

Using optimal transport to mitigate unfair predictions

Arthur Charpentier, François Hu, Agathe Fernandes-Machado & Philipp Ratz

Centre d'Économie de la Sorbonne – March 2024

CES

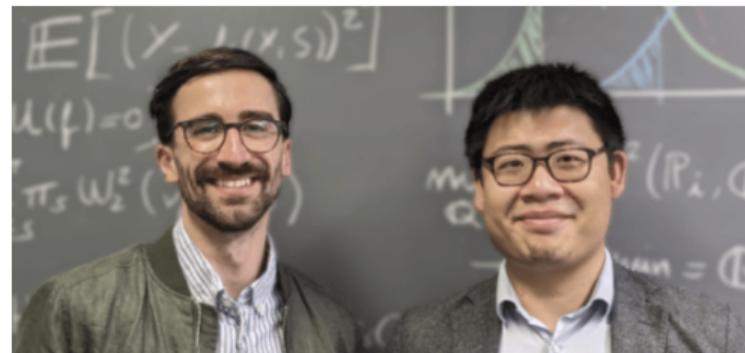
Centre d'Économie de la Sorbonne
UMR 8174

Bio (short)

François Hu Postdoctoral fellow, Université de Montréal

Philipp Ratz PhD Student, Université du Québec à Montréal

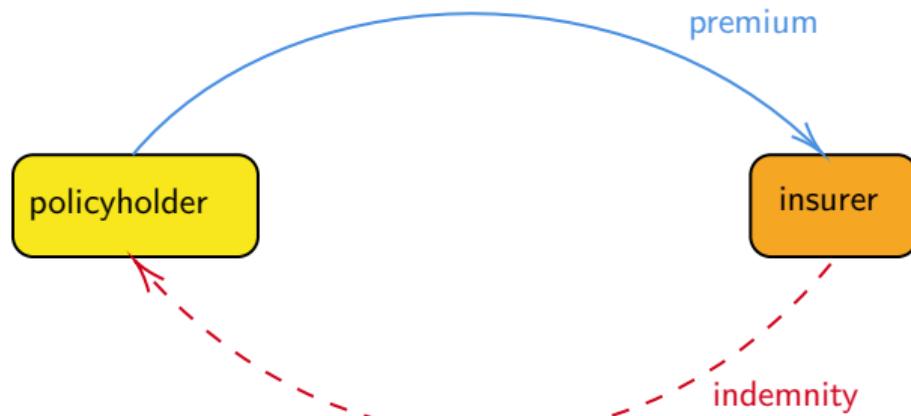
Agathe Fernandes-Machado PhD Student, Université du Québec à Montréal



- › ECML PKDD 2023 & BIAS 2023, Milano
- › AAAI Conference on Artificial Intelligence, Vancouver

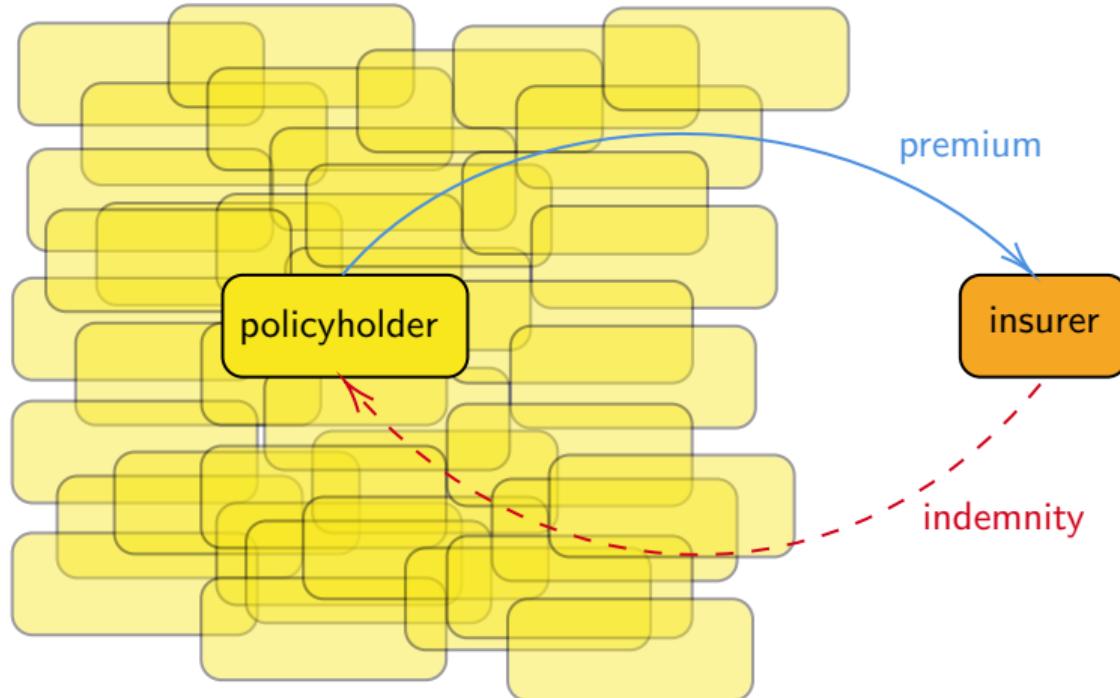
The context of insurance (very quickly)

- Insurance is a risk transfer (from a policyholder to an insurance company)



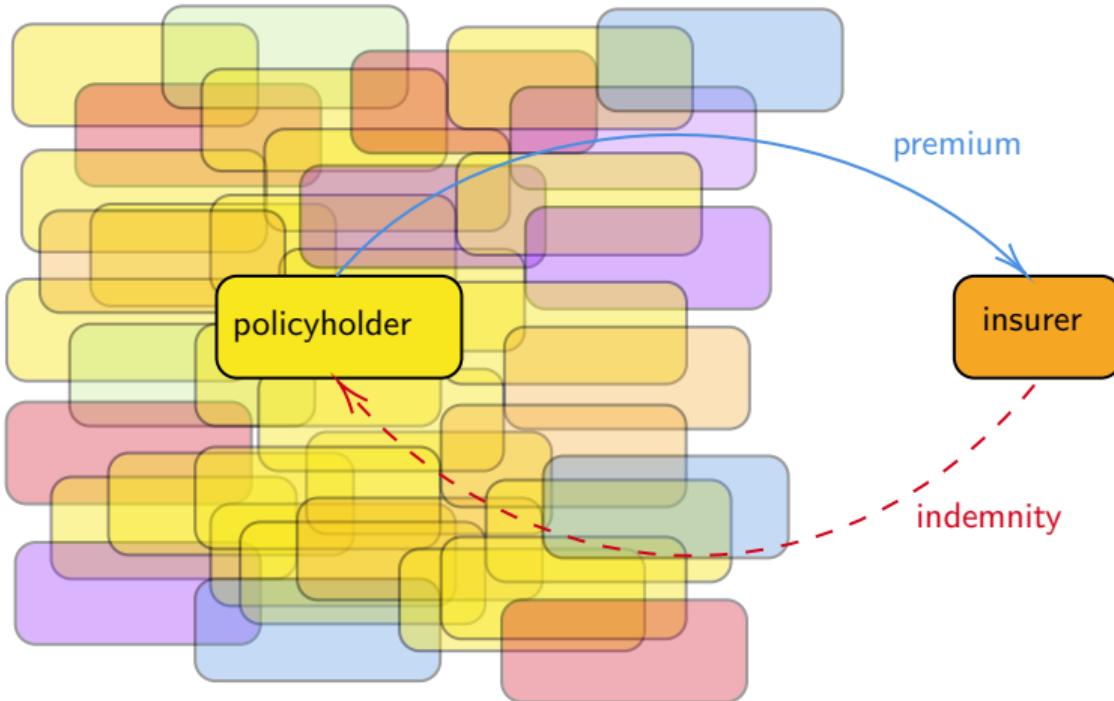
The context of insurance (very quickly)

➤ “*Insurance is the contribution of the many to the misfortune of the few*”



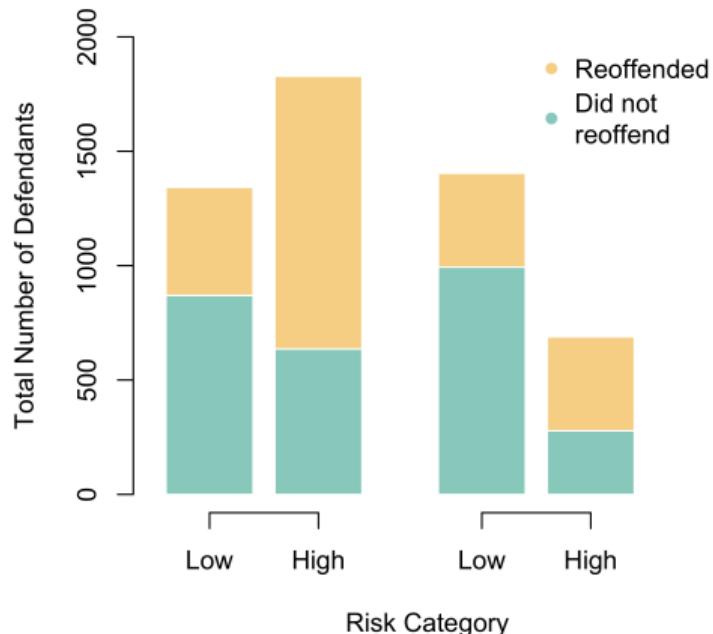
The context of insurance (very quickly)

➤ “*Insurance is the contribution of the many to the misfortune of the few*”



Motivation (1. Propublica, Actuarial Justice)

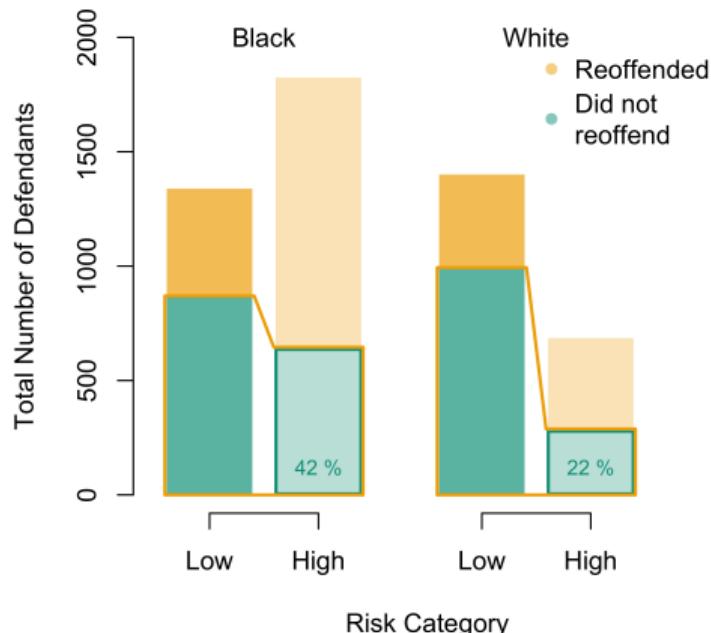
- Concept of “actuarial justice” as coined in Feeley and Simon (1994)
- Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), Perry (2013)



- <https://github.com/propublica/compas-analysis>
- Angwin et al. (2016) Machine Bias
Dressel and Farid (2018)

