Auditing Machine Learning Models for Individual Bias and Unfairness

Songkai Xue
Department of Statistics, University of Michigan

Joint work with Mikhail Yurochkin and Yuekai Sun

Introduction

High-stakes decision making involves

- Recidivism prediction (Angwin et al., 2016);
- Housing advertisement (Angwin, Tobin and Varner, 2017);
- Resume screening (Jeffrey, 2018).

Who makes the decision?

$$\text{Human} \stackrel{?}{=} \text{Bias}$$

$$\text{Machine} \neq \text{No Bias}$$

Northpointe's COMPAS Dataset

Correctional **O**ffender **M**anagement **P**rofiling for **A**lternative **S**anctions

Disparate impact on

- Minorities;
- Underprivileged groups.

Protected/Sensitive attributes include

- Race (black, white, $\cdots$);
- Gender (female, male, $\cdots$).

These attributes are protected by *federal anti-discrimination law*.

Northpointe's COMPAS Dataset (Cont.)

Prediction fails differently for black defendants.

	White	Black
Labeled higher risk, but didn't re-offend	23.5%	44.9%
Labeled lower risk, but did re-offend	47.7%	28.0%

(Source: *Machine bias*, by ProPublica.)

Algorithmic Fairness

Formal definitions of algorithmic fairness? YES.

- Dwork et al. (2012);
- Kleinberg, Mullainathan and Raghavan (2017);
- Chouldechova (2017);
- ...

Individual fairness + (statistically) inferential tools?

Lacking.

(This is what we wish to do.)

Group Fairness

Group fairness is amenable to statistical analysis, ...

- **Calibration**: equal false discovery and non-discovery rates.
- **Equalized odds**: equal false positive and negative rates.

but fails under scrutiny.

- ML models that satisfy group fairness may be blatantly unfair for individual users (Dwork et al., 2012).
- There are fundamental incompatibilities between common notions of group fairness (Kleinberg et al., 2017; Chouldechov, 2017).

Individual Fairness

Main idea:

"Treat similar users similarly".

Definition (Individual fairness, Dwork et al., 2012)

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

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

where $d_x : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ (resp. $d_y : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$) measures similarity between users (resp. outputs).

What's in the Pipeline?

1. Training individually fair ML models:
   Yurochkin, Bower, Sun, *ICLR 2020*.

2. Testing whether an ML model is individually fair or not:
   Xue, Yurochkin, Sun, *AISTATS 2020*.

Benefits of Our Methods

Main benefits are

1. Black-box:

   Observing the outputs of ML models is sufficient.

2. Computational efficiency:

   The auditor solves a convex optimization problem.

3. Interpretability:

   Specific metric leads to specific interpretation.

Mathematical Preliminaries

- The sample space:

$$\mathcal{Z} \triangleq \mathcal{X} \times \mathcal{Y}$$

- The induced metric on $\mathcal{Z}$:

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

- The Wasserstein distance on $\Delta(\mathcal{Z})$:

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

  where
  - $\Delta(\mathcal{Z})$ is the set of probability distributions on $\mathcal{Z}$;
  - $\mathcal{C}(P, Q)$ is the set of couplings between P and Q;
  - $c(\cdot, \cdot) = d_z^2(\cdot, \cdot)$ is the transportation cost function.

The Auditor's Problem

Population version of the auditor's problem:

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

where $\varepsilon \geq 0$ is a transportation budget parameter, $\ell_h : \mathcal{Z} \to \mathbb{R}_+$ is a loss function picked by the auditor.

Main idea: If there is (purely) 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.

The Auditor's Problem (Cont.)

Empirical version of the auditor's problem:

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

where P_n is the empirical distribution of the collected audit data $\{(x_i, y_i)\}_{i=1}^n$, since $P_\star$ is unknown in practice.

FaiTH statistic: We call the optimal value of this optimization problem the **Fai**r **T**ransport **H**ypothesis test statistic.

The Auditor's Problem (Cont.)

Original problem:

$$\max_{W(P,P_n)\leq\varepsilon} \mathbb{E}_{Z\sim P}[\ell_h(Z)].$$

Dual problem (Blanchet and Murthy, 2019):

$$\max_{W(P,P_n)\leq\varepsilon} \mathbb{E}_{Z\sim P}[\ell_h(Z)] = \min_{\lambda\geq 0}\{\lambda\varepsilon + \mathbb{E}_{Z\sim P_n}[\ell^c_{h,\lambda}(Z)]\},$$

$$\ell^c_{h,\lambda}(x_i, y_i) = \max_{x\in\mathcal{X}}\{\ell_h(x, y_i) - \lambda d^2_x(x, x_i)\}.$$

Pros: univariate problem; amenable to stochastic optimization.
Cons: no global convergence guarantee; hard to establish limiting distribution of test statistic.

