Transport-based Counterfactual Models

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

Transport-based Counterfactual Models

Lucas De Lara[1], Alberto González-Sanz[1], Nicholas Asher[2], Laurent Risser[1], 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 machine learning for both explaining algorithmic decisions but also defining individual notions of fairness, more intuitive than typical group fairness conditions. 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 coincide under some assumptions with causal counterfactual models. Finally, these models make counterfactual approaches to fairness feasible, and we illustrate their practicality and efficiency on fair learning. With this paper, we aim at laying out the theoretical foundations for a new, implementable approach to counterfactual thinking.

Keywords: Counterfactuals, Optimal transport, Causality, Fairness, Supervised learning

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., 2021, Rasouli and Yu, 2021, Slack et al., 2021, Bajaj et al., 2021], defense against adversarial attacks [Ribeiro et al., 2016, Moosavi-Dezfooli et al., 2016] and fairness [Kusner et al., 2017, Black et al., 2020, Plecko and Meinshausen, 2020, Asher et al., 2021].

State-of-the-art models for computing meaningful counterfactuals have mostly focused on the *nearest counterfactual explanation* 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 gender 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., 2021].

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, Mahajan et al., 2020, Karimi et al., 2021]. 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. 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 inadequate for deployment on practical cases.

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. The major benefit of this approach is that it renders doable many critical applications of counterfactual frameworks, for example in algorithmic fairness.

1.1 Outline of contributions

We make both theoretical and practical contributions in the fields of counterfactual reasoning and fair machine learning. We propose a mass-transportation framework for counterfactual reasoning and point out its similarities to the causal approach. Additionally, we show that counterfactual methods for fairness become feasible in this framework by introducing and implementing transport-based counterfactual fairness criteria. More precisely, our contributions can be outlined as follows.

1. In Section 2, we recall the necessary background on Pearl's causal modeling, while we introduce in Section 3 the basics of mass transportation and optimal transport theory. Both sections serve as the theoretical and notational toolbox that will be used throughout; they are meant to keep the paper self-contained and can be skipped by readers familiar with these subjects.

2. In Section 4, we firstly recall how to compute counterfactual quantities using causal modeling. Then, we introduce a general causality-free framework for the computation of counterfactuals through mass-transportation techniques, encompassing the approach of Black et al. [2020]. Essentially, we also propose a unified mass-transportation viewpoint of counterfactuals, be them causal-based or transport-based, through the definition *counterfactual models*, collections of couplings characterizing all possible counterfactual statements for a given feature to alter (for example the gender). We provide concrete examples of models, and discuss the limitations of the different approaches.

3. In Section 5, we leverage the unified formalism proposed in the previous section to demonstrate connections between causality and optimal transport. More precisely, after studying the implications of two general causal assumptions onto the induced counterfactual models, we demonstrate that optimal transport maps for the quadratic cost generates the same counterfactual instances as some specific causal models, including the common linear additive models. We argue that this makes optimal-transport-based counterfactual models relevant surrogates in the absence of a known causal model.

4. In Sections 6, 7 and 8, we illustrate the practicality of our approach for fairness in machine learning. We apply the mass-transportation viewpoint of structural counterfactuals by recasting the *counterfactual fairness* criterion [Kusner et al., 2017] 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. Finally, we address the training of counterfactually fair classifiers, providing statistical guarantees and numerical experiments over various datasets.

To sum-up: Sections 2 and 3 provide the prerequisites for the paper; Sections 4 and 5 introduce the concept of counterfactual models and the corresponding theory; Sections 6 to 8 address fairness applications of these models.

1.2 Related work

This work follows the paper of Black et al. [2020], which focus on building sound counterfactual quantities through optimal transport, deviating from both causal-based techniques and the nearest-counterfactual-instance principle. Our contributions in Sections 4 and 5 can be seen as the theoretical foundations of their approach, by shedding light on the link between measure-preserving counterfactuals and structural counterfactuals. Also, we note that the way we introduce the causal account for counterfactual reasoning in Section 4 concurs with [Plecko and Meinshausen, 2020] and [Bongers et al., 2021]. More precisely, we underline that the objects of interest are the joint probability distributions, or couplings, generated by manipulations of the causal model. Additionally, we propose in Section 6 a direct extension of the counterfactual fairness frameworks introduced in [Kusner et al., 2017] and [Russell et al., 2017] to transport-based counterfactual models, leading to a new method for supervised fair learning. This relates our work to the rich literature on fair learning through optimal transport [Gordaliza et al., 2019, Chiappa et al., 2020, Thibaut Le Gouic et al., 2020, Chzhen et al., 2020, Risser et al., 2022]. Finally, we note that the main result of Section 5, stating that optimal transport maps recover causal effects under specific assumptions, shares similarities with the main theorem of [Torous et al., 2021]. In contrast to our work, their assumptions are motivated by the study of heterogeneous treatment effects, which concerns counterfactual inference in the Neyman-Rubin causal framework [Rubin, 1974, Imbens and Rubin, 2015].

2 Causal modeling

Pearl's causal modeling addresses the fundamental problem of analyzing causal relations between variables, beyond mere correlations [Pearl, 2009]. It can be regarded as a mathematical formalism meant to describe associations that standard probability calculus cannot [Pearl, 2010]. This section recalls the basic theory on this modeling, borrowing the rigorous mathematical framework recently proposed by Bongers et al. [2021]. It is meant to keep the paper self-contained and can be skipped by a reader familiar with causality.

Let us fix some notations before proceeding. Throughout the paper, we consider a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. We denote respectively by $\mathcal{L}(W)$ and $\mathbb{E}[W]$ the law and expectation under $\mathbb{P}$ of any random variable W on Ω

taking values in a measurable space $(\mathbb{R}^p, \mathcal{B})$ where $p \geq 1$ and $\mathcal{B}$ is the Borel σ-algebra. Additionally, for any tuple $w := (w_i)_{i\in\mathcal{I}}$ indexed by a finite index set $\mathcal{I}$ and any subset $I \subseteq \mathcal{I}$ we write $w_I := (w_i)_{i\in I}$. Similarly, we define $\mathcal{W}_I := \prod_{i\in I} \mathcal{W}_i$ for any collection of spaces $(\mathcal{W}_i)_{i\in\mathcal{I}}$.

2.1 Definition

Causal reasoning rests on the knowledge of a *structural causal model* (SCM), which represents the causal relationships between the studied variables.

Definition 1. *Let $\mathcal{I}$ and $\mathcal{J}$ be two disjoint finite index sets, and write $\mathcal{V} := \prod_{i\in\mathcal{I}} \mathcal{V}_i \subseteq \mathbb{R}^{|\mathcal{I}|}, \mathcal{U} := \prod_{i\in\mathcal{J}} \mathcal{U}_i \subseteq \mathbb{R}^{|\mathcal{J}|}$ for two measurable product spaces. A* structural causal model *$\mathcal{M}$ is a couple $\langle U, G\rangle$ where:*

1. *$U : \Omega \to \mathcal{U}$ is a vector of random variables, sometimes called the* random seed*;*
2. *$G = \{G_i\}_{i\in\mathcal{I}}$ is a collection of measurable $\mathbb{R}$-valued functions, where for every $i \in I$ there exist two subsets of indices $\mathrm{Endo}(i) \subseteq \mathcal{I}$ and $\mathrm{Exo}(i) \subseteq \mathcal{J}$, respectively called the* endogenous *and* exogenous parents *of i, such that $G_i : \mathcal{V}_{\mathrm{Endo}(i)} \times \mathcal{U}_{\mathrm{Exo}(i)} \to \mathcal{V}_i$.*

A random vector $V : \Omega \to \mathcal{V}$ is a solution of $\mathcal{M}$ if for every $i \in \mathcal{I}$

$$V_i \overset{\mathbb{P}-a.s.}{=} G_i(V_{Endo(i)}, U_{Exo(i)}). \tag{1}$$

The collection of equations defined by (1) and characterized by G and U are called the structural equations*. By identifying G to a measurable vector function $G : \mathcal{V} \times \mathcal{U} \to \mathcal{V}$, we compactly write that V is a solution of $\mathcal{M}$ if $V \overset{\mathbb{P}-a.s.}{=} G(V, U)$.*

A structural causal model can be seen as a generative model. The variables in U are said to be *exogenous* as they are imposed *a priori* by the model. In contrast, the variables in a solution V are said to be *endogenous* as they are outputs of the model determined through the structural equations. In practice, the endogenous variables represent observed data, while the exogenous ones model latent background phenomena. Note that compared to Bongers et al. [2021], we do not assume the $(U_j)_{j\in\mathcal{J}}$ to be mutually independent.

The structural equations specify the causal dependencies between all these variables and are frequently illustrated by the directed graph defined as follows: the set of nodes is $\mathcal{I} \cup \mathcal{J}$, and a directed edge points from node k to node l if and only if $l \in \mathcal{I}$ and $k \in \mathrm{Endo}(l) \cup \mathrm{Exo}(l)$ (we say that k is a parent of l). For clarity, we often will substitute the indexes $i \in \mathcal{I}$ or $j \in \mathcal{J}$ for the variables V_i or U_j, in particular when drawing such a graph (see Figure 1). Also, similarly to Bongers et al. [2021], we will use in practice non-disjoint subsets $\mathcal{I}$ and $\mathcal{J}$ of duplicated natural integers for the sake of clarity. The example below illustrates the above notations and definitions.

Example 1. *Consider a simple SCM $\mathcal{M} := \langle U, G\rangle$ where $U := (U_1, U_2, U_3)$ is an arbitrary random vector, and such that G is defined by*

$$G_1(u_1) := u_1, \quad G_2(v_1, u_2) := v_1 + u_2, \quad G_3(v_1, v_2, u_3) := v_1 + v_2 + u_3.$$

Figure 1 represents the corresponding graph. By definition, finding a solution $V := (V_1, V_2, V_3)$ to $\mathcal{M}$ amounts to solving,

$$V_1 \overset{\mathbb{P}-a.s.}{=} U_1, \quad V_2 \overset{\mathbb{P}-a.s.}{=} V_1 + U_2, \quad V_3 \overset{\mathbb{P}-a.s.}{=} V_1 + V_2 + U_3.$$

Then, we readily obtain an almost-surely unique solution given by

$$V_1 \overset{\mathbb{P}-a.s.}{=} U_1, \quad V_2 \overset{\mathbb{P}-a.s.}{=} U_1 + U_2, \quad V_3 \overset{\mathbb{P}-a.s.}{=} 2U_1 + U_2 + U_3.$$

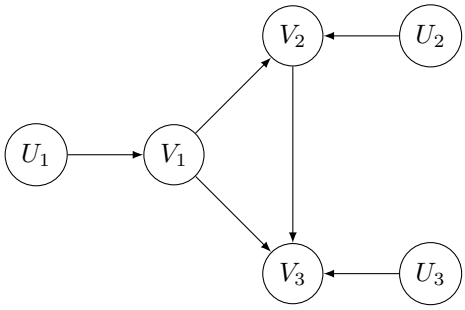


Figure 1: Example of causal graph

According to [Bongers et al., 2021, Theorem 3.3], a model $\mathcal{M} := \langle U, G \rangle$ admits a solution if and only if it is *solvable*, that is there exists a measurable function $\Gamma : \mathcal{U} \to \mathcal{V}$ such that $V \overset{\mathbb{P}-a.s.}{=} \Gamma(U) \implies V \overset{\mathbb{P}-a.s.}{=} G(V, U)$. Solvability signifies that a solution V can be expressed solely in terms of U, as in the above example. Note that SCMs are not always solvable [Bongers et al., 2021, Example 2.4]. For the sake of convenience, we make in the rest of the paper the common assumption that the considered models are *acyclic*, meaning that their graphs do not contain any cycles:

(A) *The structural causal model $\mathcal{M}$ induces a directed* acyclic *graph (DAG).*

Acyclicity entails *unique solvability* of the SCM [Bongers et al., 2021, Proposition 3.4], in the sense that Equation (1) admits a unique solution up to $\mathbb{P}$-negligible sets (as in Example 1). In particular, the generated distribution on the endogenous variables is unique. We will abusively refer to a solution as *the* solution of the SCM.

Essentially, causal structures capture the assumption that features are not independently manipulable. As we detail next, they enable to understand the downstream effect of fixing some variables to certain values onto non-intervened variables.

2.2 Do-intervention

The so-called do-calculus embodies mathematically the fundamental distinction between causation and correlation. While standard probability theory can only account for correlations through conditioning, do-calculus allows for *intervening* on variables through the do-operator. Concretely, a do-intervention is an operation mapping any model $\mathcal{M}$ to an alternative one by modifying the generative process.

Definition 2. *Let $\mathcal{M} = \langle U, G \rangle$ be an SCM, $I \subseteq \mathcal{I}$ a subset of endogenous variables, and $v_I \in \mathcal{V}_I$ a value. The action* $\mathrm{do}(I, v_I)$ *defines the modified model $\mathcal{M}_{\mathrm{do}(I,v_I)} = \langle U, \tilde{G} \rangle$ where $\tilde{G}$ is given by*

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

The model surgery described in Definition 2 consists in enforcing a state of things by substituting a set of endogenous variables by fixed values while keeping all the rest of the causal mechanism equal. By definition, do-interventions respect the exogeneity of the random seed since U remains unchanged. This transcribes the causal principle that acting upon endogenous phenomenons does not affect exogenous ones. Provided it is

solvable, the modified model $\mathcal{M}_{\mathrm{do}(I,v_I)}$ generates a new distribution of endogenous variables, describing an alternative world where every V_i for $i \in I$ is set to value v_i.

Note that do-interventions preserve acyclicity. Therefore, if an SCM $\mathcal{M}$ satisfies **(A)**, then $\mathcal{M}_{\mathrm{do}(I,v_I)}$ also satisfies **(A)**. Going further, if V is the solution of an acyclic $\mathcal{M}$, we can non-ambiguously define (up to $\mathbb{P}$-negligible sets) its intervened counterpart $V_{\mathrm{do}(I,v_I)}$ solution to $\mathcal{M}_{\mathrm{do}(I,v_I)}$. All in all, **(A)** enables to work in a convenient setting where the output of a causal model as well as the ones of its intervened counterparts are always well-defined. This implication enables to clarify the notations: in the sequel we write $\mathrm{do}(V_I = v_I)$ for the operation $\mathrm{do}(I, v_I)$, and use the subscript $V_I = v_I$ to indicate results of this operation. Crucially, intervening does not amount to conditioning in general, that is $\mathcal{L}(V \mid V_I = v_I) \neq \mathcal{L}(V_{V_I=v_I})$. This means that causal outcomes may not be observable and hence require a known causal model to be inferred, as exemplified below.

Example 2. *Let $\mathcal{M} := \langle U, G\rangle$ be the SCM from Example 1 and consider the do-intervention* $\mathrm{do}(\mathrm{V}_2 = 0)$. *This defines the intervened model $\mathcal{M}_{V_2=0} := \langle U, \tilde{G}\rangle$ where*

$$\tilde{G}_1(u_1) := u_1, \quad \tilde{G}_2(v_1, u_2) := 0, \quad \tilde{G}_3(v_1, v_2, u_3) := v_1 + v_2 + u_3.$$

Figure 2 represents the graph after surgery. The modified structural equations on a solution $\tilde{V} := (\tilde{V}_1, \tilde{V}_2, \tilde{V}_3)$ are

$$\tilde{V}_1 \overset{\mathbb{P}-a.s.}{=} U_1, \quad \tilde{V}_2 \overset{\mathbb{P}-a.s.}{=} 0, \quad \tilde{V}_3 \overset{\mathbb{P}-a.s.}{=} \tilde{V}_1 + \tilde{V}_2 + U_3.$$

Then, we readily obtain that the almost-surely unique solution is given by

$$\tilde{V}_1 \overset{\mathbb{P}-a.s.}{=} U_1, \quad \tilde{V}_2 \overset{\mathbb{P}-a.s.}{=} 0, \quad \tilde{V}_3 \overset{\mathbb{P}-a.s.}{=} U_1 + U_3.$$

Assuming that U_1, U_2, U_3 are mutually independent we have $\mathcal{L}(V_1 \mid V_2 = 0) = \mathcal{L}(U_1 \mid U_1 + U_2 = 0) = \mathcal{L}(-U_2)$ while $\mathcal{L}(\tilde{V}_1) = \mathcal{L}(U_1)$. Therefore, $\mathcal{L}(V_1 \mid V_2 = 0) \neq \mathcal{L}(\tilde{V}_1)$ in general.