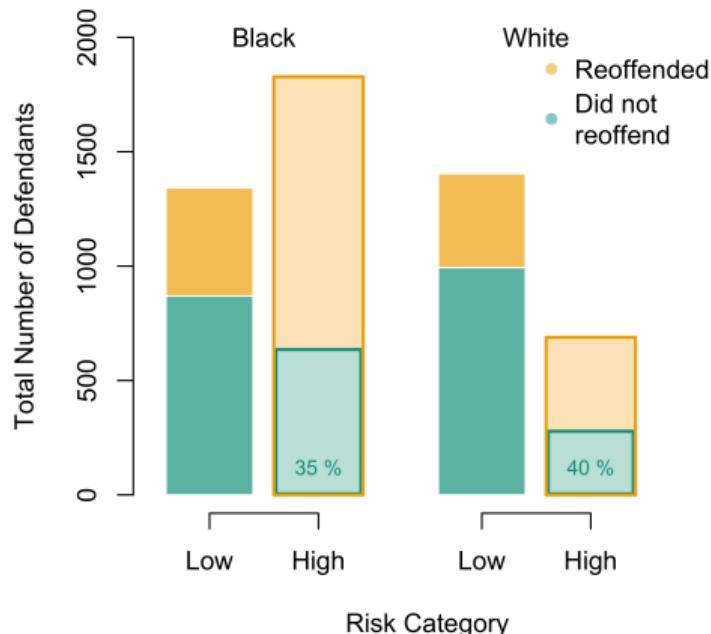
Motivation (1. Propublica, Actuarial Justice)

- From Feller et al. (2016),
 - ▶ for White people, among those who did not re-offend, 22% were wrongly classified,
 - ▶ for Black people, among those who did not re-offend, 42% were wrongly classified,
 - ▶ problem, since $42\% \gg 22\%$



Motivation (1. Propublica, Actuarial Justice)

- From Dieterich et al. (2016),
 - ▶ for White people, among those who were classified as high risk, 40% did not re-offend,
 - ▶ for Black people, among those who were classified as high risk, 35% did not re-offend,
 - ▶ no problem, since $40\% \approx 35\%$



Motivation (2. Legal Aspects)

- EU Directive ([2004/113/EC](#)), 2004 version

– Article 5 (Actuarial factors) –

1. Member States shall ensure that in all new contracts concluded after 21 December 2007 at the latest, **the use of sex as a factor in the calculation of premiums and benefits for the purposes of insurance and related financial services shall not result in differences in individuals' premiums and benefits.**
2. Notwithstanding paragraph 1, Member States may decide before 21 December 2007 to permit proportionate differences in individuals' premiums and benefits where the use of sex is a determining factor in the assessment of risk based on relevant and accurate actuarial and statistical data. The Member States concerned shall inform the Commission and ensure that accurate data relevant to the use of sex as a determining actuarial factor are compiled, published and regularly updated.



Motivation (2. Legal Aspects)

- Au Québec, Charte des droits et libertés de la personne ([C-12](#))

– Article 20.1 –

Dans un **contrat d'assurance** ou de rente, un régime d'avantages sociaux, de retraite, de rentes ou d'assurance ou un régime universel de rentes ou d'assurance, une distinction, exclusion ou préférence fondée sur l'âge, le sexe ou l'état civil est **réputée non discriminatoire lorsque son utilisation est légitime et que le motif qui la fonde constitue un facteur de détermination de risque, basé sur des données actuarielles.**



Motivation (2. Legal Aspects)

- › September 27, 2023, the Colorado Division of Insurance exposed a new proposed regulation entitled [Concerning Quantitative Testing of External Consumer Data and Information Sources, Algorithms, and Predictive Models Used for Life Insurance Underwriting for Unfairly Discriminatory Outcomes](#)
 - Section 5 (Estimating Race and Ethnicity) –

Insurers shall estimate the race or ethnicity of all proposed insureds that have applied for coverage on or after the insurer's initial adoption of the use of ECDIS, or algorithms and predictive models that use ECDIS, including a third party acting on behalf of the insurer that used ECDIS, or algorithms and predictive models that used ECDIS, in the underwriting decision-making process, by utilizing: BIFSG and the insureds' or proposed insureds' name and geolocation (...)

- › [Bayesian Improved First Name Surname Geocoding](#), or “BIFSG”
- › [External Consumer Data and Information Source](#), or “ECDIS”



Motivation (2. Legal Aspects)

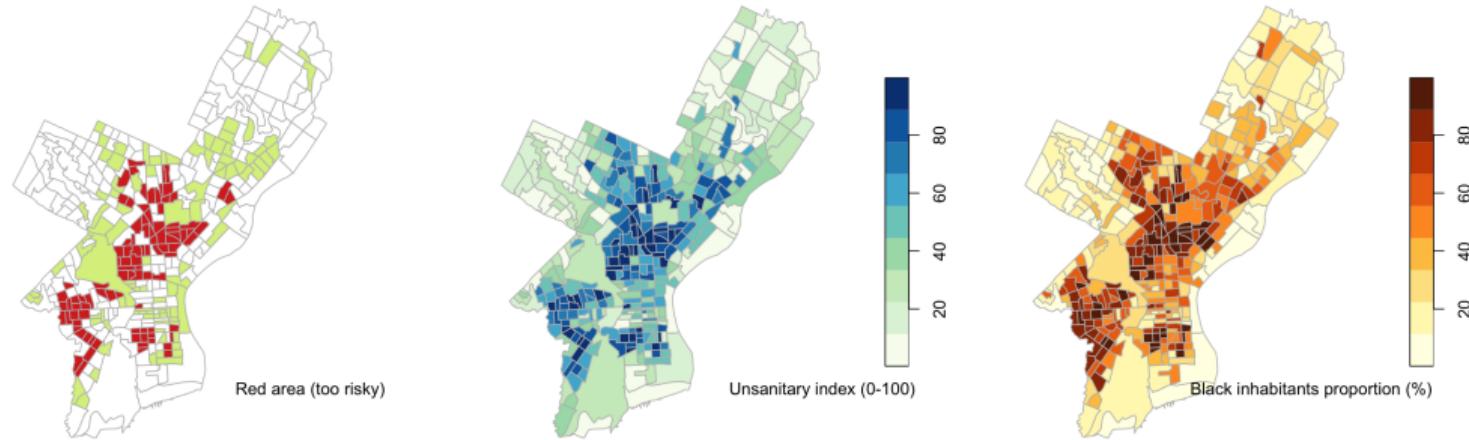
- En France, Loi n° 2008-496 du 27 mai 2008
 - Article 1 –

Constitue une **discrimination indirecte** une disposition, un critère ou une pratique neutre en apparence, mais susceptible d'entraîner, pour l'un des motifs mentionnés au premier alinéa, un désavantage particulier pour des personnes par rapport à d'autres personnes, à moins que cette disposition, ce critère ou cette pratique ne soit objectivement justifié par un but légitime et que les moyens pour réaliser ce but ne soient nécessaires et appropriés.

Extention de la "Loi n° 72-546 du 1 juillet 1972", qui supprima l'exigence de l'intention spécifique.



Motivation (3. Redlining)



(Fictitious maps, inspired by a Home Owners' Loan Corporation map from 1937)

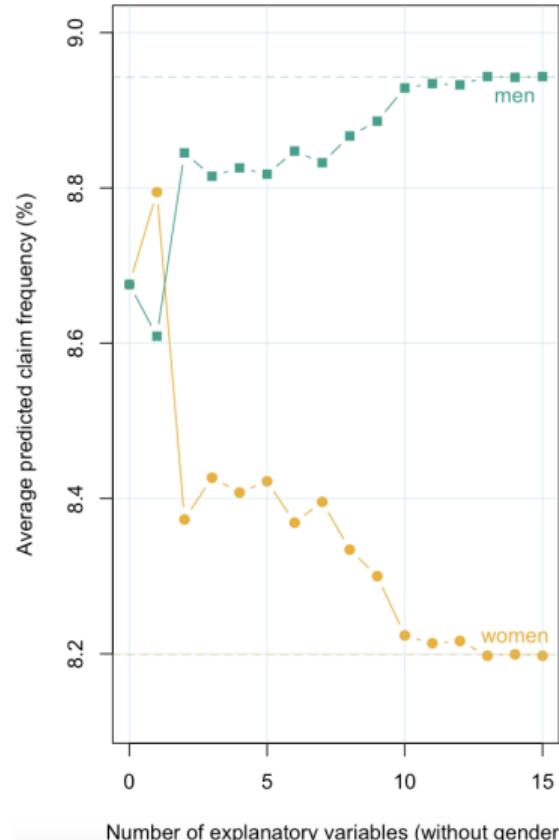
- ▶ Federal Home Loan Bank Board (FHLBB) "*residential security maps*" (for real-estate investments), [Crossney \(2016\)](#) and [Rhynhart \(2020\)](#)
- ▶ Unsanitary index and proportion of Black inhabitants

Motivation (4. Proxies)

- On a French motor dataset, average claim frequencies are 8.94% (men) 8.20% (women).
- Consider some logistic regression to estimate annual claim frequency, on k explanatory variables **excluding gender**.

	men	women
$k = 0$	8.68%	8.68%
$k = 2$	8.85%	8.37%
$k = 8$	8.87%	8.33%
$k = 15$	8.94%	8.20%
empirical	8.94%	8.20%

- See is/ough, Hume (1739)



Discrimination and Insurance

- “*What is unique about insurance is that even statistical discrimination which by definition is absent of any malicious intentions, poses significant moral and legal challenges. Why? Because on the one hand, policy makers would like insurers to treat their insureds equally, without discriminating based on race, gender, age, or other characteristics, even if it makes statistical sense to discriminate (...) On the other hand, at the core of insurance business lies discrimination between risky and non-risky insureds. But riskiness often statistically correlates with the same characteristics policy makers would like to prohibit insurers from taking into account,*” Avraham (2017)
- “*Technology is neither good nor bad; nor is it neutral,*” Kranzberg (1986)
- “*Machine learning won't give you anything like gender neutrality 'for free' that you didn't explicitly ask for,*” Kearns and Roth (2019)

Model

$$\begin{cases} \mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d : \text{'explanatory' variables} \\ s \in \mathcal{S} = \{\text{A, B}\} : \text{"sensitive variable"} \\ y \in \mathcal{Y} \subset \mathbb{R} : \text{regression problem} \\ \hat{y} \in \mathcal{Y} \subset \mathbb{R} : \text{prediction, } \hat{y} = \hat{m}(\mathbf{x}, s) \end{cases}$$

A **loss function** ℓ is a function defined on $\mathcal{Y} \times \mathcal{Y}$ such that $\ell(y, y') \geq 0$ and $\ell(y, y) = 0$, and for a fitted model \hat{m} , its **risk** is

$$\mathcal{R}(\hat{m}) = \mathbb{E}_{\mathbb{P}}[\ell(Y, \hat{m}(\mathbf{X}, S))] = \int \ell(y, \hat{m}(\mathbf{x}, s)) d\mathbb{P}(y, \mathbf{x}, s).$$

Given a dataset $\mathcal{D} = \{(y_i, s_i, \mathbf{x}_i)\}$, define the **empirical risk**

$$\widehat{\mathcal{R}}_n(\hat{m}) = \frac{1}{n} \sum_{i=1}^n \ell(\hat{m}(\mathbf{x}_i, s_i), y_i).$$

Model

- Following Vapnik (1991), the “empirical risk minimization principle” states that the learning algorithm \hat{m}^* is

$$\hat{m}^* = \operatorname{argmin}_{\hat{m} \in \mathcal{M}} \{\hat{\mathcal{R}}_n(\hat{m})\},$$

for some set \mathcal{M} of models.

