```
\typeout{Counterfactual Models: The Mass Transportation Viewpoint}



\documentclass{article}
\pdfpagewidth=8.5in
\pdfpageheight=11in
% The file ijcai21.sty is NOT the same than previous years'
\usepackage{ijcai21}
\usepackage{oursettings}
% Use the postscript times font!
\usepackage{times}
\usepackage{soul}
\usepackage{url}
\usepackage[hidelinks]{hyperref}
\usepackage[utf8]{inputenc}
%\usepackage[small]{caption}
\usepackage{graphicx}
\usepackage{amsmath}
\usepackage{amsthm}
\usepackage{booktabs}
\usepackage{algorithm}
\usepackage{algorithmic}
\urlstyle{same}

% the following package is optional:
%\usepackage{latexsym}

% See https://www.overleaf.com/learn/latex/theorems_and_proofs
% for a nice explanation of how to define new theorems, but keep
% in mind that the amsthm package is already included in this
% template and that you must *not* alter the styling.
\newtheorem{example}{Example}
\newtheorem{theorem}{Theorem}
\newtheorem{proposition}{Proposition}
\newtheorem{definition}{Definition}
\newtheorem{lemma}{Lemma}

% Following comment is from ijcai97-submit.tex:
% The preparation of these files was supported by Schlumberger Palo Alto
% Research, AT\&T Bell Laboratories, and Morgan Kaufmann Publishers.
% Shirley Jowell, of Morgan Kaufmann Publishers, and Peter F.
% Patel-Schneider, of AT\&T Bell Laboratories collaborated on their
% preparation.

% These instructions can be modified and used in other conferences as
long
% as credit to the authors and supporting agencies is retained, this
notice
% is not changed, and further modification or reuse is not restricted.
% Neither Shirley Jowell nor Peter F. Patel-Schneider can be listed as
% contacts for providing assistance without their prior permission.

% To use for other conferences, change references to files and the
% conference appropriate and use other authors, contacts, publishers, and
% organizations.
% Also change the deadline and address for returning papers and the
length and
```

```
% page charge instructions.
% Put where the files are available in the appropriate places.

%PDF Info Is REQUIRED.
\pdfinfo{
/TemplateVersion (IJCAI.2021.0)
}

\title{Counterfactual Models: The Mass Transportation Viewpoint}

\author{
Lucas De Lara$^1$
\and
Alberto Gonz\'alez-Sanz$^1$\and
Nicholas Asher$^{2,3}$\And
Jean-Michel Loubes$^1$
\affiliations
$^1$IMT\and
$^2$CNRS\and
$^3$IRIT\\
\emails
\{lucas.de\_lara, alberto.gonzalez\_sanz,loubes\}@math.univ-toulouse.fr,
nicholas.asher@irit.fr,
}

\begin{document}

\maketitle

\begin{abstract}

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

\end{abstract}

\section{Introduction}
\label{introduction}

A \emph{counterfactual} states how the world should be modified so that a
given outcome occurs. For instance, the statement \emph{had you been a
woman, you would have gotten half your salary} is a counterfactual
relating the \emph{intervention} \say{had you been a woman} to the
\emph{outcome} \say{you would have gotten half your salary}.
Counterfactuals have been used to express causal laws \cite{lewis:1973}
```

and hence have attracted the attention in the fields of explainability and robustness in machine learning, as such statements can naturally represent the dependence of a prediction on a perturbation of input data without opening the black-box.

State-of-the-art models for computing true counterfactuals have mostly focused on the \emph{nearest counterfactual instances} principle \cite{wachter2017counterfactual}, according to which one finds minimal translations, minimal changes in the features of an instance that lead to a desired outcome. However, this simple distance-based technique often fails to describe faithful alternative worlds, due to the dependence between features. Changing just the sex of a person in such a translation might convert from a typical male into an untypical female rendering true counterfactuals like the following: {\em if I were a woman I would be 190cm tall and weigh 85 kg}. According to intuition, however, such counterfactuals are false and rightly so because they are oblivious of the latent statistical distribution. As a practical consequence, such counterfactuals typically hide biases in machine learning decision rules \cite{besse2020survey}.

The intuitive link between counterfactual modality and causality motivated the use of Pearl's causal graphs and structural equations \cite{pearl2009causality} to address the aforementioned shortcoming \cite{kusner2017counterfactual,joshi2019realistic,karimi2020algorithmic,mahajan2020preserving}. Causal models capture the structural relations between variables including their dependencies and as such provide the basis for generating true \emph{structural counterfactuals}. The cost of this approach is specifying the causal model. The reliance on such a strong prior makes the causal approach appealing in theory, but limited for systematic implementation. In addition, it's not how we humans evaluate counterfactuals. Typically, we don't know the causal graph for a given situation (and we're bad at constructing them); but we have strong intuitions on alternative states of things. Intuitively, the counterfactual female counterpart of a 190cm man would not be a 190cm woman, but more more likely a shorter woman, fairly tall compared to her gender-group. Our contribution offers a mathematical theory of this intuition based on \emph{optimal transport}.

\cite{black2020fliptest} first suggested substituting causal reasoning with optimal transport but didn't justify this theoretically. We do this here. Optimal transport answers the counterfactual question \emph{had the man been a woman, how tall would have she been?} by minimizing in average a cost between all the paired instances. Interestingly, optimal transport has been used to generalize the notion of distribution function to higher dimensions \cite{delbarrio2020centeroutward}, and thus provide a statistically-faithful notion of counterpart. In addition, it recovers the causal relations in many scenarios: as our principal theoretical result, we prove that the optimal transport map for the squared euclidean cost generates the same alternative states as a large class of linear causal models. %In summary, the counterfactual models we propose [WE WERE NOT THE FIRST ONE TO PROPOSE THEM, BUT WE GIVE A BETTER THEORETICAL UNDERSTANDING] satisfy the following criteria:
%\begin{enumerate}
    %\item The generated counterfactuals respect the statistical correlation between the variables that are intervened on and the others.
    %\item Their computation requires minimal assumptions on the data generative model.

```
    %\item The notion of alternative counterpart they characterize is
intuitively justifiable.
%\end{enumerate}

We will introduce the \emph{mass transportation} viewpoint of
counterfactual models, with which we will connect causal-based methods
with optimal-transport-based methods. First, we reformulate the
structural counterfactual approach as a problem of finding distributional
correspondences, and provide a closed-form for this operation under the
\emph{single-world} assumption. On the basis of this reformulation, we
introduce a general causality-free framework for the computation of
counterfactuals through mass transportation techniques---e.g., optimal
transport. This sheds new light on how to represent counterfactual
operations, offers new perspectives to explain black-box decision rules,
and recasts attractive causal-based specifications for counterfactuals
into more practicable criteria.

Related research falls into two categories: work that represents
counterfactual interventions as operators through causal modeling
\cite{plecko2020fair,karimi2020algorithmic}, and work that moves away
from causal-based models by proposing statistically-aware data-based
methods \cite{poyiadzi2020face,black2020fliptest}. This paper gives a new
justification to the latter, by underlining a common structure with the
former, and showing that the two may even coincide.

\section{Preliminaries}

The aim of this section is to detail the mathematical notation and
concepts used in the paper. As background for two main topics here,
optimal transport and causal reasoning,
\cite{villani2003topics,villani2008optimal} provide supplementary and
precise treatments of the first topic;
\cite{scholkopf2019causality,bongers2020foundations} do the same for the
second.

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

Formally, let $P, Q$ be both probabilities on $\R^d$, whose respective
supports are denoted by $\text{supp}(P)$ and $\text{supp}(Q)$, and set a
function $c:\R^d\times \R^d\rightarrow \R$. The \emph{Kantorovich
formulation} of the optimal transport problem with cost $c$ is the
optimization problem
\begin{equation}\label{kanto}
    \min_{\pi\in \Pi(P,Q)} \int_{\R^d\times \R^d} c(x,y) d\pi(x,y).
\end{equation}
$\Pi(P,Q)\subset \mathcal{P}(\R^d\times \R^d)$ denotes the set of joint
distributions $\pi$ whose marginals coincide with $P$ and $Q$
```