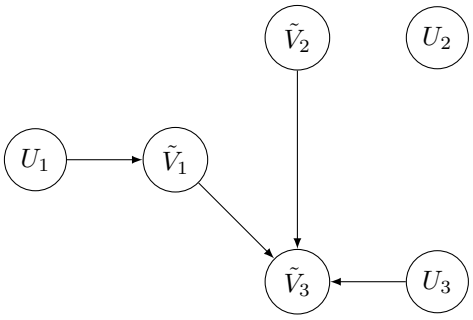


Figure 2: Intervened counterpart of Figure 1 after $\mathrm{do}(V_2 = 0)$

In Section 4, we will explain how the do-operator enables counterfactual inference from a causal model. We now turn to the second mathematical theory of interest for our work: mass transportation.

3 Mass transportation

We firstly introduce the necessary background on mass (or measure) transportation. Then, we detail the specific case of optimal transport.

3.1 Definition

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 Borel probability measures on $\mathbb{R}^d$, whose respective supports are denoted by $\text{supp}(P)$ and $\text{supp}(Q)$. We recall that the support is the set of points $x \in \mathbb{R}^d$ such that every open neighbourhood of x has a positive probability. A coupling between P and Q is a probability measure 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 $B \subseteq \mathbb{R}^d$. 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 $\mathcal{L}(Z) = P$. This coupling admits $T_{1\sharp} P$ and $T_{2\sharp} P$ as first and second marginal respectively. Thus, the deterministic coupling π between P and Q characterized by a push-forward operator T satisfying $T_\sharp P = Q$ can be written as $\pi = (I \times T)_\sharp P$ where I is the identity function on $\mathbb{R}^d$. This coupling matches a given instance $x \in \text{supp}(P)$ to $T(x) \in \text{supp}(Q)$ with probability 1.

3.2 Optimal transport

We recall here some basic knowledge on optimal transport theory, which is the mass transportation approach we focus on in this work, 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.

3.2.1 Arbitrary cost

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

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

We refer to solutions to (2) as *optimal transport maps* between P and Q with respect to c; they minimize the effort, quantified by c, of moving every elementary mass from P to Q. One mathematical complication is that the push-forward constraint renders the problem unfeasible in many general settings, in particular when P and Q are not absolutely continuous with respect to the Lebesgue measure or have unbalanced numbers of atoms.

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

$$\min_{\pi \in \Pi(P,Q)} \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x, x') \mathrm{d}\pi(x, x'). \tag{3}$$

Solutions to (3) are *optimal transport plans* (possibly non deterministic) 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. Notice that, since a push-forward operator can be identified to a coupling, the set of feasible solutions to (2) is included in the set of feasible solutions to (3).

3.2.2 Quadratic cost

Optimal transport enjoys a well-established theory, in particular when the ground cost is the squared Euclidean distance $c(x, x') := \|x - x'\|^2$ on $\mathbb{R}^d \times \mathbb{R}^d$. Suppose 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], originally proved by Cuesta and Matrán [1989] and then Brenier [1991], states that there exists a unique solution to Kantorovich's formulation of optimal transport (3), whose form is $(I \times T)_\sharp P$ where $T : \mathbb{R}^d \to \mathbb{R}^d$ solves the corresponding squared Monge problem,

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

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 Problem (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 3. *Assume that P is absolutely continuous with respect to the Lebesgue measure, and that both P and Q have finite second order moments. Then, a measurable map $T : \mathrm{supp}(P) \to \mathrm{supp}(Q)$ is a solution to (4) if and only if it satisfies 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.*

This result will play a key role in Section 5.2 to prove a link between optimal transport and causality.

3.2.3 Implementation

In practice, we do not know the measures P and Q but have access to empirical observations. This naturally raises the questions of building relevant data-driven approximations, or estimators, of the optimal transport plans, and of what should be required to ensure statistical guarantees. In this section, we briefly present the computational aspects of optimal transport, and refer to [Peyré and Cuturi, 2019] for a complete overview.

Concretely, consider two samples of i.i.d. observations $\{x_1, \dots, x_n\}$ and $\{x'_1, \dots, x'_m\}$ drawn from respectively P and Q. These samples define the empirical measures $P^n = n^{-1} \sum_{i=1}^n \delta_{x_i}$ and $Q^m = m^{-1} \sum_{i=1}^m \delta_{x'_i}$, where δ_x denotes the Dirac measure at point x. Then, the standard way to estimate an optimal transport plan between the marginals P and Q is to solve the Kantorovich formulation (3) between their empirical counterparts P^n and Q^m. By identifying a discrete coupling to a matrix, we write this problem as,

$$\min_{\pi \in \Sigma(n,m)} \sum_{i=1}^n \sum_{j=1}^m c(x_i, x'_j)\pi(i, j), \tag{5}$$

where $\Sigma(n, m) := \{\pi \in \mathbb{R}_+^{n \times m} \mid \sum_{j=1}^m \pi(i, j) = n^{-1} \text{ and } \sum_{i=1}^n \pi(i, j) = m^{-1}\}$. Note that empirical transport plans are statistically consistent. This means that if the Kantorovich problem (3) admits a unique solution π,

then a sequence $\{\pi^{n,m}\}_{n,m\in\mathbb{N}}$ of solutions to Problem (3) converges weakly to π as n and m increase to infinity [Villani, 2008, Theorem 5.19]. This property is crucial to ensure statistical guarantees in optimal-transport frameworks. We emphasize that even if a solution to Problem (5) is necessarily non-deterministic as soon as $n \neq m$, the corresponding solution to Problem (3) can be deterministic.

The main challenge when working with empirical optimal-transport solutions is that they are expensive in both computational complexity and memory: solving (5) typically requires $\mathcal{O}((n+m)nm\log(n+m))$ computer operations, and the solution is stored as an $n \times m$ matrix, which can limit the application on large datasets. Our implementation (see the experiments in Section 8) exploits the sparsity of the transport matrix to avoid overloading the memory and to speed-up the evaluation of optimal-transport-based metrics. One could also consider entropic regularization schemes to accelerate the computation of a solution to reach $\mathcal{O}(nm)$ operations [Cuturi, 2013]. However, the obtained approximation of the transport matrix is typically non sparse, hence contains many non-zero coefficients, which precludes memory-efficient implementations. This is why we address only standard optimal transport in our numerical experiments.

4 Counterfactual models

We now have all the tools to focus on the main subject of this paper: counterfactual reasoning. As mentioned in the introduction, both causality and transport techniques have been used for this purpose. However, a yet non-appreciated aspect is that these frameworks can be written in a common formalism; this is what this section addresses. More precisely, we propose the definition of *counterfactual models*, mathematical objects encoding the probabilities of all counterfactual statements with respect to modifications of one variable, and detail how to construct them with respectively causal models and mass-transportation methods.

4.1 Problem setup

Set $d \geq 1$, and define the random vector $V := (X, S) \in \mathbb{R}^{d+1}$, where the variables $X : \Omega \to \mathcal{X} \subseteq \mathbb{R}^d$ 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 $s \in \mathcal{S}$, $\mathbb{P}(S = s) > 0$. We consider the problem of computing the potential outcomes of X when changing S. Typically, S represents the sensitive, protected attribute in fairness settings, or the treatment status in the potential-outcome framework. 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?* Critically, because of correlations or structural relations between the variables, computing the alternative state does not amount to change the value of S while keeping the features X equal.

4.2 Structural counterfactuals

Answering the counterfactual question from Section 4.1 with Pearl's framework requires to assume causal dependencies between X and S. Formally, suppose that $V = (X, S) \in \mathbb{R}^{d+1}$ is the unique solution to an SCM $\mathcal{M} = \langle U, G\rangle$ satisfying the acyclicity assumption **(A)**. We recall that each *endogenous* variable V_k is then defined (up to sets of probability zero) by the structural equation

$$V_k \overset{\mathbb{P}-a.s.}{=} G_k\left(V_{\mathrm{Endo}(k)}, U_{\mathrm{Exo}(k)}\right),$$

where G_k is a real-valued measurable function, U is a vector of *exogenous* variables, while $V_{\mathrm{Endo}(k)}$ and $U_{\mathrm{Exo}(k)}$ denote respectively the endogenous and exogenous parents of V_k. In the following, we denote by U_X and U_S

the exogenous parents of respectively X and S. We write $X_{S=s}$ the intervened counterpart of X through the do-intervention $\mathrm{do}(S = s)$, that is after replacing the structural equation on S by $S = s$ while keeping the rest of the causal mechanism equal.

Then, 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, $\mu_s := \mathcal{L}(X \mid S = s)$ is the distribution of the *factual* s-instances. This observable measure describes the possible values of X such that $S = s$, and we write $\mathcal{X}_s$ for its support. 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.

4.2.1 Definition

Using the above notation, our problem can be framed as: having observed an $x \in \mathcal{X}_s$, determining the probability of the counterfactual outcome $x' \in \mathrm{supp}(\mu_{\langle s'|s\rangle})$. Pearl originally answered this question with the following *three-step procedure*: (1) set a prior $\mathcal{L}(U)$ for the SCM, (2) compute the posterior distribution $\mathcal{L}(U \mid X = x, S = s)$, and (3) solve the structural equations after the intervention $\mathrm{do}(S = s')$ with $\mathcal{L}(U \mid X = x, S = s)$ as input. This leads to the following formal definition of *structural counterfactuals*, adapted from [Pearl et al., 2016, Chapter 4].

Definition 4. *Let $\mathcal{M}$ satisfy **(A)**. For an observed evidence $\{X = x, S = s\}$ and an intervention $\mathrm{do}(S = s')$, the* structural counterfactuals *of X are characterized by the probability distribution $\mu_{\langle s'|s\rangle}(\cdot|x)$ defined as*

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

In general, 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 $\{X = x, S = s\}$. This means that, according to Pearl's causal reasoning, there is not necessarily a one-to-one correspondence between factual instances and counterfactual counterparts, but a collection of weighted correspondences described by the distribution of structural counterfactuals.

4.2.2 Mass-transportation viewpoint

While the mainstream literature on causality generally operates with the definition of structural counterfactuals given by the three-strep procedure [Kusner et al., 2017, Barocas et al., 2019], we focus in this paper on a *mass-transportation viewpoint* of counterfactuals, formalized by the following definition.

Definition 5. *Let $\mathcal{M}$ satisfy **(A)**. 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 *on X with respect to S.*

In this formalism, the quantity $\mathrm{d}\pi_{\langle s'|s\rangle}(x, x')$ 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*. As such, a counterfactual model characterizes the distribution of all the cross-world statements on X with respect to changes of S. Note that each realization of $\pi^*_{\langle s'|s\rangle}$, that is each pair of factual instance and counterfactual counterpart, is generated by a same possible value of $\mathcal{L}(U_X \mid S = s)$.

We point out that Definitions 4 and 5 characterize the exact same counterfactual statements, the formal link being $\mathrm{d}\pi^*_{\langle s'|s\rangle}(x, x') = \mu_{\langle s'|s\rangle}(x'|x)\mathrm{d}\mu_s(x)$. In particular, there is an equivalence between $\mu_{\langle s'|s\rangle}(\cdot|x)$ narrowing down to a single value for every $x \in \mathcal{X}_s$ and $\pi^*_{\langle s'|s\rangle}$ being a deterministic coupling. Assumptions rendering single-valued counterfactuals will be studied in Section 5.1.1. We also note that this joint-probability-distribution perspective of Pearl's counterfactuals concurs with the one from [Bongers et al., 2021, Section 2.5].

4.3 Transport-based counterfactuals

The main issue of structural counterfactuals, which will be widely discussed in Section 4.5, comes from the causal model being unknown in practice. Thus, the necessity to make counterfactual frameworks feasible naturally raises the question of finding good surrogates to causal counterfactuals. We have seen that the problem of assessing counterfactual statements about X with respect to interventions on S using causal models could be reduced to knowing a collection of random mappings from factual distributions $\{\mu_s\}_{s\in\mathcal{S}}$ towards counterfactual distributions $\{\mu_{\langle s'|s\rangle}\}_{s,s'\in\mathcal{S}}$. This perspective suggests that mass-transportation techniques can be natural substitutes for structural counterfactual reasoning, as they remedy to the aforementioned issues.

4.3.1 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 switching S from s to s' can be seen as a mass transportation plan, not necessarily optimal-transport based and not necessarily deterministic, between two distributions.[1] In the following, $t : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}^d, (x, x') \mapsto (x', x)$ denotes the permutation function.

Definition 6. *A* transport-based counterfactual model *is a collection of couplings $\Pi := \left\{\pi_{\langle s'|s\rangle}\right\}_{s,s'\in\mathcal{S}}$ satisfying for every $s, s' \in \mathcal{S}$,*

(i) $\pi_{\langle s'|s\rangle} \in \Pi(\mu_s, \mu_{s'})$;

(ii) $\pi_{\langle s|s\rangle} = (I \times I)_\sharp \mu_s$;

(iii) $\pi_{\langle s|s'\rangle} = t_\sharp \pi_{\langle s'|s\rangle}$.

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. Otherwise, we say that Π is a* deterministic *counterfactual model. In the deterministic case, Π can be identified almost everywhere to a collection $\mathcal{T} := \{T_{\langle s'|s\rangle}\}_{s,s'\in\mathcal{S}}$ of measurable mappings from $\mathcal{X}$ to $\mathcal{X}$ satisfying for every $s, s' \in \mathcal{S}$,*

(i) $T_{\langle s'|s\rangle\sharp}\mu_s = \mu_{s'}$;

(ii) $T_{\langle s|s\rangle} = I$;

[1] In [Asher et al., 2022, Section 7.2], we present this view of counterfactuals from a logic perspective.

(iii) $T_{\langle s'|s\rangle}$ *is invertible* μ_s*-almost everywhere such that* $T_{\langle s|s'\rangle} = T^{-1}_{\langle s'|s\rangle}$.

An element of $\mathcal{T}$ *is called a* counterfactual operator.

Similarly to structural counterfactual models, these models assign a probability to all the cross-world statements on X with respect to interventions on S. By convention, we use the superscript * to denote *structural* counterfactual models, and no superscript for *transport-based* counterfactual models. The marginal constraint (i) in Definition 6 translates the intuition that a realistic counterfactual operation on S should morph the non-intervened variables X so that their values fit the targeted distribution. In this sense, transport-based models preserve the principle that features are not independently manipulable, but without using causal relations. The symmetry constraints (ii) and (iii) cover the reciprocity intuition we have on counterfactuals counterparts. Remark that in the case of discrete measures, the operation $t_\sharp$ in condition (iii) simply amounts to transposing the associated coupling matrices. Lastly, note that this definition replaces the unobservable, SCM-dependent distributions $\{\mu_{\langle s'|s\rangle}\}_{s,s'\in\mathcal{S}}$ of structural counterfactual models by the observational $\{\mu_{s'}\}_{s'\in\mathcal{S}}$ for feasibility reasons. In Section 5.1.2, we will see that this approximation makes sense in typical fairness settings where $\mu_{\langle s'|s\rangle} = \mu_{s'}$ for every $s, s' \in \mathcal{S}$.