Example Classical ordinary (linear) least squares are obtained for the quadratic loss $\ell_2(y, \hat{y}) = (y - \hat{y})^2$, and

$$\mathcal{M} = \{m : \mathcal{X} \times \mathcal{S} \rightarrow \mathcal{Y} \text{ such that } m : (\mathbf{x}, s) \mapsto \mathbf{x}^\top \boldsymbol{\beta} + \gamma \mathbf{1}(s = \textcolor{blue}{B})\}$$

Model

For each (\mathbf{x}, s) , choose $m_{\mathbf{x}, s}^*$ that minimizes the conditional expected loss,

$$m_{\mathbf{x}, s}^* \in \operatorname{argmin}_{z \in \mathcal{Y}} \left\{ \int \ell(y, z) d\mathbb{P}_{Y|\mathbf{X}, S}(y|\mathbf{x}, s) \right\},$$

called **Bayes optimal rule**. For the ℓ_2 (quadratic) loss,

$$\mu(\mathbf{x}, s) := m_{\mathbf{x}, s}^* = \mathbb{E}[Y | \mathbf{X} = \mathbf{x}, S = s].$$

Remark In a (binary) classification problem, $\mathcal{Y} = \{\textcolor{teal}{0}, \textcolor{red}{1}\}$,

$$\hat{y} = m_{\mathbf{x}, s}^* = \mathbf{1}(\mu(\mathbf{x}, s) > 1/2) = \begin{cases} \textcolor{red}{1} & \text{if } \mu(\mathbf{x}, s) > 1/2 \\ \textcolor{teal}{0} & \text{if } \mu(\mathbf{x}, s) \leq 1/2 \end{cases}$$

where $\mu(\mathbf{x}, s) = \mathbb{E}[Y | \mathbf{X} = \mathbf{x}, S = s] = \mathbb{P}[Y = 1 | \mathbf{X} = \mathbf{x}, S = s]$, for the misclassification loss $\ell_{0/1}(y, \hat{y}) = \mathbf{1}(y \neq \hat{y})$.

Fairness for Classifiers

$$\begin{cases} \mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d : \text{"explanatory" variables} \\ s \in \{\mathbf{A}, \mathbf{B}\} : \text{"sensitive variable"} \\ y \in \{0, 1\} : \text{classification problem} \\ \hat{y} \in \{0, 1\} : \text{prediction, classically } \hat{y} = \mathbf{1}(m(\mathbf{x}, s) > t) \end{cases}$$

- Following Barocas et al. (2017), standard definitions are

A model m satisfies the **independence property** if $m(\mathbf{X}, S) \perp\!\!\!\perp S$, with respect to the distribution \mathbb{P} of the triplet (\mathbf{X}, S, Y) . \leftarrow DP demographic parity

A model satisfies the **separation property** if $m(\mathbf{X}, S) \perp\!\!\!\perp S | Y$, with respect to the distribution \mathbb{P} of the triplet (\mathbf{X}, S, Y) . \leftarrow EO equalized odds

A model satisfies the **sufficiency property** if $Y \perp\!\!\!\perp S | m(\mathbf{X}, S)$, with respect to the distribution \mathbb{P} of the triplet (\mathbf{X}, S, Y) .

Fairness (Demographic Parity) for Classifiers

- Defining “Demographic Parity”, Corbett-Davies et al. (2017) or Agarwal (2021)

Weak Demographic Parity,

Decision function \hat{y} satisfies weak demographic parity if $\hat{Y} \perp\!\!\!\perp S$, i.e.

$$\mathbb{E}[\hat{Y}|S = A] = \mathbb{E}[\hat{Y}|S = B],$$

or

$$\mathbb{E}[\mathbf{1}(m(\mathbf{X}, S) > t)|S = A] = \mathbb{E}[\mathbf{1}(m(\mathbf{X}, S) > t)|S = B].$$

- One can easily obtain weak Demographic Parity using different thresholds

$$\mathbb{E}[\mathbf{1}(m(\mathbf{X}, S) > t_A)|S = A] = \mathbb{E}[\mathbf{1}(m(\mathbf{X}, S) > t_B)|S = B].$$

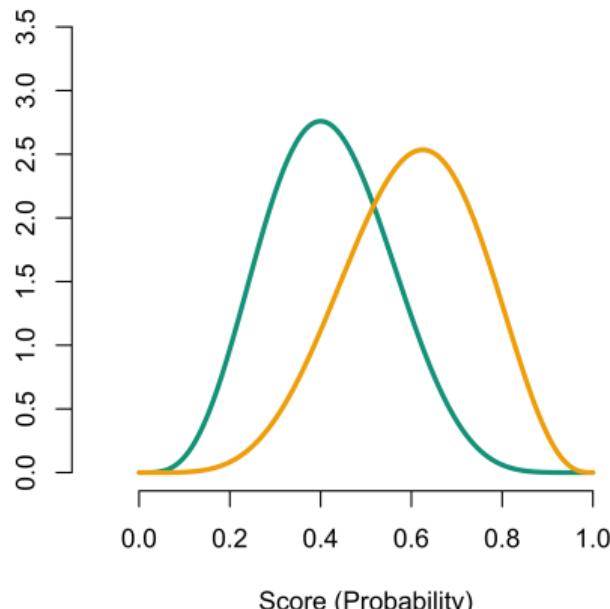
Fairness (Demographic Parity) for Scores

Strong Demographic Parity,

$$\mathbb{E}[\mathbf{1}(m(\mathbf{X}, S) \in \mathcal{E})|S = A] = \mathbb{E}[\mathbf{1}(m(\mathbf{X}, S) \in \mathcal{E})|S = B],$$

for any $\mathcal{E} \subset [0, 1]$, or $\mathbb{P}_A[\mathcal{E}] = \mathbb{P}_B[\mathcal{E}]$,

$$\begin{cases} \mathbb{P}_A[\mathcal{E}] = \mathbb{P}[m(\mathbf{X}, S) \in \mathcal{E}|S = A] \\ \mathbb{P}_B[\mathcal{E}] = \mathbb{P}[m(\mathbf{X}, S) \in \mathcal{E}|S = B] \end{cases}$$



Fairness (Demographic Parity) for Scores

- Use some "distance" between \mathbb{P}_A and \mathbb{P}_B
(TV, KL, or Wasserstein)

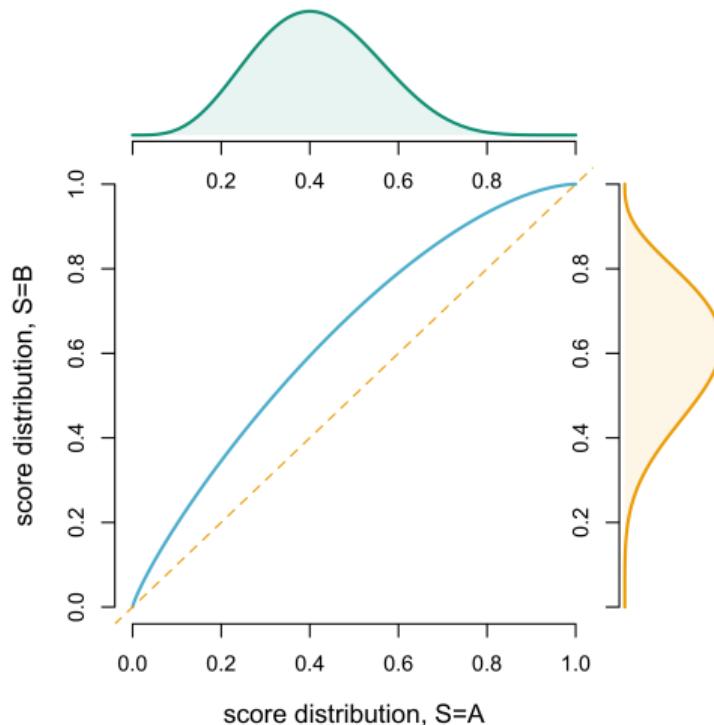
$$\inf_{\pi \in \Pi(\mathbb{P}_A, \mathbb{P}_B)} \{ \mathbb{E}[\ell(X, Y)], (X, Y) \sim \pi \}$$

or

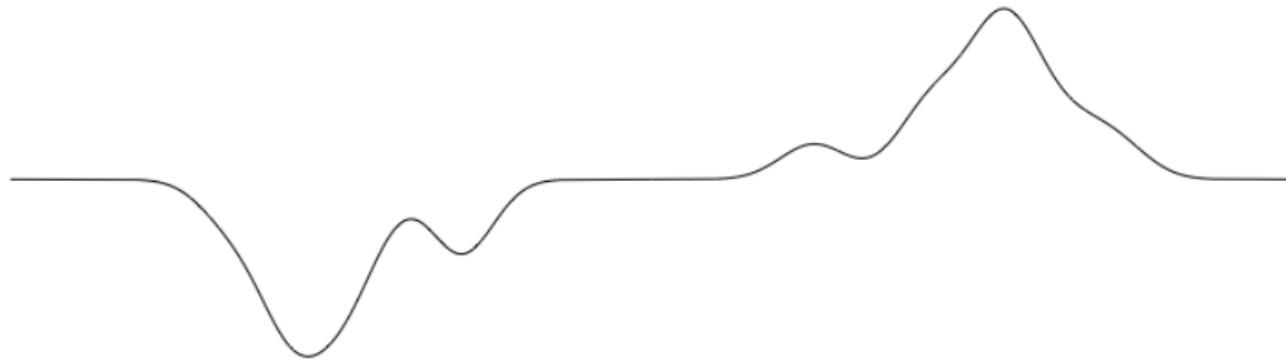
$$\inf_{\pi \in \Pi(\mathbb{P}_A, \mathbb{P}_B)} \left\{ \int \ell(x, y) \pi(dx, dy) \right\}.$$

or using a transport mapping \mathcal{T}

$$\inf_{\mathcal{T}: \mathcal{T}_{\#}\mathbb{P}_A = \mathbb{P}_B} \left\{ \int \ell(x, \mathcal{T}(x)) d\mathbb{P}_A(x) \right\}.$$



