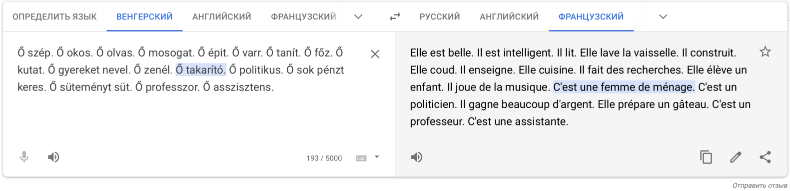
Algorithmic Fairness : regression with demographic parity constraints

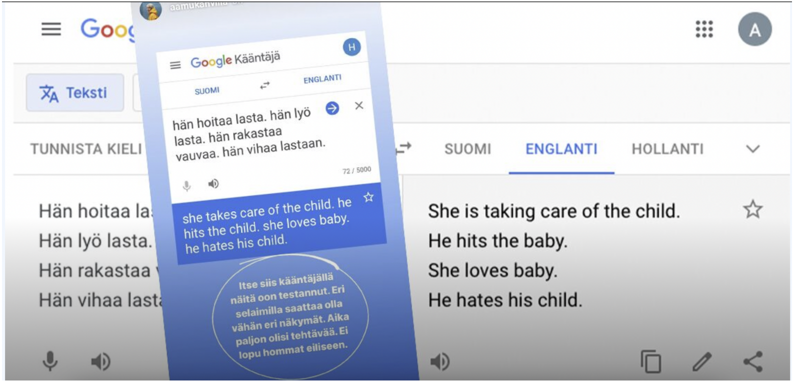
Evgenii Chzhen

Motivating examples: Google translate



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

Motivating examples: Google translate



Source https://kotiliesi.fi/ihmiset-ja-ilmiot/ilmiot/miksi-google-kaantajan-mukaan-mies-johtaa-ja-mies-tiskaa/

Motivating examples: Twitter cropping

Fact: Twitter automatically crops large images in order to fit the size of an average mobile screen.

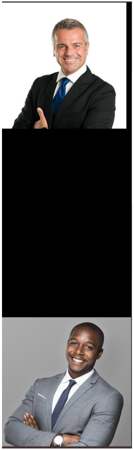
Original

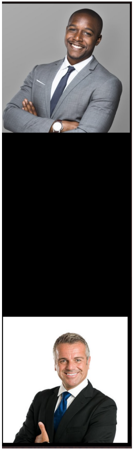
Cropped

Motivating examples: Twitter cropping

Fact: Twitter automatically crops large images in order to fit the size of an average mobile screen.

Question: How will Twitter crop these two images??





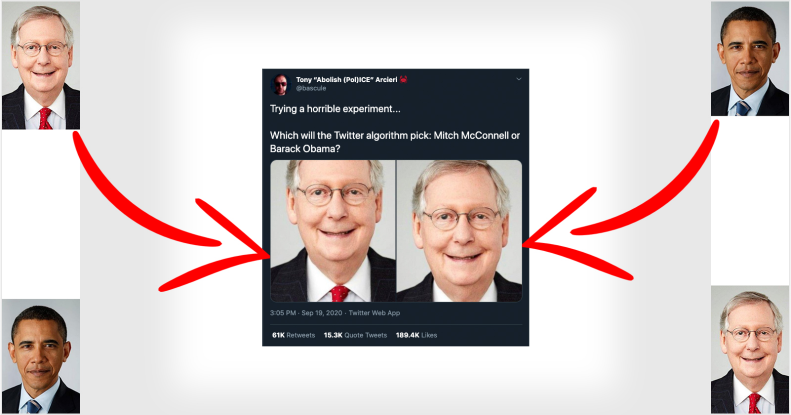
Motivating examples: Twitter cropping

Fact: Twitter automatically crops large images in order to fit the size of an average mobile screen.

Twitter's response: (https://blog.twitter.com/en_us/topics/product/2020/transparency-image-cropping.html)

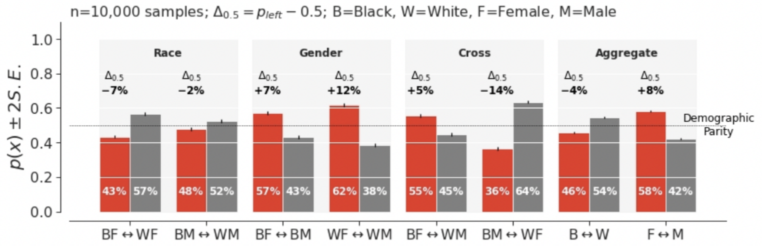
Motivating examples: Twitter cropping

Fact: Twitter automatically crops large images in order to fit the size of an average mobile screen.



