Assessing Counterfactual Fairness via (Marginally) Optimal Transport

Agathe Fernandes Machado
joint with Arthur Charpentier and Ewen Gallic

Journée de la recherche STATQAM
May 15, 2025
Université du Québec à Montréal

Table of contents

Algorithmic Fairness: Introduction

Regulation: Protected/Sensitive Attributes

Charter of Fundamental Rights of the European Union (18.12.2000, C364), Article 21

"*Any discrimination based on any ground such as **sex**, **race**, **colour**, **ethnic or social origin**, **genetic features**, **language**, **religion or belief**, **political or any other opinion**, **membership of a national minority**, **property**, **birth**, **disability**, **age** or **sexual orientation** shall be prohibited.*"

Au Québec, Charte des droits et libertés de la personne (C-12), Article 10

"*Toute personne a droit à la reconnaissance et à l'exercice, en pleine égalité, des droits et libertés de la personne, sans distinction, exclusion ou préférence fondée sur la **race**, la **couleur**, le **sexe**, l'**identité** ou l'**expression de genre**, la **grossesse**, l'**orientation sexuelle**, l'**état civil**, l'**âge** sauf dans la mesure prévue par la loi, la **religion**, les **convictions politiques**, la **langue**, l'**origine ethnique ou nationale**, la **condition sociale**, le **handicap** ou l'utilisation d'un moyen pour pallier ce handicap [...]* "

Regulation: Discrimination in Predictive Models

"*The following **AI practices shall be prohibited**: the placing on the market, the putting into service for this specific purpose, or the **use of biometric categorisation systems** that categorise individually natural persons based on their biometric data to deduce or infer their race, political opinions, trade union membership, religious or philosophical beliefs, sex life or sexual orientation*" (European Union AI Act, 2024)

"*State and federal law prohibit **insurers** from unlawfully **discriminating against certain protected classes** of individuals and from engaging in unfair discrimination, including the ability of insurers to underwrite based on certain criteria.*" (New York State Department of Financial Services, 2024)

Example: Recidivism Risk Assessment Tool

"*Our analysis of Northpointe's tool, called* ***COMPAS*** *[...] found that* ***black defendants were far more likely than white defendants to be incorrectly judged to be at a higher risk of recidivism****, while white defendants were more likely than black defendants to be incorrectly flagged as low risk.*" (Larson et al., 2016)

"*The COMPAS tool assigns defendants scores from 1 to 10 that indicate how likely they are to reoffend based on more than 100 factors, including age, sex and criminal history.* ***Notably, race is not used****.*" (Feller et al., 2016)

Due to the presence of **proxy variables** in the dataset, simply eliminating the sensitive attributes from predictive models does not guarantee fair predictions. (Upton and Cook, 2014)

Sources of Model Bias

- **Statistical bias in the data**: reproduction of past injustices, minority groups underrepresented in an imbalanced dataset;
- **Biased model**: arises from correlations between sensitive attributes and other explanatory variables (e.g., proxy variables);
- **Intentional bias**: the bias can also be the result of deliberate choices, this can be both benevolent or with malice.

What is Algorithmic Fairness?

We consider a machine learning (ML) model $m : \mathcal{X} \times \mathcal{S} \to \mathcal{Y}$ that predicts an outcome Y, e.g., the likelihood of receiving a job interview. Regulation may prohibit **discrimination** w.r.t. the **sensitive attribute** $S \in \mathcal{S}$, e.g., gender.

Approaches to **evaluate** and, if necessary, mitigate the unfairness of model predictions $\hat{Y} = m(\boldsymbol{X}, S)$ w.r.t. S:

- *Group fairness*: Compare $\hat{Y}$ between groups defined by S, e.g., the predicted scores of receiving a job interview for males *vs.* females (Barocas et al., 2023).
- Individual level: focus on a specific individual in the disadvantaged group,
  - *Individual fairness* "*any two individuals who are similar with respect to a particular task should be classified similarly*" (Dwork et al., 2012),
  - **Counterfactual fairness**: causality-based fairness (Plečko and Meinshausen, 2020; Plečko et al., 2024).

Causal Inference Framework

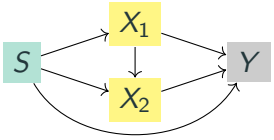
Probabilistic Graphical Models

- A Directed Acyclic Graph (**DAG**) $\mathcal{G} = (V, E)$ models relationships between variables as nodes ($V = \{X_1, \cdots, X_d\}$) and directed edges (E), such that $X_i \to X_j$ means "variable X_i causes variable X_j," (Koller and Friedman, 2009).
- Such a causal graph imposes some ordering on variables, referred to as "**topological sorting**" (Ahuja et al., 1993), where each node appears after all its parents.
- The joint distribution of $X = (X_1, \ldots, X_d)$ satisfies the **Markov property**:

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

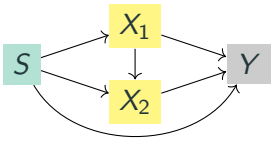
where parents(X_i) are the immediate causes of X_i.