respectively, i.e. $\pi(A\times\R^d)=P(A)$ and $\pi(\R^d\times B)=Q(B)$, for all measurable sets $A,B\in \R^d$. Solutions to \eqref{kanto} are optimal transport plans between P and Q with respect to c. They exist under very mild assumptions, like the non-negativeness of the cost.

For $T : \R^d \to \R^d$ a measurable map, we say that T \emph{pushes forward} P to Q if $Q(B) := P (T^{-1}(B)),$ for any measurable set $B\subset \R^d$. This property, denoted by $T_\sharp P = Q$, means that if the law of a random variable Z is P, then the law of $T(Z)$ is Q. This push-forward operator T characterizes a \emph{deterministic} coupling between P and Q as every instance $x \in \text{supp}(P)$ is matched to $T(x) \in \text{supp}(Q)$ with probability 1. Suppose now that the cost c is the squared euclidean distance $\norm{\cdot}^2$ in $\R^d$, that P is absolutely continuous with respect to the Lebesgue measure %\textcolor{purple}{(i.e., admits a density)}
in $\R^d$, and that both P and Q have finite second order moments. Theorem~2.12 in \cite{villani2003topics} states that there exists an unique solution to \eqref{kanto}, whose form is $(I \times T)_\sharp P$\footnote{This denotes the law of $\big(Z,T(Z)\big)$ where $Z \sim P$.} where I is the identity function on $\R^d$ and $T : \R^d \to \R^d$ is a certain function called the \emph{Brenier map}. Besides, the Brenier map coincides P-almost surely with the gradient of a convex function. Recall that \emph{P-almost surely}, or equivalently \emph{P-almost everywhere}, means that it happens for all $x\in \R^d$ except maybe in a set N such that $P(N)=0$. Then, in this quadratic case, \eqref{kanto} is equivalent to the following \emph{Monge's formulation}
\begin{equation}\label{monge}
    \min_{T:\ T_\sharp P = Q} \int_{\R^d} \norm{x-T(x)}^2 dP(x).
\end{equation}
Thanks to a famous theorem from \cite{mccann1995}, under the sole assumption that P is absolutely continuous with respect to the Lebesgue measure, there exists only one gradient of a convex function $\nabla \psi$ satisfying the push-forward condition $\nabla \psi_\sharp P = Q$. This simplifies the search for the Brenier map solving \eqref{monge}, as it suffices to find a gradient of a convex function satisfying the push-forward condition.
%In this paper, we use the terms \emph{transport-based} and \emph{mass transportation} to refer to any method defining a correspondence between two distributions through a random coupling or a push-forward operator, which includes optimal transport.

\subsection{Causal reasoning}

Causal reasoning relies on a \emph{structural causal model} (SCM) \cite{pearl2009causality}, which represents the causal relationships between variables. More precisely, an \emph{acyclic} structural causal model $\mathcal{M}$ is a triple $\langle U, V, \G \rangle$ where:

\begin{enumerate}
    \item U and V are two indexed sets of random variables. Abusing notation, we interchangeably consider U and V as sets of random variables and as random vectors;
    \item $\G = \{G_{i}\}_{V_i \in V}$ is a collection of measurable $\R$-valued functions where for every $V_i \in V$, $V_i \stackrel{a.s.}{=} G_{i}\big({V_{\text{Endo}(i)}},U_{\text{Exo}(i)}\big)$. The subsets $V_{\text{Endo}(i)} \subset V \setminus \{V_i\}$ and $U_{\text{Exo}(i)}$

\subset U$ are respectively called the \emph{endogenous} and \emph{exogenous parents} of V_i, and denote the variables that directly determine V_i through G_i.
    \item The graph whose nodes are the variables in $U \cup V$, such that an arrow is drawn from some node Z to V_i if and only if $Z \in U_{\text{Exo}(i)} \cup V_{\text{Endo}(i)}$ is a \emph{directed acyclic graph} (DAG);
\end{enumerate}

The equations in 2., the \emph{structural equations}, specify the causal dependencies between the variables. By identifying $\G$ with a measurable vector function, we compactly write: $V \stackrel{a.s.}{=} \G(V,U)$. A structural causal model can be seen as a generative model. The variables in U are said to be \emph{exogenous}, as their values are imposed on the model by an input probability distribution $\mathcal{L}(U)$. In contrast, the variables in V are said to be \emph{endogenous}, as their values are outputs of the model determined through the structural equations and the values of U. In practice, the endogenous variables represent observed events, while the exogenous ones model latent background phenomena. Note that we don't assume that the endogenous variables are mutually independent.

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

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

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

\subsection{Counterfactual questions}

Let $(\Omega, \mathcal{A}, \P)$ be a probability space, and set $d \geq 1$. Define the random vector $V := (X,S) \in \R^{d+1}$, where the variables $X : \Omega \to \mathcal{X} \subset \R^d$ represent some observed features, while the variable $S: \Omega \to \mathcal{S} \subset \R$ can be subjected to interventions. For simplicity, we assume that $\S$ is finite such that for every $s \in \S$, $\P(S=s)>0$. For every $s \in \S$, set $\mu_s := \mathcal{L}(X|S=s)$ the \emph{factual} or \emph{observational} probability distribution of s-instances, and denote by $\X_s$ its support. We consider the problem of computing the potential outcomes of X when intervening on S. Suppose for instance that the event $\{X=x,S=s\}$ is observed, and set $s' \neq s$. We aim at answering the counterfactual question: \emph{had S been equal to s'

instead of s, what would have been the value of X?} Because of structural and statistical correlations between the variables, computing the alternative state does not amount to change the value of S while keeping the features X equal.

