Sequential Conditional Transport on Probabilistic Graphs for Interpretable Counterfactual Fairness

Agathe Fernandes Machado[1], **Arthur Charpentier**[1], **Ewen Gallic**[2]

[1]Université du Québec à Montréal
[2]Aix Marseille Univ, CNRS, AMSE, Marseille, France
fernandes_machado.agathe@courrier.uqam.ca, charpentier.arthur@uqam.ca, ewen.gallic@univ-amu.fr

Abstract

In this paper, we link two existing approaches to derive counterfactuals: adaptations based on a causal graph, and optimal transport. We extend "Knothe's rearrangement" and "triangular transport" to probabilistic graphical models, and use this counterfactual approach, referred to as sequential transport, to discuss fairness at the individual level. After establishing the theoretical foundations of the proposed method, we demonstrate its application through numerical experiments on both synthetic and real datasets.

Code — https://github.com/fer-agathe/sequential_transport
Extended version — https://arxiv.org/abs/2408.03425

1 Introduction

Most applications concerning discrimination and fairness are based on "group fairness" concepts (as introduced in Hardt, Price, and Srebro (2016); Kearns and Roth (2019), or Barocas, Hardt, and Narayanan (2023)). However, in many cases, fairness should be addressed at the individual level rather than globally. As claimed in Dwork et al. (2012), "*we capture fairness by the principle that any two individuals who are similar with respect to a particular task should be classified similarly*." The concept of "counterfactual fairness" was formalized in Kusner et al. (2017), addressing questions such as "*had the protected attributes of the individual been different, other things being equal, would the decision had remain the same?*" Such a statement has clear connections with causal inference, as discussed in Pearl and Mackenzie (2018). Formally, consider observations $\{s_i, \boldsymbol{x}_i, y_i\}$, where s is a binary protected attribute (e.g., $s \in \{0, 1\}$), and $\boldsymbol{x}$ is a collection of legitimate features (possibly correlated with s). The model output is y, which is analyzed to address "algorithmic fairness" issues. Following Rubin (2005), let $y^\star(s)$ denote the potential outcome of y if s is seen as a treatment. With these notations, counterfactual fairness is achieved for individual $(s, \boldsymbol{x})$ if the average "treatment effect," conditional on $\boldsymbol{x}$ (or "CATE") is zero, i.e., $\mathbb{E}[Y^\star(1) - Y^\star(0)|\boldsymbol{X} = \boldsymbol{x}] = 0$. This quantity could be termed "*ceteris paribus* CATE" since all $\boldsymbol{x}$'s are supposed to remain unchanged for both treated and non-treated.

Following Kilbertus et al. (2017), it is possible to suppose that the protected attribute s could actually affect some explanatory variables $\boldsymbol{x}$ in a non-discriminatory way. In Charpentier, Flachaire, and Gallic (2023), the outcome y was "having a surgical intervention" during childbirth in the U.S., s was the mother's ethnic origin ("Black," or not) and $\boldsymbol{x}$ included factors such as "weight of the baby at birth." If Black mothers undergo less surgery because they tend to have smaller babies, there is no discrimination *per se*. At the very least, it should be fair, when assessing whether hospitals have discriminatory policies, to account for that difference in baby weights. Such a variable is named "resolving variable" in Kilbertus et al. (2017). Using heuristic notations, the "*ceteris paribus* CATE" $\mathbb{E}[Y^\star(1)|\boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y^\star(0)|\boldsymbol{X} = \boldsymbol{x}]$ should become a "*mutatis mutandis* CATE". For some individual $(s = 0, \boldsymbol{x})$, this indicator would be $\mathbb{E}[Y^\star(1)|\boldsymbol{X} = \boldsymbol{x}^\star(1)] - \mathbb{E}[Y^\star(0)|\boldsymbol{X} = \boldsymbol{x}]$, as coined in Charpentier, Flachaire, and Gallic (2023), to quantify discrimination, where fictitious individual $(s = 1, \boldsymbol{x}^\star(1))$ is a "counterfactual" version of $(s = 0, \boldsymbol{x})$.

Two recent approaches have been proposed to assess counterfactual fairness using this *mutatis mutandis* approach. On the one hand, Plečko and Meinshausen (2020) and Plečko, Bennett, and Meinshausen (2024) used causal graphs (DAGs) to construct counterfactuals and assess the counterfactual fairness of outcomes y based on variables $(s, \boldsymbol{x}, y)$. In network flow terminology, s acts as a "source" (only outgoing flow, or no parents), while y is a "sink" (only incoming flow). On the other hand, Black, Yeom, and Fredrikson (2020), Charpentier, Flachaire, and Gallic (2023) and De Lara et al. (2024) used optimal transport (OT) to construct counterfactuals. Moreover, using counterfactual reasoning to achieve fair machine learning (ML) models has been notably studied (Ma et al. 2023; Robertson et al. 2024). For evaluation, while De Lara et al. (2024) provided a theoretical framework, its implementation is challenging (except in the Gaussian case), and usually hard to interpret. Here, we combine the two approaches, using OT within a causal graph structure. The idea is to adapt "Knothe's rearrangement" (Bonnotte 2013), or "triangular transport" (Zech and Marzouk 2022a,b), to a general probabilistic graphical model on $(s, \boldsymbol{x}, y)$, rather than a simplistic $s \to x_1 \to x_2 \to \cdots \to x_d \to y$. The concept of "conditional OT" has been recently discussed in Bunne, Krause,

and Cuturi (2022) and Hosseini, Hsu, and Taghvaei (2023), but here, instead of learning the causal graph, we assume a known causal graph and use it to construct counterfactual versions of individuals $(s_i, \boldsymbol{x}_i, y_i)$ to address fairness issues. Additionally, since we use univariate (conditional) transport, standard classical properties of univariate transport facilitates explanations (non-decreasing mappings, and quantile based interpretations).

Main Contributions

- We use multivariate transport theory for constructing counterfactuals, as suggested in De Lara et al. (2024), and connect it to quantile preservation on causal graphs from Plečko and Meinshausen (2020) to develop a sequential transport methodology that aligns with the underlying DAG of the data.
- Sequential transport, using univariate transport maps, provides closed-form solutions for deriving counterfactuals. This allows for the development of a data-driven estimation procedure that can be applied to new out-of-samples observations without recalculating, unlike multivariate OT with non-Gaussian distributions.
- The approach's applicability is demonstrated through numerical experiments on both synthetic data and case studies, highlighting the interpretable analysis of individual counterfactual fairness when using sequential transport.

