



HAL
open science

Modèles contrefactuels pour un apprentissage machine explicable et juste : une approche par transport de masse

Lucas de Lara

► To cite this version:

Lucas de Lara. Modèles contrefactuels pour un apprentissage machine explicable et juste : une approche par transport de masse. Apprentissage [cs.LG]. Université Paul Sabatier - Toulouse III, 2023. Français. NNT : 2023TOU30138 . tel-04356522

HAL Id: tel-04356522

<https://theses.hal.science/tel-04356522v1>

Submitted on 20 Dec 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



THÈSE

**En vue de l'obtention du
DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE
Délivré par l'Université Toulouse 3 - Paul Sabatier**

**Présentée et soutenue par
Lucas DE LARA**

Le 19 juin 2023

**Modèles contrefactuels pour un apprentissage machine
explicable et juste: une approche par transport de masse**

Ecole doctorale : **EDMITT - Ecole Doctorale Mathématiques, Informatique et
Télécommunications de Toulouse**

Spécialité : **Mathématiques et Applications**

Unité de recherche :

IMT : Institut de Mathématiques de Toulouse

Thèse dirigée par

Jean-Michel LOUBES et Laurent RISSER

Jury

Mme Marianne CLAUSEL, Rapporteur

M. Arthur CHARPENTIER, Rapporteur

M. Emiliano LORINI, Examineur

M. Jean-Michel LOUBES, Directeur de thèse

M. Laurent RISSER, Co-directeur de thèse

M. Christophe GIRAUD, Président

Counterfactual Models for Fair and Explainable Machine Learning: A Mass Transportation Approach

Lucas De Lara

19th June 2023

Abstract

The spread of automatic decision rules based on machine learning has raised grave ethical concerns due to their lack of interpretability and their automatization of human discriminatory biases. This issue sparked the research field of trustworthy artificial intelligence which focuses on the development of fair and explainable machine-learning algorithms. This thesis contributes to this initiative by studying fairness and explainability through the prism of counterfactual reasoning: a modality devoted to queries such as “Had she been a man, would have he been granted the loan?”. The first chapter serves as an introduction to counterfactual frameworks in machine learning and clarifies common misconceptions. The second chapter develops the theoretical foundations for implementing this reasoning using mass-transportation methods such as optimal transport and diffeomorphic registration. In contrast to state-of-the-art methods, this approach returns counterfactual statements that are simultaneously computationally feasible and semantically realistic, crucially allowing a wider deployment of counterfactual methodologies for fairness and explainability. The third and fourth chapter are self-sufficient but motivated by the practical aspects of this mass-transportation viewpoint to counterfactual reasoning: they address the statistical estimation of transport models. The third chapter introduces a GAN estimator of Lipschitz optimal transport maps along with unique statistical guarantees for such a neural-network-based approximation. The fourth chapter deals with diffeomorphic registration, grounding theoretically and computationally diffeomorphic mass transportation driven by Sinkhorn divergences (entropy-regularized optimal-transport discrepancies).

Abstract (en français)

La propagation de règles de décision automatiques basées sur l'apprentissage machine a soulevé de graves préoccupations éthiques en raison de leur manque d'interprétabilité et de leur automatisation des biais discriminatoires humains. Ce problème a donné naissance à la recherche en l'intelligence artificielle digne de confiance, qui traite du développement d'algorithmes d'apprentissage automatique équitables et explicables. Cette thèse contribue à cette initiative en étudiant l'équité et l'explicabilité à travers le prisme du raisonnement contrefactuel : une modalité consacrée à des requêtes telles que "Si elle avait été un homme, lui aurait-on accordé le prêt ?". Le premier chapitre sert d'introduction aux méthodologies contrefactuelles dans l'apprentissage automatique et clarifie des idées reçues courantes. Le deuxième chapitre développe les fondements théoriques de la mise en œuvre de ce raisonnement à l'aide de méthodes de transport de masse telles que le transport optimal et l'appariement difféomorphique. Contrairement aux méthodes standards, cette approche renvoie des énoncés contrefactuels qui sont à la fois implémentable d'un point de vue informatique et réalistes d'un point de vue sémantique, ce qui permet un déploiement plus large des méthodologies contrefactuelles pour l'équité et l'explicabilité. Les troisième et quatrième chapitres sont autonomes mais motivés par les aspects pratiques de ce point de vue par transport de masse du raisonnement contrefactuel : ils traitent de l'estimation statistique des modèles de transport. Le troisième chapitre présente un estimateur GAN d'applications de transport optimales Lipschitz accompagné de garanties statistiques uniques pour une telle approximation basée sur des réseaux de neurones. Le quatrième chapitre traite de l'appariement difféomorphique, en fondant théoriquement et numériquement le transport de masse difféomorphique piloté par les divergences de Sinkhorn (des divergences de transport optimal entropique).

Remerciements

Je voudrais d'abord remercier mes directeurs de thèse Jean-Michel Loubes, Laurent Risser et Nicholas Asher pour l'opportunité qu'ils m'ont offerte, ainsi que Louis Béthune pour sa collaboration. Je tiens aussi à remercier les membres de mon jury, Marianne Clausel, Arthur Charpentier, Emiliano Lorini et Christophe Giraud, pour avoir accepté d'évaluer ma thèse. Je suis par ailleurs reconnaissant envers le Ministère de l'Enseignement Supérieur et de la Recherche ainsi que l'Artificial and Natural Intelligence Toulouse Institute qui m'ont financé durant ces trois années.

Par dessus tout, je remercie mon colocataire/collègue/ami Alberto González Sanz: c'est grâce à toi. *Gracias lechero.*

Contents

Abstract	i
Abstract (en français)	iii
Remerciements	v
Notations	xi
Introduction	1
Summary of the thesis	2
Introduction (en français)	13
Résumé de la thèse	14
I Counterfactual reasoning: from causality to mass transportation	25
1 Counterfactuals, explainability, fairness	27
1.1 Introduction	27
1.2 The possible-world account	28
1.2.1 Worlds and counterfactuals	29
1.2.2 Antecedents and interventions	29
1.2.3 Consequents and evaluations	31
1.2.4 Examples of similarity metrics	31
1.3 Structural causal modeling	32
1.3.1 Causal model	32
1.3.2 Do-intervention	34
1.3.3 Counterfactual inference	36
1.3.4 Discussion	37
1.4 The potential-outcome framework	39
1.4.1 Model and motivation	39
1.4.2 Estimation of causal effects	39
1.4.3 Identification of the potential outcomes	43
1.4.4 Discussion and comparison of causal frameworks	49
1.5 Counterfactual explanations in artificial intelligence	50
1.5.1 Original problem formulation	51

1.5.2	Intervention-based counterfactual explanations	53
1.6	Counterfactuals in fairness	54
1.6.1	Standard definitions of fairness	54
1.6.2	Counterfactual-based fairness	56
Appendix 1.A	Proofs of Section 1.3	59
Appendix 1.B	Proofs of Section 1.4	59
2	Transport-based counterfactual models	61
2.1	Introduction	61
2.1.1	Outline of contributions	62
2.1.2	Related work	63
2.2	Mass transportation	64
2.2.1	Definition	64
2.2.2	Optimal transport	65
2.3	Counterfactual models	67
2.3.1	Problem setup	67
2.3.2	Structural counterfactuals	67
2.3.3	Transport-based counterfactuals	69
2.3.4	Examples	72
2.3.5	Discussion	75
2.4	Theoretical results	77
2.4.1	Causal assumptions and their consequences	77
2.4.2	When optimal transport meets causality	82
2.5	Transport-based counterfactual fairness	85
2.5.1	Causal counterfactual fairness from a mass-transportation viewpoint	85
2.5.2	Extending counterfactual fairness	86
2.5.3	Counterfactual fairness as nonarbitrary statistical parity	87
2.6	Application to counterfactually fair learning	89
2.6.1	Learning problem	89
2.6.2	Consistency	90
2.7	Numerical experiments	91
2.7.1	Procedure	91
2.7.2	Datasets	93
2.7.3	Results	95
2.7.4	Discussion	97
2.8	Perspectives of extensions	97
2.8.1	Summary of the contributions	98
2.8.2	Improving transport-based counterfactual models	98
2.8.3	The shape of counterfactual fairness constraints	100
2.8.4	Interpretation and applicability of counterfactuals for fairness	104
2.8.5	Transport-based counterfactuals: from theory to practice	105
Appendix 2.A	Proofs of Section 2.4	107
Appendix 2.B	Proofs of Section 2.5	110
Appendix 2.C	Proofs of Section 2.6	115
Appendix 2.D	Proofs of Section 2.8	117

II	Statistical estimation of transport models	121
3	GAN estimation of Lipschitz optimal transport maps	123
3.1	Introduction	123
3.2	Lipschitz neural networks	126
3.2.1	Multivariate GroupSort neural networks	127
3.2.2	Approximating Lipschitz continuous functions	128
3.3	GAN estimator	129
3.3.1	Optimal transport setup	129
3.3.2	GAN setup	130
3.3.3	Main theorem	133
3.4	Numerical experiments	134
3.4.1	Implementation	135
3.4.2	Experimental results	136
3.5	Conclusion	137
3.5.1	Summary of contributions	137
3.5.2	Limitations and further research	137
	Appendix 3.A Proofs of Section 3.2	139
	Appendix 3.B Proofs of Section 3.3	141
	Appendix 3.C Proofs of Section 3.4	141
4	Diffeomorphic registration using Sinkhorn divergences	149
4.1	Introduction	149
4.2	Preliminaries and notations	151
4.2.1	Smooth functions	151
4.2.2	Actions on probability measures	152
4.3	Diffeomorphic measure transportation	152
4.3.1	Generating diffeomorphic deformations	152
4.3.2	Diffeomorphic matching of distributions	154
4.4	Entropic optimal transport	155
4.4.1	Transportation costs and Sinkhorn divergences	155
4.4.2	Regularity of the dual formulation	156
4.5	Main results	157
4.6	Implementation	159
4.6.1	Resolution procedure	159
4.6.2	Numerical experiments	162
4.7	Final remarks	169
4.7.1	Application to counterfactual reasoning	169
4.7.2	Conclusion	169
	Appendix 4.A Preliminary results	171
	4.A.1 Empirical processes	171
	4.A.2 Frechet derivative	172
	Appendix 4.B Proofs of Chapter 4	173
	Conclusion	181

Conclusion (en français)	185
References	189

Notations

This part gathers some key notations and definitions that will be used throughout the manuscript. We point out that this list is not exhaustive, and that each chapter brings some specific notations.

Sets and elements. We write $2^{\mathcal{V}}$ for the *power set* of some set \mathcal{V} , that is the set of all subsets of \mathcal{V} . Additionally, for any collection of spaces $(\mathcal{V}_i)_{i \in \mathcal{I}}$ indexed by a finite index set \mathcal{I} and any subset $I \subseteq \mathcal{I}$ we define the *product space* $\mathcal{V}_I := \prod_{i \in I} \mathcal{V}_i$. Similarly, we write $v_I := (v_i)_{i \in I}$ for any tuple $v := (v_i)_{i \in \mathcal{I}} \in \mathcal{V}_{\mathcal{I}}$. The *union* of two *disjoint* sets \mathcal{V}_1 and \mathcal{V}_2 is written as $\mathcal{V}_1 \sqcup \mathcal{V}_2$. When \mathcal{V} is finite, we denote its *cardinality* by $|\mathcal{V}|$.

Euclidean spaces. The absolute value of real numbers and the Euclidean norm of vectors are respectively given by $|\cdot|$ and $\|\cdot\|$, while \cdot denotes the Euclidean inner product. For an integer $d \geq 1$, the notation B_r refers to the centered Euclidean ball of \mathbb{R}^d with radius $r > 0$. When $\mathcal{V} \subseteq \mathbb{R}^d$, we denote by $\text{diam}(\mathcal{V})$ its *diameter*, that is $\text{diam}(\mathcal{V}) := \max_{v, v' \in \mathcal{V}} \|v - v'\|$. The *closure* of \mathcal{V} is written as $\bar{\mathcal{V}}$; its *interior* as $\overset{\circ}{\mathcal{V}}$. If additionally \mathcal{V} is closed and convex, then Proj_{Ω} stands for the *projection* onto \mathcal{V} .

Linear algebra. We write $\text{Span}(\mathcal{V})$ for the *linear span* of a set $\mathcal{V} \subseteq \mathbb{R}^d$, namely the set of all linear combinations of vectors in \mathcal{V} . If \mathcal{V} is a linear subspace of \mathbb{R}^d , then \mathcal{V}^{\perp} denotes the *orthogonal complement* of \mathcal{V} , that is the set of all vectors in \mathbb{R}^d that are orthogonal to every vector in \mathcal{V} . The set $\mathbb{R}^{p \times d}$ consists of the real-valued matrices with p rows and d columns. The *transpose* of a matrix M is written as M^T ; if M is nonsingular, then M^{-1} refers to its *inverse*. By convention, a vector $x \in \mathbb{R}^d$ is identified to a column matrix, so that $x^T y$ corresponds to the Euclidean inner product between x and $y \in \mathbb{R}^d$. The *zero vector* of \mathbb{R}^d is denoted by 0 whatever the dimension.

Probability measures. Let μ and ν be two Borel probability measures on \mathbb{R}^d . We write $\text{supp}(\mu)$ for the *support* of μ , which is defined as the largest Borel set such that every open set it intersects has positive measure. We say that a proposition $\mathcal{P}(\cdot)$ holds μ -*almost everywhere* if there exists a Borel set E such that $\mu(E) = 1$ and $\mathcal{P}(v)$ is true for any $v \in E$. The relation $\mu \ll \nu$ means that μ is *absolutely continuous* with respect to ν , formally $(\nu(E) = 0 \implies \mu(E) = 0)$ for every Borel set $E \subseteq \mathbb{R}^d$. The symbol \otimes denotes the product of measures, that is $(\mu \otimes \nu)(E_1 \times E_2) = \mu(E_1) \times \nu(E_2)$ for every Borel sets $E_1, E_2 \subseteq \mathbb{R}^d$. The *expectation* (or *mean*) under μ of a measurable function $f : \mathbb{R}^d \rightarrow \mathbb{R}^p$ for an integer $p \geq 1$ is defined as $\mu(f) := \int f d\mu$. The *push-forward measure* of μ by f is defined by

$(f_{\#}\mu)(E) := \mu \circ f^{-1}(E)$ for every Borel set $E \subseteq \mathbb{R}^p$. This operation enables to write changes of variables: for any measurable function g on \mathbb{R}^p , $\int g d(f_{\#}\mu) = \int (g \circ f) d\mu$. The set $\mathcal{T}(\mu, \nu)$ refers to all measurable mappings pushing forward μ to ν , namely all mappings $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$ such that $T_{\#}\mu = \nu$. The set $\Pi(\mu, \nu)$ consists of all joint probability distributions on $\mathbb{R}^d \times \mathbb{R}^d$ with μ and ν as respectively first and second marginals.

Random variables. A *probability space* is a triplet $(\Omega, \Sigma, \mathbb{P})$ composed of a sample space Ω , a σ -algebra $\Sigma \subseteq 2^\Omega$, and a probability measure $\mathbb{P} : \Sigma \rightarrow [0, 1]$. A *random variable* $V : \Omega \rightarrow \mathcal{V}$ is a measurable function from the sample space Ω to some measurable space $\mathcal{V} \subseteq \mathbb{R}^d$ equipped with the Borel σ -algebra. We denote respectively by $\mathcal{L}(V) := V_{\#}\mathbb{P}$ and $\mathbb{E}[V] := \int V(\omega) d\mathbb{P}(\omega)$ the *law* and *expectation* under \mathbb{P} of V . Whenever they are well-defined, the law and expectation of V *conditional to some event* $E \subseteq \Sigma$ are respectively denoted by $\mathcal{L}(V | E)$ and $\mathbb{E}[V | E]$. We write $V_1 \perp V_2$ to signify that two random variables V_1 and V_2 are *independent* under \mathbb{P} , that is $\mathcal{L}((V_1, V_2)) = \mathcal{L}(V_1) \otimes \mathcal{L}(V_2)$. The relation $V_1 \stackrel{\mathbb{P}\text{-a.s.}}{=} V_2$ means that V_1 and V_2 are *almost-surely equal*, formally $\mathbb{P}(V_1 = V_2) = 1$; while $V_1 \stackrel{\mathcal{L}}{=} V_2$ means that they are *equal in law*, that is $\mathcal{L}(V_1) = \mathcal{L}(V_2)$.

Introduction

Eighty years after the first mathematical modeling of a neural network by [McCulloch and Pitts \(1943\)](#), machine learning, the research field developing and studying methods to learn how to make new predictions from past data, has enabled artificial-intelligence systems to solve even more complex problems, from finding correlations within high-dimensional tabular data to processing languages and recognizing images. These advances triggered the proliferation of machine-learning-based algorithms in countless facets of the ordinary and professional life, such as targeted advertisement on social networks, music recommendation on streaming applications, cancer detection in health-care, and stock prices prediction in finance.

However, the massive reliance on such automatic decision rules has raised grave ethical concerns due to their growing lack of interpretability and their potential harm toward minorities. The pursuit of versatility and accuracy in supervised learning, from simple linear regression models to convoluted deep neural networks, led to the emergence of black-box models making decisions that neither the end users nor the designers can comprehend. Not only this violates the legitimate right to explanation of individuals, but this prevents from understanding whether the predictions are fair and adequate. With the prospect of artificial intelligence being involved in critical decisions such as for self-driving cars and recruitment processes, ensuring transparency of algorithmic rules has become indispensable.

This is all the more necessary that machine learning has demonstrated through many incidents over the last few years that it could not be trusted on the basis of accuracy only. Traffic-sign recognition which could be fooled by imperceptible changes,¹ the IBM facial-recognition system which proved to be racist,² the Amazon recruitment algorithm which discriminated against women,³ the Microsoft chatbot which became nazi in twenty-four hours,⁴ are just a few striking examples epitomizing the risk of letting artificial intelligence spread at every level of society without questioning the keystone assumptions of machine learning itself. Machine learning is not objective; it merely finds correlations between an outcome of interest (e.g., being hired) and some covariates (e.g., socioeconomic features, work experience) contained in a dataset inevitably shaped by structural inequalities and human biases. Therefore, it is bound to reproduce and make commonplace the biases reflected in the data, for instance that racialized people are more likely to be criminals. Moreover, learning accurate correlations does not mean learning meaningful features. Spurious or irrelevant

¹<https://spectrum.ieee.org/slight-street-sign-modifications-can-fool-machine-learning-algorithms>

²<https://www.bbc.com/news/technology-52978191>

³<https://www.bbc.com/news/technology-45809919>

⁴<https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist>

correlations abound in data, creating irregularities in the predictions, as illustrated by the notorious dog-versus-wolf classifier biased by the snow background (Ribeiro et al., 2016). Importantly, all these algorithmic biases did not require any premeditated malice from the designers: they are the natural consequence of the very concept of learning from past data through the dogma of accuracy.

These major issues spurred on the emergence of trustworthy artificial intelligence: a research initiative cross-pollinated by social sciences, mathematics and computer science, dedicated to the development of artificial-intelligence systems that are lawful, robust, explainable and fair. In less than ten years, it has become one of the trendiest research topics in the domain, with an exponentially growing number of scientific publications each year. But the need for ethical machine-learning decision rules is not only morally justified, it is also legally grounded. The European Artificial Intelligence Act⁵ will soon regulate all systems deployed in the European Union by categorizing them according to their risk, their possible prejudice on society. Should an application not satisfy the legal requirements on fairness and transparency, it would simply be forbidden. This means that the numerous machine-learning-driven companies are rapidly going to require a rich range of new learning techniques and test procedures to guarantee the trustworthiness of their products.

This thesis entirely applies to this process; the research therein ambitions to contribute to the development of fair and explainable machine-learning algorithms. The specificity of our approach is twofold. First, we study fairness and explainability through the prism of counterfactual reasoning: a modality addressing queries such as “Had she been a man, would have he been granted the loan?” that has become widely used in very recent research to generate post-hoc explanations of automatic decision rules and to uncover biases. Second, we implement this reasoning using transport methods such as optimal transport and diffeomorphic registration, which enable one to match one probability distributions to another (e.g., women to men). Moreover, in a maths-driven approach, we mind to provide theoretical guarantees throughout the manuscript, notably statistical consistency results.

Summary of the thesis

The overall structure of the manuscript reflects the two components of our research: **a first part addresses counterfactual reasoning in fairness and explainability**, providing a general theory for the implementation of counterfactual thinking through mass transportation; **a second part focuses on the statistical inference of transport models**, which furnishes tools to apply ideas from the first part. This summary aims at giving in less than ten pages a general idea of our contributions contrasted with the state of the art.

Part I, Chapter 1: Counterfactuals, explainability, fairness

For starters, we introduce counterfactual reasoning and its applications through an overview of the corresponding scientific literature. This chapter is motivated by the following observation: while most of the researchers and practitioners from machine-learning-related fields have

⁵<https://artificialintelligenceact.eu/the-act/>

become used to hearing expressions such as *counterfactual explanations*, *counterfactual outcomes*, or *counterfactually fair*, they generally lack a global understanding of what connects and distinguishes these notions. By unifying counterfactual frameworks, we shed a fresh light on counterfactual reasoning in machine learning, beyond mere reviewing.

What is a counterfactual?

In a nutshell, a counterfactual is a statement of the form “Had A occurred then B would have occurred”, for example “Had Alice been a man, he would have been granted the position”. The treatment of such statements is obviously challenging, as we cannot *observe* an alternative reality where Alice is a man. Counterfactual reasoning refers to all the theories and techniques addressing the problem of designing such alternatives, therefore enabling the assessment of counterfactual statements. Several frameworks coexist, leveraging different tools to compute alternative worlds, and suggesting different choices of what must be kept equal across them. They all stem from the seminal *possible-world account* of Lewis (1973b), well-known by the logician community, which uses a general notion of closeness between worlds to define possible alternative worlds. Throughout the chapter, we leverage this basis as a common reading grid to provide a unified understanding of counterfactual theories.

Causal modeling

Then, we present two acclaimed causal theories allowing counterfactual reasoning, which both leverage a terminology based on random variables.

