Sequential Conditional Transport on Probabilistic Graphs for Interpretable Counterfactual Fairness

Agathe FERNANDES MACHADO[a], Arthur CHARPENTIER[a], and Ewen GALLIC[b]

[a]Université du Québec à Montréal (fernandes_machado.agathe@courrier.uqam.ca), [b]AMSE, Aix-Marseille Université

MOTIVATIONS

Individual algorithmic fairness: similar individual should receive similar outcomes, regardless of the **sensitive attribute** [3].
Counterfactual fairness: evaluate whether a model's decision would have remained unchanged under a hypothetical alteration of the **sensitive attributes**.

	Ceteris Paribus	*Mutatis Mutandis* [5, 4]
Idea	Change **S** all other things equal	Some features may be influenced by **S** via **legitimate pathways**
Counterfactual of $(s=0, x)$	$(s=1, x)$	$(s=1, x^\star(1))$
Fairness	$\mathbb{E}[\,Y^\star(1) - Y^\star(0)\,\vert \boldsymbol{X}=\boldsymbol{x}] = 0$	$\mathbb{E}[\,Y^\star(1)\,\vert\,\boldsymbol{X}=\boldsymbol{x}^\star(1)\,] - \mathbb{E}[\,Y^\star(0)\,\vert \boldsymbol{X}=\boldsymbol{x}] = 0$

Where $Y^\star(1)$ and $Y^\star(0)$ are potential outcomes if $S=1$ and $S=0$, respectively. The literature looking at *mutatis mutandis* counterfactual fairness has developped two approaches based on: (i) quantile preservation on **causal graphs** [8, 9] (fairadapt), (ii) **multivariate optimal transport** (OT) [2].

Our contribution: Sequential Transport (ST), bridging the two approaches.

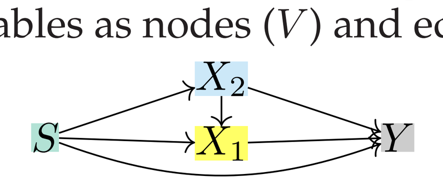
Comparison of mutatis mutandis *counterfactual fairness methods.*

Approach	Strengths	Weaknesses
Causal Graphs	• Interpretability • Aligns with causal theory	• Computationally intensive for large models • Requires a known causal graph
Optimal Transport (OT)	• Handles robust distributions • Computationally efficient	• Limited interpretability • Ignores causal relationships between variables
Sequential Transport	• Interpretability: closed-form solutions for counterfactuals using **univariate OT** • Aligns with **causal theory**	• Computationally intensive for large models • Requires a known causal graph

PROBABILISTIC GRAPHICAL MODELS

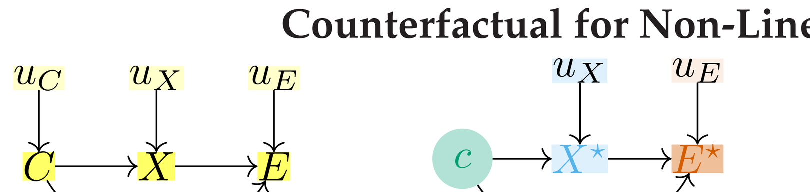
Probabilistic Graphical Models

- 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**" [1]. Here,
$$S \to X_2 \to X_1 \to Y \ .$$
- Each edge $x_i \to x_j$ represents a causal relationship, where x_i directly influences x_j.
- The joint distribution of the variables $X = (X_1, \dots, X_d)$ satisfies the **Markov property**:
$$\mathbb{P}[x_1, \cdots, x_d] = \prod_{j=1}^{d} \mathbb{P}[x_j \vert \text{parents}(x_j)],$$
where parents(x_i) are the immediate causes of x_i.

