A Sequentially Fair Mechanism for Multiple Sensitive Attributes

AAAI-24. Hu et al. [1]

Reviewer: Jihu Lee

IDEA lab
Department of Statistics
Seoul National University

April 15, 2025

Overview

- Using multi-marginal 2-Wasserstein barycenters, offers a closed-form solution of the learning problem (DP).
- Rewrite the optimal fair solution into a sequential form.

Notation

- Features: $\boldsymbol{X} \in \mathcal{X} \subset \mathbb{R}^d$, Target: $Y \in \mathcal{Y} \subset \mathbb{R}$.
- Sensitive: $\boldsymbol{A} = (A_1, \ldots, A_r) \in \mathcal{A}_1 \times \cdots \mathcal{A}_r$ where $\mathcal{A}_i = \{1, \ldots, K_i\}$ with $K_i \in \mathbb{N}$.
- $A_{i:i+k} = (A_i, \ldots, A_{i+k})$.
- Predictor: $f : \mathcal{X} \times \mathcal{A} \to \mathcal{Y}$. $\mathcal{F}$: set of predictors f.
- ν_f: p.m. of $f(\boldsymbol{X}, \boldsymbol{A})$, $\nu_{f|\boldsymbol{a}}$: p.m. of $f(\boldsymbol{X}, \boldsymbol{A})|\boldsymbol{A} = \boldsymbol{a}$.
- $F_{f|\boldsymbol{a}}(u) := \mathbb{P}(f(\boldsymbol{X}, \boldsymbol{A}) \leq u|\boldsymbol{A} = \boldsymbol{a})$.
- $Q_{f|\boldsymbol{a}}(v) := \inf\{u \in \mathbb{R} : F_{f|\boldsymbol{a}}(u) \geq v\}$.
- $\mathcal{R}(f) = \mathbb{E}(Y - f(\boldsymbol{X}, \boldsymbol{A}))^2$.
- $f^*(\boldsymbol{X}, \boldsymbol{A}) = \mathbb{E}[Y|\boldsymbol{X}, \boldsymbol{A}]$

---

p.m.=probability measure

Single Sensitive Case (previous result from [2])

- Sensitive $A \in \{1, \ldots, K\}$, $p_a = \mathbb{P}(A = a)$.
- **Fairness measure (Strong DP)**

$$\mathcal{U}(f) = \max_{a \in \mathcal{A}} \int_{u \in [0,1]} \left| Q_f(u) - Q_{f|a}(u) \right| du$$

- **Optimal fair predictor**

$$f_B = \arg\min_{f \in \mathcal{F}} \{\mathcal{R}(f) : \mathcal{U}(f) = 0\}$$

Single Sensitive Case (previous result from [2])

- Minimizing $\mathcal{R}(f)$ = minimizing $\mathbb{E}(f^*(\boldsymbol{X}, A) - f(\boldsymbol{X}, A))^2$.

$$\min_{f:\mathcal{U}(f)=0} \mathbb{E}(f^*(\boldsymbol{X}, A) - f(\boldsymbol{X}, A))^2 = \min_{\nu} \sum_{a\in[K]} p_a \mathcal{W}_2^2(\nu_{f^*|a}, \nu)$$

- A map to the minimizer $\mu_{\mathcal{A}} : \mathcal{V} \to \mathcal{V}$:

$$\mu_{\mathcal{A}}(\nu_{f^*}) := \min_{\nu} \sum_{a\in[K]} p_a \mathcal{W}_2^2(\nu_{f^*|a}, \nu) \text{ (**Wasserstein Barycenter**)}$$

- Closed-form solution of f:

$$f_B(\boldsymbol{x}, a) = \left(\sum_{a'\in\mathcal{A}} p_{a'} Q_{f^*|a'} \right) \circ F_{f^*|a}(f^*(\boldsymbol{x}, a))$$

Extension to Multiple Sensitive Case

- **Fairness measure (Strong DP)**

$$\mathcal{U}_i(f) = \max_{a_i \in \mathcal{A}_i} \int_{u \in [0,1]} \left| Q_f(u) - Q_{f|a_i}(u) \right| du$$

$$\mathcal{U}_{i,\ldots,i+k}(f) = \mathcal{U}_{i:i+k}(f) = \mathcal{U}_i(f) + \ldots + \mathcal{U}_{i+k}(f)$$

- **Optimal fair predictor (marginal)**

$$f_{B_i} = \arg\min_{f \in \mathcal{F}} \{\mathcal{R}(f) : \mathcal{U}_i(f) = 0\}$$

$$f_{B_i}(\boldsymbol{x}, \boldsymbol{a}) = \left(\sum_{a_i' \in \mathcal{A}_i} p_{a_i'} Q_{f^*|a_i'} \right) \circ F_{f^*|a_i}(f^*(\boldsymbol{x}, \boldsymbol{a}))$$

Extension to Multiple Sensitive Case

Sequentially fair mechanism

- Optimal fair predictor

$$f_B = \arg\min_{f \in \mathcal{F}} \{\mathcal{R}(f) : \mathcal{U}_{1:r}(f) = 0\}$$

- Under some assumptions, the optimal fair predictor is (Proposition 4,5):

$$\begin{aligned} f_B(\boldsymbol{x}, \boldsymbol{a}) &= (f_{B_1} \circ \cdots \circ f_{B_r})(\boldsymbol{x}, \boldsymbol{a}) \\ &= (f_{B_{\sigma(1)}} \circ \cdots \circ f_{B_{\sigma(r)}})(\boldsymbol{x}, \boldsymbol{a}) \end{aligned}$$

