Sequential Conditional (Marginally Optimal) Transport on Probabilistic Graphs for Interpretable Counterfactual Fairness

Agathe Fernandes Machado, Arthur Charpentier, Ewen Gallic

AAAI-25 / IAAI-25 / EAAI-25
FEBRUARY 25 – MARCH 4, 2025 | PHILADELPHIA, USA

UQÀM Université du Québec à Montréal
amU Aix Marseille Université
amse école d'économie d'aix-marseille aix-marseille school of economics
CRM CENTRE DE RECHERCHES MATHÉMATIQUES

SCOR FOUNDATION FOR SCIENCE
obvia

Individual Fairness

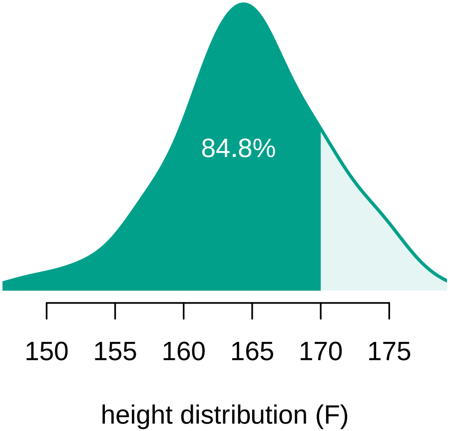
"*Had the protected attributes of the individual been different, would the decision provided by the model have remained the same?*"

- Focus on **individual fairness** (Dwork et al., 2012; Kusner et al., 2017) rather than group fairness (Barocas et al., 2023; Hardt et al., 2016).
  "*we capture fairness by the principle that any two individuals who are similar with respect to a particular task should be classified similarly.*" Dwork et al. (2012)
- Build a **counterfactual individual** and and compare the model's prediction.
- Two philosophies:
  - ***Ceteris paribus***: changing the **sensitive attribute** only, all other things equal.
  - ***Mutatis mutandis*** (Kusner et al., 2017; Kilbertus et al., 2017) (this paper): the **sensitive attribute** may influence other variables that also need to be adjusted alongside it.

Intuitive Example

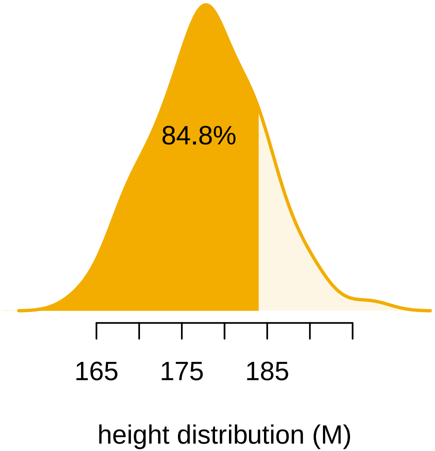
Consider the height of **females** and **males**.

- What is the counterfactual of a **female** with height 170cm (=5' 7") had she been a **male**?
- Within the distribution of **females**, this corresponds to a quantile level $\alpha = 84.8\%$.
  - $F_{\text{female}}(170) = 84.8\%$.



Intuitive Example

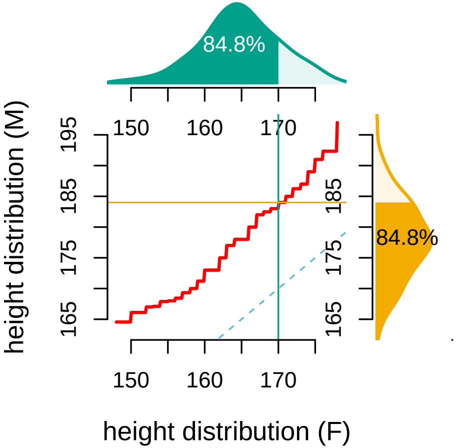
- The corresponding quantile in the height distribution of **males** is:
  - $F^{-1}_{\text{male}}(84.8\%) = 184\text{cm}\ (\approx 6')$.



