Fairness in machine learning

A study of the Demographic Parity constraint

Nicolas Schreuder

Joint work with E. Chzhen (CNRS) & S. Gaucher (CREST)

Fairness in machine learning?

Fairness in machine learning?

For scaling/financial reasons, an increasing number of *high-stakes decisions are being automated*:

Fairness in machine learning?

For scaling/financial reasons, an increasing number of *high-stakes decisions are being automated*:

- bank loans,
- job pre-screenings,
- school admissions,
- criminal sentencings,
- etc.

Fairness in machine learning?

For scaling/financial reasons, an increasing number of *high-stakes decisions are being automated*:

- bank loans,
- job pre-screenings,
- school admissions,
- criminal sentencings,
- etc.

Claim: The increasing automation of decision-making procedures critically increases the risk of simultaneously *automatising discriminations*.

Fairness in machine learning?

For scaling/financial reasons, an increasing number of *high-stakes decisions are being automated*:

- bank loans,
- job pre-screenings,
- school admissions,
- criminal sentencings,
- etc.

Claim: The increasing automation of decision-making procedures critically increases the risk of simultaneously *automatising discriminations*.

Let us see some concrete examples in the next slides.

Is Amazon sexist?

REUTERS

World Business Markets Breakingviews Video More

RETAIL OCTOBER 11, 2018 / 1:04 AM / UPDATED 4 YEARS AGO

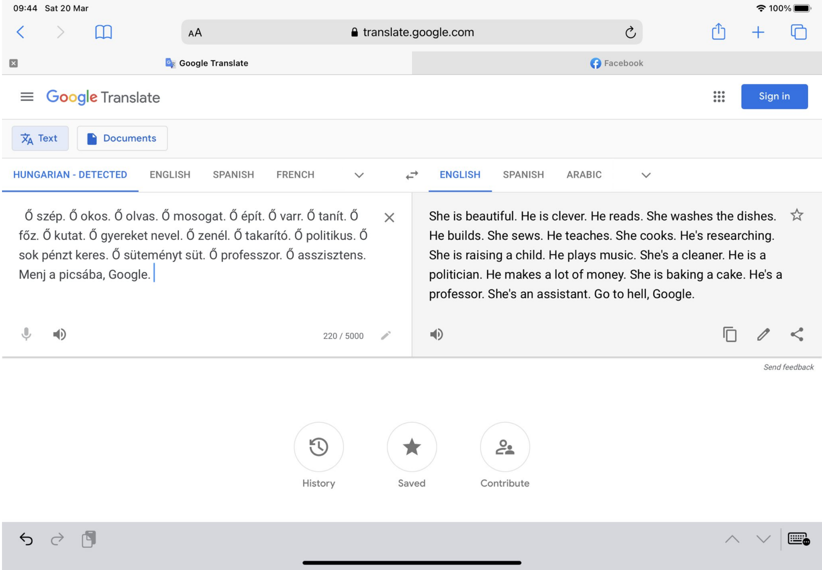
Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's AMZN.O machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

Is Google Translate sexist?



Source: Rozsa Melinda `https://www.reddit.com/r/europe/comments/m9uphb/hungarian_has_no_gendered_pronouns_so_google/`

How to formalize fairness?

Data: $(\underbrace{\text{feature}}_{\boldsymbol{X}}, \underbrace{\text{sensitive attribute}}_{S}, \underbrace{\text{label}}_{Y}) \sim \mathbb{P}$ on $\mathcal{X} \times \mathcal{S} \times \mathcal{Y}$.

How to formalize fairness?

Data: $(\underbrace{\text{feature}}_{\boldsymbol{X}}, \underbrace{\text{sensitive attribute}}_{S}, \underbrace{\text{label}}_{Y}) \sim \mathbb{P}$ on $\mathcal{X} \times \mathcal{S} \times \mathcal{Y}$.

Predictions: $f : \mathcal{Z} \to \mathcal{Y}$

- Fairness through awareness: $\mathcal{Z} = \mathcal{X} \times \mathcal{S}$ (disparate treatment);
- Fairness through UNawareness: $\mathcal{Z} = \mathcal{X}$ (legal reasons: regulations).

How to formalize fairness?

Data: $(\underbrace{\text{feature}}_{\boldsymbol{X}}, \underbrace{\text{sensitive attribute}}_{S}, \underbrace{\text{label}}_{Y}) \sim \mathbb{P}$ on $\mathcal{X} \times \mathcal{S} \times \mathcal{Y}$.

Predictions: $f : \mathcal{Z} \to \mathcal{Y}$

- Fairness through awareness: $\mathcal{Z} = \mathcal{X} \times \mathcal{S}$ (disparate treatment);
- Fairness through UNawareness: $\mathcal{Z} = \mathcal{X}$ (legal reasons: regulations).

A popular formalization of fairness is Demographic Parity (DP). We say that a prediction rule $f : \mathcal{Z} \to \mathcal{Y}$ satisfies DP if

$$f(\boldsymbol{Z}) \perp\!\!\!\perp S \ .$$

How to formalize fairness?

Data: $(\underbrace{\text{feature}}_{\boldsymbol{X}}, \underbrace{\text{sensitive attribute}}_{S}, \underbrace{\text{label}}_{Y}) \sim \mathbb{P}$ on $\mathcal{X} \times \mathcal{S} \times \mathcal{Y}$.

Predictions: $f : \mathcal{Z} \to \mathcal{Y}$

- Fairness through awareness: $\mathcal{Z} = \mathcal{X} \times \mathcal{S}$ (disparate treatment);
- Fairness through UNawareness: $\mathcal{Z} = \mathcal{X}$ (legal reasons: regulations).

A popular formalization of fairness is Demographic Parity (DP). We say that a prediction rule $f : \mathcal{Z} \to \mathcal{Y}$ satisfies DP if

$$f(\boldsymbol{Z}) \perp\!\!\!\perp S \enspace .$$

In the case of *binary classification* $\mathcal{Y} = \{0, 1\}$ and *binary sensitive attribute* $\mathcal{S} = \{0, 1\}$, it amounts to

$$\mathbb{P}(f(X, S) = 1 \mid S = 0) = \mathbb{P}(f(X, S) = 1 \mid S = 1) \enspace .$$

How to formalize fairness?

Data: $(\underbrace{\text{feature}}_{\boldsymbol{X}}, \underbrace{\text{sensitive attribute}}_{S}, \underbrace{\text{label}}_{Y}) \sim \mathbb{P}$ on $\mathcal{X} \times \mathcal{S} \times \mathcal{Y}$.