Section 2 introduces various concepts used in probabilistic graphical models from a causal perspective. Section 3, revisits classical OT covering both univariate and multivariate cases. Sequential transport is covered in Section 4. Section 5 discusses counterfactual fairness. Illustration with real data is provided in Section 6.

2 Graphical Models and Causal Networks

2.1 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). 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 $\text{parents}(x_i)$ are nodes with edges directed towards x_i, in $\mathcal{G}$. Watson et al. (2021) suggested the causal graph in Figure 1 for the German Credit dataset, where s is the "sex" (top left) and y is the "default" indicator (right). Observe that variables $\boldsymbol{x}_j$ are here sorted. As discussed in Ahuja, Magnanti, and Orlin (1993), such a causal graph 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

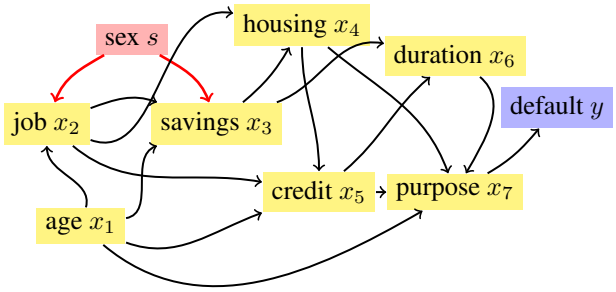


Figure 1: Causal graph in the German Credit dataset from Watson et al. (2021), or DAG.

in the graph. In our analysis, we consider a network $\mathcal{G}$ on variables $\{s, \boldsymbol{x}, y\}$ where s is the sensitive attribute, acting as a "source" (only outgoing flow, or no parents) while y is a "sink" (only incoming flow, i.e., $y \notin \text{parents}(x_i), \forall i$).

2.2 Causal Networks and Linear Structural Models

Wright (1921, 1934) used directed graphs to represent probabilistic cause-and-effect relationships among a set of variables and developed path diagrams and path analysis. Simple causal networks can be visualized on top of Figure 2. On the left is a simple model where the "cause" C directly causes ($\to$) the "effect" E. On the right, a "mediator" X is added. There is still the direct impact of C on E ($C \to E$), but there is also a mediated indirect impact ($C \to X \to E$).

Intervention in a Linear Structural Model In a simple causal graph, with two nodes, C (the cause) and E (the effect), the causal graph is $C \to E$, and the mathematical interpretation can be summarized in two (linear) assignments:

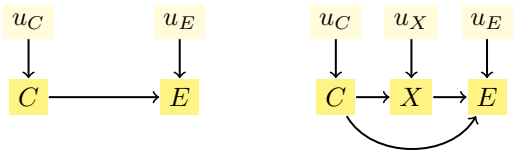
$$\begin{cases} C = a_c + U_C \\ E = a_e + b_e C + U_E, \end{cases} \tag{1}$$

where U_C and U_E are independent Gaussian random variables. That causal graph can be visualized in Figure 2, and its corresponding structural causal model (SCM) described in Equation 1 illustrates the causal relationships between variables, as in Pearl (2000). Suppose here that C is a binary variable, taking values in $\{c_0, c_1\}$. Given an observation (c_0, e), the "counterfactual outcome" if the cause had been set to c_1 (corresponding to the intervention in Figure 3), would be $e + b_e(c_1 - c_0)$. Following Pearl (2009), one can also introduce the "twin network" corresponding to a mirrored version of the initial causal graph in the counterfactual world. Plečko and Meinshausen (2020) coined this approach "fair-twin projection" when C is a binary sensitive attribute.

2.3 Non-Linear Structural Models

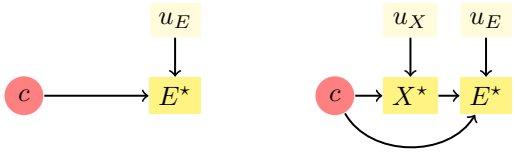
Presentation of the Model More generally, consider a non-Gaussian and nonlinear structural model, named "nonparametric structural equation model" (with independent errors) in Pearl (2000),

$$\begin{cases} C = h_c(U_C) \\ E = h_e(C, U_E), \end{cases} \tag{2}$$



$$\begin{cases} C = a_c + U_C \\ E = a_e + b_e C + U_E \end{cases} \qquad \begin{cases} C = a_c + U_C \\ X = a_x + b_x C + U_X \\ E = a_e + b_e C + \gamma_e X + U_E \end{cases}$$

Figure 2: Linear Structural Causal Model – observation.



$$\begin{cases} C = c \ \ (\text{or } \mathrm{do}(C = c)) \\ E_c^\star = a_e + b_e c + U_E \end{cases} \qquad \begin{cases} C = c \ \ (\text{or } \mathrm{do}(C = c)) \\ X_c^\star = a_x + b_x c + U_X \\ E_c^\star = a_e + b_e c + \gamma_e X_c^\star + U_E \end{cases}$$

Figure 3: Linear Structural Causal Model – intervention.

where $u \mapsto h_c(\cdot, u)$ and $u \mapsto h_e(\cdot, u)$ are strictly increasing in u; U_C and U_E are independent, and, without loss of generality, supposed to be uniform on $[0, 1]$. For a rigorous mathematical framework for non-linear non-Gaussian structural causal models, see Bongers et al. (2021) or Shpitser, Richardson, and Robins (2022).

