Counterfactual Models: The Mass Transportation Viewpoint

Lucas de Lara, Alberto González-Sanz, Nicholas Asher, Jean-Michel Loubes

Counterfactual Models: The Mass Transportation Viewpoint

Lucas De Lara[1] , **Alberto González-Sanz**[1] , **Nicholas Asher**[2,3] and **Jean-Michel Loubes**[1]
[1]IMT , [2]CNRS , [3]IRIT
{lucas.de_lara, alberto.gonzalez_sanz,loubes}@math.univ-toulouse.fr, nicholas.asher@irit.fr,

Abstract

Counterfactual reasoning aims at predicting how the world would have been *had a certain event occurred*, and as such has attracted attention from the fields of explainability and robustness in machine learning. While Pearl's causal inference provides appealing rules to calculate valid counterfactuals, it relies on a model that is unknown and hard to discover in practice. We formalize a mass transportation viewpoint of counterfactual reasoning and use distributional matching methods as a natural model-free surrogate approach. In particular, we show that optimal transport theory defines relevant counterfactuals, as they are numerically feasible, statistically-faithful, and can even coincide with counterfactuals generated by linear additive causal models. We argue this has consequences for interpretability and we illustrate the strength of the mass transportation viewpoint by recasting and generalizing the accepted counterfactual fairness condition into clearer, more practicable criteria.

1 Introduction

A *counterfactual* states how the world should be modified so that a given outcome occurs. For instance, the statement *had you been a woman, you would have gotten half your salary* is a counterfactual relating the *intervention* "had you been a woman" to the *outcome* "you would have gotten half your salary". Counterfactuals have been used to express causal laws [Lewis, 1973] and hence have attracted the attention in the fields of explainability and robustness in machine learning, as such statements can naturally represent the dependence of a prediction on a perturbation of input data without opening the black-box.

State-of-the-art models for computing true counterfactuals have mostly focused on the *nearest counterfactual instances* principle [Wachter *et al.*, 2017], according to which one finds minimal translations, minimal changes in the features of an instance that lead to a desired outcome. However, this simple distance-based technique often fails to describe faithful alternative worlds, due to the dependence between features. Changing just the sex of a person in such a translation might convert from a typical male into an untypical female rendering true counterfactuals like the following: *if I were a woman I would be 190cm tall and weigh 85 kg*. According to intuition, however, such counterfactuals are false and rightly so because they are oblivious of the latent statistical distribution. As a practical consequence, such counterfactuals typically hide biases in machine learning decision rules [Besse *et al.*, 2020].

The intuitive link between counterfactual modality and causality motivated the use of Pearl's causal graphs and structural equations [Pearl, 2009] to address the aforementioned shortcoming [Kusner *et al.*, 2017; Joshi *et al.*, 2019; Karimi *et al.*, 2020b; Mahajan *et al.*, 2020]. Causal models capture the structural relations between variables including their dependencies and as such provide the basis for generating true *structural counterfactuals*. The cost of this approach is specifying the causal model. The reliance on such a strong prior makes the causal approach appealing in theory, but limited for systematic implementation. In addition, it's not how we humans evaluate counterfactuals. Typically, we don't know the causal graph for a given situation (and we're bad at constructing them); but we have strong intuitions on alternative states of things. Intuitively, the counterfactual female counterpart of a 190cm man would not be a 190cm woman, but more more likely a shorter woman, fairly tall compared to her gender-group. Our contribution offers a mathematical theory of this intuition based on *optimal transport*.

[Black *et al.*, 2020] first suggested substituting causal reasoning with optimal transport but didn't justify this theoretically. We do this here. Optimal transport answers the counterfactual question *had the man been a woman, how tall would have she been?* by minimizing in average a cost between all the paired instances. Interestingly, optimal transport has been used to generalize the notion of distribution function to higher dimensions [**?**], and thus provide a statistically-faithful notion of counterpart. In addition, it recovers the causal relations in many scenarios: as our principal theoretical result, we prove that the optimal transport map for the squared euclidean cost generates the same alternative states as a large class of linear causal models.

We will introduce the *mass transportation* viewpoint of counterfactual models, with which we will connect causal-based methods with optimal-transport-based methods. First, we reformulate the structural counterfactual approach as a

problem of finding distributional correspondences, and provide a closed-form for this operation under the *single-world* assumption. On the basis of this reformulation, we introduce a general causality-free framework for the computation of counterfactuals through mass transportation techniques—e.g., optimal transport. This sheds new light on how to represent counterfactual operations, offers new perspectives to explain black-box decision rules, and recasts attractive causal-based specifications for counterfactuals into more practicable criteria.

Related research falls into two categories: work that represents counterfactual interventions as operators through causal modeling [Plecko and Meinshausen, 2020; Karimi *et al.*, 2020b], and work that moves away from causal-based models by proposing statistically-aware data-based methods [Poyiadzi *et al.*, 2020; Black *et al.*, 2020]. This paper gives a new justification to the latter, by underlining a common structure with the former, and showing that the two may even coincide.

2 Preliminaries

The aim of this section is to detail the mathematical notation and concepts used in the paper. As background for two main topics here, optimal transport and causal reasoning, [Villani, 2003; Villani, 2008] provide supplementary and precise treatments of the first topic; [Schölkopf, 2019; Bongers *et al.*, 2020] do the same for the second.

2.1 Optimal Transport

The mathematical theory of Optimal Transport provides a framework for constructing a joint distribution, namely a *coupling*, between two marginal probability measures. Suppose that each marginal distribution is a sand pile in the ambient space. A coupling is a *mass transport plan* transforming one pile into the other, by specifying how to move each elementary sand mass from the first distribution so as to recover the second distribution. Alternatively, we can see a coupling as a random matching which pairs start points to end points between the respective supports with a certain weight. Optimal transport defines *optimal* transport plans, obtaining a matching by minimizing a cost function between paired instances.

Formally, let P, Q be both probabilities on $\mathbb{R}^d$, whose respective supports are denoted by $\text{supp}(P)$ and $\text{supp}(Q)$, and set a function $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. The *Kantorovich formulation* of the optimal transport problem with cost c is the optimization problem

$$\min_{\pi \in \Pi(P,Q)} \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x,y) d\pi(x,y). \tag{1}$$

$\Pi(P,Q) \subset \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ denotes the set of joint distributions π whose marginals coincide with P and Q respectively, i.e. $\pi(A \times \mathbb{R}^d) = P(A)$ and $\pi(\mathbb{R}^d \times B) = Q(B)$, for all measurable sets $A, B \in \mathbb{R}^d$. Solutions to (1) are optimal transport plans between P and Q with respect to c. They exist under very mild assumptions, like the non-negativeness of the cost.

For $T : \mathbb{R}^d \to \mathbb{R}^d$ a measurable map, we say that T *pushes forward* P to Q if $Q(B) := P(T^{-1}(B))$, for any measurable set $B \subset \mathbb{R}^d$. This property, denoted by $T_\sharp P = Q$, means that if the law of a random variable Z is P, then the law of $T(Z)$ is Q. This push-forward operator T characterizes a *deterministic* coupling between P and Q as every instance $x \in \text{supp}(P)$ is matched to $T(x) \in \text{supp}(Q)$ with probability 1. Suppose now that the cost c is the squared euclidean distance $\|\cdot\|^2$ in $\mathbb{R}^d$, that P is absolutely continuous with respect to the Lebesgue measure in $\mathbb{R}^d$, and that both P and Q have finite second order moments. Theorem 2.12 in [Villani, 2003] states that there exists an unique solution to (1), whose form is $(I \times T)_\sharp P$[1] where I is the identity function on $\mathbb{R}^d$ and $T : \mathbb{R}^d \to \mathbb{R}^d$ is a certain function called the *Brenier map*. Besides, the Brenier map coincides P-almost surely with the gradient of a convex function. Recall that *P-almost surely*, or equivalently *P-almost everywhere*, means that it happens for all $x \in \mathbb{R}^d$ except maybe in a set N such that $P(N) = 0$. Then, in this quadratic case, (1) is equivalent to the following *Monge's formulation*

$$\min_{T:\, T_\sharp P = Q} \int_{\mathbb{R}^d} \|x - T(x)\|^2 dP(x). \tag{2}$$

Thanks to a famous theorem from [McCann, 1995], under the sole assumption that P is absolutely continuous with respect to the Lebesgue measure, there exists only one gradient of a convex function $\nabla\psi$ satisfying the push-forward condition $\nabla\psi_\sharp P = Q$. This simplifies the search for the Brenier map solving (2), as it suffices to find a gradient of a convex function satisfying the push-forward condition.