Fairness and Optimal Transport



Monge (1781), Mémoire sur la théorie des déblais et des remblais

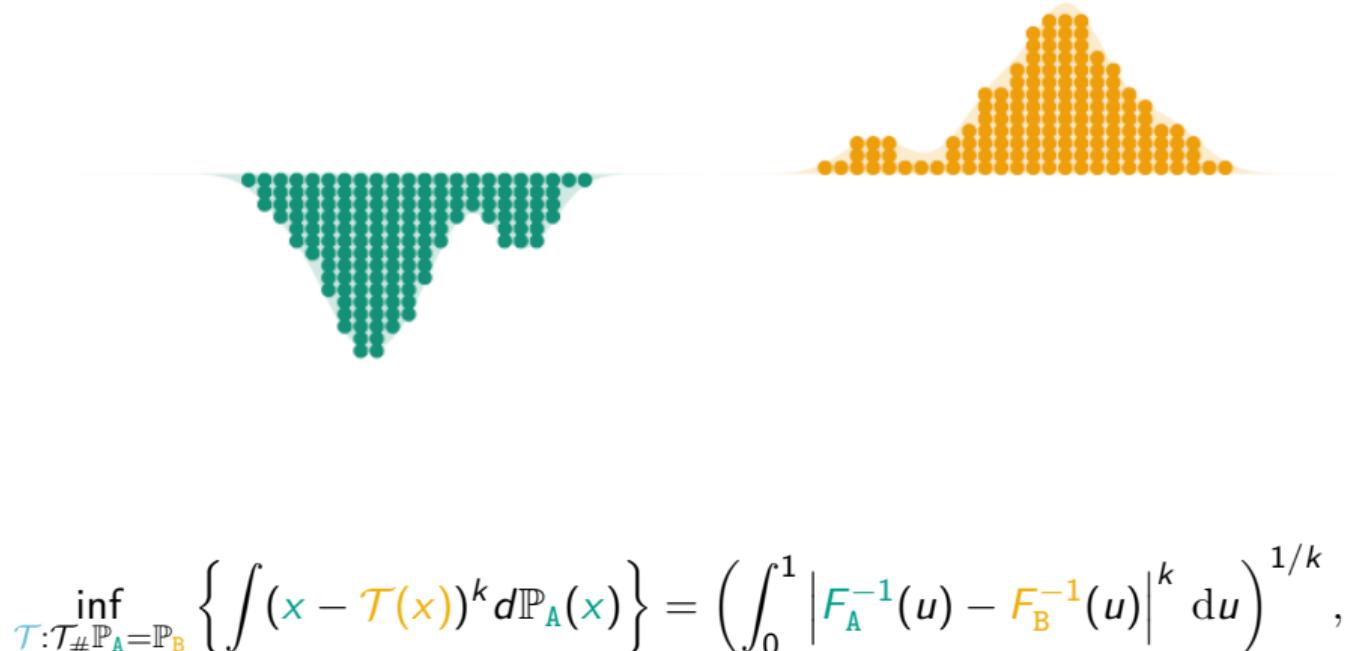
We want to **transport** optimally sand from a **hole** (with shape $-d\mathbb{P}_A$) to a **pile** (with shape $d\mathbb{P}_B$). "*Rien ne se perd, rien ne se crée, tout se transporte*": $\int d\mathbb{P}_A = \int d\mathbb{P}_B$.

Fairness and Optimal Transport

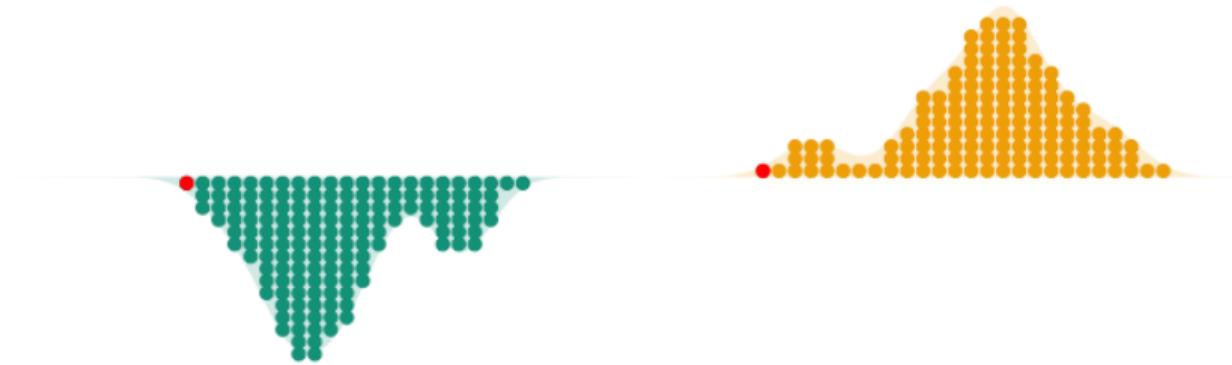


$$\inf_{\mathcal{T}: \mathcal{T}_{\#}\mathbb{P}_A = \mathbb{P}_B} \left\{ \int (\textcolor{teal}{x} - \mathcal{T}(x))^k d\mathbb{P}_A(x) \right\} = \left(\int_0^1 |F_0^{-1}(u) - F_1^{-1}(u)|^k du \right)^{1/k},$$

Fairness and Optimal Transport



Fairness and Optimal Transport



Optimal transport plan is here $\mathcal{T}^* : x \mapsto y = F_B^{-1} \circ F_A(x)$ (increasing function)

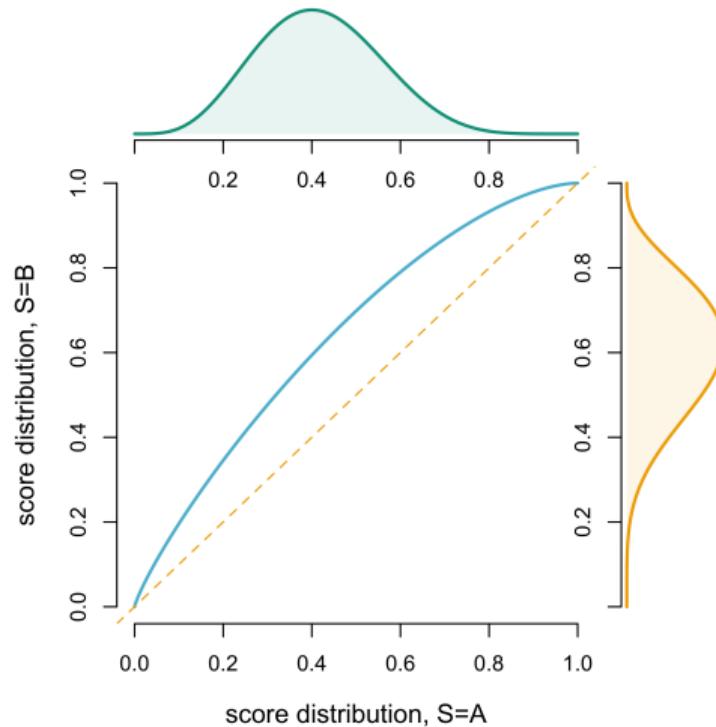
Counterfactual Fairness (and Optimal Transport)

- Can be used to quantify unfairness,

m satisfies **Strong Demographic Parity**
if $W_2 = 0$,

$$W_2 = \left(\int_0^1 \left(F_A^{-1}(u) - F_B^{-1}(u) \right)^2 du \right)^{1/2}$$

(optimal transport for \mathbb{R} -valued measures)



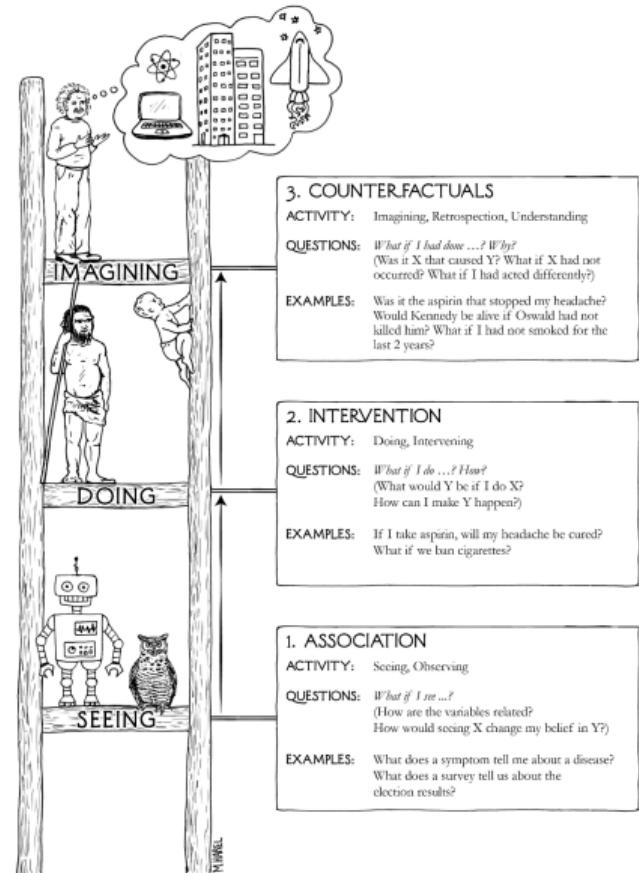
Counterfactual Fairness (and Optimal Transport)

“Ladder of causation” from Pearl et al. (2009)

- **3. Counterfactuals**
(Imagining, “*what if I had done...*”)
- **2. Intervention**
(Doing, “*what if I do...*”)
- **1. Association**
(Seeing, “*what if I see...*”)

Picture source: Pearl and Mackenzie (2018)

What would be the impact of a treatment T on a variable of interest Y ?



Counterfactual Fairness (and Optimal Transport)

- › Define individual or counterfactual fairness, Castelnovo et al. (2022)
"Individual fairness is embodied in the following principle: similar individuals should be given similar decisions. This principle deals with the comparison of single individuals rather than focusing on groups of people sharing some characteristics."
- › Following Kusner et al. (2017)

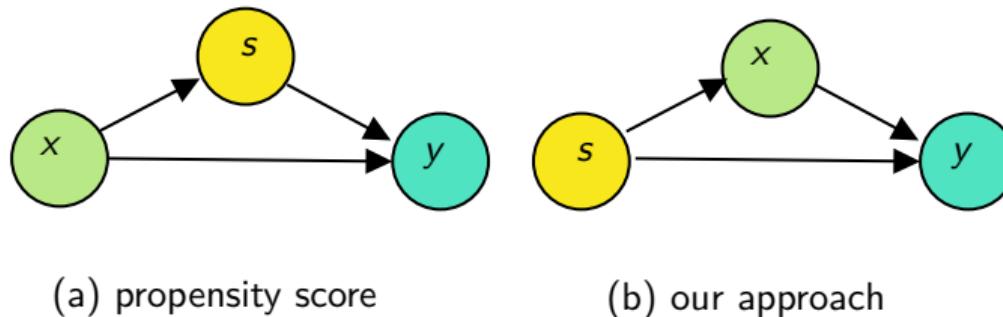
A decision is **counterfactually fair** if the prediction in the real world is the same as the prediction in the counterfactual world

$$\mathbb{E}[Y_{S \leftarrow A}^* | \mathbf{X} = \mathbf{x}] = \mathbb{E}[Y_{S \leftarrow B}^* | \mathbf{X} = \mathbf{x}], \forall \mathbf{x},$$

where $Y_{S \leftarrow A}^*$ and $Y_{S \leftarrow B}^*$ denote "potential outcomes".

- › since we use the same \mathbf{x} it is a **ceteris paribus counterfactual**.
(is the counterfactual of a man with height 190 cm a woman with height 190 cm ?)