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 $x \in \mathcal{X}_s$ after changing S from s to s' is given by $x' = T_{\langle s'|s\rangle}(x) \in \mathcal{X}_{s'}$. In contrast, a *random* model allows possibly several counterparts with probability weights. The first interest of considering random couplings is purely conceptual; rendering non necessarily unique the counterfactual counterparts of a single instance has philosophical implications [Asher et al., 2022, Section 6.3]. Besides, it is consistent with the causal approach which also authorizes non-deterministic counterfactuals. The second—and most critical benefit—is practical. While there always exist random couplings between two distributions, deterministic push-forward mappings (even causally-induced ones) may not exist when the marginals do not have densities, making this relaxation crucial for dealing with non-continuous variables. This makes the extension to random couplings necessary to tackle concrete machine-learning tasks, involving both continuous and discrete covariates. Notably, we rely on random couplings in the numerical experiments from Section 7.

4.3.2 Choosing a model

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 6, 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. To sum-up, a transport-based counterfactual model must be both *intuitively justifiable* and *computationally feasible*.

We argue that optimal-transport solutions are tailored couplings with respect to both criteria. Optimal transport has become the most popular tool in statistics-related fields to define couplings between distributions when no canonical choice is available, as in generative modeling [Balaji et al., 2020], domain adaptation [Courty et al., 2014, 2017, Redko et al., 2019, Rakotoarison et al., 2022], and transfer learning [Gayraud et al., 2017, Peterson et al., 2021] thanks to significant advances in computational schemes. Additionally, as argued by Black et al. [2020], generating a counterfactual operation by solving the optimal-transport Problem (2) leads to intuitively relevant counterfactuals, as they are obtained by minimizing a metric between paired instances

(transcribing the Lewisian most-similar-alternative-world principle) while preserving the probability distributions (ensuring distributional faithfulness). Moreover, deterministic optimal transport for the quadratic cost (see Section 3.2.2) has remarkable properties. According to Lemma 3, solutions to Problem (4) are gradients of convex functions, which extends the notion of non-decreasing function to several dimensions. In particular, the optimal transport map in dimension one is the quantile-preservation map between univariate distributions. This behaviour has notably inspired constructions of multivariate notions of quantile based on optimal transport [Chernozhukov et al., 2017, Hallin et al., 2021, Ghosal and Sen, 2022]. It also makes sense in counterfactual reasoning where, without further information on the data-generation process, preserving the quantile from on marginal to the other is an intuitive definition of the counterfactual counterpart. For the sake of illustration, Section 4.4 below provides several examples of optimal transport applied to counterfactual reasoning.

In Section 5.2 we will further justify the pertinence of *optimal-transport-based* counterfactual models by showing that they coincide with structural counterfactual models under some assumptions. However, the scope of Definition 6 goes beyond solutions to standard optimal-transport problems, allowing other transport methods and as such more possible counterfactual models. The purpose of this generalization is partly theoretic: in the future, one could propose an original matching technique and justify its relevance compared to optimal transport. In particular, the couplings mentioned in [Villani, 2008, Chapter 1] as well as diffeomorphic registration mappings [Joshi and Miller, 2000, Beg et al., 2005] are possible candidates we do not investigate in this paper. Additionally, this generalization permits the use of regularized optimal transport [Cuturi, 2013], which deviates from the original formulation of Problem (3), to accelerate computations. Note in passing that solutions to regularized optimal transport, which are non deterministic, define adequate transport-based counterfactual models thanks to Definition 6 taking into account random couplings. Lastly, we will see in Section 5.1.2 that structural counterfactual models are transport-based counterfactual models—but not necessarily optimal-transport-based—under some assumptions.

4.4 Examples

Now that we gave definitions and insights on counterfactual models, let us study two concrete examples on real data.

4.4.1 Law dataset

We start by focusing on the Law School Admission Council dataset which gathers statistics from 163 US law schools and more than 20,000 students, including four variables: the race S, the entrance-exam score X_1, the grade-point average before law school X_2, and the first-year average grade Y. The end goal is to predict the first-year grade Y from the other features (X, S). Similarly to Russell et al. [2017], we consider a fairness setting where the race plays the role of a protected, sensitive attribute which should not be discriminated against, and we restrict to only black ($S = 0$) and white ($S = 1$) students. Counterfactual reasoning has become popular in such algorithmic fairness tasks to either ensure or test that, for example, had a black student been white, the output would have been the same. This requires a model to compute the counterfactual counterparts of any students after changing their skin colors.

First, we consider a structural counterfactual model. This requires a causal model: Russell et al. [2017]

proposed the following SCM for the dataset,

$$\begin{cases} X_1 = b_1 + w_1 S + U_1, \\ X_2 = b_2 + w_2 S + U_2, \\ S = U_S, \\ U_S, U_1, U_2 \text{ independent}, \end{cases}$$

where $b := (b_1, b_2)$ and $w := (w_1, w_2)$ are deterministic $\mathbb{R}^2$ parameters obtained by adjusting linear-regression models component-wise. Let us now calculate the induced structural counterfactual model by applying Definition 5. The coupling from $S = 0$ to $S = 1$ is given by

$$\pi^*_{\langle 1|0\rangle} := \mathcal{L}\big((X, X_{S=1}) \mid S = 0\big) = \mathcal{L}\big((b + U_X, b + w + U_X)\big) = \mathcal{L}\big((X, X + w) \mid S = 0\big).$$

Conversely, the structural counterfactual coupling from $S = 1$ to $S = 0$ is

$$\pi^*_{\langle 0|1\rangle} := \mathcal{L}\big((X, X_{S=0}) \mid S = 1\big) = \mathcal{L}\big((b + w + U_X, b + U_X)\big) = \mathcal{L}\big((X, X - w) \mid S = 1\big).$$

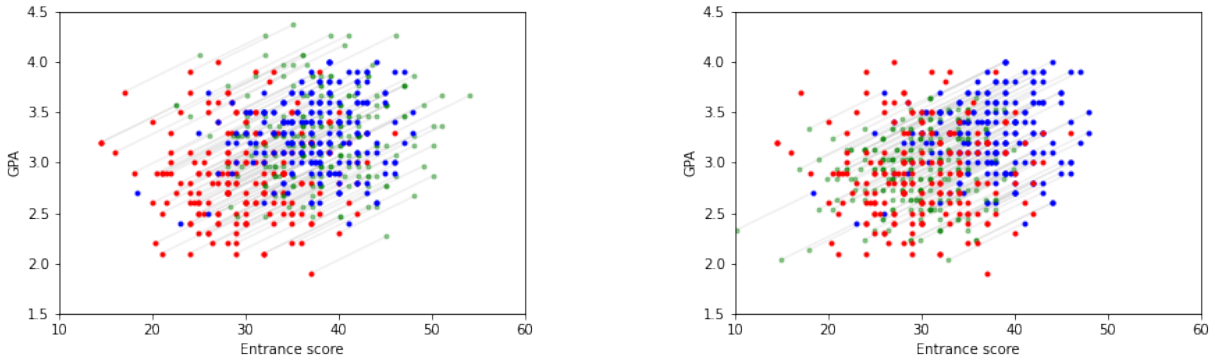
Figures 3a and 3b illustrate the computation of the corresponding counterfactual counterparts on samples. We make two important remarks.

Firstly, generating counterfactual quantities in this case amounts to translating instances of μ_0 by the constant w or conversely translating instances of μ_1 by the constant $-w$. Notably, the two couplings are deterministic: $\pi^*_{\langle 1|0\rangle}$ and $\pi^*_{\langle 0|1\rangle}$ are respectively characterized by the mappings $T^*_{\langle 1|0\rangle}(x) := x + w$ and $T^*_{\langle 0|1\rangle}(x) := x - w$. Note that there is consequently no need to specify the law of the exogenous variables to compute counterfactual quantities. Section 5.1.1 provides a general analysis of such deterministic settings.

Secondly, the causal model implies that $S \perp\!\!\!\perp U_X$. This critically entails that the counterfactual distributions are observable, since $\mu_{\langle 1|0\rangle} = \mathcal{L}(X_{S=1} \mid S = 0) = \mathcal{L}(b + w + U_X \mid S = 0) = \mathcal{L}(b + w + U_X \mid S = 1) = \mu_1$ and $\mu_{\langle 0|1\rangle} = \mu_0$ analogously. Therefore, the structural counterfactual couplings $\pi^*_{\langle 1|0\rangle}$ and $\pi^*_{\langle 0|1\rangle}$ belong respectively to $\Pi(\mu_0, \mu_1)$ and $\Pi(\mu_1, \mu_0)$. Additionally, they are transposed from one another, that is $t_\sharp \pi^*_{\langle 1|0\rangle} = \pi^*_{\langle 0|1\rangle}$. This means that the structural counterfactual model $\Pi^* := \{\pi^*_{\langle 1|0\rangle}, \pi^*_{\langle 0|1\rangle}\}$ is a transport-based counterfactual model. Mathematical justifications of these properties will be studied in Section 5.1.2.

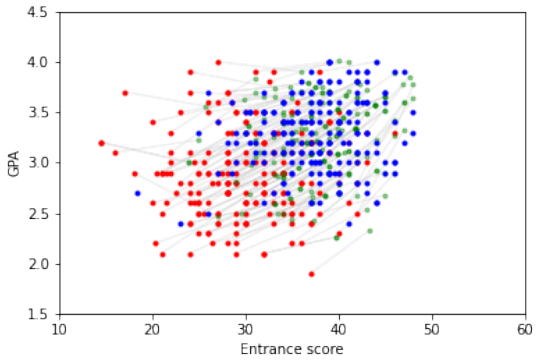
In a second time, we turn to an optimal-transport-based counterfactual model. More precisely, we learn the optimal transport map for the quadratic cost, denoted by $T_{\langle 1|0\rangle}$, from the black distribution μ_0 towards the white distribution μ_1. In practice, we rely on the Python Optimal Transport (POT) library to compute an approximation of the mapping from data [Flamary et al., 2021]. Note that solving the empirical optimal-transport problem (5) between samples provides a matching that cannot generalize to new, out-of-sample observations. This is why we employ POT's in-built non-regularized barycentric extension of the empirical solution to obtain a mapping defined everywhere. We use 800 points from each distribution to compute the estimator of $T_{\langle 1|0\rangle}$ illustrated in Figure 3a. The converse counterfactual operation $T_{\langle 0|1\rangle}$ represented in Figure 3d is produced by inversion.

We emphasize that all the couplings in Figure 3, be they causal-based or optimal-transport-based, are imperfect approximations, but for different reasons. More precisely, we assumed that a linear causal model generated the data in order to compute the structural counterfactual couplings. However, this model-class assumption is not a perfect fit: in particular, some of the produced counterfactual instances are not realistic, yielding GPA scores exceeding the upper limit of 4.0 points; more generally, while both couplings should have μ_0 and μ_1 for marginals, several counterfactual counterparts do not conform to these distributions. Besides, the

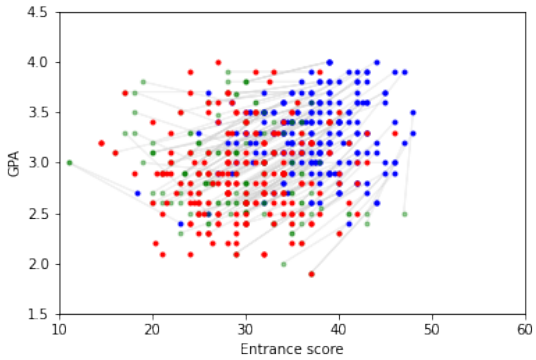


(a) $T^*_{\langle 1|0\rangle}$

(b) $T^*_{\langle 0|1\rangle}$



(c) $T_{\langle 1|0\rangle}$



(d) $T_{\langle 0|1\rangle}$

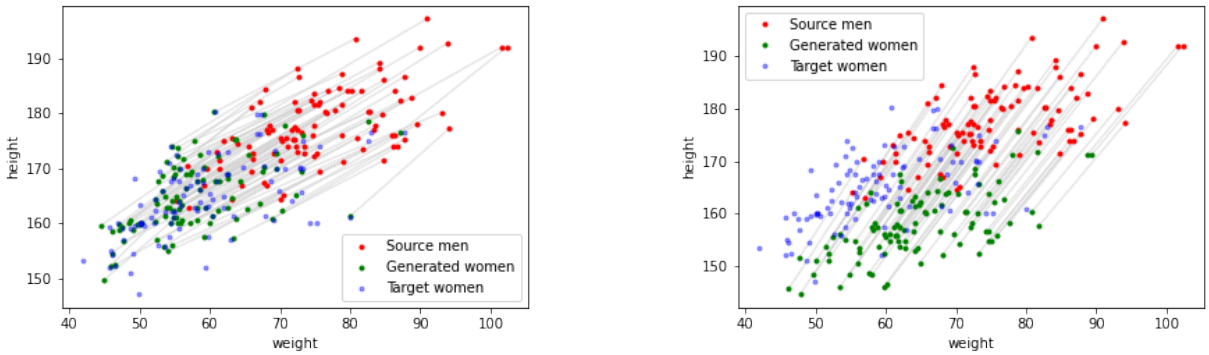
Figure 3: Counterfactual models for the Law dataset. The red sample represents 200 factual black students while the blue sample represents 200 factual white students. The green sample depicts counterfactual instances: the first column (Figures 3a and 3c) has white counterfactual students; the second column (Figures 3b and 3d) has black counterfactual students. The lightgray lines describe the coupling between factual and counterfactual instances.

translation vector w used in practice is an estimation from data, thereby an approximation of the best linear model fitting the data. The implemented optimal-transport mappings are also mere estimators of the "true" mappings between the continuous distributions. Figure 3c notably shows poor counterfactual associations for outliers of the red sample, likely due to weak estimation in low-density domains. Nevertheless, the marginal constraint of optimal transport ensures that the generated counterfactuals faithfully fit the data and are therefore plausible. Finally, despite these approximation artifacts, we remark that the causal and optimal-transport couplings have fairly similar behaviours, siding with the observations of Black et al. [2020]. This proximity will be theoretically grounded in Section 5.2.

4.4.2 Body-measurement dataset

We now further illustrate the properties of optimal-transport counterfactuals on a dataset of body measurements from $n_0 = 260$ women and $n_1 = 247$ men. The features of interest are the weight X_1 and the height X_2, while S

encodes the gender. Suppose now that Bob is a 80kg and 190cm man. What would have been Bob's height and weight had he been a woman? Since we do not know the structural relationships between X, S and possibly hidden sources of randomness U, we follow Black et al. [2020] and rely on mass-transportation techniques to answer this counterfactual question. We proceed as before to estimate the optimal transport map from the male distribution μ_1 towards the female distribution μ_0. Applying this operator to Bob, we obtain that, had he been a women, she would have been 59kg and 177cm.



(a) 2D optimal-transport matching.

(b) 2D component-wise quantile-preservation.

Figure 4: Body dataset. The red dots represent a data sample of men, while the blue dots represent a data sample of women. The green dots are the estimated counterfactual counterparts of the male sample.

Though it does not have a canonical definition when $d = 2$, optimal transport seems visually to preserve the "position" of the paired points from one marginal to another. This is due to the optimal map being the unique gradient of a convex function between distributions as previously explained. We underscore in Figure 4 that optimal transport does not amount to feature-wise quantile-preservation, making it a relevant extension of the notion of order to higher dimension. Notably, preserving the quantile along each coordinate does not satisfy the marginal constraint, yielding counterfactual women not representative of their gender's distribution.

4.5 Discussion

Counterfactuals have valuable applications in fairness and explainability. One could for example try to learn predictor h designed to make $h(x, 0)$ as close as possible to $h(x', 1)$ for every counterfactual pair (x, x'). This is what Russell et al. [2017] proposed using causal models, and what we implement in Section 7 using transport-based models. Or, one could test whether a trained predictor h is unfair by checking if $h(x, 0) = h(x', 1)$ for every counterfactual pair (x, x'), which is essentially the procedure of Black et al. [2020] leveraging optimal transport maps. However, the application of counterfactual models raises several issues. We conclude Section 4 by discussing important drawbacks of the causal account to counterfactual reasoning as well as the limitations of the transport approach.