Predictions: $f : \mathcal{Z} \to \mathcal{Y}$

- Fairness through awareness: $\mathcal{Z} = \mathcal{X} \times \mathcal{S}$ (disparate treatment);
- Fairness through UNawareness: $\mathcal{Z} = \mathcal{X}$ (legal reasons: regulations).

A popular formalization of fairness is Demographic Parity (DP). We say that a prediction rule $f : \mathcal{Z} \to \mathcal{Y}$ satisfies DP if

$$f(\boldsymbol{Z}) \perp\!\!\!\perp S \ .$$

In the case of *binary classification* $\mathcal{Y} = \{0, 1\}$ and *binary sensitive attribute* $\mathcal{S} = \{0, 1\}$, it amounts to

$$\mathbb{P}(f(X, S) = 1 \mid S = 0) = \mathbb{P}(f(X, S) = 1 \mid S = 1) \ .$$

NB: Other formalizations of fairness exist, there is no “best one”.

Popular definitions of fair classifiers

- Demographic Parity (DP) (Calders, Kamiran, and Pechenizkiy, 2009)

$$\mathbb{P}(f(\boldsymbol{Z}) = 1 \mid S = 0) = \mathbb{P}(f(\boldsymbol{Z}) = 1 \mid S = 1)$$

  1. Prediction rate is the same for two groups.
  2. Random variable $f(\boldsymbol{Z})$ is independent from S.
  3. DP (not differential privacy!) cares only about $\boldsymbol{X}|S$.

Popular definitions of fair classifiers

- Demographic Parity (DP) (Calders, Kamiran, and Pechenizkiy, 2009)

$$\mathbb{P}(f(\boldsymbol{Z}) = 1 \mid S = 0) = \mathbb{P}(f(\boldsymbol{Z}) = 1 \mid S = 1)$$

  1. Prediction rate is the same for two groups.
  2. Random variable $f(\boldsymbol{Z})$ is independent from S.
  3. DP (not differential privacy!) cares only about $\boldsymbol{X}|S$.

- Equalized Odds (M. Hardt, Price, and Srebro, 2016)

$$\mathbb{P}(f(\boldsymbol{Z}) = y \mid Y = y, S = 0) = \mathbb{P}(f(\boldsymbol{Z}) = y \mid Y = y, S = 1) \quad \forall y \in \{0, 1\}$$

  1. Equal True Positive and True Negative rates.
  2. Requires more knowledge about the distribution.

Popular definitions of fair classifiers

- Demographic Parity (DP) (Calders, Kamiran, and Pechenizkiy, 2009)

$$\mathbb{P}(f(\boldsymbol{Z}) = 1 \mid S = 0) = \mathbb{P}(f(\boldsymbol{Z}) = 1 \mid S = 1)$$

  1. Prediction rate is the same for two groups.
  2. Random variable $f(\boldsymbol{Z})$ is independent from S.
  3. DP (not differential privacy!) cares only about $\boldsymbol{X}|S$.

- Equalized Odds (M. Hardt, Price, and Srebro, 2016)

$$\mathbb{P}(f(\boldsymbol{Z}) = y \mid Y = y, S = 0) = \mathbb{P}(f(\boldsymbol{Z}) = y \mid Y = y, S = 1) \quad \forall y \in \{0, 1\}$$

  1. Equal True Positive and True Negative rates.
  2. Requires more knowledge about the distribution.

- Equal Opportunity (M. Hardt, Price, and Srebro, 2016)

$$\mathbb{P}(f(\boldsymbol{Z}) = 1 \mid Y = 1, S = 0) = \mathbb{P}(f(\boldsymbol{Z}) = 1 \mid Y = 1, S = 1)$$

  1. Equal True Positive rates.
  2. If a person $\boldsymbol{Z}$ is qualified ($Y = 1$) then positive prediction ($f(\boldsymbol{Z}) = 1$) is given with the same probability for any sensitive attribute.

Popular definitions of fair classifiers

- Demographic Parity (DP) (Calders, Kamiran, and Pechenizkiy, 2009)

$$\mathbb{P}(f(\boldsymbol{Z}) = 1 \mid S = 0) = \mathbb{P}(f(\boldsymbol{Z}) = 1 \mid S = 1)$$

  1. Prediction rate is the same for two groups.
  2. Random variable $f(\boldsymbol{Z})$ is independent from S.
  3. DP (not differential privacy!) cares only about $\boldsymbol{X}|S$.

- Equalized Odds (M. Hardt, Price, and Srebro, 2016)

$$\mathbb{P}(f(\boldsymbol{Z}) = y \mid Y = y, S = 0) = \mathbb{P}(f(\boldsymbol{Z}) = y \mid Y = y, S = 1) \quad \forall y \in \{0, 1\}$$

  1. Equal True Positive and True Negative rates.
  2. Requires more knowledge about the distribution.

- Equal Opportunity (M. Hardt, Price, and Srebro, 2016)

$$\mathbb{P}(f(\boldsymbol{Z}) = 1 \mid Y = 1, S = 0) = \mathbb{P}(f(\boldsymbol{Z}) = 1 \mid Y = 1, S = 1)$$

  1. Equal True Positive rates.
  2. If a person $\boldsymbol{Z}$ is qualified ($Y = 1$) then positive prediction ($f(\boldsymbol{Z}) = 1$) is given with the same probability for any sensitive attribute.

Question: Which one(s) should we enforce?

Incompatibility of fairness constraints[1]

1. $f(\boldsymbol{Z}) \perp\!\!\!\perp S$ - independence (DP, Statistical Parity)

2. $(f(\boldsymbol{Z}) \perp\!\!\!\perp S) \mid Y$ - separation (Equal Odds, Equal Opportunity)

3. $(Y \perp\!\!\!\perp S) \mid f(\boldsymbol{Z})$ - sufficiency (Test fairness)

[1] Taken from Chapter 2 of (Barocas, Moritz Hardt, and Narayanan, 2019).

Incompatibility of fairness constraints[1]

1. $f(\boldsymbol{Z}) \perp\!\!\!\perp S$ - independence (DP, Statistical Parity)

2. $(f(\boldsymbol{Z}) \perp\!\!\!\perp S) \mid Y$ - separation (Equal Odds, Equal Opportunity)

3. $(Y \perp\!\!\!\perp S) \mid f(\boldsymbol{Z})$ - sufficiency (Test fairness)

Proposition

If $Y \in \{0, 1\}$, $S \not\perp\!\!\!\perp Y$, and $f(\boldsymbol{Z}) \not\perp\!\!\!\perp Y$, then independence and separation cannot hold simultaneously.

[1] Taken from Chapter 2 of (Barocas, Moritz Hardt, and Narayanan, 2019).

