EquiPy: A Python Package for Sequential Fairness using Optimal Transport with Applications in Insurance

Agathe Fernandes Machado[a], Arthur Charpentier[b], François Hu[c], Philipp Ratz[b] and Suzie Grondin[b]

[a]Université du Québec à Montréal (`https://fer-agathe.github.io/`), [b]Université du Québec à Montréal, [c]Université de Montréal

Motivation: Sequential Fairness

Insurance regulation

- In certain regions of Canada and the United States (US), discrimination based on **multiple sensitive attributes** (MSA), such as ethnic origin (A_1), gender (A_2) and age (A_3), is prohibited for **insurance pricing**.
- Due to **proxy variables**, eliminating MSA from predictive models does not guarantee fair premiums.

→ There is a need for an approach to **evaluate and mitigate unfairness** in model predictions among groups with shared MSA, thereby ensuring **group fairness**.

Strong Demographic Parity Given a model f, let F_f denote the distribution of $f(\boldsymbol{X},\boldsymbol{A})$ and $F_{f|a_i}$ denote the conditional distribution of $f(\boldsymbol{X},\boldsymbol{A})|A_i = a_i$, with $\boldsymbol{X} \in \mathcal{X}$ corresponds to 'non-sensitive' features and $\boldsymbol{A} = (A_1, A_2, A_3) \in \mathcal{A}_1 \times \mathcal{A}_2 \times \mathcal{A}_3$.

1. f is strongly fair regarding a **single sensitive attribute** (SSA) A_i, if and only if: $$\mathcal{U}_i(f) = \max_{a_i \in \mathcal{A}_i} \text{distance}(F_f, F_{f|a_i}) = 0$$
2. Here, f is **strongly fair regarding MSA**, if and only if: $$\mathcal{U}(f) = \mathcal{U}_1(f) + \mathcal{U}_2(f) + \mathcal{U}_3(f) = 0$$

→ **Wasserstein distance** is employed to compute the distance between distributions.

Mitigating biases for MSA When $\mathcal{U}(f) \neq 0$, the goal is to **correct unfairness** while preserving good model performance. One effective approach is employing **Optimal Transport**, specifically using **Wasserstein barycenter**, as it maintains the order of individuals within groups in terms of risk.

Sequential fairness Hu et al. (AAAI-2024) introduce **sequential correction** of unfairness related to **MSA**, using this approach. Unlike **intersectional fairness** this sequential approach allows for:

- interpretability across sensitive features,
- easily adding sensitive attributes to meet changing regulatory demands.

1. Background: Optimal Transport

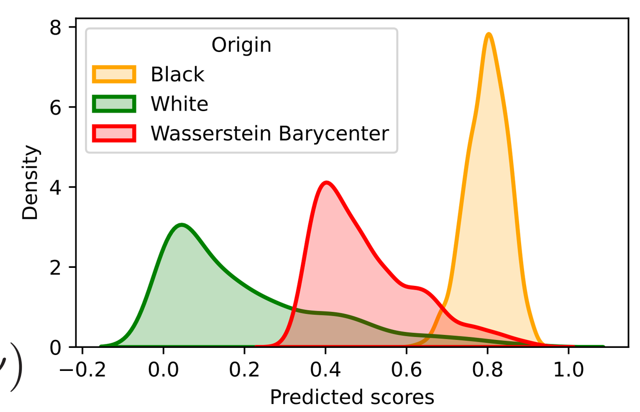
For univariate distributions ν_1 and ν_2, p-**Wasserstein distance** ($p \geq 1$) computes the minimum 'cost' required to transform ν_1 into ν_2:

$$\mathcal{W}_p(\nu_1, \nu_2) = \left(\int_{u \in [0,1]} |Q_1(u) - Q_2(u)|^p \, du \right)^{1/p}$$

In this case, the optimal transport map T^*, with $T^*_\# \nu_1 = \nu_2$, for some strictly convex 'cost', such as quadratic cost, is defined as $T^* := Q_2 \circ F_1$.

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

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



Hu et al. prove the **associativity** of the $\mathcal{W}_2$-Barycenter with univariate measures.

2. EquiPy: Methodology

Objective **EquiPy** is a Python package applying sequential fairness accross a set of MSA by transforming predictions from a **regression** or scores from a **binary classifier**, denoted $f(\boldsymbol{X},\boldsymbol{A}) \in \mathbb{R}$, into fair ones $f_B(\boldsymbol{X},\boldsymbol{A})$.