4.5.1 Shortcomings of the causal approach

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 since a causal model must be designed and tested for each possible dataset. A more straightforward approach is to directly infer the causal model from observational data. There exist for instance sound techniques to learn the causal graph, but 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. To obtain these equations, researchers often predefine the functional form of the relations between the variables on the basis of a known graph (be it assumed or inferred) and learn them through regression models [Kusner et al., 2017, Russell et al., 2017], or infer simultaneously the graph and the structural equations. However, this also becomes computationally challenging as the number of features increases. Notably, the literature mostly addresses simple linear models [Shimizu et al., 2006] or very few variables [Hoyer et al., 2008]. Finally, the approximation error implied by the choice of the functional class can lead to unrealistic, out-of-distribution counterfactuals, as exemplified in Figure 3 above. To our knowledge, the literature on causal counterfactuals has not pointed out this flaw to date.

A related issue is causal uncertainty. There exist several causal models corresponding to a same data distribution, leading to possibly different counterfactual models [see Bongers et al., 2021, Example 4.2]. 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. Assumption **(A)** is very common for both practical reasons and reasons of interpretability. In general, however, observational data can be generated through 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 [Bongers et al., 2021, Example 2.17] for an illustration. As a consequence, counterfactual quantities are ill-defined in such settings.

4.5.2 Applicability of the transport approach

Regarding transport-based counterfactual reasoning, the main practical limitation is also computational. The domain $\mathcal{S}$ of the intervened variable S must be finite for the counterfactual model to be tractable. Moreover, generating the model needs $|\mathcal{S}|\left(|\mathcal{S}|-1\right)/2$ computations of transportation plans, which can become too expensive when $|\mathcal{S}|$ is large. Therefore, this approach is tailored to settings with small $|\mathcal{S}|$, typically fairness problems where S represents gender or race.

Another inconvenience comes from the fact that one must specify a family of couplings to implement a transport-based counterfactual model. There is no quantitative rule for this choice; it is guided by intuition and feasibility reasons, and we explained above why optimal transport was a relevant option. Note that the causal approach has a similar flaw: as previously explained, structural counterfactual models are subjected to misspecification since the underlying causal model itself is uncertain. The advantage of transport methods compared to causal modeling is that they circumvent possibly wrong assumptions on the data-generative process. In particular, transport plans consistently adjust to the data (thanks to the marginal constraint) regardless of the chosen family of couplings, whereas misspecification of the SCM may lead to out-of-distribution structural

counterfactuals as aforementioned.

In the following, we derive theoretical properties of the counterfactual models introduced in this section, grounding the similarity between optimal transport and Pearl's computation of counterfactuals we evidenced in Figure 3. Interestingly, this echoes the work of Black et al. [2020], who also empirically observed that optimal transport maps generated nearly identical counterfactuals to the ones based on causal models.

5 Theoretical results

Until now, we have recalled the basics of causality and transport in Sections 2 and 3, and introduced counterfactual models, either causal-based or transport-based, in Section 4. In what follows, we demonstrate connections between both approaches. Concretely, we firstly explore in Section 5.1 the relationship between an SCM and the counterfactual model it induces, providing justifications to what we observed in Section 4.4.1. More precisely, we study the implications of typical causal assumptions onto the generated counterfactuals. Then, on the basis of these assumptions and the mass-transportation formalism proposed in Section 4, we demonstrate in Section 5.2 that optimal transport recovers structural counterfactuals in specific cases.

5.1 Causal assumptions and their consequences

We analyze in detail two standard scenarios of the causal 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. Note that none of Section 5.1 involves any specific knowledge on optimal transport theory, only on causal modeling and (general) mass transportation.

5.1.1 The deterministic case

We show that when the SCM deterministically implies the counterfactual values of X, then the counterfactual coupling is deterministic. Additionally, we provide the expression of the corresponding push-forward operator. To reformulate structural counterfactuals in deterministic transport terms, we first highlight the functional relation between an instance and its intervened counterparts.

Lemma 7. *If $\mathcal{M}$ satisfies **(A)**, then there exists a measurable function F such that $X \overset{\mathbb{P}-a.s.}{=} F(S, U_X)$ and $X_{S=s} \overset{\mathbb{P}-a.s.}{=} F(s, U_X)$ for every $s \in S$.*

The proof leverages the acyclicity of the structural equations, which implies that the system of structural equations defining X and S is triangular, enabling to express X solely in terms of U_X and S.

Now, let us set for every $s \in \mathcal{S}$ the function $f_s : u \mapsto F(s, u)$ defined $\mathcal{L}(U_X)$-almost everywhere. Using this notation, we can give a simple expression of the possible counterfactual counterparts of any factual instance. In what follows, $\overline{B}$ denotes the closure of any $B \subseteq \mathbb{R}^d$.

Proposition 8. *Let $\mathcal{M}$ satisfy **(A)**. For any $s, s' \in \mathcal{S}$ and μ_s-almost every $x \in \mathcal{X}_s$,*

$$\operatorname{supp}\left(\mu_{\langle s'|s\rangle}(\cdot|x)\right) \subseteq \overline{f_{s'} \circ f_s^{-1}(\{x\})}.$$

As a direct consequence of this proposition, all counterfactual quantities on X with respect to S are uniquely determined when the right term of the inclusion becomes a singleton, therefore when the following assumption holds.

Assumption (I) *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, **(I)** states that, conversely, U_X is deterministically determined by $\{X = x, S = s\}$. This assumption holds in particular for *additive-noise models*: classical models where the exogenous variables are additive terms of the structural equations, such as in Example 1 and Section 4.4.1.

Example 3. *An SCM* $\mathcal{M} = \langle U, G\rangle$ *is an* additive-noise model *if its causal mechanism* G *has the form*

$$G(v, u) := \phi(v) + u,$$

where $\phi : \mathcal{V} \to \mathcal{V}$ *is a measurable function. Under* ***(A)****, therefore unique solvability, each endogenous variable* V_k *is given by*

$$V_k \overset{\mathbb{P}-a.s.}{=} \phi_k\left(V_{Endo(k)}\right) + U_{Exo(k)},$$

where $\phi_k : \mathcal{V}_{\mathrm{Endo}(k)} \to \mathcal{V}_k$*. Note that the random seed* U *is fully determined by the value of* V*, meaning that for any* $v \in V$ *the posterior distribution* $\mathcal{L}(U \mid V = v)$ *narrows down to a single value. As such, whatever the do-intervention on* V*, the three-step procedure can only generate a deterministic counterfactual quantity.*

Note that in our setting, which addresses interventions on a single endogenous variable S*, satisfying* ***(I)*** *does not require a fully invertible model between* $V = (X, S)$ *and* U *but simply between* X *and* U_X *knowing* $S = s$*. As illustration, consider a partially-additive-noise model (over* X *only), namely such that* X *is generated through*

$$X \overset{\mathbb{P}-a.s.}{=} \varphi\left(S, X\right) + U_X,$$

where $\varphi : \mathcal{S}\times\mathcal{X} \to \mathcal{X}$ *is a deterministic measurable function; the equation on* S *does not matter. Assumption* ***(A)*** *entails through unique solvability that* $X \overset{\mathbb{P}-a.s.}{=} (I - \varphi(S, \cdot))^{-1}(U_X)$*. After identifying* $f_s(u) := (I - \varphi(s, \cdot))^{-1}(u)$*, we notice that Assumption* ***(I)*** *readily holds such that* $f_s^{-1}(x) = x - \varphi(s, x)$*.*

Remark that Assumption **(I)** imposes constraints on the variables and their laws to enable a deterministic correspondence between X and U_X. In particular, the two random vectors must live in spaces with same cardinal, preventing for instance a continuous U_X with a discrete X. Note also that even though it is restrictive, the mainstream literature on causality frequently assumes full invertibility. In particular, most of the causal-discovery frameworks which aim at inferring the structural equations from observational data require invertible models [Zhang and Chan, 2006, Hoyer et al., 2008] or even additive ones [Shimizu et al., 2006]. Analogously, the recent research on causal algorithmic recourse generally addresses invertible models in both theory and practice [Karimi et al., 2021, Dominguez-Olmedo et al., 2022, von Kügelgen et al., 2022]. In Section 5.2, we will use the invertibility assumption as an ideal setting to derive theoretical guarantees.

Let us finally turn to the structural counterfactual models. Assumption **(I)** implies that all the couplings between the factual and counterfactual distributions are deterministic, as written in the next proposition.

Proposition 9. *Let* $\mathcal{M}$ *satisfy* ***(A)****, suppose that* ***(I)*** *hold, and for any* $s, s' \in \mathcal{S}$ *set the mapping* $T^*_{\langle s'|s\rangle} := f_{s'} \circ f_s^{-1}|_{\mathcal{X}_s}$ *defined* μ_s*-almost everywhere, where* $f_s^{-1}|_{\mathcal{X}_s}$ *denotes the restriction of* f_s^{-1} *to* $\mathcal{X}_s$*. The following properties hold:*

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

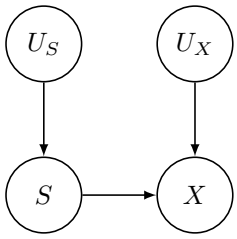


Figure 5: DAG of a structural causal model satisfying **(RE)**. The nodes U_X and X possibly represent several variables.

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 they are well-defined without any knowledge on $\mathcal{L}(U)$, meaning that the exogenous variables are not necessary to compute counterfactual quantities under **(I)**.

Lastly, remark that we framed **(I)** so that it implies that all the counterfactuals instances for *any* changes on S are deterministic, leading to a fully-deterministic counterfactual model.[2] However, according to Proposition 8, 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 **(I)** does not hold, the structural counterfactual model possibly contains both random and deterministic couplings.

5.1.2 The exogenous case

We now 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. We will see that 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* $X_{Endo(S)} = \emptyset$.

The first item, $U_S \perp\!\!\!\perp U_X$, ensures that there is no hidden confounder between X and S. The literature on causal modeling generally supposes a stronger condition known as *causal sufficiency*, which states that *all* the $(U_j)_{j\in\mathcal{J}}$ are mutually independent [Shimizu et al., 2006, Karimi et al., 2021, Bongers et al., 2021, Dominguez-Olmedo et al., 2022]. The second item, $X_{\mathrm{Endo}(S)} = \emptyset$, means that S is *ancestrally closed*: no variable in X is a direct cause (or parent) of S (see Figure 5). This holds typically in fairness problems, such as in Section 4.4.1, where the variable S to alter generally encodes someone's gender, race or age, which do not have any observable causes. As pointed out by Fawkes et al. [2021], ancestral closure is a common hypothesis in causal-fairness research, and even a requirement for many frameworks [Kusner et al., 2017, Russell et al., 2017, Nabi and Shpitser, 2018, Chiappa, 2019, Kilbertus et al., 2020, Plecko and Meinshausen, 2020].

Interestingly, relative exogeneity has critical implications on the generated counterfactuals. Assumption **(RE)** readily 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.

[2] In logic terms, this means that the model verifies the conditional excluded middle [Stalnaker, 1980].

Proposition 10. *Let $\mathcal{M}$ satisfy **(A)**. If **(RE)** holds, then for every $s, s' \in \mathcal{S}$ we have $\mu_{S=s'} = \mu_{s'} = \mu_{\langle s'|s \rangle}$.*

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 10, **(RE)** renders the causal model otiose for the purpose of generating the counterfactual distribution, as the latter coincides with the observable factual distribution $\mu_{s'}$. This is notably what occurred in the example from Section 4.4.1. However, we underline that the coupling is *still required* to determine how each instance is matched at the individual level. As such, the causal model still carries major information on the induced counterfactual quantities.

Besides, as remarked by Plecko and Meinshausen [2020] and Fawkes et al. [2021], a practical consequence of **(RE)** is that it enables to link observational and causal notions of fairness. In Section 6, we will prove a similar result through the prism of counterfactual models. The demonstration relies on the proposition below, which ensures that structural counterfactual models are transport-based counterfactual models when S is relatively exogenous to X.

Proposition 11. *Let $\mathcal{M}$ satisfy **(A)**. If **(RE)** holds, then for any $s, s' \in \mathcal{S}$,*

(i) $\pi^*_{\langle s'|s \rangle} \in \Pi(\mu_s, \mu_{s'})$;

(ii) $t_\sharp \pi^*_{\langle s'|s \rangle} = \pi^*_{\langle s|s' \rangle}$.

*Suppose additionally that **(I)** holds. Then, for any $s, s', s'' \in \mathcal{S}$,*

(iii) $T^*_{\langle s'|s \rangle \sharp} \mu_s = \mu_{s'}$;

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

Notably, this means that in classical fairness settings transport-based models can be seen as approximations, relaxations of structural models. Another meaningful consequence of Proposition 11 is that items (ii), (iv) and (v) may be false when **(RE)** does not hold. Said differently, in general contexts, there is no reciprocity between a factual instance and its structural counterfactual counterparts.

5.1.3 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 many state-of-the-art models [Bentler and Weeks, 1980, Shimizu et al., 2006, Hyttinen et al., 2012, Rothenhäusler et al., 2021].

Example 4. *Under **(RE)** and **(A)**, 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 such that $I - M$ is invertible, and $U_S \perp\!\!\!\perp U_X$. Solving the equations we get $X = (I - M)^{-1}(wS + b + U_X) =: F(S, U_X)$. Besides, note that **(I)** holds such that for any $s \in \mathcal{S}$, $f_s^{-1}(x) = (I - M)x - ws - b$. Then, for any $s, s' \in \mathcal{S}$, $T^*_{\langle s'|s \rangle}(x) = x + (I - M)^{-1} w(s' - s)$. This general expression is consistent with the example from Section 4.4.1.*

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 **(RE)**, it is presumably difficult to infer the counterfactual model from data. We address this issue the next section. Specifically, we show that optimal transport for the quadratic cost generates the same counterfactuals as a class of causal models including linear additive models.

5.2 When optimal transport meets causality

We focus on the deterministic transport-based counterfactual model $\mathcal{T} = \{T_{\langle s|s'\rangle}\}_{s,s'\in\mathcal{S}}$ defined by the solutions of Problem (4) between all pairs of factual distributions. That is, for every $s, s' \in \mathcal{S}$,

$$T_{\langle s'|s\rangle} := \underset{T:\ T_\sharp\mu_s=\mu_{s'}}{\arg\min} \int_{\mathcal{X}_s} \|x - T(x)\|^2 \mathrm{d}\mu_s(x). \tag{6}$$

As explained before in Section 4, this model provides an elegant interpretation to the obtained counterfactual statements, as they are defined by minimizing the squared Euclidean distance between paired instances, and preserve the quantile between marginals when $d = 1$. Moreover, as stated in the following theorem, this transport-based counterfactual model recovers structural counterfactuals in specific cases.

Theorem 12. *Let $\mathcal{M}$ satisfy **(A)**, **(RE)** and **(I)**. Suppose that the factual distributions $\{\mu_s\}_{s\in\mathcal{S}}$ are absolutely continuous with respect to the 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 Problem* (6).

The mass-transportation formalism of Pearl's counterfactual reasoning introduced in Section 4.2 and developed in Section 5.1 renders the proof of this theorem straightforward. The non triviality comes precisely from the reformulation of deterministic structural counterfactuals through push-forward operators. We underline that the demonstration does not require any prior knowledge on optimal transport theory except what we summarized in Lemma 3. Thus, for the sake of illustration and clarity, we reproduce it directly below.

Proof According to **(I)** and Proposition 9, 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 10 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 with respect to 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 3 it is the solution to Problem (4) between μ_s and $\mu_{s'}$, that the solution to Problem (6). ■

Understanding the strengths and limitations of Theorem 12 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 Euclidean 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 12. For instance, as the structural counterfactual operator from Example 4 is the gradient of a convex function, we obtain the following corollary.