The Auditor's Problem (Cont.)

Empirical version of the auditor's problem on finite sample space:

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

where

- $l \in \mathbb{R}^{|\mathcal{Z}|}$ is the vector of losses;
- $C \in \mathbb{R}^{|\mathcal{Z}| \times |\mathcal{Z}|}$ is the matrix of transportation costs;
- $f_{|\mathcal{Z}|} \in \Delta^{|\mathcal{Z}|}$ is the empirical distribution of the data.

Asymptotics of the FaiTH Statistic

Let

- $K = |\mathcal{Z}|$, $l \in \mathbb{R}_+^K$ and $\varepsilon \geq 0$;
- $f_\star \in \Delta^K$ and $nf_n \sim \text{Multinomial}(n; f_\star)$;
- $C \in \mathbb{R}_+^{K \times K}$ and $D \in \{0,1\}^{K \times K}$.

The **FaiTH statistic** is given by the value function

$$\psi(f_n) \triangleq \left\{ \begin{array}{ll} \max\limits_{\Pi \in \mathbb{R}_+^{K \times K}} & l^\top (\Pi^\top \mathbf{1}_K - f_n) \\ \text{subject to} & \langle C, \Pi \rangle \leq \varepsilon \\ & \langle D, \Pi \rangle = 0 \\ & \Pi \mathbf{1}_K = f_n \end{array} \right\}.$$

The **audit value** is given by $\psi(f_\star)$.

Asymptotics of the FaiTH Statistic (Cont.)

Theorem (Asymptotic distribution of the FaiTH statistic)

The asymptotic distribution of $\psi(f_n)$ is the infimum of a Gaussian process:

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

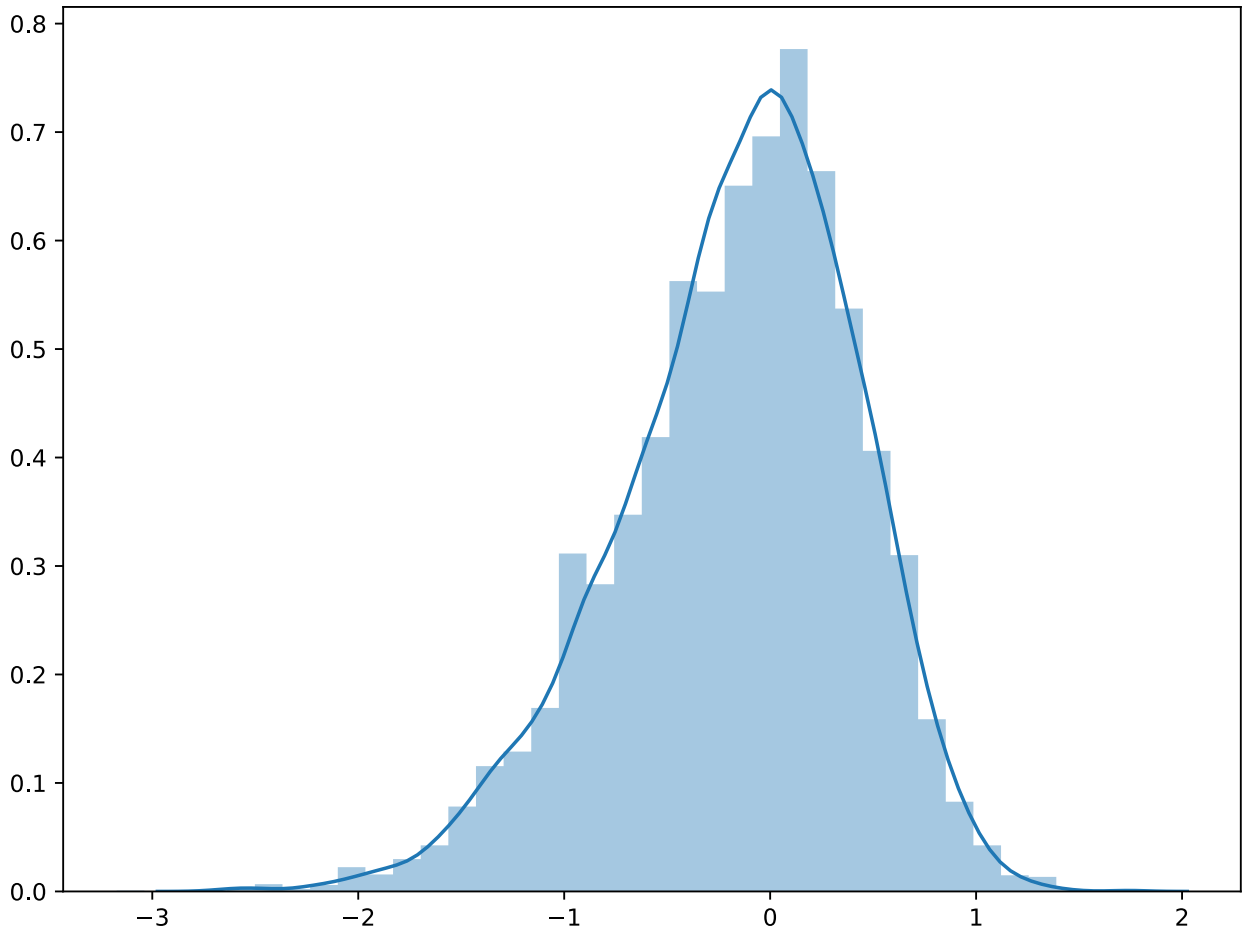
where $Z \sim \mathcal{N}(\mathbf{0}_K, \Sigma(f_\star))$, Σ is the multinomial covariance matrix of $f_\star$, and