Incompatibility of fairness constraints[1]

1. $f(\boldsymbol{Z}) \perp\!\!\!\perp S$ - independence (DP, Statistical Parity)

2. $(f(\boldsymbol{Z}) \perp\!\!\!\perp S) \mid Y$ - separation (Equal Odds, Equal Opportunity)

3. $(Y \perp\!\!\!\perp S) \mid f(\boldsymbol{Z})$ - sufficiency (Test fairness)

Proposition

If $Y \in \{0, 1\}$, $S \not\perp\!\!\!\perp Y$, and $f(\boldsymbol{Z}) \not\perp\!\!\!\perp Y$, then independence and separation cannot hold simultaneously.

Similarly, separation cannot hold simultaneously as suff./sep. in general.

[1] Taken from Chapter 2 of (Barocas, Moritz Hardt, and Narayanan, 2019).

Incompatibility of fairness constraints[1]

1. $f(\boldsymbol{Z}) \perp\!\!\!\perp S$ - independence (DP, Statistical Parity)

2. $(f(\boldsymbol{Z}) \perp\!\!\!\perp S) \mid Y$ - separation (Equal Odds, Equal Opportunity)

3. $(Y \perp\!\!\!\perp S) \mid f(\boldsymbol{Z})$ - sufficiency (Test fairness)

Proposition

If $Y \in \{0, 1\}$, $S \not\perp\!\!\!\perp Y$, and $f(\boldsymbol{Z}) \not\perp\!\!\!\perp Y$, then independence and separation cannot hold simultaneously.

Similarly, separation cannot hold simultaneously as suff./sep. in general.

Consequences: need to choose one notion of fairness (or relax?).

[1]Taken from Chapter 2 of (Barocas, Moritz Hardt, and Narayanan, 2019).

Some personal contributions on the Demographic Parity constraint

The cost of fairness/Demographic Parity

- Many works empirically studied the impact of (relaxed) fairness constraints on the risk (Bertsimas, Farias, and Trichakis, 2012; Zliobaite, 2015; Kleinberg, Mullainathan, and Raghavan, 2016; Zafar et al., 2017; Haas, 2019; Wick, Tristan, et al., 2019).
- Yet, the problem of mathematically/statistically quantifying the effect of such constraints on the risk had not been tackled.

Optimal transport and the Wasserstein-2 metric

Define, for $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$\mathsf{W}_2^2(\mu, \nu) := \inf \left\{ \mathbb{E}_{(X,Y)} \| \boldsymbol{X} - \boldsymbol{Y} \|_2^2 : \boldsymbol{X} \sim \mu, \boldsymbol{Y} \sim \nu \right\}.$$

- Metric on $\mathcal{P}_2(\mathbb{R}^d)$.
- Highly flexible/handy.
- Nice geometric features.

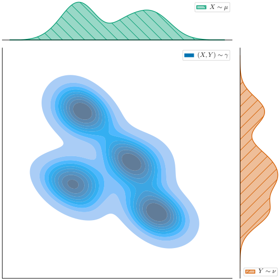


Figure: Transport plan illustration

Squared-loss regression under relaxed DP

$(\underbrace{\text{feature}}_{\boldsymbol{X}}, \underbrace{\text{sensitive attribute}}_{S}, \underbrace{\text{label}}_{Y}) \sim \mathbb{P}$ on $\mathcal{X} \times \mathcal{S} \times \mathbb{R}$.

Squared-loss regression under relaxed DP

$(\underbrace{\text{feature}}_{\boldsymbol{X}}, \underbrace{\text{sensitive attribute}}_{S}, \underbrace{\text{label}}_{Y}) \sim \mathbb{P}$ on $\mathcal{X} \times \mathcal{S} \times \mathbb{R}$.

1. Predictions: $f : \mathcal{X} \times \mathcal{S} \to \mathbb{R}$.

Squared-loss regression under relaxed DP

$(\underbrace{\text{feature}}_{\boldsymbol{X}}, \underbrace{\text{sensitive attribute}}_{S}, \underbrace{\text{label}}_{Y}) \sim \mathbb{P}$ on $\mathcal{X} \times \mathcal{S} \times \mathbb{R}$.

1. Predictions: $f : \mathcal{X} \times \mathcal{S} \to \mathbb{R}$.
2. Risk: $\mathcal{R}(f) := \mathbb{E}(Y - f(\boldsymbol{X}, S))^2$, min. by $f^*(x, s) := \mathbb{E}[Y | X{=}x, S{=}s]$).

Squared-loss regression under relaxed DP

$(\underbrace{\text{feature}}_{\boldsymbol{X}}, \underbrace{\text{sensitive attribute}}_{S}, \underbrace{\text{label}}_{Y}) \sim \mathbb{P}$ on $\mathcal{X} \times \mathcal{S} \times \mathbb{R}$.

1. Predictions: $f : \mathcal{X} \times \mathcal{S} \to \mathbb{R}$.
2. Risk: $\mathcal{R}(f) := \mathbb{E}(Y - f(\boldsymbol{X}, S))^2$, min. by $f^*(x, s) := \mathbb{E}[Y | X{=}x, S{=}s])$.
3. Relaxed Demographic parity: $\mathcal{U}(f) \leq \alpha \mathcal{U}(f^*)$, where $0 \leq \alpha \leq 1$ and

$$\mathcal{U}(f) = \min_{\nu} \sum_{s \in \mathcal{S}} w_s \mathsf{W}_2^2(\mathrm{Law}(f(\boldsymbol{X}, S) | S = s), \nu) \in [0, +\infty).$$

♣ $\mathcal{U}(f) = 0$ *if and only if* f satisfies DP.

Squared-loss regression under relaxed DP

$(\underbrace{\text{feature}}_{\boldsymbol{X}}, \underbrace{\text{sensitive attribute}}_{S}, \underbrace{\text{label}}_{Y}) \sim \mathbb{P}$ on $\mathcal{X} \times \mathcal{S} \times \mathbb{R}$.

