Assessing Counterfactual Fairness via (Marginally) Optimal Transport

Ewen Gallic, Arthur Charpentier, and Agathe Fernandes Machado

ENSAI — Séminaire Economie
November 28, 2025

amU Aix Marseille Université
amse école d'économie d'aix-marseille aix-marseille school of economics

Broad Framework

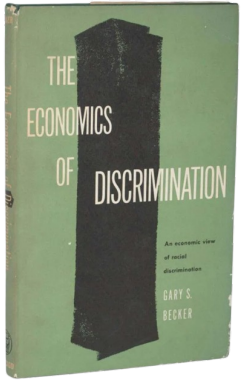
- We want to make **predictions** on an outcome variable (e.g., graduation probability, loan default risk, recidivism, claim frequency).
- To do so, we use a **statistical model**, or a **machine learning model** fed with **historical data**.
- To comply with regulations, we want to obtain a model that **does not discriminate** with respect to a **protected/sensitive attribute**.

Motivations: Regulation of Protected/Sensitive Attributes

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

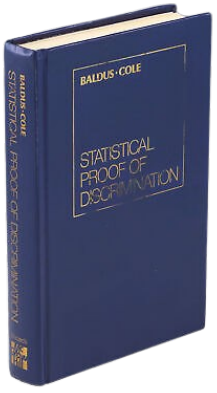
"Any discrimination based on any ground such as ***sex***, ***race***, ***colour***, ***ethnic or social origin***, ***genetic features***, ***language***, ***religion or belief***, ***political or any other opinion***, ***membership of a national minority***, ***property***, ***birth***, ***disability***,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

What is Discrimination? An Economic Perspective



- In economics, following Becker (1957), **discrimination**: situations in which individuals are treated differently based on attributes such as race, gender, etc., rather than their productivity or other relevant characteristics.
  - **Disparate treatment** (or taste-based discrimination): intentional discrimination, where individuals are treated differently explicitly because of a **protected characteristic**.
  - **Disparate impact**: policy, practice, or decision that appears neutral on the surface disproportionately affects members of a **protected group**, even without intentional discrimination.
- From a Law perspective: **direct** vs. **indirect** discrimination (Campbell and Smith, 2023)

What is Discrimination? A Statistical Perspective



- **Statistical discrimination** (see, e.g., Baldus and Cole, 1980): individuals are treated differently based on group-level statistical averages, rather than their individual characteristics. They do not arise from prejudice or bias but from **decision-makers relying on imperfect information** and using group membership as a **proxy** for individual traits.
- Some forms of discrimination are considered unacceptable (Hellman, 2008).
- Fisher (1936): separating or classifying observations into distinct groups based on measured characteristics. In this context, discrimination is purely a statistical operation with no connotation of social bias or inequality.
- However, statistical discrimination may lead to:
  - **Reinforcement of Biases** (through lack of opportunities).
  - **Legal and Ethical Concerns**.

Toy Example: Graduation Likelihood (1/2)

Assume we want to predict **the probability of graduating from university** using a logistic regression model with **three predictors**, including a **sensitive** one.

We assume that the graduation outcome $Y_i \in \{0, 1\}$ follows a Bernoulli distribution with conditional mean:

$$\mathbb{E}(Y_i \mid \boldsymbol{X}_i) = \Pr(Y_i = 1 \mid \boldsymbol{X}_i) = \frac{\exp(\boldsymbol{X}_i\beta)}{1 + \exp(\boldsymbol{X}_i\beta)}.$$

The predictor vector $\boldsymbol{X}$ contains:

- the student's **high-school grade**;
- the student's **university admission test score**;
- a binary variable indicating the student's **race** (1 = white, 0 = Black).