2.2 Causal reasoning

Causal reasoning relies on a *structural causal model* (SCM) [Pearl, 2009], which represents the causal relationships between variables. More precisely, an *acyclic* structural causal model $\mathcal{M}$ is a triple $\langle U, V, \mathbf{G} \rangle$ where:

1. U and V are two indexed sets of random variables. Abusing notation, we interchangeably consider U and V as sets of random variables and as random vectors;
2. $\mathbf{G} = \{G_i\}_{V_i \in V}$ is a collection of measurable $\mathbb{R}$-valued functions where for every $V_i \in V$, $V_i \overset{a.s.}{=} G_i\big(V_{\text{Endo}(i)}, U_{\text{Exo}(i)}\big)$. The subsets $V_{\text{Endo}(i)} \subset V \setminus \{V_i\}$ and $U_{\text{Exo}(i)} \subset U$ are respectively called the *endogenous* and *exogenous parents* of V_i, and denote the variables that directly determine V_i through G_i.
3. The graph whose nodes are the variables in $U \cup V$, such that an arrow is drawn from some node Z to V_i if and only if $Z \in U_{\text{Exo}(i)} \cup V_{\text{Endo}(i)}$ is a *directed acyclic graph* (DAG);

The equations in 2., the *structural equations*, specify the causal dependencies between the variables. By identifying $\mathbf{G}$ with a measurable vector function, we compactly write: $V \overset{a.s.}{=} \mathbf{G}(V, U)$. A structural causal model can be seen as a generative model. The variables in U are said to be *exogenous*, as their values are imposed on the model by an input probability distribution $\mathcal{L}(U)$. In contrast, the variables in V are said to be *endogenous*, as their values are outputs of the model determined through the structural equations and

[1] This denotes the law of $\big(Z, T(Z)\big)$ where $Z \sim P$.

the values of U. In practice, the endogenous variables represent observed events, while the exogenous ones model latent background phenomena. Note that we don't assume that the endogenous variables are mutually independent.

Crucially, acyclic SCMs are *uniquely solvable*[2], and so the solution V to the structural equations is well-defined. This solution also admits interventional variants under *do-interventions*. A do-intervention consists in substituting a subset of endogenous variables $V_I \subset V$ by fixed values v_I, while keeping all the rest of the causal mechanism equal. This action, denoted by $do(V_I = v_I)$, defines the modified model $\mathcal{M}_{do(V_I=v_I)} = \langle U, V_{V_I=v_I}, \tilde{\mathbf{G}} \rangle$ where $\tilde{\mathbf{G}}$ is given by

$$\tilde{G}_i := \begin{cases} v_i \text{ if } i \in I, \\ G_i \text{ if } i \notin I. \end{cases}$$

As acyclicity is preserved, it follows that the interventional solution $V_{V_I=v_I}$ is well-defined. The exogeneity of the exogenous variables is respected since U is invariant under do-interventions.

2.3 Counterfactual questions

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, and set $d \geq 1$. Define the random vector $V := (X, S) \in \mathbb{R}^{d+1}$, where the variables $X : \Omega \to \mathcal{X} \subset \mathbb{R}^d$ represent some observed features, while the variable $S : \Omega \to \mathcal{S} \subset \mathbb{R}$ can be subjected to interventions. For simplicity, we assume that $\mathcal{S}$ is finite such that for every $s \in \mathcal{S}$, $\mathbb{P}(S = s) > 0$. For every $s \in \mathcal{S}$, set $\mu_s := \mathcal{L}(X|S = s)$ the *factual* or *observational* probability distribution of s-instances, and denote by $\mathcal{X}_s$ its support. We consider the problem of computing the potential outcomes of X when intervening on S. Suppose for instance that the event $\{X = x, S = s\}$ is observed, and set $s' \neq s$. We aim at answering the counterfactual question: *had S been equal to s' instead of s, what would have been the value of X?* Because of structural and statistical correlations between the variables, computing the alternative state does not amount to change the value of S while keeping the features X equal.

3 Structural counterfactuals revisited

Causal reasoning provides a natural framework to address counterfactual questions. We assume that $V = (X, S)$ is the unique solution of an acyclic SCM, which can be defined as a 4-uplet $\mathcal{M} := \langle U, X, S, \mathbf{G} \rangle$, and set for each $s \in S$ the intervened model $\mathcal{M}_{S=s} = \langle U, X_{S=s}, S_{S=s}, \mathbf{G}_{S=s} \rangle$. For clarity, we denote by U_X and U_S the exogenous parents of respectively X and S. In this section, we recall and translate Pearl's causal modeling computation of counterfactuals into a problem of mass transportation. We describe possible instances as probability measures, and interventions as couplings.

3.1 Definitions

As introduced, a counterfactual statement is a *cross-world* statement between a factual outcome and a counterfactual outcome. Let us formalize the contrast between interventional, counterfactual and factual outcomes in terms of probabilistic distributions. For any $s \in \mathcal{S}$ the distribution of the *interventional* s-instances is defined as $\mu_{S=s} := \mathcal{L}(X_{S=s})$, and for any $s' \neq s$ the distribution of the *counterfactual* s'-instances given s is defined as $\mu_{\langle s'|s \rangle} := \mathcal{L}(X_{S=s'}|S = s)$. According to the *consistency rule* [Pearl *et al.*, 2016], for any $s \in \mathcal{S}$, the factual distribution can be written as $\mu_s = \mathcal{L}(X_{S=s}|S = s)$, which is sometimes denoted by $\mu_{\langle s|s \rangle}$ for the sake of coherence. The counterfactual distribution $\mu_{\langle s'|s \rangle}$ describes what would have been the observational instances of μ_s *had S been equal to s' instead of s*; but it does not yield specific cross-world statements on its own, as it does not specify how instances from each distribution are related. The stronger notion of a counterfactual model characterizes all the counterfactual statements w.r.t. S.

The literature proposed various approaches to characterize causality-based counterfactual models. They all concur with the principle that the counterfactual model can be identified with the joint probability distributions between observable instances and intervened counterparts, as generated by the structural equations [Imbens and Rubin, 2015; Pearl *et al.*, 2016; Bongers *et al.*, 2020]. We follow [Pearl *et al.*, 2016; Kusner *et al.*, 2017] and propose a formalization of this definition that takes into account the observed value of S before intervening on it.

Definition 1. *For every $s, s' \in \mathcal{S}$, the* structural counterfactual coupling *between μ_s and $\mu_{\langle s'|s \rangle}$ is given by*

$$\pi^*_{\langle s'|s \rangle} := \mathcal{L}\big((X, X_{S=s'})|S = s\big).$$

We call the collection of couplings $\Pi^* := \{\pi^*_{\langle s'|s \rangle}\}_{s,s' \in \mathcal{S}}$ the structural counterfactual model.

It is worth noting that, in general, the structural counterfactual couplings are *random*, because X and $X_{S=s}$ are entangled through U following a certain probability distribution. This means that, according to Pearl's causal reasoning, there is not necessarily a one-to-one deterministic correspondence between factual instances and counterfactual counterparts, but a collection of weighted correspondences described by the structural couplings. To understand how the latent SCM generates such couplings, one must address the construction of the counterfactual distributions at the individual level. A *counterfactual instance* represents a possible alternative state of X, with respect to an action on S and an observed evidence of (X, S). The following definition defines a counterfactual as a distribution rather than a random variable as in [Pearl *et al.*, 2016].

Definition 2. *For an observed evidence $\{X = x, S = s\}$ and an intervention $do(S = s')$, the* structural counterfactuals *of X are characterized by the probability distribution $\mu_{\langle s'|s \rangle}(\cdot|x)$ defined as*

$$\mu_{\langle s'|s \rangle}(\cdot|x) := \mathcal{L}(X_{S=s'}|X = x, S = s).$$

The possible outcomes $\mu_{\langle s'|s \rangle}(\cdot|x)$ are commonly generated with the so-called *three-step* procedure [Pearl *et al.*, 2016], which amounts to: (1) setting a prior $\mathcal{L}(U)$ for the model $\mathcal{M}$, (2) computing the posterior distribution $\mathcal{L}(U|X =$

[2]Rigorously, the solution is unique up to sets of probability zero w.r.t. the latent probability space.