Twitter's response: (https://blog.twitter.com/en_us/topics/product/2020/transparency-image-cropping.html)

Motivating examples: Twitter cropping



Source:

`https://blog.twitter.com/engineering/en_us/topics/insights/2021/sharing-learnings-about-our-image-cropping-algorithm`

More details in associated paper (Yee, Tantipongpipat, and Mishra, 2021)

Today's plan

1. Individual fairness
2. Group fairness
    2.1 Definitions / vocabulary for binary classification
    2.2 Types of approaches
3. Regression with demographic parity constraint

Individual fairness paradigms

"*treat like cases as like*" ($\leq$ Aristotel)
"*Ensure that similar individuals are treated similarly*" (Dwork et al., 2012)

Example. Consider binary classification problem, where observations are of the form $(\boldsymbol{x}, y) \in \mathcal{X} \times \{0, 1\}$. Individual fairness *always* considers randomized predictions $f : \mathcal{X} \to \Delta(\{0, 1\})$

Individual fairness paradigms

"*treat like cases as like*" ($\leq$ Aristotel)
"*Ensure that similar individuals are treated similarly*" (Dwork et al., 2012)

Example. Consider binary classification problem, where observations are of the form $(\boldsymbol{x}, y) \in \mathcal{X} \times \{0, 1\}$. Individual fairness *always* considers randomized predictions $f : \mathcal{X} \to \Delta(\{0, 1\})$

1. **Similarity of predictions:** $D : \Delta(\{0, 1\}) \times \Delta(\{0, 1\}) \to \mathbb{R}_+$
2. **Similarity of individuals:** $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$

A prediction $f : \mathcal{X} \to \Delta(\{0, 1\})$ is called *perfectly* (D, d)-individually fair if $\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$

$$\boxed{D(f(\boldsymbol{x}_1), f(\boldsymbol{x}_2)) \leq d(\boldsymbol{x}_1, \boldsymbol{x}_2)}$$

Individual fairness paradigms

"*treat like cases as like*" ($\leq$ Aristotel)
"*Ensure that similar individuals are treated similarly*" (Dwork et al., 2012)

Example. Consider binary classification problem, where observations are of the form $(\boldsymbol{x}, y) \in \mathcal{X} \times \{0, 1\}$. Individual fairness *always* considers randomized predictions $f : \mathcal{X} \to \Delta(\{0, 1\})$

1. **Similarity of predictions:** $D : \Delta(\{0,1\}) \times \Delta(\{0,1\}) \to \mathbb{R}_+$
2. **Similarity of individuals:** $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$

A prediction $f : \mathcal{X} \to \Delta(\{0, 1\})$ is called *perfectly* (D, d)-individually fair if $\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$

$$\boxed{D(f(\boldsymbol{x}_1), f(\boldsymbol{x}_2)) \leq d(\boldsymbol{x}_1, \boldsymbol{x}_2)}$$

A prediction $f : \mathcal{X} \to \Delta(\{0, 1\})$ is called *approximately* (D, d, α, γ)-individually fair if

$$\boxed{\mathbf{P}_{(\boldsymbol{X}_1, \boldsymbol{X}_2)}\left(D(f(\boldsymbol{X}_1), f(\boldsymbol{X}_2)) > d(\boldsymbol{X}_1, \boldsymbol{X}_2) + \gamma\right) \leq \alpha}$$

where $\boldsymbol{X}_1, \boldsymbol{X}_2$ are independent copies of $\boldsymbol{X}$ (Rothblum and Yona, 2018).

Group fairness paradigm

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

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)

Risk: $f \mapsto \mathcal{R}(f)$

- classification: $\mathcal{R}(f) = \mathbb{P}(Y \neq f(\boldsymbol{Z}))$
- regression: $\mathcal{R}(f) = \mathbb{E}(Y - f(\boldsymbol{Z}))^2$

Fairness criteria – dichotomy of prediction functions: which functions we call fair? There are a lot of definitions.

Connections of ML fairness notions with political philosophy (Heidari et al., 2019)

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 $\boldsymbol{X}|S$.
  4. Constant predictions satisfy DP

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 $\boldsymbol{X}|S$.
  4. Constant predictions satisfy DP

- Equalized Odds (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
  3. Constant predictions satisfy Equalized Odds

Popular definitions of fair classifiers

- Equal Opportunity (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

- Equal Opportunity (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

- Test fairness (Chouldechova, 2017)

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

  1. Y independent from S conditionally on $f(\boldsymbol{Z}) = 1$.
  2. Closely related to per-group calibration.

Global view on group fairness constraints

Most of the definitions of fairness fall inside or try to reflect only 3 criteria

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)

N.B. Sometimes we consider a score function $f(\boldsymbol{Z}) \in [0, 1]$. Above notions applied in this case ensure that any threshold will result in fair classification : incurs higher drop in accuracy; used in regression.

Taken from Chapter 2 of (Barocas, Hardt, and Narayanan, 2019)

Impossibilities for score functions

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)

- If S and Y are not independent, then sufficiency and independence cannot both hold.
- If $Y \in \{0, 1\}$, S and Y are not independent, $f(\boldsymbol{Z})$ is not independent from Y, then independence and separation cannot both hold.
- If S and Y are not independent, and $\mathbb{P}(Y = 1) \in (0, 1)$, then separation and sufficiency cannot both hold.

