Transport-based Counterfactual Models

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

TRANSPORT-BASED COUNTERFACTUAL MODELS

Lucas De Lara *, Alberto González-Sanz[1], Nicholas Asher[2], and Jean-Michel Loubes[1]

[1]Institut de Mathématiques de Toulouse, Université Paul Sabatier
[2]Institut de Recherche en Informatique de Toulouse, Université Paul Sabatier

ABSTRACT

Counterfactual frameworks have grown popular in explainable and fair machine learning, as they offer a natural notion of causation. However, state-of-the-art models to compute counterfactuals are either unrealistic or unfeasible. In particular, while Pearl's causal inference provides appealing rules to calculate counterfactuals, it relies on a model that is unknown and hard to discover in practice. We address the problem of designing realistic and feasible counterfactuals in the absence of a causal model. We define transport-based counterfactual models as collections of joint probability distributions between observable distributions, and show their connection to causal counterfactuals. More specifically, we argue that optimal transport theory defines relevant transport-based counterfactual models, as they are numerically feasible, statistically-faithful, and can even coincide with causal counterfactual models. We illustrate the practicality of these models by defining sharper fairness criteria than typical group fairness conditions.

Keywords: Counterfactuals, Mass transportation, Optimal transport, Causality, Fairness

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 define causation [Lewis, 1973] and hence have attracted attention in the fields of explainability and robustness in machine learning, as such statements are tailored to explain black-box decision rules. Applications abound, including algorithmic recourse [Joshi et al., 2019, Poyiadzi et al., 2020, Karimi et al., 2020], defense against adversarial attacks [Ribeiro et al., 2016, Moosavi-Dezfooli et al., 2016] and fairness [Kusner et al., 2017, Asher et al., 2020, Black et al., 2020, Plecko and Meinshausen, 2020].

State-of-the-art models for computing 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, as noted by Black et al. [2020] and Poyiadzi et al. [2020], this simple distance approach generally fails to describe realistic alternative worlds, as it implicitly assumes the features to be independent. Changing just the sex of a person in such a translation might convert from a typical male into an untypical female rendering out-of-distribution counterfactuals like the following: *if I were a woman I would be 190cm tall and weigh 85 kg*. According to intuition, such counterfactuals are false and rightly so because they are not representative of the underlying statistical distributions. As a practical consequence, such counterfactuals typically hide biases in machine learning decision rules [Lipton et al., 2018, Besse et al., 2020].

The link between counterfactual modality and causality motivated the use of Pearl's causal modeling [Pearl, 2009] to address the aforementioned shortcoming [Kusner et al., 2017, Joshi et al., 2019, Karimi et al., 2020, Mahajan et al., 2020]. Pearl's do-calculus, by enforcing a change in a set of variables while keeping the rest of the causal mechanism untouched, provides a rigorous basis for generating intuitively true counterfactuals.

*lucas.de_lara@math.univ-toulouse.fr

The cost of this approach is fully specifying the causal model, namely specifying not only the Bayesian network (or graph) capturing the causal links between variables, but also the structural equations relating them, and the law of the latent, exogenous variables. The reliance on such a strong prior makes the causal approach appealing in theory, but limited for systematic implementation.

To sum-up, research has mostly focused on two divergent frameworks to compute counterfactuals: one that proposes an easy-to-implement model that leads, however, to intuitively untrue counterfactuals; another rigorously takes into account the dependencies between variables to produce realistic counterfactuals, but at the cost of feasibility. Our contribution addresses a third way. Extending the work of Black et al. [2020], who first suggested substituting causality-based counterfactual reasoning with optimal transport, we define *transport-based counterfactual models*. Such models, by characterizing a counterfactual operation as a coupling, a mass transportation plan between two observable distributions, ensures that the generated counterfactuals are in-distribution, hence realistic. In addition, they remedy to the impracticability issues of causal modeling as they can be computed through any mass transportation techniques, for instance optimal transport. As a consequence, this renders many counterfactual frameworks for fairness feasible.

1.1 Outline of contributions

We propose a mass transportation framework for counterfactual reasoning, and point out its similarities to the causal approach. Additionally, we use these framework to derive new transport-based counterfactual fairness criteria.

1. In Section 2, we introduce a general causality-free framework for the computation of counterfactuals through mass transportation techniques (not only optimal transport), leading to the definition of *transport-based* counterfactual models.
2. In Section 3, we present the causal approach of counterfactual inference from a mass transportation perspective. The interest of this viewpoint is two-fold. First, it offers an elegant representation of counterfactual operations and common causal assumptions. In particular, we study the implications of two general causal assumptions in this formalism. Under the *single-world* assumptions, we show that the pairing between factual and counterfactual instances is given by a closed-form deterministic map; when the intervened feature is *relative exogenous*, we show that the distribution of counterfactual instances coincide with an observational distribution. Second, it provides a common formalism with transport-based counterfactual models.
3. On the basis of the formalism and assumptions presented in Section 3, we demonstrate in Section 4 that optimal transport maps for the quadratic cost generates the same counterfactual instances as a large class of causal models, including linear additive models. We argue that this makes transport-based counterfactual models relevant surrogate models in the absence of a known causal model.
4. In Section 5, we apply the mass transportation viewpoint of structural counterfactuals by recasting the *counterfactual fairness* criterion into a transport-like one. Then, we propose new causality-free criteria by substituting the causal model by transport-based models in the original criterion.

1.2 Problem setup

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 $\mu_s := \mathcal{L}(X \mid S = s)$ 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. Typically, S represents the treatment variable in treatment modeling, or the sensitive, protected attribute in fairness settings. 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?* As aforementioned, because of correlations between the variables, computing the alternative state does not amount to change the value of S while keeping the features X equal.

2 Transport-based counterfactuals

We firstly introduce the necessary background on mass (or measure) transportation. Then, we present the notion of transport-based counterfactual model. Finally, we discuss the specific case of optimal transport.