\section{Structural counterfactuals revisited}\label{structural}

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

\subsection{Definitions}

As introduced, a counterfactual statement is a \emph{cross-world} statement between a factual outcome and a counterfactual outcome. Let us formalize the contrast between interventional, counterfactual and factual outcomes in terms of probabilistic distributions. For any $s \in \S$ the distribution of the \emph{interventional} s-instances is defined as $\mu_{S=s} := \mathcal{L}(X_{S=s})$, and for any $s' \neq s$ the distribution of the \emph{counterfactual} s'-instances given s is defined as $\mu_{\langle s'|s \rangle} := \mathcal{L}(X_{S=s'}|S=s)$. According to the \emph{consistency rule} \cite{pearl2016causal}, for any $s \in \S$, the factual distribution can be written as $\mu_s = \mathcal{L}(X_{S=s}|S=s)$, which is sometimes denoted by $\mu_{\langle s|s \rangle}$ for the sake of coherence. The counterfactual distribution $\mu_{\langle s'|s \rangle}$ describes what would have been the observational instances of μ_s \emph{had S been equal to s' instead of s}; but it does not yield specific cross-world statements on its own, as it does not specify how instances from each distribution are related. The stronger notion of a counterfactual model characterizes all the counterfactual statements w.r.t. S.

The literature proposed various approaches to characterize causality-based counterfactual models. They all concur with the principle that the counterfactual model can be identified with the joint probability distributions between observable instances and intervened counterparts, as generated by the structural equations \cite{imbens2015causal,pearl2016causal,bongers2020foundations}. We follow \cite{pearl2016causal,kusner2017counterfactual} and propose a formalization of this definition that takes into account the observed value of S before intervening on it.

\begin{definition}\label{ctp} For every $s,s' \in \S$, the \emph{structural counterfactual coupling} between μ_s and $\mu_{\langle s'|s \rangle}$ is given by

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

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

\end{definition}

It is worth noting that, in general, the structural counterfactual couplings are \emph{random}, because X and $X_{S=s}$ are entangled through U following a certain probability distribution. This means that, according to Pearl's causal reasoning, there is not necessarily a one-to-one deterministic correspondence between factual instances and counterfactual counterparts, but a collection of weighted correspondences described by the structural couplings. To understand how the latent SCM generates such couplings, one must address the construction of the counterfactual distributions at the individual level. A \emph{counterfactual instance} represents a possible alternative state of X, with respect to an action on S and an observed evidence of (X,S). The following definition defines a counterfactual as a distribution rather than a random variable as in \cite{pearl2016causal}.

\begin{definition}\label{3steps}

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

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

\end{definition}

The possible outcomes $\mu_{\langle s's \rangle}(\cdot|x)$ are commonly generated with the so-called \textit{three-step} procedure \cite{pearl2016causal}, which amounts to: (1) setting a prior $\mathcal{L}(U)$ for the model $\mathcal{M}$, (2) computing the posterior distribution $\mathcal{L}(U|X=x,S=s)$, and (3) solving the structural equations of $\mathcal{M}_{S=s'}$ with $\mathcal{L}(U|X=x,S=s)$. As anticipated, the counterfactuals of an instance are not necessarily \emph{deterministic}, i.e. characterized by a degenerate distribution, but belong to a set of possible outcomes. This is due to the fact that, in general, there are several values of U consistent with an evidence $\{X=x,S=s\}$. Note that, equivalently to Definition \ref{ctp}, Definition \ref{3steps} characterizes the counterfactual semantics. In particular, the \emph{disintegrated} formulation $\mu_{\langle s's \rangle} = \int \mu_{\langle s's \rangle}(\cdot|x) d\mu_s(x)$ shows how μ_s relates to the counterfactual distribution through $\mu_{\langle s's \rangle}(\cdot|x)$.

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

\subsection{The deterministic case}\label{do}

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

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

\begin{proposition}\label{support} For any $s,s' \in \S$, $x \in \X_s$, $$\text{supp}\big(\mu_{\langle s'|s \rangle}(\cdot|x)\big) \subset f_{s'} \circ f_s^{-1}(\{x\}).$$
\end{proposition}

For any $x \in \R^d$, we denote by δ_x the distribution assigning a probability 1 to this single instance, which is called the Dirac at x. Proposition \ref{support} entails that the structural counterfactuals determine a unique counterpart, and thus the set of weighted counterfactual possibilities becomes a Dirac, if the following \emph{single-world} assumption holds:\footnote{This assumption corresponds to the logical constraint of conditional excluded middle \cite{stalnaker:1980}.} %a singleton with an instance of probability one

\begin{description}
    \item[Assumption (SW)\namedlabel{Invertibility}{\textbf{(SW)}}] \textit{The functions $\{f_s\}_{s \in \S}$ are injective.}
\end{description}

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

\begin{proposition}\label{oto} Let \ref{Invertibility} hold, and define for any $s,s' \in \S$, $T^*_{\mathsmaller{ \langle s'|s \rangle}} := f_{s'} \circ f_s^{-1} \restr{\X_s}$\footnote{$f_s^{-1} \restr{\X_s}$ denotes the restriction of f_s^{-1} to $\X_s$.}. The following properties hold:

\begin{enumerate}
    \item $\mu_{\langle s'|s \rangle}(\cdot|x) = \delta_{T^*_{\mathsmaller{ \langle s'|s \rangle}}(x)}$ for every $x \in \X_s$;
    \item $\mu_{\langle s'|s \rangle} = {T^*_{\mathsmaller{ \langle s'|s \rangle}}}_\sharp \mu_s$;
    \item $\pi^*_{\langle s'|s \rangle} = (I \times T^*_{\mathsmaller{ \langle s'|s \rangle}})_\sharp \mu_s$.
\end{enumerate}

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

\end{proposition}

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

\subsection{The exogenous case}

Let $\independent$ denote the independence between random variables. The variable S is said to be \emph{exogenous relative to} X \cite{galles1998axiomatic} if the following holds:

\begin{description}
    \item[Assumption (RE)\namedlabel{Exogeneity}{\textbf{(RE)}}]
\textit{$U_S \independent U_X$ and $X_{\text{Endo}(S)} = \emptyset$.}
\end{description}

\begin{figure}[H]
    \centering
    \begin{tikzpicture}[-latex ,auto ,node distance =2 cm ,on grid ,
    semithick ,
    state/.style ={circle ,top color =white,
    draw, minimum width =1 cm}]
    \node[state] (S) {S};
    \node[state] (X)[right= of S] {X};
    \node[state] (Ux)[above=of X]{U_X};
    \node[state] (Us)[above=of S]{U_S};
    \path (S) edge (X);
    \path (Ux) edge (X);
    \path (Us) edge (S);
    \end{tikzpicture}
    \caption{DAG satisfying \ref{Exogeneity}}
    \label{causalmodel}
\end{figure}

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

%, and enables to substitute U_S for S in the structural equations. For simplicity, we omit the exogenous variables U_S in the model, and reset $U := U_X$. Then, $S \independent U$, and we write $X = \F(S,U)$.

\begin{proposition}\label{conditioning}
If \ref{Exogeneity} holds, then for every $s,s' \in \S$ we have $\mu_{S=s'} = \mu_{s'} = \mu_{\langle s'|s \rangle}$.
\end{proposition}