Taken from Chapter 2 of (Barocas, Hardt, and Narayanan, 2019)
propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Impossibilities for score functions

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)

- If S and Y are not independent, then sufficiency and independence cannot both hold.
- If $Y \in \{0, 1\}$, S and Y are not independent, $f(\boldsymbol{Z})$ is not independent from Y, then independence and separation cannot both hold.
- If S and Y are not independent, and $\mathbb{P}(Y = 1) \in (0, 1)$, then separation and sufficiency cannot both hold.

A fact: famous example of COMPAS nearly satisfied sufficiency, but failed to satisfy separation. Due to the latter propublica published an article that extremely influenced the field of algorithmic fairness (Chouldechova, 2017).

Taken from Chapter 2 of (Barocas, Hardt, and Narayanan, 2019)
propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Three (rough) types of methods: pre-processing

Pre-processing – Fair representation
Find a mapping $\boldsymbol{Z} \mapsto \hat{\varphi}(\boldsymbol{Z})$ such that

$$\hat{\varphi}(\boldsymbol{Z}) \perp\!\!\!\perp S$$

then use any method on the representation.

A guarantee on finite sample can look like

$$\mathrm{KS}\Big(\mathrm{Law}(\hat{\varphi}(\boldsymbol{Z}) \mid S = 0, \mathrm{data}), \mathrm{Law}(\hat{\varphi}(\boldsymbol{Z}) \mid S = 1, \mathrm{data})\Big) \text{ is small}$$

Typically, (unsupervised) optimal fair representation is defined as

$$\varphi^* \in \arg\min \left\{\mathbb{E}[d(\boldsymbol{X}, \varphi(\boldsymbol{Z}))] \,:\, \varphi(\boldsymbol{Z}) \perp\!\!\!\perp S\right\} \quad .$$

Three (rough) types of methods: in-processing

In-processing (Agarwal et al., 2018; Donini et al., 2018)

$$f^*_{\mathcal{F}} \in \arg\min_{f \in \mathcal{F}} \{\mathcal{R}(f) \,:\, f(\boldsymbol{Z}) \perp\!\!\!\perp S\}$$

In-processing type method: Given data $(\boldsymbol{X}_1, S_1, Y_1), \ldots, (\boldsymbol{X}_n, S_n, Y_n)$ build an estimator $\hat{f}$ as a solution

$$\min_{f \in \mathcal{F}} \left\{\hat{\mathcal{R}}(f) + \lambda_0 \cdot \Omega_{\text{compl}}(f) + \lambda_1 \cdot \Omega_{\text{fairness}}(f)\right\}$$

1. Often methods with good guarantees are not tractable
2. Often tractable methods are not supported by guarantees

Three (rough) types of methods: in-processing

In-processing (Agarwal et al., 2018; Donini et al., 2018)

$$f^*_{\mathcal{F}} \in \operatorname*{arg\,min}_{f \in \mathcal{F}} \{\mathcal{R}(f) \,:\, f(\boldsymbol{Z}) \perp\!\!\!\perp S\}$$

In-processing type method: Given data $(\boldsymbol{X}_1, S_1, Y_1), \ldots, (\boldsymbol{X}_n, S_n, Y_n)$ build an estimator $\hat{f}$ as a solution

$$\min_{f \in \mathcal{F}} \left\{\hat{\mathcal{R}}(f) + \lambda_0 \cdot \Omega_{\text{compl}}(f) + \lambda_1 \cdot \Omega_{\text{fairness}}(f)\right\}$$

1. Often methods with good guarantees are not tractable
2. Often tractable methods are not supported by guarantees

N.B. There might be an issue of existence of non-trivial solutions, especially if $\boldsymbol{Z} = \boldsymbol{X}$. For instance if $\mathcal{F}$ is the family of linear classifiers (linear regression), and $\boldsymbol{X} \mid S$ are Gaussians we can end-up with constant $f^*_{\mathcal{F}}$, even if the underlying data comes from linear model.

Three (rough) types of methods: post-processing

Post-processing: given data, base algorithm h, find a transformation

$$h \mapsto \hat{T}(h) \ ,$$

so that $\hat{T}(h)$ satisfies your fairness constraint.

Three (rough) types of methods: post-processing

Post-processing: given data, base algorithm h, find a transformation

$$h \mapsto \hat{T}(h) \ ,$$

so that $\hat{T}(h)$ satisfies your fairness constraint.
Typical algorithm construction is based on the connection between

$$h^*_{\text{fair}} \in \operatorname*{arg\,min}_{h:\mathcal{Z}\to\mathcal{Y}} \{\mathcal{R}(h) \,:\, h \text{ is fair}\} \quad \text{and} \quad h^*_{\text{Bayes}} \in \operatorname*{arg\,min}_{h:\mathcal{Z}\to\mathcal{Y}} \mathcal{R}(h)$$