Firstly, we detail the notorious *structural account* of Pearl (2009). It rests on the knowledge of a *structural causal model*, specifying all cause-to-effect equations between observed random variables $(V_i)_{i \in \mathcal{I}}$ through a collection of assignments of the form,

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

where the variables $(U_j)_{j \in \mathcal{J}}$ represent latent primary causes of the model, while $\text{Endo}(i) \subseteq \mathcal{I}$ and $\text{Exo}(i) \subseteq \mathcal{J}$ refer respectively to the so-called *endogenous* and *exogenous* parents of the i th variable. The interest of such equations comes from the possibility of carrying out *do-interventions*: forcing an endogenous variable to take a given value while keeping the rest of the mechanism untouched. More concretely, let $V := (X, S)$ be a random vector of observed variables such that we would like to understand the downstream effect of $S : \Omega \rightarrow \mathcal{S}$ onto other features $X : \Omega \rightarrow \mathbb{R}^d$. Replacing the formula generating S by $S = s$ and propagating this change through the other equations defines altered features $X_{S=s}$, representing X had S been equal to s .

Secondly, we turn to the widely-used *potential-outcome account* of Rubin (1974) which mathematically formalizes causal inference in clinical trials. Letting S denote a binary *treatment status* (e.g., taking a drug or not) and $Y : \Omega \rightarrow \mathcal{Y}$ an *outcome* of interest (e.g., recovering or not), this framework postulates the existence of two *potential outcomes* Y_0 and Y_1 such that $Y = (1 - S) \cdot Y_0 + S \cdot Y_1$. These variables respectively represent what the outcome would be were S equal to 0 or 1. The *fundamental problem of causal inference* (Holland, 1986) refers to the fact that in practice we cannot observe simultaneously Y_0 and Y_1 , rendering unidentifiable the causal effect of S onto Y . Notably, in general, *correlation is not causation*

in the sense that $\mathcal{L}(Y_s) \neq \mathcal{L}(Y | S = s)$ for $s \in \{0, 1\}$. Nevertheless, causal inference at a global scale can still be achieved thanks to a mix of untestable assumptions and statistical tools. Adjusting on a set of available covariates X containing all possible *confounders* between the treatment and the potential outcomes, formally $(Y_0, Y_1) \perp\!\!\!\perp S | X$, enables to identify counterfactual outcome as it entails that $\mathcal{L}(Y_s | X = x) = \mathcal{L}(Y | S = s, X = x)$.

Finally, as an original contribution, **we superimpose these two causal models to derive a mathematical analysis of the similarities and differences between potential-outcome counterfactuals and structural counterfactuals by demonstrating that—in contrary to what the mainstream literature often suggests— $Y_s \neq Y_{S=s}$ in general.** More precisely, these counterfactual outcomes differ if S encodes a nonmanipulable feature that impacts the covariate X such as sex and race, but coincide if S is a treatment that can be experimentally allocated such as a drug. This critically means that the two causal approaches generate different counterfactual statements in classical fairness problems where S typically encodes sex or race, hence why we must understand the distinction. The idea is that the potential-outcome framework considers counterfactual outcomes with fixed covariates X , while structural causal modeling alters the covariates into $X_{S=s}$. We illustrate the consequences of this difference on a concrete fairness example.

Application to explainability and fairness

Lastly, we present popular applications of counterfactual reasoning in explainable artificial intelligence and algorithmic fairness, emphasizing its critical role for building trustworthy systems.

In their pioneering work, Wachter et al. (2017) propose to use counterfactual statements as a psychology-grounded approach for explaining black-box decision rules. Concretely, they look for minimal plausible changes of an input v such that the output of the machine-learning model h differs, generally by solving

$$\min_{v' \in \mathcal{V}} c(v, v') + \lambda |h(v) - h(v')|^2,$$

where c is a well-chosen cost function, λ a large-enough parameter for the decision to change, and \mathcal{V} the set of plausible worlds. This generates counterfactual statements of the form “Had the input been v' (instead of v), then output would have been different”. For the sake of clarification and unification, we show how to interface this explanation methods to the previously detailed counterfactual-inference frameworks. This can be achieved by parametrizing the artificial inputs v' as the alternative worlds of v obtained after an action, for example a do-intervention. This part notably serves to emphasize the distinction between counterfactual counterparts (i.e., alternative worlds had a certain event occurred) and counterfactual explanations (i.e., refined adversarial examples).

Turning to algorithmic fairness, we illustrate how counterfactuals can provide strong and intuitive notions of fairness. For S a *protected attribute* (e.g, race or sex) and X the other input features of a predictor h , fairness deals with rendering $h(X, S)$ independent of S , typically by enforcing $h(X, S) \perp\!\!\!\perp S$. However, as noted by Dwork et al. (2012), this so-called *statistical parity* is a group-fairness constraint that does not control for discrimination at the individual level. This is where counterfactual inference comes into play: by enabling to compute alternative inputs “had the protected attribute been changed” (e.g., man to

woman), it allows for requiring equal treatment between an individual and its *counterfactual counterparts*. This idea is at the core of the definition of *counterfactual fairness*, proposed by [Kusner et al. \(2017\)](#), which relies on Pearl’s do-calculus to ask

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

for every observations $\{X = x, S = s\}$ and any $s' \neq s$. The main drawback of this approach comes from its dependence to a structural causal model: while appealing in theory, such models are unknown in practice and can hardly be inferred from data, making causal-based methods unfeasible in most real-world tasks. This limitation motivated the work presented in the next chapter, where we introduce noncausal, *transport-based* counterfactual models providing feasible notions of individual fairness, sharper than group fairness constraints.

Part I, Chapter 2: Transport-based counterfactual models

The explainability and fairness applications reviewed in the previous chapter show that the most commonly used techniques for the computation of counterfactual counterparts (i.e., alternative worlds) in machine learning are the *closest alternative world principle* and Pearl’s causal modeling. The first approach is straightforward, but neglects correlation between features, leading to *unfaithful*, out-of-distribution counterfactuals such as “Had Bob been a woman, she would have been 190cm tall”. The second one rigorously takes into account all dependencies between variables, but requires a structural causal model that is unknown or too hard to infer in practice, making it *unfeasible* except for toy cases. This chapter, based on [De Lara et al. \(2021a\)](#), focuses on a third way by interpreting counterfactual reasoning as a problem of mass displacement from one probability distribution to another. It follows the work of [Black et al. \(2020\)](#) who first suggested the use of *optimal transport* ([Villani, 2003, 2008](#)) to design feasible and realistic counterfactuals.

Mass transportation and optimal transport

First of all, let us introduce some basic knowledge on mass transportation and optimal transport. We refer to *mass transportation* as the general problem of matching two probability distributions P and Q on $\mathcal{X} \subseteq \mathbb{R}^d$. This amounts to selecting a coupling within the set $\Pi(P, Q)$ of joint probability distributions with respectively P and Q as first and second marginals. A coupling can be seen as a random mapping, matching each instance of P to possibly several counterparts in Q with probability weights. It is said to be *deterministic* if each instance from P is matched to a unique instance from Q . In this case, the coupling is concentrated on the graph of a (P -almost every unique) deterministic mapping $T : \mathcal{X} \rightarrow \mathcal{X}$ *pushing forward* P to Q , that is $Q(E) = P(T^{-1}(E))$ for every Borel set $E \subseteq \Omega$. This property, denoted by $T_{\#}P = Q$, means that if a random variable X follows the distribution P then its image $T(X)$ follows the distribution Q .

Optimal transport theory became the most popular tool to construct such couplings when no canonical choice is available. It dates back to [Monge \(1781\)](#) who defined *optimal transport maps* as functions transforming P into Q with minimal effort according to a positive ground cost function $c : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$. Formally, these maps solve

$$\min_{T \in \mathcal{T}(P, Q)} \int_{\mathcal{X}} c(x, T(x)) dP(x),$$

where $\mathcal{T}(P, Q)$ is the set of measurable maps pushing forward P to Q . In general settings however, such a deterministic correspondence between probability distributions may not exist, in particular if P and Q are not Lebesgue absolutely continuous. This limitation motivates the so-called *Kantorovich* relaxation of Monge’s formulation of optimal transport (Kantorovich and Rubinshtein, 1958):

$$\min_{\pi \in \Pi(P, Q)} \int_{\mathcal{X}^2} c(x, x') d\pi(x, x').$$

This problem focuses on random couplings instead of deterministic mappings, and always admits solutions referred as *optimal transport plans*.

Counterfactual reasoning from a mass-transportation viewpoint

Counterfactual reasoning addresses queries of the form *had S been equal to s' instead of s , what would have been the value of X ?* The structural account answers this question through do-calculus: for a given world of reference $\{X = x, S = s\}$ it returns a distribution of counterfactual counterparts given by $\mathcal{L}(X_{S=s'} \mid X = x, S = s)$.

In contrast, Black et al. (2020) approximated an optimal transport map $T_{\langle s'|s \rangle}$ from $\mu_s := \mathcal{L}(X \mid S = s)$ to $\mu_{s'} := \mathcal{L}(X \mid S = s')$ and generated the counterfactual counterpart had S been equal to s' of an factual s -instance x by $T_{\langle s'|s \rangle}(x)$. Although this was not addressed in their paper, this idea can be naturally generalized to optimal transport plans $\pi_{\langle s'|s \rangle} \in \Pi(\mu_s, \mu_{s'})$ rather than maps. All in all, this observation-based, noncausal approach has three critical interests. Firstly, it guarantees that the generated counterfactuals are in-distribution, hence realistic. Secondly, it benefits from a growing repertoire of efficient numerical schemes to estimate such couplings from data (Peyré and Cuturi, 2019), leading to computationally feasible counterfactuals. Thirdly, on the contrary to the structural account, it obviates any assumptions on the data generation process.

As a first contribution, **we propose to unify these two frameworks under a common mass-transportation formalism** grounded by the following remark: cross-world statements “had S been equal to s' instead of s ” using Pearl’s causal modeling are characterized by the joint probability distribution $\pi_{\langle s'|s \rangle}^* := \mathcal{L}((X, X_{S=s'}) \mid S = s)$, matching any *factual* world from μ_s to its *counterfactual* counterparts in $\mu_{\langle s'|s \rangle} := \mathcal{L}(X_{S=s'} \mid S = s)$. From this perspective, structural counterfactual reasoning is, similarly to optimal transport, a problem of mass transportation. This notably means that Black et al. (2020) were implicitly mimicking structural counterfactuals. Extending their idea, **we define transport-based counterfactual models on X with respect to S as collections of couplings $\Pi := \{\pi_{\langle s'|s \rangle}\}_{s, s' \in \mathcal{S}}$ such that for every $s, s' \in \mathcal{S}$, $\pi_{\langle s'|s \rangle} \in \Pi(\mu_s, \mu_{s'})$** . This general construction encompasses couplings defined through optimal transport, but also regularized optimal transport and potentially new mass-transportation techniques. The critical interest of the unified mass-transportation formalism lies in the possibility to replace a structural counterfactual coupling $\pi_{\langle s'|s \rangle}^*$ by a transport-based one $\pi_{\langle s'|s \rangle}$ in any causal counterfactual framework to generate a transport-based alternative; this will be later illustrated with counterfactual fairness.

When optimal transport meets causality

Interestingly, **the mass-transportation formalism we introduced also enables us to mathematically demonstrate that optimal transport recovers causal changes in some settings**, which explains why [Black et al. \(2020\)](#) empirically observed that structural counterfactuals and optimal-transport counterfactuals were nearly identical. This result holds under two typical assumptions of the causal model.

The first assumption demands the structural counterfactuals to be deterministically implied by the causal model. This formally corresponds to $\mathcal{L}(X_{S=s'} \mid X = x, S = s)$ narrowing down to a Dirac distribution, which occurs in particular when the exogenous variables are additive terms of the structural equations. This entails that the causal coupling $\pi_{\langle s'|s \rangle}^*$ is deterministic: it can be identified to a mapping $T_{\langle s'|s \rangle}^*$ such that $T_{\langle s'|s \rangle}^* \mu_s = \mu_{\langle s'|s \rangle}$.

The second assumption requires the modified variable S to play the same role as an exogenous variable in the structural causal model. This holds typically in fairness problem: for example, the race is exogenous relatively to socioeconomic features. This implies that counterfactual counterparts are observable in the sense that $\mu_{\langle s'|s \rangle} = \mu_{s'}$, therefore that the *structural counterfactual model* $\Pi^* := \{\pi_{\langle s'|s \rangle}^*\}_{s,s' \in \mathcal{S}}$ is a transport-based counterfactual model.

Under both assumptions, the structural causal model induces a deterministic transport-based counterfactual model. As such, one may naturally wonder whether it could be recovered by solving a deterministic optimal-transport problem. This precisely what our main theorem states for the quadratic cost $c(x, x') := \|x - x'\|^2$: if $\mathcal{L}(X)$ is absolutely continuous with respect to the Lebesgue measure and have a finite second order moment, then

$$\pi_{\langle s'|s \rangle}^* = \arg \min_{\pi \in \Pi(\mu_s, \mu_{s'})} \int \|x - x'\|^2 d\pi(x, x')$$

if and only if $T_{\langle s'|s \rangle}^*$ is the gradient of a convex function. This condition (which probably rings a bell to people familiar with optimal-transport theory) is notably satisfied when the causal equations of the model are linear additive over X , namely of the form

$$X \stackrel{\mathbb{P}\text{-a.s.}}{=} MX + wS + U_X,$$

where $M \in \mathbb{R}^{d \times d}$ and $w \in \mathbb{R}$ are deterministic parameters. The interest of this theorem is twofold: it explains theoretically the empirical observations of [Black et al. \(2020\)](#); it further justifies that optimal transport works as a decent—noncausal—alternative to structural counterfactual reasoning in typical fairness scenarios.

Application to fairness

This analysis motivates the use of transport-based counterfactual models in the place of structural counterfactual models to derive new notions of fairness, being both sharp and feasible. We note in particular that the counterfactual fairness criterion introduced in the previous chapter can be written as: for every $s, s' \in \mathcal{S}$ and $\pi_{\langle s'|s \rangle}^*$ -almost every (x, x') ,

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

Replacing the causal couplings by transport-based ones from a model Π leads a new fairness definition that we refer as Π -counterfactual fairness. Although it trades-off causality for sound correlations, it preserves the property of ensuring fairness at the individual level, is stronger than statistical parity, and does not require assumptions on the data generation process.

Next, adapting the work of [Russell et al. \(2017\)](#), **we tackle the problem of learning a Π -counterfactually fair predictor.** This amounts to solving

$$\min_{\theta \in \Theta} \mathbb{E}[\ell(h_\theta(X, S), Y)] + \lambda \sum_{s \in \mathcal{S}} \mathbb{P}(S = s) \sum_{s' \neq s} \mathbb{E} \left[|h_\theta(X_{s'}, s') - h_\theta(X_s, s)|^2 \right],$$

where Y is the target variable, ℓ a data-fidelity loss function, $\{h_\theta\}_{\theta \in \Theta}$ a parametric class of predictors, and $\mathcal{L}((X_s, X_{s'})) = \pi_{\langle s'|s \rangle}$ for every $s, s' \in \mathcal{S}$. Theoretically, we prove in the case of quadratic optimal transport that the empirical solution converges almost-surely to its population counterpart as n , the sample size, tends to infinity; experimentally, we showcase the performances of our estimator on real datasets. To sum-up, our contribution expands the fair-learning arsenal to stronger criteria than mere group fairness condition by relaxing causality to transport methods.

Part II, Chapter 3: GAN estimation of Lipschitz optimal transport maps

In [\(Black et al., 2020\)](#), the authors employed optimal-transport counterfactuals to *audit* the fairness of black-box decision rules, while we used them to *learn* a fair predictor in the previous chapter. Whatever the situation, implementing transport-based counterfactual model requires constructing transport plans or maps from data. The richer the estimation methods at hand, the more problems we will be able to handle. This chapter, based on [\(González-Sanz et al., 2022\)](#), contributes to the growing literature on approximating optimal transport maps. It combines the generative-adversarial-network (GAN) approach of [Black et al. \(2020\)](#) with Lipschitz neural networks ([Anil et al., 2019](#); [Tanielian and Biau, 2021](#)) to design a novel neural estimator with provable statistical guarantees.

More precisely, we consider P and Q , two Lebesgue-absolutely-continuous measures such that the optimal transport map T_0 from P to Q for the quadratic cost is smooth, in particular Lipschitz. Then, on the basis of empirical distributions P_n and Q_n respectively drawn from P and Q , we approximate T_0 by the solution to the GAN problem

$$\inf_{G \in \mathcal{G}_n} \left\{ \int \|I - G\|^2 dP_n + \lambda_n \sup_{D \in \mathcal{D}_n} \int D(d(G_\# P_n) - dQ_n) \right\},$$

where \mathcal{D}_n is a class of 1-Lipschitz discriminators providing a proxy for the dual formulation of the Wasserstein-1 distance (in the Wasserstein-GAN fashion of [Arjovsky et al. \(2017\)](#)), and \mathcal{G}_n is a class of Lipschitz generators parametrizing the space of feasible mappings. The positive parameter λ_n governs the trade-off between minimizing the quadratic transportation cost, promoting the objective of the Monge problem, and minimizing the distance between the generated and the target distributions, enforcing the push-forward constraint.

Our objective is twofold: first, we aim at designing an expressive optimal transport map benefiting from a neural architecture to efficiently generalize to new out-of-sample observations; secondly, we aim at providing statistical guarantees, which was overlooked by

most of the related papers (Leygonie et al., 2019; Black et al., 2020; Makuva et al., 2020; Korotin et al., 2021; Huang et al., 2021).

Multivariate GroupSort neural networks

The problem described above involves Lipschitz neural networks, be they for the discriminators or generators. In this chapter, we leverage the recently introduced *GroupSort* activation functions to impose the Lipschitz constraint (Anil et al., 2019; Tanielian and Biau, 2021), which have proven to yield tighter estimates of 1-Lipschitz functions than previous methods such as in (Arjovsky et al., 2017; Gulrajani et al., 2017). By definition, the *GroupSort activation function* of grouping size $k \geq 2$ splits the pre-activation input into groups of size k , and then sorts each group by decreasing order. This operation is 1-Lipschitz, gradient-norm preserving and homogeneous (Anil et al., 2019). A *GroupSort neural network* is a feed-forward neural network where all activation functions (except for the first layer) are a GroupSort activation function of fixed grouping size k ; in this work we restrict to $k = 2$. Under a compactness assumption on the weights, these networks are 1-Lipschitz functions.

The specificity of the improved Wasserstein-GAN problem we tackle lies in parametrizing the *multivariate* generator as a 1-Lipschitz mapping; this was already addressed for the *univariate* discriminator in (Anil et al., 2019; Biau et al., 2021). This requires extending to multivariate GroupSort neural networks the theoretical results in (Tanielian and Biau, 2021), focusing on the approximation power of univariate GroupSort neural networks. **Notably, we demonstrate that for an arbitrary output dimension, GroupSort neural networks can approximate with given precision any bounded subclass of 1-Lipschitz functions.**

GAN estimator

On the basis of this approximation result, we define the generators \mathcal{G}_n and discriminators \mathcal{D}_n as GroupSort neural networks with well-chosen n -dependent sizes. To ensure statistical convergence, the sequence of feasible generators $\{\mathcal{G}_n\}_{n \in \mathbb{N}}$ must fill \mathcal{G} sufficiently fast, while the sequence of regularization weights $\{\lambda_n\}_{n \in \mathbb{N}}$ must tend to infinity in order to impose the push-forward condition at the limit. **We demonstrate under these assumptions the almost-surely uniform convergence of $\{\mathcal{G}_n\}_{n \in \mathbb{N}}$ to T_0 .** The proof rests on relative-compactness properties of Lipschitz functions along with the regularity of T_0 through results from (Hütter and Rigollet, 2021). Then, we illustrate the performances of our approximation on synthetic datasets.

Part II, Chapter 4: Diffeomorphic registration using Sinkhorn divergences

Optimal transport theory is not the only way to match probability distributions; although it has rarely been applied to machine-learning tasks, diffeomorphic registration enjoys well-established theory and algorithms. This fluid-mechanics-inspired framework searches an optimal velocity fields of the ambient space transferring one distribution to the other. This chapter, based on (De Lara et al., 2023), expands the mass-transportation toolbox by studying theoretically and experimentally diffeomorphic registration driven by Sinkhorn divergences, entropy-regularized optimal-transport metrics.

Diffeomorphic mass transportation

Under regularity assumptions, a velocity field $v_t(x) \in \mathbb{R}^d$ of variables $t \in [0, 1]$ and $x \in \mathbb{R}^d$ generates a family of diffeomorphisms $(\phi_t^v)_{t \in [0, 1]}$ through the flow equation:

$$\phi_t^v(x) := x + \int_0^t v_s(\phi_s^v(x)) \, ds.$$

Diffeomorphic mass transportation focuses on velocity fields v such that ϕ_1^v matches an input probability distribution α to a target β . Formally, for Λ a positive loss function between measures, this amounts to solve

$$\min_{v \in L_V^2} J_\lambda(v) \text{ with } J_\lambda(v) := \Lambda(\phi_{1\#}^v \alpha, \beta) + \lambda \|v\|_{L_V^2}^2,$$

where L_V^2 is a Hilbert space of vector fields v with finite kinetic energy $\|v\|_{L_V^2}^2$, and the regularization quantified by $\lambda > 0$ ensures that the problem is well-posed.

As explained by [Feydy et al. \(2017\)](#), the nonconvexity of this optimization program renders crucial the choice of the loss function to avoid poor local minima, whereas the gold-standard squares of maximum mean discrepancies produce many. This chapter remedies to this issue by proposing a grounded alternative.