Counterfactual for Non-Linear Structural Models [7]



$$\begin{cases} C = h_c(U_C) \\ X = h_x(C, U_X) \\ E = h_e(C, X, U_E), \end{cases} \quad \begin{cases} C = c \ \ (\text{or do}(C=c)) \\ X_c^\star = h_x(c, U_X) \\ E_c^\star = h_e(c, X_c^\star, U_E), \end{cases}$$

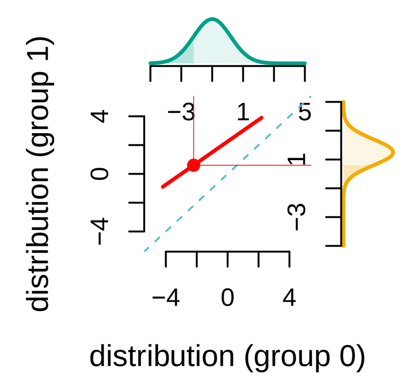
where $u \mapsto h_c(\cdot, u)$, $u \mapsto h_x(\cdot, u)$ and $u \mapsto h_e(\cdot, u)$ are strictly increasing in u, U_C, U_X and U_E are independent, supposed to be uniform on $[0, 1]$.

Let C, X, E be absolutely continuous, and consider i where $E_i = h_i(\text{parents}(E_i), U_i)$ with parents$(E_i) = \boldsymbol{x}$ fixed. Define $h_{i\vert\boldsymbol{x}}(u) = h_i(\boldsymbol{x}, u)$. Then, $e_i = h_{i\vert\boldsymbol{x}}(u_i)$ 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.

OPTIMAL TRANSPORT (OT)

Given two distributions μ_0 and μ_1 over spaces $\mathcal{X}_0$ and $\mathcal{X}_1$, OT finds a mapping $T : \mathcal{X}_0 \to \mathcal{X}_1$ that minimizes the cost of moving mass from μ_0 to μ_1. If we consider $\mathcal{X}_0 = \mathcal{X}_1$ as a compact subset of $\mathbb{R}^d$, there exists T such that $\mu_1 = T_\#\mu_0$ (push-forward of μ_0 by T) when μ_0 and μ_1 are two measures, and μ_0 is atomless. If μ_0 and μ_1 are absolutely continuous w.r.t. Lebesgue measure, we can find an "optimal" mapping T^* satisfying Monge's problem [6]. For some positive cost function $c : \mathcal{X}_0 \times \mathcal{X}_1 \to \mathbb{R}_+$,
$$T^* := \inf_{T_\#\mu_0 = \mu_1} \int_{\mathcal{X}_0} c\big(\boldsymbol{x}_0, T(\boldsymbol{x}_0)\big)\mu_0(\mathrm{d}\boldsymbol{x}_0).$$

Univariate OT for Gaussian distributions.



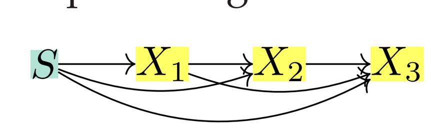
Univariate Case: the optimal Monge map $T^\star$ for some strictly convex cost c such that $T^\star_\#\mu_0 = \mu_1$ is
$$T^\star = \underbrace{F_1^{-1}}_{\text{quantile function}} \circ \underbrace{F_0}_{\text{cumul. distrib. function}}$$

Multivariate Case: with strictly convex cost in $\mathbb{R}^d \times \mathbb{R}^d$, the Jacobian matrix $\nabla T^\star$, even if not necessarily nonnegative symmetric, is diagonalizable with nonnegative eigenvalues. But, it is generally difficult to give an analytic expression for $T^\star$.

SEQUENTIAL TRANSPORT

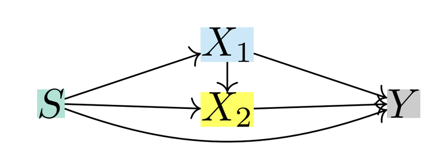
Let X_1, X_2 and X_3 be continuous variables, with continuous conditionals. We aim to **transport** an individual $(S = 0, x_1, x_2, x_3)$ from group $\{S = 0\}$ to $\{S = 1\}$, following the **topological order** of a DAG.

Knothe-Rosenblatt (KR) Conditional Transport
The KR rearrangement, inspired by the Rosenblatt chain rule, provides the "monotone lower triangular map" ("marginally optimal" [10]), sequentially mapping $\mathbf{X}\vert S=0$ to $\mathbf{X}\vert S=1$ by conditioning on each preceding node in the topological order.



$$T_{\underline{kr}}(x_1, x_2, x_3) = \begin{pmatrix} T^\star_{\underline{1}}(x_1 \vert S=0) \\ T^\star_{\underline{2}\vert 1}(x_2 \vert x_1, S=0) \\ T^\star_{\underline{3}\vert 1,2}(x_3 \vert x_2, x_1, S=0) \end{pmatrix}$$

Example of Sequential Transport (ST). ST extends the KR map to transport individuals from $\mathbf{X}\vert S = 0$ to $\mathbf{X}\vert S = 1$, while respecting any assumed underlying causal graph.