In particular, often you can show that

$$h^*_{\text{fair}} = T^*(h^*_{\text{Bayes}}) \ ,$$

treat the base algorithm h as if it were a Bayes and estimate T^*

Regression with Demographic Parity

joint works with C. Denis, M. Hebiri, L. Oneto, M. Pontil, and N. Schreuder

Regression + Demographic Parity

$$(\underbrace{\text{feature}}_{\boldsymbol{X}}, \underbrace{\text{sensitive attribute}}_{S}, \underbrace{\text{signal}}_{Y}) \sim \mathbb{P} \text{ on } \mathbb{R}^d \times \underbrace{\mathcal{S}}_{=\{1,\ldots,K\}} \times \mathbb{R}$$

Prediction: $f : \mathbb{R}^d \times \mathcal{S} \to \mathbb{R}$

Risk: $\mathcal{R}(f) = \mathbb{E}(f^*(\boldsymbol{X}, S) - f(\boldsymbol{X}, S))^2$ where $f^* = \mathbb{E}[Y \mid \boldsymbol{X}, S]$

Demographic Parity fairness

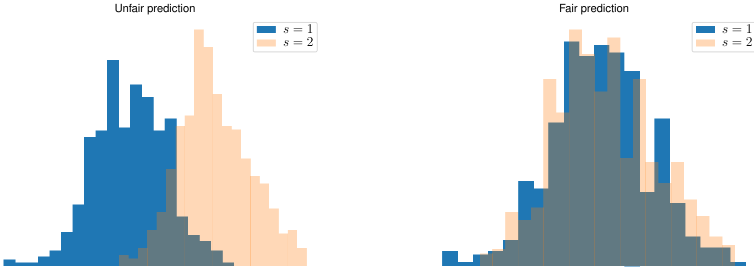
$$f(\boldsymbol{X}, S) \perp\!\!\!\perp S$$

Optimal fair prediction:

$$f_0^* \in \arg\min \{\mathcal{R}(f) \ : \ f(\boldsymbol{X}, S) \perp\!\!\!\perp S\}$$

An illustration and main assumption

$$f(\boldsymbol{X}, S) \perp\!\!\!\perp S$$



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 $\mathcal{S}$.

Main insight

Optimal fair: $f_0^* \in \underset{f:\mathbb{R}^d \times \mathcal{S} \to \mathbb{R}}{\arg\min} \left\{\mathcal{R}(f) \,:\, f(\boldsymbol{X}, S) \perp\!\!\!\perp S\right\}$

Bayes optimal: $f^* \in \underset{f:\mathbb{R}^d \times \mathcal{S} \to \mathbb{R}}{\arg\min} \ \mathcal{R}(f)$

Question: is there a link between f_0^* and f^*?

Theorem (informal with $\mathcal{S} = \{1, 2\}$)

Set $w_s = \mathbb{P}(S{=}s)$. Let Assumption (A) be satisfied, then

$$\mathrm{Law}(f_0^*(\boldsymbol{X}, S)) = \underbrace{\underset{\nu \in \mathcal{P}_2(\mathbb{R})}{\arg\min} \sum_{s \in \mathcal{S}} w_s W_2^2\Big(\mathrm{Law}(f^*(\boldsymbol{X}, S) \mid S = s),\ \nu\Big)}_{\text{Wasserstein barycenter problem}}\ ,$$

$$f_0^*(\boldsymbol{x}, 1) = w_1 f^*(\boldsymbol{x}, 1) + w_2 T^*_{1\to 2} \circ f^*(\boldsymbol{x}, 1), \qquad \forall \boldsymbol{x} \in \mathbb{R}^d\ ,$$

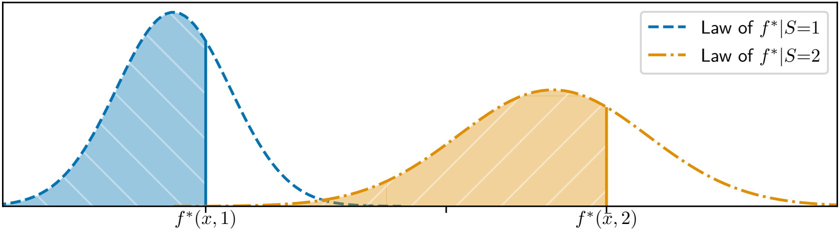
$T^*_{1\to 2}$ – optimal transport map from $\mathrm{Law}(f^* \mid S = 1)$ to $\mathrm{Law}(f^* \mid S = 2)$.

(C. et al., 2020; Le Gouic, Loubes, and Rigollet, 2020)