Entropic optimal transport

For $\varepsilon > 0$, the *entropy-regularized* transportation cost with respect to the ground cost $c : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}_+$ between probability measures α and β on $\mathcal{X} \subseteq \mathbb{R}^d$ is defined as

$$\mathcal{T}_{c, \varepsilon}(\alpha, \beta) := \min_{\pi \in \Pi(\alpha, \beta)} \int_{\mathcal{X} \times \mathcal{X}} c(x, y) d\pi(x, y) + \varepsilon \text{KL}(\pi | \alpha \otimes \beta),$$

where $\text{KL}(\mu | \nu)$ denotes the *Kullback-Leibler* divergence between probability measures μ and ν given by $\int \log\left(\frac{d\mu}{d\nu}(z)\right) d\mu(z)$ if $\mu \ll \nu$, and $+\infty$ otherwise. In [Feydy et al. \(2017\)](#), the authors leverage this cost for the loss function Λ , as it benefits from fast numerical schemes alleviating the computational burden of nonregularized optimal transport [\(Cuturi, 2013\)](#), and engenders less local minima than maximum mean discrepancies. However, this choice suffers from the so-called *entropic bias*, that is $\mathcal{T}_{c, \varepsilon}(\alpha, \alpha) \neq 0$ in general, making it an unreliable loss function. This why we propose to use Sinkhorn divergences, *unbiased* entropic transportation costs, defined as,

$$S_{c, \varepsilon}(\alpha, \beta) := \mathcal{T}_{c, \varepsilon}(\alpha, \beta) - \frac{1}{2} \mathcal{T}_{c, \varepsilon}(\alpha, \alpha) - \frac{1}{2} \mathcal{T}_{c, \varepsilon}(\beta, \beta).$$

Under regularity assumptions on c , α and β , this divergence satisfies several desirable properties for a loss function: it is nonnegative, convex in both input measures, and metrizes the convergence in law [\(Feydy et al., 2019\)](#).

Statistical convergence

As always, we do not have access to the true measures α and β in practice but to empirical counterparts α_n and β_n , which raises the question of the convergence of empirical minimizers $\{v^n\}_{n \in \mathbb{N}}$ towards a minimizer of J_λ as the sample size n tends to infinity. If Λ is the square of a maximum mean discrepancy, then according to [Glaunes et al. \(2004\)](#) there exists a minimizer of J_λ denoted by v^* such that up to the extraction of a subsequence

$$\|v^n - v^*\|_{L_V^2} \xrightarrow[n \rightarrow \infty]{a.s.} 0.$$

We demonstrate stronger statistical guarantees when Λ is a Sinkhorn divergence by specifying the convergence rate. Under regularity assumptions on the space L_V^2 , the cost c , and the measures α and β , there exists a constant $A > 0$ such that

$$\mathbb{E} [|J_\lambda(v^n) - J_\lambda(v^*)|] \leq \frac{A}{\sqrt{n}}.$$

The proof leverages the regularity of the dual formulation of entropic optimal transport, along with empirical-process arguments.

Numerical experiments

We conclude this chapter by benchmarking the use of Sinkhorn divergences for diffeomorphic matching against maximum mean discrepancies and biased transportation costs across different parameters and solving procedures. This shows the advantages of this choice compared to previous options.

As mentioned, most of these chapters are based on research papers I have written over the last three years. Although I have homogenized them for this manuscript, they preserve their original structures and are therefore self-contained.

Introduction (en français)

Quatre-vingts ans après la première modélisation mathématique d'un réseau de neurones par [McCulloch and Pitts \(1943\)](#), l'apprentissage automatique, le domaine de recherche qui développe et étudie des méthodes pour apprendre à faire de nouvelles prédictions à partir de données antérieures, a permis aux systèmes d'intelligence artificielle de résoudre des problèmes toujours plus complexes, allant de la recherche de corrélations dans des données tabulaires de grande dimension jusqu'au traitement du langage et à la reconnaissance d'images. Ces avancées ont permis la prolifération d'algorithmes basés sur l'apprentissage automatique dans d'innombrables facettes de la vie ordinaire et professionnelle, telles que la publicité ciblée sur les réseaux sociaux, la recommandation de musique sur les applications de streaming, la détection du cancer en médecine, et la prédiction des cours boursiers en finance.

Toutefois, le recours massif à ces règles de décision automatiques a soulevé de graves préoccupations éthiques en raison de leur manque croissant d'interprétabilité et de leur préjudice potentiel pour les minorités. La quête effrénée de polyvalence et de précision dans l'apprentissage supervisé, allant de simples modèles de régression linéaire jusqu'aux réseaux de neurones profonds, a conduit à l'émergence de modèles "boîte noire" prenant des décisions que ni les utilisateurs finaux ni les concepteurs ne peuvent comprendre. Non seulement cela viole le droit légitime des individus à l'explication, mais cela empêche de comprendre si les prédictions sont justes et adéquates. Dans la perspective de l'implication de l'intelligence artificielle dans des décisions cruciales telles que pour les voitures autonomes et les processus de recrutement, il est devenu indispensable de garantir la transparence des règles algorithmiques.

C'est d'autant plus nécessaire que l'apprentissage automatique a démontré à travers de nombreux incidents au cours des dernières années qu'on ne pouvait pas lui faire confiance sur la base de sa seule précision. La reconnaissance des panneaux de signalisation qui peut être trompée par des changements imperceptibles⁶, le système de reconnaissance faciale d'IBM qui s'est révélé raciste⁷, l'algorithme de recrutement d'Amazon qui a discriminé les femmes⁸, le chatbot de Microsoft qui est devenu nazi en vingt-quatre heures⁹, ne sont que quelques exemples frappants illustrant le risque de laisser l'intelligence artificielle se répandre à tous les niveaux de la société sans remettre en question les hypothèses de base de l'apprentissage automatique lui-même. L'apprentissage automatique n'est pas objectif ; il se contente de trouver des corrélations entre un phénomène d'intérêt (par exemple,

⁶<https://spectrum.ieee.org/slight-street-sign-modifications-can-fool-machine-learning-algorithms>

⁷<https://www.bbc.com/news/technology-52978191>

⁸<https://www.bbc.com/news/technology-45809919>

⁹<https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist>

être embauché) et certaines covariables (comme les caractéristiques socio-économiques ou l'expérience professionnelle) contenues dans un jeu de données inévitablement façonné par des inégalités structurelles et des préjugés humains. Par conséquent, il ne peut que reproduire et banaliser les biais reflétés dans les données, par exemple le fait que les personnes racisées sont plus susceptibles d'être des criminels. En outre, l'apprentissage de corrélations précises ne signifie pas l'apprentissage d'associations significatives. Des corrélations parasites ou non pertinentes abondent dans les données, créant des irrégularités dans les prédictions, comme l'illustre le célèbre classificateur "chien contre loup" biaisé par l'arrière-plan enneigé (Ribeiro et al., 2016). Il est important de noter que tous ces biais algorithmiques ne nécessitent aucune malveillance préméditée de la part des concepteurs : ils sont la conséquence naturelle du concept même d'apprentissage à partir de données antérieures par le dogme de la précision.

Ces problèmes majeurs ont conduit à l'émergence du domaine de l'intelligence artificielle digne de confiance : une initiative de recherche croisée entre les sciences sociales, les mathématiques et l'informatique, consacrée au développement de systèmes d'intelligence artificielle légitimes, robustes, explicables et équitables. En moins de dix ans, c'est devenue l'un des sujets de recherche les plus en vogue de l'intelligence artificielle, avec un nombre de publications scientifiques en croissance exponentielle chaque année. Mais le besoin de règles de décision automatiques de confiance n'est pas seulement justifié d'un point de vue moral, il est également fondé d'un point de vue juridique. La loi européenne sur l'intelligence artificielle¹⁰ réglera bientôt tous les systèmes déployés dans l'Union Européenne en les classant en fonction de leur risque, de leur préjudice éventuel pour la société. Si une application ne satisfait pas aux exigences légales en matière d'équité et de transparence, elle sera tout simplement interdite. Cela signifie que les nombreuses entreprises utilisant de l'apprentissage automatique vont rapidement avoir besoin d'un large éventail de nouvelles techniques d'apprentissage et de procédures de test pour garantir la fiabilité de leurs produits.

Cette thèse s'inscrit dans cette dynamique ; la recherche qu'elle porte a pour ambition de contribuer au développement d'algorithmes d'apprentissage automatique justes et explicables. La spécificité de notre approche est double. Premièrement, nous étudions l'équité et l'explicabilité à travers le prisme du raisonnement contrefactuel : une modalité qui traite des requêtes telles que "Si elle avait été un homme, lui aurait-on accordé le prêt?" qui est devenue courante dans la récente recherche pour générer des explications post-hoc des règles de décision automatiques et pour mettre en évidence leurs biais. Deuxièmement, nous mettons en œuvre ce raisonnement à l'aide de méthodes de transport telles que le transport optimal et l'appariement difféomorphe, qui permettent de faire correspondre une distribution de probabilités avec une autre (par exemple, les femmes aux hommes). De plus, dans le cadre d'une approche mathématiques, nous nous efforçons de fournir des garanties théoriques tout au long du manuscrit, notamment des résultats de convergence statistique.

Résumé de la thèse

La structure générale du manuscrit reflète les deux composantes de notre recherche : **une première partie traite du raisonnement contrefactuel dans l'équité et l'explicabilité,**

¹⁰<https://artificialintelligenceact.eu/the-act/>

fournissant une théorie générale pour la mise en œuvre du raisonnement contrefactuel par transport de masse ; **une deuxième partie se concentre sur l’inférence statistique des modèles de transport**, fournissant des outils pour appliquer les idées de la première partie. Ce résumé vise à donner en moins de dix pages une idée générale de nos contributions par rapport à l’état de l’art.

Partie I, Chapitre 1: Contrefactuels, explicabilité, équité

Pour commencer, nous présentons le raisonnement contrefactuel et ses applications à travers un aperçu de la littérature scientifique associée. Ce chapitre est motivé par l’observation suivante : bien que la plupart des chercheur·euse·s et des praticien·ne·s des domaines liés à l’apprentissage automatique ont l’habitude d’entendre des expressions telles que *explications contrefactuelles*, *résultats contrefactuels*, ou *équité contrefactuelle*, iels manquent généralement d’une compréhension globale de ce qui relie et distingue ces notions. En unifiant les méthodologies contrefactuelles, nous apportons un éclairage nouveau sur le raisonnement contrefactuel dans l’apprentissage automatique, au-delà de la simple revue.

Qu’est-ce qu’un contrefactuel?

En bref, un contrefactuel est un énoncé de la forme “Si A s’était produit, alors B se serait produit”, par exemple “Si Alice avait été un homme, il aurait obtenu le poste”. La vérification de tels énoncés est évidemment difficile, car nous ne pouvons pas observer une réalité alternative où Alice est un homme. Le raisonnement contrefactuel fait référence à toutes les théories et techniques qui traitent du problème de la conception de telles alternatives, permettant ainsi l’évaluation d’énoncés contrefactuels. Plusieurs cadres coexistent, exploitant différents outils pour calculer ces mondes alternatifs et suggérant différents choix de ce qui doit rester égal entre eux. Ils découlent tous du principe fondateur des *mondes possibles* de [Lewis \(1973b\)](#), bien connu de la communauté des logiciens, qui utilise une notion générale de proximité entre mondes pour définir les mondes alternatifs possibles. Tout au long du chapitre, nous nous appuyons sur cette base comme grille de lecture commune afin de fournir une compréhension unifiée des théories contrefactuelles.

Modélisation causale

Nous présentons ensuite deux grandes théories de la causalité qui permettent un raisonnement contrefactuel, s’écrivant toutes deux à l’aide de variables aléatoires.

Tout d’abord, nous détaillons le célèbre *point de vue structurel* de [Pearl \(2009\)](#). Il repose sur la connaissance d’un *modèle causal structurel*, spécifiant toutes les équations de cause à effet entre les variables aléatoires observées $(V_i)_{i \in \mathcal{I}}$ par une collection d’affectations de la forme,

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

où les variables $(U_j)_{j \in \mathcal{J}}$ représentent les causes primaires latentes du modèle, tandis que $\text{Endo}(i) \subseteq \mathcal{I}$ et $\text{Exo}(i) \subseteq \mathcal{J}$ réfèrent respectivement aux parents dits *endogènes* et *exogènes* de la i ème variable. L’intérêt de ces équations vient de la possibilité d’effectuer des *interventions* : forcer une variable endogène à prendre une valeur donnée tout en gardant le reste du

mécanisme intact. Plus concrètement, supposons que $V := (X, S)$ soit un vecteur aléatoire de variables observées et que nous souhaitons comprendre l'effet en aval de $S : \Omega \rightarrow \mathcal{S}$ sur d'autres caractéristiques $X : \Omega \rightarrow \mathbb{R}^d$. Le remplacement de la formule générant S par $S = s$ et la propagation de ce changement à travers les autres équations définissent une variable modifiée $X_{S=s}$, représentant X si S avait été égal à s .

Deuxièmement, nous nous tournons vers le modèle couramment utilisé des *résultats potentiels* de Rubin (1974), qui formalise mathématiquement l'inférence causale dans les essais cliniques. En supposant que S représente un *statut de traitement* binaire (par exemple, prise ou non d'un médicament) et $Y : \Omega \rightarrow \mathcal{Y}$ un *résultat* d'intérêt (par exemple, guérison ou non), ce cadre de pensée postule l'existence de deux *résultats potentiels* Y_0 et Y_1 tels que $Y = (1 - S) \cdot Y_0 + S \cdot Y_1$. Ces variables représentent respectivement ce que serait le résultat si S était égal à 0 ou à 1. Le *problème fondamental de l'inférence causale* (Holland, 1986) fait référence au fait que, dans la pratique, nous ne pouvons pas observer simultanément Y_0 et Y_1 , ce qui rend non identifiable l'effet causal de S sur Y . Notamment, en général, *corrélacion ne veut pas dire causalité* dans le sens où $\mathcal{L}(Y_s) \neq \mathcal{L}(Y | S = s)$ pour $s \in \{0, 1\}$. Néanmoins, l'inférence causale est encore possible grâce à un mélange d'hypothèses non testables et d'outils statistiques. L'ajustement sur un ensemble de covariables mesurées X contenant tous les *facteurs de confusion* possibles entre le traitement et les résultats potentiels, formellement $(Y_0, Y_1) \perp\!\!\!\perp S | X$, permet d'identifier les résultats contrefactuels car il implique que $\mathcal{L}(Y_s | X = x) = \mathcal{L}(Y | S = s, X = x)$.

Enfin, comme contribution originale, nous superposons ces deux modèles causaux pour fournir une analyse mathématique des similitudes et des différences entre les contrefactuels de résultats potentiels et les contrefactuels structurels en démontrant que - contrairement à ce que la littérature courante suggère souvent - $Y_s \neq Y_{S=s}$ en général. Plus précisément, ces variables contrefactuels diffèrent si S code une caractéristique non manipulable qui a un impact sur les covariables X , comme le sexe et la race, mais coïncident si S est un traitement qui peut être alloué de manière expérimentale, comme un médicament. Cela signifie de manière critique que les deux approches causales génèrent des énoncés contrefactuels différents dans les problèmes d'équité classiques où S code typiquement le sexe ou la race, d'où la nécessité de comprendre la distinction. L'idée est que le cadre de Rubin considère des résultats contrefactuels avec des covariables X fixées, alors que celui de Pearl modifie les covariables en $X_{S=s}$. Nous illustrons les conséquences de cette différence à l'aide d'un exemple concret d'équité.

Application à explicabilité et l'équité

Enfin, nous présentons des applications populaires du raisonnement contrefactuel dans l'intelligence artificielle explicable et l'équité algorithmique, en soulignant son rôle critique pour la construction de systèmes dignes de confiance.

Dans leur travail pionnier, Wachter et al. (2017) proposent d'utiliser les énoncés contrefactuels comme une approche psychologiquement acceptée pour expliquer les règles de décision boîte-noire. Concrètement, iels recherchent des changements minimaux plausibles d'une entrée v tels que la sortie du modèle d'apprentissage automatique h diffère, généralement en résolvant

$$\min_{v' \in \mathcal{V}} c(v, v') + \lambda |h(v) - h(v')|^2,$$

où c est une fonction de coût bien choisie, λ un paramètre suffisamment grand pour que la décision change, et \mathcal{V} l'ensemble des mondes plausibles. Cela génère des déclarations contrefactuelles de la forme “Si l'entrée avait été v' (au lieu de v), la sortie aurait été différente”. Dans un souci de clarification et d'unification, nous montrons comment interfacier ces méthodes d'explication avec les méthodes d'inférence contrefactuelle précédemment détaillées. Ceci peut être réalisé en paramétrant les entrées artificielles v' comme les mondes alternatifs de v obtenus après une action, par exemple une intervention. Cette partie permet notamment de souligner la distinction entre les alter égos contrefactuels (c'est-à-dire les mondes alternatifs si un certain événement s'était produit) et les explications contrefactuelles (c'est-à-dire des exemples antagonistes raffinés).

En ce qui concerne l'équité algorithmique, nous illustrons comment les contrefactuels peuvent fournir des notions d'équité fortes et intuitives. Pour S un *attribut protégé* (par exemple, la race ou le sexe) et X les autres caractéristiques d'entrée d'un prédicteur h , l'équité consiste à rendre $h(X, S)$ indépendant de S , généralement en imposant $h(X, S) \perp S$. Toutefois, comme le note [Dwork et al. \(2012\)](#), cette dite *parité statistique* est une contrainte d'équité de groupe qui ne contrôle pas la discrimination au niveau individuel. C'est là que l'inférence contrefactuelle entre en jeu : en permettant de calculer des entrées alternatives “si l'attribut protégé avait été modifié” (par exemple, d'un homme à une femme), elle permet d'exiger l'égalité de traitement entre un individu et ses *alter égos contrefactuels*. Cette idée est au cœur de la définition de *l'équité contrefactuelle*, proposée par [Kusner et al. \(2017\)](#), qui s'appuie sur le formalisme de Pearl pour exiger

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

pour toute observation $\{X = x, S = s\}$ et tout $s' \neq s$. Le principal inconvénient de cette approche réside dans sa dépendance à l'égard d'un modèle causal structural : bien qu'attrayants en théorie, ces modèles sont inconnus en pratique et peuvent difficilement être déduits des données, ce qui rend les méthodes fondées sur la causalité irréalisables dans la plupart des tâches du monde réel. Cette limitation a motivé les travaux présentés dans le chapitre suivant, où nous introduisons des modèles contrefactuels non causaux, basés sur du transport, qui fournissent des notions réalisables d'équité individuelle, plus fortes que les contraintes d'équité de groupe.

Partie I, Chapitre 2: Modèles contrefactuels basés sur du transport

Les applications en explicabilité et en équité examinées dans le chapitre précédent montrent que les techniques les plus couramment utilisées pour le calcul des alter égos contrefactuels (c'est-à-dire des mondes alternatifs) dans l'apprentissage automatique sont le principe du monde alternatif le plus proche et la modélisation causale de Pearl. La première approche est simple, mais néglige la corrélation entre les caractéristiques, ce qui conduit à des contrefactuels *non fiables*, hors distribution, tels que “Si Bob avait été une femme, elle aurait mesuré 190 cm”. La seconde prend rigoureusement en compte toutes les dépendances entre les variables, mais nécessite un modèle causal structural inconnu ou trop difficile à déduire dans la pratique, ce qui la rend *infaisable*, sauf pour des cas jouets. Ce chapitre, basé sur [\(De Lara et al. 2021a\)](#), se concentre sur une troisième voie en interprétant le raisonnement contrefactuel comme un problème de déplacement de masse d'une distribution de probabilité à une autre.

Il fait suite aux travaux de [Black et al. \(2020\)](#) qui ont été les premiers à suggérer l'utilisation du *transport optimal* ([Villani, 2003, 2008](#)) pour concevoir des contrefactuels réalisables et réalistes.

Transport de masse et transport optimal

Tout d'abord, introduisons quelques connaissances de base sur le transport de masse et le transport optimal. Nous appelons *transport de masse* le problème général de l'appariement de deux distributions de probabilités P et Q sur $\mathcal{X} \subseteq \mathbb{R}^d$. Cela revient à sélectionner un couplage dans l'ensemble $\Pi(P, Q)$ de distributions de probabilités jointes ayant respectivement P et Q comme première et deuxième marginales. Un couplage peut être considéré comme un appariement aléatoire, faisant correspondre chaque instance de P à d'éventuelles alter égos dans Q avec des pondérations de probabilité. On dit qu'il est *déterministe* si chaque instance de P est associée à une unique instance de Q . Dans ce cas, le couplage se concentre sur le graphe d'une (unique P -presque partout) application déterministe $T : \mathcal{X} \rightarrow \mathcal{X}$ *poussant* P vers Q , c'est-à-dire $Q(E) = P(T^{-1}(E))$ pour chaque ensemble borélien $E \subseteq \Omega$. Cette propriété, dénotée par $T_{\#}P = Q$, signifie que si une variable aléatoire X suit la distribution P , alors son image $T(X)$ suit la distribution Q .

La théorie du transport optimal est devenue l'outil le plus populaire pour construire de tels couplages lorsqu'il n'existe aucun choix canonique. Elle remonte à [Monge \(1781\)](#) qui a défini les *applications de transport optimales* comme des fonctions transformant P en Q avec un minimum d'effort selon une fonction de coût positive $c : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$. Formellement, ces applications résolvent

$$\min_{T \in \mathcal{T}(P, Q)} \int_{\mathcal{X}} c(x, T(x)) dP(x),$$

où $\mathcal{T}(P, Q)$ est l'ensemble des applications mesurables poussant P vers Q . En général, cependant, une telle correspondance déterministe entre les distributions de probabilité peut ne pas exister, en particulier si P et Q ne sont pas absolument continues par rapport à la mesure Lebesgue. Cette limitation motive ce que l'on appelle la relaxation *Kantorovich* de la formulation de Monge du transport optimal ([Kantorovich and Rubinshtein, 1958](#)) :

$$\min_{\pi \in \Pi(P, Q)} \int_{\mathcal{X}^2} c(x, x') d\pi(x, x').$$

Ce problème concerne des couplages aléatoires au lieu d'applications déterministes et admet toujours des solutions appelées *plans de transport optimaux*.

Raisonnement contrefactuel par transport de masse

Le raisonnement contrefactuel répond à des questions de la forme : "Si S avait été égal à s' au lieu de s , quelle aurait été la valeur de X ?" Le cadre structurel de Pearl répond à cette question par le biais du calcul "do" : pour un monde de référence donné $\{X = x, S = s\}$, il renvoie une distribution d'alter égos contrefactuels définie par $\mathcal{L}(X_{S=s'} | X = x, S = s)$.