Intuitive Example

Counterfactual of a 170cm (=5' 7") **female** had she been a **male**?

$$T^\star(170) = \left(F_{\text{male}}^{-1} \circ F_{\text{female}} \right)(170)$$
$$= 184 \text{ cm } (\approx 6').$$



A Few Notations

- Y: observed outcome.
  - e.g., loan approval ($Y \in \{0, 1\}$, premium ($Y \in [0, 1]$), earnings ($Y \in \mathbb{R}$).
- $S \in \{0, 1\}$: binary **sensitive attribue** requiring fairness consideration.
  - e.g., race ($S =$ **Black**, **Non Black**), sex ($S = \{$**Female**, **Male**$\}$).
- X: features that may be influenced by the sensitive.
- $Y^\star(0)$, $Y^\star(1)$: **potential** outcomes in the **protected**/**unprotected** groups.
- If we observed outome Y for some individual in group $S = 0$,

  the **counterfacual** outcomes would be $Y^\star(1)$.

Mutatis Mutandis: Two Key Approaches

- **Causal Graphs** Plečko and Meinshausen (2020); Plečko et al. (2024)
  - Based on the **causal inference** framework (Pearl, 2009; Pearl and Mackenzie, 2018; Chernozhukov et al., 2024)
  - Strong advantage: explainability

$$\text{CATE} = \mathbb{E}[\underbrace{Y^\star(1)}_{\text{potential outcomes unprotected group}} - \underbrace{Y^\star(0)}_{\text{potential outcomes protected group}} | \boldsymbol{X} = \boldsymbol{x}] \overset{?}{=} 0$$

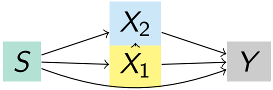
- **Optimal Transport** (De Lara et al., 2021; Charpentier et al., 2023)
  - Treat fairness adjustment as a **transport** problem in probability spaces.

$$\mathbb{E}[Y^\star(1) | \boldsymbol{X} = \boldsymbol{x}^\star(1)] - \mathbb{E}[Y^\star(0) | \boldsymbol{X} = \boldsymbol{x}] \overset{?}{=} 0$$

- Our contribution: **sequential transport** unifies these two approaches.

Graphical Models and Causal Networks

- A **Directed Acyclic Graph** (DAG) $\mathcal{G} = (V, E)$ models relationships between variables as nodes (V) and edges (E).



- Such a causal graph imposes some ordering on variables, referred to as "**topological sorting**" Ahuja et al. (1993). Here,

$$S \rightarrow X_2 \rightarrow X_1 \rightarrow Y \quad .$$

- The joint distribution of $X = (X_1, \ldots, X_d)$ satisfies the **Markov property**:

$$\mathbb{P}[x_1, \cdots, x_d] = \prod_{j=1}^{d} \mathbb{P}[x_j | \text{parents}(x_j)],$$

where $\text{parents}(x_i)$ are the immediate causes of x_i.

Counterfactual for Non Linear Models

- From Pearl (2000), let C, X, E be absolutely continuous, and consider i where $E_i = h_i(\text{parents}(E_i), U_i)$ with $\text{parents}(E_i) = \boldsymbol{x}$ fixed.
- Define $h_{i|\boldsymbol{x}}(u) = h_i(\boldsymbol{x}, u)$.
- $e_i = h_{i|\boldsymbol{x}}(u_i)$ represents the conditional quantile of E_i at probability level u_i.
- Its **counterfactual counterpart** $e_i^\star$ is the conditional quantile (conditioned on $\boldsymbol{x}^*$) at the same level u_i.

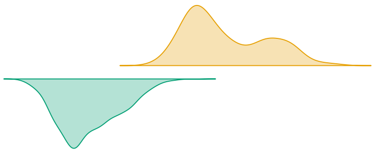
$$\begin{tikzcd} u_C \arrow[d] & u_X \arrow[d] & u_E \arrow[d] \\ C \arrow[r] \arrow[rr, bend right] & X \arrow[r] & E \end{tikzcd} \qquad \begin{cases} C = h_c(U_C) \\ X = h_x(C, U_X) \\ E = h_e(C, X, U_E), \end{cases}$$

$$\begin{tikzcd} & u_X \arrow[d] & u_E \arrow[d] \\ c \arrow[r] \arrow[rr, bend right] & X^\star \arrow[r] & E^\star \end{tikzcd} \qquad \begin{cases} C = c \quad (\text{or } \text{do}(C = c)) \\ X_c^\star = h_x(c, U_X) \\ E_c^\star = h_e(c, X_c^\star, U_E), \end{cases}$$

where $u \mapsto h_c(\cdot, u)$, $u \mapsto h_x(\cdot, u)$ and $u \mapsto h_e(\cdot, u)$ are strictly increasing in u, U_C, U_X and U_E are independent, supposed to be uniform on $[0, 1]$.