Notation

- $\boldsymbol{A} = (A_1, \ldots, A_r) = A_{1:r} \in (\mathcal{A}_1 \times \ldots \times \mathcal{A}_r)$, the sequence of r **sensitive attributes**,
- Y is the response variable.

Theoretical results Fair predictions, $f_B(\boldsymbol{X},\boldsymbol{A})$, are defined by:

$$f_B = \arg\inf_f \{\mathcal{R}(f) : \mathcal{U}(f) = 0\}$$
$$\mathcal{R}(f) := \mathbb{E}[Y - f(\boldsymbol{X},\boldsymbol{A}))]^2 \quad (\textit{risk metric})$$
$$\mathcal{U}_i(f) := \max_{a_i} \mathcal{W}_1(\nu_f, \nu_{f|a_i}) \quad (\textit{unfairness regarding } A_i)$$
$$\mathcal{U}(f) := \mathcal{U}_1(f) + \ldots + \mathcal{U}_r(f) \quad (\textit{unfairness regarding } \boldsymbol{A})$$

SSA ($r = 1$) Given the results in Chzhen et al. (2020), f_B is computed as:

$$\forall\, (\mathbf{x}, a_i) \in \mathcal{X} \times \mathcal{A}_i,\, f_B(\mathbf{x}, a_i) := \left(\sum_{a_i' \in \mathcal{A}_i} \mathbb{P}(A_i = a_i') Q_{f|a_i'} \right) \circ F_{f|a_i}(f(\mathbf{x}, a_i))$$

calculated with $\mathcal{W}_2$-Barycenter between all $f|a_i$ for $a_i \in \mathcal{A}_i$.

MSA ($r > 1$) Hu et al. establish that f_B can be expressed as:

$$\forall\, (\mathbf{x}, \mathbf{a}) \in \mathcal{X} \times \mathcal{A}_{1:r},\, f_B(\mathbf{x}, \mathbf{a}) := f_{B_1} \circ f_{B_2} \circ \cdots \circ f_{B_r}(\mathbf{x}, \mathbf{a})$$

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

with the i-th component of $\mathbf{a}$ denoted a_i.

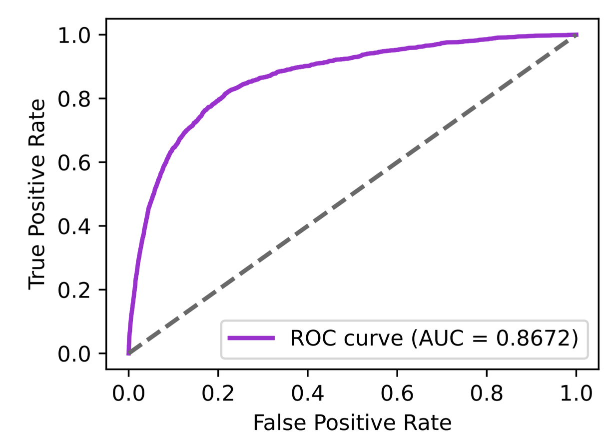
→ Fairness mitigation remains **unaffected by the order of $A_{1:r}$**, given the property of **associativity** of Wasserstein barycenters.

3. Life insurance data

Description The dataset, from the public SEER database, studies the mortality of US individuals with **melanoma skin cancer**. There are $n = 547,878$ observations from 2004 to 2018, and 16 features describing patient characteristics (age, gender, origin) and cancer attributes (tumor size, extent).

Predictive modeling Utilizing the methodology presented in Sauce et al. (2023), we convert the dataset into **survival data**. Subsequently, we predict the **one-year mortality** while accounting for exposure over the given time interval:

1. Split the data into train and test sets,
2. Fit **Logistic Regression** f using exposure as sample weights,
3. Apply f on the test set.



→ **Model-agnosticity**: EquiPy employs a **post-processing methodology** to attain fair predictions, originating from any Machine Learning model.