De façon différente, [Black et al. \(2020\)](#) ont approximé une application de transport optimale $T_{(s'|s)}$ de $\mu_s := \mathcal{L}(X | S = s)$ vers $\mu_{s'} := \mathcal{L}(X | S = s')$ et ont généré l'alter égo contrefactuel si S avait été égal à s' d'une instance factuelle x du groupe s par $T_{(s'|s)}(x)$. Bien

que cela n'ait pas été abordé dans leur article, cette idée peut être naturellement généralisée aux plans de transport optimaux $\pi_{\langle s'|s \rangle} \in \Pi(\mu_s, \mu_{s'})$ plutôt que des applications. Dans l'ensemble, cette approche non causale basée sur l'observation présente trois intérêts cruciaux. Premièrement, elle garantit que les contrefactuels générés sont dans la distribution des données, et donc réalistes. Deuxièmement, elle bénéficie d'un répertoire croissant de schémas numériques efficaces pour estimer de tels couplages à partir de données (Peyré and Cuturi, 2019), conduisant à des contrefactuels réalisables sur le plan numérique. Troisièmement, contrairement à l'approche structurelle, elle ne requiert pas d'hypothèses sur le processus de génération des données.

Comme première contribution, **nous proposons d'unifier ces deux cadres sous un formalisme commun de transport de masse** fondé sur la remarque suivante : les énoncés contrefactuels “si S avait été égal à s' au lieu de s ” utilisant la modélisation causale de Pearl sont caractérisés par la distribution de probabilité jointe $\pi_{\langle s'|s \rangle}^* := \mathcal{L}((X, X_{S=s'}) \mid S = s)$, faisant correspondre tout monde *factuel* de μ_s à ses contreparties *contrefactuelles* dans $\mu_{\langle s'|s \rangle} := \mathcal{L}(X_{S=s'} \mid S = s)$. De ce point de vue, le raisonnement contrefactuel structurel est, à l'instar du transport optimal, un problème de transport de masse. Cela signifie notamment que Black et al. (2020) imitait implicitement les contrefactuels structurels. En poussant leur idée, **nous définissons les modèles contrefactuels basés sur le transport pour X par rapport à S comme des collections de couplages $\Pi := \{\pi_{\langle s'|s \rangle}\}_{s, s' \in \mathcal{S}}$ tels que pour chaque $s, s' \in \mathcal{S}$, $\pi_{\langle s'|s \rangle} \in \Pi(\mu_s, \mu_{s'})$** . Cette construction générale englobe les couplages définis par le transport optimal, mais aussi le transport optimal régularisé et des techniques de transport de masse potentiellement nouvelles. L'intérêt essentiel du formalisme unifié de transport de masse réside dans la possibilité de remplacer un couplage contrefactuel structurel $\pi_{\langle s'|s \rangle}^*$ par un couplage basé sur le transport $\pi_{\langle s'|s \rangle}$ dans n'importe quel méthode reposant sur des contrefactuels causaux afin de générer une alternative basée sur le transport ; ceci sera illustré plus tard avec l'équité contrefactuelle.

Quand le transport optimal rencontre la causalité

De plus, **le formalisme de transport de masse que nous avons introduit nous permet également de démontrer mathématiquement que le transport optimal récupère les changements causaux dans certains contextes**, ce qui explique pourquoi Black et al. (2020) ont observé empiriquement que les contrefactuels structurels et les contrefactuels par transport optimal étaient presque identiques. Ce résultat est valable sous deux hypothèses typiques du modèle causal.

La première hypothèse exige que les contrefactuels structurels soient générés de manière déterministe par le modèle causal. Cela correspond formellement à $\mathcal{L}(X_{S=s'} \mid X = x, S = s)$ se réduisant à une distribution de Dirac, ce qui se produit en particulier lorsque les variables exogènes sont des termes additifs des équations structurelles. Cela implique que le couplage causal $\pi_{\langle s'|s \rangle}^*$ est déterministe : il peut être identifié à une application $T_{\langle s'|s \rangle}^*$ telle que $T_{\langle s'|s \rangle}^* \# \mu_s = \mu_{\langle s'|s \rangle}$.

La deuxième hypothèse exige que la variable modifiée S joue le même rôle qu'une variable exogène dans le modèle causal structurel. C'est typiquement le cas dans les problèmes d'équité : par exemple, la race est exogène par rapport aux caractéristiques socio-économiques. Cela implique que les alter égos contrefactuelles sont observables dans le sens où $\mu_{\langle s'|s \rangle} = \mu_{s'}$, et

donc que le *modèle contrefactuel structurel* $\Pi^* := \{\pi_{\langle s'|s \rangle}^*\}_{s, s' \in \mathcal{S}}$ est un modèle contrefactuel basé sur du transport.

Sous ces deux hypothèses, le modèle causal structurel induit un modèle contrefactuel déterministe basé sur du transport. On peut alors naturellement se demander s'il peut être récupéré en résolvant un problème de transport optimal déterministe. C'est précisément ce que notre théorème principal assure pour le coût quadratique $c(x, x') := \|x - x'\|^2$: si $\mathcal{L}(X)$ est absolument continu par rapport à la mesure de Lebesgue et a un moment d'ordre 2 fini, alors

$$\pi_{\langle s'|s \rangle}^* = \arg \min_{\pi \in \Pi(\mu_s, \mu_{s'})} \int \|x - x'\|^2 d\pi(x, x')$$

si et seulement si $T_{\langle s'|s \rangle}^*$ est le gradient d'une fonction convexe. Cette condition (qui dit probablement quelque chose aux personnes familières avec la théorie du transport optimal) est notamment satisfaite lorsque les équations causales du modèle pour X sont linéaires additives, c'est-à-dire de la forme

$$X \stackrel{\mathbb{P}\text{-a.s.}}{=} MX + wS + U_X,$$

où $M \in \mathbb{R}^{d \times d}$ et $w \in \mathbb{R}$ sont des paramètres déterministes. L'intérêt de ce théorème est double : il explique théoriquement les observations empiriques du [Black et al. \(2020\)](#) ; il justifie le fait que le transport optimal fonctionne comme une alternative décente - non causale - au raisonnement contrefactuel structurel dans les scénarios d'équité typiques.

Application à l'équité

Cette analyse motive l'utilisation de modèles contrefactuels basés sur le transport à la place des modèles contrefactuels structurels pour dériver de nouvelles notions d'équité, à la fois fines et réalisables. Nous notons en particulier que le critère d'équité contrefactuelle introduit dans le chapitre précédent peut être écrit comme suit : pour tout $s, s' \in \mathcal{S}$ et $\pi_{\langle s'|s \rangle}^*$ -presque tous les (x, x') ,

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

Le remplacement des couplages causaux par des couplages basés sur le transport à partir d'un modèle Π conduit à une nouvelle définition de l'équité que nous appelons l'équité contrefactuelle Π . Bien qu'elle remplace la causalité par des corrélations acceptables, elle préserve la propriété de garantir l'équité au niveau individuel, est plus forte que la parité statistique et ne nécessite pas d'hypothèses sur le processus de génération des données.

Ensuite, en adaptant les travaux de [Russell et al. \(2017\)](#), **nous nous attaquons au problème de l'apprentissage d'un prédicteur Π -contrefactuellement juste.** Cela revient à résoudre

$$\min_{\theta \in \Theta} \mathbb{E}[\ell(h_\theta(X, S), Y)] + \lambda \sum_{s \in \mathcal{S}} \mathbb{P}(S = s) \sum_{s' \neq s} \mathbb{E} \left[|h_\theta(X_{s'}, s') - h_\theta(X_s, s)|^2 \right],$$

où Y est la variable cible, ℓ une fonction de perte assurant la fidélité aux données, $\{h_\theta\}_{\theta \in \Theta}$ une classe paramétrique de prédicteurs, et $\mathcal{L}((X_s, X_{s'})) = \pi_{\langle s'|s \rangle}$ pour chaque $s, s' \in \mathcal{S}$. Théoriquement, nous prouvons dans le cas du transport optimal quadratique que la solution

empirique converge presque sûrement vers la solution réelle lorsque n , la taille de l'échantillon, tend vers l'infini ; expérimentalement, nous démontrons les performances de notre estimateur sur des jeux de données réelles. En résumé, notre contribution étend l'arsenal de l'apprentissage équitable à des critères plus forts que la simple condition d'équité de groupe en relâchant la causalité par des méthodes de transport.

Partie II, Chapitre 3: Estimation par GAN de fonctions de transport optimal Lipschitz

Dans (Black et al., 2020), les auteurs ont utilisé des contrefactuels par transport optimal pour évaluer l'équité des règles de décision boîte noire, tandis que nous les avons utilisés pour apprendre un prédicteur équitable dans le chapitre précédent. Quelle que soit la situation, la mise en œuvre d'un modèle contrefactuel basé sur le transport nécessite la construction de plans ou d'applications de transport à partir de données. Plus les méthodes d'estimation disponibles seront variées, plus nous pourrions traiter de problèmes. Ce chapitre, basé sur (González-Sanz et al., 2022), contribue à la littérature grandissante sur l'approximation des applications de transport optimales. Il combine l'approche GAN (*generative adversarial networks*) de (Black et al., 2020) avec des réseaux de neurones Lipschitz (Anil et al., 2019; Tanielian and Biau, 2021) pour concevoir un nouvel estimateur neuronal avec des garanties statistiques prouvables.

Plus précisément, nous considérons P et Q , deux mesures absolument continues par rapport à Lebesgue telles que l'application de transport optimale T_0 de P à Q pour le coût quadratique est lisse, en particulier Lipschitz. Ensuite, sur la base des distributions empiriques P_n et Q_n tirées respectivement de P et Q , nous approximations T_0 par la solution du problème de GAN

$$\inf_{G \in \mathcal{G}_n} \left\{ \int \|I - G\|^2 dP_n + \lambda_n \sup_{D \in \mathcal{D}_n} \int D(d(G\#P_n) - dQ_n) \right\},$$

où \mathcal{D}_n est une classe de discriminants 1-Lipschitz fournissant un proxy pour la formulation duale de la distance de Wasserstein-1 (à la façon Wasserstein-GAN de Arjovsky et al. (2017)), et \mathcal{G}_n est une classe de générateurs Lipschitz paramétrant l'espace des applications réalisables. Le paramètre positif λ_n régit le compromis entre la minimisation du coût de transport quadratique, favorisant l'objectif du problème de Monge, et la minimisation de la distance entre les distributions générée et cible, promouvant la contrainte de poussée.

Notre objectif est double : premièrement, nous visons à concevoir une application de transport optimale expressive bénéficiant d'une architecture neuronale pour généraliser efficacement à de nouvelles observations hors échantillon ; deuxièmement, nous visons à fournir des garanties statistiques, ce qui a été négligé par la plupart des articles connexes (Leygonie et al., 2019; Black et al., 2020; Makkuva et al., 2020; Korotin et al., 2021; Huang et al., 2021).

Réseaux de neurones GroupSort multivariées

Le problème décrit ci-dessus implique des réseaux neuronaux Lipschitz, que ce soit pour les discriminateurs ou les générateurs. Dans ce chapitre, nous tirons parti des fonctions

d'activation *GroupSort* récemment introduites pour imposer la contrainte Lipschitz (Anil et al., 2019; Tanielian and Biau, 2021), qui se sont avérées produire des estimations plus fines des fonctions 1-Lipschitz que les méthodes antérieures telles que dans (Arjovsky et al., 2017; Gulrajani et al., 2017). Par définition, la fonction d'activation *GroupSort* de taille de regroupement $k \geq 2$ divise l'entrée de pré-activation en groupes de taille k , puis trie chaque groupe par ordre décroissant. Cette opération est 1-Lipschitz, préserve la norme du gradient et est homogène (Anil et al., 2019). Un réseau de neurones *GroupSort* est un réseau neuronal feed-forward dans lequel toutes les fonctions d'activation (à l'exception de la première couche) sont une fonction d'activation *GroupSort* de taille de groupement fixe k ; dans ce travail, nous nous limitons à $k = 2$. Sous l'hypothèse de compacité des poids, ces réseaux sont des fonctions 1-Lipschitz.

La spécificité du problème de Wasserstein-GAN amélioré auquel nous nous attaquons réside dans la paramétrisation du générateur *multivarié* en tant qu'application 1-Lipschitz ; cela a déjà été abordé pour le discriminateur *univarié* dans (Anil et al., 2019; Biau et al., 2021). Il faut pour cela étendre aux réseaux neuronaux *GroupSort* multivariés les résultats théoriques de (Tanielian and Biau, 2021), qui traitent uniquement du pouvoir d'approximation des réseaux neuronaux *GroupSort* univariés.

Estimateur par GAN

Sur la base de ce résultat d'approximation, nous définissons les générateurs \mathcal{G}_n et les discriminateurs \mathcal{D}_n comme des réseaux neuronaux *GroupSort* avec des tailles bien choisies dépendant de n . Pour garantir la convergence statistique, la séquence des générateurs réalisables $\{\mathcal{G}_n\}_{n \in \mathbb{N}}$ doit remplir \mathcal{G} suffisamment rapidement, tandis que la séquence des poids de régularisation $\{\lambda_n\}_{n \in \mathbb{N}}$ doit tendre vers l'infini afin d'imposer la condition de poussée à la limite. **Nous démontrons sous ces hypothèses la convergence uniforme presque sûrement de $\{\mathcal{G}_n\}_{n \in \mathbb{N}}$ vers T_0 .** La preuve repose sur les propriétés de compacité relative des fonctions de Lipschitz ainsi que sur la régularité de T_0 grâce à des résultats de (Hütter and Rigollet, 2021). Nous illustrons ensuite les performances de notre approximation sur des jeux de données synthétiques.

Partie II, Chapitre 4: Appariement difféomorphique par divergences de Sinkhorn

La théorie du transport optimal n'est pas la seule façon de faire correspondre des distributions de probabilité ; bien qu'elle ait rarement été appliquée à des tâches d'apprentissage automatique, l'appariement difféomorphique bénéficie d'une théorie et d'algorithmes bien établis. Ce cadre inspiré de la mécanique des fluides recherche un champ de vitesse optimal dans l'espace ambiant pour transférer une distribution à l'autre. Ce chapitre, basé sur (De Lara et al., 2023), élargit la boîte à outils du transport de masse en étudiant théoriquement et expérimentalement l'appariement difféomorphique piloté par les divergences de Sinkhorn, des métriques de transport optimal entropique.

Transport de masse difféomorphique

Sous des hypothèses de régularité, un champ de vitesse $v_t(x) \in \mathbb{R}^d$ des variables $t \in [0, 1]$ et $x \in \mathbb{R}^d$ génère une famille de difféomorphismes $(\phi_t^v)_{t \in [0,1]}$ par l'équation d'écoulement :

$$\phi_t^v(x) := x + \int_0^t v_s(\phi_s^v(x)) \, ds.$$

Le transport de masse difféomorphique cherche des champs de vitesse v tels que ϕ_1^v fait correspondre une distribution d'entrée α à une distribution cible β . Formellement, pour Λ une fonction de perte positive entre les mesures, cela revient à résoudre

$$\min_{v \in L_V^2} J_\lambda(v) \text{ avec } J_\lambda(v) := \lambda(\phi_{1\#}^v \alpha, \beta) + \lambda \|v\|_{L_V^2}^2,$$

où L_V^2 est un espace de Hilbert de champs de vecteurs v avec une énergie cinétique finie $\|v\|_{L_V^2}^2$, et la régularisation quantifiée par $\lambda > 0$ garantit que le problème est bien posé.

Comme l'explique [Feydy et al. \(2017\)](#), la non-convexité de ce programme d'optimisation rend crucial le choix de la fonction de perte afin d'éviter les minima locaux peu profonds, alors que les carrés de *maximum mean discrepancies*, qui sont les métriques standard, en produisent beaucoup. Ce chapitre remédie à ce problème en proposant une alternative fondée.

Transport optimal entropique

Pour $\varepsilon > 0$, le coût global de transport entropique pour la fonction de coût local $c : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}_+$ entre les mesures de probabilité α et β sur $\mathcal{X} \subseteq \mathbb{R}^d$ est défini comme suit

$$\mathcal{T}_{c,\varepsilon}(\alpha, \beta) := \min_{\pi \in \Pi(\alpha, \beta)} \int_{\mathcal{X} \times \mathcal{X}} c(x, y) d\pi(x, y) + \varepsilon \text{KL}(\pi | \alpha \otimes \beta),$$

où $\text{KL}(\mu | \nu)$ désigne la divergence *Kullback-Leibler* entre les mesures de probabilité μ et ν donnée par $\int \log \left(\frac{d\mu}{d\nu}(z) \right) d\mu(z)$ si $\mu \ll \nu$, et $+\infty$ sinon. Dans [\(Feydy et al., 2017\)](#), les auteurs utilisent ce coût pour la fonction de perte Λ , car elle bénéficie de schémas numériques rapides qui allègent la charge de calcul du transport optimal non régularisé [\(Cuturi, 2013\)](#), et engendre moins de minima locaux que les *maximum mean discrepancies*. Cependant, ce choix souffre de ce que l'on appelle le *biais entropique*, c'est-à-dire $\mathcal{T}_{c,\varepsilon}(\alpha, \alpha) \neq 0$ en général, ce qui en fait une fonction de perte peu fiable. C'est pourquoi nous proposons d'utiliser les divergences de Sinkhorn, *unbiased* entropic transportation costs, définies comme suit,

$$S_{c,\varepsilon}(\alpha, \beta) := \mathcal{T}_{c,\varepsilon}(\alpha, \beta) - \frac{1}{2} \mathcal{T}_{c,\varepsilon}(\alpha, \alpha) - \frac{1}{2} \mathcal{T}_{c,\varepsilon}(\beta, \beta).$$

Sous des hypothèses de régularité sur c , α et β , cette divergence satisfait plusieurs propriétés souhaitables pour une fonction de perte : elle est non négative, convexe en ses deux mesures d'entrée, et métrise la convergence en loi [\(Feydy et al., 2019\)](#).

Convergence statistique

Comme à chaque fois, nous n'avons pas accès aux véritables mesures α et β dans la pratique, mais à des versions empiriques α_n et β_n , ce qui soulève la question de la convergence des minimiseurs empiriques $\{v^n\}_{n \in \mathbb{N}}$ vers un minimiseur de J_λ lorsque la taille de l'échantillon n tend vers l'infini. Si Λ est le carré d'une *maximum mean discrepancy*, alors selon [Glaunes et al. \(2004\)](#) il existe un minimiseur de J_λ noté v^* tel que à l'extraction près d'une sous-suite

$$\|v^n - v^*\|_{L_V^2} \xrightarrow[n \rightarrow \infty]{a.s.} 0.$$

Nous démontrons des garanties statistiques plus fortes lorsque Λ est une divergence de Sinkhorn en spécifiant en plus le taux de convergence. Sous des hypothèses de régularité sur l'espace L_V^2 , le coût c et les mesures α et β , il existe une constante $A > 0$ telle que

$$\mathbb{E} [|J_\lambda(v^n) - J_\lambda(v^*)|] \leq \frac{A}{\sqrt{n}}.$$

La preuve s'appuie sur la régularité de la formulation duale du transport optimal entropique, ainsi que sur des processus empiriques.

Expériences numériques

Nous concluons ce chapitre en comparant l'utilisation des divergences de Sinkhorn pour l'appariement difféomorphique aux *maximum mean discrepancies* et aux coûts de transport biaisés en fonction de différents paramètres et procédures de résolution. Cela montre les avantages de ce choix par rapport aux options utilisées antérieurement.

Comme indiqué, la plupart de ces chapitres sont basés sur des documents de recherche que j'ai rédigés au cours des trois dernières années. Bien que je les aie homogénéisés pour ce manuscrit, ils conservent leurs structures originales et sont donc autonomes.

Part I

Counterfactual reasoning: from causality to mass transportation

Chapter 1

Counterfactuals, explainability, fairness

Counterfactual thinking is thinking “contrary to the facts”, to envision alternative realities. This modality fuelled many frameworks in fair and explainable artificial intelligence, as it addresses queries such as “Would have she gotten the position had she been a man?”. However, the diversity of scientific methods and terminologies tagged “counterfactual” rendered nebulous the scope of this reasoning. For the sake of clarification, we firstly go back to the fundamental of counterfactuals: the possible-world account of Lewis. We explain how Lewis’ perspective amounts to postulating a mechanism to generate alternative worlds through interventions on factual worlds; we refer to such a mechanism as a counterfactual model. Then, we review state-of-the-art counterfactual methodologies through this reading grid. In particular, we study the distinct implications of evaluating counterfactual statements through Rubin’s causal model or Pearl’s causal model. Moreover, we motivate the use of counterfactuals in fairness and explainability, and show how counterfactual explanations and counterfactual fairness (two acclaimed counterfactual frameworks) relate to counterfactual models. This first chapter notably serves to fix ideas before proposing new counterfactual models in the second chapter.

1.1 Introduction

As mentioned at the beginning of the manuscript, the goal of this thesis is to leverage a form of counterfactual reasoning based on measure transportation theories to implement strong notions of machine-learning fairness, and explain algorithmic decision rules. But what does *counterfactual* mean? According to the Cambridge dictionary,¹ the word counterfactual can be either an adjective describing something “thinking about what did not happen but could have happened, or relating to this kind of thinking”, or a noun defined as “something such as piece of writing or an argument that considers what would have been the result if events had happened in a different way to how they actually happened”. From a propositional logic perspective, a *counterfactual* is a statement or assertion of the form “had event A occurred then event B would have occurred”, such as “had it rained today, I would have stayed at

¹<https://dictionary.cambridge.org/fr/dictionnaire/anglais/counterfactual>

home” (Lewis, 1973b). We commonly use counterfactuals to explain why event an B did not occurred. Critically, these statements relate to alternative realities, and are hence not verifiable by mere observations.