Example: Causal Graph (1/2)



- $S \in \{\textbf{male}, \textbf{female}\}$ denotes the sensitive attribute: gender,
- X_1 is a "non-protected" explanatory variable: number of years of study,
- X_2 is another "non-protected" explanatory variable: number of years of professional experience,
- Y is the outcome variable: the likelihood of receiving a job interview, which we aim to predict using a model $m : \mathcal{X} \times \mathcal{S} \to [0, 1]$.

Example: Causal Graph (2/2)



- The **topological ordering** is $S \to X_1 \to X_2 \to Y$.
- The joint distribution of this DAG can be formulated as,

$$\forall (s, x_1, x_2, y) \in \mathcal{S} \times \mathcal{X} \times \mathcal{Y}, \;\; \mathbb{P}[s, x_1, x_2, y] = \mathbb{P}[s]\, \mathbb{P}[x_1|s]\, \mathbb{P}[x_2|s, x_1]\, \mathbb{P}[y|s, x_1, x_2] \;\; .$$

Quantifying Counterfactual Fairness

Underlying Question

We assume the previous graph represents a **known DAG**. Consider a trained ML model $m : \mathcal{S} \times \mathcal{X} \to \mathcal{Y}$, and the i-th observation from our dataset given by $(S_i = \text{female}, X_{1,i} = x_1, X_{2,i} = x_2)$, with $\hat{y}_i = m(\text{female}, x_1, x_2) = 18.24\%$.

1. **Defining a counterfactual**
   "What would my chances of getting an interview have been if I had been **male**?", e.g., how to define $\hat{y}^{\star}_{i,S \leftarrow \text{male}}$?
2. **Assessing Counterfactual Fairness**
   "Would my chances of getting an interview have been the same if I had been **male**?", e.g., do we have $|\hat{y}^{\star}_{i,S \leftarrow \text{male}} - \hat{y}^{\star}_{i,S \leftarrow \text{female}}| = |\hat{y}^{\star}_{i,S \leftarrow \text{male}} - \hat{y}_i| = 0$?

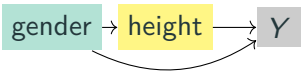
Defining a Counterfactual

How to calculate $\hat{y}^{\star}_{i,S\leftarrow \text{male}}$ for the i-th individual (female, x_1, x_2) with observed prediction $\hat{y}_i = \hat{y}^{\star}_{i,S\leftarrow \text{female}} = m(\text{female}, x_1, x_2)$?

- *Ceteris paribus*: ignoring causal relationships and simply computing $m(\text{male}, x_1, x_2)$, i.e., by changing only the value of the sensitive attribute;
- ***Mutatis mutandis*** (Kusner et al., 2017; Charpentier et al., 2023): within the causal inference framework (Pearl, 2009), explanatory variables $\boldsymbol{X}$, representing individual characteristics, must be transported if they lie in the causal descendants of the sensitive attribute S.

Mutatis Mutandis: Intuitive Example (1/3)

Consider a model m predicting Y based on gender and **height** and assume the following causal graph:



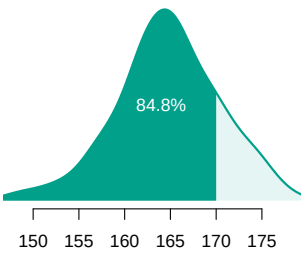
What is the counterfactual for $\hat{Y}$ of a **female** with height 170cm had she been a **male**?

$\rightarrow$ If we use the *ceteris paribus* approach, we would simply compute $\hat{Y}^{\star}_{S \leftarrow \text{male}}$ as $m(\text{male}, 170\text{cm})$. This approach completely ignores the fact that gender causally influences an individual's height. To properly compute the counterfactual $\hat{Y}$, we need to **transport the value of height** according to the change in gender, e.g., calculate the **counterfactual for height** first.

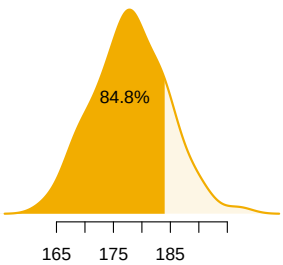
Mutatis Mutandis: Intuitive Example (2/3)

*"What is the height of a **female** of 170cm in the **counterfactual male** world?"*

Within the distribution of **females** in our dataset, this corresponds to a quantile level $\alpha = 84.8\%$, e.g., $F_{\text{female}}(170) = 84.8\%$.