1. Predictions: $f : \mathcal{X} \times \mathcal{S} \to \mathbb{R}$.
2. Risk: $\mathcal{R}(f) := \mathbb{E}(Y - f(\boldsymbol{X}, S))^2$, min. by $f^*(x, s) := \mathbb{E}[Y | X{=}x, S{=}s]$).
3. Relaxed Demographic parity: $\mathcal{U}(f) \leq \alpha \mathcal{U}(f^*)$, where $0 \leq \alpha \leq 1$ and

$$\mathcal{U}(f) = \min_{\nu} \sum_{s \in \mathcal{S}} w_s \mathsf{W}_2^2(\mathrm{Law}(f(\boldsymbol{X}, S) | S = s), \nu) \in [0, +\infty).$$

♣ $\mathcal{U}(f) = 0$ *if and only if* f satisfies DP.

$$\boxed{\alpha\textbf{-Relative Improvement:} \quad f^*_\alpha \in \arg\min \{\mathcal{R}(f) \ : \mathcal{U}(f) \leq \alpha \mathcal{U}(f^*)\}}$$

Squared-loss regression under relaxed DP

$(\underbrace{\text{feature}}_{\boldsymbol{X}}, \underbrace{\text{sensitive attribute}}_{S}, \underbrace{\text{label}}_{Y}) \sim \mathbb{P}$ on $\mathcal{X} \times \mathcal{S} \times \mathbb{R}$.

1. Predictions: $f : \mathcal{X} \times \mathcal{S} \to \mathbb{R}$.

2. Risk: $\mathcal{R}(f) \coloneqq \mathbb{E}(Y - f(\boldsymbol{X}, S))^2$, min. by $f^*(x, s) \coloneqq \mathbb{E}[Y | X{=}x, S{=}s])$.

3. Relaxed Demographic parity: $\mathcal{U}(f) \leq \alpha \mathcal{U}(f^*)$, where $0 \leq \alpha \leq 1$ and

$$\mathcal{U}(f) = \min_{\nu} \sum_{s \in \mathcal{S}} w_s \mathsf{W}_2^2(\mathrm{Law}(f(\boldsymbol{X}, S) | S = s), \nu) \in [0, +\infty).$$

♣ $\mathcal{U}(f) = 0$ *if and only if* f satisfies DP.

$$\boxed{\textbf{α-Relative Improvement:} \quad f^*_\alpha \in \arg\min \{\mathcal{R}(f) \,:\, \mathcal{U}(f) \leq \alpha \mathcal{U}(f^*)\}}$$

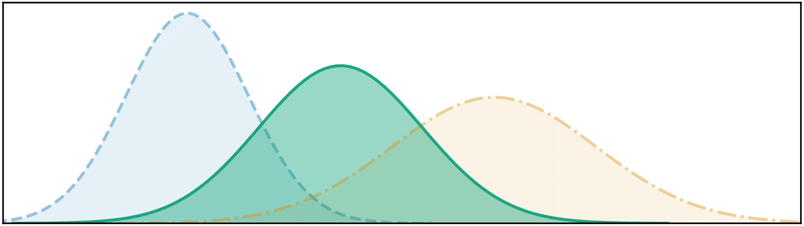
Question: What is the price –in risk– of considering fair predictors?

Unfairness through Wasserstein barycenters

$$\mathcal{U}(f) = \min_{\nu} \sum_{s \in \mathcal{S}} w_s \mathsf{W}_2^2(\mathrm{Law}(f(\boldsymbol{X}, S) | S = s), \nu).$$

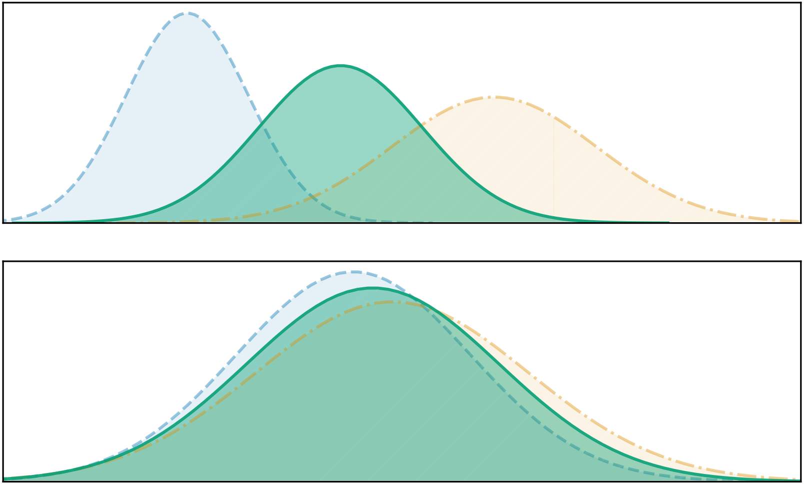
Unfairness through Wasserstein barycenters

$$\mathcal{U}(f) = \min_{\nu} \sum_{s \in \mathcal{S}} w_s \mathsf{W}_2^2(\mathrm{Law}(f(\boldsymbol{X}, S) | S = s), \nu).$$



Unfairness through Wasserstein barycenters

$$\mathcal{U}(f) = \min_{\nu} \sum_{s \in \mathcal{S}} w_s \mathsf{W}_2^2(\mathrm{Law}(f(\boldsymbol{X}, S) | S = s), \nu).$$



Main assumption

Assumption (A)

The group-wise prediction distributions $\mathrm{Law}(f^*(\boldsymbol{X}, S) \mid S = s)$ have finite second moment and are non-atomic for any s in $\mathcal{S}$.

Improving unfairness oracles

α**-Relative Improvement** $\quad f^*_\alpha \in \arg\min \left\{ \mathcal{R}(f) \,:\, \boxed{\mathcal{U}(f) \leq \alpha\, \mathcal{U}(f^*)} \right\}$

Improving unfairness oracles

α-Relative Improvement $\quad f^*_\alpha \in \arg\min \left\{ \mathcal{R}(f) \ : \ \boxed{\mathcal{U}(f) \leq \alpha \mathcal{U}(f^*)} \right\}$

Theorem

Under Assumption (A), for all $\alpha \in [0, 1]$ it holds that

$$f^*_\alpha \equiv \sqrt{\alpha} f^*_1 + (1 - \sqrt{\alpha}) f^*_0 \quad .$$

(Evgenii Chzhen and Schreuder, 2022)

Improving unfairness oracles

α-Relative Improvement $\quad f^*_\alpha \in \arg\min\left\{\mathcal{R}(f) \,:\, \boxed{\mathcal{U}(f) \leq \alpha\,\mathcal{U}(f^*)}\right\}$

Theorem

Under Assumption (A), for all $\alpha \in [0,1]$ it holds that

$$f^*_\alpha \equiv \sqrt{\alpha} f^*_1 + (1-\sqrt{\alpha}) f^*_0 \quad .$$

(Evgenii Chzhen and Schreuder, 2022)

Theorem

Under Assumption (A),

$$\mathrm{Law}(f^*_0(\boldsymbol{X}, S)) = \arg\min_{\nu} \sum_{s\in\mathcal{S}} w_s W_2^2\left(\mathrm{Law}(f^*(\boldsymbol{X}, S) \mid S = s), \nu\right) \quad ,$$

$$f^*_0(\boldsymbol{x}, s) = \left(\sum_{s'\in\mathcal{S}} w_{s'} F^{-1}_{f^*|S=s'}\right) \circ F_{f^*|S=s} \circ f^*(\boldsymbol{x}, s) \quad ,$$

where $w_s = \mathbb{P}(S{=}s)$, $F_{f^*|S=s}(t) = \mathbb{P}(f^*(\boldsymbol{X}, S){\leq}t|S{=}s)$.