4. Application in Insurance

EquiPy import

`from equipy import fairness, metrics, graphs` → **3 modules**

Fairness module

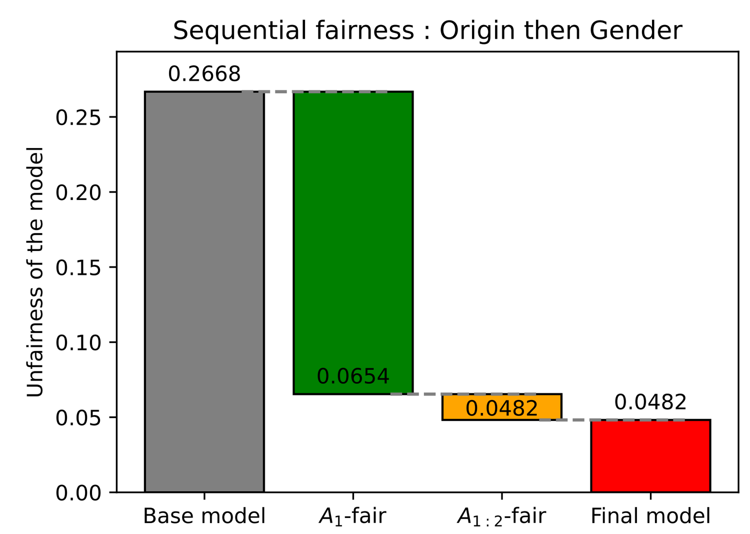
1. Split the test set into **calibration** and **test** sets,
2. **SSA**: `x_ssa_calib` and `x_ssa_test` contain values for a unique sensitive attribute (A_1: gender or origin)
```
wst = FairWasserstein()
wst.fit(scores_calib, x_ssa_calib)
y_ssa_fair = wst.transform(scores_test, x_ssa_test)
```
3. **MSA**: `x_msa_calib` and `x_msa_test` contain values for both sensitive attributes (A_1: origin, A_2: gender)
```
wst = MultiWasserstein()
wst.fit(scores_calib, x_msa_calib)
y_msa_fair = wst.transform(scores_test, x_msa_test)
```

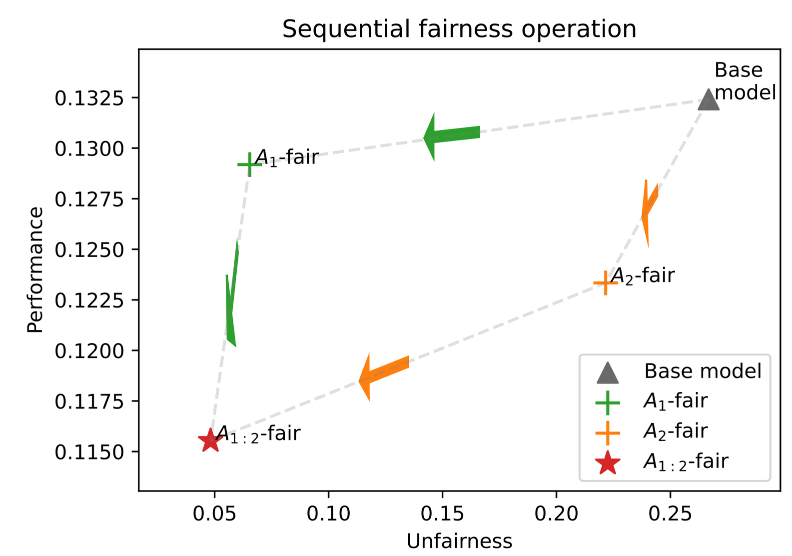
Metrics module

`unfairness(y_msa_fair, x_msa_test)` → 0.0482
`y_pred_fair = (y_msa_fair > 0.102).astype(int)`
`performance(y_pred_fair, y_true_test, f1_score)` → 0.1155

Graphs module Visualization of unfairness and metrics calculations:

- `fair_density_plot`: representation of $f|a_i$, $a_i \in \mathcal{A}_1 \cup \mathcal{A}_2$,
- `fair_waterfall_plot`: sequential gain in fairness,
- `fair_multiple_arrow_plot`: fairness-performance relationship.





Fairness step	Unfairness in origin	Unfairness in gender
Base model	**0.2371**	0.0297
Origin	**0.0345**	0.0309
Origin & Gender	**0.0469**	0.0013

→ **Prioritization accross attributes** using **approximate fairness**: `wst.transform(scores_calib, x_msa_calib, epsilon = [0,0.5])`, corresponding to exact fairness in A_1 and 0.5-approximate fairness in A_2: $f_B = 0.5 \cdot (f_{B_2} \circ f_{B_1}) + 0.5 \cdot f_{B_1}$

References

Chzhen, E., Denis, C., Hebiri, M., Oneto, L., and Pontil, M. (2020). Fair Regression with Wasserstein Barycenters. In *Advances in Neural Information Processing Systems*.

Hu, F., Ratz, P., and Charpentier, A. (2024). A Sequentially Fair Mechanism for Multiple Sensitive Attributes. *AAAI-2024*.

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.