Relative exogeneity is a critical assumption. Recall that the structural counterfactual coupling $\pi^*_{ \langle s'|s \rangle}$ represents an intervention transforming an observable distribution μ_s into an

\emph{a priori} non-observable counterfactual distribution $\mu_{\langle s'|s \rangle}$. According to Proposition \ref{conditioning}, \ref{Exogeneity} renders the causal model otiose for the purpose of generating the counterfactual distributions, as the latter coincides with the observable factual distribution $\mu_{s'}$. However, the coupling is {\em still required} to determine how each instance is matched at the individual level. Remarkably, \ref{Exogeneity} provides elegant transitivity properties to our counterfactual operators.
\begin{proposition}\label{cff} Suppose that \ref{Exogeneity} and \ref{Invertibility} hold. Then, for any $s,s',s'' \in \S$:

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

\end{proposition}

In terms of real-world modeling, \ref{Exogeneity} is intuitively satisfied in many scenarios. Let X represent the socio-economics features of individuals, and suppose for example that $\S = \{0,1\}$, where $S=0$ stands for \textit{female} while $S=1$ stands for \textit{male}. In this presumably exogenous model, any factual woman described by x is the counterfactual counterpart of her counterfactual male counterpart described by $T^*_{\mathsmaller{ \langle 1|0 \rangle}}(x)$, and changing all the factual women into their counterfactual male counterparts recovers the factual male population.

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

\begin{example}\label{ex} Under \ref{Exogeneity}, a linear additive SCM is characterized by the structural equations

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

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

\end{example}

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

\section{Transport-based counterfactuals}\label{surrogate}

\cite{black2020fliptest} mimicked the structural account of counterfactuals by computing alternative individuals using a deterministic optimal transport map, but they did not provide a mathematical or conceptual foundation for their idea. \ref{Invertibility} and \ref{Exogeneity} imply that approximating an unknown structural counterfactual model with deterministic couplings between observed data is a reasonable method. Generalizing their idea, we propose a general framework for transport-based counterfactual models that leads us to practicable SCM-free frameworks.%: a counterfactual model is an arbitrary notion of counterpart between \emph{observable} probability distributions..

\begin{definition}

\begin{enumerate}
    \item A \emph{counterfactual model} is a collection $\Pi := \{\pi_{ \langle s'|s \rangle}\}_{s,s' \in \S}$ of couplings on $\X \times \X$ such that for any $s,s' \in \S$, the first marginal of $\pi_{\mathsmaller{ \langle s'|s \rangle}}$ is μ_s, the second marginal is $\mu_{s'}$, and $\pi_{ \langle s|s \rangle} = (I \times I)_\sharp \mu_s$. An element of Π is called a \emph{counterfactual coupling}. We say that Π is a \emph{random counterfactual model} if at least one coupling for $s \neq s'$ is not deterministic.
    \item A \emph{deterministic counterfactual model} is a collection $\T := \{T_{\mathsmaller{ \langle s'|s \rangle}}\}_{s,s' \in \S}$ of mappings from $\X$ to $\X$ satisfying for any $s,s' \in \S$, ${T_{\mathsmaller{ \langle s'|s \rangle}}}_\sharp \mu_s = \mu_{s'}$ and $T_{\mathsmaller{ \langle s|s \rangle}} = \text{I}$. An element of $\T$ is called a \emph{counterfactual operator}.
\end{enumerate}

\end{definition}

%\textcolor{purple}{(I feel that this paragraph is not necessary)}\textcolor{yellow}{While most machine learning research has considered causal interventions as a mechanism to generate unobserved alternative distributions \cite{scholkopf2019causality, kusner2017counterfactual, barocas-hardt-narayanan}, this definition defines an approach motivated by designing practicable SCM-free frameworks: a counterfactual model is an arbitrary notion of counterpart between \emph{observable} probability distributions.}

%In a random model, the counterpart is not necessarily a single instance but a quantum state of weighted possible outcomes, which is consistent with the structural notion of counterfactual when \ref{Invertibility} does not hold. So \textcolor{purple}{(I do not understand why there is a \say{So} relating to the previous sentence, but we can maybe delete the previous sentence as I think we talked about the idea of multiple possible outcomes several times before)}
One challenge for this approach is to choose the model appropriately in order to define a relevant notion of counterpart. Even though the family of trivial couplings is a well-defined counterfactual model, it is not intuitively justifiable. Better suited counterfactual models can be constructed through optimal transport theory. Optimal transport with the squared euclidean cost is known to preserve quantiles in dimension one, and has been used to generalize the notion of distribution function to higher dimensions \cite{delbarrio2020centeroutward}. In this sense, it

satisfies our statistical intuitions on counterfactual reasoning. In addition, if the factual distributions are absolutely continuous w.r.t. the Lebesgue measure, then for any $s,s' \in \S$, the Brenier map between μ_s and $\mu_{s'}$ is the unique counterfactual operator that can be written as the gradient of a convex function. As the structural counterfactual operator from Example \ref{ex} is the gradient of a convex quadratic function, we obtain the following result. %Note that, while the existence of a deterministic structural counterfactual model $\T^*$ requires a strong hypothesis, the existence of a solution to \eqref{monge} only necessitates assumptions on the distributions to be mapped. In addition, for very common cases, a structural counterfactual operator is a Brenier map. %However, in some cases the method yields a unique model...

\begin{theorem}\label{linear}

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

\end{theorem}

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

The computational complexity of building an optimal transport plan between a n-sample to a m-sample is in $\mathcal{O}\big((n+m)nm \log(n+m)\big)$, but we can substantially improve on this to reach $\mathcal{O}\big(nm\big)$ with entropy-regularized versions \cite{cuturi2013sinkhorn}. As the computation is distribution-wise, not point-wise, it yields all the cross-world or counterfactual statements corresponding to a given change $s \to s'$ for the considered data-points. In contrast, computing a structural counterfactual coupling is less convenient and more challenging. First, inferring the causal graph from observational data is NP-hard, with an exponential worst-case complexity with respect to the number of nodes \cite{cooper1990computational,chickering2004large,scutari2019learning}. Second, this is not enough to compute counterfactuals, as we must still specify the structural equations. Third, even though the three-step procedure generates samples from the structural counterfactuals of a given instance through a specified SCM \cite{perov2020multiverse}, it needs to be applied at each point in order to infer the whole coupling.

%\textcolor{red}{Computing the Brenier map (necessarily n=m) as in Theorem 4.1 is in O((n3log(n)) (roughly cubic), which also provides a bound for computing a causal structure.}\textcolor{purple}{<--This I am not sure. Even though OT recovers the "true" intervention, it does not give the SCM.} In adddition, this straightforward surrogate technique dismisses the time-consuming design and inference of a causal model that would have been necessary for each new setting.