Chzhen, Denis, Hebiri et al. (2020); Le Gouic et al. (2020)

Key ingredient for the proof

Abstract geometric lemma

Let $(\mathcal{X}, d)$ be a metric space in which barycenters are well-defined. Let $\boldsymbol{a} = (a_1, \ldots, a_K) \in \mathcal{X}^K$, $\boldsymbol{w} = (w_1, \ldots, w_K)^\top \in \Delta^{K-1}$ and let $C_{\boldsymbol{a}}$ be a barycenter of $\boldsymbol{a}$ with respect to weights $\boldsymbol{w}$. For a fixed $\alpha \in [0,1]$ assume that there exists $\boldsymbol{b} = (b_1, \ldots, b_K) \in \mathcal{X}^K$ which satisfies

$$d(a_s, C_{\boldsymbol{a}}) = d(a_s, b_s) + d(b_s, C_{\boldsymbol{a}}) \ , \qquad s = 1, \ldots, K \ , \tag{P_1}$$

$$d(b_s, a_s) = (1-\sqrt{\alpha}) d(a_s, C_{\boldsymbol{a}}) \ , \qquad s = 1, \ldots, K \ . \tag{P_2}$$

Then, $\boldsymbol{b}$ is a solution of

$$\inf_{\boldsymbol{b} \in \mathcal{X}^K} \left\{ \sum_{s=1}^{K} w_s d^2(b_s, a_s) \ : \ \sum_{s=1}^{K} w_s d^2(b_s, C_{\boldsymbol{b}}) \leq \alpha \sum_{s=1}^{K} w_s d^2(a_s, C_{\boldsymbol{a}}) \right\} \ .$$

Key ingredient for the proof

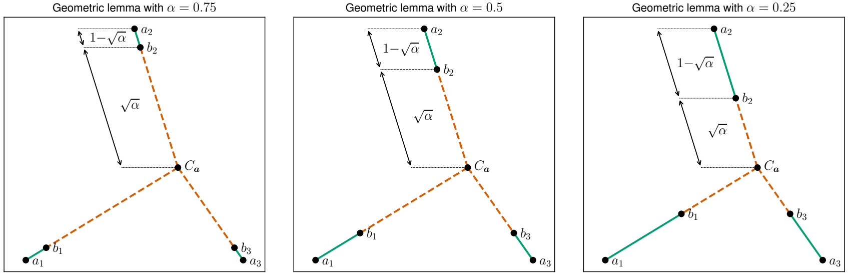
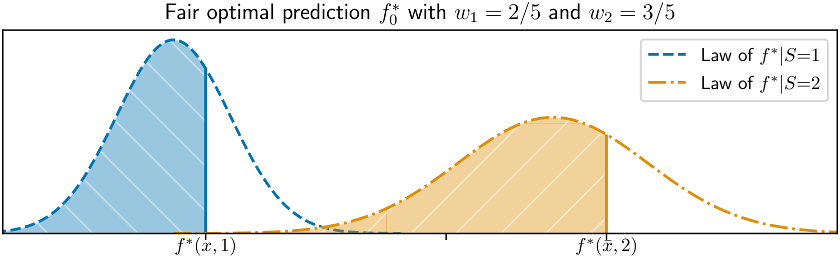
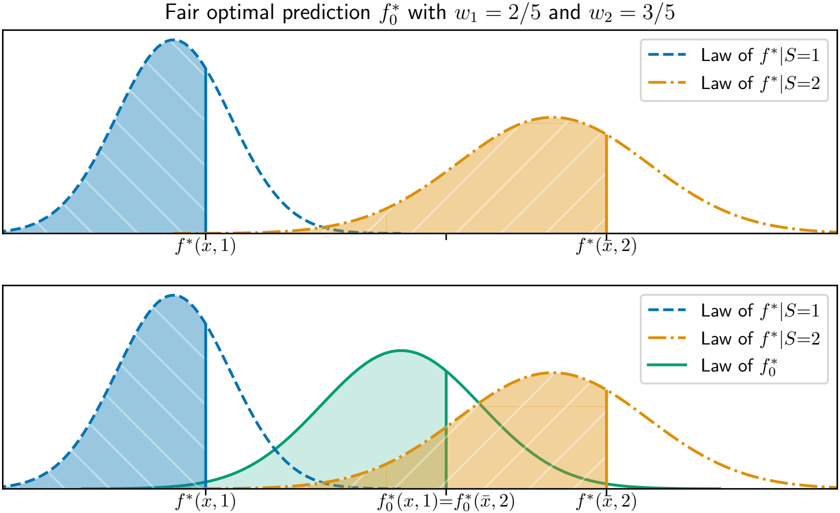


Figure: Illustration of the abstract geometric lemma for $(\mathcal{X}, d) = (\mathbb{R}^2, \|\cdot\|_2)$ and $\alpha \in \{0.25, 0.5, 0.75\}$. The initial points a_1, a_2, a_3 are the vertices of an isosceles triangle. The weights are set as follows: $w_1 = 0.1$, $w_2 = 0.4$ and $w_3 = 0.5$.

What is (exact) fair regression?



What is (exact) fair regression?



$$f_0^*(\boldsymbol{x}, s) = \Big(\sum_{s' \in \mathcal{S}} w_{s'} F^{-1}_{f^*|S=s'}\Big) \circ F_{f^*|S=s} \circ f^*(\boldsymbol{x}, s)$$

Risk/fairness trade-off

α**-Relative Improvement** $\quad f^*_\alpha \in \arg\min \left\{ \mathcal{R}(f) \, : \, \boxed{\mathcal{U}(f) \leq \alpha \mathcal{U}(f^*)} \right\}$

Risk/fairness trade-off

α-Relative Improvement $\quad f^*_\alpha \in \arg\min \left\{ \mathcal{R}(f) \ : \ \boxed{\mathcal{U}(f) \le \alpha \mathcal{U}(f^*)} \right\}$

Proposition

Under Assumption (A), for all $\alpha \in [0,1]$ it holds that

$$\mathcal{R}(f^*_\alpha) = (1-\sqrt{\alpha})^2 \boxed{\mathcal{U}(f^*)} \quad \text{and} \quad \mathcal{U}(f^*_\alpha) = \alpha \boxed{\mathcal{U}(f^*)} \ .$$