Optimal Transport and Monge Mapping

- **Optimal Transport**: how to find the best way to transport mass from **one distribution** to **another** while minimizing a given cost.
- It involves constructing a **joint distribution** (coupling) between two marginal probability measures (Villani, 2003, 2009).
- Consider a measure μ_0 (resp. μ_1) on a metric space $\mathcal{X}_0$ (resp. $\mathcal{X}_1$). The goal is to move every elementary mass from μ_0 to μ_1 in the most "efficient way."

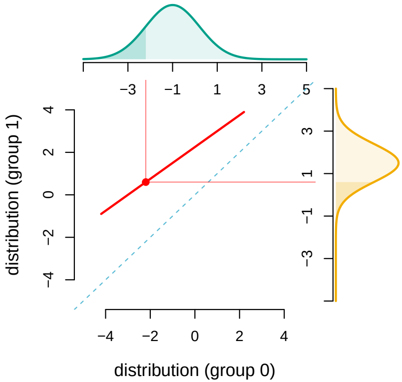


From Monge (1781): Mémoire sur la théorie des **déblais** et des **remblais**.

excavation

backfill

Univariate Optimal Transport Map



- From Santambrogio (2015), the optimal Monge map $T^\star$ for some strictly convex cost c such that $T^\star_\# \mu_0 = \mu_1$ is:

$$T^\star = \underbrace{F_1^{-1}}_{\text{quantile function}} \circ \underbrace{F_0}_{\text{cdf for } \mu_0},$$

Topological Ordering (1/4)

Step 1: Assuming a causal graph $\mathcal{G}$.
Step 2: Derive the **topological ordering** from the DAG:

- **Knothe-Rosenblatt rearrangement** (Bonnotte, 2013), inspired by the Rosenblatt chain rule: provides the "monotone lower triangular map" ("marginally optimal" Villani, 2003)

$$T_{\underline{kr}}(x_1, \cdots, x_d) = \begin{pmatrix} T^\star_{\underline{1}}(x_1) \\ T^\star_{\underline{2}}(x_2|x_1) \\ \vdots \\ T^\star_{\underline{d-1}}(x_{d-1}|x_1, \cdots, x_{d-2}) \\ T^\star_{\underline{d}}(x_d|x_1, \cdots, x_{d-1}) \end{pmatrix}.$$

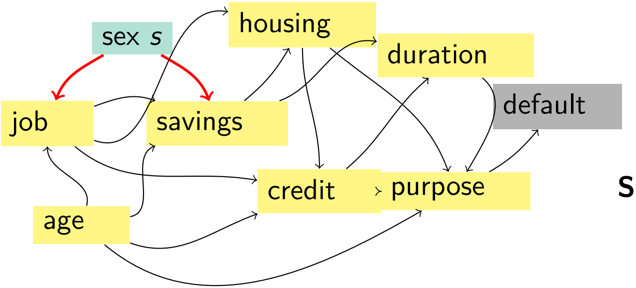
$\rightarrow$ Sequentially mapping $\mathbf{X}|S=0$ to $\mathbf{X}|S=1$ by conditioning on each preceding node in the topological order.

Topological Ordering (2/4)

- **Sequential Transport** extends the Knothe-Rosenblatt map to transport individuals from $\mathbf{X}|S=0$ to $\mathbf{X}|S=1$, while respecting any assumed underlying causal graph.
- The sequential conditional transport on graph $\mathcal{G}$ writes:

$$T^{\star}_{\mathcal{G}}(x_1, \cdots, x_d) = \begin{pmatrix} T^{\star}_1(x_1) \\ T^{\star}_2(x_2|\ parents(x_2)) \\ \vdots \\ T^{\star}_{d-1}(x_{d-1}|\ parents(x_{d-1})) \\ T^{\star}_d(x_d|\ parents(x_d)) \end{pmatrix}.$$

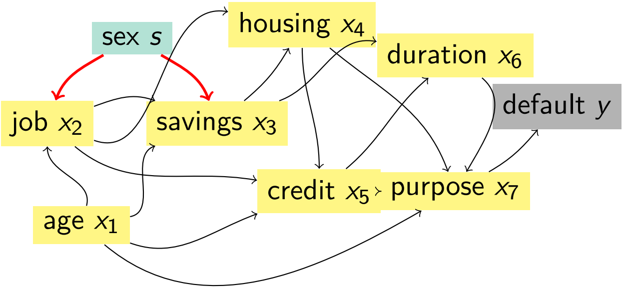
Topological Ordering (3/4)



Causal graph in the German Credit dataset from Watson et al. (2021).