2.1 Background in mass transportation

In probability theory, the problem of mass transportation amounts to 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 transportation 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. Formally, let P, Q be both probabilities on $\mathbb{R}^d$, whose respective supports are denoted by $\text{supp}(P)$ and $\text{supp}(Q)$. A coupling between P and Q is a probability π on $\mathbb{R}^d \times \mathbb{R}^d$ admitting P as first marginal and Q as second marginal, precisely $\pi(A \times \mathbb{R}^d) = P(A)$ and $\pi(\mathbb{R}^d \times B) = Q(B)$ for all measurable sets $A, B \subseteq \mathbb{R}^d$. Throughout the paper, we denote by $\Pi(P, Q) \subset \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ the set of joint distributions whose marginals coincide with P and Q respectively.

A coupling $\pi \in \Pi(P, Q)$ is said to be *deterministic* if each instance from the first marginal is paired with probability one to an instance of the second marginal. Such a coupling can be identified with a measurable map $T : \mathbb{R}^d \to \mathbb{R}^d$ that *pushes forward* P to Q, that is $Q(B) := P(T^{-1}(B))$ 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. To make the relation with random couplings, we also introduce the action of couples of functions on probability measures. For any pairs of functions $T_1, T_2 : \mathbb{R}^d \to \mathbb{R}^d$, we define $(T_1 \times T_2) : \mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}^d, x \mapsto (T_1(x), T_2(x))$. As such, $(T_1 \times T_2)_\sharp P$ denotes the law of $\big(T_1(Z), T_2(Z)\big)$ where $Z \sim P$. Thus, a push-forward operator T satisfying $T_\sharp P = Q$ characterizes the coupling $\pi = (I \times T)_\sharp P$ where I is the identity function on $\mathbb{R}^d$. This coupling matches every instance $x \in \text{supp}(P)$ to $T(x) \in \text{supp}(Q)$ with probability 1.

2.2 Definition

In [Black et al., 2020], the authors mimicked the structural account of counterfactuals by computing alternative instances using a deterministic optimal transport map. Extending their idea, we propose a more general framework where the counterfactual operation $s \mapsto s'$ can be see as a coupling, a mass transportation plan between the two observable distributions μ_s and $\mu_{s'}$. We define *counterfactual models* as models that assign a probability to all the cross-world statements on X w.r.t. interventions on S.

Definition 1. *1. A* transport-based counterfactual model *is a collection* $\Pi := \{\pi_{\langle s'|s\rangle}\}_{s,s' \in \mathcal{S}}$ *of couplings belonging to* $\Pi(\mu_s, \mu_{s'})$*, such that* $\pi_{\langle s|s\rangle} = (I \times I)_\sharp \mu_s$*. 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 transport-based 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.

In this formalism, the quantity $d\pi_{\langle s'|s\rangle}(x, x')$ is the elementary probability of the counterfactual statement *had S been equal to s' instead of s then X would have been equal to x' instead of x*. The conditions in Definition 1 translate the intuition that a realistic counterfactual operation on S should morph the non-intervened variables X so that their values fit the targeted distribution. The adjective *deterministic* refers to the fact that the model assigns to each factual instance a unique counterfactual counterpart. Formally, the counterfactual counterpart of some observation $x \in \text{supp}(\mu_s)$ for a change $s \mapsto s'$ is given by $x' = T_{\langle s'|s\rangle}(x) \in \text{supp}(\mu_{s'})$. In contrast, a *random* model allows possibly several counterparts with probability weights.

One challenge for the transport-based approach is to choose the model appropriately in order to define a relevant notion of counterpart. There possibly exists an infinite number of admissible counterfactual models in the sense of Definition 1, many of them being inappropriate. As an illustration, consider the family of trivial couplings, namely $\{\mu_s \otimes \mu_{s'}\}_{s,s' \in \mathcal{S}}$ where $\otimes$ denotes the factorization of measures. Though it is a well-defined transport-based counterfactual model, it is not intuitively justifiable as it completely decorrelates factual and counterfactual instances. A transport-based counterfactual model must be both *intuitively justifiable* and *computationally feasible*. We justify next that *optimal-transport-based* counterfactual models fit these conditions.

2.3 The case of optimal transport

We recall here some basic knowledge on optimal transport theory, and refer to [Villani, 2003, 2008] for further details. Optimal transport restricts the set of feasible couplings between two marginals by isolating ones that are optimal in some sense.

The *Monge formulation* of the optimal transport problem with general cost $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is the optimization problem

$$\min_{T:\, T_\sharp P = Q} \int_{\mathbb{R}^d} c(x, T(x)) dP(x). \tag{1}$$

We refer to solutions to (1) as *optimal transport maps* between P and Q with respect to c. One difficulty is that the push-forward constraint renders the problem unfeasible in many general settings, in particular when P and Q are not absolutely continuous w.r.t. the Lebesgue measure or have unbalanced numbers of atoms.

This issue motivated the following *Kantorovich relaxation* of the optimal transport problem with cost c,

$$\min_{\pi \in \Pi(P,Q)} \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x, x') d\pi(x, x'). \tag{2}$$

Solutions to (2) are *optimal transport plans* between P and Q with respect to c. In contrasts to optimal transport maps, they exist under very mild assumptions, like the non-negativeness of the cost.

As argued by Black et al. [2020], generating a counterfactual operation by solving (1) leads to intuitively relevant counterfactuals, as they are obtained by minimizing a metric between paired instances while preserving the probability distributions. The framework we propose generalizes the one of Black et al. [2020], restricted to optimal transport maps, by enabling random couplings. As mentioned before, using a random coupling rather than a deterministic map has philosophical implications since it renders non necessarily unique the counterfactual counterparts of a single instance. More practically, this relaxation is also crucial for dealing with non-continuous variables and for efficient implementations. In concrete settings, as practitioners do not completely know the *true* measures μ_s and $\mu_{s'}$ but have access to empirical samples, they must estimate the model from data. The computational complexity of solving (2) between an n-sample and an m-sample is in $\mathcal{O}\big((n+m)nm\log(n+m)\big)$, but one can substantially improve on this to reach $\mathcal{O}\big(nm\big)$ with entropy-regularized versions [Cuturi, 2013, Peyré et al., 2019]. However, such regularization necessarily entails non-deterministic couplings, making difficult the implementation of deterministic counterfactual models.

