

# EquiPy: A Python Package implementing Sequential Fairness with Optimal Transport

François Hu, Philipp Ratz, Arthur Charpentier, Suzie Grondin, **Agathe Fernandes Machado**

Quantil: Seminario de Matemáticas Aplicadas  
Thursday, August 8th, 2024



## 1 Introduction

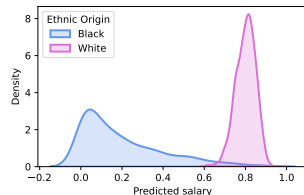
## 2 Fairness evaluation

## 3 Unfairness Mitigation

## 4 Illustrative Example

# Algorithmic Fairness

**Context** Consider a Machine Learning (ML) model  $f$ , its salary predictions on test set  $\hat{Y}$  and one sensitive attribute to which we have access in our dataset, **ethnic origin** (White/Black).



## Potential source of discrimination

- 1 **Statistical bias** in the data: reproduction of past injustices, under-represented minority in an unbalanced data set,
- 2 **Explanatory variables** of the model: proxy variables (correlation between a sensitive attribute and other explanatory variables),
- 3 **Intentional bias**: bias can be the result of deliberate choices, which can be benevolent or malicious.

# Motivations

## Legislation

- AI Act (Europe, 2024) aims to ban or limit AI systems in production that present an “**unacceptable level of risk.**”
- **Motor insurance** regulation (Zebra, 2022).

	United States									Canada				
	CA	HI	GA	NC	NY	MA	PA	FL	TX	AL	ON	NB	NL	QC
Gender	x	x	•	x	•	x	x	•	•	•	•	x	x	•
Age	x	x	•	x*	•	x	•	•	•	•	•	x	x	•
Driving experience	x	x	•	•	•	•	•	•	•	•	•	•	•	•
Credit history	x	x	•	•	•	x	•	•	•	x*	x	•	x	•
Education	x	x	x	x	x	x	•	•	•	•	•	•	•	•
Profession	x	x	x	•	x	x	•	•	•	•	•	•	•	•
Employment	x	x	x	•	x	x	•	•	•	•	•	•	•	•
Family	•	x	•	•	•	x	•	•	•	•	•	•	•	•
Housing	x	x	•	•	•	x	•	•	•	x	x	•	•	•
Address/ZIP code	•	•	•	•	•	•	•	•	•	x	x	•	•	•

• Permitted attribute    x Prohibited attribute    \* with conditions

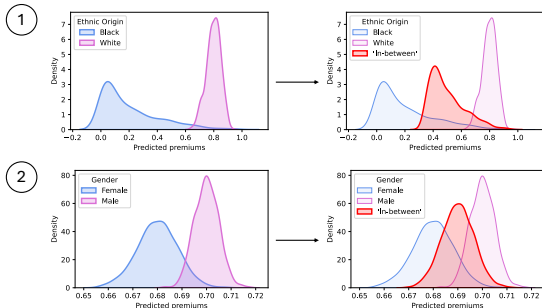
**Proxy variables** (Upton and Cook, 2014) Simply **eliminating the sensitive attributes** from predictive models does not guarantee fair premiums (Feller and Goel, 2016).

**Single Sensitive Attribute** Multiple **mitigation** approaches exist (Chzhen et al., 2020; Gouic et al., 2020; Hardt et al., 2016).

→ Approach to **evaluate and mitigate unfairness** in model predictions regarding **Multiple Sensitive Attributes** (MSA)

# Objective

Consider an **insurance pricing model**  $f$ , its predicted premiums  $\hat{Y}$  and two sensitive attributes, **ethnic origin**  $A_1$  (White/Black) and **gender**  $A_2$  (Male/Female).



We avoid selecting a **reference category** (White/Black and Male/Female) because:

- if "Black" and "Female" are the references, the total premiums would fall short of the planned amount needed to cover claims,
- if "White" and "Male" are the references, the premiums would exceed the planned amount, leading to higher costs for the insureds.

