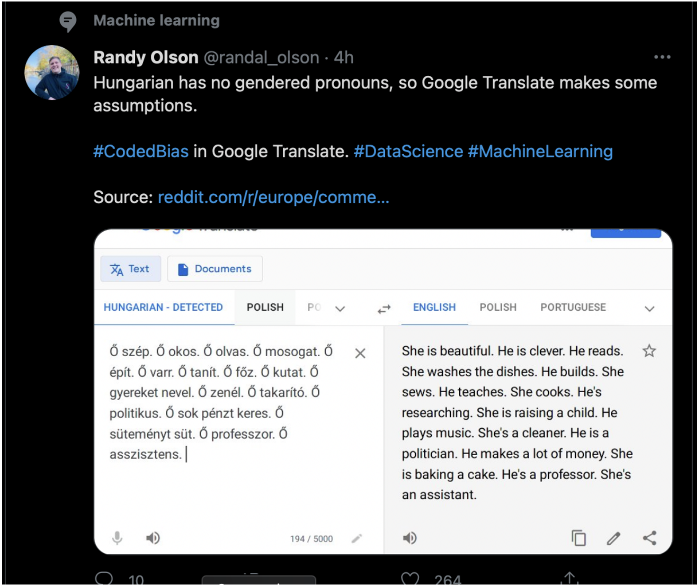
Algorithmic fairness
lecture note (short version)

Christophe Denis

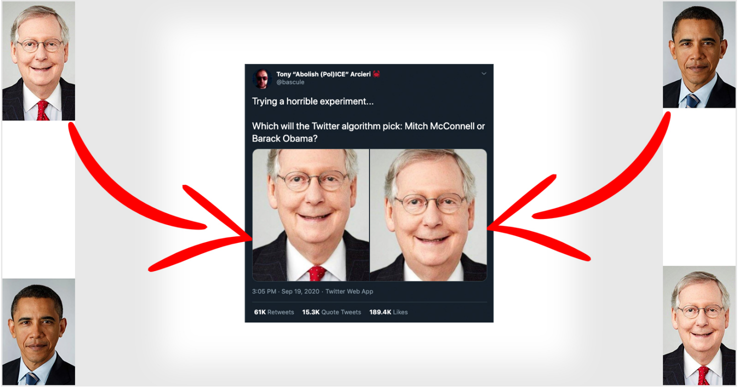
M2A and M2Stat
Sorbonne Université

Introduction and definitions

Bias in machine learning: example



Bias in machine learning: example



Historical example: COMPAS dataset

Correctional Offender Management Profiling for Alternative Sanctions

- risk-assessment software developed and owned by Northpointe
- the software is used in U.S. courts to predict recidivism risks of defendants
- investigation of `Propublica`

  "Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes."

Another example: modern redlining

- **Wikipedia**

  "Redlining is a discriminatory practice in which services (financial and otherwise) are withheld from potential customers who reside in neighborhoods classified as "hazardous" to investment; these neighborhoods have significant numbers of racial and ethnic minorities, and low-income residents."

- **Barocas *et al.* (2019)**

  " Amazon uses a data-driven system to determine the neighbourhoods in which to offer free same-day delivery. A 2016 study found stark disparities in the demographic make-up of these neighbourhoods: in many U.S. cities, white residents were more than twice as likely as black residents to live in one of the qualifying neighbourhoods."

Algorithmic Fairness

Motivation

- mitigate the bias contained in historical data
- reduce influence of a sensitive attributes in prediction
- algorithm should treat people without discrimination based on sensitive attributes
- lot of attention in recent years Calders *et al.* (2009), Zemel *et al.* (2013), Zafar *et al.*, Donini *et al.* (2018), Agarwal *et al* (2018), Barocas *et al.* (2019), . . .

Application

- social sciences (university admission)
- insurance (credit scoring)
- artificial intelligence, . . .