Interpretation for $\mathcal{S} = \{1, 2\}$

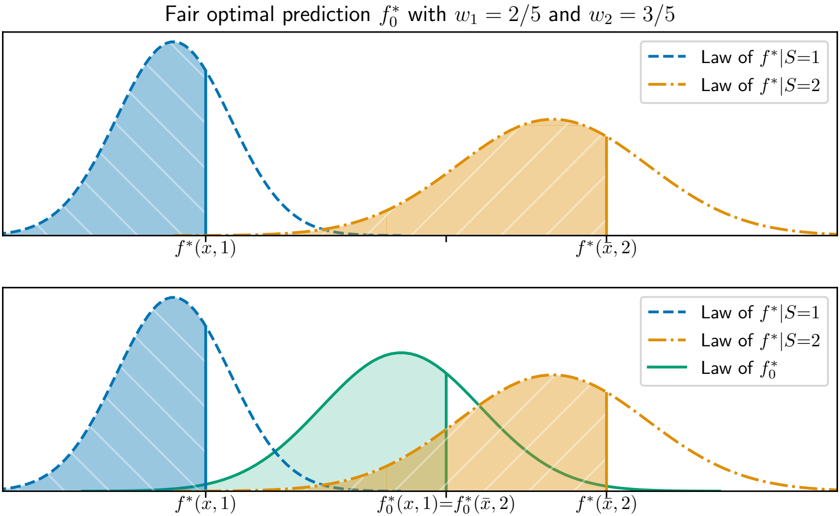
Fair optimal: $f_0^*(\boldsymbol{x}, 1) = w_1 f^*(\boldsymbol{x}, 1) + w_2 F^{-1}_{f^*|S=2} \circ F_{f^*|S=1} \circ f^*(\boldsymbol{x}, 1)$

Fair optimal prediction f_0^* with $w_1 = 2/5$ and $w_2 = 3/5$



Interpretation for $\mathcal{S} = \{1, 2\}$

Fair optimal: $f_0^*(\boldsymbol{x}, 1) = w_1 f^*(\boldsymbol{x}, 1) + w_2 F^{-1}_{f^*|S=2} \circ F_{f^*|S=1} \circ f^*(\boldsymbol{x}, 1)$



Generic post-processing estimator ($\mathcal{S} = \{1, 2\}$)

Fair optimal: $f_0^*(\boldsymbol{x}, 1) = w_1 f^*(\boldsymbol{x}, 1) + w_2 T^*_{1\to 2} \circ f^*(\boldsymbol{x}, 1)$

- Base estimator: $\hat{f} : \mathbb{R}^d \times \{1, 2\} \to \mathbb{R}$ trained independently from the following data.
- Unlabeled data: $\forall s \in \mathcal{S}$ we observe $\boldsymbol{X}_1^s, \dots, \boldsymbol{X}_{N_s}^s \overset{i.i.d.}{\sim} \mathbb{P}_{\boldsymbol{X} \mid S=s}$

Meta algo:

1. estimate $w_s = \mathbb{P}(S = s)$
2. estimate transport maps $T^*_{1\to 2}$ and $T^*_{2\to 1}$ using unlabeled data and base estimator

Generic post-processing estimator ($\mathcal{S} = \{1, 2\}$)

Fair optimal: $f_0^*(\boldsymbol{x}, 1) = w_1 f^*(\boldsymbol{x}, 1) + w_2 T^*_{1\to 2} \circ f^*(\boldsymbol{x}, 1)$

- Base estimator: $\hat{f} : \mathbb{R}^d \times \{1, 2\} \to \mathbb{R}$ trained independently from the following data.
- Unlabeled data: $\forall s \in \mathcal{S}$ we observe $\boldsymbol{X}_1^s, \dots, \boldsymbol{X}_{N_s}^s \overset{i.i.d.}{\sim} \mathbb{P}_{\boldsymbol{X}|S=s}$

Meta algo:
1. estimate $w_s = \mathbb{P}(S = s)$
2. estimate transport maps $T^*_{1\to 2}$ and $T^*_{2\to 1}$ using unlabeled data and base estimator

Put together:
3. $\hat{f}_0(\boldsymbol{x}, 1) = \hat{w}_1 \hat{f}(\boldsymbol{x}, 1) + \hat{w}_2 \hat{T}_{1\to 2} \circ \hat{f}(\boldsymbol{x}, 1)$

Theoretical guarantees

Theorem (informal)

For any joint distribution $\mathbb{P}$ of $(\boldsymbol{X}, S, Y)$, any base estimator $\hat{f}$ it holds that

$$\mathbf{E}\left[\sup_{t \in \mathbb{R}}\left|\mathbf{P}(\hat{f}_0(\boldsymbol{X}, S) \leq t | S{=}1, \mathcal{D}) - \mathbf{P}(\hat{f}_0(\boldsymbol{X}, S) \leq t | S{=}2, \mathcal{D})\right|\right] \lesssim \frac{1}{\sqrt{N_1 \wedge N_2}}$$

Under additional assumptions on $\mathbb{P}$ we have

$$\mathbf{E}\|\hat{f}_0 - f_0^*\|_1 \lesssim \underbrace{\mathbf{E}\|\hat{f} - f^*\|_1}_{\text{quality of base estimator}} \bigvee \underbrace{\sum_{s \in \mathcal{S}} p_s N_s^{-1/2}}_{\text{transport estimation}}$$

(C. et al., 2020)

Additional assumptions: $(f^*(\boldsymbol{X}, S) \mid S = s)$ admits density which is upper and lower bounded (leading constant for the risk rate depends on this upper/lower bound).

N_1 and N_2 – number of *unlabeled* samples from $\mathbb{P}_{\boldsymbol{X}|S=1}$ and $\mathbb{P}_{\boldsymbol{X}|S=2}$

How to measure unfairness ?