# Context of Multiple Sensitive Attributes

## Intersectional Fairness

MSA  $\rightarrow$  Single sensitive attribute (SSA), by intersection:

Female & White	Female & Black
Male & White	Male & Black

## Sequential Fairness (Hu et al., 2024)

$\hat{Y} \longrightarrow \hat{Y}$  fair for  $A_1 \longrightarrow \hat{Y}$  fair for  $A_1, A_2$

- **Interpretability** accross MSA,
- Easily adding sensitive attributes (SA) to meet changing **regulatory demands**.

Paper: **Sequential Fairness**



Python package: **EquiPy**



# Roadmap

## ① Introduction

## ② Fairness evaluation

Notations

Measuring Unfairness via Optimal Transport

## ③ Unfairness Mitigation

Mitigation Methods

EquiPy Mitigation Approach

## ④ Illustrative Example

## 1 Introduction

## 2 Fairness evaluation

Notations

Measuring Unfairness via Optimal Transport

## 3 Unfairness Mitigation

## 4 Illustrative Example



## 1 Introduction

## 2 Fairness evaluation

### Notations

### Measuring Unfairness via Optimal Transport

## 3 Unfairness Mitigation

## 4 Illustrative Example

# Notations

- $\mathbf{X} \in \mathcal{X}$ : 'non-sensitive' features,
- $\mathbf{A} = (A_1, \dots, A_r) \in \mathcal{A}_1 \times \dots \times \mathcal{A}_r$ :  $r$  sensitive features,
- $\hat{Y}$ : response variable (continuous or score from a binary classifier)
- $f$ : predictive model on  $(\mathbf{X}, \mathbf{A})$ , with  $f^*$  the optimal Bayes estimator  $\mathbb{E}[Y|\mathbf{X}, \mathbf{A}]$ ,
- $\nu_f$ : distribution of  $f(\mathbf{X}, \mathbf{A})$  with cumulative distribution function  $F_f$  and quantile function  $Q_f$ ,
- $\nu_{f|a_i}$ : conditional distribution of  $f(\mathbf{X}, \mathbf{A})|A_i = a_i$  with  $F_{f|a_i}$  and  $Q_{f|a_i}$ ,
- $\mathcal{R}(f) = \mathbb{E}[(Y - f(\mathbf{X}, \mathbf{A}))^2]$ : risk metric.

## 1 Introduction

## 2 Fairness evaluation

Notations

Measuring Unfairness via Optimal Transport

## 3 Unfairness Mitigation

## 4 Illustrative Example

# Demographic Parity for Group Fairness

**Demographic Parity** requires that the predictions made by a model be **independent** of a specific sensitive attribute  $A$  (such as race, gender, or age).

**Strong Demographic Parity**  $\forall a_i, a'_i \in \mathcal{A}_i, \nu_{f|a_i} = \nu_{f|a'_i}$  or  $\text{distance}(\nu_{f|a_i}, \nu_{f|a'_i}) = 0$ .

- ①  $f$  is strongly fair regarding **a single sensitive attribute** (SSA)  $A_i$ , if and only if:

$$\mathcal{U}_i(f) = \max_{a_i \in \mathcal{A}_i} \text{distance}(\nu_f, \nu_{f|a_i}) = 0$$

- ②  $f$  is strongly fair regarding **MSA**, if and only if:

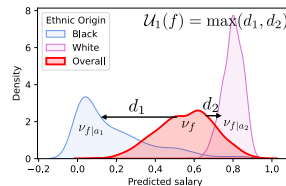
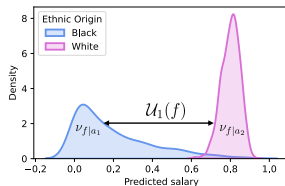
$$\mathcal{U}(f) = \mathcal{U}_1(f) + \dots + \mathcal{U}_r(f) = 0$$

→ **Wasserstein distance** from Optimal Transport (OT) theory is employed to compute the distance between distributions.

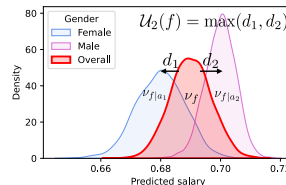
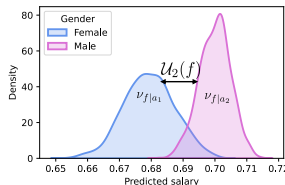
# Example

Strong Demographic Parity for MSA: **ethnic origin** ( $A_1$ ) and **gender** ( $A_2$ ).