Formalization of fairness in supervised learning (1/2)

Observation

- features: (X, S), output $Y \in \mathcal{Y}$
- $S \in \mathcal{S}$ sensitive (or protected) attribute

Sensitive attributes

- any decisions based on S are undesirable from an ethical or legal perspective
- French law number 2008-496, Article 1

  "Constitue une discrimination directe la situation dans laquelle, sur le fondement de son origine, de son sexe, de sa situation de famille, de sa grossesse, de son apparence physique, de la particulière vulnérabilité résultant de sa situation économique, apparente ou connue de son auteur, de son patronyme, de son lieu de résidence ou de sa domiciliation bancaire, . . . "

Fairness through awareness

- predictor $f \to$ prediction $f(X, S)$
- drawback: prediction relies on S

Fairness through unawarness

- awareness may induce direct discrimination
- predictor $f \to$ prediction $f(X)$

Disparate learning process

- S available at the training step but not for the prediction step

No fairness through unawareness

Remove S does not enforce fairness

- X and S are not independent
- indirect discrimination
- classic example $\rightarrow$ redlining
- French law number 2008-496, 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, ..."

Definition of fairness: *group fairness approach*

Definition of fairness: two approaches

- group fairness $\rightarrow$ defines fairness at population level
- other approach individual fairness: *similar* people should be treated *similarly*

Definition for group fairness ($Z = (X, S)$ or $Z = X$)

- independence $\rightarrow f(Z) \perp\!\!\!\perp S$
- separation $\rightarrow (f(Z) \perp\!\!\!\perp S) \mid Y$
- sufficiency $\rightarrow (Y \perp\!\!\!\perp S) \mid f(Z)$
- in general only one of the constraint could be achieve

Independence and separation

- assume that f satisfies independence and sufficiency
- then $Y \perp\!\!\!\perp S$

Independence and separation ($Y \in \{0, 1\}$)

- assume that f satisfies independence and separation
- then either $Y \perp\!\!\!\perp S$ or $f(Z) \perp\!\!\!\perp Y$ or both

Sufficiency and separation ($Y \in \{0, 1\}$)

- assume that f satisfies sufficiency and separation
- then either $Y \perp\!\!\!\perp S$ or $f(Z) \perp\!\!\!\perp Y$ or $\mathbb{P}(f(Z) = 1|Y = 1) = 0$

Fairness in binary classification: *group fairness approach*

Framework

- observation $Z = (X, S)$ or $Z = X$, and $Y \in \{0, 1\}$,
- classifier $f \rightarrow$ prediction $f(Z)$

Definition of fairness

- Demographic parity (DP), for each $s \in \mathcal{S}$

$$\mathbb{P}\left(f(X,S) = 1 | S = s\right) = \mathbb{P}\left(f(X,S) = 1\right)$$

- Equalized odds (EOd), for each $s \in \mathcal{S}$, and $y \in \{0, 1\}$

$$\mathbb{P}\left(f(X,S) = y | S = s, Y = y\right) = \mathbb{P}\left(f(X,S) = y | Y = y\right)$$

- Equal opportunity (EO), for each $s \in \mathcal{S}$

$$\mathbb{P}\left(f(X,S) = 1 | S = s, Y = 1\right) = \mathbb{P}\left(f(X,S) = 1 | Y = 1\right)$$

Main approaches to enforce fairness

Pre-processing

- find a feature representation $z \mapsto \phi(z)$
- such that $\phi(Z)$ independent on S

In-processing

- Based on the empirical risk minimization
- given a set of predictor $\mathcal{F}$, solve

$$f \in \arg\min_{f \in \mathcal{F}} \hat{R}(f) + \lambda \hat{C}(f),$$

with $\hat{R}(f)$ empirical risk, $\hat{C}(f)$ empirical fairness constraints

Post-processing

- given a pre-built predictor f, not necessary fair
- find a transformation $\hat{T}$
- *s.t.* $\hat{T}(f)$ satisfies a desired fairness constraint

Classification through awareness under DP constraint

Binary classification under DP constraint

Notations

- $\mathcal{S} = \{-1, 1\}$, and $\mathcal{Y} = \{0, 1\}$
- $\pi_s = \mathbb{P}(S = s) > 0$, and $\eta(X, S) = \mathbb{P}(Y = 1 | X, S)$
- classifier $f \to$ prediction $f(X, S) \in \{0, 1\}$

Problem

- DP constraint

$$\sum_{s \in \mathcal{S}} s \mathbb{P}\left(f(X, S) = 1 | S = s\right) = 0$$

- $f^* \in \arg\min_f \{\mathbb{P}(f(X, S) \neq Y), f \text{ satisfies DP}\}$
- lagrangian associated to the minimization problem

$$\mathcal{L}(f, \lambda) = \mathbb{P}\left(f(X, S) \neq Y\right) + \lambda \sum_{s \in \mathcal{S}} s \mathbb{P}(f(X, S) = 1 | S = s)$$