Demographic Parity: $f(\boldsymbol{X}, S) \perp\!\!\!\perp S$

- **Problem:** too stiff — either fair or unfair.
- **Question:** how to quantify unfairness *i.e.*, violation of DP?
- **Question:** how to trade accuracy for fairness?

Popular measure is based on KS distance (Agarwal, Dudik, and Wu, 2019; Oneto, Donini, and Pontil, 2019)

$$\mathcal{U}_{\mathrm{KS}}(f) := \sum_{s \in \mathcal{S}} \mathrm{KS}\left(\mathrm{Law}(f(\boldsymbol{X}, S) \mid S = s), \mathrm{Law}(f(\boldsymbol{X}, S))\right)$$

How to measure unfairness ?

Demographic Parity: $f(\boldsymbol{X}, S) \perp\!\!\!\perp S$

- **Problem:** too stiff — either fair or unfair.
- **Question:** how to quantify unfairness *i.e.*, violation of DP?
- **Question:** how to trade accuracy for fairness?

Popular measure is based on KS distance (Agarwal, Dudik, and Wu, 2019; Oneto, Donini, and Pontil, 2019)

$$\mathcal{U}_{\mathrm{KS}}(f) := \sum_{s \in \mathcal{S}} \mathrm{KS}\left(\mathrm{Law}(f(\boldsymbol{X}, S) \mid S = s), \mathrm{Law}(f(\boldsymbol{X}, S))\right)$$

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

From previous result: $\mathcal{R}(f_0^*) = \mathcal{U}(f^*)$

Improving unfairness oracles

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

- f^*_α – ${}^1/_\alpha$ times fairer than f^*.
- f^*_0 – optimal DP fair prediction.
- $f^*_1 \equiv f^*$ – Bayes optimal prediction.

Improving unfairness oracles

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

- f^*_α – ${}^1\!/\!_\alpha$ times fairer than f^*.
- f^*_0 – optimal DP fair prediction.
- $f^*_1 \equiv f^*$ – Bayes optimal prediction.

Theorem

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

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

$$\alpha\text{-RI} \equiv \sqrt{\alpha} \cdot \text{Bayes optimal} + (1 - \sqrt{\alpha}) \cdot \text{Fair optimal}$$

(C. and Schreuder, 2020)

N.B. We can use previous algorithm to estimate f^*_0 and *any* standard algorithm for estimation of f^*

Idea of the proof

Goal: $$\min_{f:\mathcal{Z}\to\mathbb{R}} \left\{ \sum_{s=1}^{K} w_s \mathbb{E}[(f(\boldsymbol{X},S) - f^*(\boldsymbol{X},S))^2 \mid S=s] \;:\; \mathcal{U}(f) \leq \alpha\,\mathcal{U}(f^*) \right\}$$

LB: $$\sum_{s=1}^{K} w_s \mathsf{W}_2^2 \left(\mathrm{Law}(f(\boldsymbol{X},S)|S=s), \mathrm{Law}(f^*(\boldsymbol{X},S)|S=s)\right)$$

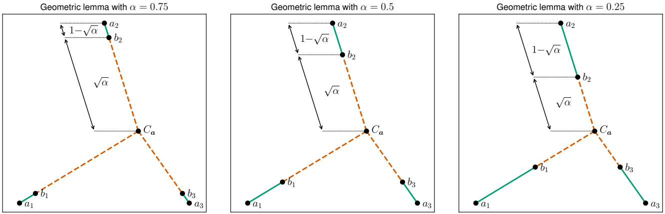
Idea of the proof

Goal: $$\min_{f:\mathcal{Z}\to\mathbb{R}} \left\{ \sum_{s=1}^{K} w_s \mathbb{E}[(f(\boldsymbol{X},S) - f^*(\boldsymbol{X},S))^2 \mid S=s] \;:\; \mathcal{U}(f) \le \alpha\,\mathcal{U}(f^*) \right\}$$

LB: $$\sum_{s=1}^{K} w_s \mathsf{W}_2^2\left(\mathrm{Law}(f(\boldsymbol{X},S)|S=s), \mathrm{Law}(f^*(\boldsymbol{X},S)|S=s)\right)$$

New problem

$$\min_{\boldsymbol{b}\in\mathcal{P}_2^K(\mathbb{R})} \left\{ \sum_{s=1}^{K} w_s \mathsf{W}_2^2\left(b_s, a_s\right) \;:\; \sum_{s=1}^{K} w_s \mathsf{W}_2^2(b_s, C_{\boldsymbol{b}}) \le \alpha \sum_{s=1}^{K} w_s \mathsf{W}_2^2(a_s, C_{\boldsymbol{a}}) \right\}$$



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

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

(C. and Schreuder, 2020)