1



2



$$\mathcal{U}(f) = \mathcal{U}_1(f) + \mathcal{U}_2(f)$$

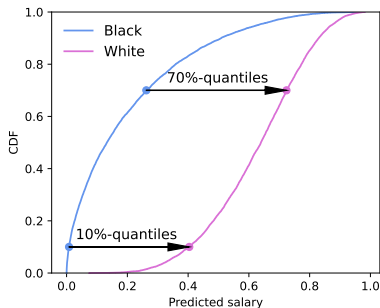
# Optimal Transport

The objective of OT is to minimize the overall cost of moving one mass distribution ( $\nu_A$ ) onto another one ( $\nu_B$ ). We are searching for the most efficient mapping  $T$  to move mass between  $\nu_A$  and  $\nu_B$ , s.t.  $\nu_B = T_{\#}\nu_A$ , by solving (Monge, 1781)

$$\inf_{T_{\#}\nu_A=\nu_B} \int_A c(x, T(x)) d\nu_A(x)$$

For some strictly convex 'cost'  $c$ , such as quadratic cost, and univariate distributions  $\nu_A$  and  $\nu_B$ , the **optimal transport map**  $T^*$  is (Santambrogio, 2015)

$$T^* = Q_B \circ F_A$$

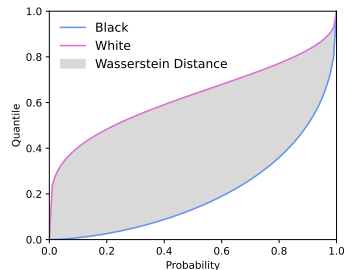


# Optimal Transport and Wasserstein distance

For univariate distributions  $\nu_A$  and  $\nu_B$ ,  $p$ -**Wasserstein distance** ( $p \geq 1$ ) corresponds to the value of the minimum 'cost' required to transform  $\nu_A$  into  $\nu_B$  (Wasserstein, 1969):

$$\mathcal{W}_p(\nu_A, \nu_B) = \left( \int_{u \in [0,1]} |Q_A(u) - Q_B(u)|^p du \right)^{1/p}$$

→ **Fairness criterion**:  $\mathcal{U}_i(f) = \max_{a_i \in \mathcal{A}_i} \mathcal{W}_1(\nu_f, \nu_{f|a_i})$ .



## 1 Introduction

## 2 Fairness evaluation

## 3 Unfairness Mitigation

Mitigation Methods

EquiPy Mitigation Approach

## 4 Illustrative Example



1 Introduction

2 Fairness evaluation

3 Unfairness Mitigation  
Mitigation Methods  
EquiPy Mitigation Approach

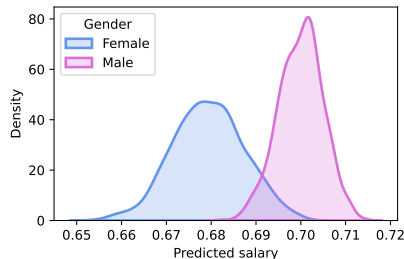
4 Illustrative Example

# Multiple methods

**Objective:** Transform model predictions  $f(\mathbf{X}, \mathbf{A}) \in \mathbb{R}$  into fair ones  $f_B(\mathbf{X}, \mathbf{A})$ , while **preserving good performance**  $\mathcal{R}(f)$ .

- Pre-processing: transform multivariate distribution of  $\mathbf{X}$ ,
- In-processing: add a 'fairness' penalty in the objective function,

- **Post-processing:** transform univariate distribution of  $\hat{Y} = f(\mathbf{X}, \mathbf{A})$ .