Toy Example: Graduation Likelihood (2/2)

The predicted *odds* of graduating are:

$$\begin{cases} \widehat{\text{odds}}(\text{white}) = \exp\left[\hat{\beta}_0 + \hat{\beta}_1 \,\text{HS grade} + \hat{\beta}_2 \,\text{admission score} + \hat{\beta}_3\right], \\ \widehat{\text{odds}}(\text{Black}) = \exp\left[\hat{\beta}_0 + \hat{\beta}_1 \,\text{HS grade} + \hat{\beta}_2 \,\text{admission score}\right]. \end{cases}$$

Hence:

$$\widehat{\text{odds}}(\text{white}) = \exp\left[\hat{\beta}_0 + \hat{\beta}_1 \mathbf{1}_{\text{HS grade}} + \hat{\beta}_2 \,\text{adm. score} + \hat{\beta}_3 \,\mathbf{1}_{\text{man}}\right] = \widehat{\text{odds}}(\text{Black}) \cdot \exp[\beta_3]$$

$\times e^{\beta_3}$ ceteris paribus

If $\hat{\beta}_3 = 0.2$, then $e^{0.2} \approx 1.22$:

- the odds of graduating are about **22% higher** for white students, *ceteris paribus*.

From Prediction to Discrimination Measurement

- Our model suggests that race is associated with different graduation odds.
  - But this difference does not tell us whether it reflects **legitimate academic factors** or **unfair discrimination**.
- With such insight from the data, should a University trying to screen applicants **discriminate** by race?
- In other words, does it make **statistical** sense to discriminate?
  - discrimination w.r.t. a **sensitive attribute**.

Fair Discrimination (in Insurance): an Oxymoron

Avraham (2017)

"what is unique about insurance is that even statistical discrimination (the act by which an insurer uses a characteristic of an insured or potential insured as a statistic for the risk it poses to an insurer), which by definition is absent any malicious intentions, poses significant moral and legal challenges. Why? Because ***on the one hand, policy makers*** *would like insurers to* ***treat their insureds equally, without discriminating*** *based on* ***race, gender, age, or other characteristics****,* ***even if it makes statistical sense*** *to discriminate.* [...] ***On the other hand****, at the core of* ***insurance business lies discrimination between risky and non-risky insureds****.* ***But riskiness often statistically correlates with the same characteristics policy makers would like to prohibit insurers from taking into account****."*

Individual Characteristics

- In our example, **race** may be a statistical predictor, but from the European legislation perspective using it leads to a **direct discrimination**.
- Here, **race** is not a **causal predictor**. It does not reflect **individual behaviour**.
- In the era of **big data** and **artificial intelligence**, a naive solution consists in **hiding the sensitive attribute**, and use a **machine learning model** trained on additional (hopefully behavioural) data:
  - explicability issues
  - proxy discrimination issues (Pedreshi et al., 2008; Dwork et al., 2012).

Motivations: Regulation Regarding Discrimination in Predictive Models

European Union AI Act (2024)

*"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"*

Example: Recidivism Risk Assessment Tool

- On the one hand: "*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)
- On the other hand: "*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: Wrap-up

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

What is Algorithmic Fairness?

- Let $m : \mathcal{X} \to \mathcal{Y}$ be a predictive model that predicts an outcome Y (e.g., claims) w.r.t. a **sensitive attribute** $S \in \mathcal{S}$ (e.g., gender, race) using features $\boldsymbol{X}$.
- Regulations may prohibit **discrimination** on the sensitive attribute, requiring m to be fair w.r.t. to S.
- **Approaches** to **evaluate** and, if necessary, **mitigate** the unfairness of model predictions $\hat{Y} = m(\boldsymbol{X})$ for S:
  - **Group fairness**: compare $\hat{Y}$ between groups defined by S, e.g., graduation for Black students vs. graduation for white students (Barocas et al., 2023; Hardt et al., 2016).
  - **Individual fairness**: focus on a specific individual in the disadvantaged group. "*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)

Our Approach

Our framework determines, given a trained model and the data used to train it, whether the model discriminates with respect to a **sensitive attribute**.

1. Assume a **causal structure**.
2. Construct **counterfactual individuals**.
   - **Intervene** on the protected attribute (e.g., set race = white),
   - **propagate changes** through the structural causal model to generate counterfactual versions of each individual (using **Optimal Transport**).
3. **Compare model predictions** with the original individual and its counterfactual.

Since we follow **causal pathways**, our methodology offers **interpretable individual-level explanation**.