$$\Lambda = \underset{\nu,\mu \geq 0, \lambda \in \mathbb{R}^K}{\arg\max} \{\varepsilon\nu + f_\star^\top \lambda : \nu C + \mu D + \lambda \mathbf{1}_n^\top \preceq_{\mathbb{R}_+^{K \times K}} -\mathbf{1}_n l^\top\}.$$

Proof: Canonical perturbation theory $\implies$ Hadamard directional differentiability $\implies$ Delta method.

Asymptotics of the FaiTH Statistic (Cont.)

A non-Gaussian example:



Boostrapping the Audit Value

Efron's n-out-of-n bootstrap is not consistent because ψ is not smooth enough. Instead, we use m-out-of-n bootstrap.

Theorem (Consistency of m-out-of-n bootstrap)

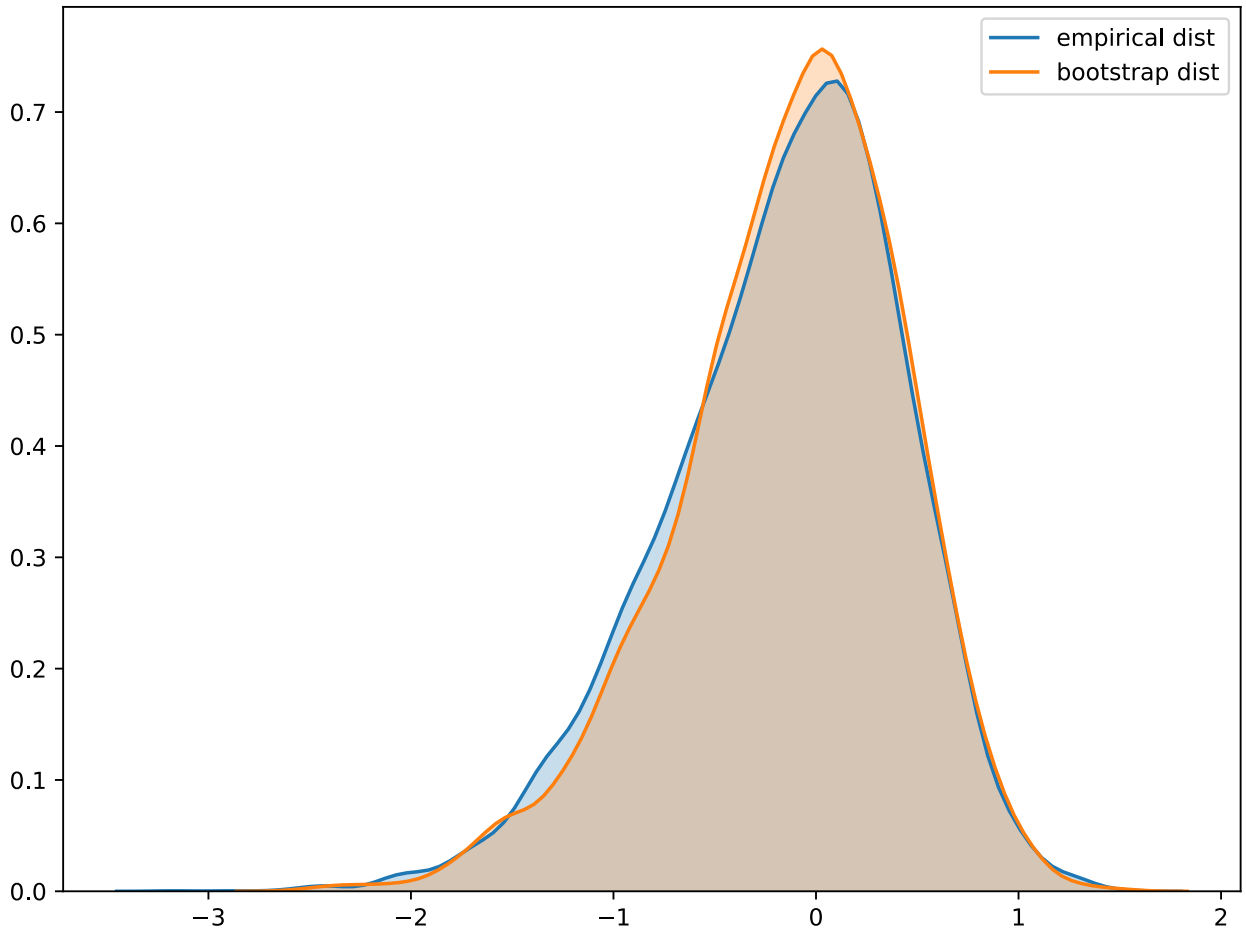
Let $mf^*_{n,m} \sim \text{Multinomial}(m; f_n)$. *As long as* $m = m(n) \to \infty$ *and* $m/n \to 0$, *we have*

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

where $\text{BL}_1(\mathbb{R})$ *is the* 1*-Lipschitz function subset of the* $\|\cdot\|_\infty$ *ball.*

Boostrapping the Audit Value (Cont.)

A non-Gaussian example:



Fair Transport Hypothesis Test

Definition (δ-fairness)

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

Fair **T**ransport **H**ypothesis Test (**FaiTH** test):

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

The auditor considers this hypothesis testing problem in order to test whether or not an ML system is δ-fair.

Inference for the Audit Value

Two-sided confidence interval for the audit value $\psi(f_\star)$:

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

where c^*_q be the q-th quantile of the bootstrap distribution.

Theorem (Asymptotic coverage of two-sided CI)

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

Inference for the Audit Value (Cont.)

One-sided confidence interval for the audit value $\psi(f_\star)$:

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

We reject the null hypothesis H_0 if

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

Theorem (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 $\lim_{n\to\infty} \mathbb{P}\left(\delta \notin \mathrm{CI}_{\text{one-sided}}\right) = 1$.