Interestingly, Black et al. [2020] observed that optimal transport maps generated nearly identical counterfactuals to the ones based on causal models. In the following, we show that the transport-based approach can be seen as an approximation of Pearl's computation of counterfactuals. Specifically, we reframe the generation of structural counterfactuals as a problem of mass transportation in Section 3, and prove that it identifies (possibly deterministic) transport-based counterfactual models in several settings. In Section 4, we demonstrate that a relevant choice of the transportation cost c makes optimal-transport counterfactuals coincide with a large class of causal counterfactuals.

3 Structural counterfactuals, revisited

Pearl's causal reasoning provides a natural framework to address the counterfactual problem introduced in Section 1.2. We consider a causal framework where $V = (X, S) \in \mathbb{R}^{d+1}$ is the unique solution to an acyclic *structural causal model* (SCM). More precisely, each *endogenous* variable V_i is defined (up to sets of probability zero) by the structural equation

$$V_i \overset{\mathbb{P}-a.s.}{=} G_i(V_{\text{Endo}(i)}, U_{\text{Exo}(i)}), \tag{3}$$

where G_i is a real-valued measurable function, U is a set of *exogenous* variables, while $V_{\text{Endo}(i)}$ and $U_{\text{Exo}(i)}$ denote respectively the endogenous and exogenous parents of V_i. Throughout the paper, we denote by U_X and U_S the exogenous parents of respectively X and S. The word "parent" refers to the fact that causal dependencies are generally represented by a *directed acyclic graph* (DAG).

We write $X_{S=s}$ the random vector defined as the solution to (3) under the do-intervention $do(S = s)$, that is after replacing the structural equation on S by $S = s$ while keeping the rest of the causal mechanism equal. This framework enables to define counterfactual statements as *had S been equal to s, then X would have been equal to $X_{S=s}$*. Following Karimi et al. [2020], we will refer to these counterfactuals as *structural* counterfactuals throughout the paper. In this section, we underline the mass transportation formalism of this computation of counterfactuals.

3.1 Definition

We introduce the following notations to formalize the contrast between interventional, counterfactual and factual outcomes. For $s, s' \in \mathcal{S}$ we define three probability distributions. Firstly, we recall that $\mu_s := \mathcal{L}(X \mid S = s)$ is the distribution of the *factual* s-instances. This observable measure describes the possible values of X such that $S = s$. Secondly, we denote by $\mu_{S=s} := \mathcal{L}(X_{S=s})$ the distribution of the *interventional* s-instances. It describes the alternative values of X in a world where S is forced to take the value s. On the contrary to the factual distribution, the interventional distribution is in general not observational, in the sense that we cannot draw empirical observations from it. Finally, we define by $\mu_{\langle s'|s\rangle} := \mathcal{L}(X_{S=s'} \mid S = s)$ the distribution of the *counterfactual* s'-instances given s. It describes what would have been the factual instances of μ_s *had* S *been equal to* s' *instead of* s. According to the *consistency rule* [Pearl et al., 2016], the factual and counterfactual distributions coincide when $s = s'$, that is $\mu_s = \mu_{\langle s|s\rangle}$. However, when $s \neq s'$, the counterfactual distribution $\mu_{\langle s'|s\rangle}$ is generally not observable. This is a key difference compared to the transport-based approach which relates observable distributions.

As for the transport-based approach, we need a counterfactual model on X w.r.t. interventions on S to characterize the distribution of all the cross-world statements. We refer to the model derived from Pearl's causal reasoning as the *structural counterfactual model*.

Definition 2. *For every* $s, s' \in \mathcal{S}$*, the* structural counterfactual coupling *between* μ_s *and* $\mu_{\langle s'|s\rangle}$ *is given by*

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

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

This definition shares similarities with Definition 1 by characterizing a counterfactual model as a collection of joint probability distributions. However, a structural couterfactual model is *a priori* not a transport-based counterfactual model since a coupling $\pi^*_{\langle s'|s\rangle}$ belongs to $\Pi(\mu_s, \mu_{\langle s'|s\rangle})$ and not $\Pi(\mu_s, \mu_{s'})$. By convention, we use the superscript * to denote the *structural* counterfactual models defined a causal model, and no superscript for the *transport-based* counterfactual models. Moreover, the structural counterfactuals couplings are not necessarily deterministic. This is due to the fact that, in general, there are several values of $X_{S=s}$ consistent with an evidence $\{X = x, S = s\}$.

In this section, we introduced Pearl's counterfactuals from a joint probability distribution perspective, which concurs with the work of Bongers et al. [2020]. However, some readers may be more familiar with the definition based on the so-called *three-step procedure* [Pearl et al., 2016]. This is for instance how Kusner et al. [2017] and Barocas et al. [2019] presented them. Crucially, these two viewpoints characterize the exact same counterfactual statements. We provide further insight on this topic in Appendix A.

In what follows, we study two specific scenarios of this counterfactual framework: first, when the counterfactuals are deterministic— then the computation can be written as an explicit push-forward operation; second, when S can be considered exogenous—then the counterfactual distribution is observable.

3.2 The deterministic case

We show that when the SCM deterministically implies the counterfactual values of X, 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 functional 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 $f_s : u \mapsto \mathbf{F}(s, u)$. Note that the s'-counterfactual counterparts of an s-instance x belong to the set $f_{s'} \circ f_s^{-1}(\{x\})$. This means that the counterfactual quantities are uniquely determined when the following assumption holds.

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

While the unique solvability of acyclic models ensures that (X, S) is deterministically determined by U, **(SW)** states that, conversely, U_X is deterministically determined by $\{X = x, S = s\}$. This *single-world* assumption implies that all the couplings between the factual and counterfactual distributions are deterministic, as written in the next proposition.