Connections With Conditional Quantiles Consider now some general DAG, $\mathcal{G}$, on $\boldsymbol{X} = (X_1, \cdots, X_d)$, supposed to be absolutely continuous. With previous notations, $X_i = h_i(\text{parents}(X_i), U_i)$, a.s., for all variables, representing the structural equations. We can write this compactly as $\boldsymbol{X} = h(\text{parents}(\boldsymbol{X}), \boldsymbol{U})$, a.s., by considering h as a vector function. Solving the structural model means finding a function g such that $\boldsymbol{X} = g(\boldsymbol{U})$, a.s. To illustrate, consider a specific i, and $X_i = h_i(\text{parents}(X_i), U_i)$. If $\text{parents}(X_i) = \boldsymbol{x}$ is fixed, define $h_{i|\boldsymbol{x}}(u) = h_i(\boldsymbol{x}, u)$. Let U be a uniform random variable, and let $F_{i|\boldsymbol{x}}$ be the cumulative distribution of $h_{i|\boldsymbol{x}}(U)$, $F_{i|\boldsymbol{x}}(x) = \mathbb{P}[h_{i|\boldsymbol{x}}(U) \leq x]$. Since X_i is absolutely continuous, $F_{i|\boldsymbol{x}}$ is invertible, and $F_{i|\boldsymbol{x}}^{-1}$ is a conditional quantile function (conditional on $\text{parents}(X_i) = \boldsymbol{x}$). Let $V = F_{i|\boldsymbol{x}}(h_{i|\boldsymbol{x}}(U))$, then $X_i = F_{i|\boldsymbol{x}}^{-1}(V)$ and V is uniformly distributed on $[0, 1]$. This means that $x_i = h_{i|\boldsymbol{x}}(u_i)$ corresponds to the quantile of variable X_i, conditional on the values of its parents, $\text{parents}(X_i)$, with probability level u_i. In the observational world, u_i represents the (conditional) probability level associated with observation x_i, and its counterfactual counterpart is $x_i^\star$ corresponding to the (conditional) quantile associated with the same probability level u_i.

This representation has been used in Plečko and Meinshausen (2020) and Plečko, Bennett, and Meinshausen (2024), where $X_i = F_{i|\boldsymbol{x}}^{-1}(V)$ is simply the probabilistic representation of "quantile regression," as introduced by Koenker and Bassett Jr (1978) (and further studied in Koenker (2005) and Koenker et al. (2017)). This could be extended to "quantile regression forests," as in Meinshausen and Ridgeway (2006), or any kind of ML model, as Cannon (2018) or Pearce et al. (2022). Observe that Ma and Koenker (2006) considered some close "recursive structural equation models," characterized by a system of equations where each endogenous variable is regressed on other endogenous and exogenous variables in a hierarchical manner. They used some sequential quantile regression approach to solve those recursive SEMs. An alternative we consider here is to use the connection between quantiles and OT (discussed in Chernozhukov, Fernández-Val, and Melly (2013) or Hallin and Konen (2024)) to define some "conditional transport" that relates to those conditional quantiles.

3 Optimal Transport

Given two metric spaces $\mathcal{X}_0$ and $\mathcal{X}_1$, consider a measurable map $T : \mathcal{X}_0 \to \mathcal{X}_1$ and a measure μ_0 on $\mathcal{X}_0$. The push-forward of μ_0 by T is the measure $\mu_1 = T_\# \mu_0$ on $\mathcal{X}_1$ defined by $T_\#\mu_0(B) = \mu_0\big(T^{-1}(B)\big), \forall B \subset \mathcal{X}_1$. For all measurable and bounded $\varphi : \mathcal{X}_1 \to \mathbb{R}$,

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

For our applications, if we consider measures $\mathcal{X}_0 = \mathcal{X}_1$ as a compact subset of $\mathbb{R}^d$, then there exists T such that $\mu_1 = T_\#\mu_0$, when μ_0 and μ_1 are two measures, and μ_0 is atomless, as shown in Villani (2003) and Santambrogio (2015). In that case, and if we further suppose that measures μ_0 and μ_1 are absolutely continuous, with densities f_0 and f_1 (w.r.t. Lebesgue measure), a classical change of variable expression can be derived. Specifically, the previous integral

$$\int_{\mathcal{X}_1} \varphi(\boldsymbol{x}_1) f_1(\boldsymbol{x}_1) \mathrm{d}\boldsymbol{x}_1$$

is simply (if ∇T is the Jacobian matrix of mapping T):

$$\int_{\mathcal{X}_0} \varphi\big(T(\boldsymbol{x}_0)\big) \underbrace{f_1(T(\boldsymbol{x}_0)) \det \nabla T(\boldsymbol{x}_0)}_{=f_0(\boldsymbol{x}_0)} \mathrm{d}\boldsymbol{x}_0.$$

Out of those mappings from μ_0 to μ_1, we can be interested in "optimal" mappings, satisfying Monge problem, from Monge (1781), i.e., solutions of

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

for some positive ground cost function $c : \mathcal{X}_0 \times \mathcal{X}_1 \to \mathbb{R}_+$.

In general settings, however, such a deterministic mapping T between probability distributions may not exist (in particular if μ_0 and μ_1 are not absolutely continuous, with respect to Lebesgue measure). This limitation motivates the Kantorovich relaxation of Monge's problem, as considered in Kantorovich (1942),

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

with our cost function c, where $\Pi(\mu_0, \mu_1)$ is the set of all couplings of μ_0 and μ_1. This problem focuses on couplings rather than deterministic mappings It always admits solutions referred to as OT plans.

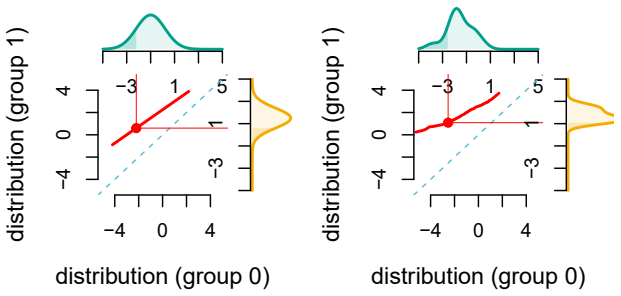


Figure 4: Univariate OT, with Gaussian distributions (left) and general marginal distributions (right). The transport curve ($T^\star$) is shown in red.