The word counterfactual has also been extensively employed in scientific research, such as in mathematics and computer science, to describe a variety of objects, techniques or frameworks meant to understand causation. Let us mention three of the top-cited scientific papers having the word “counterfactual” in their title: “Inference on Counterfactual Distributions” (Chernozhukov et al., 2013), “Counterfactual Fairness” (Kusner et al., 2017), “Counterfactual Visual Explanations” (Goyal et al., 2019). The first one comes from the field of econometrics and addresses the statistical estimation of *counterfactual* effects in regression models, which describe how an outcome variable of interest is affected by a change of some covariates or a change in the relationship between the outcome and the covariates. The second one introduces a causal notion of machine-learning fairness, requiring equal treatment between any individual (e.g., a woman) and their *counterfactual* counterparts “had they belonged to a different group” (e.g., men) generated through Pearl’s causal modeling (Pearl, 2009). The third one focuses on explaining neural-network-based image classifiers by uncovering *counterfactual* visual explanations: the pixels of an image which made the classifier predicting a certain class instead of another. Observe that, while all these counterfactual notions relate in some sense to alternative states of things—making relevant the use of the adjective counterfactual—they differ in the type of objects they refer to, the mathematical frameworks they are based-on, and the tasks they are meant to be applied for, leading to misunderstanding across research fields. Should you mention that you work on “counterfactuals” at a machine-learning conference, people would understand something different depending on their backgrounds.

The goal of this chapter is specifically to clarify the similarities and differences between the various state-of-the-art counterfactual frameworks that blossomed in academic research over the past years, in order to make precise the role of the *counterfactual models* at the heart of this thesis. We firstly present the main frameworks for counterfactual inference in light of Lewis’s pioneering work (Lewis, 1973b), and then review the usage of counterfactual techniques in *explainability* and *fairness*: the two facets of trustworthy artificial intelligence. In passing, we introduce several key notations and definitions that will be used throughout the manuscript.

1.2 The possible-world account

Counterfactual inference addresses the complex problem of evaluating the truth of counterfactual statements. Imagine you have a headache but have no medicine left. Hence, you complain: “Had I taken medicine I would have felt better”. While such a statement might seem intuitively true, there is no evidence that the headache would have stopped for sure, since we cannot observe the alternative reality where you took the pill. Lewis (1973b) proposed a general formal framework to verify counterfactuals, which relies on the notion of *world*.

1.2.1 Worlds and counterfactuals

A world is a conjunction of statements characterizing the state of things. For the sake of illustration, imagine a bank deciding to grant a loan or not to individuals on the basis of their profiles (containing typically socioeconomic indicators). In this setting, a world is identified to the conjunction of features defining the profile; each world describes a possible application scenario. Throughout, we represent a world by a vector $v := (v_1, \dots, v_p) \in \mathbb{R}^p$ where each $i \in \{1, \dots, p\}$ represents a feature variable (e.g., yearly salary in dollars) and v_i indicates its value (e.g., $5 \cdot 10^4$). For simplicity, we assume that the bank follows a deterministic rule $h : \mathbb{R}^p \rightarrow \{0, 1\}$ delivering 1 if the loan is granted to the application and 0 otherwise. Now, consider a specific world v describing someone who earns 50K per year and whose application got rejected by the bank: this will be our *world of reference*. The bank could provide the following counterfactual to explain its decision:

Had your income been 100K per year, you would have been granted the loan. (1.1)

This statement relates the *antecedent* **A** “Earning 100K per year” to the *consequent* **B** “The loan is granted”. It is formally written as $v \models \mathbf{A} \square \rightarrow \mathbf{B}$. Note that counterfactuals are *local*: a same antecedent can entail different outcomes depending on the world of reference. As such it is crucial to keep track of v in the notation. Counterfactuals are also *partial*, because they do not specify the minimally sufficient conditions for the outcome to be true. For example, it could be that earning 80K per year is enough to get the loan in v . To verify a counterfactual, Lewis proposes to use a general notion of similarity between worlds, the only requirement being that a world be closest to itself. Then, $v \models \mathbf{A} \square \rightarrow \mathbf{B}$ is true just in case **B** is true for every possible worlds v' satisfying **A** which are “similar” to v . This implies the following abstract procedure to evaluate a given counterfactual $v \models \mathbf{A} \square \rightarrow \mathbf{B}$: (1) using the notion of similarity and the antecedent **A**, compute the alternative worlds v' satisfying **A** similar to v ; (2) check whether **B** holds in every v' . In what follows we detail each of these two steps.

1.2.2 Antecedents and interventions

The key step of counterfactual inference is the computation of alternative worlds; the evaluation of the outcome generally amounts to a simple check. To properly explain this process, we need to precise how to formally write the possible antecedents of counterfactual statements. We emphasize that the formalism introduced in this subsection does not come from the mainstream literature; it is an original proposition meant to provide a unified framework for the various counterfactual approaches presented in this chapter.

For simplicity, we assume that an antecedent **A** can always be framed in terms of explicit feature modification. Let $\mathcal{I} \subseteq \{1, \dots, p\}$ enumerate the feature variables that can be modified. In particular, \mathcal{I} is not necessarily the entire $\{1, \dots, p\}$ because depending on the task one could decide a number of features to be immutable, as we detail later in the manuscript. Now, define the set

$$\mathcal{A} := \{(I, \tilde{v}_I) \in 2^{\mathcal{I}} \times \mathcal{V}_I\}$$

where $I \subseteq \mathcal{I}$ denotes a subset of actionable features while \mathcal{V}_i describes the possible values of the i th feature. Any couple $a = (I, \tilde{v}_I) \in \mathcal{A}$ represents the action of setting each feature $i \in I$ to the chosen value \tilde{v}_i and is referred as an *intervention*. An intervention mathematically

encodes an antecedent. For example, if $i = 1$ is the salary, then the antecedent A “Earning 100K per year” corresponds to the intervention $a = (\{1\}, 10^5)$.

Computing the counterparts of the world of reference v had A occurred requires to translate the effect of intervention a as a mathematical operation on v . The design of this operation is directly related to the notion of similarity. For example, in the canonical *ceteris paribus* approach, being similar signifies keeping all features identical except the ones concerned by A. Concretely, this means that verifying (1.1) following this principle amounts to checking whether the loan is granted for the application obtained by changing the yearly income in the original application to 100K while keeping all other variables equal. Said differently, it amounts to checking if $h(v') = 1$ for the v' defined by $v'_1 = 10^5$ and $v'_i = v_i$ for every $i \neq 1$.

Choosing a different notion of similarity would lead to different alternative worlds v' , hence to possibly different outcomes and counterfactuals. In particular, even though the *ceteris-paribus* viewpoint seems legitimate, it is often considered unfaithful as it neglects correlations between features: a different income could indicate a different occupation or could induce a different home location, making the *ceteris-paribus* counterparts of v possibly outside the set of plausible worlds. Therefore, it can be relevant to change the notion of similarity into one capturing the dependencies between A and the other features. The approach consisting in including more changes than those explicitly mentioned is sometimes referred as *mutatis mutandis*, which means *after changing what should be changed*.

We emphasize that this partly where the variety of counterfactual frameworks stems from: there are as many ways of thinking counterfactually as there are ways of computing alternative worlds, or equivalently, as there are notions of similarity. Defining the notion of similarity amounts to designing a model specifying the downstream effect of any intervention $a = (I, \tilde{v}_I) \in \mathcal{A}$ onto the features not specified by I . In all generality, we propose to define such a model by a collection of transformations $\{T_a\}_{a \in \mathcal{A}}$, such that $T_a(v)$ yields the counterparts of any world v under any action a . As an example, the *ceteris paribus* account of counterfactuals induces the following deterministic transformation for $a = (I, \tilde{v}_I)$:

$$T_a(v) := \begin{cases} \tilde{v}_i & \text{if } i \in I, \\ v_i & \text{otherwise.} \end{cases}$$

There are many degrees of freedom for specifying the model: $T_a(v)$ could be always single-valued, implying that alternative worlds are deterministically determined by v and a ,² or could otherwise render probability distributions to include unmodeled sources of randomness. For instance, we could consider that were Bob a woman, there is a 40% chance that she would be writer and a 60% chance that she would be an astronaut. The literature also considers *sets* of alternative worlds, which can be represented by a uniform distribution. Note that we particularly appreciate the distribution-based mathematical representation of alternative worlds as it encompasses all situations: a Dirac distribution for a single counterpart; a uniform distribution for multiple counterparts; a nonuniform distribution for truly random counterparts.

In the rest of this thesis, we call any model of interventions $\{T_a\}_{a \in \mathcal{A}}$ for some set of actions \mathcal{A} a *counterfactual model*, and refer to the instances in $T_a(v)$ as the *counterfactual*

²In logic, this corresponds to the *conditional excluded middle* assumption of Stalnaker (1980).

counterparts of v had a occurred. All in all, the choice of the counterfactual model is always arbitrary to some extent. It must reflect the implicit meaning of the counterfactual statements (e.g., *ceteris paribus* or not, deterministic or not), but could also be guided by feasibility reasons (this aspect will be discussed later).

1.2.3 Consequents and evaluations

Now that we formalized the computation of alternative worlds, we can turn to the evaluation of a counterfactual $v \models A \square \rightarrow B$. Assume that A corresponds to an intervention $a \in \mathcal{A}$ whose effect is given by a transformation T_a . Recall that $v \models A \square \rightarrow B$ is true just in case B is true in all worlds v' similar to v and satisfying A . For this definition to be operational, B must be associated to some proposition $\mathcal{P}(\cdot)$ specifiable on any world. In our running example, we would define $\mathcal{P}(v')$ as “ $h(v') = 1$ ”, namely “the loan is granted”. Summing everything up, $v \models A \square \rightarrow B$ is true just in case $\mathcal{P}(T_a(v))$ is true.

Moreover, for this definition to be nonambiguous, we must take into account the fact that both $T_a(v)$ and \mathcal{P} can be random (i.e., defined as probability distributions or random variables). Conceptually, this does not raise particular issues: whenever one of the two objects is nondeterministic, we can either evaluate the truth of the counterfactual statement for a certain realization, or give a probability of truth. In the set-valued scenario, $v \models A \square \rightarrow B$ is true just in case $\mathcal{P}(v')$ is true for all $v' \in T_a(v)$. Recall that $T_a(v)$ can be defined as a uniform distribution in such cases, which then requires the probability of truth of \mathcal{P} to be 100% for the counterfactual to be declared true.

1.2.4 Examples of similarity metrics

Let us conclude this section by a short illustration. As mentioned, the verification process of counterfactuals fundamentally rests on the choice of a similarity metric only requiring a world to be closest to itself. This makes distances on the space of worlds \mathcal{V} natural candidates to compute alternative worlds.

Formally, let $D : \mathcal{V} \times \mathcal{V} \rightarrow \mathbb{R}_+$ be a distance. Then—for this specific notion of similarity—the counterfactual counterparts of v had $a = (I, (\tilde{v}_i)_{i \in I})$ occurred are given by

$$T_a(v) := \arg \min_{v' \in \mathcal{V}} D(v, v') \text{ s.t. } v'_i = \tilde{v}_i \text{ for all } i \in I. \quad (1.2)$$

This definition returns the D -closest worlds satisfying A . Note that if $\mathcal{V} = \mathbb{R}^p$, it provides *ceteris paribus* counterparts. In general however, the features I live in different, possibly nonindependent ranges $(\mathcal{V}_i)_{i \in I}$. Alternatively, we could relax the notion of similarity by defining for some radius $r > 0$,

$$T_a(v) := \{v' \in \mathcal{V} \mid D(v, v') \leq r \text{ and } v'_i = \tilde{v}_i \text{ for all } i \in I\}. \quad (1.3)$$

Here, the alternative A -worlds are given by the intersection of the worlds accessible from v (defined by a ball centered at v) and the worlds satisfying A .

Unfortunately, this simple distance-based viewpoint generally shares the drawbacks of the *ceteris paribus* approach: by neglecting the latent probability distribution over worlds, it cannot faithfully capture dependencies between variables. For example, let Bob be a

man whose height is 190cm and whose weight is 85kg. A distance-based model renders true the counterfactual “Had Bob been a woman, she would be 190cm tall and weigh 85kg”, as such women exist (even though they are rarer). According to intuition, such counterfactuals are false and rightly so because they are not representative of the underlying statistical distributions. The sex of a person influences their physical features, as such we would expect Bob’s height and weight to be modified after intervention. In Sections [1.3](#) and [1.4](#), we present other perspectives leveraging more sophisticated notions of similarity.

To sum-up, counterfactual reasoning amounts to mapping a world of reference to some alternative worlds satisfying a desired property, and then checking whether an outcome of interest holds in these worlds. In order to properly understand a counterfactual-based methodology, we recommend to systematically answer the following questions:

1. What is the space of considered worlds?
2. What is the studied antecedent, or equivalently the features to modify?
3. What is the proposition defining the outcome of interest?
4. How are the interventions defined, or equivalently the notion of similarity?

In the following, we unify several counterfactual frameworks from the machine-learning-related literature through this common reading grid.

1.3 Structural causal modeling

As aforementioned, the *ceteris paribus* and distance-based approaches of counterfactual reasoning generally fail to describe realistic alternative worlds, as they implicitly assume the features to be independent. This limitation motivated the use of Pearl’s causal modeling ([Pearl, 2009](#)) to take into account the fact that variables are not independently manipulable.

1.3.1 Causal model

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, 2010b](#)). This section recalls the basic theory on this modeling, borrowing the rigorous mathematical framework recently proposed by [Bongers et al. \(2021\)](#).

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

Definition 1.3.1: Structural causal model

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

1. $U : \Omega \rightarrow \mathcal{U}$ 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 $i \in \mathcal{I}$ there exist two subsets of indices $\text{Endo}(i) \subseteq \mathcal{I}$ and $\text{Exo}(i) \subseteq \mathcal{J}$, respectively called the *endogenous* and *exogenous parents* of i , such that G_i is from $\mathcal{V}_{\text{Endo}(i)} \times \mathcal{U}_{\text{Exo}(i)}$ to \mathcal{V}_i .

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

$$V_i \stackrel{\mathbb{P}\text{-a.s.}}{=} G_i(V_{\text{Endo}(i)}, U_{\text{Exo}(i)}). \quad (1.4)$$

The collection of equations defined by (1.4) and characterized by G and U are called the *structural equations*. By identifying G to a measurable vector function $G : \mathcal{V} \times \mathcal{U} \rightarrow \mathcal{V}$, we compactly write that V is a solution of \mathcal{M} if

$$V \stackrel{\mathbb{P}\text{-a.s.}}{=} G(V, U).$$

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 $k \in \text{Endo}(l) \cup \text{Exo}(l)$ (we say that k is a parent of l). Slightly abusing notation, 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 1.1). Also, similarly to Bongers et al. (2021), we will use in practice nondisjoint 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 1.3.1: SCM and solution

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

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

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

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

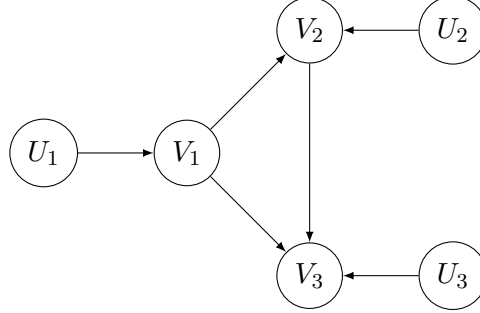


Figure 1.1: Example of causal graph

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 manuscript the common assumption that the considered models are *acyclic*, meaning that their graphs do not contain any cycles:

Assumption (A): Acyclicity

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

Acyclicity entails *unique solvability* of the SCM, in the sense that Equation (1.4) admits a unique solution up to \mathbb{P} -negligible sets (Bongers et al., 2021, Proposition 3.4). We will abusively refer to such a solution as *the* solution of the SCM. Notably, the SCM from Example 1.3.1 satisfies (A). Besides, the absence of cycles allows for clearer interpretation of the causal dependencies.

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 nonintervened variables.

1.3.2 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 1.3.2: Do-intervention

Let $\mathcal{M} = \langle U, G \rangle$ be an SCM, $I \subset \mathcal{I}$ a subset of endogenous variables, and $v_I \in \mathcal{V}_I$ a value. The action $\text{do}(I, v_I)$ defines the modified model $\mathcal{M}_{\text{do}(I, v_I)} = \langle U, \tilde{G} \rangle$ 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 1.3.2 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 principle that acting upon endogenous phenomena does not affect exogenous ones. Provided it is solvable, the modified model $\mathcal{M}_{\text{do}(I, v_I)}$ 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 $\mathcal{M}_{\text{do}(I, v_I)}$ also satisfies (A). Going further, if V is the solution of an acyclic \mathcal{M} , we can nonambiguously define (up to \mathbb{P} -negligible sets) its intervened counterpart $V_{\text{do}(I, v_I)}$ solution to $\mathcal{M}_{\text{do}(I, v_I)}$. 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 $\text{do}(V_I = v_I)$ for the operation $\text{do}(I, v_I)$, 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 1.3.2: Intervening is not conditioning in general

Let $\mathcal{M} := \langle U, G \rangle$ be the SCM from Example 1.3.1 and consider the do-intervention $\text{do}(V_2 = 0)$. 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 1.2 represents the graph after surgery. The modified structural equations on a solution $\tilde{V} := (\tilde{V}_1, \tilde{V}_2, \tilde{V}_3)$ are

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

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

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

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.

Working directly with Definition 1.3.2 to express variables after intervention can be burdensome as it requires to solve a modified causal model. Throughout this manuscript, we rely on the next proposition which provides a general expression of the solution before and after intervention.

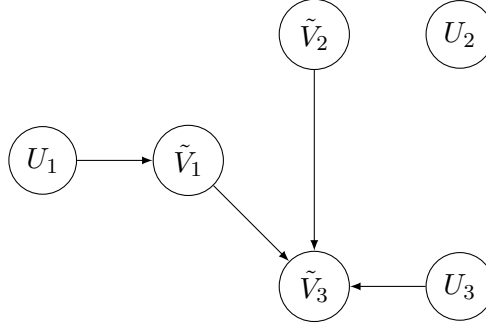


Figure 1.2: Intervened counterpart of Figure 1.1 after $\text{do}(V_2 = 0)$

Proposition 1.3.1: Do-calculus on variables

Let $\mathcal{M} = \langle U, G \rangle$ be an SCM satisfying **(A)** with solution V , and consider a partition $I \sqcup J = \mathcal{I}$. There exists a deterministic measurable function F_J such that

$$V_J \stackrel{\mathbb{P}\text{-a.s.}}{=} F_J(V_{\text{Endo}(J) \setminus J}, U_{\text{Exo}(J)}).$$

Moreover, for any intervention $\text{do}(V_I = v_I)$ the solution \tilde{V} of $\mathcal{M}_{V_I = v_I}$ verifies

$$\tilde{V}_J \stackrel{\mathbb{P}\text{-a.s.}}{=} F_J(v_{\text{Endo}(J) \setminus J}, U_{\text{Exo}(J)}),$$

$$\tilde{V}_I \stackrel{\mathbb{P}\text{-a.s.}}{=} v_I.$$

Importantly, this is the same deterministic function F_J that generates V_J and its intervened counterpart \tilde{V}_J , the only change being the assignment $V_I = v_I$. Note that in our notations, we will often artificially extend the input variables of F_J to write $V_J \stackrel{\mathbb{P}\text{-a.s.}}{=} F_J(V_I, U_{\text{Exo}(J)})$ and $\tilde{V}_J \stackrel{\mathbb{P}\text{-a.s.}}{=} F_J(v_I, U_{\text{Exo}(J)})$.

1.3.3 Counterfactual inference

Do-calculus provides a natural framework to address the computation of alternative worlds. Consider for instance the event $\{V = v\}$ where V is the solution to an SCM $\mathcal{M} := \langle U, G \rangle$ satisfying **(A)**, and set $v'_I \neq v_I$. We aim at answering the counterfactual question: *had V_I been equal to v'_I (instead of v_I), what would have been the value of V ?* Pearl answers this question using the so-called *three-step procedure*:

1. **Abduction:** Deduce the posterior distribution of U given the world of reference $\{V = v\}$;
2. **Action:** Carry out do-calculus on \mathcal{M} to obtain the intervened causal mechanism $G_{V_I = v'_I}$ of $\mathcal{M}_{V_I = v'_I}$;
3. **Prediction:** Pass the posterior distribution $\mathcal{L}(U \mid V = v)$ through $G_{V_I = v'_I}$ to generate the distribution $\mathcal{L}(V_{V_I = v'_I} \mid V_I = v_I)$ of alternative worlds.

Note that when $\mathcal{L}(U \mid V = v)$ is not degenerate, this approach does not define a single alternative world given an intervention and a factual world, but instead a set of alternative worlds with probability weights. This entails that in general the induced counterfactual semantic is stochastic: a counterfactual statement evaluated through this model is not true or false, but has a certain probability of occurrence. We refer to the values taken by $\mathcal{L}(V_{V_I=v'_I} \mid V = v)$ as the *structural counterfactual counterparts* of $V = v$ under the intervention $\text{do}(V_I = v'_I)$. By construction, a factual instance v and its counterfactual counterparts share the same possible values of exogenous variables, distributed as $\mathcal{L}(U \mid V = v)$, but differ from the causal mechanism generating them. This implicitly defines a notion of similarity, such that “being similar” means “can be generated by a same U value”.

Bridging this definition to the formalism introduced in Section [1.2](#), we define for any intervention $a = (I, v'_I)$ and world of reference v the transformation $T_a(v)$ as the result of $\text{do}(V_I = v'_I)$ on V conditional to $\{V = v\}$. Concretely, this leads to

$$T_{\text{do}(V_I=v'_I)}(v) = \mathcal{L}(V_{V_I=v'_I} \mid V = v).$$

This model of interventions suggests that the true intrinsic features of a world, the “all other things” that must be kept equal, are its U -value. Considering an alternative world after intervention amounts to process the same U -value through a different generative mechanism. Thus, cross-world counterfactual statements according to Pearl compare a same object but through two different prisms, a same entity but within two different realities.

1.3.4 Discussion

Structural causal models are powerful tools; they enable to analyze the causal effects of any studied variables, furnishing a reliable framework for counterfactual inference. However, they also raise natural concern about their applicability. Beyond toy examples generating artificial data through a fabricated mechanism, we generally do not know the SCMs governing observed phenomena. *Causal discovery* (also known as *structure learning*) refers to the field of research addressing this issue by investigating methods to derive causal models from observable data. In this section, we discuss the challenges of this initiative and the limitations of structural causal modeling for counterfactual reasoning.