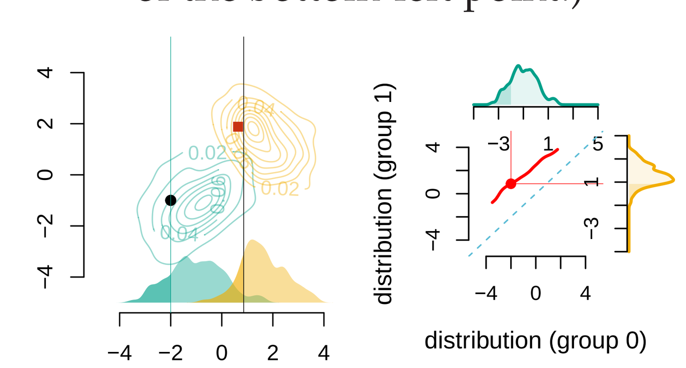


$$T_{\underline{st}}(x_1, x_2) = \begin{pmatrix} T^\star_{\underline{1}}(x_1 \vert S=0) \\ T^\star_{\underline{2}\vert 1}(x_2 \vert x_1, S=0) \end{pmatrix}$$

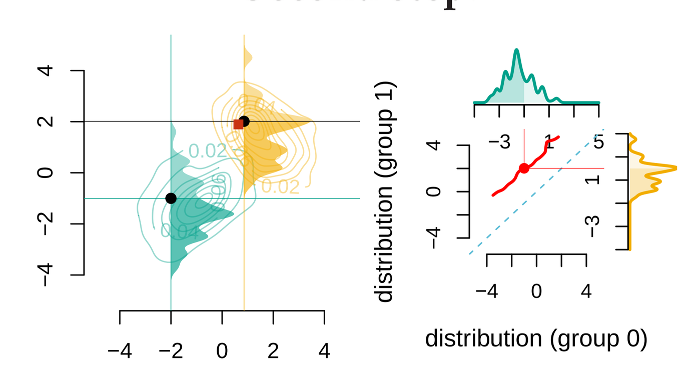
Algorithm 1: Sequential transport on causal graph

Require: graph $\mathcal{G}$ on $(s, \boldsymbol{x})$, with adjacency matrix $\boldsymbol{A}$
Require: dataset $(s_i, \boldsymbol{x}_i)$ and one individual $(s = 0, \boldsymbol{a})$
Require: bandwidths $\boldsymbol{h}$ and $\boldsymbol{b}_j$'s
  $(s, \boldsymbol{v}) \leftarrow \boldsymbol{A}$ the topological ordering of vertices (DFS)
  $T_s \leftarrow$ identity
  for $j \in \boldsymbol{v}$ **do**
    $\boldsymbol{p}(j) \leftarrow$ parents(j)
    $T_j(\boldsymbol{a}_{\boldsymbol{p}(j)}) \leftarrow (T_{\boldsymbol{p}(j)_1}(\boldsymbol{a}_{\boldsymbol{p}(j)}), \cdots, T_{\boldsymbol{p}(j)_{k_j}}(\boldsymbol{a}_{\boldsymbol{p}(j)}))$
    $(x_{i,j\vert s}, \boldsymbol{x}_{i,\boldsymbol{p}(j)\vert s}) \leftarrow$ subsets when $s \in \{0, 1\}$
    $w_{i,j\vert 0} \leftarrow \phi(\boldsymbol{x}_{i,\boldsymbol{p}(j)\vert 0}; \boldsymbol{a}_{\boldsymbol{p}(j)}, \boldsymbol{b}_j)$ (Gaussian kernel)
    $w_{i,j\vert 1} \leftarrow \phi(\boldsymbol{x}_{i,\boldsymbol{p}(j)\vert 1}; T_j(\boldsymbol{a}_{\boldsymbol{p}(j)}), \boldsymbol{b}_j)$
    $\hat{f}_{h_j\vert s} \leftarrow$ density estimator of $x_{\cdot,j\vert s}$, weights $w_{\cdot,j\vert s}$.
    $\hat{F}_{h_j\vert s}(\cdot) \leftarrow \int_{-\infty}^{\cdot} \hat{f}_{h_j\vert s}(u)\mathrm{d}u$, c.d.f.
    $\hat{Q}_{h_j\vert s} \leftarrow \hat{F}^{-1}_{h_j\vert s}$, quantile
    $\hat{T}_j(\cdot) \leftarrow \hat{Q}_{h_j\vert 1} \circ \hat{F}_{h_j\vert 0}(\cdot)$
  end for
  $\boldsymbol{a}^\star \leftarrow (T_1(\boldsymbol{a}_1), \cdots, T_d(\boldsymbol{a}_d))$
  return $(s = 1, \boldsymbol{a}^\star)$, counterfactual of $(s = 0, \boldsymbol{a})$