## 1 Introduction

## 2 Fairness evaluation

## 3 Unfairness Mitigation

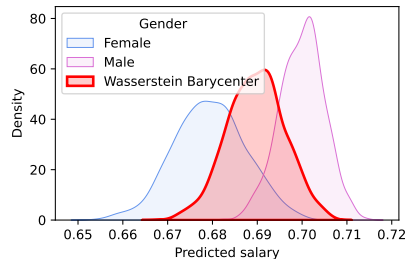
- Mitigation Methods
- EquiPy Mitigation Approach

## 4 Illustrative Example

# Wasserstein barycenter

The **Wasserstein Barycenter** finds a representative distribution that lies between  $K$  given distributions  $(\nu_1, \dots, \nu_K)$ , and weights  $(w_1, \dots, w_K) \in \mathbb{R}_+^K$ . The  $\mathcal{W}_2$ -Barycenter is the minimizer:

$$\text{Bar}\{(w_k, \nu_k)_{k=1}^K\} = \underset{\nu}{\operatorname{argmin}} \sum_{k=1}^K w_k \cdot \mathcal{W}_2^2(\nu_k, \nu)$$



Constructing  $f_B$  with Wasserstein barycenter, Gouic et al. (2020) prove  $f_B = \operatorname{arginf}_f \{\mathcal{R}(f) : \mathcal{U}(f) = 0\}$ .

# Achieving Fairness via Optimal Transport

## Single Sensitive Attribute ( $r = 1$ ) (Chzhen et al., 2020)

$$\nu_{f_B} = \mu_{\mathcal{A}_1}(\nu_{f^*}) = \inf_f \sum_{a_1 \in \mathcal{A}_1} \mathbb{P}(A_1 = a_1) \cdot \mathcal{W}_2^2(\nu_{f^*|a_1}, \nu_f)$$

$$f_B(\mathbf{x}, a_1) = \left( \sum_{a'_1 \in \mathcal{A}_1} \mathbb{P}(A_1 = a'_1) Q_{f^*|a'_1} \right) \circ F_{f^*|a_1}(f^*(\mathbf{x}, a_1))$$

$$\forall (\mathbf{x}, a_1) \in \mathcal{X} \times \mathcal{A}_1.$$

→ **EquiPy**: This approach is implemented in the function **FairWasserstein** of **fairness** module.

## Example

Consider ML model predictions  $\hat{y} = \hat{f}(\mathbf{x}, a_1)$  where  $a_1 \in \mathcal{A}_1$  corresponds to the observations of the SSA,  $A_1$ : **ethnic origin** (White/Black).

### Mitigation approach

$$\begin{aligned}\hat{f}_{B_1}(\mathbf{x}, a_1 = \text{White}) &= \mathbb{P}[A_1 = \text{White}] \cdot \hat{f}(\mathbf{x}, a_1 = \text{White}) \\ &\quad + \mathbb{P}[A_1 = \text{Black}] \cdot Q_{\text{Black}} \circ F_{\text{White}}(\hat{f}(\mathbf{x}, a_1 = \text{White}))\end{aligned}$$

$$\begin{aligned}\hat{f}_{B_1}(\mathbf{x}, a_1 = \text{Black}) &= \mathbb{P}[A_1 = \text{Black}] \cdot \hat{f}(\mathbf{x}, a_1 = \text{Black}) \\ &\quad + \mathbb{P}[A_1 = \text{White}] \cdot Q_{\text{White}} \circ F_{\text{Black}}(\hat{f}(\mathbf{x}, a_1 = \text{Black}))\end{aligned}$$

# Sequential fairness

## Multiple Sensitive Attributes ( $r \geq 1$ ) (Hu et al., 2024)

$$f_B(\mathbf{x}, \mathbf{a}) := f_{B_1} \circ f_{B_2} \circ \dots \circ f_{B_r}(\mathbf{x}, \mathbf{a})$$
$$f_{B_i} \circ f_{B_j}(\mathbf{x}, \mathbf{a}) = \left( \sum_{a'_i \in \mathcal{A}_i} \mathbb{P}(A_i = a'_i) Q_{f_{B_j}|a'_i} \right) \circ F_{f_{B_j}|a_i}(f_{B_j}(\mathbf{x}, \mathbf{a}))$$

$\forall (\mathbf{x}, \mathbf{a}) \in \mathcal{X} \times \mathcal{A}_{1:r}$ , with the  $i$ -th component of  $\mathbf{a}$  denoted  $a_i$ .

Hu et al. prove the **associativity** of Wasserstein barycenters:

$\mu_{\mathcal{A}_1} \circ \mu_{\mathcal{A}_2}(\nu_{f^*}) = \mu_{\mathcal{A}_2} \circ \mu_{\mathcal{A}_1}(\nu_{f^*})$ . Fairness mitigation remains **unaffected by the order** of  $\mathcal{A}_{1:r}$ .