Road Map

1. Introduction
2. Causal Inference Framework
3. Quantifying Counterfactual Fairness
4. Sequential Transport for Evaluating Counterfactual Fairness
5. Counterfactuals for Categorical Data
6. Conclusion

2. Causal Inference Framework

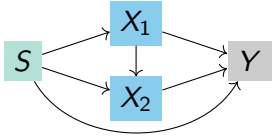
Probabilistic Graphical Models

- 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**:

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

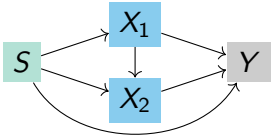
where $\text{parents}(X_i)$ are the immediate causes of X_i.

Example: Causal Graph (1/2)



- $S \in \{\textbf{white}, \textbf{Black}\}$ denotes the **sensitive attribute**: eg., race,
- X_1 is a "non-protected" explanatory variable: e.g., **high school grades**,
- X_2 is another "non-protected" explanatory variable: e.g., the **university admission test score**,
- Y is the outcome variable: e.g., the likelihood of graduating from university, 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,

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

- Using the Markov property, the joint distribution factorizes as one conditional term per node, conditioned only on its parents.

3. 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 $(S_i = \text{Black}, X_{1,i} = x_1, X_{2,i} = x_2)$, with $\hat{y}_i = m(\text{Black}, x_1, x_2) = 18.24\%$.

1. **Defining a counterfactual**

   "What would this student's chances of graduating have been if they had been **white**?", i.e., how to define $\hat{y}^{\star}_{i,S \leftarrow \text{white}}$?

2. **Assessing Counterfactual Fairness**

   "Would this student's chances of graduating have been the same if they had been **white**?", i.e., do we have $|\hat{y}^{\star}_{i,S \leftarrow \text{white}} - \hat{y}^{\star}_{i,S \leftarrow \text{Black}}| = |\hat{y}^{\star}_{i,S \leftarrow \text{white}} - \hat{y}_i| = 0$?

Counterfactual Fairness

1. **Affirmative actions**: "*The contractor will not discriminate against any employee or applicant for employment because of race, creed, color, or national origin. The contractor will **take affirmative action** to ensure that applicants are employed, and that employees are treated during employment, without regard to their race, creed, color, or national origin.*" (John F. Kennedy, EO #10925, March 6, 1961)
   "*In order to get beyond racism, we must first take account of race. There is no other way. And **in order to treat some persons equally, we must treat them differently**.*" (Justice Harry Blackmun, Regents of Univ. of Cal. v. Bakke, 438 U.S. 265, 407, via Scalia (1979))
2. **Blindness**: "*The way to stop discrimination on the basis of race is to **stop discriminating on the basis of race**.*" (Chief Justice John G. Roberts, Jr, Parents Involved in Community Schools v. Seattle School District No. 1, via Turner (2015))

Defining a Counterfactual

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

- ***Ceteris paribus***: ignoring causal relationships and simply computing $m(\text{white}, x_1, x_2)$, 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 simpler model m predicting Y based on **sex** and **height** and assume the following causal graph:

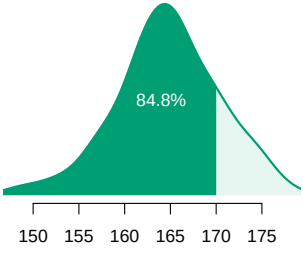
What is the counterfactual for $\hat{Y}$ *of a* ***female*** *with height 170cm had she been a* ***male****?*

- If we use the *ceteris paribus* approach, we would simply compute $\hat{Y}^{\star}_{S \leftarrow \text{male}}$ as $m(\text{male}, 170\text{cm})$.
- $\rightarrow$ completely ignores the fact that sex "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 sex
  - i.e., calculate the **counterfactual for height** first.

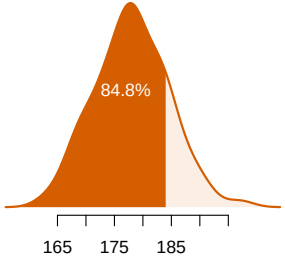
Mutatis Mutandis: Intuitive Example (2/3)

*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 $\alpha = 84.8\%$, i.e., $F_{\text{female}}(170) = 84.8\%$.