$x, S = s$), and (3) solving the structural equations of $\mathcal{M}_{S=s'}$ with $\mathcal{L}(U|X = x, S = s)$. As anticipated, the counterfactuals of an instance are not necessarily *deterministic*, i.e. characterized by a degenerate distribution, but belong to a set of possible outcomes. This is due to the fact that, in general, there are several values of U consistent with an evidence $\{X = x, S = s\}$. Note that, equivalently to Definition 1, Definition 2 characterizes the counterfactual semantics. In particular, the *disintegrated* formulation $\mu_{\langle s'|s\rangle} = \int \mu_{\langle s'|s\rangle}(\cdot|x)d\mu_s(x)$ shows how μ_s relates to the counterfactual distribution through $\mu_{\langle s'|s\rangle}(\cdot|x)$.

To sum-up, we have shown how to see a counterfactual coupling $\pi^*_{\langle s'|s\rangle}$ as a transport plan between an observed world and an alternative world, where all the elementary correspondences are given by the structural counterfactuals $\{\mu_{\langle s'|s\rangle}(\cdot|x)\}_{x\in\mathcal{X}_s}$. In what follows, we study, from the mass transportation perspective, two specific scenarios mitigating the involvement of SCMs when computing counterfactuals: first, when the correspondences are deterministic— then the computation can be written as an explicit push-forward operation; second, when S can be considered exogenous—then the alternative world is observable.

3.2 The deterministic case

Interestingly, when the SCM entails that the structural counterfactuals for each antecedent (or instance) determine a unique counterfactual possibility, then the counterfactual coupling is deterministic, and can be identified with a push-forward operator. To reformulate structural counterfactuals in deterministic transport terms, we first highlight the relation between an individual and its intervened counterparts.

From the acyclicity of the causal model, we can recursively substitute for the X_i their functional form to obtain a measurable function $\mathbf{F}$ such that $\mathbb{P}$-almost surely $X = \mathbf{F}(S, U_X)$ and $X_{S=s} = \mathbf{F}(s, U_X)$ for any $s \in \mathcal{S}$. Now, let us define for every $s \in \mathcal{S}$ the function $f_s : u \mapsto \mathbf{F}(s, u)$. The next proposition specifies the range of the possible outcomes.

Proposition 1. *For any $s, s' \in \mathcal{S}$, $x \in \mathcal{X}_s$,*

$$supp\big(\mu_{\langle s'|s\rangle}(\cdot|x)\big) \subset f_{s'} \circ f_s^{-1}(\{x\}).$$

For any $x \in \mathbb{R}^d$, we denote by δ_x the distribution assigning a probability 1 to this single instance, which is called the Dirac at x. Proposition 1 entails that the structural counterfactuals determine a unique counterpart, and thus the set of weighted counterfactual possibilities becomes a Dirac, if the following *single-world* assumption holds:[3]

Assumption (SW) *The functions $\{f_s\}_{s\in\mathcal{S}}$ are injective.*

While the unique solvability of acyclic models ensures that (X, S) is completely determined by U, **(SW)** states that, conversely, U_X is determined by $\{X = x, S = s\}$. This implies that the coupling between the factual and counterfactual distributions is deterministic.

Proposition 2. *Let **(SW)** hold, and define for any $s, s' \in \mathcal{S}$, $T^*_{\langle s'|s\rangle} := f_{s'} \circ f_s^{-1}|_{\mathcal{X}_s}$[4]. The following properties hold:*

1. *$\mu_{\langle s'|s\rangle}(\cdot|x) = \delta_{T^*_{\langle s'|s\rangle}(x)}$ for every $x \in \mathcal{X}_s$;*
2. *$\mu_{\langle s'|s\rangle} = T^*_{\langle s'|s\rangle\sharp}\mu_s$;*
3. *$\pi^*_{\langle s'|s\rangle} = (I \times T^*_{\langle s'|s\rangle})\sharp\mu_s$.*

We say that $T^*_{\langle s'|s\rangle}$ is a structural counterfactual operator*, and identify $\mathcal{T}^* := \{T^*_{\langle s'|s\rangle}\}_{s,s'\in\mathcal{S}}$ to the structural counterfactual model Π^*.*

The operators in $\mathcal{T}^*$ describe the effect of causal interventions on factual distributions, without assuming any knowledge of $\mathcal{L}(U)$.

3.3 The exogenous case

Let $\perp\!\!\!\perp$ denote the independence between random variables. The variable S is said to be *exogenous relative to X* [Galles and Pearl, 1998] if the following holds:

Assumption (RE) *$U_S \perp\!\!\!\perp U_X$ and $X_{Endo(S)} = \emptyset$.*

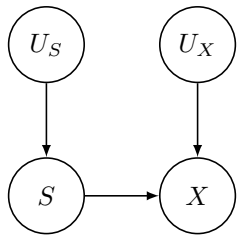

Figure 1: DAG satisfying **(RE)**

This represents a scenario where: (1) there is no hidden confounder between X and S, (2) no variable in X is a direct cause of S. Note that **(RE)** entails that $S \perp\!\!\!\perp U_X$. Then, it is easy to see that at the distributional level, intervening on S amounts to conditioning X by a value of S.

Proposition 3. *If **(RE)** holds, then for every $s, s' \in \mathcal{S}$ we have $\mu_{S=s'} = \mu_{s'} = \mu_{\langle s'|s\rangle}$.*

Relative exogeneity is a critical assumption. Recall that the structural counterfactual coupling $\pi^*_{\langle s'|s\rangle}$ represents an intervention transforming an observable distribution μ_s into an *a priori* non-observable counterfactual distribution $\mu_{\langle s'|s\rangle}$. According to Proposition 3, **(RE)** renders the causal model otiose for the purpose of generating the counterfactual distributions, as the latter coincides with the observable factual distribution $\mu_{s'}$. However, the coupling is *still required* to determine how each instance is matched at the individual level. Remarkably, **(RE)** provides elegant transitivity properties to our counterfactual operators.

Proposition 4. *Suppose that **(RE)** and **(SW)** hold. Then, for any $s, s', s'' \in \mathcal{S}$:*

1. *The operator $T^*_{\langle s'|s\rangle}$ is invertible, such that $\mu_{s'}$-almost everywhere $T^{*-1}_{\langle s'|s\rangle} = T^*_{\langle s|s'\rangle}$;*
2. *μ_s-almost everywhere, $T^*_{\langle s''|s'\rangle} \circ T^*_{\langle s'|s\rangle} = T^*_{\langle s''|s\rangle}$.*

In terms of real-world modeling, **(RE)** is intuitively satisfied in many scenarios. Let X represent the socio-economics

[3] This assumption corresponds to the logical constraint of conditional excluded middle [Stalnaker, 1980].

[4] $f_s^{-1}|_{\mathcal{X}_s}$ denotes the restriction of f_s^{-1} to $\mathcal{X}_s$.

features of individuals, and suppose for example that $\mathcal{S} = \{0, 1\}$, where $S = 0$ stands for *female* while $S = 1$ stands for *male*. In this presumably exogenous model, any factual woman described by x is the counterfactual counterpart of her counterfactual male counterpart described by $T^*_{\langle 1|0\rangle}(x)$, and changing all the factual women into their counterfactual male counterparts recovers the factual male population.

We conclude Section 3 by illustrating how our notation and assumptions apply to the case of *linear additive* structural models, which account for most of the state-of-the-art models.

Example 1. *Under **(RE)**, a linear additive SCM is characterized by the structural equations*

$$X = MX + wS + b + U_X,$$

*where $w, b \in \mathbb{R}^d$ and $M \in \mathbb{R}^{d\times d}$ are deterministic parameters. Acyclicity implies that $I - M$ is invertible, so that $X = (I-M)^{-1}(wS+b+U_X) =: \mathbf{F}(S, U_X)$. Note that **(SW)** holds such that for any $s \in S$, $f_s^{-1}(x) = (I-M)x - ws - b$. Then, for any $s, s' \in \mathcal{S}$, $T^*_{\langle s'|s\rangle} = x + (I - M)^{-1}w(s' - s)$.*

The transport viewpoint of structural counterfactual reasoning suggests that transport-based method can be natural substitutes for causal modeling, a topic we explore next.

4 Transport-based counterfactuals

[Black *et al.*, 2020] mimicked the structural account of counterfactuals by computing alternative individuals using a deterministic optimal transport map, but they did not provide a mathematical or conceptual foundation for their idea. **(SW)** and **(RE)** imply that approximating an unknown structural counterfactual model with deterministic couplings between observed data is a reasonable method. Generalizing their idea, we propose a general framework for transport-based counterfactual models that leads us to practicable SCM-free frameworks.