Proposition 3. *Let* ***(SW)*** *hold, and define for any* $s, s' \in \mathcal{S}$, $T^*_{\langle s'|s\rangle} := f_{s'} \circ f_s^{-1}|_{\mathcal{X}_s}$.[2] *The following properties hold:*

1. $\mu_{\langle s'|s\rangle} = T^*_{\langle s'|s\rangle \sharp}\mu_s$;
2. $\pi^*_{\langle s'|s\rangle} = (I \times T^*_{\langle s'|s\rangle})_\sharp \mu_s$.

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 deterministic structural counterfactual model* $\Pi^* = \{(I \times T^*_{\langle s'|s\rangle})_\sharp \mu_s\}_{s,s'\in\mathcal{S}}$.

Similarly to the structural counterfactual couplings, the operators in $\mathcal{T}^*$ describe the effect of causal interventions on factual distributions. We highlight that in our setting, they are well-defined without any knowledge on $\mathcal{L}(U)$.

Assumption **(SW)** is frequent in practice. It holds in particular for every *additive noise models*, models where the exogenous variables are additive terms of the structural equations, which accounts for most of the state-of-the-art SCMs. In [Karimi et al., 2020] for example, the authors considered such a setting for the purpose of computing algorithmic recourse. Note also that assumption **(SW)** imposes constraints on the variables and their laws to enable a deterministic correspondence between X and U_X. First, it entails that X and U_X have the same dimension, meaning that one exogenous variable U_i corresponds to exactly one endogenous variable X_i. Second, it requires for every pair (X_i, U_i) that both variables are either continuous or discrete with same-size domains. Remark that we framed **(SW)** so that it implies that all the counterfactuals instances for *any* changes on S are deterministic, leading to a fully deterministic counterfactual model. However, it suffices that one f_s be injective for some $s \in \mathcal{S}$ to render all the counterfactual couplings $\{\pi^*_{\langle s'|s\rangle}\}_{s'\in\mathcal{S}}$ deterministic. Therefore, when **(SW)** does not hold, the structural counterfactual model possibly contains both random and deterministic couplings.

3.3 The exogenous case

We discuss the counterfactual implications of the position of S in the causal graph. More specifically, we focus on the case where S can be considered as a root node. This entails that the structural counterfactual model is a transport-based counterfactual model.

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* $X_{Endo(S)} = \emptyset$.

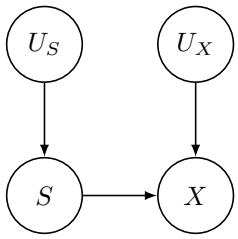

Figure 1: DAG satisfying **(RE)**

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 4. *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 4, **(RE)** renders the causal model otiose for the purpose of generating the counterfactual distribution, as the latter coincides with the observable factual distribution $\mu_{s'}$.

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

However, the coupling is *still required* to determine how each instance is matched at the individual level. We note that **(RE)** provides transitivity properties to our counterfactual operators.

Proposition 5. *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 features of individuals. Figure out a situtation where $\mathcal{S} = \{0, 1\}$ such that $S = 0$ stands for *female* while $S = 1$ stands for *male*, and assume for simplicity that **(SW)** holds. 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. As noted by Plecko and Meinshausen [2020], a practical consequence of **(RE)** is that it enables to link observational and causal notions of fairness, as addressed in Section 5. We remark that when both **(SW)** and **(RE)** hold, then Π^* is a deterministic transport-based counterfactual model.

3.4 The example of linear additive SCMs

We illustrate 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*

$$\begin{aligned} X &= MX + wS + b + U_X, \\ S &= U_S, \end{aligned}$$

*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 $s, s' \in \mathcal{S}$, $T^*_{\langle s'|s\rangle}(x) = x + (I - M)^{-1}w(s' - s)$.*

Remarkably, in the specific case of linear additive SCMs fitting **(RE)**, computing counterfactual quantities amounts to applying translations between factual distributions. Therefore, should an oracle reveal that the SCM belongs to this class without providing the structural equations, it would suffice to compute the mean translation between sampled points from μ_s and $\mu_{s'}$ to obtain an estimator of the counterfactual operator $T^*_{\langle s'|s\rangle}$. For more complex SCMs satisfying both **(SW)** and **(RE)**, it is presumably difficult to infer the counterfactual model from data. We address this issue in Section 4.1.

3.5 Shortcomings

We conclude Section 3 by discussing important drawbacks of the causal approach to counterfactual reasoning.

The main limitation, as for any causal-based framework, is its feasibility. Assuming a known causal model, in particular a fully-specified causal model, is a too strong assumption in practice. It requires experts to reach a consensus on the causal graph, the structural equations, the distribution of the input exogenous variables, and to test the validity of their model on available data. This is not a realistic scenario, especially when dealing with a high-number of features and possibly complex structural relations. Besides, this is not practical as a causal model must be designed and tested for each possible dataset. A more straighforward approach is to directly infer the causal graph from observational data. While there exist sound techniques to do so, they suffer from being 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]. In addition, this is not enough to compute counterfactual quantities, as the structural equations would still be lacking.

A related issue is causal uncertainty. There exist several causal models corresponding to a same data distribution, leading to possibly different counterfactual models (see Example 4.2 in [Bongers et al., 2020]). It cannot be tested whether the adjusted model is the *true* one, making the modeling inherently uncertain. Moreover, for non-deterministic structural counterfactual models, the computation of counterfactual quantities requires to know the law of the exogenous variables, which is not observable. While it is common to assume a prior distribution on U, this also adds uncertainty in the causal modeling, hence on the induced counterfactuals.

Perhaps more surprisingly, counterfactual quantities are sometimes nonexistent in Pearl's causal framework. The causal modeling we introduced is very general: we do not assume the exogenous variables to be mutually independent, and only suppose that the equations are acyclic. The DAG assumption is very common for both practicality and interpretability reasons. In general, however, observational data can be generated though an acyclical mechanism. Critically, (solvable) acyclic models do not always admit solutions under do-interventions, implying that $X_{S=s}$ may not be defined. We refer to Example 2.17 in [Bongers et al., 2020] for an illustration. As a consequence, counterfactual quantities are ill-defined in such settings.