COMPAS Results

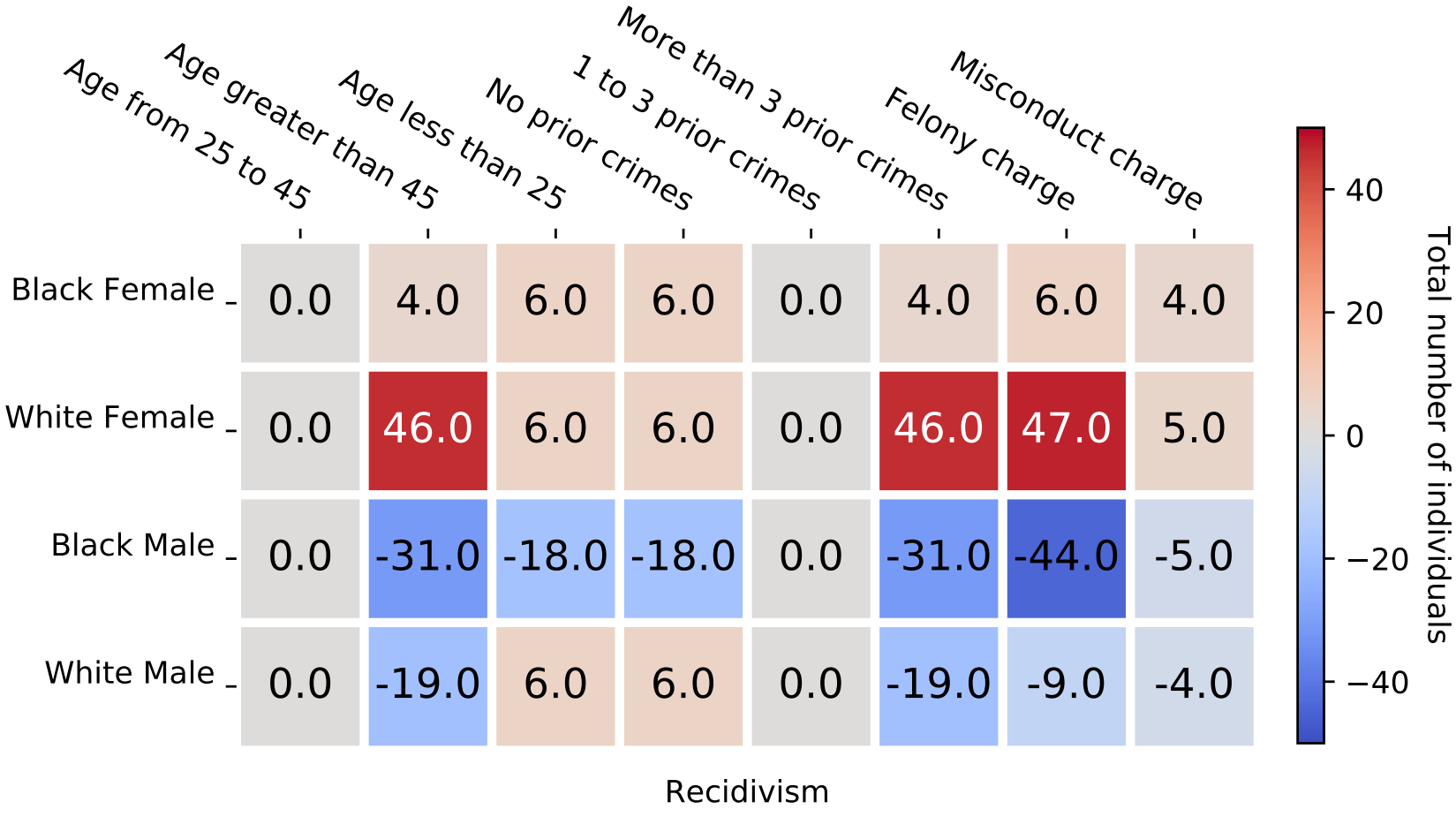
Experiment setup:

- Total number of data points: 5278;
- 70% for training and 30% for auditing ($n = 1584$);
- Discrete space $\mathcal{Z}$ with $|\mathcal{Z}| = 144$;
- Two samples which only differ in race or gender are free to move;
- $0 - 1$ loss, and $\delta = 0.0365$.