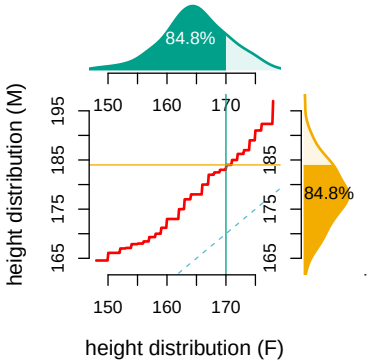
The corresponding quantile in the height distribution of **males** is $F^{-1}_{\text{male}}(84.8\%) = 184\text{cm}$.



Counterfactual for $\hat{Y}$ of a 170cm **female** had she been a **male**?

1. $S \leftarrow$ male,
2. Calculate the counterfactual for height, $\text{height}^{\star}_{S \leftarrow \text{male}}$, as

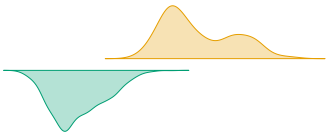
$$T^{\star}(170) = (F^{-1}_{\text{male}} \circ F_{\text{female}})(170)$$
$$= 184\text{cm} \ .$$



We obtain the counterfactual for $\hat{Y}$, $\hat{y}^{\star}_{S \leftarrow \text{male}} = m(\text{male}, 184)$.

Optimal Transport and Monge Mapping

- **Optimal Transport** (OT): how to find the best way to transport mass from **one distribution** to **another** while minimizing a given cost.
- 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." (Villani, 2003, 2008)



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

Optimal Transport Map

Proposition

If $\mathcal{X}_0 = \mathcal{X}_1$ is a compact subset of $\mathbb{R}^d$ and μ_0 is atomless, then there exists T such that $\mu_1 = T_{\#}\mu_0$ (push-forward measure).

Definition: Monge problem, (Monge, 1781)

If we further assume μ_0 and μ_1 are absolutely continuous w.r.t. Lebesgue measure, then we can find an "optimal" mapping, satisfying

$$\inf_{T_{\#}\mu_0 = \mu_1} \int_{\mathcal{X}_0} c\big(x_0, T(x0)\big)\, d\mu_0(x_0),$$

for a general cost function $c : \mathcal{X}_0 \times \mathcal{X}_1 \to \mathbb{R}^+$.

Closed-Form Transport Map for a Univariate Distribution

OT map for continuous univariate distributions (Santambrogio, 2015)

The optimal Monge map T^* for some strictly convex cost c, such that $T_{\#}\mu_0 = \mu_1$, is given by

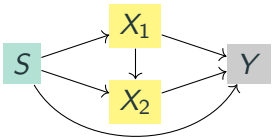
$$T^* := {F_1}^{-1} \circ F_0 \ ,$$

where F_0 and F_1 are the cumulative distribution functions associated with μ_0 and μ_1, respectively.

Sequential Transport for Evaluating Counterfactual Fairness

Approach Summary

Let's go back to our toy example.



To compute the counterfactual value $\hat{y}^{\star}_{i,S\leftarrow\text{male}}$ for the i-th individual (female, x_1, x_2), we need to transport the covariates $\boldsymbol{X} = (X_1, X_2)$, since both are descendants of S.

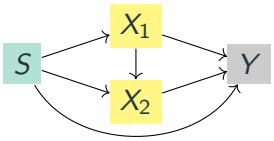
Existing approaches:

1. Plečko and Meinshausen (2020) uses a structural causal model framework,
2. De Lara et al. (2024) uses multivariate OT without a closed-form solution.

We link these methods to derive counterfactuals within the algorithmic fairness framework by **applying sequential transport on a presumed causal graph** (Cheridito and Eckstein, 2023), extending Knothe's rearrangement from OT (Carlier et al., 2008).

Applying Sequential Transport

1. Assumed DAG,
2. Joint distribution:
   $\mathbb{P}[s, x_1, x_2, y] = \mathbb{P}[s]\mathbb{P}[x_1|s]\mathbb{P}[x_2|s, x_1]\mathbb{P}[y|s, x_1, x_2]$,
3. Topological ordering : $S \to X_1 \to X_2 \to Y$.