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



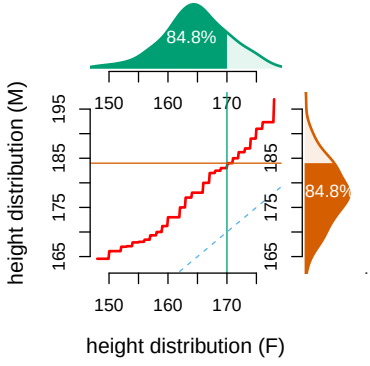
Mutatis Mutandis: Intuitive Example (3/3)

*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

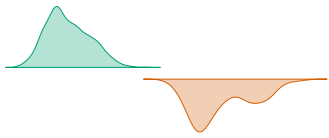
$$T^{\star}(170) = (F^{-1}_{\text{male}} \circ F_{\text{female}})(170)$$
$$= 184\text{cm}.$$

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, 2009)



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)

We want to find an "optimal" mapping, satisfying

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

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

Optimal Transport Map for Univariate Distribution

OT map for continuous univariate distributions (Santambrogio, 2015)

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

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

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

In the **multivariate case**, it is generally difficult to obtain an analytic expression for $T^\star$, except in the Gaussian case.

Optimal Transport Coupling

In the **general setting**, such a deterministic mapping may not exist, in particular if μ_0 and μ_1 are not continuous w.r.t. the Lebesgue measure.

Focusing on **couplings** rather than deterministing mappings leads to Kantorovich (1942) problem, which always admits a solution:

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

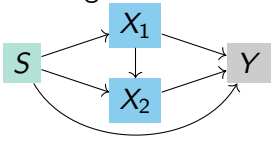
where $\Pi(\mu_0, \mu_1)$ is the set of all couplings of μ_0 and μ_1.

- It involves constructing a **joint distribution** (coupling) between two marginal probability measures (Villani, 2003, 2009).
- In practice, the optimization problem is solved using **numerical algorithms**.

4. Sequential Transport for Evaluating Counterfactual Fairness

The Approach in a Nutshell

Let us go back to our toy example of the likelihood to graduate from University.



To compute the counterfactual value $\hat{y}^{\star}_{i,S\leftarrow\text{white}}$ for the i-th individual (Black, 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 assuming a causal graph.

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

Applying Multivariate Optimal Transport

With **Multivariate OT**: "*There is no quantitative rule for this choice* [of transport coupling]*; it is guided by intuition and feasibility reasons.*" (De Lara et al., 2024)

The **Multivariate OT plan**:

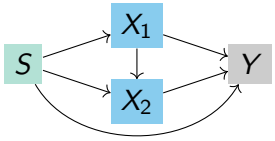
$$\pi_{\underline{ot}}(\text{Black}, x_1, x_2) = \begin{pmatrix} \text{white} \\ \pi^\star(x_1, x_2 | S = \text{Black}) \end{pmatrix}$$

The **Counterfactual prediction**, had they been **white**: $\hat{y}^{\star,\underline{ot}}_{i,S\leftarrow\text{white}} = m\left(\text{white}, \pi^\star(\boldsymbol{x})\right)$.

Counterfactual fairness $|\hat{y}^{\star,\underline{ot}}_{i,S\leftarrow\text{white}} - 18.24\%| \neq 0$?

Applying Sequential Transport (Our Approach)

1. Assumed DAG,
2. Topological ordering : $S \to X_1 \to X_2 \to Y$.



Sequential Transport map (Fernandes Machado et al., 2025a):

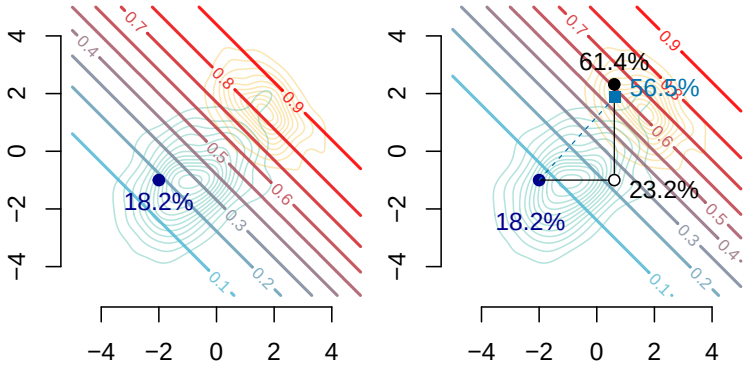
$$T_{\underline{st}}(\text{Black}, x_1, x_2) = \begin{pmatrix} \text{white} \\ T_{\underline{1}}^{\star}(x_1 | S = \text{Black}) \\ T_{\underline{2}|1}^{\star}(x_2 | x_1, S = \text{Black}) \end{pmatrix} = \begin{pmatrix} \text{white} \\ {F_{X_1,\text{white}}}^{-1}\Big(F_{X_1,\text{Black}}(x_1)\Big) \\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