Step 1: Asusming a causal graph $\mathcal{G}$.

Topological Ordering (4/4)



Causal graph in the German Credit dataset from Watson et al. (2021).

- **Step 2**: sequential conditional transport based on a topological ordering:

$$T^{\star}_{\mathcal{G}}(x_1, \cdots, x_7) = \begin{pmatrix} T^{\star}_1(x_1) \\ T^{\star}_2(x_2|x_1) \\ T^{\star}_3(x_3|x_1, x_2) \\ T^{\star}_4(x_4|x_2, x_3) \\ T^{\star}_5(x_5|x_1, x_2, x_4) \\ T^{\star}_6(x_6|x_3, x_5) \\ T^{\star}_7(x_7|x_1, x_4, x_5, x_6) \end{pmatrix}.$$

Example With Simulated Data

We generate a sample $\{(S_i, X_{1i}, X_{2i}, Y_i)\}_{i=1}^{200}$, with $S \in \{0,1\}$, and the covariates $\boldsymbol{X} = (X_1, X_2)$ are drawn from two bivariate Gaussian distributions with **group-specific parameters**.

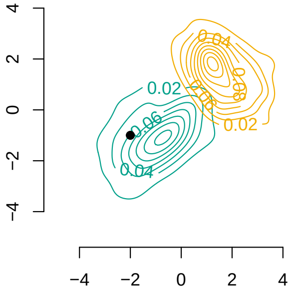
$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$, $\mu_s = \begin{pmatrix} \mu_{s,X_1} \\ \mu_{s,X_2} \end{pmatrix}$, $\Sigma_s = \begin{pmatrix} \sigma^2_{s,X_1} & \rho_{s,X_1,X_2} \\ \rho_{s,X_1,X_2} & \sigma^2_{s,X_2} \end{pmatrix}$, for $s = \{0, 1\}$.

Each outcome Y is drawn from a $\mathcal{B}er(p_s)$ with

$$p_s = exp(eta_s)/(1 + exp(eta_s))$$

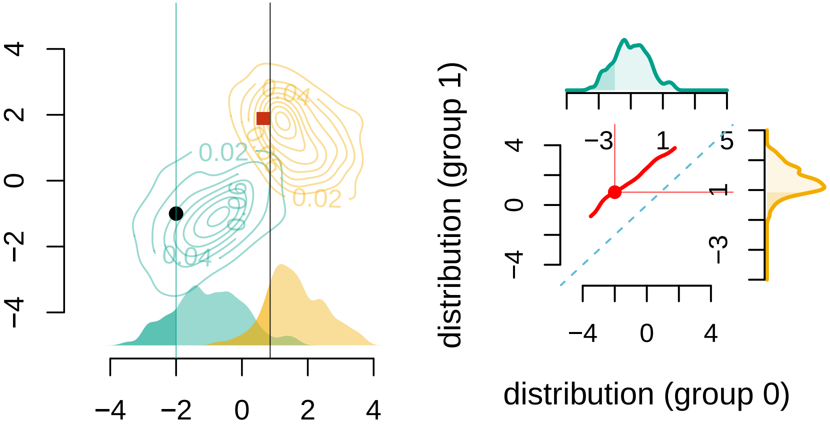
$$\text{where } \begin{cases} \eta_0 = 0.6X1 + 0.2X2 \\ \eta_1 = 0.4X_1 + 0.3X2. \end{cases}$$

Let us focus on **individual** $(s = 0, x_1 = -2, x_2 = -1)$.