First step. (Red square: multivariate OT of the bottom-left point.)



Second step.



INTERPRETABLE COUNTERFACTUAL FAIRNESS

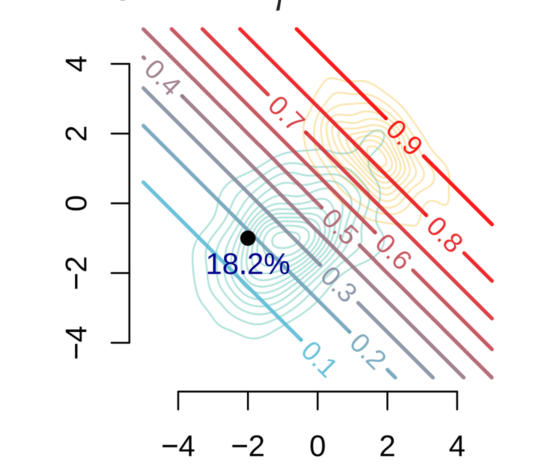
Consider a predictive model m with iso scores shown in the figures on the right for **group 0** (top) and **group 1** (bottom):
$$m(s, x_1, x_2) = \big(1 + \exp\big[-\big((x_1 + x_2)/2 + s\big)\big]\big)^{-1} \ .$$

Observation ($\mathbf{s{=}0}$, $x_1 = -2, x_2 = -1$) with $m(0, -2, -1) = 18.24\%$.
Counterfactual prediction $m(s = 1, x_1^\star, x_2^\star)$ is constructed using Algo. 1, assuming either $X_1 \to X_2$ (bottom right path, predicted 61.4%) or $X_2 \to X_1$ (top left path, predicted 56.3%).
The ***mutatis mutandis* difference** can be decomposed, using the ***ceteris paribus* difference**, **the change in x_1**, and **the change in x_2 conditional on the change in x_1**:

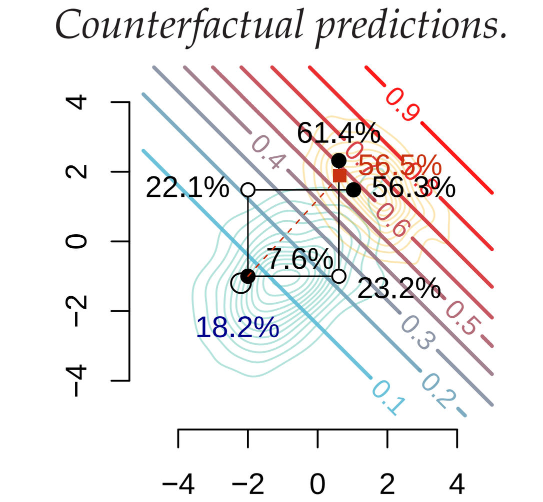
$$\begin{aligned} & m(s=1, x_1^\star, x_2^\star) - m(s=0, x_1, x_2) = +43.16\% \\ = \ & m(s=1, x_1, x_2) - m(s=0, x_1, x_2) \quad : -10.66\% \\ + \ & m(s=1, x_1^\star, x_2) - m(s=1, x_1, x_2) \quad : +15.63\% \\ + \ & m(s=1, x_1^\star, x_2^\star) - m(s=1, x_1^\star, x_2) \quad : +38.18\% . \end{aligned}$$

Fairness metric. **Demographic Parity** can be extended to **Counterfactual Demographic Parity**, allowing fairness assessment within subgroup $s = 0$ (more fairness criteria in the paper):
$$\text{CDP} = \frac{1}{n_0} \sum_{i \in \mathcal{D}_0} m(1, \boldsymbol{x}_i^\star) - m(0, \boldsymbol{x}_i),$$
i.e., "**average treatment effect of the treated**" in the classical causal literature.