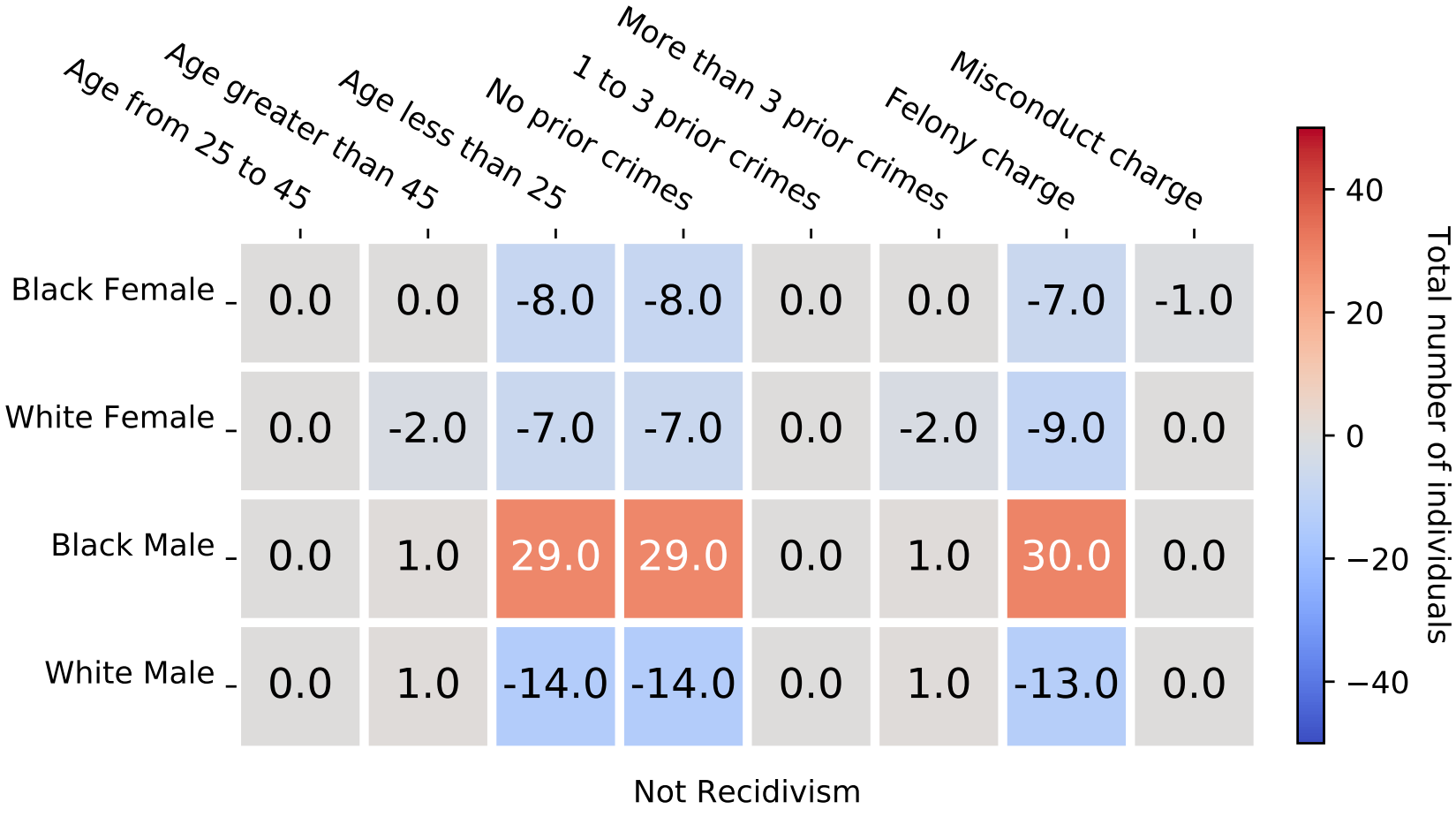
FaiTH value can be interpreted as misclassification rates induced by the solution of the auditor's problem.

3.65% is the midpoint of the proportion of innocent prisoners in the United States. (Source: *Miscarriage of justice*, by B. A. Garner)