(Evgenii Chzhen and Schreuder, 2022)

Risk/fairness trade-off

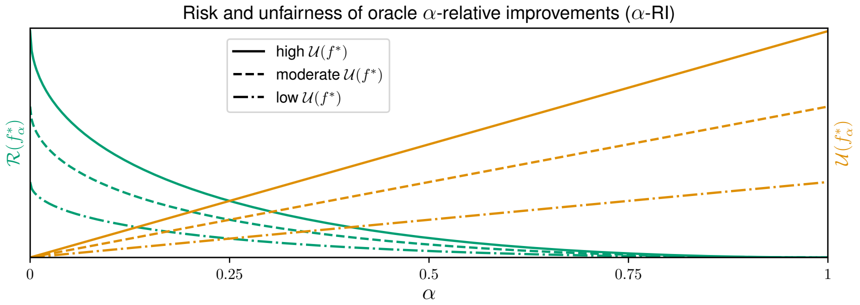
α-Relative Improvement $\quad f^*_\alpha \in \arg\min \left\{ \mathcal{R}(f) \; : \; \boxed{\mathcal{U}(f) \leq \alpha\,\mathcal{U}(f^*)} \right\}$

Proposition

Under Assumption (A), for all $\alpha \in [0,1]$ it holds that

$$\mathcal{R}(f^*_\alpha) = (1-\sqrt{\alpha})^2 \boxed{\mathcal{U}(f^*)} \quad \text{and} \quad \mathcal{U}(f^*_\alpha) = \alpha \boxed{\mathcal{U}(f^*)} \ .$$

(Evgenii Chzhen and Schreuder, 2022)



Minimax statistical framework

Data: $(\boldsymbol{X}_1, S_1, Y_1), \dots, (\boldsymbol{X}_n, S_n, Y_n) \overset{i.i.d.}{\sim} \mathbf{P}_{(f^*, \boldsymbol{\theta})}$, $(f^*, \boldsymbol{\theta}) \in \mathcal{F} \times \Theta$
Given $\alpha \in [0, 1]$ and $t > 0$, the goal of the statistician is to construct an estimator $\hat{f}$, which simultaneously satisfies

1. Uniform fairness guarantee:

$$\forall (f^*, \boldsymbol{\theta}) \in \mathcal{F} \times \Theta \quad \mathbf{P}_{(f^*, \boldsymbol{\theta})} \left(\mathcal{U}(\hat{f}) \leq \alpha \mathcal{U}(f^*) \right) \geq 1 - t \enspace ,$$

2. Uniform risk guarantee:

$$\forall (f^*, \boldsymbol{\theta}) \in \mathcal{F} \times \Theta \quad \mathbf{P}_{(f^*, \boldsymbol{\theta})} \left(\mathcal{R}(\hat{f}) \leq r_{n, \alpha, f^*}(\mathcal{F}, \Theta, t) \right) \geq 1 - t \enspace .$$

Application to linear model with systematic bias

Linear regression with systematic group-dependent bias model:

$$Y_i = \left\langle \boldsymbol{X}_i, \bar{\beta}^* \right\rangle + b^*_{S_i} + \sigma\xi_i, \quad i = 1, \ldots, n \ ,$$

where $\{\xi_i\}_{i=1}^n \overset{i.i.d.}{\sim} \mathcal{N}(0,1) \perp\!\!\!\perp \{\boldsymbol{X}_i\}_{i=1}^n \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \perp\!\!\!\perp \{S_i\}_{i=1}^n$.
We assume that σ is known and $\boldsymbol{\Sigma} > 0$.

We propose an estimator $\hat{f}$ which, with probability at least $1-\delta$, satisfies $\mathcal{U}(\hat{f}) \leq \alpha\,\mathcal{U}(f^*)$ and achieves the minimax optimal rate

$$\mathcal{R}(\hat{f}) \asymp \left\{\sigma^2\left(\frac{p+K}{n} + \frac{\log(1/\delta)}{n}\right)\right\} \bigvee \left\{(1-\sqrt{\alpha})^2\,\mathcal{U}(f^*)\right\} \ .$$

Recently, (Fukuchi and Sakuma, 2022) proposed an extension of our model to allow correlations between X and S.

Proposed estimator (1/2)

Oracle α-RI

Under the considered model it holds that

$$f^*_\alpha(\boldsymbol{x}, s) = \left\langle \boldsymbol{x}, \bar{\beta}^* \right\rangle + \sqrt{\alpha} b^*_s + (1-\sqrt{\alpha}) \sum_{s=1}^{K} w_s b^*_s, \qquad \forall \alpha \in [0, 1] \ .$$

Plug-in estimated parameters

$$(\hat{\beta}, \hat{\boldsymbol{b}}) \in \operatorname*{arg\,min}_{(\bar{\beta}, \boldsymbol{b}) \in \mathbb{R}^p \times \mathbb{R}^K} \sum_{s=1}^{K} w_s \left\| \boldsymbol{Y}_s - \mathbf{X}_s \bar{\beta} - b_s \mathbf{1}_{n_s} \right\|^2_{n_s} \ .$$

to get a family of linear estimators

$$\hat{f}_\tau(\boldsymbol{x}, s) = \langle \boldsymbol{x}, \hat{\beta} \rangle + \sqrt{\tau} \hat{b}_s + (1-\sqrt{\tau}) \sum_{s=1}^{K} w_s \hat{b}_s, \qquad (\boldsymbol{x}, s) \in \mathbb{R}^p \times [K] \ .$$

Proposed estimator (2/2)

Define

$$\delta_n := \delta_n(p, K, t) = 8\left(\frac{p}{n} + \frac{K}{n}\right) + 16\left(\sqrt{\frac{p}{n}} + \sqrt{\frac{K}{n}}\right)\sqrt{\frac{t}{n}} + \frac{32t}{n} \ .$$

Upper bound theorem

Let $\alpha \in [0, 1]$. For n "large enough", setting

$$\hat{\tau} = \alpha \left(1 + \frac{\sigma \delta_n^{1/2}}{\mathcal{U}^{1/2}(\hat{f}_1) - \sigma \delta_n^{1/2}}\right)^{-2} \mathbb{1}\left(\mathcal{U}^{1/2}(\hat{f}_1) > \sigma \delta_n^{1/2}\right) \ ,$$

it holds with probability at least $1 - 4e^{-t/2}$ that

$$\mathcal{U}(\hat{f}_{\hat{\tau}}) \leq \alpha \mathcal{U}(f^*) \quad \text{and} \quad \mathcal{R}^{1/2}(\hat{f}_{\hat{\tau}}) \leq 2\sigma(1+\sqrt{\alpha})\delta_n^{1/2} + (1-\sqrt{\alpha})\mathcal{U}^{1/2}(f^*) \ .$$