Duality

- weak duality (always holds)

$$\inf_f \sup_{\lambda \in \mathbb{R}} \mathcal{L}(f, \lambda) \geq \sup_{\lambda \in \mathbb{R}} \inf_f \mathcal{L}(f, \lambda)$$

- strong duality

$$f^* = \inf_f \sup_{\lambda \in \mathbb{R}}, \mathcal{L}(f, \lambda) = \sup_{\lambda \in \mathbb{R}} \inf_f \mathcal{L}(f, \lambda)$$

Mathematical tools: *subgradient*

Subgradient (h defined on $\mathbb{R}^d$)

- $g \in \partial h(x)$ iff $h(z) - h(x) \geq g^T(z - x), \;\; \forall z \in \mathbb{R}^d$
- h subdifferentiable at x if $\partial h(x) \neq \emptyset$
- if h subdiffentiable then $x^* \in \arg\min_{x \in \mathbb{R}^d} h(x)$ iff $0 \in \partial h(x^*)$

Pointwise maximum

- $z \mapsto h(z) = \max_{i=1,\ldots,M} h_i(z)$, h_i convex and differentiable
- $\partial h(x) = \mathbf{Conv}\left\{\nabla h_i(x), \;\; h_i(x) = h(x)\right\}$

Useful property

- $z \mapsto h(z, W)$ convex, $F(z) = \mathbb{E}\left[h(z, W)\right]$
- F convex and $\partial F(x) = \mathbb{E}\left[\partial h(x, W)\right]$

Optimal predictor under DP constraint: *result*

Continuity assumption

- $t \mapsto \mathbb{P}(\eta(X, s) \le t | S = s)$ is continuous

Optimal predictor

- the optimal fair classifier f^* can be characterized as

$$f^*(x, s) = \mathbb{1}_{\left\{\eta(x,s) \ge \frac{1}{2} + \frac{s\lambda^*}{2\pi_s}\right\}}$$

- λ^* are lagrange multiplier characterized as

$$\lambda^* \in \arg\min_{\lambda \in \mathbb{R}} \sum_{s \in \mathcal{S}} \mathbb{E}_{X|S=s}\left[\max\left(\pi_s\left(2\eta(X, S) - 1\right) - s\lambda), 0\right)\right]$$

Optimal predictor: *sketch of the proof (1/2)*

- for each $\lambda \in \mathbb{R}$, consider the Lagrangian $\mathcal{L}(f, \lambda)$ defined as

$$\mathcal{L}(f, \lambda) = \mathbb{P}(f(X, S) \neq Y) + \lambda \sum_{s \in \{-1,1\}} s\mathbb{P}_{X|S=s}\left(f(X, S) = 1\right)$$

- we have that $\mathcal{L}(f, \lambda)$ can be expressed as

$$\mathbb{E}\left[Y\right] - \sum_{s \in \{-1,1\}} \mathbb{E}_{X|S=s}\left[\left(\pi_s(2\eta(X, S) - 1) - s\lambda\right) f(X, S)\right]$$

- we deduce that $f^*_\lambda \in \arg\min_f \mathcal{L}(f, \lambda)$ is characterized as

$$f^*_\lambda(x, s) = \mathbb{1}_{\{\pi_s(2\eta(x,s)-1)-s\lambda\}},$$

and

$$\mathcal{L}(f^*_\lambda, \lambda) = \mathbb{E}\left[Y\right] - \sum_{s \in \mathcal{S}} \mathbb{E}_{X|S=s}\left[\max\left(\pi_s(2\eta(X, S) - 1) - s\lambda, 0\right)\right]$$

Optimal predictor: *sketch of the proof (2/2)*

- consider $\lambda^* \arg\min_\lambda H(\lambda)$ with

$$H(\lambda) = \sum_{s \in \mathcal{S}} \mathbb{E}_{X|S=s}\left[\max\left(\pi_s(2\eta(X,S) - 1) - s\lambda, 0\right)\right]$$

- observe that $\lambda^* \in \arg\max \mathcal{L}\left(\lambda, f^*_\lambda\right)$
- under continuity assumption, $\lambda \mapsto H(\lambda)$ is differentiable and the first order condition shows that $f^*_{\lambda^*}$ satisfies DP
- therefore, with the weak duality, we obtain that $f^* = f^*_{\lambda^*}$