Counterfactual Fairness (and Optimal Transport)



- Charpentier et al. (2023a) defined **mutatis mutandis** counterfactual fairness,

$$\mathbb{E}[Y_{S \leftarrow A}^* | X = x] = \mathbb{E}[Y_{S \leftarrow B}^* | X = T^*(x)], \forall x.$$

(probability to get surgery when delivering a baby for Black / non-Black mother)

Back to Actuarial Justice

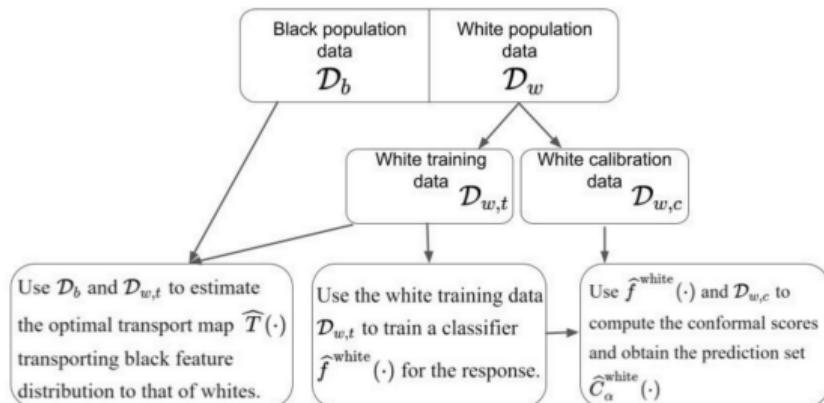
➤ See Berk et al. (2021)

Improving Fairness in Criminal Justice Algorithmic Risk Assessments Using Optimal Transport and Conformal Prediction Sets*

Richard A. Berk
University of Pennsylvania
Arun Kumar Kuchibhotla
Carnegie Mellon University
Eric Tchetgen Tchetgen
University of Pennsylvania

Abstract

In the United States and elsewhere, risk assessment algorithms are being used to help inform criminal justice decision-makers. A common intent is to forecast an offender's "future dangerousness." Such algorithms have been correctly criticized for potential unfairness, and there is an active cottage industry trying to make repairs. In this paper, we use counterfactual reasoning to consider the prospects for improved fairness when members of a disadvantaged class are treated by a risk algorithm as if they are members of an advantaged class. We combine a machine learning classifier trained in a novel manner with an optimal transport adjustment for the relevant joint probability distributions, which together provide a constructive response to claims of bias-in-bias-out. A key distinction is made between fairness claims that are empirically testable and fairness claims that are not. We then use confusion tables and conformal prediction sets to evaluate achieved fairness for estimated risk. Our data are a random sample of 300,000 offenders at their arraignments for a large metropolitan area in the United States during which decisions to release or detain are made. We show that substantial improvement in fairness can be achieved consistent with a Pareto improvement for legally protected classes.



*Cary Coglianese and Sandra Mayson provided many insightful suggestions for legal conceptions of fairness and the prospect for criminal justice reform. Emanuele Candès offered several very instructive insights when commenting on this work at the Stanford/Berkeley Online Causal Inference Seminar. We also received very helpful feedback from a group of researchers at MIT and Harvard who work on causal inference. In that regard, a special thanks go to Devavrat Shah. Thanks also go to three thoughtful reviewers.

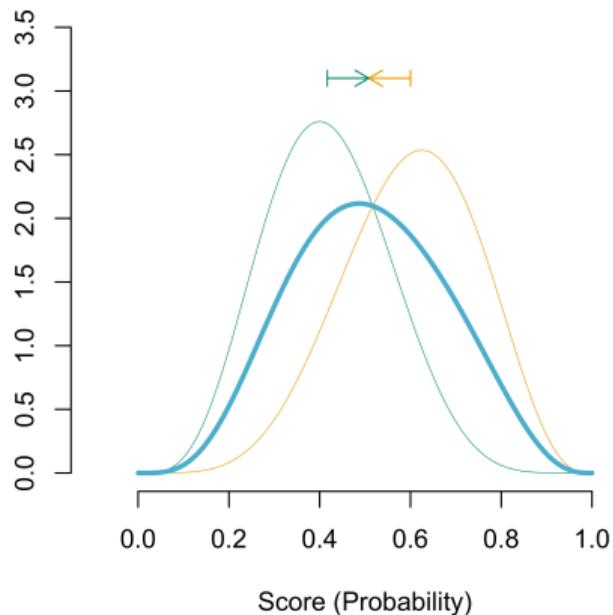
Mitigation with Wasserstein Barycenter

- › If $W_2 \neq 0$ can we mitigate discrimination ?
- › Use of Wasserstein Barycenter
see Charpentier et al. (2023b)
- › In Euclidean spaces

$$\mathbf{z}^* = \operatorname{argmin}_{\mathbf{z}} \left\{ \sum_{i=1}^n \omega_i d(\mathbf{z}, \mathbf{z}_i)^2 \right\},$$

- › For probability measures

$$\mathbb{P}^* = \operatorname{argmin}_{\mathbb{Q}} \left\{ \sum_{i=1}^n \omega_i d(\mathbb{Q}, \mathbb{P}_i)^2 \right\},$$



We have defined the risk of a model $m \in \mathcal{M}$ as $\mathcal{R}(m) = \mathbb{E}[\ell(Y, m(\mathbf{X}))]$.

Mitigation with Wasserstein Barycenter

- › Define the classes of fair models,

$$\begin{cases} \mathcal{M}_{\text{DP}} = \{m \in \mathcal{M} \text{ s.t. } m(\mathbf{X}) \perp\!\!\!\perp S\} \\ \mathcal{M}_{\text{EO}} = \{m \in \mathcal{M} \text{ s.t. } m(\mathbf{X}) \perp\!\!\!\perp S \mid Y\} \end{cases}$$

- › Fairness is achieved by projection onto a fair subspace

$$\hat{m}_{\text{fair}} \in \underset{m \in \mathcal{M}_{\text{fair}}}{\operatorname{argmin}} \{\hat{\mathcal{R}}_n(m)\}$$

Given a risk \mathcal{R} , a class \mathcal{M} and the fair-subclass $\mathcal{M}_{\text{fair}}$, the **price of fairness**

$$\mathcal{E}_{\text{fair}}(\mathcal{M}) = \min_{m \in \mathcal{M}_{\text{fair}}} \{\mathcal{R}(m)\} - \min_{m \in \mathcal{M}} \{\mathcal{R}(m)\}.$$

Mitigation with Wasserstein Barycenter

Recall that Bayes estimator is the best model, for the ℓ_2 loss,

$$\mu(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}] \text{ and set } \begin{cases} \mu_{\textcolor{teal}{A}}(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}, S = \textcolor{teal}{A}] \\ \mu_{\textcolor{blue}{B}}(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}, S = \textcolor{blue}{B}] \end{cases}$$

From the definition of Wasserstein distance,

$$W_2(p, q) = \left(\inf_{\pi \in \Pi(p, q)} \int |x - y|^2 d\pi(x, y) \right)^{1/2}$$

Thus,

$$\mathbb{E}[|m(\mathbf{X}, S) - \mu_S(\mathbf{X})|^2 | S = s] \geq W_2(\mathbb{P}_m, \mathbb{P}_s)^2$$

Mitigation with Wasserstein Barycenter

Price of fairness and Wasserstein Barycenter

$$\mathcal{E}_{\text{fair}}(\mathcal{M}) = \min_{m \in \mathcal{M}_{\text{fair}}} \{\mathcal{R}(m)\} - \min_{m \in \mathcal{M}} \{\mathcal{R}(m)\} \geq \min_{g \in \mathcal{M}} \{ \mathbb{E} \left(W_2(\mathbb{P}_S, \mathbb{P}_{S,g})^2 \right) \}$$

where \mathbb{P}_S is the condition distribution of $\mu(\mathbf{X}, S)$, given S , and $\mathbb{P}_{S,g}$ is the condition distribution of $g(\mathbf{X}, S)$, given S . Moreover, if $\mathcal{M}_{\text{fair}} = \mathcal{M}_{\text{DP}}$, and if \mathbb{P}_s is absolutely continuous (w.r.t. Lebesgue measure),

$$\mathcal{E}_{\text{DP}}(\mathcal{M}) = \min_{g \in \mathcal{M}} \{ \mathbb{E} \left(W_2(\mathbb{P}_S, \mathbb{P}_{S,g})^2 \right) \} = \min_{g \in \mathcal{M}} \left\{ \sum_s \mathbb{P}[S = s] \cdot W_2(\mathbb{P}_s, \mathbb{P}_{s,g})^2 \right\}$$

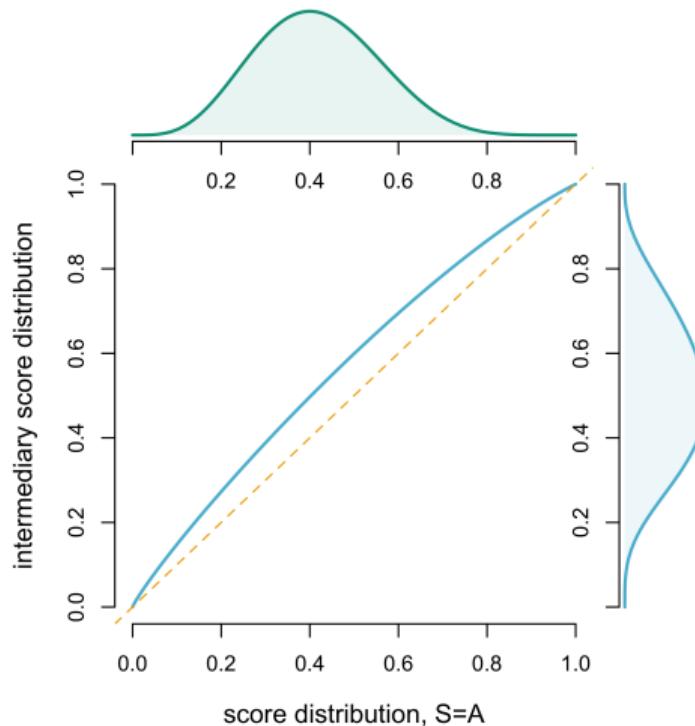
See [Gouic et al. \(2020\)](#) for a complete proof.

We recognize on the right the barycenter, with weights $\mathbb{P}[S = s]$ and distance W_2 .