3.1 Univariate Optimal Transport

Suppose here that $\mathcal{X}_0 = \mathcal{X}_1$ is a compact subset of $\mathbb{R}$. The optimal Monge map $T^\star$ for some strictly convex cost c such that $T^\star_\# \mu_0 = \mu_1$ is $T^\star = F_1^{-1} \circ F_0$, where $F_i : \mathbb{R} \to [0,1]$ is the cumulative distribution function associated with μ_i, $F_i(x) = \mu_i((-\infty, x])$, and F_i^{-1} is the generalized inverse (corresponding to the quantile function), $F_i^{-1}(u) = \inf\{x \in \mathbb{R} : F_i(x) \geq u\}$. Observe that $T^\star$ is an increasing mapping (which is the univariate definition of being the gradient of a convex function, from Brenier (1991)). This is illustrated in Figure 4, with a Gaussian case on the left ($T^\star$ affine), and general densities on the right.

3.2 Multivariate Optimal Transport

In a multivariate setting, when $\mathcal{X}_0 = \mathcal{X}_1$ is a compact subset of $\mathbb{R}^d$, from Brenier (1991), with a quadratic cost, the optimal Monge map $T^\star$ is unique, and it is the gradient of a convex mapping $\psi : \mathbb{R}^d \to \mathbb{R}^d$, $T^\star = \nabla\psi$. Therefore, its Jacobian matrix $\nabla T^\star$ is nonnegative and symmetric. More generally, 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, as proved in Cordero-Erausquin (2004) and Ambrosio, Gigli, and Savaré (2005). 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, as in Appendix A.[1]

4 Sequential Transport

4.1 Knothe-Rosenblatt Conditional Transport

As explained in Villani (2003); Carlier, Galichon, and Santambrogio (2010); Bonnotte (2013), the Knothe-Rosenblatt (KR) rearrangement is directly inspired by the Rosenblatt chain rule, from Rosenblatt (1952), and some extensions obtained on general measures by Knothe (1957). Using notations of Section 2.3 in Santambrogio (2015), let $\mu_{0:d}$ 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. Suppose that the μ_0-conditionals, corresponding to the measures $\mu_{0:d}$, $\mu_{0:d-1|d}$, etc., are atomless (satisfied as soon as μ_0 is absolutely continuous with respect to the Lebesgue measure). For the first two,

$$\mu_0(\mathbb{R}^{d-1} \times \mathrm{d}x_d) = \mu_{0:d}(\mathrm{d}x_d)$$
$$\mu_0(\mathbb{R}^{d-2} \times \mathrm{d}x_{d-1} \times \mathrm{d}x_d) = \mu_{0:d}(\mathrm{d}x_d)\mu_{0:d-1|d}(\mathrm{d}x_{d-1}|x_d)$$

and iterate. Define conditional (univariate) cumulative distribution functions:

$$\begin{cases} F_{0:d}(x_d) = \mu_{0:d}((-\infty, x_d]) = \mu_0(\mathbb{R}^{d-1} \times (-\infty, x_d]) \\ F_{0:d-1|d}(x_{d-1}|x_d) = \mu_{0:d-1|d}((-\infty, x_{d-1}]|x_d), \end{cases}$$

etc. And similarly for μ_1. For the first component, let $T_d^\star$ denote the monotone nondecreasing map transporting from $\mu_{0:d}$ to $\mu_{1:d}$, defined as $T_{\underline{d}}^\star(\cdot) = F_{1:d}^{-1}(F_{0:d}(\cdot))$. For the second component, let $T_{\overline{d-1}}^\star(\cdot|x_d)$ denote the monotone nondecreasing map transporting from $\mu_{0:d-1|d}(\cdot|x_d)$ to $\mu_{1:d-1|d}(\cdot|T_{\underline{d}}^\star(x_d))$, $T_{\overline{d-1}}^\star(\cdot|x_d) = F_{1:d-1|d}^{-1}(F_{0:d-1|d}(\cdot|x_d)|T_{\underline{d}}^\star(x_d))$. We can then repeat the construction, and finally, the KR rearrangement is

$$T_{\overline{kr}}(x_1, \cdots, x_d) = \begin{pmatrix} T_{\overline{1}}^\star(x_1|x_2, \cdots, x_d) \\ T_{\overline{2}}^\star(x_2|x_3, \cdots, x_d) \\ \vdots \\ T_{\overline{d-1}}^\star(x_{d-1}|x_d) \\ T_{\overline{d}}^\star(x_d) \end{pmatrix}.$$

As proved in Santambrogio (2015) and Carlier, Galichon, and Santambrogio (2010), $T_{\overline{kr}}$ is a transportation map from μ_0 to μ_1, in the sense that $\mu_1 = T_{\overline{kr}\#}\mu_0$. Following Bogachev, Kolesnikov, and Medvedev (2005) and Backhoff et al. (2017), $T_{\overline{kr}}$ is the "monotone upper triangular map" uniquely defined when the μ_1-conditionals are atomless for a chosen coordinate order. Bogachev, Kolesnikov, and Medvedev (2005) defined the "monotone lower triangular map,"

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

The map $x_i \mapsto T_{\underline{i}}^\star(x_i|x_1, \cdots, x_{i-1})$ is monotone (nondecreasing) for all $(x_1, \cdots, x_{i-1}) \in \mathbb{R}^{i-1}$. Further, by construction, this KR transport map has a triangular Jacobian matrix $\nabla T_{\underline{kr}}$ with nonnegative entries on its diagonal, making it suitable for various geometric applications. However, this mapping does not satisfy many properties; for example, it is not invariant under isometries of $\mathbb{R}^d$ as mentioned in Villani (2009). Carlier, Galichon, and Santambrogio (2010) proved that the KR transport maps could be seen as limits of quadratic OTs. A direct interpretation is that this iterative sequential transport can be seen as "marginally optimal." Some explicit formulas can be obtained in the Gaussian case, as discussed in Appendix A.

[1] Our appendix is available at https://arxiv.org/abs/2408.03425.

4.2 Sequential Conditional Transport on a Probabilistic Graph

The "monotone lower triangular map," introduced in Bogachev, Kolesnikov, and Medvedev (2005) could be used when dealing with time series, since there is a natural ordering between variables, indexed by the time, as discussed in Backhoff et al. (2017) or Bartl, Beiglböck, and Pammer (2021). In the general non-temporal case of time series X_t, it is natural to extend that approach to acyclical probabilistic graphic models, following Cheridito and Eckstein (2023). Instead of two general measures μ_0 and μ_1 on $\mathbb{R}^d$, we use only measures "factorized according to $\mathcal{G}$," some probabilistic graphical model, as defined in Lauritzen (2020).