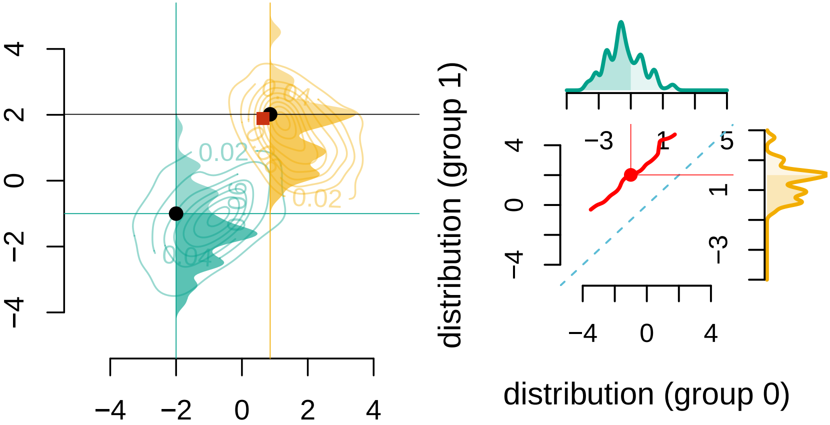
Estimated Densities of the Simulated Data in Both Groups.

Transport $x_1 \mid s$ From Group 0 to Group 1



Sequential Transport (simulated data). Red square: multivariate OT. **transport** $x_1 \mid s$.

Transport $x_2 \mid x_1, s$ From Group 0 to Group 1



Sequential Transport (simulated data). Red square: multivariate OT. **transport** $x_2 \mid x_1, s$

Code

This can be easily done with our functions from our small package:

```r
remotes::install_github(
  repo = "fer-agathe/sequential_transport", subdir = "seqtransfairness")
library(seqtransfairness)
sim_dat <- simul_dataset() # Simulate data
variables <- c("S", "X1", "X2", "Y")
adj <- matrix(
  # S  X1 X2 Y
  c(0, 1, 1, 1,# S
    0, 0, 1, 1,# X1
    0, 0, 0, 1,# X2
    0, 0, 0, 0  # Y
  ),
  ncol = length(variables), byrow = TRUE
  dimnames = rep(list(variables), 2))
# Sequential transport according to the causal graph
transported <- seq_trans(data = sim_dat, adj = adj, s = "S", S_0 = 0, y = "Y")
predict(transported) # Transp. values from S=0 to S=1, using the causal graph.
```

Interpretable Counterfactual Fairness

Now, assume a logistic regression model was fitted on the simulated data and returned scores according to:

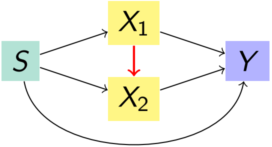
$$m(x_1, x_2, s) = \left(1 + \exp\left[-\left((x_1 + x_2)/2 + \mathbf{1}(s = 1)\right)\right]\right)^{-1}.$$

Observation: $(\mathbf{s=0}, x_1 = -2, x_2 = -1)$

Prediction : $m(\mathbf{0}, -2, -1) = 18.24\%.$

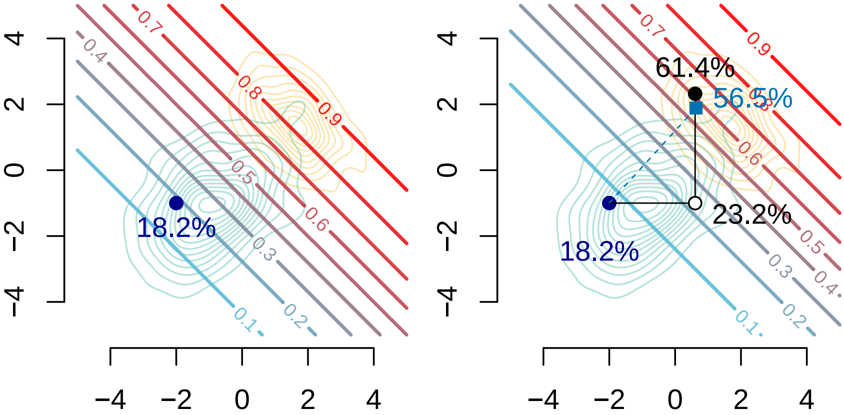
Pred. with Seq. T : $m(s = 1, x_1^\star, x_2^\star) = 61.4\%$

Pred with OT : $m(s = 1, x_1^\star, x_2^\star) = 56.5\%$



Assumed causal structure.