Sequential transport map:

$$T_{\underline{st}}(\text{female}, x_1, x_2) = \begin{pmatrix} \text{male} \\ T^\star_{\underline{1}}(x_1|S = \text{female}) \\ T^\star_{\underline{2}|1}(x_2|x_1, S = \text{female}) \end{pmatrix} = \begin{pmatrix} \text{male} \\ {F_{X_1,\text{male}}}^{-1}\Big(F_{X_1,\text{female}}(x_1)\Big) \\ {F_{X_2|X_1,\text{male}}}^{-1}\Big(F_{X_2|X_1,\text{female}}(x_2|x_1)\big|x_1^\star\Big) \end{pmatrix}$$

Counterfactual prediction, had she been **male**: $\hat{y}^\star_{i,S\leftarrow\text{male}} = m(\text{male}, x_1^\star, x_2^\star)$.

Counterfactual fairness $|\hat{y}^\star_{i,S\leftarrow\text{male}} - 18.24\%| \neq 0$?

$$\begin{aligned}
\text{Observation:} \quad & (\text{female}, x_1, x_2) \\
\text{Prediction:} \quad & m(\text{female}, x_1, x_2) = 18.24\% \\
\text{Pred. with Seq. T:} \quad & m(\text{male}, x_1^\star, x_2^\star) = 61.40\%
\end{aligned}$$

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

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

Conclusion

- For more details on the theoretical foundations of the approach, see Fernandes Machado et al. (2025b). For implementation details, visit our website: `https://fer-agathe.github.io/sequential_transport`.
- So far, to derive **counterfactuals**, we have discussed a quantile-based interpretation when the characteristics to be transported, $\boldsymbol{X}$, are continuous. But how can we handle **categorical data**? *"What would have been this woman's **marital status**, had she been a man?"* (Fernandes Machado et al., 2025a).
- The next step is to develop a **post-processing mitigation method** to correct unfairness, when detected through counterfactual fairness, **for all individuals in the disadvantaged group**.

Appendix

References

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

D. Barber. *Bayesian reasoning and machine learning*. Cambridge University Press, 2012.

S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. Adaptive Computation and Machine Learning series. MIT Press, 2023. ISBN 9780262376525. URL `https://books.google.ca/books?id=HuGwEAAAQBAJ`.

D. Bertsekas and J. Tsitsiklis. *Introduction to Probability*. Athena Scientific optimization and computation series. Athena Scientific, 2008. ISBN 9781886529236.

Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

References ii

G. Carlier, A. Galichon, and F. Santambrogio. From knothe's transport to brenier's map and a continuation method for optimal transport, 2008. URL https://arxiv.org/abs/0810.4153.

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

P. Cheridito and S. Eckstein. Optimal transport and Wasserstein distances for causal models, 2023.

L. De Lara, A. González-Sanz, N. Asher, L. Risser, and J.-M. Loubes. Transport-based counterfactual models. *Journal of Machine Learning Research*, 25(136):1–59, 2024.

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012. doi: 10.1145/2090236.2090255.

European Union AI Act. Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence (artificial intelligence act). *Official Journal of the European Union*, 2024. URL https://artificialintelligenceact.eu/article/5/. Article 5 – Prohibited AI Practices.

References iii

A. Feller, E. Pierson, S. Corbett-Davies, and S. Goel. A computer program used for bail and sentencing decisions was labeled biased against blacks. it's actually not that clear., Oct. 2016. URL https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/a-computer-program-used-for-bail-and-sentencing-decisions-was-labeled-biased-against-blacks-its-

A. Fernandes Machado, A. Charpentier, and E. Gallic. Optimal transport on categorical data for counterfactuals using compositional data and dirichlet transport. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*. International Joint Conferences on Artificial Intelligence Organization, 2025a. URL https://arxiv.org/abs/2501.15549. Main Track.

A. F. Fernandes Machado, A. Charpentier, and E. Gallic. Sequential conditional transport on probabilistic graphs for interpretable counterfactual fairness. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(18):19358–19366, Apr. 2025b. doi: 10.1609/aaai.v39i18.34131. URL https://ojs.aaai.org/index.php/AAAI/article/view/34131.

N. J. Higham. *Functions of matrices: theory and computation*. SIAM, 2008.

L. V. Kantorovich. On the translocation of masses. In *Doklady Akademii Nauk USSR*, volume 37, pages 199–201, 1942.

References iv

R. Koenker and K. F. Hallock. Quantile regression. *Journal of Economic Perspectives*, 15(4):143–156, 2001. doi: 10.1257/jep.15.4.143.

D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

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

J. Larson, S. Mattu, L. Kirchner, and J. Angwin. How we analyzed the compas recidivism algorithm. *ProPublica*, May 2016. URL `https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm`.