Definition. *Consider some acyclical causal graph $\mathcal{G}$ on $(s, \boldsymbol{x})$ where variables are topologically sorted, where $s \in \{0, 1\}$ is a binary variable, defining two measures μ_0 and μ_1 on $\mathbb{R}^d$, by conditioning on $s = 0$ and $s = 1$, respectively, factorized according to $\mathcal{G}$. Define*

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

This mapping will be called "sequential conditional transport on the graph $\mathcal{G}$, "or shortly "sequential transport."[2]

A classical algorithm for topological sorting is Kahn (1962)'s "Depth First Search" (DFS), and other algorithms are discussed in Section 20.4 in Cormen et al. (2022). For the causal graphs of Figure 5:

$$T^\star_{\mathcal{G}}(x_1, x_2) = \begin{pmatrix} T^\star_1(x_1) \\ T^\star_2(x_2|x_1) \end{pmatrix}, \text{ for Figure 5a,}$$

$$T^\star_{\mathcal{G}}(x_1, x_2) = \begin{pmatrix} T^\star_1(x_1|x_2) \\ T^\star_2(x_2) \end{pmatrix}, \text{ for Figure 5b.}$$

In that simple case, for Figure 5a, we recognize the "monotone lower triangular map," and the "monotone upper triangular map," for 5b (see Section 4.1). Finally, for the causal graph on the German Credit dataset of Figure 1, variables are sorted, and

$$T^\star_{\mathcal{G}}(x_1, \cdots, x_7) = \begin{pmatrix} T^\star_1(x_1) \\ T^\star_2(x_2|x_1) \\ T^\star_3(x_3|x_1, x_2) \\ T^\star_4(x_4|x_2, x_3) \\ T^\star_5(x_5|x_1, x_2, x_4) \\ T^\star_6(x_6|x_3, x_5) \\ T^\star_7(x_7|x_1, x_4, x_5, x_6) \end{pmatrix}.$$

Alternatively, using the "monotone lower triangular map" for the German Credit dataset to compute counterfactuals would imply that the assumed DAG contains more edges than the DAG illustrated in Figure 5. In this case, the edges are specified as $E = \{(i, j) \in V^2 : i < j\}$, with $V = \{s, x_1, x_2, \ldots, x_7\}$. Notably, multivariate OT corresponds to a fully connected graph, with $E = \{(i, j) \in V^2 : i \neq j\}$ (Cheridito and Eckstein 2023). The impact of edge misspecifications on sequential transport is examined in Appendix E. If the graph is entirely unknown, one could infer it as discussed in (Zheng et al. 2018; Yu et al. 2019; Cai et al. 2023), or use Bayes'rule to compute posterior DAGs, incorporating uncertainty quantification for counterfactuals (Toth et al. 2022).

For the fairness application in the next section, s is treated as a "source" with no parents, allowing it to be the first vertex in the topological ordering of the network on $(s, \boldsymbol{x})$. The counterfactual value is then derived by propagating "downstream" in the causal graph as s changes from 0 to 1

[2]Given the topological order of the graph and assuming the μ_0, μ_1-conditionals are atomless, the existence and unicity of the sequential transport map are guaranteed, as it involves fewer conditioning variables compared to the KR transport map.

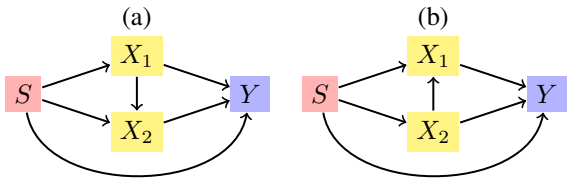


Figure 5: Two simple causal networks, with two legitimate mitigating variables, x_1 and x_2.

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)
    $T_j(\boldsymbol{a}_{\boldsymbol{p}(j)}) \leftarrow (T_{\boldsymbol{p}(j)_1}(\boldsymbol{a}_{\boldsymbol{p}(j)}), \cdots, T_{\boldsymbol{p}(j)_{k_j}}(\boldsymbol{a}_{\boldsymbol{p}(j)}))$
    $(x_{i,j|s}, \boldsymbol{x}_{i,\boldsymbol{p}(j)|s}) \leftarrow$ subsets when $s \in \{0, 1\}$
    $w_{i,j|0} \leftarrow \phi(\boldsymbol{x}_{i,\boldsymbol{p}(j)|0}; \boldsymbol{a}_{\boldsymbol{p}(j)}, \boldsymbol{b}_j)$ (Gaussian kernel)
    $w_{i,j|1} \leftarrow \phi(\boldsymbol{x}_{i,\boldsymbol{p}(j)|1}; T_j(\boldsymbol{a}_{\boldsymbol{p}(j)}), \boldsymbol{b}_j)$
    $\hat{f}_{h_j|s} \leftarrow$ density estimator of $x_{\cdot,j|s}$, weights $w_{\cdot,j|s}$.
    $\hat{F}_{h_j|s}(\cdot) \leftarrow \int_{-\infty}^{\cdot} \hat{f}_{h_j|s}(u)\mathrm{d}u$, c.d.f.
    $\hat{Q}_{h_j|s} \leftarrow \hat{F}^{-1}_{h_j|s}$, quantile
    $\hat{T}_j(\cdot) \leftarrow \hat{Q}_{h_j|1} \circ \hat{F}_{h_j|0}(\cdot)$
  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 $(s = 0, \boldsymbol{a})$

4.3 Algorithm

Algorithm 1 describes this sequential approach, which can be illustrated using the DAG in Figure 5a, as shown in Figure 6. The preliminary step is to determine the topological order of the causal graph. In Figure 5a the order is $(s, (x_1, x_2))$. The first step is to estimate densities $\widehat{f}_{1|s}$ of x_1 in the two groups (s being either 0 or 1) as shown in