Counterfactual prediction: $\hat{y}^{\star}_{i,S\leftarrow\text{white}} = m\Big(\text{white}, T_{\underline{1}}^{\star}(x_1), T_{\underline{2}|1}^{\star}(x_2|x_1)\Big)$.

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

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 $X_1 \to X_2$) and* ***Optimal Transport***



Interpretable Counterfactual Fairness

$$
\begin{aligned}
\text{Observation:} \quad & (\text{Black}, x_1, x_2) \\
\text{Prediction:} \quad & m(\text{Black}, x_1, x_2) && = 18.24\% \\
\text{Pred. } \textit{cet. par.} \quad & m(\text{white}, x_1, x_2) && = 7.58\% \\
\text{Pred. with Seq. T:} \quad & m(\text{white}, x_1^\star, x_2^\star) && = 61.40\%
\end{aligned}
$$

With the **ST map**, the ***mutatis mutandis* difference can be decomposed** :

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

5. Counterfactuals for Categorical Data

What About Transporting Categorical Data?

What would have been the ***marital status*** *of this woman, had she been a man?*

- Classical OT methods for continuous X are not directly applicable:
  - **categorical features lack a canonical distance**.
  - OT requires a cost.
- We might be tempted to use **random matching**, but it is unstable.
- In Fernandes Machado et al. (2025b), we suggest a method based on **transporting** the values of categorical data represented in the **simplex**.

Random Matching

To transport $\mathbf{x}$, a possibility is to perform **random matching** between $\{\boldsymbol{x}_{0,1}, \cdots, \boldsymbol{x}_{0,n_0}\}$ and $\{\boldsymbol{x}_{1,1}, \cdots, \boldsymbol{x}_{1,n_1}\}$ using cost function $c(\boldsymbol{x}_{0,i}, \boldsymbol{x}_{1,j}) = 1_{\{\boldsymbol{x}_{0,i} \neq \boldsymbol{x}_{1,j}\}}$.

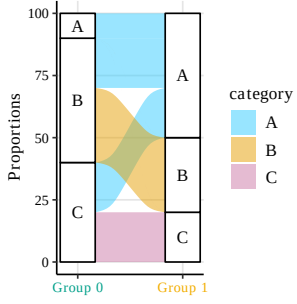
Example of a random matching within categories, $n_0 = n_1 = 6$

		7	8	9	10	11	12
		A	B	B	B	C	C
1	A						x
2	A			x			
3	A	x					
4	B				x		
5	B		x				
6	C					x	

Random Matching with Arbitrary Ordering Over Categories

We can also perform **random matching by assuming an ordering of categories**, using cost function $c(\boldsymbol{x}_{0,i}, \boldsymbol{x}_{1,j}) = \tilde{x}_{0,i} - \tilde{x}_{1,j}$, where $\tilde{x}$ represents the category order.

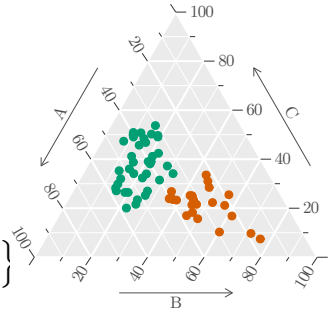
Example of a random matching within ordered categories



From Category To Probabilities

- Let **x** be a **categorical variable** taking values in $\{x_1, \cdots, x_d\}$ (d categories).
- For example, consider $d = 3$, with **x** taking values in $\{A, B, C\}$.
- To apply Optimal Transport, we embed each category into the **probability simplex**:

$$\boldsymbol{p} = (p_A, p_B, p_C) \in \mathcal{S}_2 = \left\{ \mathbf{x} \in [0,1]^3 : x_1 + x_2 + x_3 = 1 \right\}$$



This turns the categorical feature into a **continuous compositional vector** on which OT can be applied.