Post-processing method: *a plug-in approach*

Objective

- estimate $f^*(x,s) = \mathbb{1}_{\left\{\eta(x,s) \geq \frac{1}{2} + \frac{s\lambda^*}{2\pi_s}\right\}}$

Plug-in approach

- labeled sample $\mathcal{D}_n$ $\rightarrow$ estimate η
- unlabeled sample $(X_1, S_1), \ldots, (X_N, S_N)$
- $\{S_1, \ldots, S_N\}$ $\rightarrow$ estimate π_s by their empirical frequencies
- $\{X_1, \ldots, X_N\}$ $\rightarrow$ estimate parameter λ^*

Randomization

- fairness guarantee requires continuity assumption
- introduce $\zeta \sim \mathcal{U}_{[0,u]}$ independent of (X,S), $u \rightarrow 0$
- $\bar{\eta}(X,S,\zeta) = \eta(X,S) + \zeta$

Randomized classifier

Randomized fair classifier

- $(X_1, \ldots, X_N) \to (X_1^s, \ldots, X_{N_s}^s)$ i.i.d. from $X|S = s$
- $(\zeta_1^s, \ldots, \zeta_{N_s}^s)$ i.i.d from ζ
- estimator $\hat{\lambda}$

$$\hat{\lambda} \in \arg\min_{\lambda \in \mathbb{R}} \sum_{s \in \mathcal{S}} \frac{1}{N_s} \sum_{i=1}^{N_s} \max\left(\pi_s(2\bar{\eta}(X_i^s, s, \zeta_i^s) - 1) - s\lambda, 0\right)$$

- resulting classifier

$$\hat{f}(x, s) = \mathbb{1}_{\left\{\hat{\eta}(x,s) \geq \frac{1}{2} + \frac{s\hat{\lambda}}{2\hat{\pi}_s}\right\}}$$

Theoretical guarantees: *fairness guarantee*

Unfairness measure

$$\mathcal{U}(f) = \left| \sum_{s \in \mathcal{S}} s \mathbb{P}\left(f(X,S) = 1 | S = s\right) \right|$$

Distribution free-result

There exists C depending only on π_s such that for any estimator $\hat{\eta}$

$$\mathbb{E}\left[\mathcal{U}(\hat{f})\right] \leq C N^{-1/2}$$

Theoretical guarantees: *consistency*

Measure of performance

- $f^* \in \arg\min_f \mathcal{R}_{\lambda^*}(f) = \mathcal{L}(f, \lambda^*)$

$$\mathcal{R}_{\lambda^*}(f) = \mathbb{P}\left(f(X,S) \neq Y\right) + \lambda^* \sum_{s \in \mathcal{S}} s\mathbb{P}(f(X,S) = 1 | S = s)$$

Theorem

Under continuity assumption

$$\mathbb{E}\left[\mathcal{R}_{\lambda^*}(\hat{f}) - \mathcal{R}_{\lambda^*}(f^*)\right] \lesssim \mathbb{E}\left[|\hat{\eta}(X,S) - \eta(X,S)|\right] + u + N^{-1/2}$$

- assume that $\hat{\eta}$ are consistent and $u \to 0$
  $\hookrightarrow$ $\hat{f}$ is consistent

Observations

- only one *labeled* sample $(X_i, S_i, Y_i), i = 1, \ldots, n$
- $(X_1, \ldots, X_n) \to (X_1^s, \ldots, X_{n_s}^s)$ i.i.d. from $X|S = s$

Fair E.R.M.

- let $\mathcal{F}$ a class of classifier
- empirical unfairness constraint, for $\varepsilon > 0$, $f \in \mathcal{F}$

$$\hat{\mathcal{U}}(f) = \left| \sum_{s \in \mathcal{S}} s \frac{1}{n_s} \sum_{i=1}^{n_s} f(X_i^s, s) \right| \leq \varepsilon$$

- empirical risk $\hat{R}(f) = \dfrac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{f(X_i, S_i) \neq Y_i\}}$
- empirical risk minimizer

$$\hat{f} \in \arg\min_f \{\hat{R}(f),\ \hat{\mathcal{U}}(f) \leq \hat{\varepsilon}\}$$

In-Processing approach: *properties*

Assumptions

- $\mathcal{F}$ with finite VC dimension $V(\mathcal{F})$, $f^* \in \mathcal{F}$
- classical result

