency of the Bootstrap When a Parameter is on the Boundary of the Parameter Space. *Econometrica*, 68(2):399–405, 2000. ISSN 0012-9682.

Julia Angwin and Terry Parris Jr. Facebook Lets Advertisers Exclude Users by Race. https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race, October 2016.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine Bias. www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, May 2016.

Julia Angwin, Ariana Tobin, and Madeleine Varner. Facebook (Still) Letting Housing Advertisers Exclude Users by Race. https://www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin, November 2017.

Solon Barocas and Andrew D. Selbst. Big Data's Disparate Impact. *SSRN Electronic Journal*, 2016. ISSN 1556-5068. doi: 10.2139/ssrn.2477899.

Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, October 2018. URL https://arxiv.org/abs/1810.01943.

Marianne Bertrand and Sendhil Mullainathan. Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review*, 94(4): 991–1013, September 2004. ISSN 0002-8282. doi: 10.1257/0002828042002561.

P. J. Bickel, F. Götze, and W. R. van Zwet. Resampling Fewer Than n Observations: Gains, Losses, and Remedies for Losses. In Sara van de Geer and Marten Wegkamp, editors, *Selected Works of Willem van Zwet*, pages 267–297. Springer New York, New York, NY, 2012. ISBN 978-1-4614-1313-4 978-1-4614-1314-1. doi: 10.1007/978-1-4614-1314-1_17.

Peter J Bickel and Anat Sakov. On the choice of m in the m out of n bootstrap and confidence bounds for extrema. *Statistica Sinica*, pages 967–985, 2008.

Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

Jose Blanchet, Karthyek Murthy, and Nian Si. Confidence Regions in Wasserstein Distributionally Robust Estimation. *arXiv:1906.01614 [math, stat]*, June 2019.

Joseph Frédéric Bonnans and Alexander Shapiro. *Perturbation Analysis of Optimization Problems*. Springer Series in Operations Research. Springer, New York, NY, 2000. ISBN 978-1-4612-7129-1 978-0-387-98705-7. OCLC: 247674137.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *arXiv:1703.00056 [cs, stat]*, February 2017.

Jeffrey Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*, October 2018.

Lutz Dümbgen. On nondifferentiable functions and the bootstrap. *Probability Theory and Related Fields*, 95 (1):125–140, 1993.

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

B Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

Zheng Fang and Andres Santos. Inference on directionally differentiable functions. *The Review of Economic Studies*, 86(1):377–412, 2019.

Han Hong and Jessie Li. The numerical delta method. *Journal of Econometrics*, 206(2):379–394, 2018.

Han Hong and Jessie Li. The numerical bootstrap. *The Annals of Statistics*, 48(1):397–412, 2020.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1): 1–33, 2012.

Marcel Klatt, Carla Tameling, and Axel Munk. Empirical Regularized Optimal Transport: Statistical

Theory and Applications. *arXiv:1810.09880 [math, stat]*, October 2018.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. *arXiv:1609.05807 [cs, stat]*, September 2016.

Jaeho Lee and Maxim Raginsky. Minimax statistical learning with wasserstein distances. In *Advances in Neural Information Processing Systems*, pages 2687–2696, 2018.

Dimitris N Politis, Joseph P Romano, and Michael Wolf. *Subsampling*. Springer Science & Business Media, 1999.

ProPublica. How we analyzed the compas recidivism algorithm.

Werner Römisch. Delta method, infinite dimensional. *Wiley StatsRef: Statistics Reference Online*, 2014.

Jun Shao. Bootstrap Sample Size in Nonregular Cases. *Proceedings of the American Mathematical Society*, 122(4):1251–1262, 1994. ISSN 0002-9939. doi: 10.2307/2161196.

Alexander Shapiro. Asymptotic analysis of stochastic programs. *Annals of Operations Research*, 30(1): 169–186, 1991.

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying Some Distributional Robustness with Principled Adversarial Training. *arXiv:1710.10571 [cs, stat]*, October 2017.

Max Sommerfeld and Axel Munk. Inference for empirical wasserstein distances on finite spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):219–238, 2018.

Ariana Tobin. HUD Sues Facebook Over Housing Discrimination and Says the Company's Algorithms Have Made the Problem Worse. https://www.propublica.org/article/hud-sues-facebook-housing-discrimination-advertising-algorithms, March 2019a.

Ariana Tobin. New York Is Investigating Whether Facebook Lets Advertisers Discriminate. https://www.propublica.org/article/new-york-is-investigating-whether-facebook-lets-advertisers-discriminate, July 2019b.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, October 1998. doi: 10.1017/CBO9780511802256.

Wikipedia. Miscarriage of justice.

Mikhail Yurochkin, Amanda Bower, and Yuekai Sun. Training individually fair ML models with sensitive subspace robustness. In *International Conference on Learning Representations*, Addis Ababa, Ethiopia, 2020.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning Fair Representations. In *International Conference on Machine Learning*, pages 325–333, February 2013.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.