Definition 3. *1. A* counterfactual model *is a collection $\Pi := \{\pi_{\langle s'|s\rangle}\}_{s,s'\in\mathcal{S}}$ of couplings on $\mathcal{X} \times \mathcal{X}$ such that for any $s, s' \in \mathcal{S}$, the first marginal of $\pi_{\langle s'|s\rangle}$ is μ_s, the second marginal is $\mu_{s'}$, and $\pi_{\langle s|s\rangle} = (I \times I)_\sharp \mu_s$. An element of Π is called a* counterfactual coupling. *We say that Π is a* random counterfactual model *if at least one coupling for $s \neq s'$ is not deterministic.*

2. A deterministic counterfactual model *is a collection $\mathcal{T} := \{T_{\langle s'|s\rangle}\}_{s,s'\in\mathcal{S}}$ of mappings from $\mathcal{X}$ to $\mathcal{X}$ satisfying for any $s, s' \in \mathcal{S}$, $T_{\langle s'|s\rangle\sharp}\mu_s = \mu_{s'}$ and $T_{\langle s|s\rangle} = I$. An element of $\mathcal{T}$ is called a* counterfactual operator.

One challenge for this approach is to choose the model appropriately in order to define a relevant notion of counterpart. Even though the family of trivial couplings is a well-defined counterfactual model, it is not intuitively justifiable. Better suited counterfactual models can be constructed through optimal transport theory. Optimal transport with the squared euclidean cost is known to preserve quantiles in dimension one, and has been used to generalize the notion of distribution function to higher dimensions [**?**]. In this sense, it satisfies our statistical intuitions on counterfactual reasoning. In addition, if the factual distributions are absolutely continuous w.r.t. the Lebesgue measure, then for any $s, s' \in \mathcal{S}$, the Brenier map between μ_s and $\mu_{s'}$ is the unique counterfactual operator that can be written as the gradient of a convex function. As the structural counterfactual operator from Example 1 is the gradient of a convex quadratic function, we obtain the following result.

Theorem 1. *Let $\mathcal{M}$ be a linear additive SCM satisfying **(RE)** (see Example 1). If the factual distributions are absolutely continuous w.r.t. Lebesgue measure, then for any $s, s' \in \mathcal{S}$, the structural counterfactual operator $T^*_{\langle s'|s\rangle}$ is the Brenier map between μ_s and $\mu_{s'}$.*

Whether or not elements of the structural counterfactual model Π^* are solutions to a Kantorovich or Monge problem for a certain cost function is presumably difficult to prove for more complex SCMs. Theorem 1 supports the intuition that substituting Π^* with a surrogate Π from optimal transport provides a decent approximation of the do-calculus. Using a model close to Π^* would be ideal in terms of interpretability of a decision-making process, but an expert can always propose and defend a different notion of similarity Π.

The computational complexity of building an optimal transport plan between a n-sample to a m-sample is in $\mathcal{O}\big((n + m)nm \log(n + m)\big)$, but we can substantially improve on this to reach $\mathcal{O}\big(nm\big)$ with entropy-regularized versions [Cuturi, 2013]. As the computation is distribution-wise, not point-wise, it yields all the cross-world or counterfactual statements corresponding to a given change $s \to s'$ for the considered data-points. In contrast, computing a structural counterfactual coupling is less convenient and more challenging. First, inferring the causal graph from observational data is NP-hard, with an exponential worst-case complexity with respect to the number of nodes [Cooper, 1990; Chickering *et al.*, 2004; Scutari *et al.*, 2019]. Second, this is not enough to compute counterfactuals, as we must still specify the structural equations. Third, even though the three-step procedure generates samples from the structural counterfactuals of a given instance through a specified SCM [Perov *et al.*, 2020], it needs to be applied at each point in order to infer the whole coupling.

5 Applications

In this section we look at two applications of transport based counterfactuals—explicability or interpretability and fairness of a *black-box* algorithm. Counterfactuals have been used already in both areas [Wachter *et al.*, 2017; Kusner *et al.*, 2017; Karimi *et al.*, 2020a]. [Karimi *et al.*, 2020a] exploit automated reasoning based methods to find counterfactuals that can explain program behavior, and these methods have computational complexity problems given that they must test for satisfiability or unsatisfiability that is at least NP hard (depending on the logic fragment used). On the other hand a transport based method by aligning two entire probability distributions can provide a set of explanatory counterfactuals that mimic the causal approach in polynomial time, which means that the transport based approach can apply to the interpretability of programs for which a SAT based approach is not practically possible. In addition, the transport based

method capturing as it does in some cases the causal structure of the phenomenon has a firmer conceptual basis as an explanatory tool than standard heuristically guided approaches relying either on local approximation by simpler linear models as LIME [Ribeiro *et al.*, 2016] or relying on the computation of indices measuring the contribution of each variable and its importance as in SHAP [Lundberg and Lee, 2017] or in [Bachoc *et al.*, 2020].

We now turn to the fairness application. Suppose that the random variable S encodes the observed *sensitive* or *protected attribute* (e.g., race, gender) which divides the population into different classes in a machine learning prediction task. The counterfactual framework, by capturing the structural or statistical links between the features and the protected attribute, proposes sharper notions of fairness than *statistical parity*, which only gives a notion of *group fairness*, and does not control discrimination at a subgroup or an individual level: a conflict illustrated by [Dwork *et al.*, 2012]. We first use the mass transportation formalism introduced in Section 3 to reformulate the *counterfactual fairness* [Kusner *et al.*, 2017] condition, which is achieved when individuals and their structural counterfactual counterparts are treated equally.

Definition 4. *A predictor $\hat{Y} = h(X, S)$ is* counterfactually fair *if for every $s, s' \in \mathcal{S}$ and μ_s-almost every x,*

$$\mathcal{L}(\hat{Y}_{S=s}|X = x, S = s) = \mathcal{L}(\hat{Y}_{S=s'}|X = x, S = s),$$

where $\hat{Y}_{S=s} := h(X_{S=s}, s)$.

For each individual, this condition guarantees the truth of the counterfactual statement *had the protected attribute been changed, the outcome would have been the same*. The structural counterfactual transport plans allow for simpler characterizations of counterfactual fairness.

Proposition 5. *1. A predictor $h(X, S)$ is counterfactually fair if and only if for every $s, s' \in \mathcal{S}$ and $\pi^*_{\langle s'|s\rangle}$-almost every (x, x'),*

$$h(x, s) = h(x', s').$$

*2. If **(SW)** holds, then a predictor $h(X, S)$ is counterfactually fair if and only if for every $s, s' \in \mathcal{S}$ and μ_s-almost every x,*

$$h(x, s) = h(T^*_{\langle s'|s\rangle}(x), s').$$

*3. If **(SW)** and **(RE)** hold, then a predictor $h(X, S)$ is counterfactually fair if and only if for every $s, s' \in \mathcal{S}$ such that $s < s'$ and μ_s-almost every x,*

$$h(x, s) = h(T^*_{\langle s'|s\rangle}(x), s').$$

The condition **(SW)** has two main advantages in terms of clarity and practicability of the formulation. First, it highlights the clear relationship between factual and counterfactual quantities. Second, testing counterfactual fairness requires only the knowledge of the structural equations, but not the one of $\mathcal{L}(U)$. Note that, if **(RE)** holds, then counterfactual fairness is a stronger criterion than the statistical parity across groups.

Proposition 6. *Suppose that **(RE)** holds. If the predictor $h(X, S)$ satisfies counterfactual fairness, then it satisfies statistical parity, namely $h(X, S) \perp\!\!\!\perp S$. The converse does not hold in general.*

One can think of being counterfactually fair as being invariant by counterfactual operations w.r.t. the protected attribute. In order to define SCM-free criteria, we generalize this idea to the models introduced in Section 4.

Definition 5. *1. Let $\Pi = \{\pi_{\langle s'|s\rangle}\}_{s,s' \in \mathcal{S}}$ be a random counterfactual model. A predictor $h(X, S)$ is* Π-counterfactually fair *if for every $s, s' \in \mathcal{S}$ and $\pi_{\langle s'|s\rangle}$-almost every (x, x'),*

$$h(x, s) = h(x', s').$$

2. Let $\mathcal{T} = \{T_{\langle s'|s\rangle}\}_{s,s' \in \mathcal{S}}$ be a deterministic counterfactual model. A predictor $h(X, S)$ is $\mathcal{T}$-counterfactually fair *if for every $s, s' \in \mathcal{S}$ and μ_s-almost every x,*

$$h(x, s) = h(T_{\langle s'|s\rangle}(x), s').$$

Because the proof of Proposition 6 only relies on the assumption that the couplings are transport plans between the factual distributions, the following proposition holds.