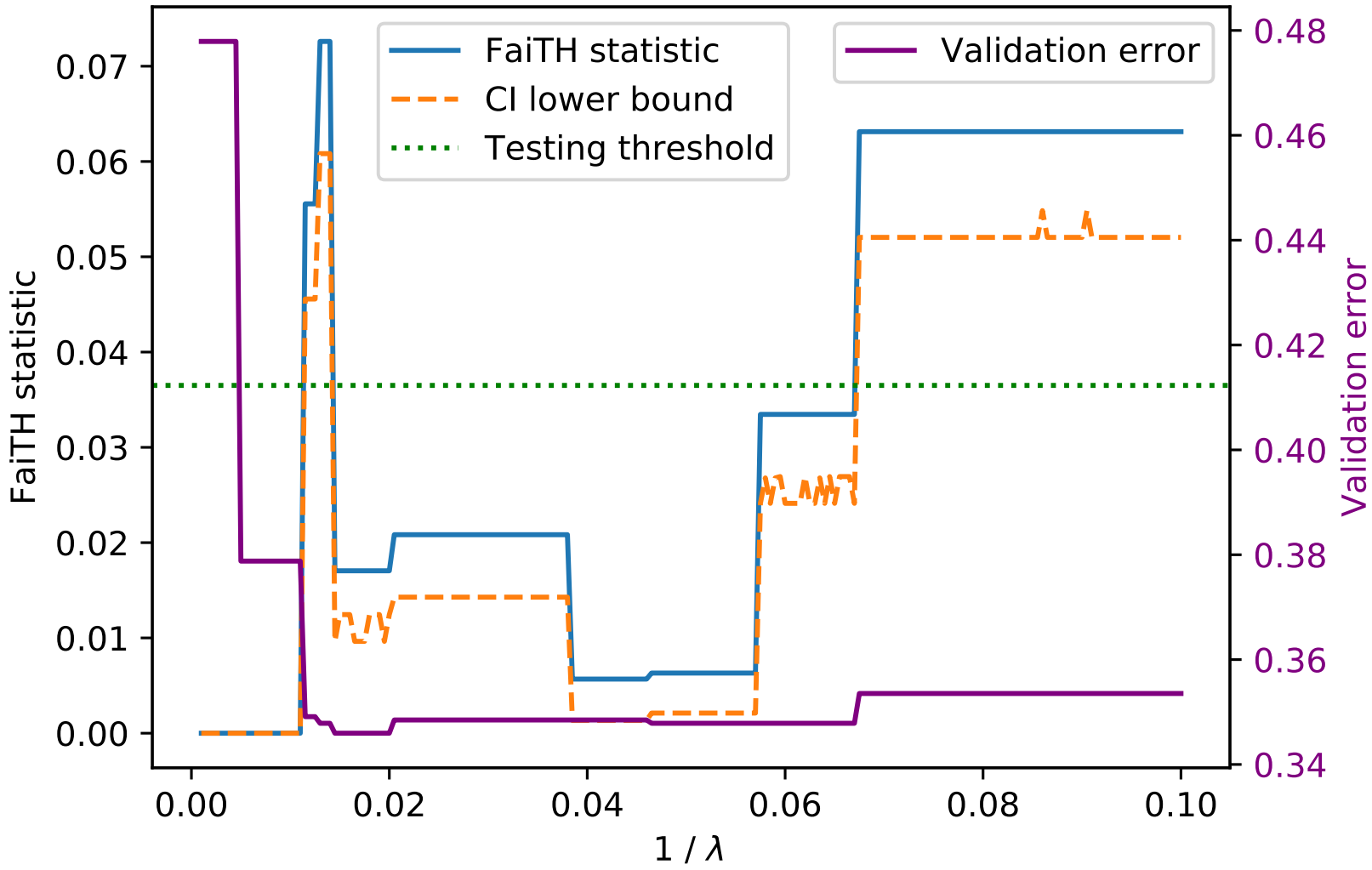
COMPAS Results (Cont.)