%{\color{purple}What I think we could say according to the references from Stefan Bauer. There exist several results regarding causal discovery, which is the problem of inferring the Bayesian network from data. This is in general a NP-hard problem with an exponential at-worst complexity, but a quadratic complexity can be attained should the graph be sparse \cite{claassen2013learning, Chickering1996,cooper1990computational,scutari2019learning}. However, this is not enough to compute counterfactuals, as the model won't be fully specified (i.e. the structural equations remain unknown). (Maybe one could fit a model once the graph is known to have an approximation of the SCM.) When the SCM if fully specified, we can compute counterfactuals through the 3steps procedure. For a given factual instance, the procedure gives a sample of the counterfactual counterparts, and can be achieved in a linear time with respect to the desired size of the counterfactual sample \cite{perov2020multiverse}. It is not directly comparable to optimal transport as the latter does not produce a sample, but matches instances. OT does not require to do the procedure for each factual instance, but instead directly align the two distribution samples, which is a much more convenient may of storing the counterfactual coupling than having samples related to every factual instance.}

%{\color {cyan} in the case of theorem 4.1, does the complexity of computing the unique optimal transport plan at O(nm)? What can we say about the complexity of computing the Brenier map?---c'etait Ã§a ma question. Bon peut etre une question bete}

\section{Applications}\label{applications}

In this section we look at two applications of transport based counterfactuals---explicability or interpretability and fairness of a \textit{black-box} algorithm. Counterfactuals have been used already in both areas \cite{wachter2017counterfactual,kusner2017counterfactual,karimi:etal:2020}. \cite{karimi:etal:2020} exploit automated reasoning based methods to find counterfactuals that can explain program behavior, and these methods have computational complexity problems given that they must test for satisfiability or unsatisfiability that is at least NP hard (depending on the logic fragment used). On the other hand a transport based method by aligning two entire probability distributions can provide a set of explanatory counterfactuals that mimic the causal approach in polynomial time, which means that the transport based approach can apply to the interpretability of programs for which a SAT based approach is not practically possible. In addition, the transport based method capturing as it does in some cases the causal structure of the phenomenon has a firmer conceptual basis as an explanatory tool than standard heuristically guided approaches relying either on local approximation by simpler linear models as LIME \cite{ribeiro2016should} or relying on the computation of indices measuring the contribution of each variable and its importance as in SHAP \cite{lundberg2017shap} or in \cite{me2020}.

We now turn to the fairness application. Suppose that the random variable S encodes the observed \emph{sensitive} or \emph{protected attribute} (e.g., race, gender) which divides the population into different classes in a machine learning prediction task. The counterfactual framework, by capturing the structural or statistical links between the features and the protected attribute, proposes sharper notions of fairness than \emph{statistical parity}, which only gives a notion of \emph{group

```
fairness}, and does not control discrimination at a subgroup or an
individual level: a conflict illustrated by \cite{dwork2012fairness}. We
first use the mass transportation formalism introduced in Section
\ref{structural} to reformulate the \textit{counterfactual fairness}
\cite{kusner2017counterfactual} condition, which is achieved when
individuals and their structural counterfactual counterparts are treated
equally. %The predictor is defined as $\hat{Y} := h(X,S)$, where $h :
\mathcal{X} \times \S \to \{0,1\}$ is deterministic. For every $s \in \S$
we introduce the intervened counterpart of the predictor as
$\hat{Y}_{S=s} := h(X_{S=s},s)$.

%The most common notion of fairness is the so-called \emph{statistical
parity}, which is satisfied when the rate of positive outcomes is equal
across protected classes: for any $s,s' \in \S$

%$$
%%    \P(h(X,S)=1|S=s) = \P(h(X,S)=1|S=s').
%$$

%Note that this corresponds to $h(X,S) \independent S$ in the case of
binary classification. However, this criterion provides limited
information for analyzing how unfair an algorithm is. In particular, it
only gives a notion of \emph{group fairness}, and does not control
discrimination at a subgroup or an individual level: a conflict
illustrated by \cite{dwork2012fairness}. **

%\subsection{Structural counterfactual fairness}

\begin{definition}\label{cf} A predictor $\hat{Y}=h(X,S)$ is
\emph{counterfactually fair} if for every $s,s' \in \S$ and $\mu_s$-
almost every $x$,

$$
   \mathcal{L}(\hat{Y}_{S=s}|X=x,S=s) =
\mathcal{L}(\hat{Y}_{S=s'}|X=x,S=s),
$$
where $\hat{Y}_{S=s} := h(X_{S=s},s)$.

\end{definition}

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

\begin{proposition}\label{rcf}

\begin{enumerate}
    \item A predictor $h(X,S)$ is counterfactually fair if and only if
for every $s,s' \in \S$ and $\pi^*_{ \langle s'|s \rangle}$-almost every
$(x,x')$,
    $$
    h(x,s) = h(x',s').
    $$
    \item If \ref{Invertibility} holds, then a predictor $h(X,S)$ is
counterfactually fair if and only if for every $s,s' \in \S$ and $\mu_s$-
almost every $x$,
```

```
    $$
    h(x,s) = h(T^*_{\mathsmaller{ \langle s'|s \rangle}}(x),s').
    $$
    \item If \ref{Invertibility} and \ref{Exogeneity} hold, then a
predictor $h(X,S)$ is counterfactually fair if and only if for every
$s,s' \in \S$ such that $s<s'$ and $\mu_s$-almost every $x$,
    $$
    h(x,s) = h(T^*_{\mathsmaller{ \langle s'|s \rangle}}(x),s').
    $$
\end{enumerate}

\end{proposition}

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

\begin{proposition}\label{stronger} Suppose that \ref{Exogeneity} holds.
If the predictor $h(X,S)$ satisfies \textit{counterfactual fairness},
then it satisfies \textit{statistical parity}, namely $h(X,S)
\independent S$. The converse does not hold in general.
\end{proposition}

%\subsection{Generalized counterfactual fairness}

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

\begin{definition}\label{tcounter}

\begin{enumerate}
    \item Let $\Pi = \{\pi_{\langle s'|s \rangle}\}_{s,s' \in \S}$ be a
random counterfactual model. A predictor $h(X,S)$ is \emph{$\Pi$-
counterfactually fair} if for every $s,s' \in \S$ and $\pi_{\langle s'|s
\rangle}$-almost every $(x,x')$,
    $$
    h(x,s) = h(x',s').
    $$
    \item Let $\T = \{T_{\langle s'|s \rangle}\}_{s,s' \in \S}$ be a
deterministic counterfactual model. A predictor $h(X,S)$ is \emph{$\T$-
counterfactually fair} if for every $s,s' \in \S$ and $\mu_s$-almost
every $x$,
    $$
    h(x,s) = h(T_{\mathsmaller{ \langle s'|s \rangle}}(x),s').
    $$
\end{enumerate}

\end{definition}

Because the proof of Proposition \ref{stronger} only relies on the
assumption that the couplings are transport plans between the factual
distributions, the following proposition holds.
```

\begin{proposition}\label{Tcf} Let Π be a counterfactual model (deterministic or not). If a predictor $h(X,S)$ satisfies Π-counterfactual fairness, then it satisfies statistical parity, namely $h(X,S) \independent S$. The converse does not hold in general.
\end{proposition}

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

\section{Conclusion}

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

%% The file named.bst is a bibliography style file for BibTeX 0.99c
\bibliographystyle{named}
{\small \bibliography{references,more-refs}}

\newpage

\appendix

This supplementary material addresses the mathematical proofs of the paper.