Postulating the model

It might be tempting to postulate the causal relationships on the basis of intuitions and prior knowledge. While this could be reasonable for a small number of variables regulated by established mechanisms such as laws of physics, this is not realistic for typical machine-learning problems dealing with a high-number of features and possibly complex structural relations. Assuming a fully-specified causal model 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. Moreover, this is not practical since a causal model must be designed and tested for each possible dataset.

Causal discovery

As mentioned, 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 without restrictive assumptions, with an exponential worst-case complexity with respect to the number of nodes (Chickering et al., 2004; Scutari et al., 2019). In addition, 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 later exemplified in this manuscript.

Apart from these computational challenges, the fundamental constraint of causal discovery is *causal uncertainty*: there exist several causal models corresponding to a same data distribution (see Bongers et al., 2021, Example 4.2). Therefore, it cannot be tested through observations only whether the adjusted model is the “true” one. This makes the modeling inherently uncertain not only for the relationships between variables G , but also for the law of the exogenous variables U . A general principle to keep in mind is that causal inference will always require untestable assumptions about the world’s functioning. Making strong hypothesis (e.g., postulating the graph or the form of the equations) tends to facilitate the estimation of the model and can exclude spurious possibilities, but increases the model-approximation error. Overall, causal discovery demands a subtle mix of plausible assumptions and efficient estimation procedures.

Existence of 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 $V_{I=v_I}$ 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.

Structural causal models are powerful tools; but they are expensive in assumptions and computations, which drastically limits their applicability on real-world tasks. In the following section, we present a lighter (but less flexible) framework for counterfactual inference.

1.4 The potential-outcome framework

The potential-outcome framework, also known as *Neyman-Rubin causal modeling* (Rubin, 1974), was designed to understand the causal effect of a treatment onto an outcome of interest, for instance when one aims at assessing the contribution of a drug to recovering from some disease. It has become the most widely used framework for causal inference in social sciences and medicine due to its intuitive formalism, mathematically much lighter than structural causal models.

1.4.1 Model and motivation

Let $S : \Omega \rightarrow \{0, 1\}$ represent a binary *treatment status*, typically such that $S = 0$ indicates the absence of treatment and $S = 1$ indicates a treatment. More generally, it can encode any distinction between some groups (e.g., men and women). Assuming no interference between units, this framework postulates *potential outcomes* $Y_0 : \Omega \rightarrow \mathbb{R}$ and $Y_1 : \Omega \rightarrow \mathbb{R}$ under each treatment status. These potential outcomes as well as the treatment may depend on some covariates $X : \Omega \rightarrow \mathbb{R}^d$ such as the weight, the height, or historical data. Critically, we cannot observe simultaneously Y_0 and Y_1 for a single unit: a problem referred as the *fundamental problem of causal inference* (Holland, 1986). We only have access to the realized *outcome variable* $Y : \Omega \rightarrow \mathbb{R}$ which is supposed to be *consistent* with (Y_0, Y_1) , that is satisfying $Y = (1 - S) \cdot Y_0 + S \cdot Y_1$. Concretely, if $S(\omega) = 1$ for some $\omega \in \Omega$, then $Y(\omega) = Y_1(\omega)$, and $Y_0(\omega)$ becomes unidentifiable by mere observations. In this case, $Y_1(\omega)$ is called the *factual outcome* while $Y_0(\omega)$ is called the *counterfactual outcome*.

Understanding the causal relationship between the treatment and the outcome in this framework amounts to estimating the difference $Y_1 - Y_0$ from observed data. In practice, people commonly focus on the *average treatment effect* $\mathbb{E}[Y_1 - Y_0]$ or the *conditional average treatment effect* $\mathbb{E}[Y_1 - Y_0 \mid X = x]$. The main challenge lies in the fact that *correlation is not causation* in general. In particular, the observable quantity $\mathbb{E}[Y \mid S = s]$ does not necessarily coincide with the unobservable quantity $\mathbb{E}[Y_s]$ for $s \in \{0, 1\}$. Typically, if some medical treatment is more likely to be taken by weaker patients, we may observe a lower rate of recovery among the treated group compared to the nontreated group due to the health condition even though the medicine does increase recovery all other things being kept equal: we would observe $\mathbb{E}[Y \mid S = 1] < \mathbb{E}[Y \mid S = 0]$ while $\mathbb{E}[Y_1] > \mathbb{E}[Y_0]$ (a phenomenon referred as *Simpson's paradox*). In this case, the health condition is called a *confounder*: a variable associated with both the distribution of the treatment and the outcome. However, causal inference from observational data is still possible, as detailed next.

1.4.2 Estimation of causal effects

We say that a treatment effect is *identifiable* if it can be expressed with observational quantities only, that is in terms of X , S and Y . Identifiability requires two fundamental assumptions. The first one goes by many names through the literature: *conditional ignorability*, *conditional exchangeability*, *conditional exogeneity*, and *unconfoundedness*. It states that the potential outcomes are independent of the treatment conditional to the covariates, that is $S \perp\!\!\!\perp (Y_0, Y_1) \mid X$. Said differently, it prevents the existence of unmodeled confounders between the treatment and the potential outcomes. Note that this assumption is untestable, as it would

require to observe simultaneously the two potential outcomes. The second key hypothesis is *positivity*, which ensures that all individual can be exposed to both treatment statuses, that is $0 < \mathbb{P}(S = 1 | X) < 1$. It readily follows from conditional ignorability and positivity that $\mathcal{L}(Y | X, S = s)$ coincides with $\mathcal{L}(Y_s | X)$ for $s \in \{0, 1\}$, meaning that observable outcomes have a causal meaning. In the following, we briefly present several methods to identify the average causal effect which all build upon this implication.

Controlled trials

The first method is not a mathematical computation but a physical act. As previously mentioned, causal effects cannot be estimated from the conditional distributions $\mathcal{L}(Y | S = s)$ as long as there exist confounders, variables correlated with both the treatment and the potential outcomes. In a *randomized controlled trial*, a practitioner supervises the treatment allocation (the treatment status of each unit is not predetermined but decided according to some probability law) to rule out confounder. Supposing no confounder, or *ignorability*, mathematically corresponds to $S \perp\!\!\!\perp (Y_0, Y_1)$. It implies that correlation is causation, that is $\mathcal{L}(Y | S = s) = \mathcal{L}(Y_s)$, making the average treatment effect trivially identifiable. Note that in general, the potential outcome framework only assumes *conditional* ignorability, namely $S \perp\!\!\!\perp (Y_0, Y_1) | X$. While ignorability guarantees the complete absence of confounders, modeled or not, conditional ignorability ensures that all possible confounders are included in the covariates X .

The gold standard for causal inference is a *completely* (or *fully*) randomized experiment, which assigns the treatment independently to all experimental conditions. This is canonically achieved by tossing a same coin to decide each unit's status. This design blinds the treatment status to any experimental conditions (be they observed or not), thereby ensuring $(X, Y_0, Y_1) \perp\!\!\!\perp S$ without additional assumptions (Zhang and Zhao, 2023). In case the treatment assignment is controlled but not fully randomized, practitioners can at most ensure $S \perp\!\!\!\perp X$ for a set of *observed* covariates. This condition, referred as *covariates balance*, entails ignorability if conditional ignorability and positivity hold: note that $S \perp\!\!\!\perp X$ along with $S \perp\!\!\!\perp (Y_0, Y_1) | X$ and $0 < \mathbb{P}(S = 1 | X) < 1$ implies that $S \perp\!\!\!\perp (Y_0, Y_1)$. Intuitively, making the weaker and the stronger patients evenly likely to receive treatment disentangles the contribution of the health condition from the one of the treatment onto the recovery, rendering causal the association between the observed outcomes and the treatment.

To sum-up, in the case of experimental studies, practitioners can ensure identifiability by randomizing their protocol.³ Nevertheless, causal inference outside the setting of randomized controlled trials is still needed for several reasons. Depending on the treatment, enforcing randomization can easily become immoral: thoroughly testing the effect of parental violence onto mental health would require asking a large number of parents to hit their children. Additionally, organizing large-scale experiments costs time and money. This is why in practice we frequently rely on observational studies for which randomization does not hold. Next, we detail purely mathematical techniques to reach identifiability in this context.

³We point out that randomized controlled trials are not the panacea though, as on the contrary to observational studies they suffer from eligibility biases. This has motivated a growing research on combining observational and randomized data, as reviewed by Colnet et al. (2020).

Re-weighting

A first approach to simulate the properties of randomization in nonrandomized observational studies is to manipulate the dataset to artificially enforce covariate balance. The idea is to modify the weights of the samples to render the treatment independent to the covariates. For example, *inverse probability weighting* leverages the propensity score (Rosenbaum and Rubin, 1983), defined as

$$e(x) := \mathbb{P}(S = 1 \mid X = x),$$

to balance covariates between groups. Intuitively, dividing the weight of every unit by how likely they were to have their actual treatment status makes every unit evenly likely to be treated. Mathematically, this corresponds to a change of probability: we modify the underlying \mathbb{P} to define a new probability \mathbb{Q} on Ω through

$$\frac{d\mathbb{Q}}{d\mathbb{P}} := \frac{1}{2} \left(\frac{S}{e(X)} + \frac{1-S}{1-e(X)} \right).$$

The probability \mathbb{Q} is well defined under *positivity*, that is $0 < e(X) < 1$, and satisfies $S \perp\!\!\!\perp X$ plus $\mathbb{Q}(S = 1) = \mathbb{Q}(S = 0) = 1/2$. Said differently, the pseudo population obtained by re-weighting looks as if the treatment was randomly allocated by tossing an unbiased coin. This enables to compute averaged potential outcomes under from observational quantities. Let $\mathbb{E}_{\mathbb{Q}}$ denote the expectation under \mathbb{Q} . Assuming conditional ignorability (under \mathbb{P}) we have

$$\begin{aligned} \mathbb{E}_{\mathbb{Q}}[Y \mid S = 1] &= \frac{\mathbb{E}_{\mathbb{Q}}[Y \cdot S]}{\mathbb{Q}(S = 1)} = 2 \cdot \frac{1}{2} \cdot \mathbb{E} \left[Y \cdot S \cdot \left(\frac{S}{e(X)} + \frac{1-S}{1-e(X)} \right) \right] \\ &= \mathbb{E} \left[\frac{S}{e(X)} \cdot Y \right] = \mathbb{E} \left[\mathbb{E} \left[\frac{S}{e(X)} \cdot Y \mid X \right] \right] = \mathbb{E} \left[\frac{\mathbb{E}[S \cdot Y \mid X]}{e(X)} \right] \\ &= \mathbb{E} \left[\frac{\mathbb{E}[S \cdot Y_1 \mid X]}{e(X)} \right] = \mathbb{E} \left[\frac{e(X) \cdot \mathbb{E}_{\mathbb{P}}[Y_1 \mid X]}{e(X)} \right] = \mathbb{E} [\mathbb{E}[Y_1 \mid X]] \\ &= \mathbb{E}[Y_1]. \end{aligned}$$

Similarly, $\mathbb{E}_{\mathbb{Q}}[Y \mid S = 0] = \mathbb{E} \left[\frac{1-S}{1-e(X)} \cdot Y \right] = \mathbb{E}[Y_0]$, leading to

$$\mathbb{E}[Y_1 - Y_0] = \mathbb{E} \left[\frac{S}{e(X)} \cdot Y \right] - \mathbb{E} \left[\frac{1-S}{1-e(X)} \cdot Y \right].$$

This shows that the average treatment effect under \mathbb{P} is identifiable, as it can be computed from the observed X , S and Y . However, “there ain’t no such thing as a free lunch”; there is a price to pay for such a trick. Notably, this approach requires estimating a functional form of e from observable data, which is generally achieved through logistic regression. As shown by Smith and Todd (2005) and Kang and Schafer (2007), estimands of average treatment effects based on the propensity score are extremely sensitive to the score-estimation quality. Moreover, the inverse probability weighting suffers from instabilities when the propensity score approaches 0 or 1 since the weights tend to infinity. Several papers tried to mitigate these practical issues by constructing refined estimates of the propensity score (Imai and Ratkovic, 2014) or the treatment effects (Funk et al., 2011) that are robust to misspecification. We do not detail these methods here for the sake of simplicity.

Stratification

Causal inference from observational data can also be achieved via *stratification*. Recall that conditional ignorability enables to identify causal effects through observable quantities *within levels of X*, leading to:

$$\begin{aligned}\mathbb{E}[Y_1 - Y_0 | X] &= \mathbb{E}[Y_1 | S = 1, X] - \mathbb{E}[Y_0 | S = 0, X] \\ &= \mathbb{E}[Y | S = 1, X] - \mathbb{E}[Y | S = 0, X].\end{aligned}$$

Then, integrating on the covariates gives the so-called *adjustment formula* on the average treatment effect:

$$\mathbb{E}[Y_1 - Y_0] = \mathbb{E}[\mathbb{E}[Y | S = 1, X] - \mathbb{E}[Y | S = 0, X]].$$

In practice, stratification follows the adjustment formula by: (1) splitting the units into groups with equal values of confounders X , (2) estimating the conditional effect $\mathbb{E}[Y | S = 1, X] - \mathbb{E}[Y | S = 0, X]$ for each group, (3) marginalizing over each group. This method enabled to solve Simpson's paradox in the notorious 1973 Berkeley's admission process (Bickel et al., 1975). At first glance, it seemed that female applicants were disadvantaged compared to male applicants since the overall proportion of accepted women was inferior. However, figures showed that their rate of acceptance conditional to the course choice was sometimes higher than men. This contradiction can be clarified by adjusting the sex (playing the role of the treatment) on the course choice (playing the role of the confounder). The analysis after stratification did not evidence any disparate success between male and female applicants, meaning that there was a positive correlation between being a female and being rejected but no direct cause-to-effect relationship.

Note that stratification is better tailored to confounders with discrete values. Another limitation comes from the fact that some adjustment groups may contain very few or even no samples, leading to an inconsistent estimation of the conditional averaged effect. A way to mitigate this issue is to stratify on the propensity score $e(X)$. The propensity score works as a confounding variable: adjusting on it makes the treatment independent to the covariates. Additionally, it generally leads to more units within adjustment strata, hence to better estimates of the conditional effects. Nevertheless, this would suffer from the propensity-score estimation issue we mentioned in the paragraph on re-weighting.

Matching

The last commonly used strategy is called *matching*. For illustration, let $\{(x^{(i)}, s^{(i)}, y^{(i)})\}_{i=1}^n$ be a dataset of n i.i.d. *units* sampled from (X, S, Y) . Each unit i represents a treatment experiment where either $y_0^{(i)}$ or $y_1^{(i)}$ is observed. If we could fill the missing entries represented by question marks in Table 1.1, we could compute treatment effects. An intuitive solution for this missing-data-imputation problem would be to look for the units with the most similar covariates belonging to a different treatment group, then to collect the outcomes of these alternative units, and finally to fill the missing entry with a combination of these outcomes. *Matching* techniques follow this principle by leveraging two ingredients: a metric quantifying the closeness between two units; an algorithm specifying how to match units between different treatment groups on the basis of the similarity score.

unit i	treatment S	covariates X	outcome Y_0	outcome Y_1
1	0	$x^{(1)}$	$y_0^{(1)}$?
2	1	$x^{(2)}$?	$y_1^{(2)}$
...
n	1	$x^{(n)}$?	$y_1^{(n)}$

Table 1.1: Illustration of the missing-data-imputation problem. It amounts to fill the question marks, which represent unobserved quantities.

Although it may seem natural to choose distances on the space of covariates for the similarity metric, it often leads to inefficient matchings when the number of covariates becomes high. In practice, most metrics quantify the similarity not between the covariates but between the propensity scores to reduce the dimension of the problem. Concretely, two units i and j from different treatment groups are deemed similar if $e(x^{(i)}) \approx e(x^{(j)})$, that is if they are evenly likely to be treated. The *exact* matching algorithm corresponds to the *ceteris paribus approach*: it matches a unit i to the units j in the opposite group with identical propensity scores. Obviously, most of the units are likely to remain unmatched. In contrast, the *nearest-neighbor* matching algorithm leverages a distance to match unit i to the units j with the closest propensity scores (up to a threshold referred as the caliper). We also mention the *kernel* matching algorithm, constructing the counterfactual outcome of i using a linear combination of every $y^{(j)}$ for $j \neq i$ weighted by the kernel value between $x^{(i)}$ and $x^{(j)}$. We emphasize that while the propensity score may seem magical, as it reduces a multi-dimensional matching problem on the covariates X to a uni-dimensional matching problem on the score $e(X)$, the curse of the dimension has simply been transferred into the estimation (through regression) of $e(X)$ from X . Obviously, the matching approach requires a high-quality estimate of the propensity score to be accurate.

A second key idea behind the use of the propensity score comes from the fact that the distribution of covariates is the same for treated and nontreated units with equal propensity score, that is $S \perp\!\!\!\perp X \mid e(X)$. As such, aligning according to this score tends to balance covariates between groups among the matched units. Naturally, unmatched units, for which there is no sufficiently similar unit with an alternative treatment status, are discarded from the computation of the average treatment effect.

As a final remark, note that while both stratification and re-weighting enable to recover causal effect even outside the randomized-control-trial setting, they are limited to the inference of *averaged* effects. In contrast, matching techniques also estimate the counterfactual outcome at the *unit* level.

1.4.3 Identification of the potential outcomes

At this stage, one may naturally wonder whether structural causal modeling and the potential-outcome framework produce the same counterfactuals. We believe the literature on causal inference to be strongly misleading on this matter. A plethora of scientific books and papers interchangeably use Pearl's do-notation and the potential-outcome subscript notation to write outcomes under interventions, suggesting that the corresponding definitions of

counterfactuals are identical and differ only from theirs perspectives (Colnet et al., 2020; Imbens, 2020; Makhlouf et al., 2020; Neal, 2020). To justify this, they often refer to Pearl, who claimed that “the two frameworks can be used interchangeably and symbiotically”.⁴ However, to our knowledge, works on equivalences between the two causal frameworks miss the point of actually proving whether counterfactual outcomes are equal across models or implicitly address specific cases. Notably, both (Pearl, 2009, Chapter 7) and (Richardson and Robins, 2013)—acclaimed references on unifications of causal frameworks—consider ex nihilo the *mathematical* equivalence between the two notations. In this section, we provide a mathematical analysis of the similarities and differences between approaches by precisely identifying the law of the potential outcomes.

Models

Let $N, d, p \geq 1$ be integers, and define three random variables $S : \Omega \rightarrow \mathcal{S} := \{0, 1, \dots, N\}$, $X : \Omega \rightarrow \mathbb{R}^d$, and $Y : \Omega \rightarrow \mathbb{R}^p$. In order to study the consistency of counterfactual statements between the Neyman-Rubin causal model and Pearl’s causal model, we consider a superimposed construction where (S, X, Y) is concurrently governed by a potential-outcome model and a structural causal model.

On the one hand, we assume that Y is the outcome of interest, S the treatment status, and X some covariates in a potential-outcome framework. This amounts to postulating N random vectors $(Y_s)_{s \in \mathcal{S}}$ satisfying the *consistency rule*:

$$Y \stackrel{\mathbb{P}\text{-a.s.}}{=} \sum_{s \in \mathcal{S}} \mathbf{1}_{\{S=s\}} Y_s.$$

Note that we address a more general framework than before, considering a nonbinary treatment and a multivariate outcome. In this setting, the two fundamental assumptions for causal inference can be written as:

Assumption (P): Positivity

$$0 < \mathbb{P}(S = s \mid X) < 1, \text{ for all } s \in \mathcal{S}.$$

Assumption (CI): Conditional ignorability

$$(Y_s)_{s \in \mathcal{S}} \perp\!\!\!\perp S \mid X.$$

On the other hand, we assume that these variables are generated by a latent, unknown structural causal model: the random vector $V := (S, X, Y)$ is the solution to an acyclical SCM $\mathcal{M} = \langle U, G \rangle$ where U_S, U_X and U_Y denote the exogenous parents of respectively S, X , and Y . Moreover, we suppose that \mathcal{M} satisfies:

⁴<http://causality.cs.ucla.edu/blog/index.php/2012/12/03/judea-pearl-on-potential-outcomes/>

Assumption (O): Outcome

$$U_Y \perp\!\!\!\perp (U_S, U_X) \text{ and } Y_{\text{Endo}(S)} = Y_{\text{Endo}(X)} = \emptyset.$$

This assumption captures two major characteristics of the modeling. The random-noise condition states that all potential confounders between S and Y are included in X ; this may seem weaker than **(CI)**, hence redundant, but expressing it as a separate assumption on \mathcal{M} will simplify the demonstrations. The graphical condition formally defines the variable Y as the *outcome*; it changes in response to X and S but not the contrary. Through Proposition **1.3.1**, this item permits to write

$$\begin{aligned} X &\stackrel{\mathbb{P}\text{-a.s.}}{=} F_X(S, U_X), \\ S &\stackrel{\mathbb{P}\text{-a.s.}}{=} F_S(X, U_S), \\ Y &\stackrel{\mathbb{P}\text{-a.s.}}{=} F_Y(S, X, U_Y), \end{aligned}$$

where F_X, F_S and F_Y are deterministic measurable functions derived from G . The artificial cycle in these formulas (i.e., X and S are both functions of each other) merely serves to consider all configurations of causal links between S and X (see Figure **1.3**); strictly, \mathcal{M} satisfies **(A)**. Proposition **1.3.1** also enables to define for every $s \in \mathcal{S}$ the post-intervention outcome under $\text{do}(S = s)$ as

$$Y_{S=s} \stackrel{\mathbb{P}\text{-a.s.}}{=} F_Y(s, X_{S=s}, U_Y),$$

where the altered covariates are $X_{S=s} \stackrel{\mathbb{P}\text{-a.s.}}{=} F_X(s, U_X)$.

Critically, the two considered causal models differ fundamentally in their constructions of counterfactual outcomes. As noted by **Pearl (2010a)**, the potential outcomes $(Y_s)_{s \in \mathcal{S}}$ are “undefined *primitives*” of the Neyman-Rubin causal model, not related to any formal of measurable quantities, while the intervened outcomes $(Y_{S=s})_{s \in \mathcal{S}}$ are “*derivatives*” of the structural causal model by application of do-calculus. However, Pearl uses the same notation for both constructions. Are they truly equivalent in the sense that $Y_s \stackrel{\mathbb{P}\text{-a.s.}}{=} Y_{S=s}$, or at least $Y_s \stackrel{\mathcal{L}}{=} Y_{S=s}$? This is what the address next.