(Evgenii Chzhen and Schreuder, 2022)

Lower bound

Define

$$\bar{\delta}_n \coloneqq \bar{\delta}_n(p, K, t) \coloneqq \left(\sqrt{(p+K)/n} + \sqrt{32t/n}\right)^2 / (3 \cdot 2^9) \ . \tag{1}$$

Lower bound

For all $n, p, K \in \mathbb{N}$, $t \geq 0$, $\sigma > 0$, $\alpha \in [0, 1]$ it holds for all $t \geq 0$ and all $t' \leq 1 - e^{-t}/12$ that any estimator $\hat{f}$ satisfying

$$\inf_{(f^*, \boldsymbol{\theta}) \in \mathcal{F} \times \Theta} \mathbf{P}_{(f^*, \boldsymbol{\theta})} \left(\mathcal{U}(\hat{f}) \leq \alpha \mathcal{U}(f^*) \right) \geq 1 - t' \ .$$

verifies

$$\sup_{(\bar{\beta}^*, \boldsymbol{b}^*), \boldsymbol{\Sigma} \succ 0} \mathbf{P}_{(\bar{\beta}^*, \boldsymbol{b}^*)} \left(\mathcal{R}^{1/2}(\hat{f}) \geq \sigma \bar{\delta}_n^{1/2} \vee (1 - \sqrt{\alpha}) \mathcal{U}^{1/2}(f^*) \right) \geq \frac{1}{12} e^{-t} \ .$$

(Evgenii Chzhen and Schreuder, 2022)

Numerical experiments

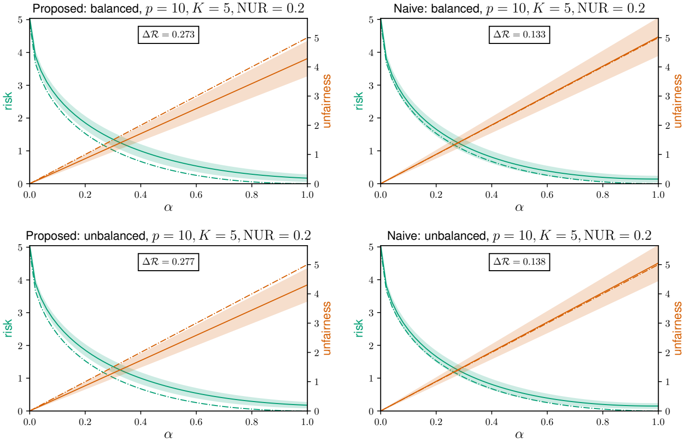


Figure: Dashed green and brown lines correspond to the risk and unfairness of f^*_α respectively. Solid green and brown lines correspond to the average risk and unfairness of $\hat{f}_{\tau(\alpha)}$ and the shaded region shows three standard deviations over 50 repetitions.

General post-processing procedure: definition

For each $f : \mathbb{R}^p \times [K] \to \mathbb{R}$, $s \in [K]$ and $i \in [2N_s]$, define the following r.v.

$$\tilde{f}_i^s := f(\boldsymbol{X}_i^s, s) + \mathcal{U}([-\sigma, \sigma]) \quad \text{and} \quad \tilde{f}(\boldsymbol{x}, s) := f(\boldsymbol{x}, s) + \mathcal{U}([-\sigma, \sigma]) \quad \forall\, \boldsymbol{x} \in \mathbb{R}^p \enspace .$$

Using the above quantities, we build the following estimators: for all $t \in \mathbb{R}$

$$\hat{F}_{1,\nu_s^f}(t) := \frac{1}{N_s+1}\left(\sum_{i=1}^{N_s} \mathbb{1}\left\{\tilde{f}_i^s < t\right\} + \mathcal{U}([0,1])\left(1 + \sum_{i=1}^{N_s} \mathbb{1}\left\{\tilde{f}_i^s = t\right\}\right)\right) \enspace ,$$

$$\hat{F}_{2,\nu_s^f}(t) := \frac{1}{N_s} \sum_{i=N_s+1}^{2N_s} \mathbb{1}\left\{\tilde{f}_i^s \leq t\right\} \enspace .$$

Finally, for each $f : \mathbb{R}^p \times [K] \to \mathbb{R}$ we define an estimator of f_0^*,

$$\hat{\Pi}(f)(\boldsymbol{x}, s) = \sum_{s'=1}^{K} w_{s'} \hat{F}^{-1}_{2,\nu^f_{s'}} \circ \hat{F}_{1,\nu_s^f} \circ \tilde{f}(\boldsymbol{x}, s), \quad \forall (\boldsymbol{x}, s) \in \mathbb{R}^p \times [K] \enspace .$$

How fair/accurate is it?

General post-processing: fairness guarantees

Theorem (Demographic parity guarantee)

For any $f : \mathbb{R}^p \times [K] \to \mathbb{R}$*, any joint distribution* $\mathbb{P}$ *of* $(\boldsymbol{X}, S, Y)$ *and any* $\sigma > 0$*, it holds that*

$$\mathrm{Law}\left(\hat{\Pi}(f)(\boldsymbol{X}, S) \mid S = s\right) = \mathrm{Law}\left(\hat{\Pi}(f)(\boldsymbol{X}, S) \mid S = s'\right) \qquad \forall s, s' \in [K] \ .$$

General post-processing: estimation guarantees

Assumption

For all $s \in [K]$, the measures ν_s^* are supported on an interval in $\mathbb{R}$, admit density w.r.t. *Lebesgue measure which is lower and upper bounded by $\underline{\lambda}_s > 0$ and $\overline{\lambda}_s > 0$ respectively.*

Theorem (Estimation guarantee)

Let the above assumption and Assumption (A) hold. Then, for any base prediction rule $f : \mathbb{R}^p \times [K] \to \mathbb{R}$, any $\sigma \in (0, 1)$ and any $q \in [1, \infty)$,

$$\mathbf{E}\|\hat{\Pi}(f) - f_0^*\|_q \leq C_{\underline{\boldsymbol{\lambda}}}^q \Bigg(\|f - f^*\|_q + \min\left\{ \|f - f^*\|_{q-1}^{1/p} + \sigma^{1/p}, \|f - f^*\|_\infty + \sigma \right\} \mathbb{1}\{q > 1\} + \left\{ \sum_{s=1}^K w_s N_s^{-1/2} \right\} + \left\{ \sum_{s=1}^K w_s N_s^{-q/2} \right\}^{1/q} + \sigma \Bigg) ,$$

where $C_{\underline{\boldsymbol{\lambda}}}^q$ depends only on $(\underline{\lambda}_s)_s, (\overline{\lambda}_s)_s$, $q \in [1, \infty)$ and $1/p + 1/q = 1$.