Mitigation with Wasserstein Barycenter

Given scores $m(\mathbf{x}, s = \text{A})$ and $m(\mathbf{x}, s = \text{B})$,
the “fair barycenter score” is

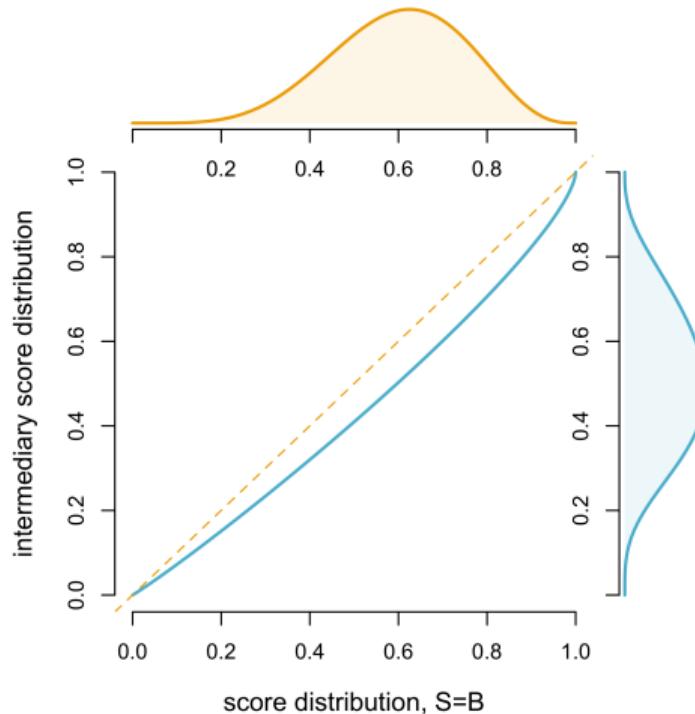
$$\begin{aligned} & m^*(\mathbf{x}, s = \text{A}) \\ = & \mathbb{P}[S = \text{A}] \cdot m(\mathbf{x}, s = \text{A}) \\ + & \mathbb{P}[S = \text{B}] \cdot F_{\text{B}}^{-1} \circ F_{\text{A}}(m(\mathbf{x}, s = \text{A})) \end{aligned}$$



Mitigation with Wasserstein Barycenter

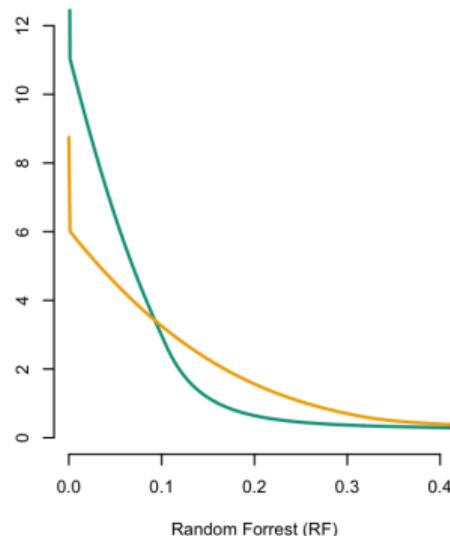
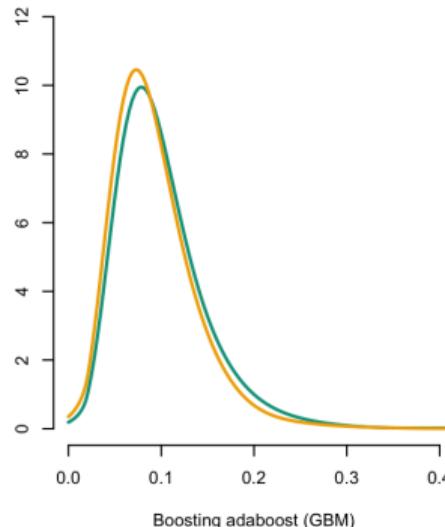
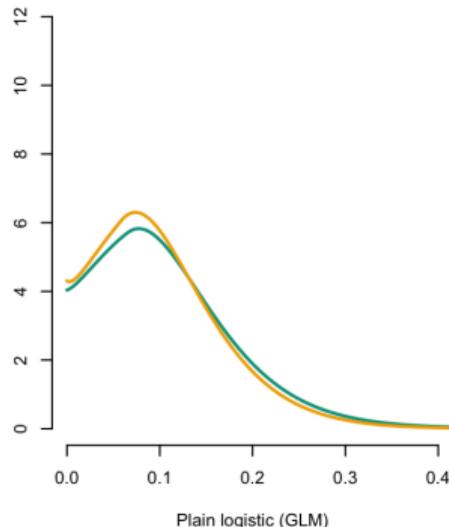
Given scores $m(\mathbf{x}, s = \text{A})$ and $m(\mathbf{x}, s = \text{B})$,
the “fair barycenter score” is

$$\begin{aligned} & m^*(\mathbf{x}, s = \text{B}) \\ = & \mathbb{P}[S = \text{A}] \cdot F_{\text{A}}^{-1} \circ F_{\text{B}}(m(\mathbf{x}, s = \text{B})) \\ + & \mathbb{P}[S = \text{B}] \cdot m(\mathbf{x}, s = \text{B}) \end{aligned}$$



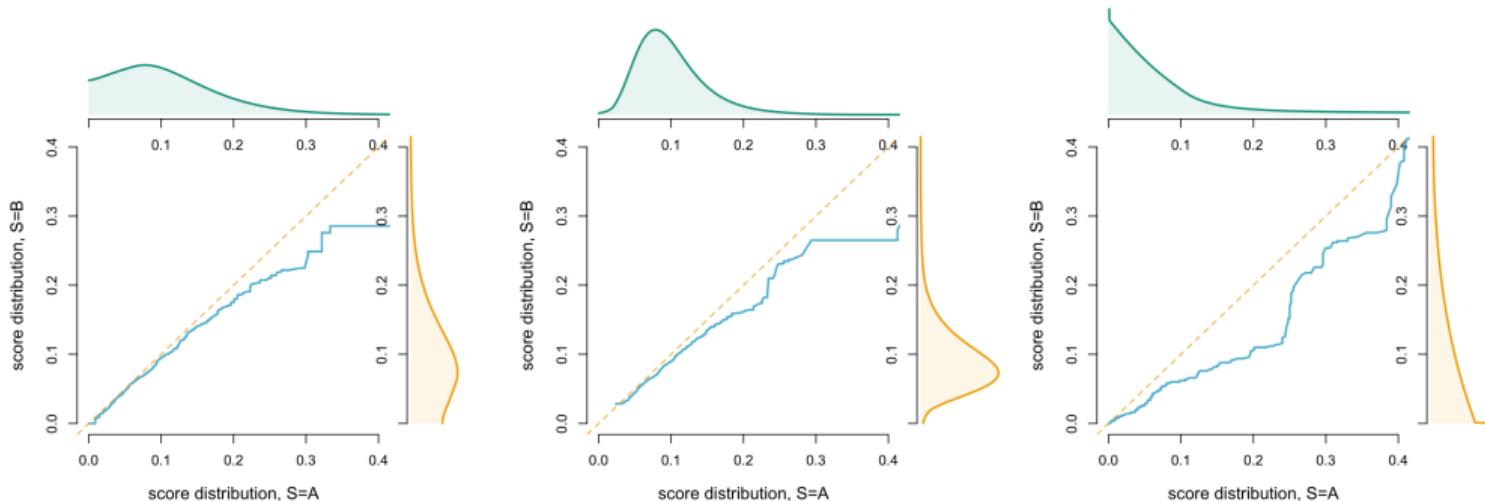
Mitigation with Wasserstein Barycenter

- › If the two models are balanced, m^* is also balanced.
- › Annual claim occurrence (motor insurance, Charpentier et al. (2023b))
- › Three models (plain GLM, GBM, Random Forest)



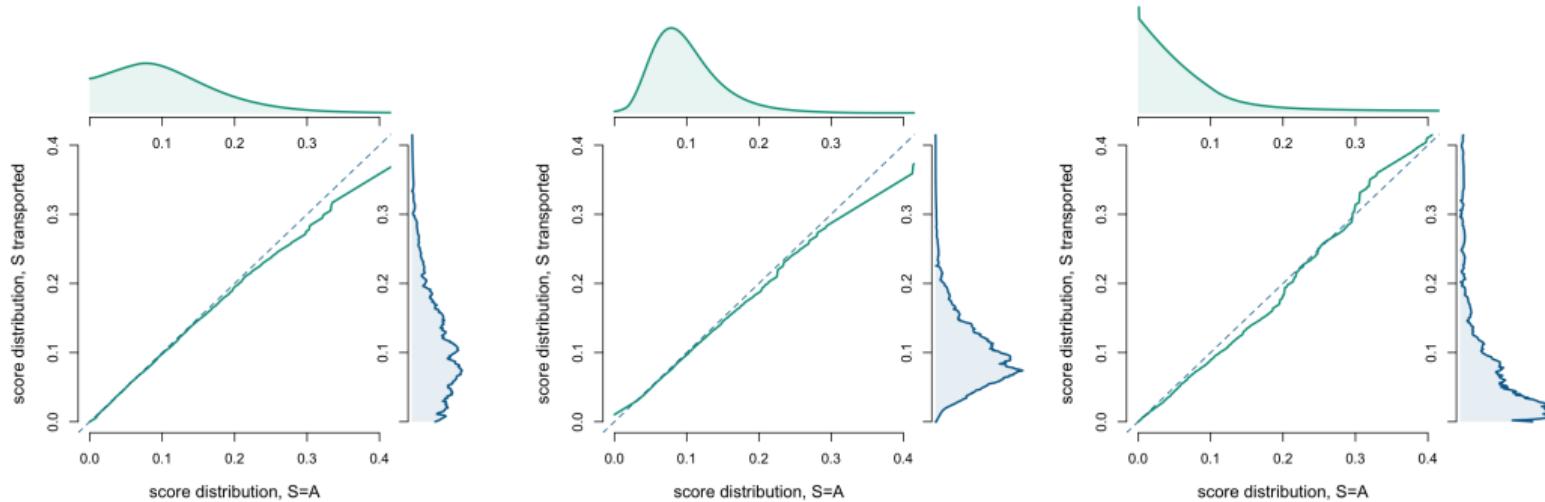
Mitigation with Wasserstein Barycenter

- › Predictions are different for men ($= A$) and women ($S = B$)



- › since $W_2 \neq 0$ consider post processing mitigation

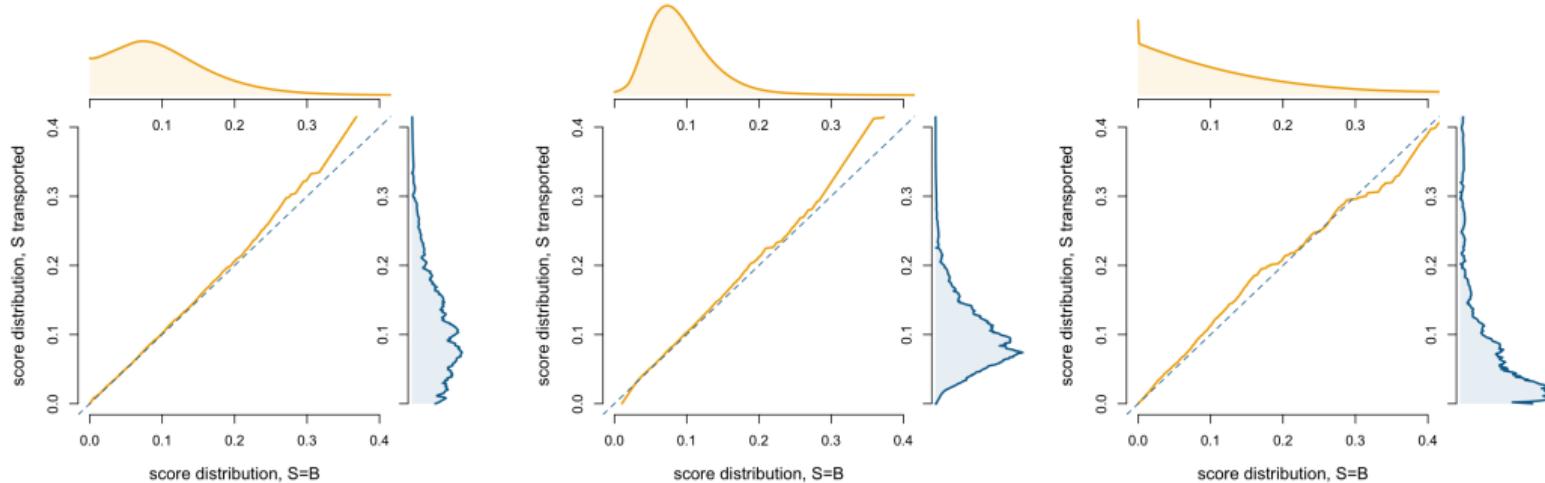
Mitigation with Wasserstein Barycenter



- Given scores $m(\mathbf{x}, s = \text{A})$ and $m(\mathbf{x}, s = \text{B})$, the “fair barycenter score” is

$$m^*(\mathbf{x}, s = \text{A}) = \mathbb{P}[S = \text{A}] \cdot m(\mathbf{x}, s = \text{A}) + \mathbb{P}[S = \text{B}] \cdot F_{\text{B}}^{-1} \circ F_{\text{A}}(m(\mathbf{x}, s = \text{A}))$$