Note that the mass transportation viewpoint of structural counterfactuals we introduced in this section does not resolve these impracticability issues. This viewpoint is a formulation, not a trick. Specifically, we still need a fully-specified causal model to determine $\{\pi^*_{\langle s'|s\rangle}\}_{s,s'\in\mathcal{S}}$. Nevertheless, this perspective suggests that transport-based methods can be natural substitutes for structural counterfactual reasoning, as they remedy to the aforementioned issues. Moreover, basic optimal-transport counterfactual models can sometimes coincide with structural counterfactuals as explained in the following section.

4 When optimal transport meets causality

Firstly, we introduce the key properties of optimal transport for the squared euclidean cost. Then, we show that it generates the same counterfactuals as a large class of causal models.

4.1 Squared euclidean optimal transport

Suppose that the ground cost c is the squared euclidean distance $c(x, x') := \|x - x'\|^2$ on $\mathbb{R}^d \times \mathbb{R}^d$, 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], Brenier's theorem, states that there exists a unique solution to Kantorovich's formulation of optimal transport (2), whose form is $(I \times T)_\sharp P$ where $T : \mathbb{R}^d \to \mathbb{R}^d$ solves the corresponding squared Monge problem,

$$\min_{T:\, T_\sharp P = Q} \int_{\mathbb{R}^d} \|x - T(x)\|^2 dP(x). \tag{4}$$

Although it may not be unique, this optimal transport map T is uniquely determined P-almost everywhere, and we will abusively refer to it as *the* solution to (4). Crucially, this map coincides P-almost everywhere with the gradient of a convex function. Moreover, according to McCann [1995], under the sole assumption that P is absolutely continuous with respect to the Lebesgue measure, there exists only one (up to P-negligible sets) gradient of a convex function $\nabla\phi$ satisfying the push-forward condition $\nabla\phi_\sharp P = Q$. We combine Brenier's and McCann's theorems into the following lemma, which simplifies the search for the solutions to (4).

Lemma 6. *Assume that P is absolutely continuous w.r.t. the Lebesgue measure, and that both P and Q have finite second order moments. If $T : supp(P) \to supp(Q)$ is a measurable map satisfying the two following conditions,*

1. *$T_\sharp P = Q$,*
2. *there exists a convex function $\phi : \mathbb{R}^d \to \mathbb{R}$ such that $T = \nabla\phi$ P-almost everywhere,*

then T is the solution to (4)*.*

This result will play a key role in the next subsection.

4.2 Main theorem

We focus on the deterministic transport-based counterfactual model defined by the collection of solutions to (4) between the pairs of measures $\{(\mu_s, \mu_{s'})\}_{s,s'\in\mathcal{S}}$. This model gives an elegant interpretation to the obtained counterfactual statements, as they are defined by minimizing the squared euclidean distance between paired instances, and can be seen as *non-decreasing* models. Note that being the gradient of a convex function generalizes the notion of non-decreasing function to several dimensions. Moreover, it recovers structural counterfactuals in specific cases.

Theorem 7. *Consider an SCM satisfying **(SW)** and **(RE)**. Suppose that the factual distributions are absolutely continuous w.r.t. Lebesgue measure and have finite second order moments. If for $s, s' \in \mathcal{S}$, the structural*

counterfactual operator $T^*_{\langle s'|s\rangle}$ is the gradient of some convex function, then it is the solution to (4) *between μ_s and $\mu_{s'}$.*

The mass transportation formalism of Pearl's counterfactual reasoning introduced in Section 3 renders the proof of this theorem straightforward. We underline that it does not require any prior knowledge on optimal transport theory except what we summarized in Lemma 6. Thus, for the sake of illustration and clarity, we reproduce the demonstration directly below.

Proof. According to **(SW)** and Proposition 3, the SCM defines a structural counterfactual operator $T^*_{\langle s'|s\rangle}$ between μ_s and $\mu_{\langle s'|s\rangle}$. Additionally, **(RE)** implies through Proposition 4 that $\mu_{\langle s'|s\rangle} = \mu_{s'}$. Therefore, $T^*_{\langle s'|s\rangle\sharp}\mu_s = \mu_{s'}$. Assume now that μ_s is absolutely continuous w.r.t. the Lebesgue measure, and that both μ_s and $\mu_{s'}$ have finite second order moments. If $T^*_{\langle s'|s\rangle}$ is the gradient of some convex function, then according to Lemma 6 it is the solution to (4) between μ_s and $\mu_{s'}$. □

Understanding the strength and limits of Theorem 7 requires understanding how rich is the class of SCMs fitting its assumptions. The larger the class, the more likely optimal transport maps for the squared eculidean cost will provide (nearly) identical counterfactuals to causality. Finding explicit conditions on f_s and $f_{s'}$ so that $f_{s'} \circ f_s^{-1}$ is the gradient of a convex potential requires tedious computations as soon as $d > 1$, which renders the identification of the relevant SCMs difficult. Nevertheless, we can find specific sub-classes of causal models fitting Theorem 7. For instance, as the structural counterfactual operator from Example 1 is the gradient of a convex function, we obtain the following corollary.

Corollary 8. *Consider a linear additive SCM satisfying **(RE)** (see Example 1). If the factual distributions are absolutely continuous w.r.t. Lebesgue measure and have finite second order moments, then for any $s, s' \in \mathcal{S}$, the structural counterfactual operator $T^*_{\langle s'|s\rangle}$ is the solution to* (4) *between μ_s and $\mu_{s'}$.*

The scope of Theorem 7 goes beyond linear additive SCMs, as shown in the following non-linear non-additive example.