Fairness Level Control

Single sensitive case [2]

$$f_{B_1}^{\varepsilon_1} \in \arg\min_{f \in \mathcal{F}} \{\mathcal{R}(f) : \mathcal{U}_1(f) \leq \varepsilon_1 \cdot \mathcal{U}_1(f^*)\}$$

$$f_{B_1}^{\varepsilon_1}(\boldsymbol{X}, \boldsymbol{A}) := (1 - \varepsilon_1) \cdot f_{B_1}(\boldsymbol{X}, \boldsymbol{A}) + \varepsilon_1 \cdot f^*(\boldsymbol{X}, \boldsymbol{A})$$

Multiple extension (This paper)

$$f_B^{\boldsymbol{\varepsilon}} = \arg\min_{f \in \mathcal{F}} \left\{ \mathcal{R}(f) : \mathcal{U}(f) \leq \sum_{i=1,\ldots,r} \varepsilon_i \cdot \mathcal{U}_i(f^*) \right\}$$

$$f_B^{\varepsilon}(\boldsymbol{X}, \boldsymbol{A}) = \left(f_{B_{\sigma(1)}}^{\varepsilon_{\sigma(1)}} \circ \cdots \circ f_{B_{\sigma(r)}}^{\varepsilon_{\sigma(r)}} \right)(\boldsymbol{X}, \boldsymbol{A})$$

Calculation

- $f^* \leftarrow \hat{f}$: plug-in with any unfair ML model.

$$\widehat{f_{B_i}}(\boldsymbol{x}, \boldsymbol{a}) = \left(\sum_{a_i' \in \mathcal{A}_i} \hat{p}_{a_i'} \right) \circ \hat{Q}_{\hat{f}|a_i} \left(\hat{f}(\boldsymbol{x}, \boldsymbol{a}) \right)$$

$$\widehat{f_B} = \widehat{f_{B_1}} \circ \cdots \circ \widehat{f_{B_r}}$$

Experiments

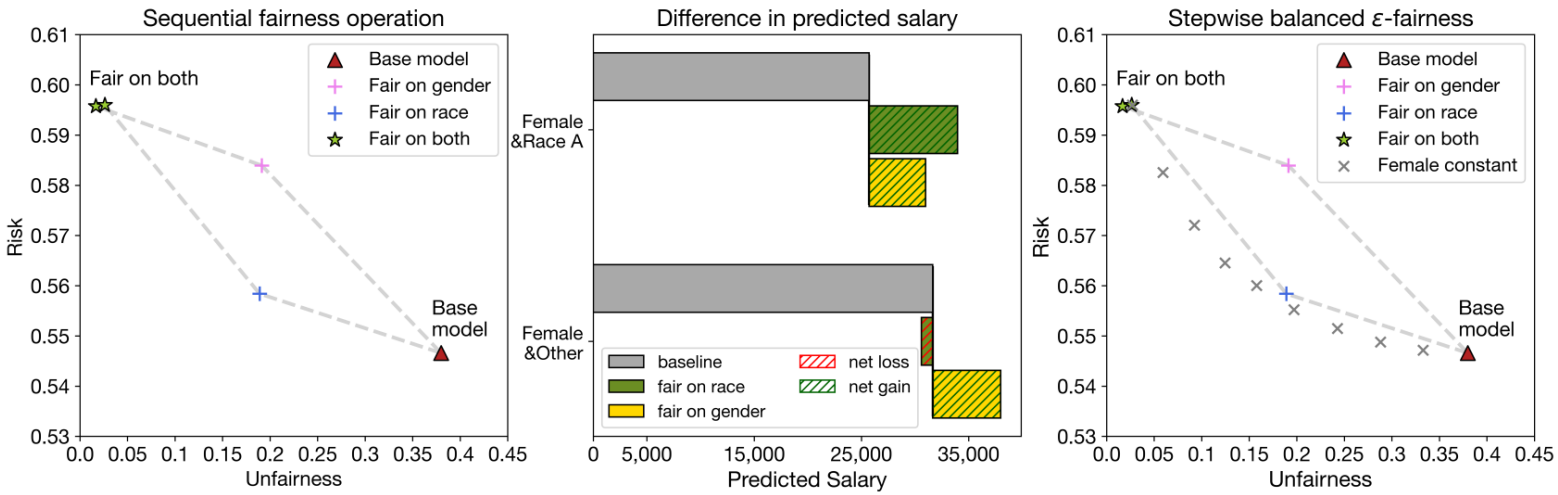

Figure 3: Applications on the *folktables* data set. Left pane, visualisation of the combined unfairness across two sensitive attributes and intermediate solutions rendering predictions fair on only one of them. Center pane, marginal changes to predicted income when rendering fair the predictions w.r.t. a single variable and the baseline predictions. Right pane, visualization of global metrics when correcting the score first for race, but keeping the average predicted salary of female individuals constant.

References

[1] Hu, François, Philipp Ratz, and Arthur Charpentier. "A sequentially fair mechanism for multiple sensitive attributes." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. No. 11. 2024.

[2] Chzhen, Evgenii, et al. "Fair regression with wasserstein barycenters." Advances in Neural Information Processing Systems 33 (2020): 7321-7331.


THANK
You!