Proposition 7. *Let Π be a counterfactual model (deterministic or not). If a predictor $h(X, S)$ satisfies Π-counterfactual fairness, then it satisfies statistical parity, namely $h(X, S) \perp\!\!\!\perp S$. The converse does not hold in general.*

Using Definition 5 as an individual-level fairness criterion has several practical advantages. In contrast to Definitions 4 and Proposition 5, it relies on a well-defined counterfactual model that obviates any assumptions about the causal model. This alternative approach to counterfactual fairness alleviates the impracticability of causal reasoning, trading the detection of structural links between variables for the discovery of statistical correlations. Besides, as Definition 4 amounts to Π^*-counterfactual fairness when **(RE)** holds, one can think of Definition 5 as an approximation of counterfactual fairness.

6 Conclusion

We focused on the challenge of designing sound counterfactuals when the causal model is unknown. We framed the computation of counterfactuals through causal models as a problem of mass transportation, and studied two key scenarios of counterfactual reasoning through this viewpoint. On the basis of this reformulation, we introduced a general formalism for the computation of counterfactual counterparts based on any distributional matching technique. In particular, we showed that optimal transport defines relevant counterfactual models, as it is tailored for numerical implementation, satisfies statistical intuitions, and can even recover the structural dependencies of linear additive SCMs. On the strength of this alternative counterfactual modeling, we proposed original counterfactual fairness conditions, free of prior assumptions on the data-generation process. This offered new conceptual and practical perspectives for counterfactual reasoning.

References

[Bachoc *et al.*, 2020] François Bachoc, Fabrice Gamboa, Max Halford, Jean-Michel Loubes, and Laurent Risser. Explaining machine learning models using entropic variable projection, 2020.

[Besse *et al.*, 2020] Philippe Besse, Eustasio del Barrio, Paula Gordaliza, Jean-Michel Loubes, and Laurent Risser. A survey of bias in machine learning through the prism of statistical parity for the adult data set. *arXiv preprint arXiv:2003.14263*, 2020.

[Black *et al.*, 2020] Emily Black, Samuel Yeom, and Matt Fredrikson. Fliptest: Fairness testing via optimal transport. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 111–121, New York, NY, USA, 2020. Association for Computing Machinery.

[Bongers *et al.*, 2020] Stephan Bongers, Patrick Forré, Jonas Peters, Bernhard Schölkopf, and Joris M Mooij. Foundations of structural causal models with cycles and latent variables. *arXiv preprint arXiv:1611.06221*, 2020.

[Chickering *et al.*, 2004] David Maxwell Chickering, David Heckerman, and Christopher Meek. Large-sample learning of bayesian networks is np-hard. *Journal of Machine Learning Research*, 5(Oct):1287–1330, 2004.

[Cooper, 1990] Gregory F. Cooper. The computational complexity of probabilistic inference using bayesian belief networks (research note). *Artif. Intell.*, 42(2–3):393–405, March 1990.

[Cuturi, 2013] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26:2292–2300, 2013.

[Dwork *et al.*, 2012] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[Galles and Pearl, 1998] David Galles and Judea Pearl. An axiomatic characterization of causal counterfactuals. *Foundations of Science*, 3(1):151–182, 1998.

[Imbens and Rubin, 2015] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[Joshi *et al.*, 2019] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. Towards realistic individual recourse and actionable explanations in black-box decision making systems, 2019.

[Karimi *et al.*, 2020a] Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. Model-agnostic counterfactual explanations for consequential decisions. In *International Conference on Artificial Intelligence and Statistics*, pages 895–905. PMLR, 2020.

[Karimi *et al.*, 2020b] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: from counterfactual explanations to interventions. *arXiv preprint arXiv:2002.06278*, 2020.

[Kusner *et al.*, 2017] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 4066–4076. Curran Associates, Inc., 2017.

[Lewis, 1973] David Lewis. Causation. *Journal of Philosophy*, 70(17):556–567, 1973.

[Lundberg and Lee, 2017] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 4765–4774. Curran Associates, Inc., 2017.

[Mahajan *et al.*, 2020] Divyat Mahajan, Chenhao Tan, and Amit Sharma. Preserving causal constraints in counterfactual explanations for machine learning classifiers, 2020.

[McCann, 1995] Robert J. McCann. Existence and uniqueness of monotone measure-preserving maps. *Duke Math. J.*, 80(2):309–323, 11 1995.

[Pearl *et al.*, 2016] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[Pearl, 2009] Judea Pearl. *Causality*. Cambridge university press, 2009.

[Perov *et al.*, 2020] Yura Perov, Logan Graham, Kostis Gourgoulias, Jonathan Richens, Ciaran Lee, Adam Baker, and Saurabh Johri. Multiverse: Causal reasoning using importance sampling in probabilistic programming. In Cheng Zhang, Francisco Ruiz, Thang Bui, Adji Bousso Dieng, and Dawen Liang, editors, *Proceedings of The 2nd Symposium on Advances in Approximate Bayesian Inference*, volume 118 of *Proceedings of Machine Learning Research*, pages 1–36. PMLR, 08 Dec 2020.

[Plecko and Meinshausen, 2020] Drago Plecko and Nicolai Meinshausen. Fair data adaptation with quantile preservation. *Journal of Machine Learning Research*, 21:1–44, 2020.

[Poyiadzi *et al.*, 2020] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. Face: feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 344–350, 2020.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[Schölkopf, 2019] Bernhard Schölkopf. Causality for machine learning, 2019.

[Scutari *et al.*, 2019] Marco Scutari, Claudia Vitolo, and Allan Tucker. Learning bayesian networks from big data with greedy search: computational complexity and efficient implementation. *Statistics and Computing*, 29(5):1095–1108, 2019.

[Stalnaker, 1980] Robert C Stalnaker. A defense of conditional excluded middle. In *Ifs*, pages 87–104. Springer, 1980.

[Villani, 2003] Cédric Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

[Villani, 2008] Cédric Villani. *Optimal Transport: Old and New*. Number 338 in Grundlehren der mathematischen Wissenschaften. Springer, Berlin, 2008. OCLC: ocn244421231.

[Wachter *et al.*, 2017] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

This supplementary material addresses the mathematical proofs of the paper.

A Lemmas

We start by proving two key results we mentioned in Section 3. The first one specifies formulas for X and its interventional variants.

Lemma 1. *There exists a measurable function* $\mathbf{F}$ *such that* $\mathbb{P}$*-almost surely* $X = \mathbf{F}(S, U_X)$ *and* $X_{S=s} = \mathbf{F}(s, U_X)$ *for any* $s \in \mathcal{S}$.

Proof. Recall that, rigorously, the structural equations hold almost surely. Throughout this proof, we implicitly work with a fixed input ω for the random variables, where ω belongs to some measurable set $\Omega_0 \subset \Omega$ such that $\mathbb{P}(\Omega_0) = 1$ and

$$X_i = G_{X_i}\big(X_{\text{Endo}(X_i)}, S_{\text{Endo}(X_i)}, U_{X_i}\big),$$
$$S = G_S\big(X_{\text{Endo}(S)}, U_S\big).$$

Because the graph of $\mathcal{M}$ is a DAG, it has a topological ordering on the variables in X. Then, we can recursively substitute the X_i according to this ordering to obtain

$$X = \tilde{\mathbf{F}}\big(S_{\text{Endo}(X)}, U_X\big),$$

where $\tilde{\mathbf{F}}$ is a measurable function. Remark that either $S_{\text{Endo}(X)} = \{S\}$ or $S_{\text{Endo}(X)} = \emptyset$, depending on whether S is a parent of X in the graph. Then, without loss of generality, we can define $\mathbf{F}$ such that $\mathbf{F}(S, U_X) := \tilde{\mathbf{F}}\big(S_{\text{Endo}(X)}, U_X\big)$. Consequently, $X = \mathbf{F}(S, U_X)$. Now, recall that $do(S = s)$ preserves the structural equations of X, and does not impact U. Then, using the exact same procedure for $(X_{S=s}, S_{S=s})$ instead of (X, S) we get $X_{S=s} = \mathbf{F}(S_{S=s}, U_X) = \mathbf{F}(s, U_X)$.

□

The second result is the consistency rule.