Counterfactual assuming X_2 is caused by X_1



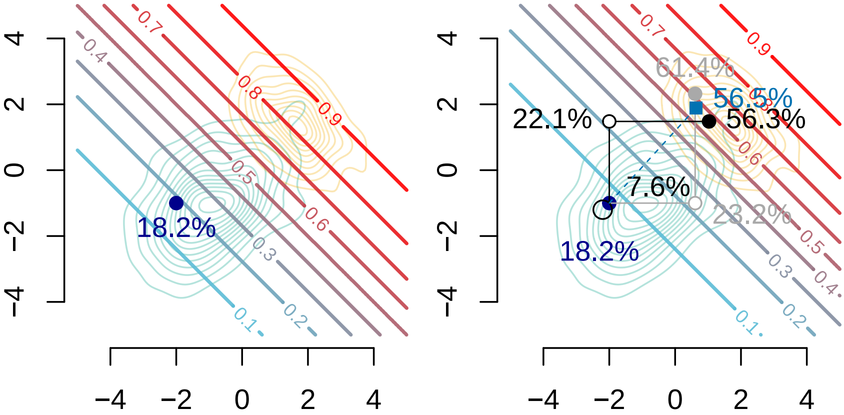
Predictions by m of: the **observation** using factual (left), counterfactual (right): **counterfactual by Seq. T.** (assuming $X_1 \rightarrow X_2$) and **optimal. transport**.

Decomposition of the *mutatis mutandis* difference

The ***mutatis mutandis* difference can be decomposed:**

$$
\begin{aligned}
& m(s=1, x_1^\star, x_2^\star) - m(s=0, x_1, x_2) = +43.16\% \ (\textit{\textbf{mutatis mutandis}}\ \textbf{diff.}) \\
= \quad & m(s=1, x_1, x_2) - m(s=0, x_1, x_2) \quad : -10.66\% \ (\textit{\textbf{cet. par. diff.}}) \\
+ \quad & m(s=1, x_1^\star, x_2) - m(s=1, x_1, x_2) \quad : +15.63\% \ (\textbf{change in } x_1) \\
+ \quad & m(s=1, x_1^\star, x_2^\star) - m(s=1, x_1^\star, x_2) \quad : +38.18\% \ (\textbf{change in } x_2 | x_1^\star) .
\end{aligned}
$$

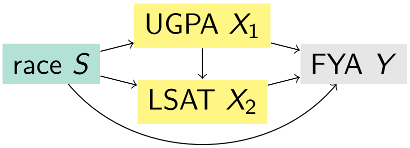
Counterfactual assuming X_1 is caused by X_2



Predictions by m of: the **observation** using factual (left), counterfactual (right): **counterfactual by Seq. T.** (assuming $X_2 \rightarrow X_1$) and **optimal. transport** .

Application on Real Data

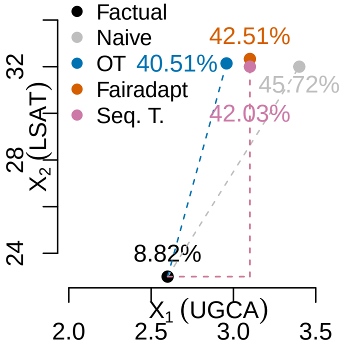
- Law School Admission Council Dataset (Wightman, 1998)
- 1st year law school grade (FYA) > median? ($Y \in \{0, 1\}$)
- Race ($s \in \{$**Black**, **White**$\}$)
- Undergrad. GPA before law school (x_1, UGPA)
- Law School Admission Test (x_2, LSAT)
- Logistic model (**aware**, i.e., including **S**)



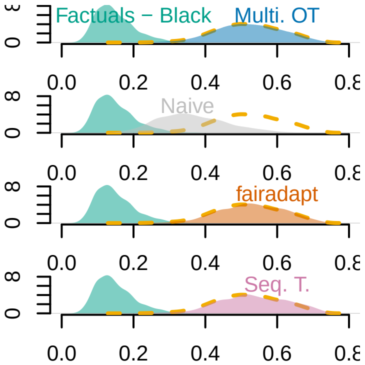
Assumed causal graph.