→ **EquiPy**: This approach is implemented in the function **MultiWasserstein** of **fairness** module.

## Example

Consider transformed model predictions fair regarding **ethnic origin**  $\hat{f}_{B_1}(\mathbf{x}, \mathbf{a})$  where  $\mathbf{a} = (a_1, a_2) \in \mathcal{A}_1 \times \mathcal{A}_2$  corresponds to the observations of the MSA,  $A_1$  and  $A_2$ : **ethnic origin** and **gender** (Male/Female).

### Mitigation approach

$$\begin{aligned}\hat{f}_{B_2}(\mathbf{x}, a_1, a_2 = \text{Male}) &= \mathbb{P}[A_2 = \text{Male}] \cdot \hat{f}_{B_1}(\mathbf{x}, a_1, a_2 = \text{Male}) \\ &\quad + \mathbb{P}[A_2 = \text{Female}] \cdot Q_{\text{Female}} \circ F_{\text{Male}}(\hat{f}_{B_1}(\mathbf{x}, a_1, a_2 = \text{Male}))\end{aligned}$$

$$\begin{aligned}\hat{f}_{B_2}(\mathbf{x}, a_1, a_2 = \text{Female}) &= \mathbb{P}[A_1 = \text{Female}] \cdot \hat{f}_{B_1}(\mathbf{x}, a_1, a_2 = \text{Female}) \\ &\quad + \mathbb{P}[A_2 = \text{Male}] \cdot Q_{\text{Male}} \circ F_{\text{Female}}(\hat{f}_{B_1}(\mathbf{x}, a_1, a_2 = \text{Female}))\end{aligned}$$



1 Introduction

2 Fairness evaluation

3 Unfairness Mitigation

4 Illustrative Example

# Life insurance dataset

- Public SEER dataset: <https://seer.cancer.gov>,
- Prediction of **one-year mortality** of US individuals with melanoma skin cancer,  
→ Utilizing the methodology presented in Sauce et al. (2023), we convert the dataset into survival data, by accounting for **exposure** over a given time interval.
- Sample size  $n = 547,878$  from 2004 to 2018,
- Explanatory variables: 16 features describing patient characteristics (age, **gender** male/female, **ethnic origin**) and cancer attributes (tumor size, extent).

→ **MSA** framework: use of the function **MultiWasserstein**.

# Model fitting

- ① Split the data into train and test sets,
- ② Fit **Logistic Regression**\*  $f$ ,
- ③ Apply  $f$  on the test set to obtain  $\hat{y}_{\text{test}}$ .

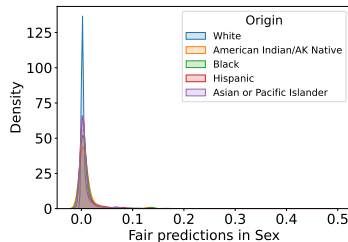
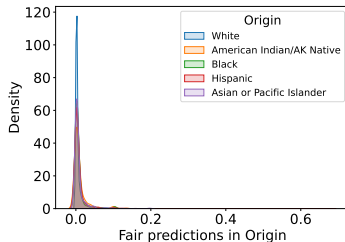
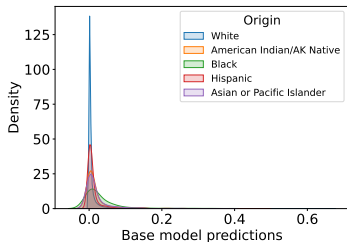
We consider different model fitting scenarios, in which we include or exclude sensitive attributes as explanatory variables:

Ethnic origin	Gender	AUC	Unfairness
No	No	0.87	0.22
<b>Yes</b>	<b>Yes</b>	0.87	0.27

\* **Model-agnosticity** of EquiPy:  $f$  can be any ML model.