\section{Lemmas}

We start by proving two key results we mentioned in Section \ref{structural}. The first one specifies formulas for X and its interventional variants.

\begin{lemma}\label{F}
There exists a measurable function $\F$ such that $\P$-almost surely $X = \F(S,U_X)$ and $X_{S=s} = \F(s,U_X)$ for any $s \in \S$.
\end{lemma}

\begin{proof}

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

\begin{align*}
    X_i &=
G_{X_i}\big(X_{\text{Endo}(X_i)},S_{\text{Endo}(X_i)},U_{X_i}\big),\\
    S &= G_{S}\big(X_{\text{Endo}(S)},U_S\big).
\end{align*}

Because the graph of $\mathcal{M}$ is a DAG, it has a topological
ordering on the variables in $X$. Then, we can recursively substitute the
$X_i$ according to this ordering to obtain
$$
    X = \tilde{\F}\big(S_{\text{Endo}(X)},U_X\big),
$$
where $\tilde{\F}$ is a measurable function. Remark that either
$S_{\text{Endo}(X)}=\{S\}$ or $S_{\text{Endo}(X)} = \emptyset$, depending
on whether $S$ is a parent of $X$ in the graph. Then, without loss of
generality, we can define $\F$ such that $\F(S,U_X) :=
\tilde{\F}\big(S_{\text{Endo}(X)},U_X\big)$. Consequently, $X =
\F(S,U_X)$. Now, recall that $do(S=s)$ preserves the structural equations
of $X$, and does not impact $U$. Then, using the exact same procedure for
$(X_{S=s},S_{S=s})$ instead of $(X,S)$ we get $X_{S=s} = \F(S_{S=s},U_X)
= \F(s,U_X)$.

\end{proof}

The second result is the consistency rule.

\begin{lemma}\label{consistency}
For any $s \in \S$, $\mu_{\langle s|s\rangle} = \mu_s$
\end{lemma}

\begin{proof}

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

\begin{align*}
    \mu_s &= \mathcal{L}(X|S=s)\\
    &= \mathcal{L}(\F(S,U_X)|S=s)\\
    &= \mathcal{L}(\F(s,U_X)|S=s),
\end{align*}

and

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

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

\end{proof}
```

```
\section{Proofs of Section \ref{structural}}

\noindent Proof of Proposition \ref{support}.

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

\begin{align*}
    &\P(X_{S=s'} \in B|X=x,S=s)\\
    &= \P(f_{s'}(U_X) \in B|X=x,S=s)\\
    &= \P(f_{s'}(U_X) \in B, U_X \in f_s^{-1}(\{x\})|X=x,S=s)\\
    &= \P(f_{s'}(U_X) \in B, f_{s'}(U_X) \in f_{s'} \circ f_s^{-
1}(\{x\})|X=x,S=s) \\
    &= \P(X_{S=s'} \in \big[ B \cap f_{s'} \circ f_s^{-1}(\{x\})\big]
|X=x,S=s).
\end{align*}
Consequently, $\mathcal{L}(X_{S=s'}|X=x,S=s)$ does not put mass outside
$f_{s'} \circ f_s^{-1}(\{x\})$.
\end{proof}

\noindent Proof of Proposition \ref{oto}.

\begin{proof}

Let $s,s' \in \S$ and $x \in \X_s$. From Lemma \ref{F} we know that $X =
f_S(U_X)$, and according to \ref{Invertibility} we additionally have $U_X
= f^{-1}_S(X)$. We address each point separately.

\paragraph{Proof of 1.} By definition of the structural counterfactuals,

\begin{align*}
    \mathcal{L}(X_{S=s'}|X=x,S=s) &= \mathcal{L}(f_{s'}(U_X)|X=x,S=s)\\
    &= \mathcal{L}(f_{s'}(f^{-1}_S(X))|X=x,S=s)\\
    &= \mathcal{L}(f_{s'}\circ f^{-1}_s(x)|X=x,S=s)\\
    &= \mathcal{L}(f_{s'} \circ f^{-1}_s(x))\\
    &= \delta_{f_{s'} \circ f^{-1}_s(x)}.
\end{align*}
This proves the first point of the proof.

\paragraph{Proof of 2.} By definition of the counterfactual distribution,

\begin{align*}
    \mu_{\langle s'|s \rangle} &= \mathcal{L}(X_{S=s'}|S=s)\\
    &= \mathcal{L}(f_{s'}(U_X)|S=s)\\
    &= \mathcal{L}(f_{s'}\circ f^{-1}_S(X)|S=s)\\
    &= \mathcal{L}(f_{s'}\circ f^{-1}_s(X)|S=s)\\
    &= \big(f_{s'} \circ f^{-1}_s\big)_\sharp \mu_s.\\
\end{align*}
This proves the second point of the proposition.

\paragraph{Proof of 3.} By definition of the structural counterfactual
coupling,

\begin{align*}
    \pi_{\langle s'|s \rangle} &= \mathcal{L}\big((X,X_{S=s'})|S=s\big)\\
```

```
    &= \mathcal{L}\big((X,f_{s'}(U_X))|S=s\big)\\
    &= \mathcal{L}\big((X,f_{s'}(f_s^{-1}(X)))|S=s\big)\\
    &= \mathcal{L}\big((X_s,f_{s'} \circ f_s^{-1}(X_s))\big),
\end{align*}
where $X_s \sim \mu_s$. This concludes the proof.

\end{proof}

\noindent Proof of Proposition \ref{conditioning}.

\begin{proof}
To show this, set $s \in \S$ and invoke Lemma \ref{F} once again to write
$X = \F(S,U_X)$ and $X_{S=s} = \F(s,U_X)$. Recall that \ref{Exogeneity}
implies that $S \independent U_X$. Then,

\begin{align*}
    \mathcal{L}(X|S=s) &= \mathcal{L}\big(\F(S,U_X)|S=s\big),\\
&=\mathcal{L}\big(\F(s,U_X)|S=s\big),\\
    &=\mathcal{L}\big(\F(s,U_X)\big),\\
    &=\mathcal{L}(X_{S=s}).
\end{align*}
This means that $\mu_s = \mu_{S=s}$. Similarly, for $s,s' \in \S$ we have

\begin{align*}
    \mathcal{L}(X_{S=s'}|S=s) &= \mathcal{L}\big(\F(s',U_X)|S=s\big),\\
&=\mathcal{L}\big(\F(s',U_X)\big),\\
    &=\mathcal{L}\big(\F(s',U_X)|S=s'\big),\\
    &=\mathcal{L}\big(\F(S,U_X)|S=s'\big),\\
    &=\mathcal{L}(X|S=s').
\end{align*}
This means that $\mu_{\mathsmaller{ \langle s'|s \rangle}} = \mu_{s'}$,
which concludes the proof.