Predictions with: **factuals**, **naive** (cet. par.), **optimal transport**, **fairadapt**, **sequential transport**

Application on Real Data



Pred. for a **Black indiv.** using their factual and counterfactual characteristics



Densities of predicted scores. Yellow line: **density for White indiv.**

Global Fairness Metrics

A model m satisfies the **independence property** if $m(\boldsymbol{X}, S) \perp\!\!\!\perp S$, with respect to the distribution $\mathbb{P}$ of the triplet $(\boldsymbol{X}, S, Y)$ (Barocas et al., 2017).

$$\textbf{Demographic Parity} \rightarrow \mathbb{E}[\underbrace{\hat{Y}}_{\text{score } \hat{y}} \mid S = A] \overset{?}{=} \mathbb{E}[\underbrace{\hat{Y}}_{\text{score } \hat{y}} \mid S = B]$$

Demographic Parity can be extended to **Counterfactual Demographic Parity**, allowing fairness assessment within subgroup $s = 0$:

$$\mathrm{CDP} = \frac{1}{n_0} \sum_{i \in \mathcal{D}_0} m(1, \boldsymbol{x}_i^\star) - m(0, \boldsymbol{x}_i),$$

i.e., "**average treatment effect of the treated**" in the classical causal literature.

Global Fairness Metrics

	Naive	Fairadapt	multi. OT	seq. T
Aware model	0.22	0.38	0.37	0.37
Unaware model	0	0.19	0.18	0.18

Table 1: Counterfactual Demographic Parity comparing predictions using $(s = 0, \boldsymbol{x})$ (factuals) and using $(x = 1, \boldsymbol{x}^{\star})$ (counterfactuals), for the aware model (which includes S) and the unaware model (which does not).

Conclusion

- We introduced **sequential transport** as a novel approach to individual fairness:
  - Linking causal graph approach to optimal transport approach.
- Provides an **interpretable closed-form solution**.

arXiv:2408.03425 fer-agathe/sequential_transport

Agathe Fernandes Machado Ewen Gallic Arthur Charpentier

Comments are welcome: fernandes_machado.agathe@courrier.uqam.ca

References I

Ahuja, R. K., Magnanti, T. L., and Orlin, J. B. (1993). *Network flows: Theory, algorithms, and applications*. Prentice Hall.

Barocas, S., Hardt, M., and Narayanan, A. (2017). Fairness in machine learning. *Nips tutorial*, 1:2017.

Barocas, S., Hardt, M., and Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT press.

Bonnotte, N. (2013). From Knothe's rearrangement to Brenier's optimal transport map. *SIAM Journal on Mathematical Analysis*, 45(1):64–87.

Charpentier, A., Flachaire, E., and Gallic, E. (2023). Optimal transport for counterfactual estimation: A method for causal inference. In *Optimal Transport Statistics for Economics and Related Topics*, pages 45–89. Springer.

Chernozhukov, V., Hansen, C., Kallus, N., Spindler, M., and Syrgkanis, V. (2024). Applied causal inference powered by ml and ai. *arXiv preprint arXiv:2403.02467*.

De Lara, L., González-Sanz, A., Asher, N., and Loubes, J.-M. (2021). Transport-based counterfactual models.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

References II

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323.

Kilbertus, N., Rojas Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. *Advances in neural information processing systems*, 30.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4066–4076. NIPS.

Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*.

Pearl, J. (2000). Comment. *Journal of the American Statistical Association*, 95(450):428–431.

Pearl, J. (2009). *Causality*. Cambridge university press.

Pearl, J. and Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic books.

Plečko, D. and Meinshausen, N. (2020). Fair data adaptation with quantile preservation. *Journal of Machine Learning Research*, 21(242):1–44.

References III

Plečko, D., Bennett, N., and Meinshausen, N. (2024). fairadapt: Causal reasoning for fair data preprocessing. *Journal of Statistical Software*, 110(4):1–35.

Santambrogio, F. (2015). *Optimal transport for applied mathematicians*. Springer.

Villani, C. (2003). *Topics in optimal transportation*, volume 58. American Mathematical Society.

Villani, C. (2009). *Optimal Transport*. Springer Berlin Heidelberg.

Watson, D. S., Gultchin, L., Taly, A., and Floridi, L. (2021). Local explanations via necessity and sufficiency: unifying theory and practice. In de Campos, C. and Maathuis, M. H., editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 1382–1392. PMLR.

Wightman, L. F. (1998). Lsac national longitudinal bar passage study. lsac research report series. Technical report, Law School Admission Council, Newtown, PA.