N. Meinshausen. Quantile regression forests. *J. Mach. Learn. Res.*, 7:983 – 999, 2006. ISSN 1532-4435.

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

New York State Department of Financial Services. Insurance Circular Letter No. 7: Use of Artificial Intelligence Systems and External Consumer Data and Information Sources in Insurance Underwriting and Pricing, July 2024. URL `https://www.dfs.ny.gov/industry-guidance/circular-letters/cl2024-07`.

References v

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

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

D. Plečko, N. Bennett, and N. Meinshausen. fairadapt: Causal reasoning for fair data preprocessing. *Journal of Statistical Software*, 110(4):1 – 35, 2024. doi: 10.18637/jss.v110.i04. URL https://www.jstatsoft.org/index.php/jss/article/view/v110i04.

M. Rosenblatt. Remarks on a multivariate transformation. *The annals of mathematical statistics*, 23(3): 470–472, 1952.

D. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66, 10 1974. doi: 10.1037/h0037350.

F. Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Progress in Nonlinear Differential Equations and Their Applications. Springer International Publishing, 2015. ISBN 9783319208282. URL https://books.google.ca/books?id=UOHHCgAAQBAJ.

M. Susser. What is a cause and how do we know one? a grammar for pragmatic epidemiology. *American Journal of Epidemiology*, 133(7):635–648, 1991.

G. Upton and I. Cook. *A Dictionary of Statistics 3e*. Oxford Paperback Reference. OUP Oxford, 2014. ISBN 9780199679188. URL `https://books.google.ca/books?id=4WygAwAAQBAJ`.

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

C. Villani. *Optimal transport – Old and new*, volume 338, pages xxii+973. 01 2008. doi: 10.1007/978-3-540-71050-9.

Wasserstein. Markov processes over denumerable products of spaces, describing large systems of automata. *Probl. Peredachi Inf.*, 5, 1969.

Optimal map for continuous multivariate distributions (Brenier, 1991)

With a quadratic cost, the optimal Monge map T^* is unique, and it is the gradient of a convex function, $T^* = \nabla\varphi$.

Unfortunately, it is generally difficult to give an analytic expression for the optimal mapping T^*, unless additional assumptions are made, such as assuming that both distributions are Gaussian.

Optimal Transport and Monge mapping (1/2)

In probability theory, the mass transportation problem involves constructing a joint distribution, known as a coupling, between two marginal probability measures (Villani, 2003, 2008).

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."

Definition

Suppose $T : \mathcal{X}_0 \to \mathcal{X}_1$. The push-forward of μ_0 by T is the measure $\mu_1 = T_{\#}\mu_0$ on $\mathcal{X}_1$ s.t. $\forall B \subset \mathcal{X}_1, \quad T_{\#}\mu_0(B) = \mu_0(T^{-1}(B))$.

Proposition

For all measurable and bounded $\varphi : \mathcal{X}_1 \to \mathbb{R}$,

$$\int_{\mathcal{X}_1} \varphi(x_1)\, dT_{\#}\mu_0(x_1) = \int_{\mathcal{X}_0} \varphi(T(x_0))\, d\mu_0(x_0) \ .$$

Optimal Transport and Monge mapping (2/2)

Proposition

If $\mathcal{X}_0 = \mathcal{X}_1$ is a compact subset of $\mathbb{R}^d$ and μ_0 is atomless, then there exists T such that $\mu_1 = T_{\#}\mu_0$.

Definition: Monge problem (Monge, 1781)

If we further assume μ_0 and μ_1 are absolutely continuous w.r.t. Lebesgue measure, then we can find an "optimal" mapping, satisfying

$$\inf_{T_{\#}\mu_0 = \mu_1} \int_{\mathcal{X}_0} c\big(x_0, T(x_0)\big) d\mu_0(x_0),$$

for a general cost function $c : \mathcal{X}_0 \times \mathcal{X}_1 \to \mathbb{R}^+$.

The optimal mapping is denoted T^*.

Optimal Transport plans

In general settings, however, such a deterministic mapping T^* between probability distributions may not exist.

Kantorovich relaxation (Kantorovich, 1942)

The Kantorovich relaxation of Monge mapping is defined as

$$\inf_{\pi \in \Pi(\mu_0, \mu_1)} \int_{\mathcal{X}_0 \times \mathcal{X}_1} c(\boldsymbol{x}_0, \boldsymbol{x}_1) \pi(\mathrm{d}\boldsymbol{x}_0, \mathrm{d}\boldsymbol{x}_1),$$

for a general cost function $c : \mathcal{X}_0 \times \mathcal{X}_1 \to \mathbb{R}^+$ and $\Pi(\mu_0, \mu_1)$ the set of all couplings of μ_0 and μ_1.