Lemma 2. *For any* $s \in \mathcal{S}$, $\mu_{\langle s|s\rangle} = \mu_s$

Proof. From Lemma 1, $\mathbb{P}$-almost surely $X = \mathbf{F}(S, U_X)$ and $X_{S=s} = \mathbf{F}(s, U_X)$ for any $s \in \mathcal{S}$. Then,

$$\begin{aligned}\mu_s &= \mathcal{L}(X|S = s)\\ &= \mathcal{L}(\mathbf{F}(S, U_X)|S = s)\\ &= \mathcal{L}(\mathbf{F}(s, U_X)|S = s),\end{aligned}$$

and

$$\begin{aligned}\mu_{\langle s|s\rangle} &= \mathcal{L}(X_{S=s}|S = s)\\ &= \mathcal{L}(\mathbf{F}(s, U_X)|S = s).\end{aligned}$$

Consequently, $\mu_s = \mu_{\langle s|s\rangle}$.

□

B Proofs of Section 3

Proof of Proposition 1.

Proof. According to Lemma 1 we can write that $X = \mathbf{F}(S, U_X)$ $\mathbb{P}$-almost surely. This implies that $\{X = x, S = s\} \subset \{U_X \in f_s^{-1}(\{x\})\}$. Besides, $X_{S=s'} = f_{s'}(U_X)$. Then, write for B an arbitrary measurable set of $\mathcal{X}$

$$\begin{aligned}&\mathbb{P}(X_{S=s'} \in B|X = x, S = s)\\ &= \mathbb{P}(f_{s'}(U_X) \in B|X = x, S = s)\\ &= \mathbb{P}(f_{s'}(U_X) \in B, U_X \in f_s^{-1}(\{x\})|X = x, S = s)\\ &= \mathbb{P}(f_{s'}(U_X) \in B, f_{s'}(U_X) \in f_{s'} \circ f_s^{-1}(\{x\})|X = x, S = s)\\ &= \mathbb{P}(X_{S=s'} \in \big[B \cap f_{s'} \circ f_s^{-1}(\{x\})\big]|X = x, S = s).\end{aligned}$$

Consequently, $\mathcal{L}(X_{S=s'}|X = x, S = s)$ does not put mass outside $f_{s'} \circ f_s^{-1}(\{x\})$. □

Proof of Proposition 2.

Proof. Let $s, s' \in \mathcal{S}$ and $x \in \mathcal{X}_s$. From Lemma 1 we know that $X = f_S(U_X)$, and according to **(SW)** we additionally have $U_X = f_S^{-1}(X)$. We address each point separately.

Proof of 1. By definition of the structural counterfactuals,

$$\begin{aligned}\mathcal{L}(X_{S=s'}|X = x, S = s) &= \mathcal{L}(f_{s'}(U_X)|X = x, S = s)\\ &= \mathcal{L}(f_{s'}(f_S^{-1}(X))|X = x, S = s)\\ &= \mathcal{L}(f_{s'} \circ f_s^{-1}(x)|X = x, S = s)\\ &= \mathcal{L}(f_{s'} \circ f_s^{-1}(x))\\ &= \delta_{f_{s'} \circ f_s^{-1}(x)}.\end{aligned}$$

This proves the first point of the proof.

Proof of 2. By definition of the counterfactual distribution,

$$\begin{aligned}\mu_{\langle s'|s\rangle} &= \mathcal{L}(X_{S=s'}|S = s)\\ &= \mathcal{L}(f_{s'}(U_X)|S = s)\\ &= \mathcal{L}(f_{s'} \circ f_S^{-1}(X)|S = s)\\ &= \mathcal{L}(f_{s'} \circ f_s^{-1}(X)|S = s)\\ &= \big(f_{s'} \circ f_s^{-1}\big)_\sharp \mu_s.\end{aligned}$$

This proves the second point of the proposition.

Proof of 3. By definition of the structural counterfactual coupling,

$$\begin{aligned}\pi_{\langle s'|s\rangle} &= \mathcal{L}\big((X, X_{S=s'})|S = s\big)\\ &= \mathcal{L}\big((X, f_{s'}(U_X))|S = s\big)\\ &= \mathcal{L}\big((X, f_{s'}(f_s^{-1}(X)))|S = s\big)\\ &= \mathcal{L}\big((X_s, f_{s'} \circ f_s^{-1}(X_s))\big),\end{aligned}$$

where $X_s \sim \mu_s$. This concludes the proof.

□

Proof of Proposition 3.

Proof. To show this, set $s \in \mathcal{S}$ and invoke Lemma 1 once again to write $X = \mathbf{F}(S, U_X)$ and $X_{S=s} = \mathbf{F}(s, U_X)$. Recall that **(RE)** implies that $S \perp\!\!\!\perp U_X$. Then,

$$\begin{aligned}\mathcal{L}(X|S=s) &= \mathcal{L}\big(\mathbf{F}(S,U_X)|S=s\big),\\ &= \mathcal{L}\big(\mathbf{F}(s,U_X)|S=s\big),\\ &= \mathcal{L}\big(\mathbf{F}(s,U_X)\big),\\ &= \mathcal{L}(X_{S=s}).\end{aligned}$$

This means that $\mu_s = \mu_{S=s}$. Similarly, for $s, s' \in \mathcal{S}$ we have

$$\begin{aligned}\mathcal{L}(X_{S=s'}|S=s) &= \mathcal{L}\big(\mathbf{F}(s',U_X)|S=s\big),\\ &= \mathcal{L}\big(\mathbf{F}(s',U_X)\big),\\ &= \mathcal{L}\big(\mathbf{F}(s',U_X)|S=s'\big),\\ &= \mathcal{L}\big(\mathbf{F}(S,U_X)|S=s'\big),\\ &= \mathcal{L}(X|S=s').\end{aligned}$$

This means that $\mu_{\langle s'|s\rangle} = \mu_{s'}$, which concludes the proof. □

Proof of Proposition 4.

Proof. We address each point separately.

Proof of 1. Set $s, s' \in \mathcal{S}$. By definition $T^*_{\langle s'|s\rangle} = f_{s'} \circ f_s^{-1}|_{\mathcal{X}_s}$, which induces a bijection from $\mathcal{X}_s$ to $\mathrm{Im}(T^*_{\langle s'|s\rangle})$. Let us denote $\mathrm{Im}(T^*_{\langle s'|s\rangle})$ by $\mathcal{X}_{\langle s'|s\rangle}$, so that ${T^*}^{-1}_{\langle s'|s\rangle} = f_s \circ f_{s'}^{-1}|_{\mathcal{X}_{\langle s'|s\rangle}}$.

Now, recall that $\mathbb{P}$-almost surely $X_{S=s} = f_s(U_X)$ and $X_{S=s'} = f_{s'}(U_X)$. Besides, from **(RE)** and Proposition 3, it follows that $\mu_s = \mathcal{L}(X_{S=s})$ and $\mu_{s'} = \mathcal{L}(X_{S=s'})$. This implies that there exists a measurable set $\Omega_0 \subset \Omega$ such that for every $\omega \in \Omega_0$,

$$\begin{aligned}X_{S=s}(\omega) &= f_s(U_X(\omega)) \subset \mathcal{X}_s,\\ X_{S=s'}(\omega) &= f_{s'}(U_X(\omega)) \subset \mathcal{X}_{s'}.\end{aligned}$$

In the rest of the proof, we implicitely work with an arbitraty $\omega \in \Omega_0$. Write $U_X = f_s^{-1}(X_{S=s})$ so that $X_{S=s'} = (f_{s'} \circ f_s^{-1})(X_{S=s})$. Since $X_{S=s} \in \mathcal{X}_s$, this leads to $X_{S=s'} = (f_{s'} \circ f_s^{-1}|_{\mathcal{X}_s})(X_{S=s}) = T^*_{\langle s'|s\rangle}(X_{S=s})$, and consequently $X_{S=s'} \in \mathcal{X}_{\langle s'|s\rangle}$. Then, we can apply ${T^*}^{-1}_{\langle s'|s\rangle}$ on $X_{S=s'}$ to obtain

$$\begin{aligned}{T^*}^{-1}_{\langle s'|s\rangle}(X_{S=s'}) &= f_s \circ f_{s'}^{-1}|_{\mathcal{X}_{\langle s'|s\rangle}}(X_{S=s'})\\ &= f_s \circ f_{s'}^{-1}|_{\mathcal{X}_{s'}}(X_{S=s'})\\ &= {T^*}_{\langle s|s'\rangle}(X_{S=s'}).\end{aligned}$$

This means that the equality ${T^*}^{-1}_{\langle s'|s\rangle} = {T^*}_{\langle s|s'\rangle}$ holds on $X_{S=s'}(\Omega_0)$ where $\mathbb{P}(\Omega_0) = 1$. Thus, it holds $\mu_{s'}$-almost everywhere as $\mu_{s'}\big(X_{S=s'}(\Omega_0)\big) = \mathbb{P}(\Omega_0) = 1$. This concludes the first part of the proof.