Identification

Let us start with a crucial remark: the potential outcomes are ill-defined in the sense that there is no unique choice of $(Y_s)_{s \in \mathcal{S}}$ satisfying the consistency rule. More precisely, while necessarily $Y_s \stackrel{\mathbb{P}\text{-a.s.}}{=} Y$ on $\{S = s\}$ for $s \in \mathcal{S}$, there is no restriction on Y_s over $\Omega \setminus \{S = s\}$; it could take any value on without violating the consistency rule. Consequently, it is mathematically impossible to associate Y_s —well-identifiable on the event $\{S = s\}$ only—to $Y_{S=s}$ —defined (almost) everywhere through the structural causal model \mathcal{M} . Without further assumptions, we only have identification of the *observed outcomes*, namely $Y_s \stackrel{\mathbb{P}\text{-a.s.}}{=} Y_{S=s}$ on $\{S = s\}$, as a direct consequence of the proposition below.

Proposition 1.4.1: Consistency of structural counterfactual outcomes

Let (S, X, Y) be the solution of an SCM \mathcal{M} satisfying **(A)** and **(O)**. Then, $(Y_{S=s})_{s \in \mathcal{S}}$ verifies the consistency rule,

$$Y \stackrel{\mathbb{P}\text{-a.s.}}{=} \sum_{s \in \mathcal{S}} \mathbf{1}_{\{S=s\}} Y_{S=s}.$$

Nevertheless, although the two fundamental assumptions of causal inference cannot fully solve the identification of the potential outcomes, as they do not constrain the variables almost surely, they permit to identify the *law* of the potential outcomes in terms of observable quantities: they entail that $\mathcal{L}(Y_s | X = x) = \mathcal{L}(Y | X = x, S = s)$ for any $s \in \mathcal{S}$. As our main mathematical result, we propose a different kind of identification under the same assumptions. The theorem below identifies the law of the potential outcomes through the latent SCM \mathcal{M} , thereby enabling us to compare $(Y_s)_{s \in \mathcal{S}}$ with $(Y_{S=s})_{s \in \mathcal{S}}$.

Theorem 1.4.1: Identification of potential outcomes

Let $(Y_s)_{s \in \mathcal{S}}$ be random variables such that

$$Y \stackrel{\mathbb{P}\text{-a.s.}}{=} \sum_{s \in \mathcal{S}} \mathbf{1}_{\{S=s\}} Y_s,$$

and suppose that S, X , and $(Y_s)_{s \in \mathcal{S}}$ verify **(P)** along with **(CI)**. Additionally, assume that $V := (S, X, Y)$ is the solution to an SCM $\mathcal{M} = \langle U, G \rangle$ satisfying **(A)** and **(O)**. This notably entails that there exists a deterministic function F_Y such that $Y \stackrel{\mathbb{P}\text{-a.s.}}{=} F_Y(S, X, U_Y)$. Then,

$$\mathcal{L}((S, X, (Y_s)_{s \in \mathcal{S}})) = \mathcal{L}((S, X, (F_Y(s, x, U_Y))_{s \in \mathcal{S}})).$$

This means in particular that under the assumptions of Theorem **1.4.1**, we concurrently have

$$\begin{aligned} (Y_s)_{s \in \mathcal{S}} &\stackrel{\mathcal{L}}{=} (F_Y(s, X, U_Y))_{s \in \mathcal{S}}, \\ (Y_{S=s})_{s \in \mathcal{S}} &\stackrel{\mathbb{P}\text{-a.s.}}{=} (F_Y(s, X_{S=s}, U_Y))_{s \in \mathcal{S}}. \end{aligned}$$

Therefore, $(Y_s)_{s \in \mathcal{S}}$ and $(Y_{S=s})_{s \in \mathcal{S}}$ are not necessarily equal in law since $\mathcal{L}(X) \neq \mathcal{L}(X_{S=s})$ in general (we provide an example a few paragraphs below). This critically signifies that counterfactual inference is not equivalent between frameworks since the joint probability distributions $\mathcal{L}((S, X, (Y_{S=s})_{s \in \mathcal{S}}))$ and $\mathcal{L}((S, X, (Y_s)_{s \in \mathcal{S}}))$ are not always equal. As consequence, in contrast to what many papers suggest, *the potential-outcome subscript notation and the do notation are not equivalent*, be they as subscripts for random variables or law-dependent quantities.

Actually, we do have equality in law if X is not altered by do-interventions on S , that is if S is not a parent of X in \mathcal{M} . Notably, this configuration encompasses various typical

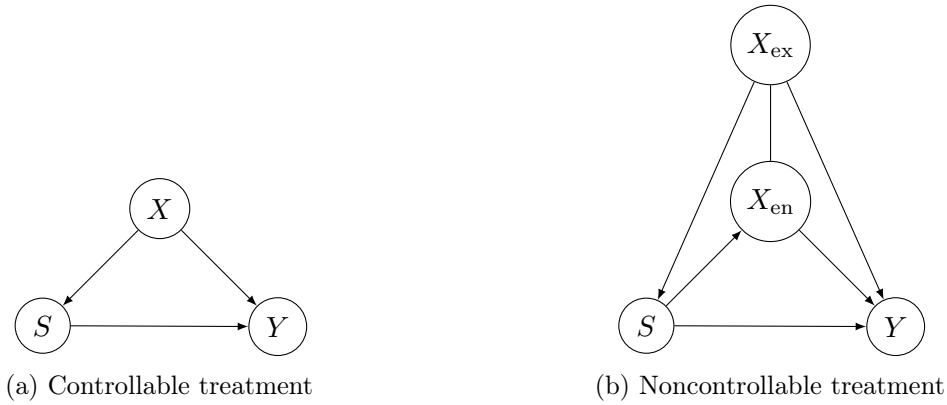


Figure 1.3: The two possible configurations of the treatment. X_{ex} denotes the parents of S in X while X_{en} are the remaining covariates. Exogenous variables are not represented. In (a), S does not cause X ; in (b), S can impact X .

causal-inference scenarios: in clinical trials, the covariates X may influence the treatment allocation S but never the contrary. In such cases, the covariates X play the same role as exogenous variables in \mathcal{M} , as illustrated in Figure [1.3a](#). Thus, both the Neyman-Rubin causal model and Pearl's causal model produce the same counterfactuals in common situations. However, people also rely on causal inference outside the scope of clinical trials, in settings where the treatment cannot be manipulated and impacts the covariates (see Figure [1.3b](#)). In particular, S engenders X in emblematic causal problems such as the aforementioned Berkeley's admission paradox where S represented the sex. In these cases, confusion between the two causal approaches can lead to misleading results: the Neyman-Rubin causal model considers counterfactual outcomes at fixed X , whereas Pearl's causal model alters the covariates into $X_{S=s}$. Note that what distinguishes the two configurations is the nature of the so-called treatment: if the treatment is an intrinsic feature of units, such as race or sex, then structural counterfactuals and potential-outcomes counterfactuals are not equal; if the treatment can be assigned *a posteriori* to units, then the two notions of counterfactuals coincide in law under the two fundamental assumptions of causal inference. Another way of reconciliation comes from remarking that potential-outcome counterfactuals can be derived from the latent SCM by intervening on *both* S and X instead of S only. According to the rules of do-calculus, $Y_{S=s, X=x} \stackrel{\mathbb{P}^{-a.s.}}{=} F_Y(s, x, U_Y)$ whose law coincides with $\mathcal{L}(Y_s | X = x, S = s)$. This signifies that potential outcomes implicitly intervene on X to keep it at a certain level, hence why covariates are often referred as *control* variables.

Before turning to a concrete illustration, let us connect Rubin's account for counterfactuals to the formalism of Section [1.2](#). Importantly, the potential-outcome framework only authorizes interventions of the treatment status S , hence why it only considers actions of the form $a = (S, (s))$. For such an action a and a world of reference $v := (s, x, y)$ the transformation $T_{s'}(s, x, y)$ can be defined as the random variable $(s', X, Y_{s'})$ conditional to $\{S = s, X = x, Y = y\}$. Concretely, this leads to

$$T_{s'}(s, x, y) := \mathcal{L}((s', X, Y_{s'}) | S = s, X = x, Y = y).$$

This formulation clearly shows that cross-world counterfactual statements in this framework compare two different worlds sharing the same observed features X but differing in S . In contrast, Pearl's account induced the transformation

$$T_{S=s'}(s, x, y) := \mathcal{L}((s', X_{S=s'}, Y_{S=s'}) \mid S = s, X = x, Y = y).$$

Observe that potential-outcome counterfactuals are *ceteris paribus* counterfactuals with respect to the covariates, whereas structural counterfactuals are *mutatis mutandis* counterfactuals. We emphasize that both definitions are perfectly legitimate, but convey distinct meanings. Therefore, they should not be employed for the same purpose. In the sequel, we illustrate their implications on a concrete case.

Illustration

This example generalizes and circumstantiates the discussion from (Kusner et al., 2017, Appendix S1). The treatment status S indicates the gender, $S = 0$ standing for women and $S = 1$ standing for men; the covariate X quantifies the level of work experience, a higher score encoding a richer experience; the outcome Y evaluates a candidate's application for some position, a better score giving a higher probability of acceptance. Suppose that these three variables S , X and Y are ruled by the following SCM satisfying **(O)**,

$$\begin{aligned} X &\stackrel{\mathbb{P}\text{-a.s.}}{=} \alpha S + U_X, \\ Y &\stackrel{\mathbb{P}\text{-a.s.}}{=} X + \beta S + U_Y, \end{aligned}$$

where α and β are deterministic parameters quantifying the causal influence of S onto respectively X and Y , and U_X represents the hidden merit or effort of an individual. Typically, a positive parameter α describes the societal inequalities leading women to have a lower level of work experience than men with equal merit U_X . Moreover, we suppose that **(P)** holds and set two potential outcomes (Y_0, Y_1) verifying the consistency rule and **(CI)**. We consider the problem of assessing the counterfactual outcome of a woman described by $\{S = 0, X = x, Y = y\}$, *had she been a man*. The potential-outcome approach evaluates the following *treatment effect*:

$$\begin{aligned} \text{TE}_1(0, x, y) &:= \mathbb{E}[Y_1 - Y_0 \mid S = 0, X = x, Y = y] \\ &= \mathbb{E}[Y_1 \mid S = 0, X = x, Y = y] - y \\ &= \mathbb{E}[X + \beta + U_Y \mid S = 0, X = x, Y = y] - y \\ &= x + \beta + \mathbb{E}[U_Y \mid S = 0, X = x, Y = y] - y \\ &= x + \beta + y - x - y \\ &= \beta. \end{aligned}$$

Observe that this first effect completely ignores the dependence of Y on S through X , as it involves only β . This is due to the fact that TE_1 keeps the covariate X fixed, comparing two *distinct* individuals with identical profiles but different genders. In contrast, Pearl's approach

assesses the following structural counterfactual effect,

$$\begin{aligned}
\text{SCE}_1(0, x, y) &:= \mathbb{E}[Y_{S=1} - Y_{S=0} \mid S = 0, X = x, Y = y] \\
&= \mathbb{E}[Y_{S=1} \mid S = 0, X = x, Y = y] - y \\
&= \mathbb{E}[X_{S=1} + \beta + U_Y \mid S = 0, X = x, Y = y] - y \\
&= \mathbb{E}[\alpha S + U_X \mid S = 0, X = x] + \beta + \mathbb{E}[U_Y \mid S = 0, X = x, Y = y] - y \\
&= \alpha + \mathbb{E}[U_X \mid S = 0, X = x] + \beta + y - x - y \\
&= \alpha + x + \beta + y - x - y \\
&= \alpha + \beta.
\end{aligned}$$

Remark that this second effect takes into account the whole path of influence of S onto Y , involving both α and β . This comes from the fact that the SCE_1 fixes the random seed U and not the covariate, comparing a *same* individual in two alternative realities where the gender is switched. Most importantly, $\text{CATE} \neq \text{SCE}$ if $\alpha \neq 0$, and consequently $\mathcal{L}((S, X, Y_{S=0}, Y_{S=1})) \neq \mathcal{L}((S, X, Y_0, Y_1))$.

From a fairness perspective, the treatment effect TE_1 says that if $\beta = 0$, that is if S is not a *direct* cause of Y , then the application process is fair; whether it is unfair towards men or women when $\beta \neq 0$ depends on the sign of β . In contrast, the structural counterfactual effect SCE_1 says that if $\beta = -\alpha$, that is the decision rule Y compensates the discrepancy of work experiences X across genders S , then the application process is fair. Each analysis points out a different notion of fairness: considering the SCE_1 as a fairness criterion suggests that recruiters should correct societal inequalities by preferring women with potentially lower work experience but higher merit whereas relying on the TE_1 suggests it is only explicitly including the gender in the decision-rule pipeline that is unfair.

1.4.4 Discussion and comparison of causal frameworks

We finish our analysis of the Neyman-Rubin causal model by discussing its assumptions, and underlining other divergences with structural causal models.

Firstly, the identification we provided in Theorem [1.4.1](#) as well as the causal inference techniques presented in Section [1.4.2](#) critically require two fundamental assumptions. While conditional ignorability is intuitively plausible in most cases, the positivity assumption raises more issues. It basically states that the distributions $\mathcal{L}(X \mid S = s)$ for $s \in \mathcal{S}$ share the same support, which is violated as soon as the groups represented by S bear unique properties. Consider for example that S encodes the gender, and that the covariates X specifies the position (among other attributes). Positivity would forbid the computation of the counterfactual outcome *had she been a man* of every woman occupying a women-only job. In contrast, a structural causal model allows for matching individuals that would be deemed incomparable in the potential-outcome framework by making comparisons within a latent space of exogenous features.

Secondly, we emphasize that whatever the chosen framework, causal inference will always require untestable assumptions. As aforementioned, a structural causal model is always at least partly a postulate, because a same observational distribution can be generated by different models, precluding identification by mere observations. Analogously, conditional ignorability from the potential-outcome framework cannot be experimentally

verified. Moreover, note that while it rests on arguably lighter assumptions than a fully specified SCM, the potential-outcome framework lacks versatility: it can only intervene on one discrete variable, and always distinguishes invariant covariate features X from variant outcome features Y among endogenous variables.

Thridly, to conclude on a more philosophical note, we point out that considering the causal effect of modifying immutable features such as a person’s sex or race raises important concerns. One cannot manipulate the world to modify such variables as they could for assigning a drug. Therefore, many people consider that such causal effects are ill-defined by nature, claiming that there is “no causation without manipulation” (Holland, 1986). This principle goes against the whole literature on causal fairness, which typically intervenes on sex and race.

In this section, we studied the verification of counterfactual whose key challenge is the modeling of alternative worlds, and detailed two popular frameworks to carry out this crucial inference step. In the next section, we focus on the role of counterfactuals in explainable artificial intelligence (XAI), underlining their importance for building trustworthy algorithms.

1.5 Counterfactual explanations in artificial intelligence

The ability of modern machine-learning algorithms, especially deep neural networks, to learn accurate decision rules on high-dimensional data has made them widely deployed to tackle various real-world problems, sometimes involving critical decisions for high-risk systems as in medicine, transportation, or security. However, the complexity of these models generally prevents from understanding how they arrive at their decisions even when their internal structures are accessible. Moreover, several research papers demonstrated their vulnerability to adversarial attacks (Moosavi-Dezfooli et al., 2016; Huang et al., 2017), underlining how unreliable accuracy was to measure the trustworthiness of artificial intelligence (AI) systems. This issue triggered the emergence of the field of explainable AI (XAI) which addresses the problem of making AI decisions understandable by humans through post-hoc explainability techniques for black-box models (Ribeiro et al., 2016; Lundberg and Lee, 2017; Petsiuk et al., 2018; Fel et al., 2021) or complex ones (Zeiler and Fergus, 2014; Shrikumar et al., 2016; Selvaraju et al., 2017; Shrikumar et al., 2017; Sundararajan et al., 2017). These *feature attribution methods* aim at interpreting individual decisions in an intuitive manner, typically by assigning an importance score to each feature (e.g., the yearly income explains the loan refusal at 80%).

On the one hand, many of these methods provide explanations that can be regarded as *factual* explanations, pointing out why the decision was made (e.g., you did not get the loan because your salary was 50K/year). On the other hand, *counterfactual* explanations (e.g., had your income been 100K/year you would have gotten the loan) have become a cornerstone of XAI since the seminal work of Wachter et al. (2017), looking for minimal changes in the input’s features so that the output differs. The interest of counterfactual explanations, in contrast to standard attribution methods, is twofold. First, they are more intuitive than factual explanations, as evidenced by philosophical and psychological studies (Lewis, 1973b; Byrne, 2019). This aspect is critical, as building trustworthy AI systems is contingent to human acceptance. Second, they enable to provide *algorithmic recourse*, namely actionable guidelines for an end user to change the algorithm’s decision (e.g., increasing the salary by

50K/year). In what follows we detail the formulation of counterfactual explanations in XAI, stress out the possible confusions with counterfactual inference, and discuss the technical advantages and limitations.

1.5.1 Original problem formulation

The counterfactual explanation framework in XAI tackles the problem of explaining the decision of a black-box binary classifier by looking for minimal changes in the features so that the outcome differs. Formally, let $h : \mathbb{R}^p \rightarrow [0, 1]$ be a predictor returning the estimated probability of belonging to the positive class, and set a *focal point* v . Finding a counterfactual explanation of h 's behaviour at v amounts to solving

$$\min_{v' \in \mathbb{R}^p} c(v, v') + \lambda \cdot |h(v') - h(v)|^2 \quad (1.5)$$

where c a positive cost function on \mathbb{R}^p (e.g., a distance function) and $\lambda > 0$ is a large-enough trade-off parameter forcing the output to change. The difference between the focal point v and the returned input v' sheds light on which features mattered in the decision making process.

Parameter-tuning

The counterfactual explanation problem (1.5) can be tuned through the choice of the cost function c and the trade-off parameter λ . Different choices lead to different explanations, which raises the question of what a “good” counterfactual explanation is.

To some extent, this depends on the purpose of the explanation. For example, if we were to provide recourse to a person who has been denied a loan, we would not furnish a counterfactual explanation involving a change in nonactionable features such as the age, the race, or the sex. However, this would become relevant if we were to uncover discriminatory biases of the decision rule. This distinction between *acting* and *understanding* can be achieved in practice by well-choosing the cost function or by adding constraints to the optimization problem (Ustun et al., 2019).

It also became commonly agreed that counterfactual explanations should be *sparse*, that is implying changes in a limited number of features. We refer to (Keane et al., 2021, Section 5) for further insights on this aspect. The intuition is that sparsity keeps explanations simple and intelligible to humans. In their seminal work, (Wachter et al., 2017) leveraged a scaled version of Manhattan distance to ensure sparsity without, however, any control on the sparsity level.

Lastly, many papers pointed out that Problem (1.5), by being oblivious of the underlying probability distribution of the data, often returned out-of-distribution examples, leading to unrealistic explanations and unfeasible recourse. This issue cannot be addressed by simply modifying the cost function, but by more radical reformulations restricting the set of attainable worlds. Proposed solutions include the use of generative models (Liu et al., 2019) or high density path (Poyiadzi et al., 2020) to force minimal changes to end up within the dataset, but also causal models (Karimi et al., 2021) to take into account dependencies beyond mere correlations.

As explained by Kuhl et al. (2022), refinements such as choosing a sophisticated cost function, changing the trade-off parameter, or including a plausibility constraint are what distinguish a counterfactual-explanation solution v' from a mere *adversarial example*, defined as imperceptible changes in the input switching the model's output (Papernot et al., 2017).

The logic behind counterfactual explanations

Let us analyze this framework through the prism of Lewis' original definition of counterfactuals. For clarity, suppose that the focal point belongs to the negative class, that is $h(v) < 0.5$, and consider v' a solution to (1.5). Trivially, the following counterfactual statement holds:

$$\text{Had the input been equal to } v', \text{ then the output class would have been 1.} \quad (1.6)$$

This justifies the name of *counterfactual* explanations. However, we emphasize that solving (1.5) does not return any counterfactual explanations but specific ones. Recall that in general the antecedent of a counterfactual does not provide minimally sufficient conditions for its consequence to hold. Notably, any input v' such that $h(v') > 0.5$ provides a counterfactual statement with the same consequence as (1.6). Therefore, encouraging the proximity between v and v' in (1.5) is critical for the explanation to be relevant as it excludes noninformative antecedents that would be far from the decision boundary. As such, Problem (1.5) can be seen as a process to pick out "good" counterfactual explanations. In light of this, we discuss next two frequent confusions about this framework.

Counterfactual explanations and counterfactual counterparts

The returned input v' is often lazily referred as the *counterfactual* of v . We believe this denomination to be abusive and misleading, even for refined versions of Problem (1.5). For given world of reference and consequent, there are never counterfactuals in themselves, but counterfactuals *had a given event occurred*. However, Statement (1.6) does not specify such an event; the antecedent "Had v been v' " does not trace back the input v' to a meaningful feature modification $a \in \mathcal{A}$ of the world of reference v , hence why we rather call v' an *adversarial example* or a *counterfactual explanation* to avoid confusion with the counterfactual counterparts presented in Sections 1.3 and 1.4.

To further underline the distinction, let us evaluate the truthiness of Statement (1.6) in a logician fashion. The world of reference is the focal point v , the antecedent is "The input is equal to v' ", and the outcome is "The output class is 1". In a first time, we must compute the most similar worlds to v satisfying the antecedent. This operation is straightforward: v' is the only world being equal to v' , thereby satisfying this property. Then, because $h(v') > 0.5$ by construction, the outcome holds and the counterfactual statement is true. Observe that the main challenge of counterfactual verification, that is applying faithful interventions to attain the antecedent, is trivialized the context of vanilla counterfactual explanations. Naturally, v' could be the result of some intervention T_a on v , but this hypothetical step is completely ignored by default; we explain a few paragraphs below how to integrate interventions to explanations. To sum-up, in the standard counterfactual-explanation framework, only the *explanation* is counterfactual, not the *generated input*.