Our Contribution: From Categorical to Compositional Data

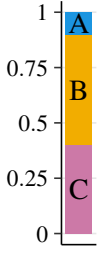
To obtain a counterfactual for each observation from **group 0** to **group 1**, we follow a three steps procedure:

1. **Representation** (from categorical to compositional data): encode the categorical variable as a point in the probability simplex.
2. **Transport** (coupling on the simplex): Optimal Transport within the simplex to learn a mapping from **group 0** to **group 1**.
3. **Reassignment** (from composition to categorical data): assign a category to each matched individual by transporting mass from a continuous distribution to a discrete one, using optimal transport theory where the target is concentrated at the simplex's vertices.

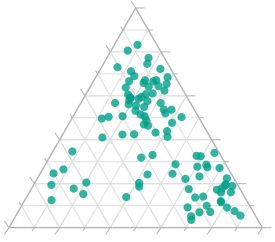
Step 1: Representation

- We convert the categorical variable $\mathbf{x} \in [\![d]\!]$ into a compositional vector $\hat{\boldsymbol{p}} \in \mathcal{S}_{d-1}$.
- To do so, we train a **probabilistic classifier** (a Multinomial Logistic Regression).

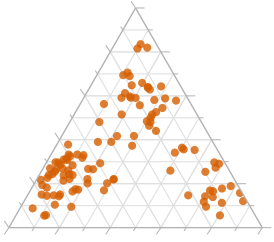
Initial freq. in **group 0**



Representation in group 0



Representation in group 1

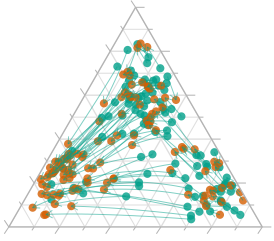


Step 2: Transport

- We need to compute the **distance** between the estimated probabilities $\hat{\boldsymbol{p}}_0$ and $\hat{\boldsymbol{p}}_1$.
- The distance metric is based on the optimal transport cost between two probability vectors in the unit simplex (called "Dirichlet transport" in Baxendale and Wong (2022)):

$$\forall \mathbf{x}, \mathbf{y} \in \mathcal{S}_d, \quad c(\mathbf{x}, \mathbf{y}) = \log\left(\frac{1}{d}\sum_{i=1}^{d}\frac{y_i}{x_i}\right) - \frac{1}{d}\sum_{i=1}^{d}\log\left(\frac{y_i}{x_i}\right) \quad .$$

We can then solve the OT problem with that cost function.



Step 3: Reassignment

- After transport, each individual from **group 0** has a **counterfactual composition**

$$\hat{\boldsymbol{p}}^\star \in \mathcal{S}_2.$$

- But the original variable is **categorical**: $\mathbf{x} \in \{A, B, C\}$.
- We must convert the transported composition $\hat{\boldsymbol{p}}^\star$ back into a **single category**.
- This conversion must:
  - respect the **geometry** of the simplex,
  - ensure the **counterfactual distribution matches** the distribution in group 1,
  - be **deterministic**.
- This is achieved using **semi-discrete Optimal Transport**: a continuous distribution (simplex) is mapped to a discrete one (the simplex's vertices).

Step 3: Reassignment via Power Diagram

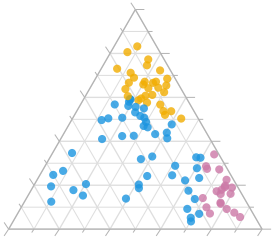
- Semi-discrete OT maps each point $\hat{\boldsymbol{p}}^\star \in \mathcal{S}_2$ to one of the three vertices:

$$\boldsymbol{u}_A = (1,0,0),\ \boldsymbol{u}_B = (0,1,0),\ \boldsymbol{u}_C = (0,0,1).$$

- OT with quadratic cost produces a **Laguerre–Voronoi (power) diagram** partitioning the simplex into convex regions: $R_A,\ R_B,\ R_C$.
- The counterfactual category is determined by the region containing the transported composition:

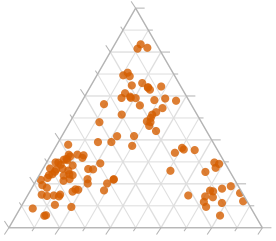
$$\hat{\boldsymbol{p}}^\star \in R_i \quad \Rightarrow \quad \mathbf{x}^\star = i.$$

- This way, we obtain a category assignment with mass preservation (same proportions as in **group 1**).