Corollary 13. *Consider a linear additive SCM satisfying **(RE)** (see Example 4). If the factual distributions $\{\mu_s\}_{s\in\mathcal{S}}$ are absolutely continuous with respect to the 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 $\mu_{s'}$.*

Therefore, up to a linear approximation of the data-generation process, employing optimal transport maps for counterfactual reasoning in fairness contexts recovers causal changes, as in the example from Section 4.4.1. Besides, the scope of Theorem 12 goes beyond linear additive SCMs, as shown in the following non-linear non-additive example.

Example 5. *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 $\alpha > 0$, $U_1 > 0$, and $U_S \perp\!\!\!\perp (U_1, U_2)$. It satisfies **(A)**, **(I)** and **(RE)**, such that for any $s, s' \in \mathcal{S}$, the associated structural counterfactual operator is given by,*

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

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

Note that the converse of the implication in Theorem 12 does not hold. This comes from the fact that many functions (even continuous ones) cannot be written as gradients when $d > 1$, as illustrated in the following example.

Example 6. *Consider the following SCM,*

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

*where $U_S \perp\!\!\!\perp (U_1, U_2)$. It satisfies **(A)**, **(I)** and **(RE)**, such that for any $s, s' \in \mathcal{S}$, the associated structural counterfactual operator is given by,*

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

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

Through Section 5, we aimed notably at justifying the pertinence of optimal transport in counterfactual frameworks on top of the insights and illustrations given in Section 4. To sum-up, the main requisite for transport-based methods, typically optimal transport, to be used as substitutes for causal counterfactual reasoning is Assumption **(RE)**, ensuring that structural counterfactual models are transport-based counterfactual models. As previously explained, this condition is almost systematically verified in fairness problems, making the proposed surrogate approach relevant in various essential tasks. The more specific assumptions from Theorem 12, which

include **(I)**, describe an ideal setting meant to derive theoretical guarantees; optimal transport remains an arguably relevant alternative even outside this context. Altogether, Theorem 12 and Corollary 13 support the intuition that computing a Π from optimal transport provides a suitable approximation of the unknown structural Π^*. In the sequel, we apply this approach by extending causal counterfactual frameworks for fairness to transport-based models.

6 Transport-based counterfactual fairness

The strength of the unified mass-transportation viewpoint of counterfactual reasoning we proposed in Section 4 and further studied in Section 5 lies in the fact that all definitions and frameworks implicitly based on a structural counterfactual model have a transport-based analogue, and can therefore be made feasible. In this section, we apply this process 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 $h : \mathcal{X} \times \mathcal{S} \mapsto \mathbb{R}$ 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 4 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.

6.1 Causal counterfactual fairness from a mass-transportation viewpoint

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

Definition 14. *Let $\mathcal{M}$ satisfy **(A)**. A predictor $\hat{Y} = h(X, S)$ is* counterfactually fair *if for every $s, s' \in \mathcal{S}$ and μ_s-almost every x in $\mathcal{X}_s$,*

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

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

However, the above definition does not clearly emphasize the pairing between factual and counterfactual values. Interestingly, the mass-transportation viewpoint allows for pair-wise characterizations of counterfactual fairness.

Proposition 15. *Let $\mathcal{M}$ satisfy **(A)**.*

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* ***(RE)*** *holds, 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 $\pi^*_{\langle s'|s\rangle}$-almost every (x, x'),*

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

3. *If* ***(I)*** *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,*

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

4. *If* ***(I)*** *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,*

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

Items 2 to 4 in Proposition 15 are variations of the first item under the implications of **(RE)** and **(I)** through respectively Propositions 11 and 9. Note that they have practical interests. Assumption **(I)** highlights the deterministic relationship between factual and counterfactual quantities and makes unnecessary the knowledge of $\mathcal{L}(U)$ to test counterfactual fairness. Assumption **(RE)** entails by symmetry that only half of the couplings are necessary to check the condition. Additionally, if **(RE)** holds, then counterfactual fairness is a stronger criterion than the statistical parity across groups, as shown in the following proposition.

Proposition 16. *Let $\mathcal{M}$ satisfy* ***(A)*** *and 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.*

6.2 Extending counterfactual fairness

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

Definition 17. 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,*

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

Note that it follows from the symmetry of the transport-based counterfactual models (see items (ii) and (iii) in Definition 6) that only half of the couplings are truly required in the above conditions. Besides, because the proof of Proposition 16 only relies on the assumption that the couplings have factual distributions for marginals, the following proposition automatically holds.

Proposition 18. *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 result 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 SCM 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 14 and Proposition 15, Definition 17 relies on computationally feasible counterfactual models that obviate any assumptions about the data-generation process. In addition, as Definition 14 amounts to Π^*-counterfactual fairness (when **(RE)** holds), one can as well think of Definition 17 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. While Black et al. [2020] focused on explaining discriminatory biases in binary decision rules, we address the training of a Π-counterfactually fair predictor in Section 7.

6.3 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 19. *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.

7 Application to counterfactually fair learning

We now address an application of transport-based counterfactual models to fairness. More precisely, we introduce a supervised learning procedure trading-off between Π-counterfactual fairness and accuracy, and provide statistical guarantees.

7.1 Learning problem

In [Russell et al., 2017], the authors considered a learning problem involving a penalization controlling the pair-wise difference in decision between the training inputs and their structural counterfactual counterparts. While they gave empirical evidence of the efficiency of their training method, they had to assume a known causal model and did not provide consistency guarantees on the estimated predictor. In this sub-section, we illustrate that this counterfactual approach can naturally be made both feasible and statistically consistent by replacing the structural counterparts by transport-based counterparts. Note that in contrast to Russell et al. [2017], we do not optimize over several counterfactual models.

Let $Y : \Omega \to \mathcal{Y} \subseteq \mathbb{R}$ denote the so-called *ground-truth* variable to predict, and denote by $\mathcal{D}$ the law of the data (X, S, Y). We consider a parametric class of predictors $\{h_\theta\}_{\theta\in\Theta}$ from $\mathcal{X} \times \mathcal{S}$ to $\mathcal{Y}$, indexed by $\Theta \subseteq \mathbb{R}^p$ where $p > 1$. For a given counterfactual model $\Pi := \left\{\pi_{\langle s'|s\rangle}\right\}_{s,s'\in\mathcal{S}}$ and a given weight $\lambda > 0$, we define the following *expected* risk on the predictors,

$$\mathcal{R}_{\mathcal{D},\Pi,\lambda}(\theta) := \mathbb{E}[\ell(h_\theta(X, S), Y)] + \lambda \sum_{s\in\mathcal{S}} \mathbb{P}(S = s) \sum_{s'\neq s} \mathbb{E}[r_\theta(X_s, s, X_{s'}, s')], \tag{7}$$

where $\mathcal{L}((X_s, X_{s'})) = \pi_{\langle s'|s\rangle}$ for every $s, s' \in \mathcal{S}$. The application ℓ denotes a data-loss function, continuous with respect to each of its input variables, while $r_\theta(x, s, x', s')$ is a penalty promoting counterfactual fairness by enforcing the difference between the outputs of the algorithm for an individual and its counterfactual, namely $|h_\theta(x, s) - h_\theta(x', s')|$, to be small. For instance, in [Russell et al., 2017], the authors considered the tightest convex relaxation of ϵ-*approximate counterfactual fairness*, that is $r_\theta(x, s, x', s') := \max\left\{0, |h_\theta(x, s) - h_\theta(x', s')| - \epsilon\right\}$ for some $\epsilon > 0$. In this paper, we rather work with the penalty $r_\theta(x, s, x', s') := |h_\theta(x, s) - h_\theta(x', s')|^2$ which is smoother. Through λ, the risk $\mathcal{R}_{\mathcal{D},\Pi,\lambda}$ quantifies a trade-off between accuracy and counterfactual fairness. Importantly, when $\Pi = \Pi^*$, it corresponds precisely to the expected risk of the learning problem proposed by Russell et al. [2017] reframed using the mass-transportation viewpoint. In what follows, we will simply write $\mathcal{R}_{\mathcal{D},\Pi,\lambda}$ as $\mathcal{R}$.

In practice, we learn a predictor by minimizing an empirical version of $\mathcal{R}$. To this end, we need an empirical counterfactual model. Concretely, consider a training set $\{(x_i, s_i, y_i)\}_{i=1}^n$ composed of n i.i.d. observations drawn from $\mathcal{D}$. We divide this collection into $\mathcal{S}$ protected categories by defining for any $s \in \mathcal{S}$ the index $\mathcal{I}_s^n := \{1 \leq i \leq n \mid s_i = s\}$ of length $n_s := |\mathcal{I}_s^n|$. Then, the empirical versions of the factual distributions are for every $s \in \mathcal{S}$, $\mu_s^n := n_s^{-1} \sum_{i\in\mathcal{I}_s^n} \delta_{x_i}$. In our case, the counterfactual pairs between μ_s^n and $\mu_{s'}^n$ are estimated *within* the training dataset through an empirical transport plan $\{\pi^n_{\langle s|s'\rangle}(i, j)\}_{i\in\mathcal{I}_s, j\in\mathcal{I}_{s'}}$, typically by solving Problem (5) as explained in Section 3.2.3. Then, we define the following *empirical* risk,

$$\mathcal{R}_n(\theta) := \frac{1}{n}\sum_{i=1}^n \ell(h_\theta(x_i, s_i), y_i) + \lambda n_{s_i} \sum_{s'\neq s_i} \sum_{j\in\mathcal{I}_{s'}^n} \pi^n_{\langle s'|s_i\rangle}(i, j) r_\theta(x_i, s_i, x_j, s'). \tag{8}$$

The learning procedure amounts to carrying out a gradient-descent-based routine to minimize $\mathcal{R}_n$. We underline that this procedure, as the original one from [Russell et al., 2017], is tailored to both regression and multi-class classification. It also works for more than two protected groups, but requires the domain of the sensitive variable to be finite.

7.2 Consistency

In this part, we focus on the counterfactual model constructed with quadratic optimal transport, and prove the statistical consistency of the learning procedure. Set a sequence $\{\theta_n\}_{n\in\mathbb{N}^*}$ defined by $\theta_n \in \arg\min_{\theta\in\Theta} \mathcal{R}_n(\theta)$. The next theorem ensures the convergence to zero of the excess risk $\mathcal{R}(\theta_n) - \min_{\theta\in\Theta} \mathcal{R}(\theta)$ for ball-constrained linear predictions.

Theorem 20. *Suppose that for every pair of factual distributions, the Kantorovich problem* (3) *with cost* $c(x, x') := \|x - x'\|^2$ *admits a unique solution. Thus, we can define the counterfactual model* $\Pi = \{\pi_{\langle s'|s\rangle}\}_{s,s'\in\mathcal{S}}$ *and its empirical counterpart* $\Pi^n = \{\pi^n_{\langle s'|s\rangle}\}_{s,s'\in\mathcal{S}}$ *as, for every* $s, s' \in \mathcal{S}$,

$$\pi_{\langle s'|s\rangle} := \underset{\pi\in\Pi(\mu_s,\mu_{s'})}{\arg\min} \int_{\mathcal{X}_s\times\mathcal{X}_{s'}} \|x-x'\|^2 \mathrm{d}\pi, \tag{9}$$

$$\pi^n_{\langle s'|s\rangle} \in \underset{\pi\in\Sigma(n_s,n_{s'})}{\arg\min} \sum_{i\in\mathcal{I}_s}\sum_{j\in\mathcal{I}_{s'}} \|x_i-x_j\|^2 \pi(i,j). \tag{10}$$

Now, let $\Phi : \mathcal{X}\times\mathcal{S}\to\mathbb{R}^p$ be a feature map such that for every $s\in S$ and $x_1,x_2\in\mathcal{X}$, $\|\Phi(x_1,s)-\Phi(x_2,s)\|\leq L_s\|x_1-x_2\|$ where $L_s>0$. Consider for $\Theta\subseteq\mathbb{R}^p$ the class of linear predictors $\{h_\theta\}_{\theta\in\Theta}$ defined as $h_\theta(x,s) := \theta^T\Phi(x,s)$. If the following assumptions hold,

(i) there exists $D>0$ such that $\Theta=\{\theta\in\mathbb{R}^p \mid \|\theta\|\leq D\}$,

(ii) there exists $R>0$ such that $\mathcal{X}\subseteq\left\{x\in\mathbb{R}^d \mid \|x\|\leq R\right\}$,

(iii) there exists $b>0$ such that $\mathcal{Y}\subseteq\{y\in\mathbb{R}\mid |y|\leq b\}$,

(iv) there exists $L>0$ such that for any $(x,s,y)\in\mathcal{X}\times\mathcal{S}\times\mathcal{Y}$, the function $\theta\in\Theta\mapsto\ell(\theta^T\Phi(x,s),y)$ is L-Lipschitz,

then,

$$\mathcal{R}(\theta_n) - \min_{\theta\in\Theta}\mathcal{R}(\theta) \xrightarrow[n\to+\infty]{a.s.} 0.$$

The proof analyzes separately the accuracy term from the regularization term. The demonstration for the former follows classical results from the statistical-learning literature; the demonstration for the latter is original: we firstly show that each penalty contribution can be bounded by a distance between the empirical and the true coupling, and then invoke the convergence in law. We gather some additional remarks below.

Remark 21.

1. *Uniqueness of the solution* (9) *holds when the factual distributions are Lebesgue absolutely-continuous, or uniform over a same number of atoms.*
2. *Typically, Φ is defined as $(x,s)\mapsto(x,s,1)$ in order to add an intercept, or corresponds to the feature map of a kernel when aiming for non-linear decision boundaries.*
3. *Assumptions (i) to (iv) are common for supervised learning problems. The sets $\mathcal{X}$ and $\mathcal{Y}$ are usually bounded spaces, as well as Θ the set of parameters defining the algorithm. The Lipschitz conditions for the loss functions and the feature map can be directly assumed or are direct consequences of smoothness properties and compactness assumptions of the spaces on which they are defined.*
4. *The assumption on the second-order moments of the factual distributions is automatically satisfied under (ii).*
5. *If the risks $\mathcal{R}_n$ and $\mathcal{R}$ are strictly convex, then $\theta_n = \arg\min_{\theta\in\Theta}\mathcal{R}_n(\theta)$ and $\theta^* = \arg\min_{\theta\in\Theta}\mathcal{R}(\theta)$ are well-defined, and it follows that $\theta_n \xrightarrow[n\to+\infty]{a.s.} \theta^*$ (this additional step is detailed in the proof of Theorem 20).*

8 Numerical experiments

In this section, we present the implementation of our counterfactually fair learning procedure on real data, and show that it has the expected behaviour. The code is available at `https://github.com/lucasdelara/PI-Fair`.

8.1 Procedure

Whatever the dataset, the general procedure is the following: after dividing the studied dataset into a training set and a testing set, we learn one empirical counterfactual models for each set. The first one implements the penalty of the training loss function; the second enables to evaluate the counterfactual fairness of the trained predictors. We compute the corresponding optimal transport plans using the default (non-regularized) POT solver. Then, we train several predictors for various values of the weight λ to study the model's ability to trade-off between accuracy and fairness. Finally, we assess the performances of the learnt algorithms according to three criteria: accuracy, group fairness and counterfactual fairness, and we benchmark them against baselines.