Proof of 2. Set $s, s', s'' \in \mathcal{S}$. Following the same principle as before, we implicitly work on a set Ω_0 such that $\mathbb{P}(\Omega_0) = 1$ and for every $\omega \in \Omega_0$,

$$\begin{aligned}X_{S=s}(\omega) &= f_s(U_X(\omega)) \subset \mathcal{X}_s,\\ X_{S=s'}(\omega) &= f_{s'}(U_X(\omega)) \subset \mathcal{X}_{s'}.\end{aligned}$$

Then, we write

$$\begin{aligned}{T^*}_{\langle s''|s\rangle}(X_{S=s}) &= f_{s''} \circ f_s^{-1}|_{\mathcal{X}_s}(X_{S=s})\\ &= (f_{s''} \circ f_{s'}^{-1}) \circ (f_{s'} \circ f_s^{-1}|_{\mathcal{X}_s})(X_{S=s}).\end{aligned}$$

Note that $(f_{s'} \circ f_s^{-1}|_{\mathcal{X}_s})(X_{S=s}) = X_{S=s'} \in \mathcal{X}_{s'}$. Hence,

$$\begin{aligned}{T^*}_{\langle s''|s\rangle}(X_{S=s}) &= (f_{s''} \circ f_{s'}^{-1}|_{\mathcal{X}_{s'}}) \circ (f_{s'} \circ f_s^{-1}|_{\mathcal{X}_s})(X_{S=s})\\ &= {T^*}_{\langle s''|s'\rangle} \circ {T^*}_{\langle s'|s\rangle}(X_{S=s}).\end{aligned}$$

Similarly to the previous point, this means that the equality ${T^*}_{\langle s''|s\rangle} = {T^*}_{\langle s''|s'\rangle} \circ {T^*}_{\langle s'|s\rangle}$ holds on $X_{S=s}(\Omega_0)$ where $\mathbb{P}(\Omega_0) = 1$. Thus, it holds μ_s-almost everywhere as $\mu_s\big(X_{S=s}(\Omega_0)\big) = \mathbb{P}(\Omega_0) = 1$. This concludes the proof. □

C Proofs of Section 4

Proof of Theorem 1.

Proof. We address the structural equations

$$X = MX + wS + b + U_X,$$

where $w, b \in \mathbb{R}^d$ and $M \in \mathbb{R}^{d\times d}$ are deterministic parameters. Acyclicity imposes that $I - M$ is invertible, which enables to write

$$X = (I-M)^{-1}(wS + b + U_X) =: \mathbf{F}(S, U_X).$$

Using our previous notations, we have that for any $s \in \mathcal{S}$, $f_s(u) = (I-M)^{-1}(ws + b + u)$. Remark that **(SW)** holds such that $f_s^{-1}(x) = (I-M)x - ws - b$. Now, set $s, s' \in \mathcal{S}$, and use the definition of $T^*_{\langle s'|s\rangle}$ to obtain

$$\begin{aligned}T^*_{\langle s'|s\rangle}(x) &= (I-M)^{-1}\big(w(s'-s) + (I-M)x\big)\\ &= x + (I-M)^{-1}w(s'-s).\end{aligned}$$

According to Section 2.1, it suffices to show that $T^*_{\langle s'|s\rangle}$ coincides μ_s-almost everywhere with the gradient of a convex function to conclude that it is the Brenier map between μ_s and $\mu_{s'}$. This is clearly the case, as $T^*_{\langle s'|s\rangle}$ is the gradient of the convex function $x \mapsto \frac{1}{2}\|x\|^2 + \big[(I-M)^{-1}w(s'-s)\big]^T x$. □

D Proofs of Section 5

Proof of Proposition 5.

Proof. We address each point separately.

Proof of 1. We aim at showing that counterfactual fairness is equivalent to:

(Goal) *For every $s, s' \in \mathcal{S}$, there exists a measurable set $C := C(s, s') \subset \mathcal{X} \times \mathcal{X}$ satisfying $\pi^*_{\langle s'|s\rangle}(C) = 1$ such that for every $(x, x') \in C$*

$$h(x, s) = h(x', s').$$

A direct reformulation of the counterfactual fairness condition is:

(CF) *For every $s, s' \in \mathcal{S}$, there exists a measurable set $A := A(s)$ satisfying $\mu_s(A) = 1$, such that for every $x \in A$ and every measurable set $M \subset \mathbb{R}$*

$$\begin{aligned}\mathbb{P}(\hat{Y}_{S=s} \in M | X = x, S = s) \\ = \mathbb{P}(\hat{Y}_{S=s'} \in M | X = x, S = s).\end{aligned} \tag{3}$$

To show that **(CF)** is equivalent to **(Goal)**, we first rewrite **(CF)** into the following intermediary formulation:

(IF) *For every $s, s' \in \mathcal{S}$, there exists a measurable set $A := A(s)$ satisfying $\mu_s(A) = 1$, such that for every $x \in A$ and every measurable $M \subset \mathbb{R}$ there exists a measurable set $B := B(s, s', x, M)$ satisfying $\mu_{\langle s'|s\rangle}(B|x) = 1$ and such that for every $x' \in B$,*

$$\mathbf{1}_{\{h(x,s)\in M\}} = \mathbf{1}_{\{h(x',s')\in M\}}.$$

Proof that (CF) $\iff$ (IF). Suppose that $s, s', x \in A$ and $M \subset \mathbb{R}$ are fixed. According to Lemma 2, $\mathcal{L}(X|S = s) = \mathcal{L}(X_{S=s}|S = s)$, so that we can rewrite the left-term of (3) as

$$\begin{aligned}\mathbb{P}(\hat{Y}_{S=s} \in M | X = x, S = s) \\ &= \mathbb{P}(h(X_{S=s}, s) \in M | X = x, S = s) \\ &= \mathbb{P}(h(X, s), s) \in M | X = x, S = s) \\ &= \mathbb{P}(h(x, s) \in M) \\ &= \mathbf{1}_{\{h(x,s)\in M\}}.\end{aligned}$$

Then, using the distributions of the structural counterfactuals, express the right-term of (3) as

$$\begin{aligned}\mathbb{P}(\hat{Y}_{S=s'} \in M | X = x, S = s) \\ &= \mathbb{P}(h(X_{S=s'}, s') \in M | X = x, S = s) \\ &= \int \mathbf{1}_{\{h(x',s')\in M\}} d\mu_{\langle s'|s\rangle}(x'|x).\end{aligned}$$

Because the indicator functions equal either 0 or 1, the condition $\mathbf{1}_{\{h(x,s)\in M\}} = \int \mathbf{1}_{\{h(x',s')\in M\}} d\mu_{\langle s'|s\rangle}(x'|x)$ is equivalent to $\mathbf{1}_{\{h(x,s)\in M\}} = \mathbf{1}_{\{h(x',s')\in M\}}$ for $\mu_{\langle s'|s\rangle}(\cdot|x)$-almost every x'. This means that there exists a measurable set $B := B(s, s', x, M)$ such that $\mu_{\langle s'|s\rangle}(B|x) = 1$ and for every $x' \in B$,

$$\mathbf{1}_{\{h(x,s)\in M\}} = \mathbf{1}_{\{h(x',s')\in M\}}$$

This proves that **(CF)** is equivalent to **(IF)**.

Proof that (IF) $\implies$ (Goal). As **(IF)** is true for any arbitrary measurable set $M \subset \mathbb{R}$, we can apply this result with $M = \{h(x, s)\}$ to obtain a measurable set $B := B(s, s', x)$ such that $\mu_{\langle s'|s\rangle}(B|x) = 1$ and for every $x' \in B$, $h(x', s') = h(x, s)$. To sum-up, for every $s, s' \in \mathcal{S}$, there exists a measurable set $A := A(s)$ satisfying $\mu_s(A) = 1$ such that for every $x \in A$, there exists a measurable set $B := B(s, s', x)$ satisfying $\mu_{\langle s'|s\rangle}(B|x) = 1$, such that for every $x' \in B$, $h(x', s') = h(x, s)$. Now, we must show that the latter equality holds for $\pi^*_{\langle s'|s\rangle}$-almost every (x, x').