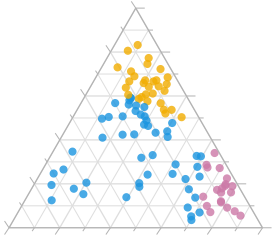


Step 3: Illustration of the Reassignment

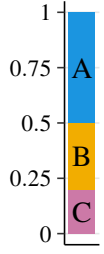
Matched compositional representation in **group 1**



Categorical label from the power diagram

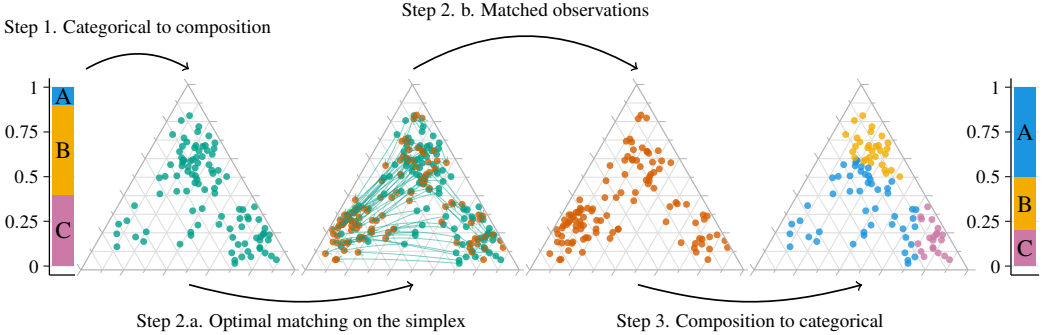


Classes of the counterfactual individuals



Wrap Up

*Generation of counterfactual values from **group 0** to **group 1** using "Dirichlet OT"*



6. Conclusion

Conclusion

- Without addressing algorithmic fairness issues: having fair model is illusive.
- Addressing fairness using a sequential approach (which requires to know the causal structure *a priori* provides an **explainable method**.
- We suggest using **optimal transport** on the simplex to build **counterfactuals** for **categorical data**.


Agathe
Fernandes Machado


Arthur
Charpentier


Ewen
Gallic

7. Appendix

References I

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

Avraham, R. (2017). *Discrimination and Insurance*, page 335–347. Routledge.

Baldus, D. C. and Cole, J. W. (1980). Statistical proof of discrimination. *(No Title)*.

Barocas, S., Hardt, M., and Narayanan, A. (2023). *Fairness and Machine Learning: Limitations and Opportunities*. Adaptive Computation and Machine Learning series. MIT Press.

Baxendale, P. and Wong, T.-K. L. (2022). Random concave functions. *The Annals of Applied Probability*, 32(2):812–852.

Becker, G. S. (1957). *The economics of discrimination*. University of Chicago press.

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

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

Campbell, C. and Smith, D. (2023). Distinguishing between direct and indirect discrimination. *The Modern Law Review*, 86(2):307–330.

Carlier, G., Galichon, A., and Santambrogio, F. (2008). From knothe's transport to brenier's map and a continuation method for optimal transport.

References II

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.

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

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

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*, ITCS '12, page 214–226. ACM.

European Union AI Act (2024). 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*. Article 5 – Prohibited AI Practices.

Feller, A., Pierson, E., Corbett-Davies, S., and Goel, S. (2016). A computer program used for bail and sentencing decisions was labeled biased against blacks. it's actually not that clear.

Fernandes Machado, A., Charpentier, A., and Gallic, E. (2025a). Sequential conditional transport on probabilistic graphs for interpretable counterfactual fairness. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(18):19358–19366.

Fernandes Machado, A., Gallic, E., and Charpentier, A. (2025b). 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.

References III

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188.

Hajian, S. and Domingo-Ferrer, J. (2013). A methodology for direct and indirect discrimination prevention in data mining. *IEEE Transactions on Knowledge and Data Engineering*, 25(7):1445–1459.

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

Hellman, D. (2008). When is discrimination wrong?