8.1.1 Evaluation metrics

In what follows, $h : \mathcal{X} \times \mathcal{S} \to \mathcal{Y}$ denotes either a binary classifier ($\mathcal{Y} = \{0, 1\}$) or a regression function ($\mathcal{Y} = \mathbb{R}$), and $\mathcal{D}$ denotes a dataset (X, S, Y). Let us properly define the different metrics we employ:

- To evaluate the data fidelity of a classifier, we compute the *accuracy* (Acc), defined as

$$\mathrm{Acc}(h, \mathcal{D}) := \mathbb{P}(h(X, S) = Y).$$

  For a regression function, we compute the *mean square error* (MSE), defined as

$$\mathrm{MSE}(h, \mathcal{D}) := \mathbb{E}\left[\|h(X, S) - Y\|^2\right].$$

- To assess the statistical parity of a binary classifier when the protected attribute is binary, we compute the the *parity gap* (PG), defined as

$$\mathrm{PG}(h, \mathcal{D}) := |\mathbb{P}(h(X, S) = 1 \mid S = 0) - \mathbb{P}(h(X, S) = 1 \mid S = 1)|.$$

  It quantifies the violation to group fairness, and equals zero when statistical parity is achieved. For a regression function, we use the *Kolmogorov-Smirnov distance* (KS) between $\mathcal{L}(\hat{Y} \mid S = 0)$ and $\mathcal{L}(\hat{Y} \mid S = 1)$, defined as

$$\mathrm{KS}(h, \mathcal{D}) := \sup_{y \in \mathbb{R}} \left|\mathbb{E}[\mathbf{1}_{\{h(X,S)>y\}} \mid S = 0] - \mathbb{E}[\mathbf{1}_{\{h(X,S)>y\}} \mid S = 1]\right|.$$

  Note that this extends the parity gap to the continuous case. The purpose of these two group-fair indicators is to illustrate Proposition 18, stating that counterfactual fairness implies statistical parity.

- Finally, we need a metric to evaluate counterfactual fairness. We extend the notion of (ϵ, δ)-*approximate counterfactual fairness* introduced by Russell et al. [2017] to transport-based counterfactual models. For a counterfactual model Π and a tolerance $\epsilon > 0$, we define the probability for the disparate treatment by h between (x, s) and its s'-counterfactual counterpart to be lower than ϵ as

$$\mathrm{CFT}_\epsilon(h, x, s, s', \Pi) := \int_{x' \in \mathcal{X}_{s'}} \mathbf{1}_{\{|h(x,s)-h(x',s')| \leq \epsilon\}} \frac{\mathrm{d}\pi_{\langle s'|s \rangle}}{\mathrm{d}\mu_s}(x'|x).$$

  Then, for a probability threshold $0 \leq \delta \leq 1$, we say that a predictor h is (ϵ, δ)-*approximately counterfactually fair* if for every $s \in \mathcal{S}$, for μ_s-almost every $x \in \mathcal{X}_s$, and for every $s' \neq s$,

$$\mathrm{CFT}_\epsilon(h, x, s, s', \Pi) \geq 1 - \delta. \tag{11}$$

Dataset	Adult	COMPAS	Law	Crimes
Task	Classification	Classification	Regression	Regression
$S : 0/1$	Woman/Man	Black/White	Black/White	Black/Non-black
d	35	6	2	97
n_{train}	32,724	4,120	13,109	1,335
n_{test}	16,118	2,030	6,458	659

Table 1: Datasets

We make two remarks: firstly, if h is a classifier, then the only relevant value for ϵ is 0; secondly, if the counterfactual model is deterministic, then the only relevant value for δ is 0. As the empirical counterfactual models we use are non-deterministic—although their continuous counterparts may be deterministic—we set $\delta = 0.1$ whatever the prediction task. In practice, we quantify counterfactual fairness through the (ϵ, δ)-*counterfactual fairness rate* (CFR),

$$\mathrm{CFR}_{\epsilon,\delta}(h, \mathcal{D}, \Pi) := \sum_{s \in \mathcal{S}} \mathbb{P}(S = s) \int_{x \in \mathcal{X}_s} \left(\prod_{s' \neq s} \mathbf{1}_{\{\mathrm{CFT}_\epsilon(h,x,s,s',\Pi) \geq 1-\delta\}} \right) \mathrm{d}\mu_s(x).$$

This corresponds to the proportion of points satisfying Condition 11. In the classification setting we set $\epsilon = 0$ while in the regression setting we work with $\epsilon = \frac{1}{2}\mathbb{E}\left[|Y - Y'|\right]$ where Y' is an independent copy of Y.

8.1.2 Baselines

We aim at applying our regularized approach for several values of the weight λ to study the model's ability to trade-off between accuracy and fairness. For classification tasks, we consider logistic models; for regression tasks, we consider linear regression models. Theses choices will be useful in particular to benchmark our method against the one of Zafar et al. [2017], tailored to such models. For a given λ, we write Π-**Fair**(λ) for the corresponding regularized predictor. We compare the obtained results to three baseline algorithms: the best constant predictor **Const**, which achieves perfect fairness; the group-fair predictor **Z** developed by Zafar et al. [2017], which is meant to maximize accuracy under an exact-fairness constraint; the unaltered ($\lambda = 0$) predictor **U**, which is presumably the most accurate but also the most unfair predictor.

8.2 Datasets

We carry out the experiments on four datasets: the first two for classification and the last two for regression. Note that in all the considered settings, the sensitive variable S is binary and relatively exogenous to X. Table 1 summarizes information about each dataset after preprocessing.

8.2.1 Adult

The Adult Data Set from the UCI Machine Learning Repository [Dua and Graff, 2019] has become a gold reference dataset to evaluate and benchmark fairness frameworks. The *classification* task is to predict whether the income of an individual exceeds 50K USD per year based on census data. Concretely, the dataset contains $n = 48,842$ instances with 14 attributes (numerical and categorical). The ground-truth variable Y equals 1 whenever the incomes exceeds 50K, and 0 otherwise. In this work, we set the sensitive variable S to be the

gender: $S = 0$ stands for *female*, while $S = 1$ stands for *male*. The potential sources of algorithmic bias against women have been widely studied by Besse et al. [2021]. They mainly amount to an under representation of women in the dataset, as well as a high correlation between being a woman and having a lower income. Any standard algorithms, optimizing only for accuracy, are bound to be unfair towards women. Before training any models, we process the data using a one-hot-encoding of the categorical attributes. The processing is the exact same as in [Besse et al., 2021]. This leads to a dataset of dimension $d + 1 = 36$ (without the outcome). We divide it into a training set of size $n_{train} = 32,724$ and a testing set of size $n_{test} = 16,118$.

8.2.2 COMPAS

The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is an infamous score used by US court officers to assess the risk of criminal recidivism. ProPublica analyzed more than 10,000 of cases from Florida, and concluded that black defendants tended to be predicted riskier than they actually were whereas white defendants were often predicted at lower risk than they were.[3] In this part, we follow Kusner et al. [2017] and try to predict the risk of recidivism while avoiding discrimination against the race, using the same data. Keeping only black and white defendants, we get $n = 6,150$ instances with $d + 1 = 7$ attributes such as the number of prior offenses and the type of crime they committed. The ground-truth variable Y equals 1 if the individual recidivated and 0 otherwise. We set the sensitive variable S to be the *race*: $S = 0$ stands for *black*, while $S = 1$ stands for *white*. Finally, we divide the data into a training set of size $n_{train} = 4,120$ and a testing set of size $n_{test} = 2,030$.

8.2.3 Law School

This is the dataset used in Section 4.4.1, gathering statistics from 163 US law schools and more than 20,000 students. Here again we follow Kusner et al. [2017], and try to predict the first-year average grade of individuals Y on the basis of the race (black or white) S, the entrance-exam score X_1, and the grade-point average before law school X_2. All in all, we have $d = 2$ features excluding the outcome and the protected attributes, and work with $n_{train} = 13,109$ training entries and $n_{test} = 6,458$ testing entries.

8.2.4 Communities and crimes

The Communities and Crimes dataset can also be found in the UCI Machine Learning Repository [Dua and Graff, 2019]. It contains socio-economics, law enforcement and crime data from communities across the United States. Similarly to Chzhen et al. [2020], we consider the problem of predicting the rate of violent crime per 10^5 of population Y with $S = 0$ indicating that at least 50% of the population is black and $S = 1$ otherwise. After processing the 128 numerical and categorical attributes composing the dataset, we obtain $d + 1 = 98$ features over $n_{train} = 1,335$ training instances and $n_{testing} = 659$ testing instances.

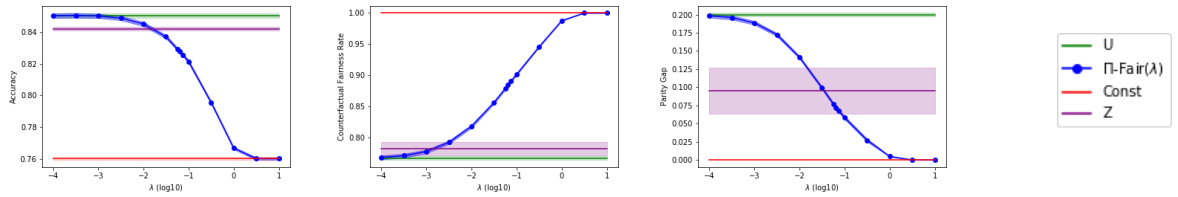
8.3 Results

The regularization weight λ takes successively all the values in a grid $\left\{10^{-4}, 10^{-3.5}, \dots, 10^{1}\right\}$. We repeat the training and evaluation processes of our models together with the baselines across 10 repeats for every datasets. As all learning techniques are deterministic, the randomness of the experiments comes uniquely from the division of the dataset into a training and testing sets. The results are reported in the figures below.

[3] `https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm`

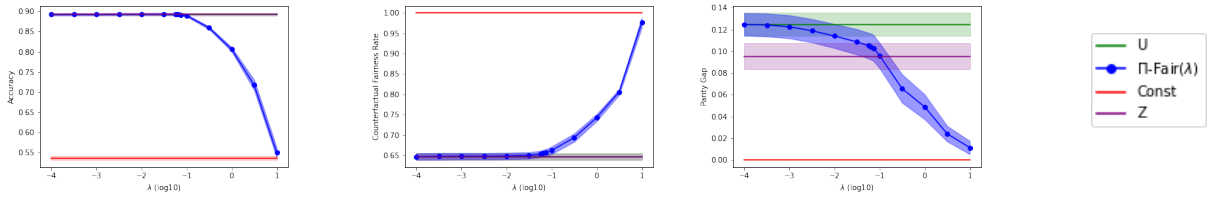
8.3.1 Trade-off between accuracy and fairness

Figures 6 to 9 show the evolution with respect to λ of the accuracy, the counterfactual fairness rate, and the statistical-parity metric. The solid line represents the mean value of the evaluation metric, while the vertical length of the shaded area corresponds to the standard deviation.



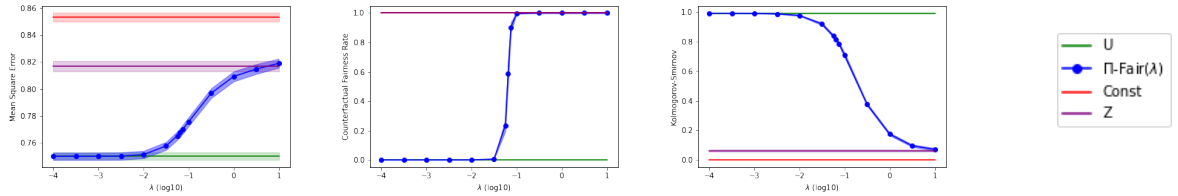
(a) Acc (b) CFR (c) PG (d) Classifiers

Figure 6: Evaluation metrics on the Adult dataset.



(a) Acc (b) CFR (c) PG (d) Classifiers

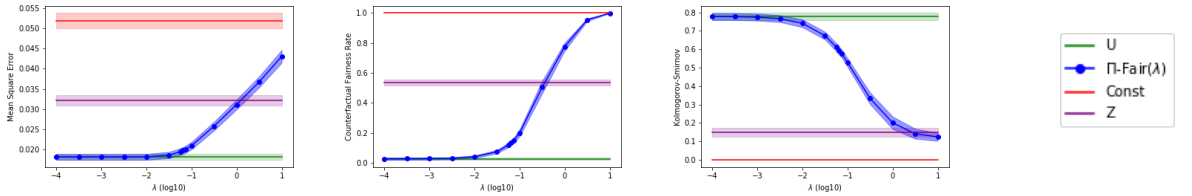
Figure 7: Evaluation metrics on the COMPAS dataset.



(a) MSE (b) CFR (c) KS (d) Predictors

Figure 8: Evaluation metrics on the Law dataset.

We observe that our learning algorithm is able to reliably trade-off accuracy for counterfactual fairness as λ increases, confirming the relevancy of the approach. Additionally, the evaluation metrics remain stable across the different repeats. As anticipated from Proposition 18, the regularization also tends to improve group fairness. Overall, the group-fair learning technique of Zafar et al. [2017] sacrifices less accuracy than our method to reach the same level of statistical parity, but our method performs better at encouraging counterfactual fairness. We conclude that the prevailing technique depends on the specific type of fairness one wants to achieve. Note that the group-fair predictor $\mathbf{Z}$ on the Law dataset (Figure 8) behaves similarly to the perfectly counterfactually-fair predictor. This is likely due to the use of simple linear models on such a low-dimensional dataset ($d + 1 = 3$) limiting the space of feasible algorithms. We leave the in-depth analysis of this phenomenon for further research.

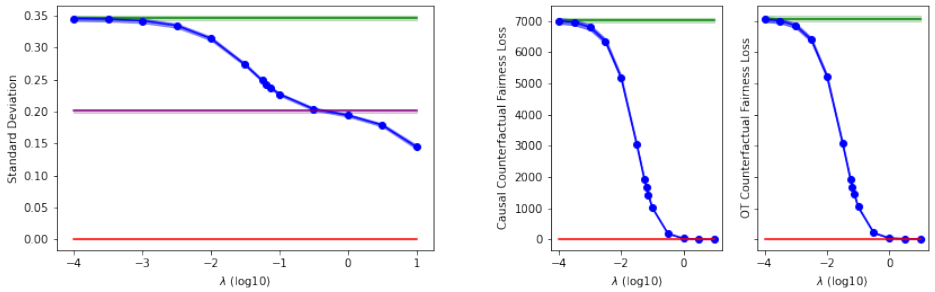


(a) MSE (b) CFR (c) KS (d) Predictors

Figure 9: Evaluation metrics on the Crimes dataset.

8.3.2 Recovering causal effects

To conclude these numerical experiments, let us verify that our optimal-transport counterfactual loss enforces causal counterfactual fairness in the adequate setting. We address the Law dataset for which a plausible causal model is known (see Section 4.4.1) and satisfies the assumptions of Corollary 13. Figure 10b displays the evolution of the two counterfactual losses, one based on the structural counterfactual model and the other on the optimal-transport counterfactual model, for predictors trained according to the optimal-transport counterfactual model. Figure 10a serves as a sanity check: it plots the normalized-variance indicator $\sqrt{\frac{\mathbb{E}\left[\left(\hat{Y}-\mathbb{E}[\hat{Y}]\right)^2\right]}{\mathbb{E}\left[(Y-\mathbb{E}[Y])^2\right]}}$ to control how close a predictor is to being constant.



(a) Standard deviation of $\hat{Y}$ (b) Counterfactual fairness losses

Figure 10: Promotion of causal counterfactual fairness on the Law dataset.

As anticipated by theory, the training process does promote causal counterfactual fairness: the two curves in Figure 10b are almost identical. Crucially, this is not a consequence of the predictors merely becoming constant, since the sequence of predictions in Figure 10a have variations that remain significantly higher than the best constant predictor.

8.4 Discussion

To sum-up, our learning procedure enables to increase counterfactual fairness while limiting the loss in accuracy, and is both theoretically sound and computationally efficient. This simple approach expands the fair learning arsenal to stronger fairness criteria than group fairness conditions, and so without requiring any additional knowledge on the data-generation process.

Regarding limitations, we note that the current procedure is not tailored to mini-batch learning. Using mini-batches would require to compute a new empirical counterfactual model for each one, which increases the computational complexity, especially since the batch-size should be chosen large enough for the empirical transport plans to make sense. This opens new lines of inquiry for leveraging recent advances on computational optimal transport in order to improve counterfactual learning problems. In particular, we could take advantage of entropic regularization schemes to speed-up the computation of optimal transport plans [Cuturi, 2013, Peyré and Cuturi, 2019]. This would make the produced counterfactual model blurry, but still close to the desired solution, allowing a trade-off between precision of the counterfactuals and numerical efficiency. Additionally, we could use the growing literature on plug-in estimations of optimal transport maps [Beirlant et al., 2020, Hallin et al., 2021, Manole et al., 2021, Pooladian and Niles-Weed, 2021] to construct empirical counterfactual models not as a matrices, but as a mappings able to generalize to out-of-sample observations, reusable on new datasets and batches. We leave these directions for future work.