This problem always admits solutions and focuses on couplings rather than deterministic mappings.

Wasserstein distance (Wasserstein, 1969)

Consider two measures μ_0 and μ_1 on $\mathbb{R}^d$, with a norm $\|\cdot\|$ on $\mathbb{R}^d$. Then define with $p \geq 1$

$$W_p(\mu_0, \mu_1) = \left(\inf_{\pi \in \Pi(\mu_0, \mu_1)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x_0 - x_1\|^p \, d\pi(x_0, x_1) \right)^{1/p},$$

where $\Pi(\mu_0, \mu_1)$ is the set of all couplings of μ_0 and μ_1.

The Wasserstein distance corresponds to the minimum value of Kantorovich relaxation formulation of Optimal Transport problem with a norm $\|\cdot\|$ as cost function c.

Monge/Kantorovich formulations

Proposition

If $\mathcal{X}_0 = \mathcal{X}_1$ is a compact subset of $\mathbb{R}^d$ and μ_0 is atomless,

$$\min \{\text{Monge problem}\} = \min \{\text{Kantorovich relaxation}\} \quad .$$

Conditional transport (1/3)

Let denote $\mu_{0:d}$ denote the marginal d-th measure, $\mu_{0:d-1|d}$ the conditional $d-1$-th measure given x_d, $\mu_{0:d-2|d-1,d}$ the conditional $d-2$-th measure given x_{d-1} and x_d, etc. And, let $T_d^\star$ denote the univariate optimal transport map from $\mu_{0:d}$ to $\mu_{1:d}$, $T_{\overline{d-1}}^\star(\cdot|x_d)$ denote the monotone nondecreasing map transporting from $\mu_{0:d-1|d}(\cdot|x_d)$ to $\mu_{1:d-1|d}(\cdot|T_d^\star(x_d))$, etc.

Conditional transport (2/3)

The Knothe-Rosenblatt rearrangement is directly inspired by the Rosenblatt chain rule, from Rosenblatt (1952).

"Monotone lower triangular map" from Knothe-Rosenblatt rearrangement (Santambrogio, 2015)

If measure μ_0 is absolutely continuous on $\mathbb{R}^d$, then $T_{\underline{kr}}$ is a transportation map from μ_0 to μ_1

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

Conditional transport (3/3)

Mapping on an acyclical causal graph $\mathcal{G}$ (Cheridito and Eckstein, 2023; Fernandes Machado et al., 2025b)

If measure μ_0 is absolutely continuous on $\mathbb{R}^d$, then $T_{\overline{ST}}$ is a transportation map from μ_0 to μ_1

$$T^\star_{\mathcal{G}}(x_1, \cdots, x_d) = \begin{pmatrix} T^\star_1(x_1) \\ T^\star_2(x_2|\ parents(x_2)) \\ \vdots \\ T^\star_{d-1}(x_{d-1}|\ parents(x_{d-1})) \\ T^\star_d(x_d|\ parents(x_d)) \end{pmatrix}.$$

This mapping will be called "sequential conditional transport on the graph $\mathcal{G}$."

Gaussian transport (1/3)

Univariate Optimal Gaussian Transport

The optimal mapping, from a $\mathcal{N}(\mu_0, \sigma_0^2)$ to a $\mathcal{N}(\mu_1, \sigma_1^2)$ distribution is (linear)

$$x_1 = T^\star(x_0) = \mu_1 + \frac{\sigma_1}{\sigma_0}(x_0 - \mu_0),$$

which is a nondecreasing linear transformation.

Multivariate Optimal Gaussian Transport

If $\boldsymbol{X}_0 \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and $\boldsymbol{X}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, the optimal mapping is (linear)

$$\boldsymbol{x}_1 = T^\star(\boldsymbol{x}_0) = \boldsymbol{\mu}_1 + \boldsymbol{A}(\boldsymbol{x}_0 - \boldsymbol{\mu}_0),$$

where $\boldsymbol{A}$ is a symmetric positive matrix that satisfies $\boldsymbol{A}\boldsymbol{\Sigma}_0\boldsymbol{A} = \boldsymbol{\Sigma}_1$, which has a unique solution given by $\boldsymbol{A} = \boldsymbol{\Sigma}_0^{-1/2}\left(\boldsymbol{\Sigma}_0^{1/2}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_0^{1/2}\right)^{1/2}\boldsymbol{\Sigma}_0^{-1/2}$, where $\boldsymbol{M}^{1/2}$ is the square root of the square (symmetric) positive matrix $\boldsymbol{M}$ based on the Schur decomposition ($\boldsymbol{M}^{1/2}$ is a positive symmetric matrix), as described in Higham (2008).