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

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

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

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.

Larson, Jeff ans Mattu, S., Kirchner, L., and Angwin, J. (2016). How we analyzed the compas recidivism algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pages 666–704.

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

References IV

Pedreshi, D., Ruggieri, S., and Turini, F. (2008). Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD08, page 560–568. ACM.

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

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

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

Santambrogio, F. (2015). *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Progress in Nonlinear Differential Equations and Their Applications. Springer International Publishing.

Scalia, A. (1979). The disease as cure: In order to get beyond racism, we must first take account of race. *Wash. ULQ*, page 147.

Turner, R. (2015). The way to stop discrimination on the basis of race. *Stan. JCR & CL*, 11:45.

Upton, G. and Cook, I. (2014). *A Dictionary of Statistics 3e*. Oxford Paperback Reference. OUP Oxford.

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

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

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

8.1. Fairness Quantification

Mitigation

Some techniques can be used to prevent models from perpetuating biases with respect to the sensitive attribute. These techniques can be applied at several stages (Hajian and Domingo-Ferrer, 2013)

1. **Preprocessing**: transform source data to remove biases before model training.
2. **In-processing**: modify algorithms to embed fairness constraints during training.
3. **Postprocessing**: alter models after training to correct unfair outcomes.

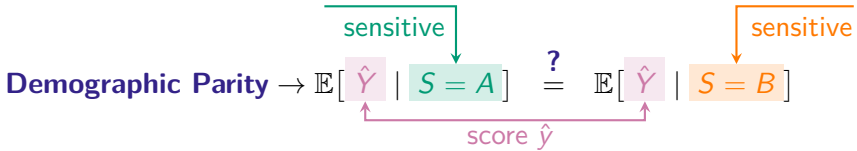
How Can Fairness be Quantified?

We would like to **quantify unfairness** of a **supervised model** $\hat{m}(\cdot)$ trained on a set $\{(y_i, \boldsymbol{x}_i, s_i)\}_{i=1}^n$, where y is the value to predict (i.e., the outcome), $\boldsymbol{x}$ is a set of (unprotected) predictors, s is a **protected attribute**, and $i \in \{1, \ldots, n\}$ denotes an individual.

The outcome may be:

- **Binary** (classification task):
  - $\hat{y}_i = \mathbf{1}(\hat{m}(\boldsymbol{x}_i, s_i) > \text{threshold}) \in \{0, 1\}$
- **Continuous** (regression task):
  - $\hat{y}_i = \hat{m}(\boldsymbol{x}_i, s_i) \in [0, 1]$: a score
  - $\hat{y}_i = \hat{m}(\boldsymbol{x}_i, s_i) \in \mathbb{R}$: a premium

Group Fairness Metrics in a Nutshell



8.2. Optimal Transport Theory

Multivariate Optimal Transport

Optimal map for continuous multivariate distributions (Brenier, 1991)

With a quadratic cost and suppose that μ is absolutely continuous w.r.t. the Lebesgue measure in $\mathbb{R}^d$, the optimal Monge map $T^\star$ is unique, and it is the gradient of a convex function, $T^\star = \nabla\varphi$.

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

Optimal Transport and Monge Mapping

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 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.

Optimal Transport and Wasserstein distance

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 $\mu_{1:d}$, $T_{d-1}^\star(\cdot|x_d)$ 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

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

Conditional transport (3/3)

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

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

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

This mapping will be called "sequential 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 $\boldsymbol{A} = \boldsymbol{\Sigma}_0^{-1/2}(\boldsymbol{\Sigma}_0^{1/2}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_0^{1/2})^{1/2}\boldsymbol{\Sigma}_0^{-1/2}$, 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).

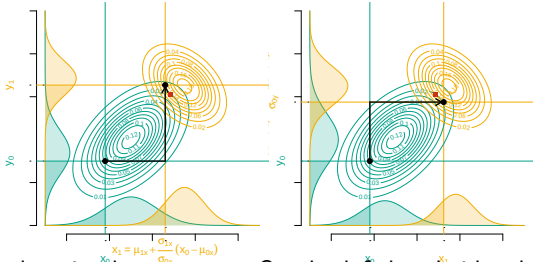
Gaussian transport (2/3)

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. (2025a).

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)



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."