To this end, set $C := C(s, s') = \{(x, x') \in \mathcal{X} \times \mathcal{X} | x \in A(s), x' \in B(s, s', x)\}$. Remark that by definition of A and B, for every $(x, x') \in C$, $h(x, s) = h(x', s')$. To conclude, let us prove that $\pi^*_{\langle s'|s\rangle}(C) = 1$.

$$\begin{aligned}\pi^*_{\langle s'|s\rangle}(C) &= \int_A \mathbb{P}(X_{S=s'} \in B | X = x, S = s) d\mu_s(x) \\ &= \int_A \mu_{\langle s'|s\rangle}(B|x) d\mu_s(x) \\ &= \int_A 1 d\mu_s(x) \\ &= \mu_s(A) \\ &= 1.\end{aligned}$$

This proves that **(IF)** implies **(Goal)**.

Proof that (Goal) $\implies$ (IF). Using **(Goal)**, consider a measurable set $C := C(s, s')$ satisfying $\pi^*_{\langle s'|s\rangle}(C) = 1$ and such that for every $(x, x') \in C$, $h(x, s) = h(x', s')$. Then, define for any $x \in \mathcal{X}$, the measurable set $B(s, s', x) := \{x' \in \mathcal{X} | (x, x') \in C\}$. According to the disintegrated formula of $\pi^*_{\langle s'|s\rangle}$,

$$1 = \int \mu_{\langle s'|s\rangle}(B|x) d\mu_s(x).$$

Since $0 \leq \mu_{\langle s'|s\rangle}(B|x) \leq 1$, this implies that for μ_s-almost every x, $\mu_{\langle s'|s\rangle}(B|x) = 1$. Said differently, there exists a measurable set $A := A(s)$ satisfying $\mu_s(A) = 1$ such that for every $x \in A$, the measurable set $B(s, s', x)$ satisfies $\mu_{\langle s'|s\rangle}(B|x) = 1$. By construction of B and by definition of C, for every $x \in A$ and every $x' \in B$, $h(x, s) = h(x', s')$. To obtain **(IF)**, it suffices to take any measurable $M \in \mathbb{R}$ and to note that the latter equality implies that $\mathbf{1}_{\{h(x,s)\in M\}} = \mathbf{1}_{\{h(x',s')\in M\}}$.

Proof of 2. Consider **(CF)**, and recall that for every $s, s' \in S$, μ_s-almost every x and every measurable $M \subset \mathbb{R}$ the left term of (3) is $\mathbf{1}_{\{h(x,s)\in M\}}$. Let us now reframe the right-term of (3). If **(SW)** holds, using that $U_X = f_S^{-1}(X)$ we obtain

$$
\begin{aligned}
&\mathbb{P}(\hat{Y}_{S=s'} \in M | X = x, S = s) \\
&\quad = \mathbb{P}(h(X_{S=s'}, s') \in M | X = x, S = s) \\
&\quad = \mathbb{P}(h(\mathbf{F}(s', U_X), s') \in M | X = x, S = s) \\
&\quad = \mathbb{P}(h(f_{s'}(f_S^{-1}(X)), s') \in M | X = x, S = s) \\
&\quad = \mathbb{P}(h(f_{s'} \circ f_s^{-1}(x), s') \in M) \\
&\quad = \mathbb{P}(h(T^*_{\langle s'|s \rangle}(x), s') \in M) \\
&\quad = \mathbf{1}_{\{h(T^*_{\langle s'|s \rangle}(x), s') \in M\}}.
\end{aligned}
$$

Consequently, **(CF)** holds if and only if, for every measurable $M \in \mathbb{R}$

$$
\mathbf{1}_{\{h(x,s) \in M\}} = \mathbf{1}_{\{h(T^*_{\langle s'|s \rangle}(x), s') \in M\}}.
$$

Using the same reasoning as before, we take $M = \{h(x, s)\}$ to prove that this condition is equivalent to $h(x, s) = h(T^*_{\langle s'|s \rangle}(x), s')$. This concludes the second part of the proof.

Proof of 3. From 2. and Proposition 3, it follows that counterfactual fairness can be written as: for every $s, s' \in \mathcal{S}$ such that $s' < s$, for μ_s-almost every x

$$
h(x, s) = h(T^*_{\langle s'|s \rangle}(x), s'),
$$

and for $\mu_{s'}$-almost every x

$$
h(x, s') = h(T^*{}_{\langle s|s' \rangle}(x), s').
$$

Set $s, s' \in \mathcal{S}$ such that $s' < s$. To prove 3. we must show that these two conditions are equivalent. Set A a measurable subset of $\mathcal{X}_s$ such that $\mu_s(A) = 1$, and $h(x, s) = h(T^*_{\langle s'|s \rangle}(x), s')$ for any $x \in A$. Then, make the change of variable $x' = T^*_{\langle s'|s \rangle}(x)$ so that $h(T^{*-1}_{\langle s'|s \rangle}(x'), s') = h(x', s')$ for every $x' \in T^*_{\langle s'|s \rangle}(A)$. By Propositions 2 and 3, $T^*_{\langle s'|s \rangle \sharp} \mu_s = \mu_{s'}$, which implies that $\mu_{s'}(T^*_{\langle s'|s \rangle}(A)) = 1$. Therefore, the equality $h(T^{*-1}_{\langle s'|s \rangle}(x'), s) = h(x', s')$ holds for $\mu_{s'}$-almost every x'. Finally, recall that according to Proposition 4, $T^{*-1}_{\langle s'|s \rangle} = T^*{}_{\langle s|s' \rangle}$ $\mu_{s'}$-almost everywhere. As the intersection of two sets of probability one is a set of probability one, $h(T^*{}_{\langle s|s' \rangle}(x'), s) = h(x', s')$ holds for $\mu_{s'}$-almost every x'.

To prove the converse, proceed similarly by switching s to s'.

□

Proof of Proposition 6.

Proof. According to Proposition 5, h is counterfactually fair if and only if for any $s, s' \in \mathcal{S}$ and for $\pi^*_{\langle s'|s \rangle}$-almost every (x, x'), $h(x, s) = h(x', s')$ or equivalently $\mathbf{1}_{\{h(x,s) \in M\}} = \mathbf{1}_{\{h(x',s') \in M\}}$ for every measurable $M \in \mathbb{R}$. Set $s, s' \in \mathcal{S}$. Recall that from **(RE)**, $\pi^*_{\langle s'|s \rangle}$ admits μ_s for first marginal, and $\mu_{s'}$ for second marginal. Let us integrate this equality w.r.t. $\pi^*_{\langle s'|s \rangle}$ to obtain, for every measurable $M \subset \mathbb{R}$

$$
\int \mathbf{1}_{\{h(x,s) \in M\}} d\mu_s(x) = \int \mathbf{1}_{\{h(x',s') \in M\}} d\mu_{s'}(x).
$$

This can be written as, for every measurable $M \in \mathbb{R}$

$$
\mathbb{P}(h(X, s) \in M | S = s) = \mathbb{P}(h(X, s') \in M | S = s'),
$$

which means that

$$
\mathcal{L}(h(X, S) | S = s) = \mathcal{L}(h(X, S) | S = s').
$$

As this holds for any $s, s' \in \mathcal{S}$, we have that $h(X, S) \perp\!\!\!\perp S$.

One can easily convince herself that the converse is not true. As a counterexample, consider the following causal model,

$$
X = S \times U_X + (1 - S) \times (1 - U_X).
$$

Where S follows an arbitrary law and does not depend on U_X. Observe that **(RE)** is satisfied so that

$$
\begin{aligned}
\mathcal{L}(X_{S=0}) &= \mathcal{L}(X | S = 0), \\
\mathcal{L}(X_{S=1}) &= \mathcal{L}(X | S = 1), \\
\mathcal{L}(X | S = 0) &= \mathcal{L}(X | S = 1).
\end{aligned}
$$

In particular, whatever the chosen predictor, statistical parity will hold since the observational distributions are the same. By definition of the structural counterfactual operator, we have $T^*_{\langle 1|0 \rangle}(x) = 1 - x$. Now, set the *unaware* predictor (i.e., which does not take the protected attribute as an input), $h(X) := \text{sign}(X - 1/2)$. Clearly,

$$
h(T^*_{\langle 1|0 \rangle}(x)) = -h(x) \neq h(x).
$$

□