Example 2. *Consider the following SCM,*

$$\begin{cases} X_1 = \alpha(S)U_1 + \beta_1(S), \\ X_2 = -\alpha(S)\ln^2\left(\frac{X_1-\beta_1(S)}{\alpha(S)}\right)U_2 + \beta_2(S), \\ S = U_S, \end{cases}$$

*where α, β_1, β_2 are $\mathbb{R}$-valued functions such that $\alpha > 0$, and $U_1 > 0$. It satisfies **(SW)** and **(RE)**, such that for any $s, s' \in \mathcal{S}$, the associated structural counterfactual operator is given by,*

$$T^*_{\langle s'|s\rangle}(x) = \frac{\alpha(s')}{\alpha(s)}x + \big[\beta(s') - \beta(s)\big],$$

where $\beta = (\beta_1, \beta_2)$ is $\mathbb{R}^2$-valued. This is the gradient of the convex function $x \mapsto \frac{\alpha(s')}{2\alpha(s)}\|x\|^2 + \big[\beta(s') - \beta(s)\big]^T x$. Then, if the factual distributions are absolutely continuous w.r.t. the Lebesgue measure, $T^*_{\langle s'|s\rangle}$ is the solution to (4) *between μ_s and $\mu_{s'}$.*

On the contrary, one can effortlessly find counterexamples to Theorem 7. This comes from the fact that many functions (even continuous ones) cannot be written as gradients when $d > 1$.

Example 3. *Consider the following SCM,*

$$\begin{cases} X_1 = U_1, \\ X_2 = SX_1^2 + U_2, \\ S = U_S. \end{cases}$$

*It satisfies **(SW)** and **(RE)**, such that for any $s, s' \in \mathcal{S}$, the associated structural counterfactual operator is given by,*

$$T^*_{\langle s'|s\rangle}(x_1, x_2) = (x_1,\ x_2 + (s' - s)x_1^2).$$

Note that this cannot be written as the gradient of a function. Consequently, it is not a solution to (4)*.*

Theorem 7 and Corollary 8 support the intuition that substituting Π^* with a surrogate Π from optimal transport provides a decent approximation of causal changes. Using a model close to Π^* would be ideal for interpretability reasons, but an expert can always propose and defend a different notion of similarity Π (built with optimal transport or not). The main interest lies in the feasibility of transport-based solutions, as they do not require any causal assumptions.

5 Application to fairness

We now turn to the application of the mass transportation viewpoint of counterfactual reasoning to fairness in machine learning. Suppose that the random variable S encodes a so-called *sensitive* or *protected attribute* (for example race or gender) which divides the population into different classes in a machine learning prediction task. We denote by $h : \mathbb{R}^d \times \mathcal{S} \mapsto \mathbb{R}$ an arbitrary predictor defining the random variable of predicted output $\hat{Y} := h(X, S)$. Fairness addresses the question of the dependence of $\hat{Y}$ on the protected attribute S. The most classical fairness criterion is the so-called *demographic* or *statistical parity*, which is achieved when $\hat{Y} \perp\!\!\!\perp S$.

However, this criterion is notoriously limited, as it 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]. The counterfactual framework, by capturing the structural or statistical links between the features and the protected attribute, allows for sharper notions of fairness. We first use the mass transportation formalism introduced in Section 3 to reformulate the accepted *counterfactual fairness* condition [Kusner et al., 2017]. On the basis of the reformulation, we then propose new fairness criteria derived from transport-based counterfactual models.

5.1 Generalizing counterfactual fairness

Counterfactual fairness is achieved when individuals and their structural counterfactual counterparts are treated equally.

Definition 9. *A predictor $\hat{Y} = h(X, S)$ is* counterfactually fair *if for every $s, s' \in \mathcal{S}$ and μ_s-almost every x,*

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

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

The above definition does not emphasize the pairing between factual and counterfactual values. Interestingly, the structural counterfactual transport plans allow for pair-wise characterizations of counterfactual fairness.

Proposition 10. *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,*

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

Assumption **(SW)** highlights the deterministic relationship between factual and counterfactual quantities. In addition, it makes unnecessary the knowledge of $\mathcal{L}(U)$ to test counterfactual fairness. Note that, if **(RE)** holds, then counterfactual fairness is a stronger criterion than the statistical parity across groups.

Proposition 11. *Suppose that **(RE)** holds. If the predictor $h(X, S)$ satisfies counterfactual fairness, then it satisfies statistical parity. 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 2.

Definition 12. *1. Let $\Pi = \{\pi_{\langle s'|s\rangle}\}_{s,s' \in \mathcal{S}}$ be a (random) transport-based 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 transport-based 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 11 only relies on the assumption that the couplings have factual distributions for marginals, the following proposition holds.

Proposition 13. *Let Π be a transport-based counterfactual model (deterministic or not). If a predictor $h(X, S)$ satisfies Π-counterfactual fairness, then it satisfies statistical parity. The converse does not hold in general.*

This proposition has interesting consequences. Consider that, for the purpose of computing counterfactual quantities, some practitioners designed a candidate SCM fitting the data and satisfying **(RE)**. Even if the candidate is misspecified, it would still characterize a transport-based counterfactual model controlling statistical parity. The fair data processing transformation proposed by Plecko and Meinshausen [2020] is an illustrative example.

More generally, the conceptual interest of transport-based fairness criteria is the same as the original counterfactual fairness criterion: they offer notions of individual fairness while still controlling for discrimination against protected groups. The added value is their feasibility. In contrast to Definition 9 and Proposition 10, Definition 12 relies on computationally feasible counterfactual models that obviate any assumptions about the data-generation process. In addition, as Definition 9 amounts to Π^*-counterfactual fairness (when **(RE)** holds), one can as well think of Definition 12 as an approximation of counterfactual fairness.

Crucially, these new criteria can naturally be applied in classical explainability and fairness machine learning frameworks based on counterfactual reasoning. We note that the explanability framework of Black et al. [2020] can be seen as controlling for a transport-based notion of counterfactual fairness.

5.2 Ethical risk

We conclude this section by discussing a potentially negative impact of our work. As aforementioned, the transport-based approach allows for many counterfactual models, but they do not all define legitimate notions of counterparts. Consequently, transport-based notions of counterfactual fairness could be used for unethical fair-washing. The next proposition formalizes this risk.