Mitigation with Wasserstein Barycenter

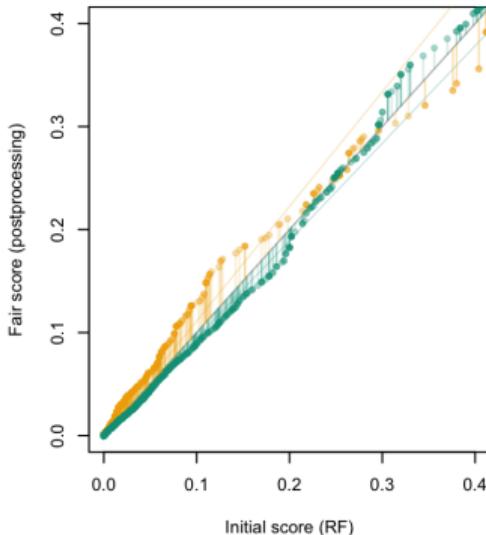
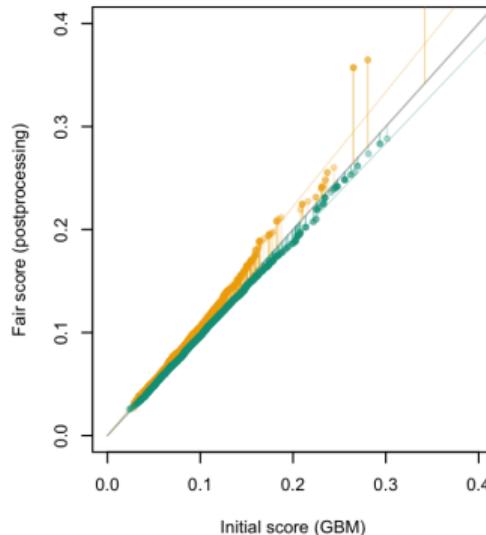
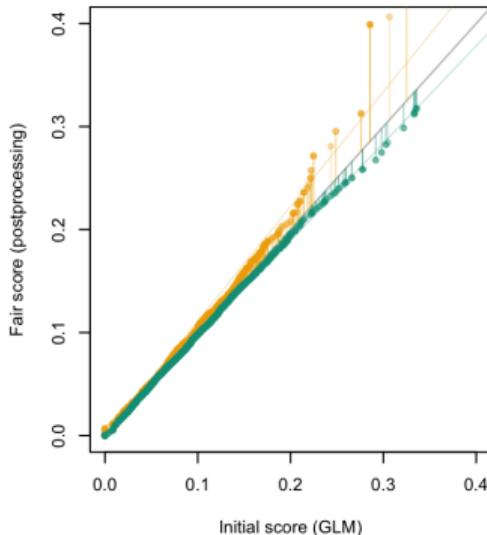


- Given scores $m(\mathbf{x}, s = \textcolor{teal}{A})$ and $m(\mathbf{x}, s = \textcolor{blue}{B})$, the “fair barycenter score” is

$$m^*(\mathbf{x}, s = \textcolor{blue}{B}) = \mathbb{P}[S = \textcolor{teal}{A}] \cdot F_{\textcolor{teal}{A}}^{-1} \circ F_{\textcolor{blue}{B}}(m(\mathbf{x}, s = \textcolor{blue}{B})) + \mathbb{P}[S = \textcolor{blue}{B}] \cdot m(\mathbf{x}, s = \textcolor{blue}{B})$$

Mitigation with Wasserstein Barycenter

- › We can plot $\{(m(\mathbf{x}_i, \mathbb{A}), m^*(\mathbf{x}_i, \mathbb{A})\}$ and $\{(m(\mathbf{x}_i, \mathbb{B}), m^*(\mathbf{x}_i, \mathbb{B})\}$



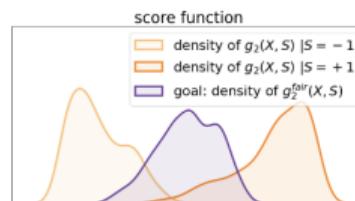
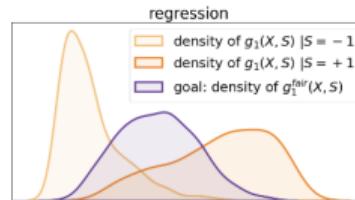
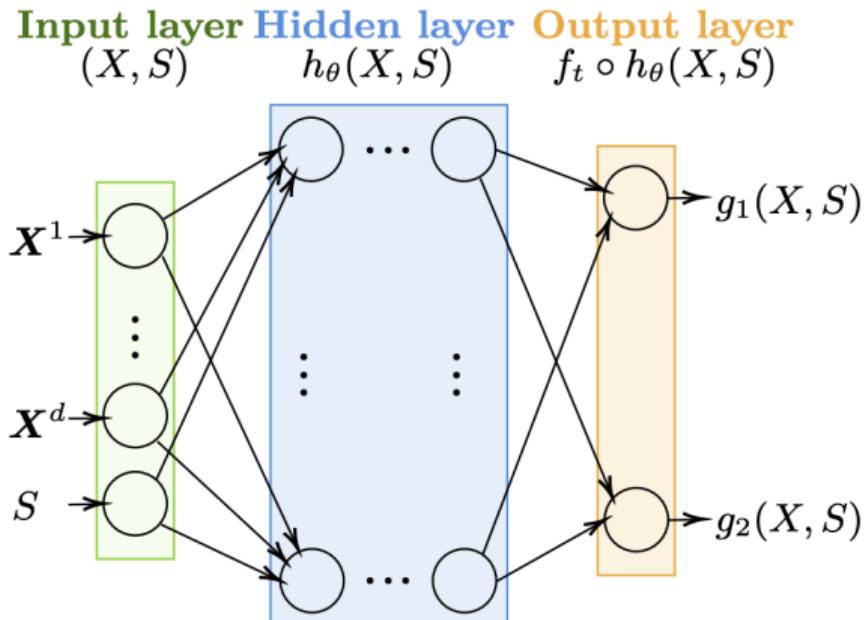
Mitigation with Wasserstein Barycenter

- › Numerical values, for initial occurrence probability of 5%, 10% and 20%, we have

	A (men)				B (women)			
	$\times 0.94$	GLM	GBM	RF	$\times 1.11$	GLM	GBM	RF
$m(\mathbf{x}) = 5\%$	4.73%	4.94%	4.80%	4.42%	5.56%	5.16%	5.25%	6.15%
$m(\mathbf{x}) = 10\%$	9.46%	9.83%	9.66%	8.92%	11.12%	10.38%	10.49%	12.80%
$m(\mathbf{x}) = 20\%$	18.91%	19.50%	18.68%	18.26%	22.25%	20.77%	21.63%	21.12%

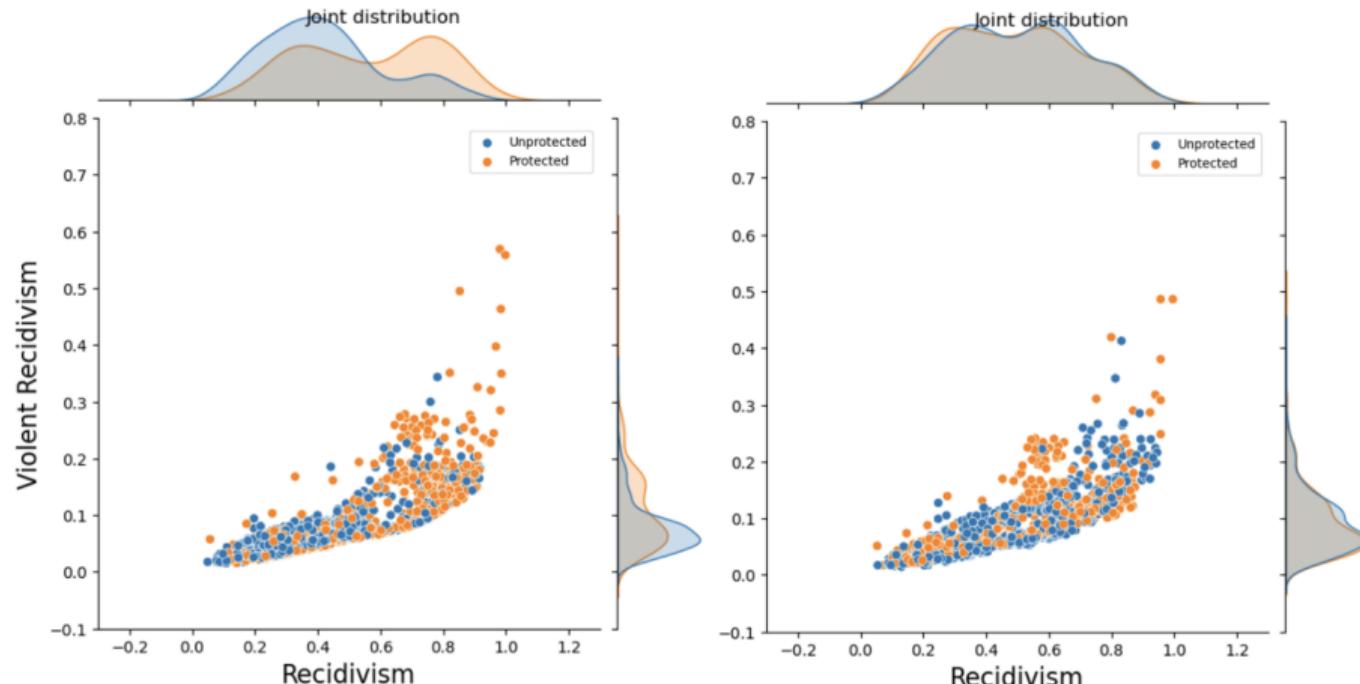
- › Recent work on the use of Wasserstein Barycenter, in [Charpentier et al. \(2023b\)](#) and [Hu et al. \(2023a,b,c\)](#), and optimal transport for counterfactual fairness in [Charpentier et al. \(2023a\)](#).