Gaussian transport (2/3)

Details on Conditional Transport (Cholesky decomposition) and Sequential Transport (based on a DAG $\mathcal{G}$) for Gaussian distribution in Appendix of Fernandes Machado et al. (2025b).

The idea is that if $\boldsymbol{X} = (X_1, X_2) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\rho \in [0, 1]$, thanks to the properties of Gaussian vectors, we have: (Bertsekas and Tsitsiklis, 2008)

$X_2|X_1 = x_1 \sim \mathcal{N}\left(\mu_2 + \rho\sigma_2 \frac{x_1 - \mu_1}{\sigma_1}, (1-\rho^2)\sigma_2^2\right)$ and

$X_1|X_2 = x_2 \sim \mathcal{N}\left(\mu_1 + \rho\sigma_1 \frac{x_2 - \mu_2}{\sigma_2}, (1-\rho^2)\sigma_1^2\right)$.

Therefore we can apply univariate optimal transport map sequentially to X_1 then $X_2|X_1$, or to X_2 then $X_1|X_2$.

Gaussian transport (3/3)

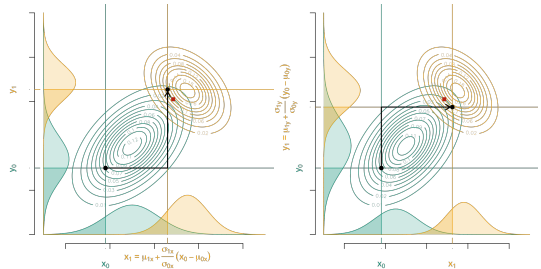


Figure 1: Two Gaussian conditional optimal transports. On the left-hand side, the process begins with a univariate transport along the x axis (using $T_x^\star$), followed by a transport along the y axis on the conditional distributions (using $T_{y|x}^\star$), corresponding to the "lower triangular affine mapping." On the right-hand side, the sequence is reversed: it starts with a univariate transport along the y axis (using $T_y^\star$) followed by transport along the x axis on the conditional distributions (using $T_{x|y}^\star$). The red square is the multivariate OT of the point in the bottom left, corresponding to the "upper triangular affine mapping."

Ladder of Causation

According to Pearl (2009), there are three levels of reasoning about causality, representing an ascending hierarchy of increasingly complex questions.

1. **Association**: *What is happening?* Observing correlations and statistical dependencies.
2. **Intervention**: *What is the effect of implementing a new law on traffic accidents?* We take an action (an intervention) in the future, that does not contradict the actual situation, to measure the impact on the outcome. It is related to randomized/controlled experiments (we control T assignment), to evaluate impact of policies at the population level (we consider the entire distribution of exogeneous variables $\boldsymbol{U}$ and not a given unit/indiv.).
3. **Counterfactual**: *What would have happened if I had take a different course of action in a specific situation given that I know what actually happened?* We draw alternate worlds for an individual (represented by the values of $\boldsymbol{U}$) that contradict the observed one.

It is noteworthy that framework used in Rubin (1974) does not differentiate between levels 2 and 3.

Probabilistic Graphical Models

Following standard notations in probabilistic graphical models (see Koller and Friedman (2009) or Barber (2012)), given a random vector $\boldsymbol{X} = (X_1, \cdots, X_d)$, consider a directed acyclic graph (DAG) $\mathcal{G} = (V, E)$, where $V = \{x_1, x_2, \cdots, x_d\}$ are the vertices (corresponding to each variable), and E are directed edges, such that $x_i \to x_j$ means "variable x_i causes variable x_j," in the sense of Susser (1991).

Global Markov property

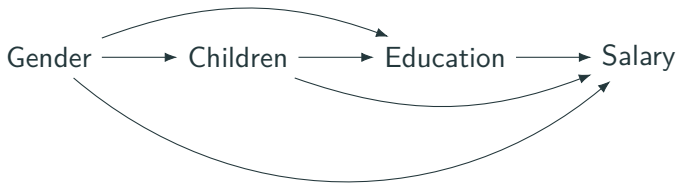
The joint distribution of $\boldsymbol{X}$ satisfies the (global) Markov property w.r.t. $\mathcal{G}$:

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

where parents(x_i) are nodes with edges directed towards x_i, in $\mathcal{G}$.