# Transforming predictions

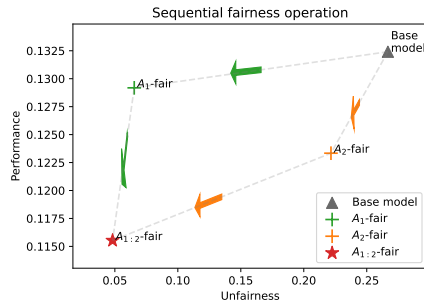
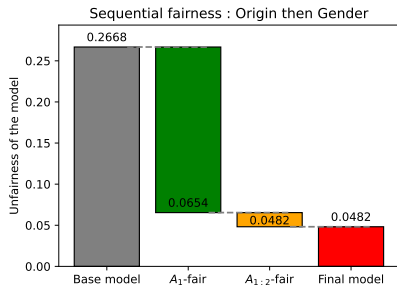
- 1 Split the test data into calibration and test sets,
- 2 Specify an order to sequentially correct:  $A_1$  corresponds to **ethnic origin** and  $A_2$  corresponds to **gender**,
- 3 Fit and transform your test predictions using MultiWasserstein from fairness module.



# Visualizations

**Unfairness** and **metric** calculations with graphs module:

- `fair_waterfall_plot`: sequential gain in fairness for the specified order  $A_1$  then  $A_2$ ,
- `fair_multiple_arrow_plot`: fairness-performance relationship for all potential pathways.



## Additional results: Approximate fairness

When correcting biases related to **gender**, we reduce fairness regarding **origin**:

Fairness step	Unfairness in <b>origin</b>	Unfairness in <b>gender</b>
Base model	<b>0.2371</b>	0.0297
<b>Origin</b>	<b>0.0345</b>	0.0309
<b>Origin &amp; Gender</b>	<b>0.0469</b>	0.0013

We can **prioritize fairness accross attributes** by specifying  $\epsilon = [0, 0.5]$  corresponding to exact fairness in  $A_1$  and 0.5-approximate fairness in  $A_2$ .

$$f_B = 0.5 \cdot (f_{B_2} \circ f_{B_1}) + 0.5 \cdot f_{B_1}$$

## Wrap up

- The novel approach of **Sequential Fairness**, introduced in Hu et al. (2024), allows to mitigate unfairness regarding **Multiple Sensitive Attributes**.
- The Python package **EquiPy** implements the Sequential Fairness approach and is applicable to **any continuous Machine Learning predictions**.

Code available: [https://github.com/fer-agathe/quantil\\_seminar](https://github.com/fer-agathe/quantil_seminar)

Comments are welcome: [fernandes\\_machado.agathe@courrier.uqam.ca](mailto:fernandes_machado.agathe@courrier.uqam.ca)

Documentation **EquiPy**



## 5 Appendix



# References I

- Chzhen, E., Denis, C., Hebiri, M., Oneto, L. and Pontil, M. (2020). Fair Regression with Wasserstein Barycenters. In *Advances in Neural Information Processing Systems*.
- Feller, P. E. C.-D. S., A. and Goel, S. (2016). A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.
- Gouic, T. L., Loubes, J.-M. and Rigollet, P. (2020). Projection to fairness in statistical learning.
- Hardt, M., Price, E., Price, E. and Srebro, N. (2016). Equality of Opportunity in Supervised Learning. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I. and Garnett, R. (eds), *Advances in Neural Information Processing Systems*, 29. Curran Associates, Inc.
- Hu, F., Ratz, P. and Charpentier, A. (2024). A sequentially fair mechanism for multiple sensitive attributes. *Proceedings of the AAAI Conference on Artificial Intelligence* 38: 12502–12510, doi:10.1609/aaai.v38i11.29143.
- Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*.
- Santambrogio, F. (2015). *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Progress in Nonlinear Differential Equations and Their Applications. Springer International Publishing.

## References II

- Sauce, M., Chancel, A. and Ly, A. (2023). AI and Ethics in Insurance: a new solution to mitigate proxy discrimination in risk modeling. *arXiv* 2307.13616.
- Upton, G. and Cook, I. (2014). *A Dictionary of Statistics* 3e. Oxford Paperback Reference. OUP Oxford.
- Wasserstein (1969). Markov processes over denumerable products of spaces, describing large systems of automata. *Probl. Peredachi Inf.* 5.
- Zebra, T. (2022). Car insurance rating factors by state. <https://www.thezebra.com/>.