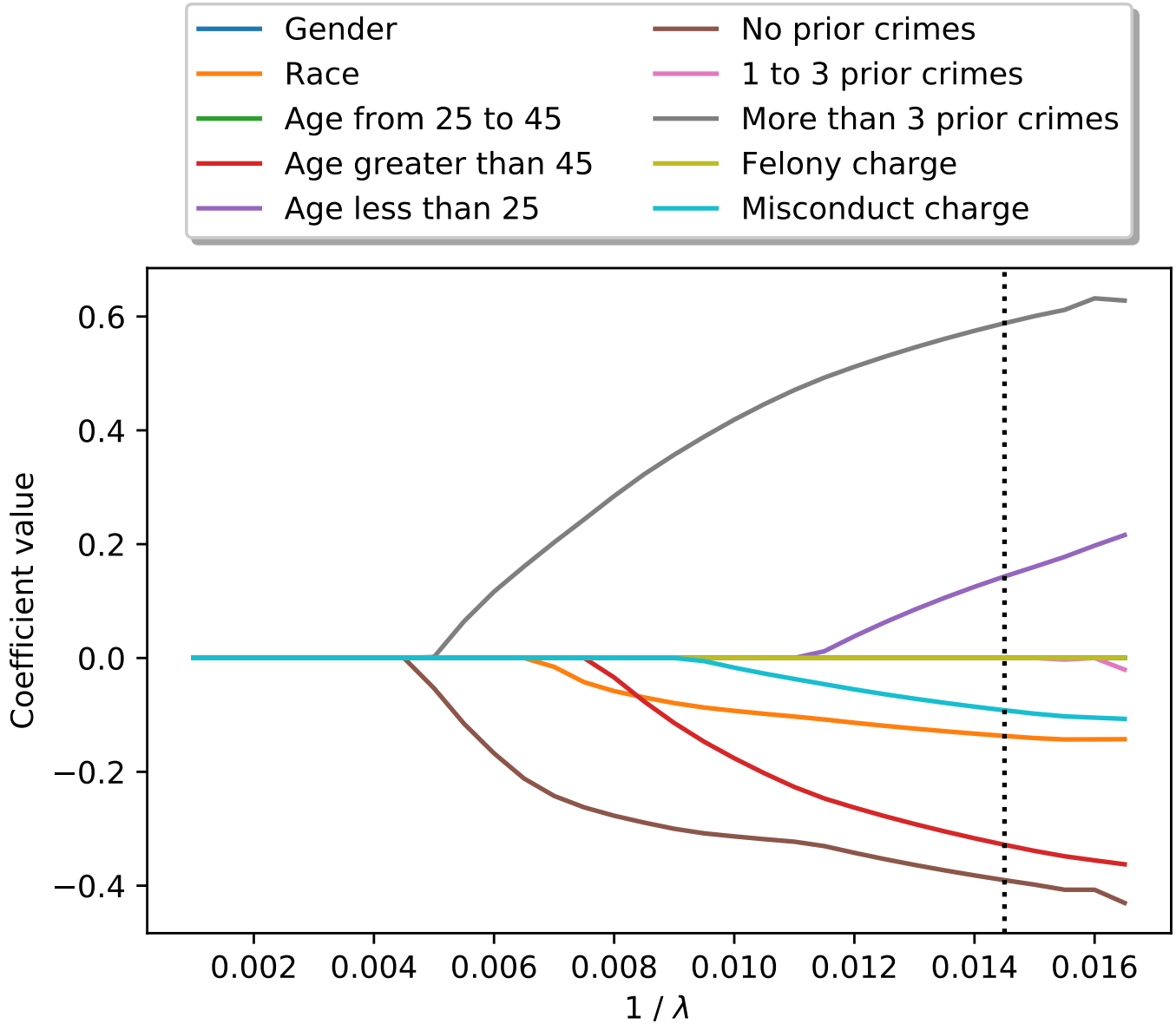
COMPAS Results (Cont.)



COMPAS Results (Cont.)



COMPAS Results (Cont.)



COMPAS Results (Cont.)

	FaiTH	$\text{CI}^{(2)}_{\text{lower}}$	$\text{CI}^{(2)}_{\text{upper}}$	$\text{CI}^{(1)}_{\text{lower}}$
LR	$.06 \pm .02$	$.05 \pm .02$	$.07 \pm .03$	$.05 \pm .02$
ADB	$.18 \pm .06$	$.16 \pm .05$	$.20 \pm .06$	$.16 \pm .05$
RWT	$.15 \pm .02$	$.13 \pm .02$	$.17 \pm .02$	$.14 \pm .02$
LFR	$.07 \pm .05$	$.06 \pm .04$	$.08 \pm .05$	$.06 \pm .05$
RLR	$\mathbf{.02} \pm .02$	$\mathbf{.01} \pm .02$	$\mathbf{.02} \pm .02$	$\mathbf{.01} \pm .02$

	Accuracy	AOD	EOD	SPD
LR	$\mathbf{.67} \pm .01$	$-.23 \pm .04$	$-.19 \pm .04$	$-.26 \pm .03$
ADB	$.65 \pm .01$	$-.05 \pm .13$	$-.01 \pm .12$	$-.08 \pm .13$
RWT	$.66 \pm .01$	$-\mathbf{.02} \pm .04$	$\mathbf{.01} \pm .04$	$-\mathbf{.06} \pm .04$
LFR	$.66 \pm .01$	$-.09 \pm .09$	$-.06 \pm .07$	$-.13 \pm .08$
RLR	$.66 \pm .01$	$-.19 \pm .03$	$-.15 \pm .03$	$-.22 \pm .03$

Fair classification techniques. ADB: adversarial debiasing; RWT: reweighting; LFR: learning fair representation; RLR: regularized logistic regression.
Group fairness metrics. AOD: average odds difference; EOD: equal opportunity difference; SPD: statistical parity difference.

Summary and Discussion

Summaries:

- Individual fairness is a restricted form of robustness: robustness to certain sensitive perturbations.
- Our inferential tools only require black-box access to the ML model, are computationally efficient, and allow auditors to control the false alarm rate and provide asymptotically exact certificates of fairness.

Future directions:

- Continuous sample space $\mathcal{X} \times \mathcal{Y}$;
- Scale invariant for losses.

THE END

Email: sxue@umich.edu