Risk/fairness trade-off

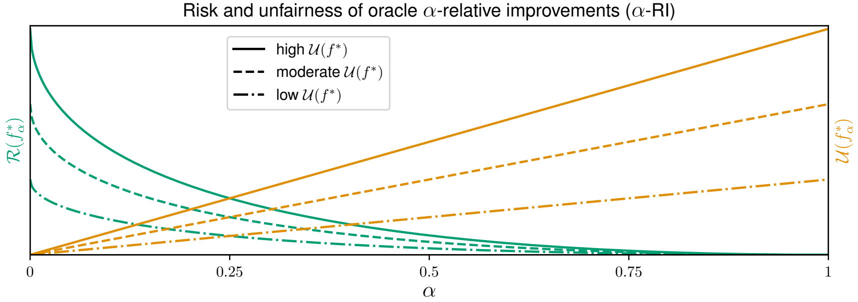
α-Relative Improvement $\quad f^*_\alpha \in \arg\min\left\{\mathcal{R}(f) \,:\, \boxed{\mathcal{U}(f) \leq \alpha\,\mathcal{U}(f^*)}\right\}$

Proposition

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

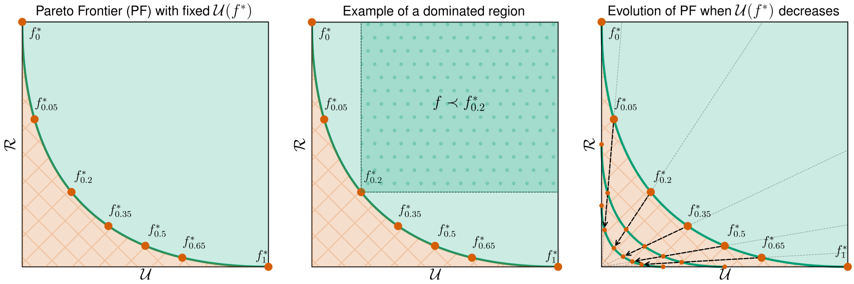
$$\mathcal{R}(f^*_\alpha) = (1-\sqrt{\alpha})^2\boxed{\mathcal{U}(f^*)} \quad \text{and} \quad \mathcal{U}(f^*_\alpha) = \alpha\boxed{\mathcal{U}(f^*)}$$

(C. and Schreuder, 2020)



Pareto efficiency

- Multi-objective optimization: $\min_{f:\mathcal{Z}\to\mathbb{R}} \Big(\mathcal{U}(f), \mathcal{R}(f)\Big)$.
- Each prediction f defines a point $(\mathcal{U}(f), \mathcal{R}(f))$.
- f is dominated by f' iff $\mathcal{R}(f') \leq \mathcal{R}(f)$ and $\mathcal{U}(f') \leq \mathcal{U}(f)$.



Minimax statistical framework

Data: $(\boldsymbol{X}_1, S_1, Y_1), \ldots, (\boldsymbol{X}_n, S_n, Y_n) \overset{i.i.d.}{\sim} \mathbf{P}_{(f^*,\boldsymbol{\theta})}$, $(f^*, \boldsymbol{\theta}) \in \mathcal{F} \times \Theta$
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:

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

2. Uniform risk guarantee:

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

Problem-dependent lower bound

For $t \in (0,1)$, let $\delta_n(\mathcal{F}, \Theta, t)$ be a sequence that verifies

$$\inf_{\hat{f}} \sup_{(f^*,\boldsymbol{\theta}) \in \mathcal{F}\times\Theta} \mathbf{P}_{(f^*,\boldsymbol{\theta})} \left(\mathcal{R}(\hat{f}) \geq \delta_n(\mathcal{F}, \Theta, t) \right) \geq t$$

Theorem

Any estimator $\hat{f}$ satisfying

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

verifies

$$\boxed{\sup_{\substack{f^* \in \mathcal{F} \\ \boldsymbol{\theta} \in \Theta}} \mathbf{P}_{(f^*,\boldsymbol{\theta})} \left(\mathcal{R}^{1/2}(\hat{f}) \geq \delta_n^{1/2}(\mathcal{F}, \Theta, t) \vee \underbrace{(1-\sqrt{\alpha})\mathcal{U}^{1/2}(f^*)}_{=\mathcal{R}^{1/2}(f^*_\alpha)} \right) \geq t \wedge (1 - t')}$$

Conclusions

1. Individual fairness – predict with Lipschitz functions

$$D(f(\boldsymbol{x}), f(\boldsymbol{x}')) \leq d(\boldsymbol{x}, \boldsymbol{x}')$$

2. Group fairness – enforce some independence criterion

$$f(\boldsymbol{Z}) \perp\!\!\!\perp S, \qquad (f(\boldsymbol{Z}) \perp\!\!\!\perp S) \mid Y, \qquad (Y \perp\!\!\!\perp S) \mid f(\boldsymbol{Z})$$

3. Regression with demographic parity ($f(\boldsymbol{Z}) \perp\!\!\!\perp S$) can be characterized by Wasserstein barycenter problem

$$\mathcal{R}(f_0^*) = \mathcal{U}(f^*)$$

4. Risk/fairness trade-off can be characterized explicitly for introduced notion of unfairness

Thank you for your attention

EUROPEAN COMMISSION

Brussels, 21.4.2021

COM(2021) 206 final

2021/0106(COD)

Proposal for a

REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS

Questions?