9 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 couplings between factual and counterfactual instances. This novel perspective enabled us to generalize sharp but unfeasible causal criteria of fairness by actionable transport-based ones. We illustrated that the use optimal transport was a competitive approach to implement these criteria, as it can recover causal changes and can be computed efficiently. In particular, we proposed an new easy-to-implement method to train accurate classifiers with a counterfactual fairness regularization. We provided statistical guarantees, and showed empirically the relevancy of our method. In doing this article, we hope to shed a new light on counterfactual reasoning, and to open lines for strengthening the explainability and fair-learning arsenal in artificial intelligence.

Acknowledgements

The authors thank the AI interdisciplinary institute ANITI, grant agreement ANR-19-PI3A-0004 under the French investing for the future PIA3 program.

A Proofs of Section 5

Proof of Lemma 7 As a direct consequence of Assumption **(A)**, there exists a topological ordering on the nodes of the graph induced by $\mathcal{M}$. Therefore, starting with the components X_k for which $\mathrm{Endo}(k) = \emptyset$ or $\mathrm{Endo}(k) = \{S\}$, we can recursively replace the terms $X_{\mathrm{Endo}(k)}$ in the formulas $G_k(X_{\mathrm{Endo}(k)}, S_{\mathrm{Endo}(k)}, U_{\mathrm{Exo}(k)})$ by expressions depending only on U_X and S. This yields a measurable function F such that $X \overset{\mathbb{P}-a.s.}{=} F(S, U_X)$. The same computation but changing S to s for some $s \in \mathcal{S}$ leads to $X_{S=s} \overset{\mathbb{P}-a.s.}{=} F(s, U_X)$.

■

Proof of Proposition 8 Recall that $X \overset{\mathbb{P}-a.s.}{=} F(S, U_X)$. This implies that, $\mathbb{P}$-almost surely, $(X = x, S = s) \implies U_X \in f_s^{-1}(\{x\})$. Besides, $X_{S=s'} \overset{\mathbb{P}-a.s.}{=} f_{s'}(U_X)$ according to Lemma 7. Then, let $B \subseteq \mathcal{X}$ be an arbitrary measurable subset and compute:

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

Therefore, $\mathcal{L}(X_{S=s'} \mid X = x, S = s)$ does not put mass outside $f_{s'} \circ f_s^{-1}(\{x\})$. The definition of the support—the set of points $x \in \mathbb{R}^d$ such that every open neighborhood of x has a positive probability—thus implies that $\mathrm{supp}(\mu_{\langle s'|s\rangle}(\cdot|x)) \subseteq \overline{f_{s'} \circ f_s^{-1}(\{x\})}$. ■

Proof of Proposition 9 Set $s, s' \in \mathcal{S}$ and $x \in \mathcal{X}_s$. Note that, according to **(I)**, $U_X \overset{\mathbb{P}-a.s.}{=} f_S^{-1}(X)$. Let us address each item of the proposition separately.

• **Item 1.** Proposition 8 states that $\mathrm{supp}(\mu_{\langle s'|s\rangle}(\cdot|x)) \subseteq \overline{f_{s'} \circ f_s^{-1}(\{x\})}$ for μ_s-almost every $x \in \mathcal{X}_s$. This means according to **(I)** that $\mathrm{supp}(\mu_{\langle s'|s\rangle}(\cdot|x)) \subseteq \{f_{s'} \circ f_s^{-1}(x)\}$. Since the support of a probability distribution cannot be empty, we have equality. This proves the first item.

• **Item 2.** By definition of the counterfactual distribution, we find that

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

This proves the second item.

• **Item 3.** Similarly, by definition of the structural counterfactual coupling we obtain

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

where $\mathcal{L}(X_s) = \mu_s$. This completes the proof.

■

Proof of Proposition 10 Set $s \in \mathcal{S}$ and recall that $X \overset{\mathbb{P}-a.s.}{=} F(S, U_X)$ while $X_{S=s} \overset{\mathbb{P}-a.s.}{=} F(s, U_X)$. Thanks to Assumption **(RE)**, we have that $S \perp\!\!\!\perp U_X$. Therefore,

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

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

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

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

■

Proof of Proposition 11 We address each item separately.

• **Item** (i)**.** It is a direct consequence of $\pi^*_{\langle s'|s\rangle} \in \Pi(\mu_s, \mu_{\langle s'|s\rangle})$ by definition and $\mu_{\langle s'|s\rangle} = \mu_{s'}$ from Proposition 10.

• **Item** (ii)**.** Recall that **(RE)** implies that $S \perp\!\!\!\perp U_X$. Then, by definition we have

$$\begin{aligned}
\pi^*_{\langle s|s'\rangle} &= \mathcal{L}((X, X_{S=s}) \mid S = s') \\
&= \mathcal{L}((f_{s'}(U_X), f_s(U_X)) \mid S = s') \\
&= \mathcal{L}((f_{s'}(U_X), f_s(U_X)) \mid S = s) \\
&= \mathcal{L}((X_{S=s'}, X) \mid S = s) \\
&= t_\sharp \mathcal{L}((X, X_{S=s'}) \mid S = s) \\
&= t_\sharp \pi^*_{\langle s'|s\rangle}.
\end{aligned}$$

• **Item** (iii)**.** It is a direct consequence of $T^*_{\langle s'|s\rangle \sharp}\mu_s = \mu_{\langle s'|s\rangle}$ from Proposition 9 and $\mu_{\langle s'|s\rangle} = \mu_{s'}$ from Proposition 10.

• **Item** (iv)**.** We know according to Lemma 7 that $X_{S=s} \overset{\mathbb{P}-a.s.}{=} f_s(U_X)$ and $X_{S=s'} \overset{\mathbb{P}-a.s.}{=} f_{s'}(U_X)$. Furthermore, it follows from **(RE)** and Proposition 10 that $\mu_s = \mathcal{L}(X_{S=s})$ and $\mu_{s'} = \mathcal{L}(X_{S=s'})$. Wrapping this up, there exists a measurable set $\Omega_0 \subseteq \Omega$ with $\mathbb{P}(\Omega_0) = 1$ such that for every $\omega \in \Omega_0$,

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

In the rest of the proof we implicitly work on an $\omega \in \Omega_0$. Assumption **(I)** ensures that $U_X = f_s^{-1}(X_{S=s})$ so that $X_{S=s'} = \left(f_{s'} \circ f_s^{-1}\right)(X_{S=s})$. Noting that $X_{S=s} \in \mathcal{X}_s$, we obtain $X_{S=s'} = \left(f_{s'} \circ f_s^{-1}|_{\mathcal{X}_s}\right)(X_{S=s}) = T^*_{\langle s'|s\rangle}(X_{S=s})$. Following the same computation after switching s and s', we additionally get that $X_{S=s} = \left(f_s \circ f_{s'}^{-1}|_{\mathcal{X}_{s'}}\right)(X_{S=s'}) = T^*_{\langle s|s'\rangle}(X_{S=s'})$.

Therefore, $T^*{}_{\langle s'|s\rangle}$ is invertible on $X_{S=s'}(\Omega_0)$ such that $T^{*-1}_{\langle s'|s\rangle} = T^*{}_{\langle s|s'\rangle}$ on $X_{S=s}(\Omega_0)$. Since $\mu_s\left(X_{S=s}(\Omega_0)\right) = \mathbb{P}(\Omega_0) = 1$ and $\mu_{s'}\left(X_{S=s'}(\Omega_0)\right) = \mathbb{P}(\Omega_0) = 1$, this means that $T^*{}_{\langle s'|s\rangle}$ is invertible μ_s-almost everywhere such that $T^{*-1}_{\langle s'|s\rangle} = T^*{}_{\langle s|s'\rangle}$ $\mu_{s'}$-almost everywhere. This completes the proof.

■

Proof of Corollary 13 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 $x \mapsto \frac{1}{2}\|x\|^2 + \left[(I - M)^{-1}w(s' - s)\right]^T x$. As **(RE)** holds and μ_s is Lebesgue-absolutely continuous with finite second order moment, it follows from Theorem 12 that $T^*_{\langle s'|s\rangle}$ is the solution to (4) between μ_s and $\mu_{s'}$.

■

B Proofs of Section 6

Proof of Proposition 15 We address each item separately.

• **Item 1.** We claim that counterfactual fairness is equivalent to

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

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

Note that a direct reformulation of the original 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 $M \subseteq \mathbb{R}$*

$$\mathbb{P}\left(\hat{Y}_{S=s} \in M \mid X = x, S = s\right) = \mathbb{P}\left(\hat{Y}_{S=s'} \in M \mid X = x, S = s\right). \tag{12}$$

We aim at showing that **(CF)** is equivalent to **(Goal)**. To do so, we first prove that one can 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 $M \subseteq \mathbb{R}$ 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).** Set $s, s', x \in A$ and $M \subseteq \mathbb{R}$. According to the consistency rule, $\mathcal{L}(X \mid S = s) = \mathcal{L}(X_{S=s} \mid S = s)$, we can rewrite the left term of (12) as

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

Then, using Definition 4, we reframe the right term of (12) as

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

Remark now that because the indicator functions take either the value 0 or 1, the condition

$$\mathbf{1}_{\{h(x,s)\in M\}} = \int \mathbf{1}_{\{h(x',s')\in M\}} \mathrm{d}\mu_{\langle s'|s\rangle}(x'|x)$$

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 $M \subseteq \mathbb{R}$, 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 $\pi^*_{\langle s'|s\rangle}(C) = 1$.

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

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

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

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

• **Item 2.** Recall that $\pi^*_{\langle s|s\rangle} = (I \times I)_\sharp \mu_s$. Therefore, it follows from the previous item that counterfactual fairness can be written as: for every $s, s' \in \mathcal{S}$ such that $s' < s$, and $\pi^*_{\langle s'|s\rangle}$-almost every (x,x')

$$h(x,s) = h\left(x',s'\right),$$

and for $\pi^*_{\langle s|s'\rangle}$-almost every (x,x')

$$h(x,s') = h\left(x',s\right).$$

Moreover, **(RE)** implies through Proposition 11 that $\pi^*_{\langle s|s'\rangle} = t_\sharp \pi^*_{\langle s'|s\rangle}$. Therefore, the second condition above can be written as: for $\pi^*_{\langle s'|s\rangle}$-almost every (x,x')

$$h(x',s') = h\left(x,s\right),$$

which is exactly the first condition. This means that only the first condition is necessary, proving this item.

• **Item 3.** Consider **(CF)**, and recall that for every $s, s' \in S$, μ_s-almost every x and every measurable $M \subseteq \mathbb{R}$ the left term of (12) is $\mathbf{1}_{\{h(x,s)\in M\}}$. Let us now reframe the right-term of (12). If **(I)** holds, using that $U_X \overset{\mathbb{P}-a.s.}{=} f_S^{-1}(X)$ we obtain

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

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

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

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 third part of the proof.

• **Item 4.** From the previous item and Proposition 10, 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

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

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

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

Set $s, s' \in \mathcal{S}$ such that $s' < s$. To prove the fourth item, we show as for item 2 that the two above 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 9 and 10, $T^*_{\langle s'|s\rangle\sharp}\mu_s = \mu_{s'}$, 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 11, $T^{*-1}_{\langle s'|s\rangle} = T^*{}_{\langle s|s'\rangle}$ $\mu_{s'}$-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'.

■

Proof of Proposition 16 According to Proposition 15, 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 with respect to $\pi^*_{\langle s'|s\rangle}$ to obtain, for every measurable $M \subseteq \mathbb{R}$

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

This can be written as,

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

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,

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

■

Proof of Proposition 19 Suppose that the classifier $h(X,S)$ takes values in the finite set $\mathcal{Y} \subset \mathbb{R}$, and define 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 $y \in \mathcal{Y}$,

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

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,

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

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

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

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,

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

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

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

■

C Proofs of Section 7

Proof of Theorem 20 The outline of the proof is typical for such supervised learning problems, though some parts require basic knowledge on optimal transport. It mainly amounts to show the uniform convergence of $\{\mathcal{R}_n\}_{n\in\mathbb{N}^*}$ to $\mathcal{R}$, to then use the following classical deviation inequality,

$$\mathcal{R}(\theta_n)-\min_{\theta\in\Theta}\mathcal{R}(\theta)\leq 2\sup_{\theta\in\Theta}|\mathcal{R}_n(\theta)-\mathcal{R}(\theta)|. \tag{13}$$

For any measure P and any measurable function g, we will use the notation $P(g):=\int g\mathrm{d}P$ throughout the proof.

• **Step 1. Uniform convergence of the risk.** By the triangle inequality,

$$\begin{aligned}
\sup_{\theta\in\Theta}|\mathcal{R}_n(\theta)-\mathcal{R}(\theta)| \leq &\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n\ell(h_\theta(x_i,s_i),y_i)-\mathbb{E}\left[\ell(h_\theta(X,S),Y)\right]\right| \\
&+\lambda\sum_{s\in\mathcal{S}}\sum_{s'\neq s}\sup_{\theta\in\Theta}\left|\left(\frac{n_s}{n}\pi^n_{\langle s'|s\rangle}-\mathbb{P}(S=s)\pi_{\langle s'|s\rangle}\right)\left(r_\theta(\cdot,s,\cdot,s')\right)\right|.
\end{aligned}$$

The first term corresponds to the standard uniform risk deviation of supervised learning problems for Lipschitz losses and linear predictions. Under Assumptions (i) to (iv), for $0 < \delta < 1$ it follows from [Shalev-Shwartz and Ben-David, 2014, Theorem 26.5] that with probability greater than $1 - \delta$,

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \ell(h_\theta(x_i, s_i), y_i) - \mathbb{E}\left[\ell(h_\theta(X, S), Y)\right] \right| \leq \frac{\ell_0 + LD}{\sqrt{n}} \left(2 + \sqrt{2 \log \frac{1}{\delta}} \right),$$

where $\ell_0 = \sup_{|y| \leq b} |\ell(0, y)|$. Then, by taking $\delta_n := \frac{1}{n^2}$, we apply Borel-Cantelli lemma so that for every $\omega \in \Omega$, there exists a threshold $N(\omega)$ such that for any $n \geq N(\omega)$,

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \ell(h_\theta(x_i, s_i), y_i) - \mathbb{E}\left[\ell(h_\theta(X, S), Y)\right] \right| \leq \frac{\ell_0 + LD}{\sqrt{n}} \left(2 + \sqrt{4 \log n} \right).$$

The upper bound tends to zero as n tends to infinity, and consequently

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \ell(h_\theta(x_i, s_i), y_i) - \mathbb{E}\left[\ell(h_\theta(X, S), Y)\right] \right| \xrightarrow[n \to +\infty]{a.s.} 0.$$

The critical part is dealing with the counterfactual penalization. Let $s, s' \in \mathcal{S}$ such that $s' \neq s$. In the following of this step, we aim at showing that,

$$\sup_{\theta \in \Theta} \left| \left(\frac{n_s}{n} \pi^n_{\langle s'|s \rangle} - \mathbb{P}(S = s) \pi_{\langle s'|s \rangle} \right) (r_\theta(\cdot, s, \cdot, s')) \right| \xrightarrow[n \to +\infty]{a.s.} 0.$$

To do so, we use the triangle inequality again, leading to,

$$\begin{aligned}
&\sup_{\theta \in \Theta} \left| \left(\frac{n_s}{n} \pi^n_{\langle s'|s \rangle} - \mathbb{P}(S = s) \pi_{\langle s'|s \rangle} \right) \left(r_\theta(\cdot, s, \cdot, s') \right) \right| \\
&\leq \left| \frac{n_s}{n} - \mathbb{P}(S = s) \right| \sup_{\theta \in \Theta} \int r_\theta(x, s, x', s') \mathrm{d}\pi_{\langle s'|s \rangle}(x, x') \qquad (14) \\
&+ \mathbb{P}(S = s) \sup_{\theta \in \Theta} \left| \int r_\theta(x, s, x', s') \left(\mathrm{d}\pi^n_{\langle s'|s \rangle}(x, x') - \mathrm{d}\pi_{\langle s'|s \rangle}(x, x') \right) \right|. \qquad (15)
\end{aligned}$$

The terms (14) tends to zero almost surely as n increases to infinity. We now turn to the convergence of the term (15).