Thank you for your attention!

Questions?

Thank you for your attention! Questions?

- Our unfairness measure puts two conflicting quantities on the same scale: the risk-fairness trade-off is described by only one quantity.
- Minimax framework to study (relaxed) DP-fair estimators.
- Derived general problem-depend lower bound in this framework.
- Lower bound is tight for linear regression with systematic bias model.
- Only fair regression matters.

Thank you for your attention! Questions?

- Our unfairness measure puts two conflicting quantities on the same scale: the risk-fairness trade-off is described by only one quantity.
- Minimax framework to study (relaxed) DP-fair estimators.
- Derived general problem-depend lower bound in this framework.
- Lower bound is tight for linear regression with systematic bias model.
- Only fair regression matters.

Open questions: what about

- other models?
- other fairness constraints/relaxations?

Thank you for your attention! Questions?

- Our unfairness measure puts two conflicting quantities on the same scale: the risk-fairness trade-off is described by only one quantity.
- Minimax framework to study (relaxed) DP-fair estimators.
- Derived general problem-depend lower bound in this framework.
- Lower bound is tight for linear regression with systematic bias model.
- Only fair regression matters.

Open questions: what about

- other models?
- other fairness constraints/relaxations?

For more details:

- Evgenii Chzhen and Nicolas Schreuder (2022). “A minimax framework for quantifying risk-fairness trade-off in regression”. In: *The Annals of Statistics* 50.4, pp. 2416–2442
- Solenne Gaucher, Nicolas Schreuder, and Evgenii Chzhen (2023). “Fair learning with Wasserstein barycenters for non-decomposable performance measures”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 2436–2459

Bibliography I

Barocas, Solon, Moritz Hardt, and Arvind Narayanan (2019). *Fairness and Machine Learning*. http://www.fairmlbook.org. fairmlbook.org.

Bertsimas, Dimitris, Vivek F Farias, and Nikolaos Trichakis (2012). "On the efficiency-fairness trade-off". In: *Management Science* 58.12, pp. 2234–2250.

Calders, T., F. Kamiran, and M. Pechenizkiy (2009). "Building classifiers with independency constraints". In: *IEEE international conference on Data mining*.

Chzhen, E et al. (2020). "Fair Regression with Wasserstein Barycenters". In: *NeurIPS 2020*.

Chzhen, Evgenii and Nicolas Schreuder (2022). "A minimax framework for quantifying risk-fairness trade-off in regression". In: *The Annals of Statistics* 50.4, pp. 2416–2442.

Fukuchi, Kazuto and Jun Sakuma (2022). "Minimax Optimal Fair Regression under Linear Model". In: *arXiv preprint arXiv:2206.11546*.

Gaucher, Solenne, Nicolas Schreuder, and Evgenii Chzhen (2023). "Fair learning with Wasserstein barycenters for non-decomposable performance measures". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 2436–2459.

Bibliography II

Haas, Christian (2019). "The price of fairness-A framework to explore trade-offs in algorithmic fairness". In.

Hardt, M., E. Price, and N. Srebro (2016). "Equality of opportunity in supervised learning". In: *Neural Information Processing Systems*.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan (2016). "Inherent trade-offs in the fair determination of risk scores". In: *arXiv preprint arXiv:1609.05807*.

Le Gouic, T., J.-M. Loubes, and P. Rigollet (2020). "Projection to Fairness in Statistical Learning". In: *arXiv preprint arXiv:2005.11720*.

Wick, Michael, Jean-Baptiste Tristan, et al. (2019). "Unlocking fairness: a trade-off revisited". In: *Advances in neural information processing systems* 32.

Zafar, Muhammad Bilal et al. (2017). "Fairness constraints: Mechanisms for fair classification". In: *Artificial intelligence and statistics*. PMLR, pp. 962–970.

Zliobaite, Indre (2015). "On the relation between accuracy and fairness in binary classification". In: *arXiv preprint arXiv:1505.05723*.

Barocas, Solon, Moritz Hardt, and Arvind Narayanan (2019). *Fairness and Machine Learning*. `http://www.fairmlbook.org`. fairmlbook.org.

Bibliography III

Bertsimas, Dimitris, Vivek F Farias, and Nikolaos Trichakis (2012). "On the efficiency-fairness trade-off". In: *Management Science* 58.12, pp. 2234–2250.

Calders, T., F. Kamiran, and M. Pechenizkiy (2009). "Building classifiers with independency constraints". In: *IEEE international conference on Data mining*.

Chzhen, E et al. (2020). "Fair Regression with Wasserstein Barycenters". In: *NeurIPS 2020*.

Chzhen, Evgenii and Nicolas Schreuder (2022). "A minimax framework for quantifying risk-fairness trade-off in regression". In: *The Annals of Statistics* 50.4, pp. 2416–2442.

Fukuchi, Kazuto and Jun Sakuma (2022). "Minimax Optimal Fair Regression under Linear Model". In: *arXiv preprint arXiv:2206.11546*.

Gaucher, Solenne, Nicolas Schreuder, and Evgenii Chzhen (2023). "Fair learning with Wasserstein barycenters for non-decomposable performance measures". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 2436–2459.

Haas, Christian (2019). "The price of fairness-A framework to explore trade-offs in algorithmic fairness". In.

Bibliography IV

Hardt, M., E. Price, and N. Srebro (2016). "Equality of opportunity in supervised learning". In: *Neural Information Processing Systems.*

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan (2016). "Inherent trade-offs in the fair determination of risk scores". In: *arXiv preprint arXiv:1609.05807.*

Le Gouic, T., J.-M. Loubes, and P. Rigollet (2020). "Projection to Fairness in Statistical Learning". In: *arXiv preprint arXiv:2005.11720.*

Wick, Michael, Jean-Baptiste Tristan, et al. (2019). "Unlocking fairness: a trade-off revisited". In: *Advances in neural information processing systems* 32.

Zafar, Muhammad Bilal et al. (2017). "Fairness constraints: Mechanisms for fair classification". In: *Artificial intelligence and statistics.* PMLR, pp. 962–970.

Zliobaite, Indre (2015). "On the relation between accuracy and fairness in binary classification". In: *arXiv preprint arXiv:1505.05723.*