Proposition 14. *If $h(X, S)$ is a classifier satisfying statistical parity, then there exists a transport-based counterfactual model Π such that $h(X, S)$ satisfies Π-counterfactual fairness.*

Practitioners could take advantage of the weak notion of statistical parity to construct counterfactual models such that their trained classifiers are counterfactually fair, while still discriminating at the subgroup or individual level. This is why we argue that practitioners must always be able to justify the counterfactual models when not imposed by legal experts of the prediction task.

6 Conclusion

In this paper, we focused on the challenge of designing sound and feasible counterfactuals. Our work showed that the causal account for counterfactual modeling can be written in a mass transportation formalism, where implying either deterministic or random counterfactuals has a direct formulation in terms of the deterministic or random nature of the coupling. This perspective enabled us to generalize sharp but unfeasible causal criteria of fairness by transport-based ones. In doing so, we hope to have shed a new light on counterfactual reasoning, and fairness techniques.

A Further background in structural counterfactual reasoning

In this appendix, we further detail counterfactual reasoning through Pearl's causal modeling. More specifically, we make the connection between the joint-probability viewpoint of structural counterfactuals we introduced and the more typical three-step procedure approach.

Recall that our problem is: having observed an $x \in \text{supp}(\mu_s)$, determining the probability of the counterfactual outcome $x' \in \text{supp}(\mu_{\langle s'|s \rangle})$. Pearl et al. [2016] originally answered this question with the following procedure: (1) set a prior $\mathcal{L}(U)$ for the model $\mathcal{M}$, (2) compute the posterior distribution $\mathcal{L}(U \mid X = x, S = s)$, and (3) solve the structural equations of $\mathcal{M}_{S=s'}$ with $\mathcal{L}(U \mid X = x, S = s)$. This leads to the following formal definition of *structural counterfactuals*, adapted from [Pearl et al., 2016], which will play a role in the proof of Proposition 10.

Definition 15. *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*

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

As previously mentioned, this definition characterizes the exact same counterfactuals as Definition 2, the formal link being the following disintegrated formulation, $\pi^*_{\langle s'|s\rangle} = \int \big(\delta_x \otimes \mu_{\langle s'|s\rangle}(\cdot|x)\big) d\mu_s(x)$, where δ_x denotes the Dirac measure at point x. As anticipated, the structural counterfactuals of a single instance are not necessarily *deterministic*, that is characterized by a degenerate distribution, but belong to a set of possible outcomes with probability weights. This comes from the fact that several values of U can generate a same observation $\{X = x, S = s\}$. The next proposition specifies the range of the counterfactual outcomes.

Proposition 16. *For any $s, s' \in \mathcal{S}$, $x \in \mathcal{X}_s$,*

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

Proof. Recall 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 $\mathcal{X}$

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

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

This readily entails the following result, consistent with Proposition 3.

Proposition 17. *If **(SW)** holds, then $\mu_{\langle s'|s\rangle}(\cdot|x) = \delta_{T^*_{\langle s'|s\rangle}(x)}$ for every $x \in \mathcal{X}_s$.*

In particular, we recover that in a deterministic setting, the knowledge of $\mathcal{L}(U)$ is not necessary to compute counterfactual quantities.

B Proofs of the main results

This section addresses the proofs of the paper.

B.1 Proofs of Section 3

B.1.1 Proposition 3

Proof. Set $s, s' \in \mathcal{S}$ and $x \in \mathcal{X}_s$. Note that, according to **(SW)**, $U_X = f_S^{-1}(X)$ $\mathbb{P}$-almost surely. Let us address each point of the proposition separately.

Proof of 1. By definition of the counterfactual distribution, and as being equal almost surely implies being equal in law, we find

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

This proves the first point of the proposition.

Proof of 2. Similarly, by definition of the structural counterfactual coupling we obtain

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

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

□

B.1.2 Proposition 4

Proof. Set $s \in \mathcal{S}$ and recall that $X = \mathbf{F}(S, U_X)$ and $X_{S=s} = \mathbf{F}(s, U_X)$ $\mathbb{P}$-almost surely. Thanks to Assumption **(RE)**, we have that $S \perp\!\!\!\perp U_X$. Then,

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

This means that $\mu_s = \mu_{S=s}$. Similarly, for $s, s' \in \mathcal{S}$ the counterfactual distribution becomes

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

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

□

B.1.3 Proposition 5

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 4, 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,

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

As for 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 completes the proof.

□

B.2 Proofs of Section 4

B.2.1 Corollary 8

Proof. We address the structural equations

$$
\begin{aligned}
X &= MX + wS + b + U_X, \\
S &= U_S,
\end{aligned}
$$

where $w, b \in \mathbb{R}^d$ and $M \in \mathbb{R}^{d\times d}$ are deterministic parameters. We showed that for any $s, s' \in \mathcal{S}$,

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

Notice that $T^*_{\langle s'|s\rangle}$ is the gradient of the convex function $x \mapsto \frac{1}{2}\|x\|^2 + \big[(I - M)^{-1}w(s' - s)\big]^T x$. As **(RE)** holds and μ_s is absolutely continuous w.r.t. the Lebesgue measure, it follows from Theorem 7 that $T^*_{\langle s'|s\rangle}$ is the solution to (4) between μ_s and $\mu_{s'}$.

□

B.3 Proofs of Section 5

B.3.1 Proposition 10

Proof. We address each point separately.

Proof of 1. We claim 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 $M \subset \mathbb{R}$*

$$\mathbb{P}(\hat{Y}_{S=s} \in M \mid X = x, S = s) = \mathbb{P}(\hat{Y}_{S=s'} \in M | X = x, S = s). \tag{5}$$

We aim at showing that **(CF)** is equivalent to **(Goal)**. To do so, 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 the consistency rule $\mathcal{L}(X \mid S = s) = \mathcal{L}(X_{S=s} \mid S = s)$, so that we can rewrite the left term of (5) as

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

Then, using Definition 15, we reframe the right term of (5) as

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

Remark now that because the indicator functions take either the value 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 $B(s, s', x) := \{x' \in \mathcal{X} \mid (x, x') \in C\}$. We use disintegrated formula of $\pi^*_{\langle s'|s\rangle}$ to write

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

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 (5) is $\mathbf{1}_{\{h(x,s)\in M\}}$. Let us now reframe the right-term of (5). If **(SW)** holds, using that $U_X = f_S^{-1}(X)$ we obtain

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

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 the previous point and Proposition 4, 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 the third point, 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 3 and 4, $T^*_{\langle s'|s\rangle \sharp}\mu_s = \mu_{s'}$, 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 5, $T^{*-1}_{\langle s'|s\rangle} = T^*{}_{\langle s|s'\rangle}$ $\mu_{s'}$-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, we can proceed similarly by switching s to s'.