Observed prediction.



Counterfactual predictions.

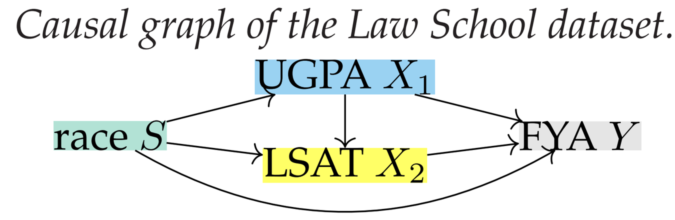


Red dot: multivariate OT.

APPLICATION ON REAL DATA

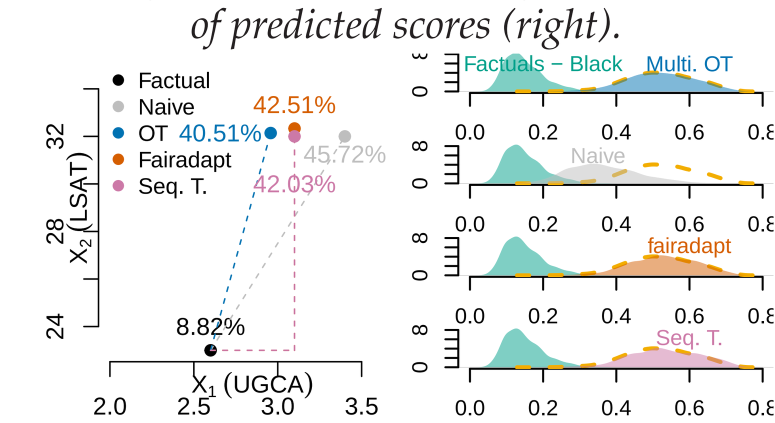
- Law School Admission Council Dataset
- 1st year law school grade (FYA) > median?
- Race ($s \in \{$**Black**, **White**$\}$)
- Undergrad. GPA before law school (x_1, UGPA), Law School Admission Test (x_2, LSAT).
- Logistic model (aware, i.e., including **S**)

Causal graph of the Law School dataset.



We compare predicted values using **factuals**, ***ceteris paribus* counterfactuals**, **optimal transport**, **fairadapt**, and **sequential transport**. The left figure shows results for **a Black individual** (black dot). The right figure shows the densities of estimated scores.

Counterfactual calculations (left) and densities of predicted scores (right).



CDP for Black individuals comparing classifier predictions over original features $\boldsymbol{x}$ (resp. $(s=0, \boldsymbol{x})$) and their counterfactuals $\boldsymbol{x}^\star$ (resp. $(s=1, \boldsymbol{x}^\star)$).

	Fairadapt	multi. OT	seq. T
Aware model	0.3810	0.3727	0.3723
Unaware model	0.1918	0.1821	0.1817

REFERENCES

[1] Ahuja, R. K., Magnanti, T. L., and Orlin, J. B. (1993). *Network flows: Theory, algorithms, and applications*. Prentice Hall.
[2] De Lara, L., González-Sanz, A., Asher, N., and Loubes, J.-M. (2021). Transport-based counterfactual models.
[3] 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.
[4] 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.
[5] 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.
[6] Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*.
[7] Pearl, J. (2000). Comment. *Journal of the American Statistical Association*, 95(450):428–431.
[8] Plečko, D. and Meinshausen, N. (2020). Fair data adaptation with quantile preservation. *Journal of Machine Learning Research*, 21(242):1–44.
[9] Plečko, D., Bennett, N., and Meinshausen, N. (2024). fairadapt: Causal reasoning for fair data preprocessing. *Journal of Statistical Software*, 110(4):1–35.
[10] Villani, C. (2003). *Topics in optimal transportation*, volume 58. American Mathematical Society.

AAAI-25 / IAAI-25 / EAAI-25