\end{proof}

\noindent Proof of Proposition \ref{cff}.

\begin{proof}

We address each point separately.

\paragraph{Proof of 1.} Set $s,s' \in \S$. By definition
$T^*_{\mathsmaller{ \langle s'|s \rangle}} = f_{s'} \circ f^{-
1}_s\restr{\X_s}$, which induces a bijection from $\X_s$ to
$\text{Im}(T^*_{\mathsmaller{ \langle s'|s \rangle}})$. Let us denote
$\text{Im}(T^*_{\mathsmaller{ \langle s'|s \rangle}})$ by
$\X_{\mathsmaller{ \langle s'|s \rangle}}$, so that ${T^*}^{-
1}_{\mathsmaller{ \langle s'|s \rangle}} = f_{s} \circ f^{-
1}_{s'}\restr{\X_{\mathsmaller{ \langle s'|s \rangle}}}$.

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

\begin{align*}
```

```
    X_{S=s}(\omega) &= f_s(U_X(\omega)) \subset \X_s,\\
    X_{S=s'}(\omega) &= f_{s'}(U_X(\omega)) \subset \X_{s'}.
\end{align*}

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

\begin{align*}
    {T^*}^{-1}_{\mathsmaller{ \langle s'|s \rangle}}(X_{S=s'}) &= f_{s}
\circ f^{-1}_{s'}\restr{\X_{\mathsmaller{ \langle s'|s
\rangle}}}(X_{S=s'})\\
    &= f_{s} \circ f^{-1}_{s'}\restr{\X_{s'}}(X_{S=s'})\\
    &= {T^*}_{\mathsmaller{ \langle s|s' \rangle}}(X_{S=s'}).
\end{align*}

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

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

\begin{align*}
    X_{S=s}(\omega) &= f_s(U_X(\omega)) \subset \X_s,\\
    X_{S=s'}(\omega) &= f_{s'}(U_X(\omega)) \subset \X_{s'}.
\end{align*}

Then, we write
\begin{align*}
    {T^*}_{\mathsmaller{ \langle s''|s \rangle}}(X_{S=s}) &= f_{s''}
\circ f^{-1}_{s}\restr{\X_{s}}(X_{S=s})\\
    &= (f_{s''} \circ f^{-1}_{s'}) \circ (f_{s'} \circ f^{-
1}_{s}\restr{\X_{s}}) (X_{S=s}).
\end{align*}

Note that $(f_{s'} \circ f^{-1}_{s}\restr{\X_{s}})(X_{S=s}) = X_{S=s'}
\in \X_{s'}$. Hence,

\begin{align*}
    {T^*}_{\mathsmaller{ \langle s''|s \rangle}}(X_{S=s}) &= (f_{s''}
\circ f^{-1}_{s'}\restr{\X_{s'}}) \circ (f_{s'} \circ f^{-
1}_{s}\restr{\X_{s}}) (X_{S=s})\\
    &= {T^*}_{\mathsmaller{ \langle s''|s' \rangle}} \circ
{T^*}_{\mathsmaller{ \langle s'|s \rangle}} (X_{S=s}).
\end{align*}

Similarly to the previous point, this means that the equality
${T^*}_{\mathsmaller{ \langle s''|s \rangle}} = {T^*}_{\mathsmaller{
```

\langle s''|s' \rangle}} \circ {T^*}_{\mathsmaller{ \langle s'|s \rangle}}$ holds on $X_{S=s}(\Omega_0)$ where $\P(\Omega_0)=1$. Thus, it holds μ_{s}-almost everywhere as $\mu_{s}\big(X_{S=s}(\Omega_0)\big)=\P(\Omega_0)=1$. This concludes the proof.

\end{proof}

\section{Proofs of Section \ref{surrogate}}

Proof of Theorem \ref{linear}.

\begin{proof}

We address the structural equations

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

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

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

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

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

\end{proof}

\section{Proofs of Section \ref{applications}}

\noindent Proof of Proposition \ref{rcf}.

\begin{proof}

We address each point separately.

\paragraph{Proof of 1.} We aim at showing that counterfactual fairness is equivalent to:

```
\begin{description}
    \item[(Goal)\namedlabel{Goal}{\textbf{(Goal)}}] {\it For every $s,s'
\in \S$, there exists a measurable set $C := C(s,s') \subset \X \times
\X$ satisfying $\pi^*_{\langle s'|s \rangle}(C) = 1$ such that for every
$(x,x') \in C$
$$
    h(x,s) = h(x',s').
$$}
\end{description}

A direct reformulation of the counterfactual fairness condition is:

\begin{description}
    \item[(CF)\namedlabel{CF}{\textbf{(CF)}}] {\it For every $s,s' \in
\S$, there exists a measurable set $A := A(s)$ satisfying $\mu_s(A)=1$,
such that for every $x \in A$ and every measurable set $M \subset \R$
\begin{equation}\label{cf_eq}
\begin{split}
    \P(\hat{Y}_{S=s} \in M|&X=x,S=s)\\ &=\P(\hat{Y}_{S=s'} \in
M|X=x,S=s).
\end{split}
\end{equation}
}
\end{description}

To show that \ref{CF} is equivalent to \ref{Goal}, we first rewrite
\ref{CF} into the following intermediary formulation:

\begin{description}
    \item[(IF)\namedlabel{IF}{\textbf{(IF)}}] {\it For every $s,s' \in
\S$, there exists a measurable set $A := A(s)$ satisfying $\mu_s(A)=1$,
such that for every $x \in A$ and every measurable $M \subset \R$ 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 $x' \in B$,
$$
    \mathbf{1}_{\{h(x,s)\in M\}} = \mathbf{1}_{\{h(x',s')\in M\}}.
$$}
\end{description}

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

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

Then, using the distributions of the structural counterfactuals, express
the right-term of \eqref{cf_eq} as

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

\end{align*}

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

$$
    \mathbf{1}_{\{h(x,s)\in M\}} = \mathbf{1}_{\{h(x',s')\in M\}}
$$
This proves that \ref{CF} is equivalent to \ref{IF}.

\paragraph{Proof that \ref{IF} $\implies$ \ref{Goal}.} As \ref{IF} is true for any arbitrary measurable set $M \subset \R$, 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 $s,s' \in \S$, 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{align*}
    \pi^*_{\langle s'|s \rangle}(C) &= \int_A \P(X_{S=s'} \in B | X=x,S=s) d\mu_s(x)\\
    &= \int_A \mu_{\langle s'|s \rangle}(B|x) d\mu_s(x)\\
    &= \int_A 1 d\mu_s(x)\\
    &= \mu_s(A)\\
    &= 1.
\end{align*}

This proves that \ref{IF} implies \ref{Goal}.