□

B.3.2 Proposition 11

Proof. According to Proposition 10, 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 $M \in \mathbb{R}$

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

which means that

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

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

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

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,

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

□

B.3.3 Proposition 14

Proof. Suppose that the classifier $h(X, S)$ takes values in the finite set $\mathcal{Y} \subset \mathbb{R}$, and denote for any $s \in \mathcal{S}$ and $y \in \mathcal{Y}$, the sets $\mathcal{H}(s, y) := \{x \in \mathbb{R}^d \mid h(x, s) = y\}$. Statistical parity can be written as, for any $s \in \mathcal{S}$ and any $y \in \mathcal{Y}$,

$$\mu_s(\mathcal{H}(s, y)) = p_y,$$

where $\{p_y\}_{y\in\mathcal{Y}}$ is a probability on $\mathcal{Y}$ that does not depend on s.

Now, set $s, s' \in \mathcal{S}$. We aim at constructing a coupling $\pi_{\langle s'|s\rangle}$ between μ_s and $\mu_{s'}$ such that,

$$\pi_{\langle s'|s\rangle}\big(\{(x,x') \in \mathbb{R}^d \times \mathbb{R}^d \mid h(x,s) = h(x',s')\}\big) = 1.$$

We define our candidate $\pi_{\langle s'|s\rangle}$ as,

$$d\pi_{\langle s'|s\rangle}(x,x') := \sum_{y\in\mathcal{Y}} \frac{\mathbf{1}_{\{x\in\mathcal{H}(s,y)\}}\mathbf{1}_{\{x'\in\mathcal{H}(s',y)\}}}{p_y} d\mu_s(x) d\mu_{s'}(x').$$

First, let's show that it admits respectively μ_s and $\mu_{s'}$ as first and second marginals. Let $A \subseteq \mathbb{R}^d$ be a measurable set,

$$\begin{aligned}
\pi_{\langle s'|s\rangle}(A \times \mathbb{R}^d) &= \sum_{y\in\mathcal{Y}} \int_{\mathbb{R}^d}\int_A \frac{\mathbf{1}_{\{x\in\mathcal{H}(s,y)\}}\mathbf{1}_{\{x'\in\mathcal{H}(s',y)\}}}{p_y} d\mu_s(x) d\mu_{s'}(x') \\
&= \sum_{y\in\mathcal{Y}} \frac{p_y}{p_y} \int_A \mathbf{1}_{\{x\in\mathcal{H}(s,y)\}} d\mu_s(x) \\
&= \sum_{y\in\mathcal{Y}} \mu_s(A \cap \mathcal{H}(s,y)) \\
&= \mu_s(A).
\end{aligned}$$

One can follow the same computation for the second marginal. To conclude, compute

$$\begin{aligned}
&\pi_{\langle s'|s\rangle}\big(\{(x,x') \in \mathbb{R}^d \times \mathbb{R}^d \mid h(x,s) = h(x',s')\}\big) \\
&\qquad = \pi_{\langle s'|s\rangle}\Big(\bigsqcup_{y\in\mathcal{Y}} \mathcal{H}(s,y) \times \mathcal{H}(s',y)\Big) \\
&\qquad = \sum_{y\in\mathcal{Y}} \pi_{\langle s'|s\rangle}(\mathcal{H}(s,y) \times \mathcal{H}(s',y)) \\
&\qquad = \sum_{y\in\mathcal{Y}} \frac{1}{p_y} \int \mathbf{1}_{\{x\in\mathcal{H}(s,y)\}} d\mu_s(x) \int \mathbf{1}_{\{x\in\mathcal{H}(s',y)\}} d\mu_{s'}(x) \\
&\qquad = \sum_{y\in\mathcal{Y}} \frac{1}{p_y} p_y \times p_y \\
&\qquad = 1.
\end{aligned}$$

□

References

Nicholas Asher, Soumya Paul, and Chris Russell. Adequate and fair explanations, 2020.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. `http://www.fairmlbook.org`.

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.

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. ISBN 9781450369367. doi: 10.1145/3351095.3372845. URL `https://doi.org/10.1145/3351095.3372845`.

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.

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.

Gregory F. Cooper. The computational complexity of probabilistic inference using bayesian belief networks (research note). *Artif. Intell.*, 42(2–3):393–405, March 1990. ISSN 0004-3702. doi: 10.1016/0004-3702(90)90060-D. URL https://doi.org/10.1016/0004-3702(90)90060-D.

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

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.

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

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.

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

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. URL https://proceedings.neurips.cc/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

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

Zachary Lipton, Julian McAuley, and Alexandra Chouldechova. Does mitigating ml's impact disparity require treatment disparity? In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/8e0384779e58ce2af40eb365b318cc32-Paper.pdf.

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

Robert J. McCann. Existence and uniqueness of monotone measure-preserving maps. *Duke Math. J.*, 80(2):309–323, 11 1995. doi: 10.1215/S0012-7094-95-08013-2. URL https://doi.org/10.1215/S0012-7094-95-08013-2.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016.

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

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

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

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

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. ISBN 9781450342322. doi: 10.1145/2939672.2939778. URL https://doi.org/10.1145/2939672.2939778.

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.

Cédric Villani. *Topics in optimal transportation*. Number 58 in Graduate Studies in Mathematics. American Mathematical Soc., 2003.

Cédric Villani. *Optimal Transport: Old and New*. Number 338 in Grundlehren der mathematischen Wissenschaften. Springer, Berlin, 2008. ISBN 978-3-540-71049-3. OCLC: ocn244421231.

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.