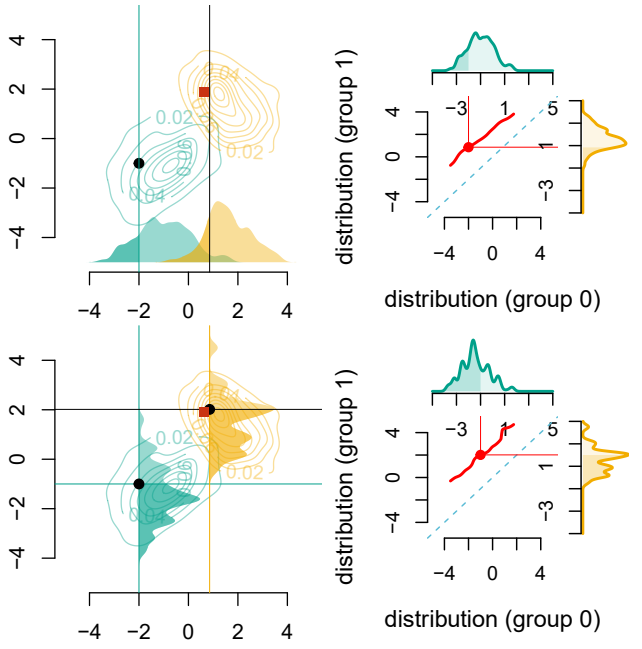


Figure 6: Illustration of Algorithm 1 for DAG in Figure 5a, with simulated data; first step at the top, second step at the bottom. The red square represents the multivariate OT of the bottom-left point.

the top left of Figure 6. Next, numerical integration and inverse are used to compute the cumulative distributions $\widehat{F}_{1|s}$ and quantile functions $\widehat{Q}_{1|s}$. To compute the counterfactual for $(s = 0, \boldsymbol{a})$, $a_1^\star$ is calculated as $\hat{T}_1(a_1)$, where $\hat{T}_1(\cdot) = \hat{Q}_{1|1} \circ \hat{F}_{1|0}(\cdot)$. The second step involves considering the second variable in the topological order, conditional on its parents. Suppose x_2 is the second variable, and for illustration that x_1 is the (only) parent of x_2. The densities $\widehat{f}_{2|s}$ of x_2 are then estimated in the two groups, conditional on their parents: either conditional on $x_1 = a_1$ (subgroup $s = 0$), or conditional on $x_1 = a_1^\star$ (subgroup $s = 1$). This is feasible since all transports of parents were computed in an earlier step. This can be visualized in the bottom left of Figure 6. As in the previous step, the conditional cumulative distributions $\widehat{F}_{2|s}$ and conditional quantile functions $\widehat{Q}_{2|s}$ (conditional on the parents) are computed. Then $a_2^\star$ is determined as $\hat{T}_2(a_2)$ where $\hat{T}_2(\cdot) = \hat{Q}_{2|1} \circ \hat{F}_{2|0}(\cdot)$. This process is repeated until all variables have been considered. At the end, starting from an individual with features $\boldsymbol{x} = \boldsymbol{a}$, in group $s = 0$, the counterfactual version in group $s = 1$ is obtained, with transported features, *mutatis mutandis*, $\boldsymbol{a}^\star$. As the number of parents per variable in the DAG increases, calculating conditional distributions for a variable becomes complex and less robust. Handling categorical variables in counterfactuals is detailed in Appendix B, enabling the application of sequential transport to datasets like `adult income` and `COMPAS` in Appendix D.

5 Interpretable Counterfactual Fairness

5.1 Individual Counterfactual Fairness

General Context Following Dwork et al. (2012), a fair decision means that "similar individuals" are treated similarly. As discussed in the introduction, Kusner et al. (2017) and Russell et al. (2017) considered a "counterfactual fairness" criterion. Based on the approach discussed above, it is possible to quantify unfairness, for a single individual, of a model m, trained on features $(s, \boldsymbol{x})$ to predict an outcome y. If $y \in \{0, 1\}$ is binary, then m represents the underlying score, corresponding to the conditional probability that $y = 1$.

Illustration With Simulated Data Consider the causal graphs in Figure 5, with one sensitive attribute s, two legitimate features x_1 and x_2 and one outcome y. Here, y is the score obtained from a logistic regression, specifically,

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

Iso-scores can be visualized at the top of Figure 7, with group 0 on the left, 1 on the right. Consider an individual $(s, x_1, x_2) = (s = 0, -2, -1)$ in group 0, with a score of 18.24% (bottom left of Figure 7). Using Algorithm 1, its counterfactual counterpart $(s = 1, x_1^\star, x_2^\star)$ can be constructed. The resulting score varies depending on the causal assumption. The score would be 61.40% assuming the causal graph of Figure 5a, and 56.34% assuming causal graph 5b. In the first case, the *mutatis mutandis* difference $m(s = 1, x_1^\star, x_2^\star) - m(s = 0, x_1, x_2)$, i.e., $+43.15\%$, is:

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

The first term is the *ceteris paribus* difference, the second one the change in x_1 and the third one the change in x_2, conditional on the change in x_1. If, instead, we assume the causal graph of Figure 5b, the score of the same individual would become 56.34% and the *mutatis mutandis* difference $m(s = 1, x_1^\star, x_2^\star) - m(s = 0, x_1, x_2)$, i.e., $+38.09\%$, is:

$$\begin{aligned} & m(s = 1, x_1, x_2) - m(s = 0, x_1, x_2) && : -10.66\% \\ + \; & m(s = 1, x_1, x_2^\star) - m(s = 1, x_1, x_2) && : +14.51\% \\ + \; & m(s = 1, x_1^\star, x_2^\star) - m(s = 1, x_1, x_2^\star) && : +34.24\%. \end{aligned}$$

At the bottom right of Figure 7, the *mutatis mutandis* impact on the scores can be visualized.