PROHIBITED ARTIFICIAL INTELLIGENCE PRACTICES

Article 5

1. The following artificial intelligence practices shall be prohibited:

   (a) the placing on the market, putting into service or use of an AI system that deploys subliminal techniques beyond a person's consciousness in order to materially distort a person's behaviour in a manner that causes or is likely to cause that person or another person physical or psychological harm;

   (b) the placing on the market, putting into service or use of an AI system that exploits any of the vulnerabilities of a specific group of persons due to their age, physical or mental disability, in order to materially distort the behaviour of a person pertaining to that group in a manner that causes or is likely to cause that person or another person physical or psychological harm;

   (c) the placing on the market, putting into service or use of AI systems by public authorities or on their behalf for the evaluation or classification of the trustworthiness of natural persons over a certain period of time based on their social behaviour or known or predicted personal or personality characteristics, with the social score leading to either or both of the following:

      (i) detrimental or unfavourable treatment of certain natural persons or whole groups thereof in social contexts which are unrelated to the contexts in which the data was originally generated or collected;

      (ii) detrimental or unfavourable treatment of certain natural persons or whole groups thereof that is unjustified or disproportionate to their social behaviour or its gravity;

Bibliography I

Agarwal, A. et al. (2018). “A reductions approach to fair classification”. In: *arXiv preprint arXiv:1803.02453*.

Agarwal, Alekh, Miroslav Dudik, and Zhiwei Steven Wu (2019). “Fair regression: Quantitative definitions and reduction-based algorithms”. In: *International Conference on Machine Learning*. PMLR, pp. 120–129.

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

C., E. and N. Schreuder (2020). “A minimax framework for quantifying risk-fairness trade-off in regression”. In: *arXiv preprint arXiv:2007.14265*.

C., E et al. (2020). “Fair Regression with Wasserstein Barycenters”. In: *NeurIPS 2020*.

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

Chouldechova, Alexandra (2017). “Fair prediction with disparate impact: A study of bias in recidivism prediction instruments”. In: *Big data* 5.2, pp. 153–163.

Donini, M. et al. (2018). “Empirical risk minimization under fairness constraints”. In: *Neural Information Processing Systems*.

Bibliography II

Dwork, Cynthia et al. (2012). "Fairness through awareness". In: *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226.

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

Heidari, Hoda et al. (2019). "A moral framework for understanding fair ML through economic models of equality of opportunity". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 181–190.

Le Gouic, Thibaut, Jean-Michel Loubes, and Philippe Rigollet (2020). "Projection to fairness in statistical learning". In: *arXiv e-prints*, arXiv–2005.

Oneto, L., M. Donini, and M. Pontil (2019). "General Fair Empirical Risk Minimization". In: *arXiv preprint arXiv:1901.10080*.

Rothblum, Guy N and Gal Yona (2018). "Probably Approximately Metric-Fair Learning". In: *arXiv e-prints*, arXiv–1803.

Yee, K., U. Tantipongpipat, and S. Mishra (2021). *Image Cropping on Twitter: Fairness Metrics, their Limitations, and the Importance of Representation, Design, and Agency*. arXiv: `2105.08667 [cs.CY]`.

Bibliography III

Agarwal, A. et al. (2018). “A reductions approach to fair classification”. In: *arXiv preprint arXiv:1803.02453.*

Agarwal, Alekh, Miroslav Dudik, and Zhiwei Steven Wu (2019). “Fair regression: Quantitative definitions and reduction-based algorithms”. In: *International Conference on Machine Learning*. PMLR, pp. 120–129.

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

C., E. and N. Schreuder (2020). “A minimax framework for quantifying risk-fairness trade-off in regression”. In: *arXiv preprint arXiv:2007.14265.*

C., E et al. (2020). “Fair Regression with Wasserstein Barycenters”. In: *NeurIPS 2020.*

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

Chouldechova, Alexandra (2017). “Fair prediction with disparate impact: A study of bias in recidivism prediction instruments”. In: *Big data* 5.2, pp. 153–163.

Donini, M. et al. (2018). “Empirical risk minimization under fairness constraints”. In: *Neural Information Processing Systems.*

Bibliography IV

Dwork, Cynthia et al. (2012). "Fairness through awareness". In: *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226.

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

Heidari, Hoda et al. (2019). "A moral framework for understanding fair ML through economic models of equality of opportunity". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 181–190.

Le Gouic, Thibaut, Jean-Michel Loubes, and Philippe Rigollet (2020). "Projection to fairness in statistical learning". In: *arXiv e-prints*, arXiv–2005.

Oneto, L., M. Donini, and M. Pontil (2019). "General Fair Empirical Risk Minimization". In: *arXiv preprint arXiv:1901.10080.*

Rothblum, Guy N and Gal Yona (2018). "Probably Approximately Metric-Fair Learning". In: *arXiv e-prints*, arXiv–1803.

Yee, K., U. Tantipongpipat, and S. Mishra (2021). *Image Cropping on Twitter: Fairness Metrics, their Limitations, and the Importance of Representation, Design, and Agency.* arXiv: `2105.08667 [cs.CY]`.