Recent work



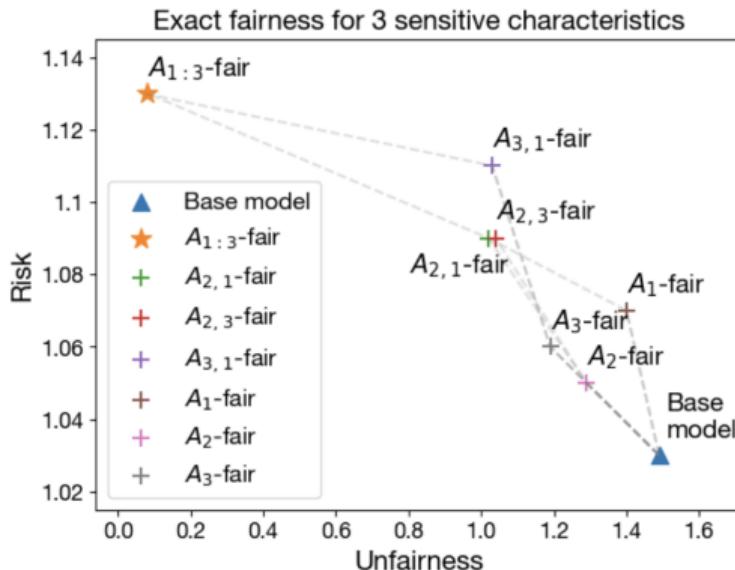
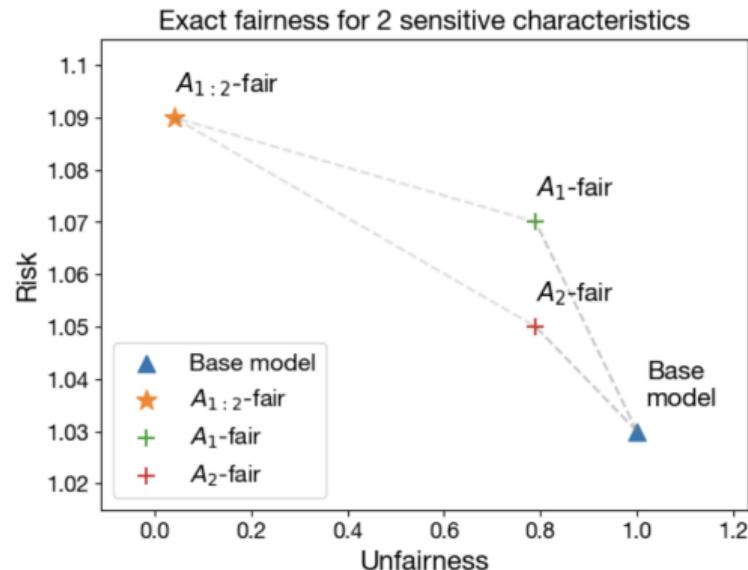
- Fairness in Multi-Task Learning via Wasserstein Barycenters, [Hu et al. \(2023a\)](#)

Recent work



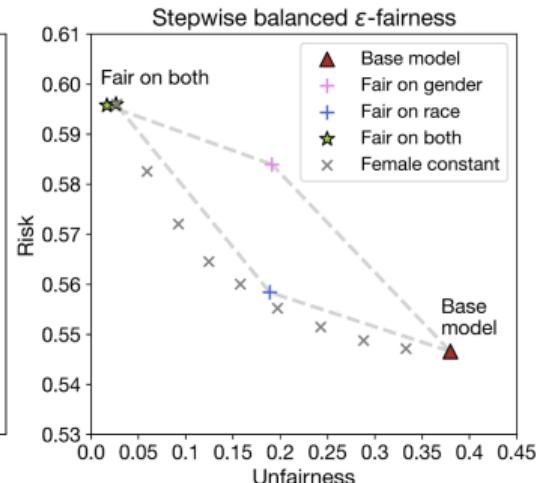
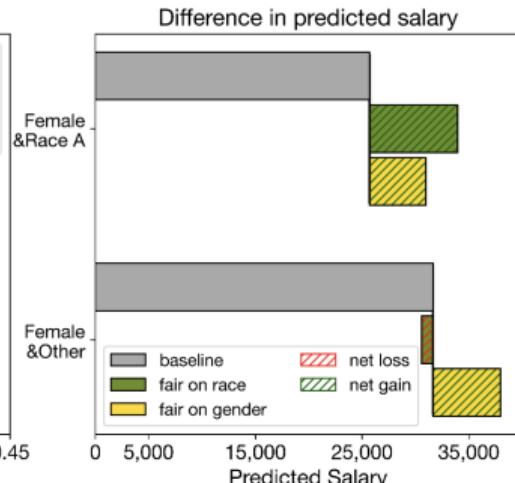
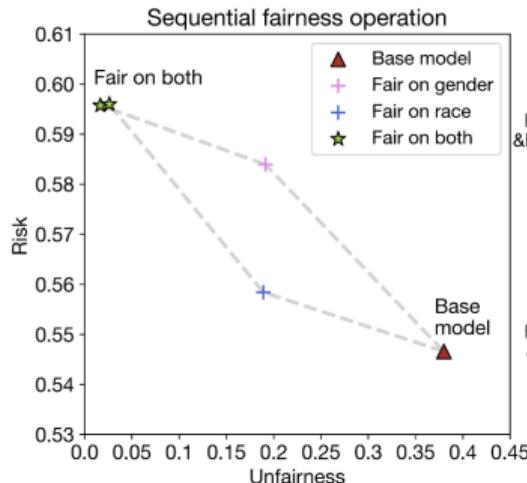
- Fairness in Multi-Task Learning via Wasserstein Barycenters, [Hu et al. \(2023a\)](#)

Recent work



- A Sequentially Fair Mechanism for Multiple Sensitive Attributes, Hu et al. (2023c)

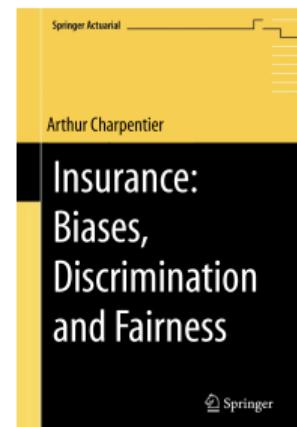
Recent work



- A Sequentially Fair Mechanism for Multiple Sensitive Attributes, Hu et al. (2023c)

Mitigation ? (brief conclusion)

- If it is mandatory to mitigate, there are robust techniques that can guarantee fairness
- Supreme Court Justice Harry Blackmun stated, in 1978,
“In order to get beyond racism, we must first take account of race. There is no other way. And in order to treat some persons equally, we must treat them differently.”
Knowlton (1978), cited in Lippert-Rasmussen (2020)
- In 2007, John G. Roberts of the U.S. Supreme Court submits
“The way to stop discrimination on the basis of race is to stop discriminating on the basis of race”
Sabbagh (2007) and Turner (2015)
- To go further,
Charpentier (2024) Insurance: Biases, Discrimination and Fairness.



Mitigation ? (brief conclusion)



Institut intelligence et données (IID)

ULaval nouvelles



EN

À propos

Expertises

Services

Projets

Actualités

Évènements

Journée sur l'équité et la discrimination en assurance 2024

16 mai 2024, 8h30 à 17h | En mode hybride
En direct de l'Université Laval, pavillon Alexandre-Vachon + Zoom

Réservez la date!

Ouverture des inscriptions à venir.

Workshop on Fairness and Discrimination in Insurance 2024

May 16th 2024, 8:30AM – 5PM | Hybrid Event
Live from Université Laval, pavillon Alexandre-Vachon + Zoom

Save the date!

Registrations opening soon.

Workshop on Fairness and Discrimination in Insurance

Journée sur l'équité et la discrimination en assurance

May 16th 2024,
8:30AM - 5PM
Hybrid Event

16 mai 2024,
8h30 à 17h
Événement hybride



References

- Agarwal, S. (2021). Trade-offs between fairness and interpretability in machine learning. In *IJCAI 2021 Workshop on AI for Social Good*.
- Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals and it's biased against blacks. *ProPublica*, May 23.
- Avraham, R. (2017). Discrimination and insurance. In Lippert-Rasmussen, K., editor, *Handbook of the Ethics of Discrimination*, pages 335–347. Routledge.
- Barocas, S., Hardt, M., and Narayanan, A. (2017). Fairness in machine learning. *Nips tutorial*, 1:2017.
- Berk, R. A., Kuchibhotla, A. K., and Tchetgen, E. T. (2021). Improving fairness in criminal justice algorithmic risk assessments using optimal transport and conformal prediction sets. *arXiv*, 2111.09211.
- Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., and Cosentini, A. C. (2022). A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12(1):1–21.
- Charpentier, A. (2024). *Insurance: biases, discrimination and fairness*. Springer Verlag.
- Charpentier, A., Flachaire, E., and Gallic, E. (2023a). Causal inference with optimal transport. In Thach, N. N., Kreinovich, V., Ha, D. T., and Trung, N. D., editors, *Optimal Transport Statistics for Economics and Related Topics*. Springer Verlag.

References

- Charpentier, A., Hu, F., and Ratz, P. (2023b). Mitigating discrimination in insurance with wasserstein barycenters. *BIAS, 3rd Workshop on Bias and Fairness in AI, International Workshop of ECML PKDD*.
- Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. *arXiv*, 1701.08230.
- Crossney, K. B. (2016). Redlining. <https://philadelphiaencyclopedia.org/essays/redlining/>.
- Dieterich, W., Mendoza, C., and Brennan, T. (2016). Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*, 7(7.4):1.
- Dressel, J. and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580.
- Feeley, M. and Simon, J. (1994). Actuarial justice: The emerging new criminal law. *The futures of criminology*, 173:174.
- Feller, A., Pierson, E., Corbett-Davies, S., and Goel, S. (2016). A computer program used for bail and sentencing decisions was labeled biased against blacks. it's actually not that clear. *The Washington Post*, October 17.
- Gouic, T. L., Loubes, J.-M., and Rigollet, P. (2020). Projection to fairness in statistical learning. *arXiv*, 2005.11720.

References

- Hu, F., Ratz, P., and Charpentier, A. (2023a). Fairness in multi-task learning via wasserstein barycenters. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases – ECML PKDD*.
- Hu, F., Ratz, P., and Charpentier, A. (2023b). Parametric Fairness with Statistical Guarantees. *ArXiv*, 2310.20508.
- Hu, F., Ratz, P., and Charpentier, A. (2023c). A sequentially fair mechanism for multiple sensitive attributes. *ArXiv*, 2309.06627.
- Hume, D. (1739). *A Treatise of Human Nature*. Cambridge University Press Archive.
- Kearns, M. and Roth, A. (2019). *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press.
- Knowlton, R. E. (1978). Regents of the university of california v. bakke. *Arkansas Law Review*, 32:499.
- Kranzberg, M. (1986). Technology and history:" kranzberg's laws". *Technology and culture*, 27(3):544–560.
- Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4066–4076. NIPS.
- Lippert-Rasmussen, K. (2020). *Making sense of affirmative action*. Oxford University Press.

References

- Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*.
- Pearl, J. et al. (2009). Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146.
- Pearl, J. and Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic books.
- Perry, W. L. (2013). *Predictive policing: The role of crime forecasting in law enforcement operations*. Rand Corporation.
- Rhynhart, R. (2020). Mapping the legacy of structural racism in philadelphia. *Philadelphia, Office pf the Controller*.
- Sabbagh, D. (2007). *Equality and transparency: A strategic perspective on affirmative action in American law*. Springer.
- Turner, R. (2015). The way to stop discrimination on the basis of race. *Stanford Journal of Civil Rights & Civil Liberties*, 11:45.
- Vapnik, V. (1991). Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4.