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}f(X_i) - \mathbb{E}\left[f(X_i)\right]\right|\right] \leq C\sqrt{\frac{\log(n)}{n}}$$

Theoretical guarantees

- let $\hat{\varepsilon} \propto \sum_{s\in\mathcal{S}}\sqrt{\frac{\log(n)}{n_s}}$
- unfairness $\mathbb{E}\left[\mathcal{U}(\hat{f})\right] \leq C\sqrt{\frac{\log(n)}{n}}$
- risk bound $\mathbb{E}\left[R_{\lambda^*}(\hat{f}) - R_{\lambda^*}(f^*)\right] \leq C\sqrt{\frac{\log(n)}{n}}$

In-Processing approach: *algorithm*

Convexification

- convex surrogate of both risk and fairness constraint
- E.R.M. with convex loss Donini *et al* (2018)

Randomized classifier

- define a set of distributions over the class of classifier
- sample according to a given distribution μ
- randomized classifier
- E.R.M. with randomized classifiers Agarwal *et al* (2018)

Regression through awareness under DP constraint

Fair regression problem

- observation (X, S, Y), $Y \in \mathbb{R}$
- $Y = \eta(X, S) + \varepsilon$ with $\mathbb{E}\left[\varepsilon | X, S\right] = 0$
- prediction rule: $f : \mathbb{R}^d \times \mathcal{S} \to \mathbb{R}$
- risk $R(f) = \mathbb{E}\left[(Y - f((X, S))^2\right]$
- *exact* DP constraint

$$\sup_{t \in \mathbb{R}} |\mathbb{P}(f(X, S) \leq t | S = 1) - \mathbb{P}(f(X, S) \leq t | S = -1)| = 0$$

- *optimal fair* predictor f^* defined as

$$f^* \in \arg\min_f \{R(f), f \ \text{ satisfies } \ \mathrm{DP}\}$$

First approach: *discretization*

Discretization

- assume that $|Y| \leq 1$
- consider a grid $\mathcal{G}_L = \{\frac{l}{L},\ l = -L, \ldots, L\}$, $L > 0$
- discretized predictor $f_L(x, s) \in \mathcal{G}_L$

DP constraint for discretized predictor

- f_L^* statisfies DP *iff*

$$\max_{l \in \{-L,\ldots,L\}} \sum_{s \in \mathcal{S}} s\mathbb{P}_{X|S=s}(f_L(X,S) = \frac{l}{L}) = 0$$

- $f_L^* \arg\min_{f_L} \{R(f_L),\ \ f_L \ \ \text{satisfies} \ \ \text{DP}\}$

Approximation property

- we have

$$R(f_L^*) \leq R(f^*) + 2\frac{\sqrt{\mathrm{Var}(Y)}}{L} + \frac{1}{L^2}$$

- proposal : estimate f_L^* rather than f^*

Optimal discretized fair predictor

Continuity assumption

- $t \mapsto \mathbb{P}(\eta(X, s) \leq t | S = s)$ is continuous

Optimal predictor

- f_L^* can be characterized as

$$f_L^* \in \arg\min_l \pi_s \left(\eta(X, S) - \frac{l}{L} \right) - s\lambda_l^*,$$

with $\lambda^* = (\lambda_{-L}^*, \ldots \lambda_L^*)$

$$\lambda^* \in \arg \min_{\lambda \in \mathbb{R}^{2L+1}} \sum_{s \in \mathcal{S}} \mathbb{E}_{X|S=s} \max_l \left(s\lambda - \pi_s \left(\eta(x, s) - \frac{l}{L} \right) \right)$$

Estimation

- similar to the post-processing procedure in classification

Wasserstein distance

- let μ, ν two probability distribution on $\mathbb{R}$
- Wasserstein-2 distance

$$\mathcal{W}_2^2\left(\mu,\nu\right) = \inf_{\gamma \in \Gamma_{\mu,\nu}} \int |x-y|^2 \, d\gamma(\mu,\nu),$$

s.t. $\forall \gamma \in \Gamma(\mu,\nu)$, $\gamma(A \times \mathbb{R}) = \mu(A)$, and $\gamma(\mathbb{R} \times B) = \nu(B)$