5.2 Global Fairness Metrics

Instead of focusing on a single individual, it is possible to quantify the fairness of a model m on a global scale. For example, the Demographic Parity criterion can be extended to Counterfactual Demographic Parity (CDP), allowing fairness assessment within a population subgroup with $s = 0$. Consider the empirical version of "counterfactual fairness" in Kusner et al. (2017)

$$\text{CDP} = \frac{1}{n_0} \sum_{i \in \mathcal{D}_0} m(1, \boldsymbol{x}_i^\star) - m(0, \boldsymbol{x}_i), \qquad (3)$$

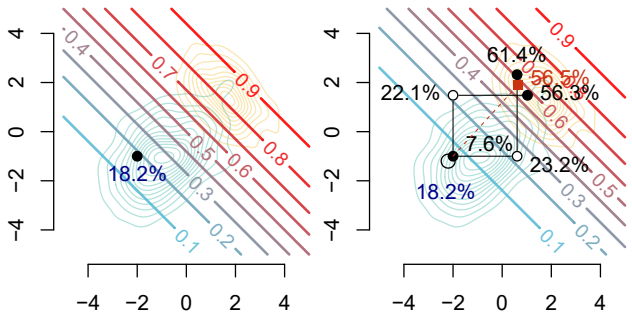


Figure 7: In the background, level curves for $(x_1, x_2) \mapsto m(0, x_1, x_2)$ and $m(1, x_1, x_2)$ respectively on the left and on the right. Then, on the left, individual $(s, x_1, x_2) = (s = 0, -2, -1)$ (predicted 18.2% by model m), and on the right, visualization of two counterfactuals $(s = 1, x_1^\star, x_2^\star)$ according to causal graphs 5a (bottom right path, predicted 61.4%) and 5b (top left path, predicted 56.3%). The red dot is the counterfactual obtained with multivariate OT (predicted 56.5%).

which corresponds to the "average treatment effect of the treated" in the classical causal literature. This can be computed more efficiently using Algorithm 2 in Appendix B, which offers a faster alternative compared to Algorithm 1. Other group fairness metrics, based on Equalized Odds, can be extended to aggregated counterfactual fairness measures (see Appendix C).

6 Application on Real Data

We analyze the Law School Admission Council dataset (Wightman 1998), focusing on four variables: race $s \in \{\text{Black}, \text{White}\}$ (corresponding to 0 and 1), undergraduate GPA before law school (x_1, UGPA), Law School Admission Test (x_2, LSAT), and a binary response (y) indicating whether the first-year law school grade (FYA) is above the median, as described in Black, Yeom, and Fredrikson (2020). Unlike De Lara et al. (2024); Black, Yeom, and Fredrikson (2020); Kusner et al. (2017), we assume the causal graph in Figure 8, where UGPA influences LSAT. We aim to evaluate counterfactual fairness for Black individuals in logistic regression predictions $(\hat{y}|s = 0)$, comparing an "aware" classifier, i.e., that includes s among the explanatory variables, with an "unaware" model that considers only $\boldsymbol{x} = (x_1, x_2)$. Fairness is measured using CDP (see Eq. 3). We apply the sequential transport method from Algorithm 2 to compute counterfactuals $\hat{y}^\star(s = 1)|s = 0$ following the network's topological order in Figure 8. These results are compared with those obtained from multivariate OT (De Lara et al. 2024) and quantile regressions (Plečko, Bennett, and Meinshausen 2024), namely Fairadapt.

Figure 9 illustrates the similarity between Fairadapt and sequential transport, both assuming a DAG, as shown by the counterfactual pathways for a Black individual (left) and the alignment of counterfactual predicted score densities (right). The density of multivariate OT counterfactuals resembles factual White outcome distribution due to its matching process. Overall, the three methods yield similar results, as reflected in the aggregated counterfactual fairness metric in Table 1. Lastly, the "aware" model, which directly incorporates s into its covariates, is less counterfactually fair than the "unaware" model.

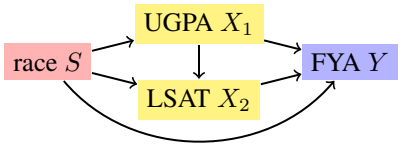


Figure 8: Causal graph of the Law School dataset.

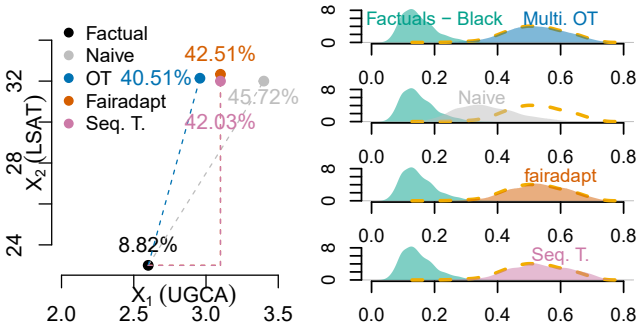


Figure 9: Counterfactual calculations for a Black individuals on the left (percentages indicate predicted scores), and densities of predicted scores (aware model) for all Black individuals with factuals and counterfactuals on the right. The dashed line represents the density of predicted scores for the observed White individuals.

Conclusion

In this paper, we propose a sequential transport approach for constructing counterfactuals based on OT theory while respecting the underlying causal graph of the data. By using conditional univariate transport maps, we derive closed-form solutions for each coordinate of an individual's characteristics, which facilitates the interpretation of both individual counterfactual fairness of our predictive model, and global fairness through "counterfactual demographic parity." Future work could extend counterfactual fairness evaluation to mitigation by applying pre-processing or in-processing methods using sequential transport for counterfactual generation.

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

Table 1: CDP for Black individuals from Eq. 3 comparing classifier predictions over original features $\boldsymbol{x}$ (resp. $(s = 0, \boldsymbol{x})$) and their counterfactuals $\boldsymbol{x}^\star$ (resp. $(s = 1, \boldsymbol{x}^\star)$), using Fairadapt, multivariate OT, and sequential transport.

Acknowledgments

Agathe Fernandes Machado acknowledges that the project leading to this publication has received funding from OBVIA. Arthur Charpentier acknowledges funding from the SCOR Foundation for Science and the National Sciences and Engineering Research Council (NSERC) for funding (RGPIN-2019-07077). Ewen Gallic acknowledge funding from the French government under the "France 2030" investment plan managed by the French National Research Agency (reference: ANR-17-EURE-0020) and from Excellence Initiative of Aix-Marseille University – A*MIDEX.

References

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

Ambrosio, L.; Gigli, N.; and Savaré, G. 2005. *Gradient flows: in metric spaces and in the space of probability measures*. Springer.

Backhoff, J.; Beiglbock, M.; Lin, Y.; and Zalashko, A. 2017. Causal transport in discrete time and applications. *SIAM Journal on Optimization*, 27(4): 2528–2562.