Firstly, let us show that the functions $\{r_\theta(\cdot, s, \cdot, s')\}_{\theta \in \Theta}$ are uniformly Lipschitz on $\mathcal{X} \times \mathcal{X}$. For any $(x_1, x_1'), (x_2, x_2') \in \mathcal{X} \times \mathcal{X}$, we have,

$$\begin{aligned}
|r_\theta(x_1, s, x_1', s') - r_\theta(x_2, s, x_2', s')| &\leq \left| \theta^T \left(\Phi(x_1, s) - \Phi(x_1', s') - \Phi(x_2, s) + \Phi(x_2', s') \right) \right|^2 \\
&\leq \left| \theta^T \left(\Phi(x_1, s) - \Phi(x_2, s) \right) \right|^2 \\
&+ \left| \theta^T \left(\Phi(x_1', s') - \Phi(x_2', s') \right) \right|^2, \\
&\leq \|\theta\|^2 \|\Phi(x_1, s) - \Phi(x_2, s)\|^2 \\
&+ \|\theta\|^2 \|\Phi(x_1', s') - \Phi(x_2', s')\|^2, \\
&\leq D^2 \left\{ L_s^2 \|x_1 - x_2\|^2 + L_{s'}^2 \|x_1' - x_2'\|^2 \right\}, \\
&\leq D^2 \max_{s \in \mathcal{S}} L_s^2 \|(x_1, x_1') - (x_2, x_2')\|^2, \\
&\leq 4D^2 \max_{s \in \mathcal{S}} L_s^2 R^2 \|(x_1, x_1') - (x_2, x_2')\|.
\end{aligned}$$

Let us set $\Lambda := 4D^2(\max_{s\in\mathcal{S}} L_s)^2R^2$, so that the functions $\{r_\theta(\cdot,s,\cdot,s')\}_{\theta\in\Theta}$ are Λ-Lipschitz.

Secondly, we know from [Villani, 2008, Theorem 5.19] that $\pi^n_{\langle s'|s\rangle}$ converges almost-surely weakly to $\pi_{\langle s'|s\rangle}$ as $n_s, n_{s'} \to +\infty$. Moreover, for any $s \in \mathcal{S}$, we have $\frac{n_s}{n} \xrightarrow[n\to+\infty]{} \mathbb{P}(S = s) > 0$, hence $n_s \xrightarrow[n\to+\infty]{} +\infty$. As a consequence, $\pi^n_{\langle s'|s\rangle}$ converges almost-surely weakly to $\pi_{\langle s'|s\rangle}$ as $n \to +\infty$. Additionally, since $\mathcal{X}_s \times \mathcal{X}_{s'} \subseteq \mathcal{X} \times \mathcal{X}$, it follows from Assumption (ii) that $\pi_{\langle s'|s\rangle}$ is compactly supported. According to Remark 7.13 in [Villani, 2003], this implies that $W_1(\pi^n_{\langle s'|s\rangle}, \pi_{\langle s'|s\rangle}) \xrightarrow[n\to+\infty]{a.s.} 0$, where W_1 denotes the Wasserstein-1 distance. Using the dual formulation of the Wasserstein distance, this convergence can be written as,

$$W_1(\pi^n_{\langle s'|s\rangle}, \pi_{\langle s'|s\rangle}) = \sup_{r\in \mathrm{Lip}_1(\mathcal{X}\times\mathcal{X},\mathbb{R})} \int r \left(\mathrm{d}\pi^n_{\langle s'|s\rangle} - \mathrm{d}\pi_{\langle s'|s\rangle}\right) \xrightarrow[n\to+\infty]{a.s.} 0.$$

Noting that for any $\theta \in \Theta$, $r_\theta(\cdot,s,\cdot,s')/\Lambda$ is 1-Lipschitz, we have

$$\frac{1}{\Lambda}\sup_{\theta\in\Theta}\left|\int r_\theta(x,s,x',s')\left(\mathrm{d}\pi^n_{\langle s'|s\rangle}(x,x') - \mathrm{d}\pi_{\langle s'|s\rangle}(x,x')\right)\right| \leq W_1(\pi^n_{\langle s'|s\rangle}, \pi_{\langle s'|s\rangle}) \xrightarrow[n\to+\infty]{a.s.} 0.$$

This entails that the term (15) converges almost surely to 0, and completes the proof.

• **Step 2. Consistency of the mimimum.** For this additional step, we assume that $\mathcal{R}_n$ and $\mathcal{R}$ have unique minimizers, and we denote $\theta^* := \arg\min_{\theta\in\Theta} \mathcal{R}(\theta)$. Note that the sequence $\{\theta_n\}_{n\in\mathbb{N}^*}$ is bounded by D, and as such we can extract a sub-sequence $\{\theta_{\sigma(n)}\}_{n\in\mathbb{N}^*}$ converging to some $\theta_\sigma \in \Theta$. Let us prove that $\theta_\sigma = \theta^*$ regardless of the choice of the subsequence $\{\sigma(n)\}_{n\in\mathbb{N}}$. According to the deviation inequality (13) and by continuity of $\mathcal{R}$, we have at the limit,

$$\mathcal{R}(\theta_\sigma) \leq \mathcal{R}(\theta^*).$$

This means that θ_σ is a minimizer of $\mathcal{R}$. Therefore, by uniqueness, $\theta_\sigma = \theta^*$. This completes the proof, as this implies that,

$$\theta_n \xrightarrow[n\to+\infty]{a.s.} \theta^*.$$

■

References

N. Asher, S. Paul, and C. Russell. Fair and adequate explanations. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 79–97. Springer, 2021.

N. Asher, L. De Lara, S. Paul, and C. Russell. Counterfactual models for fair and adequate explanations. *Machine Learning and Knowledge Extraction*, 4(2):316–349, 2022.

M. Bajaj, L. Chu, Z. Y. Xue, J. Pei, L. Wang, P. C.-H. Lam, and Y. Zhang. Robust counterfactual explanations on graph neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

Y. Balaji, R. Chellappa, and S. Feizi. Robust optimal transport with applications in generative modeling and domain adaptation. *Advances in Neural Information Processing Systems*, 33:12934–12944, 2020.

S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. `http://www.fairmlbook.org`.

M. F. Beg, M. I. Miller, A. Trouvé, and L. Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International journal of computer vision*, 61(2):139–157, 2005.

J. Beirlant, S. Buitendag, E. del Barrio, M. Hallin, and F. Kamper. Center-outward quantiles and the measurement of multivariate risk. *Insurance: Mathematics and Economics*, 95:79–100, 2020.

P. M. Bentler and D. G. Weeks. Linear structural equations with latent variables. *Psychometrika*, 45(3):289–308, 1980.

P. Besse, E. del Barrio, P. Gordaliza, J.-M. Loubes, and L. Risser. A survey of bias in machine learning through the prism of statistical parity. *The American Statistician*, pages 1–11, 2021.

E. Black, S. Yeom, and M. Fredrikson. Fliptest: Fairness testing via optimal transport. In *Conference on Fairness, Accountability, and Transparency*, page 111–121, New York, USA, 2020. Association for Computing Machinery.

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

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

V. Chernozhukov, A. Galichon, M. Hallin, and M. Henry. Monge–kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, 45(1):223–256, 2017.

S. Chiappa. Path-specific counterfactual fairness. In *AAAI Conference on Artificial Intelligence*, volume 33, pages 7801–7808, 2019.

S. Chiappa, R. Jiang, T. Stepleton, A. Pacchiano, H. Jiang, and J. Aslanides. A general approach to fairness with optimal transport. In *AAAI Conference on Artificial Intelligence*, pages 3633–3640, 2020.

D. M. Chickering, D. Heckerman, and C. Meek. Large-sample learning of bayesian networks is np-hard. *Journal of Machine Learning Research*, 5(Oct):1287–1330, 2004.

E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil. Fair regression with wasserstein barycenters. In *Advances in Neural Information Processing Systems*, volume 33, pages 7321–7331. Curran Associates, Inc., 2020.

G. F. Cooper. The computational complexity of probabilistic inference using bayesian belief networks (research note). *Artificial Intelligence*, 42(2–3):393–405, Mar. 1990.

N. Courty, R. Flamary, and D. Tuia. Domain adaptation with regularized optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 274–289. Springer, 2014.

N. Courty, R. Flamary, A. Habrard, and A. Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. *Advances in Neural Information Processing Systems*, 30, 2017.

J. A. Cuesta and C. Matrán. Notes on the wasserstein metric in hilbert spaces. *The Annals of Probability*, pages 1264–1276, 1989.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 26:2292–2300, 2013.

R. Dominguez-Olmedo, A. H. Karimi, and B. Schölkopf. On the adversarial robustness of causal algorithmic recourse. In *International Conference on Machine Learning*, pages 5324–5342. PMLR, 2022.

D. Dua and C. Graff. UCI machine learning repository, 2019. URL `http://archive.ics.uci.edu/ml`.

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science Conference*, pages 214–226, 2012.

J. Fawkes, R. Evans, and D. Sejdinovic. Selection, ignorability and challenges with causal fairness. In *Conference on Causal Learning and Reasoning*, 2021.

R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, L. Gautheron, N. T. Gayraud, H. Janati, A. Rakotomamonjy, I. Redko, A. Rolet, A. Schutz, V. Seguy, D. J. Sutherland, R. Tavenard, A. Tong, and T. Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

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

N. T. Gayraud, A. Rakotomamonjy, and M. Clerc. Optimal transport applied to transfer learning for p300 detection. In *BCI 2017-7th Graz Brain-Computer Interface Conference*, page 6, 2017.

P. Ghosal and B. Sen. Multivariate ranks and quantiles using optimal transport: Consistency, rates and nonparametric testing. *The Annals of Statistics*, 50(2):1012–1037, 2022.

P. Gordaliza, E. Del Barrio, F. Gamboa, and J.-M. Loubes. Obtaining fairness using optimal transport theory. In *International Conference on Machine Learning*, pages 2357–2365. PMLR, 2019.

M. Hallin, E. del Barrio, J. Cuesta-Albertos, and C. Matrán. Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach. *The Annals of Statistics*, 49(2):1139 – 1165, 2021.

P. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21, 2008.

A. Hyttinen, F. Eberhardt, and P. O. Hoyer. Learning linear cyclic causal models with latent variables. *The Journal of Machine Learning Research*, 13(1):3387–3439, 2012.

G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

S. Joshi, O. Koyejo, W. Vijitbenjaronk, B. Kim, and J. Ghosh. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *arXiv preprint arXiv:1907.09615*, 2019.

S. C. Joshi and M. I. Miller. Landmark matching via large deformation diffeomorphisms. *IEEE transactions on image processing*, 9(8):1357–1370, 2000.

A.-H. Karimi, B. Schölkopf, and I. Valera. Algorithmic recourse: from counterfactual explanations to interventions. In *ACM Conference on Fairness, Accountability, and Transparency*, pages 353–362, 2021.

N. Kilbertus, P. J. Ball, M. J. Kusner, A. Weller, and R. Silva. The sensitivity of counterfactual fairness to unmeasured confounding. In *Conference on Uncertainty in Artificial Intelligence*, pages 616–626. PMLR, 2020.

M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, volume 30, pages 4066–4076. Curran Associates, Inc., 2017.

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

Z. Lipton, J. McAuley, and A. Chouldechova. Does mitigating ml's impact disparity require treatment disparity? In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

D. Mahajan, C. Tan, and A. Sharma. Preserving causal constraints in counterfactual explanations for machine learning classifiers. *arXiv preprint arXiv:1912.03277*, 2020.

T. Manole, S. Balakrishnan, J. Niles-Weed, and L. Wasserman. Plugin estimation of smooth optimal transport maps. *arXiv preprint arXiv:2107.12364*, 2021.

R. 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.

S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2574–2582, 2016.

R. Nabi and I. Shpitser. Fair inference on outcomes. In *AAAI Conference on Artificial Intelligence*, volume 32, 2018.

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

J. Pearl. The mathematics of causal relations. *Causality and Psychopathology: Finding the Determinants of Disorders and their Cures (P. Shrout, K. Keyes and K. Ornstein, eds.). Oxford University Press, Corvallis, OR*, pages 47–65, 2010.

J. Pearl, M. Glymour, and N. P. Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

V. Peterson, N. Nieto, D. Wyser, O. Lambercy, R. Gassert, D. H. Milone, and R. D. Spies. Transfer learning based on optimal transport for motor imagery brain-computer interfaces. *IEEE Transactions on Biomedical Engineering*, 69(2):807–817, 2021.

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

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

A.-A. Pooladian and J. Niles-Weed. Entropic estimation of optimal transport maps. *arXiv preprint arXiv:2109.12004*, 2021.

R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie, and P. Flach. Face: feasible and actionable counterfactual explanations. In *AAAI/ACM Conference on AI, Ethics, and Society*, pages 344–350, 2020.

H. Rakotoarison, L. Milijaona, A. Rasoanaivo, M. Sebag, and M. Schoenauer. Learning meta-features for automl. In *International Conference on Learning Representations*, 2022.

P. Rasouli and I. C. Yu. Care: Coherent actionable recourse based on sound counterfactual explanations. *arXiv preprint arXiv:2108.08197*, 2021.

I. Redko, N. Courty, R. Flamary, and D. Tuia. Optimal transport for multi-source domain adaptation under target shift. In *International Conference on Artificial Intelligence and Statistics*, pages 849–858. PMLR, 2019.

M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

L. Risser, A. G. Sanz, Q. Vincenot, and J.-M. Loubes. Tackling algorithmic bias in neural-network classifiers using wasserstein-2 regularization. *Journal of Mathematical Imaging and Vision*, pages 1–18, 2022.

D. Rothenhäusler, N. Meinshausen, P. Bühlmann, J. Peters, et al. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society Series B*, 83(2):215–246, 2021.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

C. Russell, M. J. Kusner, J. Loftus, and R. Silva. When worlds collide: Integrating different counterfactual assumptions in fairness. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

M. Scutari, C. Vitolo, and A. Tucker. Learning bayesian networks from big data with greedy search: computational complexity and efficient implementation. *Statistics and Computing*, 29(5):1095–1108, 2019.

S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen, and M. Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

D. Slack, S. Hilgard, H. Lakkaraju, and S. Singh. Counterfactual explanations can be manipulated. *Advances in Neural Information Processing Systems*, 34, 2021.

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

Thibaut Le Gouic, Jean-Michel Loubes, and Philippe Rigollet. Projection to fairness in statistical learning. *arXiv preprint arXiv:2005.11720*, 2020.

W. Torous, F. Gunsilius, and P. Rigollet. An optimal transport approach to causal inference. *arXiv preprint arXiv:2108.05858*, 2021.

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

C. Villani. *Optimal Transport: Old and New*. Number 338 in Grundlehren der mathematischen Wissenschaften. Springer, Berlin, 2008.

J. von Kügelgen, A.-H. Karimi, U. Bhatt, I. Valera, A. Weller, and B. Schölkopf. On the fairness of causal algorithmic recourse. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9584–9594, 2022.

S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *International Conference on Artificial Intelligence and Statistics*, pages 962–970. PMLR, 2017.

K. Zhang and L.-W. Chan. Extensions of ica for causality discovery in the hong kong stock market. In *International Conference on Neural Information Processing*, pages 400–409. Springer, 2006.