Useful characterizations

- if X admits a density ν, there exists a mapping T such that

$$\mathcal{W}_2^2\left(\mu,\nu\right) = \mathbb{E}\left[(X - T(X))^2\right],$$

with $T = F_\mu^{-1} \circ F_\nu$

- we have also

$$\mathcal{W}_2^2\left(\mu,\nu\right) = \int_0^1 \left|F_\mu^{-1}(t) - F_\nu^{-1}(t)\right|^2 dt$$

Characterization of fair optimal predictor

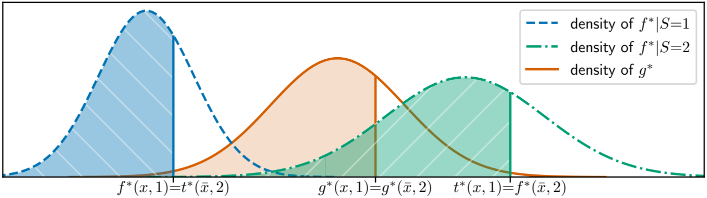
Optimal fair predictor (*Chzhen et al. (2020)*)

Assume that $\nu_{\eta_{|s}} := \mathcal{L}(\eta(X,S)|S=s)$ has a density, $s \in \mathcal{S}$. Then

$$\min_{f \text{ is fair}} \mathbb{E}\left[\eta(X,S) - f(X,S))^2\right] = \min_{\nu} \sum_{s \in \mathcal{S}} \pi_s \mathcal{W}_2^2(\nu_{\eta_{|s}}, \nu)$$

besides $f^*(x,s) = \pi_s \eta(x,s) + (1-\pi_s) t^*(x,s)$,
with $t^*(x,s) = F^{-1}_{\eta_{|s'}}(F_{\eta_{|s}}(\eta(x,s)),\ s' \neq s$

Fair optimal prediction g^* with $p_1 = 2/5$ and $p_2 = 3/5$



Data-driven procedure

Plug-in procedure

- labeled sample $\mathcal{D}_n \to \hat{\eta}$ + randomization $\to \bar{\eta}$
- unlabeled sample $(X_1, S_1), \ldots, (X_N, S_N)$
- $(S_1, \ldots, S_N) \to \hat{\pi}_s$
- $\mathcal{U}_{N_s} = \{X_1^s, \ldots, X_{N_s}^s\} = \mathcal{U}_{N_s}^0 \cup \mathcal{U}_{N_s}^1$
- $\mathcal{U}_{N_s}^0 \to \hat{F}_{\bar{\eta}_{|s}}^{-1}, \mathcal{U}_{N_s}^1 \to \hat{F}_{\bar{\eta}_{|s}}$

Resulting estimator

- $\hat{f}(x, s) = \hat{\pi}_s \hat{f}(x, s) + 1 - \hat{\pi}_s \hat{F}_{\hat{\eta}_{|s'}}^{-1}(\hat{F}_{\hat{\eta}_{|s}}(\hat{\eta}(x, s))$

Theoretical properties

- same guarantees as in classification.

Some references

- Hartz *et al.*, Equality of opportunity in supervised learning., NeurIPS (2016)
- Barocas *et al*, Fairness and Machine Learning (2019)
- Donini *et al*, Empirical Risk Minimization Under Fairness Constraints, NeurIPS (2018)
- Agarwal *et al*, A Reductions Approach to Fair Classification, ICML (2018)
- Chzhen *et al.*, Leveraging Labeled and Unlabeled Data for Consistent Fair Binary Classification , NeurIPS (2019)
- Chzhen *et al.*, Fair regression with Wasserstein barycenters ,NeurIPS (2020)
- Chzhen *et al.*, Fair Regression via Plug-in Estimator and Recalibration With Statistical Guarantees , NeurIPS (2020)

Papers for project

- Agarwal *et al.*, A Reductions Approach to Fair Classification, ICML (2018)
- Calmon *et al.*, Optimized Pre-Processing for Discrimination Prevention, NeurIPS (2017)
- Jiang *et al.*, Wasserstein Fair Classification, UAI (2020)
- Chzhen *et al.*, Fair regression with Wasserstein barycenters, NeurIPS (2020)
- Alghamandi *et al.*, Beyond Adult and COMPAS: Fairness in Multi-Class Prediction, NeurIPS (2022)
- Denis *et al.*, Fairness guarantee in multi-class classification, preprint (2023)
- Xian *et al.*, Fair and Optimal Classification via Post-Processing, ICML (2023)