Topological order

As discussed in Ahuja et al. (1993), a DAG imposes some ordering on variables. In this "topological sorting," a vertex must be selected before its adjacent vertices, which is feasible because each edge is directed such that no cycle exists in the graph. Here is an example of topological order of a DAG:



Defining counterfactuals

Consider the non-linear strucural model associated with a directed acyclic graph (DAG) $\mathcal{G}$ and independent errors $\boldsymbol{U}$, supposed to be uniform on [0,1].

$$\begin{cases} S = h_s(U_S), \text{ the gender,} \\ X = h_x(S, U_X), \text{ the education level,} \\ Y = h_y(S, X, U_Y), \text{ the salary outcome,} \end{cases}$$

Following path from S to Y in $\mathcal{G}$,

$$\begin{cases} S = \text{male}, \\ X^*(S = \text{male}) = h_x(\text{male}, U_X), \\ Y^*(S = \text{male}) = h_y(\text{male}, X^*, U_Y). \end{cases}$$

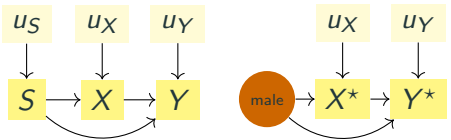


Figure 2: Individual with observed $S =$ female and errors $\boldsymbol{u} = (u_S, u_X, u_Y)$ on $\mathcal{G}$ and its "twin projection" in the **counterfactual world** (Pearl, 2009).

Quantile regression

$Q_\alpha(x) = \inf\{y : F(y|\boldsymbol{X} = \boldsymbol{x}) \geq \alpha\}$

Quantile loss function $L_\alpha(y, q) = \begin{cases} \alpha|y - q| & \text{si } y > q \\ (1-\alpha)|y - q| & \text{si } y \leq q \end{cases}$

$\hat{Q}_\alpha(\boldsymbol{x}) = \arg\min_q \mathbb{E}[L_\alpha(y, q)|\boldsymbol{X} = \boldsymbol{x}]$

Linear quantile regression (Koenker and Hallock, 2001):
$\hat{Q}_\alpha(\boldsymbol{x}) = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}_\alpha$ with $\hat{\boldsymbol{\beta}}_\alpha = \arg\min_\beta \sum_{i=1}^n L_\alpha(y_i, \boldsymbol{x}_i^T\boldsymbol{\beta})$

Quantile regression forests (Meinshausen, 2006):
$\hat{F}(y|\boldsymbol{X} = \boldsymbol{x}_i) = \sum_{i=1}^n w_i(\boldsymbol{x}_i)1_{y_i \leq y}$ with $w_i(\boldsymbol{x}_i) = \frac{1}{M}\sum_{m=1}^M w_i(\boldsymbol{x}_i, \theta_m)$, M is the number of trees, θ_m denotes parameters of tree m and $w_i(\boldsymbol{x}_i, \theta_m) = \frac{1_{\boldsymbol{x}_i \in R_m}}{\#\{\boldsymbol{x} \in R_m\}}$ with R_m denoting a leaf in tree m.

Counterfactual Fairness Metric

Definition (Kusner et al., 2017)

A predictor $\hat{Y}$ is counterfactually fair if under any context $S = s$ and $\boldsymbol{X} = \boldsymbol{x}$,

$$\hat{Y}(S = s)|\boldsymbol{X} = \boldsymbol{x}, S = s \quad = \quad \hat{Y}(S = s')|\boldsymbol{X} = \boldsymbol{x}, S = s$$

for any value $s' \in \mathcal{S}$.

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

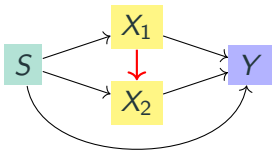


Figure 3: Assumed causal structure.

Counterfactual assuming X_2 is caused by X_1

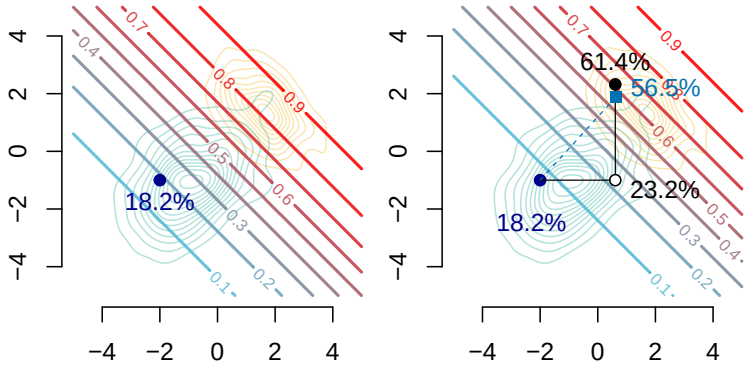


Figure 4: Predictions by m of: the **observation** using factual (left), counterfactual (right): **counterfactual by Seq. T.** (assuming $X_1 \to X_2$) and **optimal. transport**.

Decomposition of the *mutatis mutandis* difference

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

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