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

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

Bartl, D.; Beiglböck, M.; and Pammer, G. 2021. The Wasserstein space of stochastic processes. arXiv:2104.14245.

Black, E.; Yeom, S.; and Fredrikson, M. 2020. Fliptest: fairness testing via optimal transport. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 111–121.

Bogachev, V. I.; Kolesnikov, A. V.; and Medvedev, K. V. 2005. Triangular transformations of measures. *Sbornik: Mathematics*, 196(3): 309.

Bongers, S.; Forré, P.; Peters, J.; and Mooij, J. M. 2021. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5): 2885–2915.

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

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

Bunne, C.; Krause, A.; and Cuturi, M. 2022. Supervised training of conditional Monge maps. *Advances in Neural Information Processing Systems*, 35: 6859–6872.

Cai, H.; Wang, Y.; Jordan, M.; and Song, R. 2023. On Learning Necessary and Sufficient Causal Graphs. arXiv:2301.12389.

Cannon, A. J. 2018. Non-crossing nonlinear regression quantiles by monotone composite quantile regression neural network, with application to rainfall extremes. *Stochastic Environmental Research and Risk Assessment*, 32: 3207–3225.

Carlier, G.; Galichon, A.; and Santambrogio, F. 2010. From Knothe's transport to Brenier's map and a continuation method for optimal transport. *SIAM Journal on Mathematical Analysis*, 41(6): 2554–2576.

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*, 45–89. Springer.

Cheridito, P.; and Eckstein, S. 2023. Optimal transport and Wasserstein distances for causal models. arXiv:2303.14085.

Chernozhukov, V.; Fernández-Val, I.; and Melly, B. 2013. Inference on counterfactual distributions. *Econometrica*, 81(6): 2205–2268.

Cordero-Erausquin, D. 2004. Non-smooth differential properties of optimal transport. *Contemporary Mathematics*, 353: 61–72.

Cormen, T. H.; Leiserson, C. E.; Rivest, R. L.; and Stein, C. 2022. *Introduction to algorithms*. MIT press.

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*, 214–226.

Hallin, M.; and Konen, D. 2024. Multivariate Quantiles: Geometric and Measure-Transportation-Based Contours. arXiv:2401.02499.

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

Hosseini, B.; Hsu, A. W.; and Taghvaei, A. 2023. Conditional Optimal Transport on Function Spaces. arXiv:2311.05672.

Kahn, A. B. 1962. Topological sorting of large networks. *Commun. ACM*, 5(11): 558–562.

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

Kearns, M.; and Roth, A. 2019. *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press.

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.

Knothe, H. 1957. Contributions to the theory of convex bodies. *Michigan Mathematical Journal*, 4(1): 39–52.

Koenker, R. 2005. *Quantile regression*, volume 38. Cambridge university press.

Koenker, R.; and Bassett Jr, G. 1978. Regression quantiles. *Econometrica: journal of the Econometric Society*, 33–50.

Koenker, R.; Chernozhukov, V.; He, X.; and Peng, L. 2017. *Handbook of quantile regression*. CRC Press.

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., eds., *Advances in Neural Information Processing Systems 30*, 4066–4076. NIPS.

Lauritzen, S. L. 2020. *Lectures on graphical models*. University of Copenhagen.

Ma, J.; Guo, R.; Zhang, A.; and Li, J. 2023. Learning for Counterfactual Fairness from Observational Data. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, 1620–1630. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701030.

Ma, L.; and Koenker, R. 2006. Quantile regression methods for recursive structural equation models. *Journal of Econometrics*, 134(2): 471–506.

Meinshausen, N.; and Ridgeway, G. 2006. Quantile regression forests. *Journal of machine learning research*, 7(6).

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

Pearce, T.; Jeong, J.-H.; jia, y.; and Zhu, J. 2022. Censored Quantile Regression Neural Networks for Distribution-Free Survival Analysis. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 7450–7461. Curran Associates, Inc.

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

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

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

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

Plečko, D.; Bennett, N.; and Meinshausen, N. 2024. fairadapt: Causal Reasoning for Fair Data Preprocessing. *Journal of Statistical Software*, 110(4): 1–35.

Robertson, J.; Hollmann, N.; Awad, N.; and Hutter, F. 2024. FairPFN: Transformers Can do Counterfactual Fairness. arXiv:2407.05732.

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

Rubin, D. B. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469): 322–331.

Russell, C.; Kusner, M. J.; Loftus, J.; and Silva, R. 2017. When worlds collide: integrating different counterfactual assumptions in fairness. *Advances in neural information processing systems*, 30.

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

Shpitser, I.; Richardson, T. S.; and Robins, J. M. 2022. *Multivariate Counterfactual Systems and Causal Graphical Models*, 813–852. New York, NY, USA: Association for Computing Machinery, 1 edition. ISBN 9781450395861.

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

Toth, C.; Lorch, L.; Knoll, C.; Krause, A.; Pernkopf, F.; Peharz, R.; and von Kügelgen, J. 2022. Active Bayesian Causal Inference. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 16261–16275. Curran Associates, Inc.

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

Villani, C. 2009. *Optimal transport: old and new*, volume 338. Springer.

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

Wightman, L. F. 1998. LSAC National Longitudinal Bar Passage Study. LSAC Research Report Series. Technical report, Law School Admission Council, Newtown, PA.

Wright, S. 1921. Correlation and causation. *Journal of Agricultural Research*, 20.

Wright, S. 1934. The method of path coefficients. *The annals of mathematical statistics*, 5(3): 161–215.

Yu, Y.; Chen, J.; Gao, T.; and Yu, M. 2019. DAG-GNN: DAG Structure Learning with Graph Neural Networks. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 7154–7163. PMLR.

Zech, J.; and Marzouk, Y. 2022a. Sparse approximation of triangular transports, part I: The finite-dimensional case. *Constructive Approximation*, 55(3): 919–986.

Zech, J.; and Marzouk, Y. 2022b. Sparse approximation of triangular transports, part II: The infinite-dimensional case. *Constructive Approximation*, 55(3): 987–1036.

Zheng, X.; Aragam, B.; Ravikumar, P.; and Xing, E. P. 2018. DAGs with NO TEARS: continuous optimization for structure learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, 9492–9503. Red Hook, NY, USA: Curran Associates Inc.