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

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

Since $0 \leq \mu_{\langle s'|s \rangle}(B|x) \leq 1$, this implies that for μ_s-almost every x, $\mu_{\langle s'|s \rangle}(B|x)=1$. Said differently, there exists a measurable set $A := A(s)$ satisfying $\mu_s(A) = 1$ such that for every $x \in A$, the measurable set $B(s,s',x)$ satisfies $\mu_{\langle s'|s \rangle}(B|x)=1$. By

construction of B and by definition of C, for every $x \in A$ and every $x' \in B$, $h(x,s) = h(x',s')$. To obtain \ref{IF}, it suffices to take any measurable $M \in \R$ and to note that the latter equality implies that $\mathbf{1}_{\{h(x,s)\in M\}} = \mathbf{1}_{\{h(x',s')\in M\}}$.

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

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

Consequently, \ref{CF} holds if and only if, for every measurable $M \in \R$

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

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

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

$$
h(x,s) = h(T^*_{\mathsmaller{ \langle s'|s \rangle}}(x),s'),
$$
and for $\mu_{s'}$-almost every x
$$
h(x,s') = h({T^*}_{\mathsmaller{ \langle s|s' \rangle}}(x),s').
$$

Set $s,s' \in \S$ such that $s' < s$. To prove 3. we must show that these two conditions are equivalent. Set A a measurable subset of $\X_s$ such that $\mu_s(A) = 1$, and $h(x,s) = h(T^*_{\mathsmaller{ \langle s'|s \rangle}}(x),s')$ for any $x \in A$. Then, make the change of variable $x' = T^*_{\mathsmaller{ \langle s'|s \rangle}}(x)$ so that $h({T^*}^{-1}_{\mathsmaller{ \langle s'|s \rangle}}(x'),s') = h(x',s')$ for every $x' \in T^*_{\mathsmaller{ \langle s'|s \rangle}}(A)$. By Propositions \ref{oto} and \ref{conditioning}, ${T^*_{\mathsmaller{ \langle s'|s \rangle}}}_\sharp \mu_s = \mu_{s'}$, which implies that $\mu_{s'}(T^*_{\mathsmaller{ \langle s'|s \rangle}}(A))=1$. Therefore,

the equality $h({T^*}^{-1}_{\mathsmaller{ \langle s'|s \rangle}}(x'),s) = h(x',s')$ holds for $\mu_{s'}$-almost every x'. Finally, recall that according to Proposition \ref{cff}, ${T^*}^{-1}_{\mathsmaller{ \langle s'|s \rangle}} = {T^*}_{\mathsmaller{ \langle s|s' \rangle}}$ $\mu_{s'}$-almost everywhere. As the intersection of two sets of probability one is a set of probability one, $h({T^*}_{\mathsmaller{ \langle s|s' \rangle}}(x'),s) = h(x',s')$ holds for $\mu_{s'}$-almost every x'.

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

\end{proof}

\noindent Proof of Proposition \ref{stronger}.

\begin{proof}

According to Proposition \ref{rcf}, h is counterfactually fair if and only if for any $s,s' \in \S$ 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 $M \in \R$. Set $s,s' \in \S$. Recall that from \ref{Exogeneity}, $\pi^*_{\langle s'|s \rangle}$ admits μ_s for first marginal, and $\mu_{s'}$ for second marginal. Let us integrate this equality w.r.t. $\pi^*_{\langle s'|s \rangle}$ to obtain, for every measurable $M \subset \R$

$$
    \int \mathbf{1}_{\{h(x,s)\in M\}} d\mu_s(x) = \int \mathbf{1}_{\{h(x',s')\in M\}} d\mu_{s'}(x).
$$
This can be written as, for every measurable $M \in \R$

$$
    \P(h(X,s) \in M|S=s) = \P(h(X,s') \in M|S=s'),
$$
which means that
$$
    \mathcal{L}(h(X,S)|S=s) = \mathcal{L}(h(X,S)|S=s').
$$
As this holds for any $s,s' \in \S$, we have that $h(X,S) \independent S$.

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 \ref{Exogeneity} is satisfied so that

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

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 $T^*_{\mathsmaller{
\langle 1|0 \rangle}}(x) = 1-x$.
Now, set the \textit{unaware} predictor (i.e., which does not take the
protected attribute as an input), $h(X) := \text{sign}(X-1/2)$. Clearly,

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

\end{proof}

\end{document}


\subsection{Individual fairness with counterfactuals}

\emph{Individual fairness} states that feature-wise similar individuals
should receive similar outcomes. However, counterfactual reasoning states
that distance and group-wise similarities are not necessarily related.
Consider for instance two persons applying for the same job, with similar
features, but belonging to different protected groups. A \emph{fair}
comparison of the two individuals should take into account the struggle
the disadvantaged one had to face in order to reach the same
achievements. Using counterfactual modeling, we propose a discrepancy
satisfying this principle. For this, consider a deterministic
counterfactual model $\T = \{T_{\langle s'|s' \rangle}\}_{s,s' \in \S}$,
and set the following mappings

$$
    T_{s'}(x,s) := T_{\langle s'|s \rangle}(x) \mathbf{1}_{s \neq s'} + x
\mathbf{1}_{s=s'},
$$
which give a deterministic representation of any individual in any group.
Then, define

$$
    d_{\T}\big((x,s),(x',s')\big) := \frac{1}{|\S|} \sum_{s'' \in \S}
\norm{T_{s''}(x,s)-T_{s''}(x',s')}.
$$
This notion of similarity $d_{\T}$ never compares directly individuals
from different classes: if necessary, it transports beforehand a point to
its counterpart in another group. This enables to combine individual
fairness and statistical fairness in a way consistent with counterfactual
fairness. Set $\gamma$ the joint distribution of $(X,S)$.

\begin{defn}\label{alberto} Let $\T$ be a deterministic counterfactual
model, $f : \X \times \S \to \R$ be a regression function, and $L>0$ be a
constant. The predictor $f$ is \emph{$\T$-individually fair across
groups} if there exists $L>0$ such that for $\gamma$-almost every $(x,s)$
and $(x',s')$,

$$
    |f(x,s)-f(x,s')| \leq L \times d_{\T}((x,s),(x',s')).
$$
\end{defn}

%%JM
```

When dealing with the Adult Dataset, the machine learning model trained on the whole dataset is denoted by $f(X,S)$. As pointed out in \cite{besse2020survey}, the correlations between the sensitive variable {\it sex} S, and the decision on the acceptation of the loan by the bank is learnt by the model and transformed into a causal decision. The bias of the algorithm is the same whether we use the model $f(X,S)$ or a model who would choose the best decision for all possible sex of the individual ${\rm max}(f(X,1),f(X,0))$. This highlights that changing only the value of the sensitive variable $S=0$ changed into $S=1$ is not enough to properly define a counterfactual. Rather the whole distribution must be transformed ...

```
@article{besse2020survey,
  title={A survey of bias in Machine Learning through the prism of
Statistical Parity for the Adult Data Set},
  author={Besse, Philippe and del Barrio, Eustasio and Gordaliza, Paula
and Loubes, Jean-Michel and Risser, Laurent},
  journal={arXiv preprint arXiv:2003.14263},
  year={2020}
}
```