We emphasize that the issue we are pointing out is more about denomination than utility, since knowing a reachable target v' is always useful, especially in a recourse scenario.

Nevertheless, the sole information of the target feels shallow in a fairness-auditing context, where one would like to understand whether the change from v to v' suggests a change in the person's sex or race for example.

Input modification versus output modification

Keep in mind that the computation of counterfactual explanations does not correspond to an intervention on the classifier's output. The counterfactual-explanation principle, looking for the closest instances so that the output differs, inopportunately resembles Lewis's computation of alternatives worlds, looking for the most similar instances so that the antecedent holds. Should we intervene on the output, the alternative example v' would correspond to the most similar alternative world such that $h(v') > 0.5$ holds, rendering the counterfactual statement:

$$\text{Had the output class been equal to 1, then the input would have been } v'. \quad (1.7)$$

However, since we aim at explaining the output class, it must characterize the consequent in the counterfactual statement, not the antecedent. Moreover, this viewpoint would be irrelevant even for explaining the features: it is the input that causes the decision, not the contrary. Therefore, there cannot be downstream effects of the modification of the output onto the features.

To sum-up, the standard formulation of the counterfactual explanation problem is critically not related to counterfactual inference. It simply selects the most relevant counterfactual explanation among all possible ones. The misunderstanding in the broad area of counterfactuals for machine learning largely comes from the lack of clarity on this matter. We believe that the key for transparency lies in analyzing any counterfactual framework through the prism of Lewis's original formulation presented in Section 1.2. That being said, we now turn to the limitations of (1.5) and possible solutions.

1.5.2 Intervention-based counterfactual explanations

As aforementioned, solving (1.5) with a classical distance often generates unfaithful explanations or unfeasible recourse, as the adversarial examples may not be the result of a realistic intervention. Additionally, it does not accompany by default the generated example by an explanatory action on the focal point. A way of addressing this issue is to clearly define the counterfactual model first, and then restrict the possible counterexamples v' to the points attainable by applying interventions on the focal point v .

We use the same notations as in Section 1.2 where \mathcal{A} represents the set of possible interventions over a list $\mathcal{I} \subseteq \{1, \dots, p\}$ of manipulable features. The choice of \mathcal{I} heavily depends on the considered objective. If for instance the goal is to compute algorithmic recourse, then any interventions on immutable attributes such as the race should be excluded from this list. In contrast, they are relevant if the goal is to uncover discriminatory biases. Assuming for simplicity that alternative worlds are deterministically implied, we denote by $T_a(v)$ the unique counterpart of any factual world $v \in \mathcal{V}$ after applying $a \in \mathcal{A}$, namely the counterfactual counterpart of v had a occurred. Typically, for $a := (I, \tilde{v}_I)$ the transformation T_a can be induced by the do-intervention $\text{do}(I, \tilde{v}_I)$ on a presumed causal model generating

the data, as proposed in [Karimi et al. \(2021\)](#). Then, letting $\mathcal{F}(v) = \{T_a(v) \mid a \in \mathcal{A}\}$, we can define the following refined counterfactual-explanation optimization problem:

$$\min_{v' \in \mathcal{F}(v)} c(v, v') + \lambda \cdot |h(v') - h(v)|^2, \quad (1.8)$$

which can be recast as

$$\min_{a \in \mathcal{A}} c(v, T_a(v)) + \lambda \cdot |h(T_a(v)) - h(v)|^2. \quad (1.9)$$

The benefit of this last formulation is that it poses the problem in terms of finding meaningful actions rather than finding adversarial examples. Thus, it produces statements of the form,

$$\text{Had } a \text{ occurred, then the output class would have been } 1, \quad (1.10)$$

adding a layer of information to [\(1.6\)](#). The pitfall comes from the set \mathcal{A} being infinite in general, for instance as soon as one actionable feature is continuous. In particular, this imposes to discretize the space of feasible actions to practically solve [\(1.9\)](#), leading to an approximated explanation. In addition, looking for the less-costly action within a fine-grained space can be expensive.

In this section, we showed how to interface counterfactuals with explicability techniques for artificial-intelligence systems. In the sequel, we focus on the interest of counterfactual inference in algorithmic fairness.

1.6 Counterfactuals in fairness

In the introduction of this manuscript, we mentioned a number of notoriously unfair algorithms and gave insights on the origin of such biases. We also considered fairness-inspired examples of causal inference in [Section 1.4](#). However, we still have not properly formalized what it means for an algorithmic decision rule to be fair. This is the role of this section, which briefly reviews classical fairness criteria.

1.6.1 Standard definitions of fairness

What does it mean for a decision to be fair? For example, should someone’s application receive a negative appraisal while a very similar one gets positive evaluation, the assessment process would be naturally deemed unfair. This intuition motivated the guiding principle of *treating similar individuals similarly* formalized by [Dwork et al. \(2012\)](#): a decision rule is individually fair if it renders similar outputs to similar inputs, letting $l > 0$, a (deterministic) decision rule $h : \mathcal{V} \rightarrow \mathbb{R}$ is *l-individually fair* if for (almost) every $(v, v') \in \mathcal{V}$, $|h(v) - h(v')| \leq l\|v - v'\|$.

In addition to this principle, we generally consider that a fair rule must not discriminate, that is must not make distinction on the basis of so-called *protected* or *sensitive attributes* such as race, sex, gender, sexual orientation, religious belief, political opinion and age. Many research papers proposed formal criteria to encode different notions of “excluding” or “ignoring” protected attributes from the decision making process. The most straightforward criteria

called *fairness through unawareness* requires the decision rule not to take the protected attribute, encoded by a variable $s \in \mathcal{S}$, as an entry of its formula. More precisely, a deterministic decision rule $h : \mathcal{X} \times \mathcal{S} \rightarrow \mathbb{R}$ satisfies *fairness through unawareness* if there exists some $\tilde{h} : \mathcal{X} \rightarrow \mathbb{R}$ such that for every $s \in \mathcal{S}$, $h(\cdot, s) = \tilde{h}(\cdot)$.

For example, a fair hiring process satisfying individual fairness and unawareness is expected to give similar reviews to individuals with similar skills while ignoring the difference in sex or race. However, even though this first notion of fairness through unawareness seems legitimate, it does not prevent an algorithm to discriminate on the basis of a protected attribute encoded by a random variable $S : \Omega \rightarrow \mathcal{S}$. This is due to the other input variables $X : \Omega \rightarrow \mathcal{X}$ being correlated with S (e.g., the height and salary of a person are correlated with their sex), enabling the algorithm to infer the protected attribute through the features to make unfair decisions. This issue led to the definition of *statistical* or *demographic parity*: a fairness condition requiring the statistical independence between decisions and protected attributes. In the sequel, \hat{Y} denotes a predictor of some target variable $Y : \Omega \rightarrow \mathcal{Y} \subseteq \mathbb{R}$, defined as a deterministic function of either X and S (aware) or solely X (unaware). Then, \hat{Y} satisfies statistical parity if $\hat{Y} \perp\!\!\!\perp S$.

For a binary classifier, that is when $\mathcal{Y} = \{0, 1\}$, this simply signifies equal rates of positive decisions across protected groups. If additionally $\mathcal{S} = \{0, 1\}$, then statistical parity is commonly evaluated through the *disparate impact*,

$$\text{DI}(\hat{Y}, S) := \frac{\min \left\{ \mathbb{P}(\hat{Y} = 1 \mid S = 1), \mathbb{P}(\hat{Y} = 1 \mid S = 0) \right\}}{\max \left\{ \mathbb{P}(\hat{Y} = 1 \mid S = 1), \mathbb{P}(\hat{Y} = 1 \mid S = 0) \right\}}, \quad (1.11)$$

which equals 1 when fairness is perfectly achieved and approaches 0 as the predictor becomes more unfair. Importantly, the main interest of statistical parity is to prevent the use of unfair proxies: irrelevant variables in X for the prediction task that are correlated to the protected attributes S . Typically, a few-month gap in the work experience likely corresponds to a maternity leave, and could be used to unfavor women even without explicit mention of the gender in the resumes. While this provides a well-founded notion of fairness, complying with parity is arguably undesirable in some cases where some differences in the distributions of X conditional on $S = s$ are relevant for the decision-making process. In certain fields, women tend to have less adequate profiles experience-wise than men due to societal inequalities. Therefore, asking for parity means recruiting less-experienced women than some rejected men, which would be deemed unfair according to individual fairness. This raises the question of whether companies (or other institutions) should carry out affirmative actions to address structural inequalities they are not directly responsible for. For example, a US federal law constraints employers to have a disparate impact above 80% except if they can justify the discrimination by economic interests ([Civil Rights Act, 1964](#)). Note that the fairness-inspired illustration of potential-outcome counterfactuals from Section [1.4](#) also called attention to this question.

Critically, statistical parity is tailored to prediction tasks where the outcomes can be graduated into desirable ones and undesirable ones (e.g., a higher or lower salary, acceptance of rejection) and becomes irrelevant for prediction tasks such as image labelling. Imagine a classifier meant to automatically recognize individuals on pictures. It would not make sense to ask the classifier to give similar outputs in average between white and black individuals.

However, we would expect it to be equally accurate across race groups. This led to the definition of *equality of odds*, the accuracy pendant of statistical parity (Hardt et al., 2016): a predictor \hat{Y} satisfies equality of odds if $\hat{Y} \perp\!\!\!\perp S \mid Y$.

While there exist other fairness conditions with respect to sensitive attributes, such as *avoiding disparate treatment* ($\hat{Y} \perp\!\!\!\perp S \mid X$) and *predictive parity* ($Y \perp\!\!\!\perp S \mid \hat{Y}$), statistical parity and equality of odds have become the gold-standard criteria in the fair learning literature. Note in passing that these two conditions are often incompatible: no algorithm can simultaneously satisfy statistical parity and equality of odds, except if Y is independent of S (Kleinberg et al., 2017). Thus, in the very common situation where the outcome itself is biased by the protected attributes, one cannot impose both fairness conditions. More generally, tensions and incompatibilities between fairness conditions prevent practitioners from piling-up fairness guarantees when designing predictors. In practice, trade-offs must be made according to the prediction task, but also feasibility and legal constraints.

Regarding ethical and legal aspects, criteria such as statistical parity and equality of odds are notoriously limited since they only provide notions of *group* fairness, and do not control discrimination at a subgroup or an individual level: a conflict illustrated by Dwork et al. (2012). In particular, the current French law only recognizes discrimination at the individual level, making group fairness conditions nonoperational. This justifies the need for sharper definitions, ensuring the protection of sensitive attributes at every decisions. Next, we explain how counterfactuals provide a rigorous basis to address this problem.

1.6.2 Counterfactual-based fairness

While it is straightforward to test group fairness conditions, for instance by computing the disparate impact on a testing set and verifying its closeness to 1, there is no natural way to control for discrimination at the individual level. An interesting procedure is the one adopted for human-based decisions by the Observatoire des Discriminations⁵ in France. When an individual complains that they were denied a position because of their presumed origin, the organization constructs a fake application with equal relevant skills and work experience, but changing all origin indicators (e.g., name, residence) to white-connoted ones. Should the company consider this application, the decision would be judged racist, hence unfair. Observe that this test tries to answer the counterfactual query “Had the applicant been white, would have they been accepted?” using a *ceteris-paribus* approach. This provides an intuitively legitimate condition to uncover occurrences of discrimination.

However, especially in the case of machine learning, simply switching the protected attribute while keeping the other features equal is notoriously inefficient for three reasons. First, as mentioned before, because of strong correlations between the features, the machine often learns a proxy instead of the explicit attribute. This is why it is hard to test their fairness: some algorithms will appear to be fair because they pass the attribute-switch test, all the while being discriminatory. Second, also because of statistical correlations between features, an individual with a flipped attribute *ceteris paribus* can easily become an outlier of the targeted population. For example, a tall well-paid man can be representative of the men distribution, whereas a tall well-paid woman may not be statistically representative of her gender group. Because the algorithm is trained on realistic data-sets, it lacks sufficient

⁵<https://www.observatoiredesdiscriminations.fr/testing>

information to make well-founded decisions on such irregularities. Third, as most of the algorithms do not even take protected attributes as entry variables (they are fair through unawareness), simply switching them cannot impact the decision. Yet, they can still be discriminatory because of the aforementioned proxy problem.

Definition(s)

The above discussion contextualizes into fairness-related problems what we previously explained on the necessity for sophisticated counterfactual models taking into account dependencies between features. For instance, [Kusner et al. \(2017\)](#) proposed to rely on SCMs to alter the protected attribute, leading to the accepted notion of *counterfactual fairness* which requires the algorithm to treat equally individuals and their structural counterfactual counterparts. Mathematically, a predictor $\hat{Y} := h(X, S)$ is *counterfactually fair* if for every possible observation $\{X = x, S = s\}$ and modification $s' \in \mathcal{S}$,

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

where $\hat{Y}_{S=s} := h(X_{S=s}, s)$. We could generalize this definition to SCM-free constructions of counterfactual counterparts, to other counterfactual models. This idea is the crux of Chapter [2](#), where we build counterfactuals using mass-transportation techniques, leading to an original noncausal individual fairness condition (see Definition [2.5.2](#)). In passing, note also that adapting counterfactual fairness to ceteris-paribus counterparts recovers fairness through unawareness.

Fairness auditing

As explained at the beginning of this discussion, whatever the considered model of interventions, counterfactual reasoning also has promising applications in fairness auditing of classifiers. It permits to compare a factual outcome with its counterfactual outcome “had S been flipped”, therefore to test whether S mattered in the decision-making process at the individual scale. In [\(Alvarez and Ruggieri, 2023\)](#), the authors developed such a testing procedure, making it robust to random occurrences of discrimination by studying output variations over neighborhoods of the factual and counterfactual inputs.

But more interestingly, it permits to *explain* the role of S in the decision. The FlipTest of [Black et al. \(2020\)](#) records the changes in X of all the counterfactual pairs “had S been flipped” for which there is discrimination (i.e., \hat{Y} differs). These changes shed light on the features tainted by the protected attribute that mattered the most in the decision making process. Crucially, this enables experts of the prediction task to assess whether the classifier actually relied on unfair proxies. More concretely, let S be the sex and X contain the height among other variables. Due to the correlation between height and sex, intervening on the sex will likely incur a height modification. However, the unfairness of using height in the decision-making process seems to depend on the scenario. For illustration, figure out algorithmic recruitment procedures: in the case of an office job recruitment, taking height into account is obviously unfair; in the case of casting actors or models, not necessarily. This emphasizes that fairness and explainability are intertwined; one cannot rigorously assess or ensure algorithmic fairness if they cannot understand the basis on which decisions are made.

For curious readers, we prove the statistical consistency of the FlipTest in (De Lara et al., 2021b) (which we did not include in this manuscript for the sake of succinctness). Note also that the FlipTest leverages optimal transport instead of SCMs to compute counterfactuals, which greatly motivated the work presented in the next chapter.

Appendix 1.A Proofs of Section 1.3

Proof of Proposition 1.3.1 Since \mathcal{M} is acyclical, there exists a topological ordering on the indices in \mathcal{I} , and therefore on the subset J . This means in particular that there exist some $j \in J$ such that G_j takes only variables in V_I as endogenous inputs. Starting from these indices, and recursively substituting along the topological ordering produces a measurable F_J such that

$$V_J \stackrel{\mathbb{P}\text{-a.s.}}{=} F_J(V_{\text{Endo}(J) \setminus J}, U_{\text{Exo}(J)}).$$

Note that $\text{Endo}(J) \setminus J \subseteq I$. Carrying out the same substitution on the intervened model $\mathcal{M}_{V_I=v_I}$ with solution \tilde{V} gives

$$\tilde{V}_J \stackrel{\mathbb{P}\text{-a.s.}}{=} F_J(v_{\text{Endo}(J) \setminus J}, U_{\text{Exo}(J)}),$$

while by definition $\tilde{V}_I \stackrel{\mathbb{P}\text{-a.s.}}{=} v_I$. ■

Appendix 1.B Proofs of Section 1.4

Proof of Proposition 1.4.1 Let $s \in \mathcal{S}$ and recall that according to Proposition 1.3.1,

$$\begin{aligned} Y_{S=s} &\stackrel{\mathbb{P}\text{-a.s.}}{=} F_Y(s, X_{S=s}, U_Y) \\ X_{S=s} &\stackrel{\mathbb{P}\text{-a.s.}}{=} F_X(s, U_X). \end{aligned}$$

This means that these identities hold on a measurable set $\Omega^* \subseteq \Omega$ such that $\mathbb{P}(\Omega^*) = 1$.

Next, observe that for any $\omega \in \Omega^*$ such that $S(\omega) = s$, we have

$$X(\omega) = F_X(s, U_X(\omega)) = X_{S=s}(\omega),$$

therefore

$$Y_{S=s}(\omega) = F_Y(s, X(\omega), U_Y(\omega)) = F_Y(S(\omega), X(\omega), U_Y(\omega)) = Y(\omega).$$

This concludes the proof. ■

Proof of Theorem 1.4.1 Let us compute the conditional joint distribution $\mathcal{L}((Y_{s''})_{s'' \in \mathcal{S}} \mid X = x, S = s)$ which is well-defined for all $x \in X(\Omega)$ and $s \in \mathcal{S}$ by positivity. The consistency rule entails that

$$\mathcal{L}((Y_{s''})_{s'' \in \mathcal{S}} \mid X = x, S = s) = \mathcal{L}((Y_0, \dots, Y_{s-1}, Y, Y_{s+1}, \dots, Y_N) \mid X = x, S = s).$$

Moreover, according to \mathcal{M} the observed outcome can be written as $Y \stackrel{\mathbb{P}\text{-a.s.}}{=} F_Y(S, X, U_Y)$, leading to

$$\mathcal{L}((Y_{s''})_{s'' \in \mathcal{S}} \mid X = x, S = s) = \mathcal{L}((Y_0, \dots, Y_{s-1}, F_Y(s, x, U_Y), Y_{s+1}, \dots, Y_N) \mid X = x, S = s).$$

Next, remark that Assumption **(O)** entails that $U_Y \perp\!\!\!\perp (S, X)$. Therefore, it follows from $(Y_s)_{s \in \mathcal{S}} \perp\!\!\!\perp X \mid S$ that for any $s' \neq s$ the above equality is equivalent to

$$\mathcal{L}((Y_{s''})_{s'' \in \mathcal{S}} \mid X = x) = \mathcal{L}((Y_0, \dots, Y_{s-1}, F_Y(s, x, U_Y), Y_{s+1}, \dots, Y_N) \mid X = x, S = s').$$

Then, using once the again the consistency rule we obtain

$$\mathcal{L}((Y_{s''})_{s'' \in \mathcal{S}} \mid X = x) = \mathcal{L}((Y_0, \dots, F_Y(s, x, U_Y), \dots, Y_{s'-1}, Y, Y_{s'+1}, \dots) \mid X = x, S = s'),$$

and the expression of Y through F_Y yields

$$\mathcal{L}((Y_{s''})_{s'' \in \mathcal{S}} \mid X = x) = \mathcal{L}((Y_0, \dots, F_Y(s, x, U_Y), \dots, F_Y(s', x, U_Y), \dots) \mid X = x, S = s').$$

We repeat this step by conditioning on all possible values of S to finally obtain

$$\mathcal{L}((Y_{s''})_{s'' \in \mathcal{S}} \mid X = x, S = s) = \mathcal{L}((F_Y(s'', x, U_Y))_{s'' \in \mathcal{S}} \mid X = x, S = s).$$

Therefore, since $U_Y \perp\!\!\!\perp (S, X)$, marginalizing on (S, X) yields

$$\mathcal{L}((S, X, (Y_s)_{s \in \mathcal{S}})) = \mathcal{L}((S, X, (F_Y(s, X, U_Y))_{s \in \mathcal{S}})).$$

■

Chapter 2

Transport-based counterfactual models

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 chapter, we aim at laying out the theoretical foundations for a new, implementable approach to counterfactual thinking.

2.1 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, 1973a) 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; Slack et al., 2021; Bajaj et al., 2021; Rasouli and Chieh Yu, 2022), 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 sex of a person in such a translation might convert from a typical male into an untypical female, rendering out-of-distribution counterfactuals like the following: *if I were a woman I would be 190cm tall and weigh 85 kg*. According to intuition, such counterfactuals are false and rightly so because they are not representative of the underlying statistical distributions. As a practical consequence, such counterfactuals typically hide biases in machine learning decision rules (Lipton et al., 2018; Besse et al., 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.

2.1.1 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. Section 2.2 introduces the basics of mass transportation and optimal transport theory, while we refer to Section 1.3 from Chapter 1 for the necessary background on Pearl’s causal modeling. Both sections serve as the theoretical and notational toolbox that

will be used throughout; they are meant to keep the chapter self-contained and can be skipped by readers familiar with these subjects.

2. In Section 2.3, 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 2.4, 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 2.5, 2.6 and 2.7, 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 1.3 and 2.2 provide the prerequisites for the chapter; Sections 2.3 and 2.4 introduce the concept of counterfactual models and the corresponding theory; Sections 2.5 to 2.7 address fairness applications of these models.

2.1.2 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 2.3 and 2.4 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 2.3 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 2.5 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 2.4, 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).

2.2 Mass transportation

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

2.2.1 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 $\pi(E_1 \times \mathbb{R}^d) = P(E_1)$ and $\pi(\mathbb{R}^d \times E_2) = Q(E_2)$ for all Borel sets $E_1, E_2 \subseteq \mathbb{R}^d$. Throughout the manuscript, we denote by $\Pi(P, Q)$ the set of joint distributions over $\mathbb{R}^d \times \mathbb{R}^d$ 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 $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$ that *pushes forward* P to Q , that is $Q(E) := P(T^{-1}(E))$ for any Borel set $E \subseteq \mathbb{R}^d$. This property, denoted by $T_{\#}P = Q$, 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 $T_1, T_2 : \mathbb{R}^d \rightarrow \mathbb{R}^d$, we define $(T_1 \times T_2) : \mathbb{R}^d \rightarrow \mathbb{R}^d \times \mathbb{R}^d, x \mapsto (T_1(x), T_2(x))$. As such, $(T_1 \times T_2)_{\#}P$ denotes the law of $(T_1(Z), T_2(Z))$ where $\mathcal{L}(Z) = P$. This coupling admits $T_{1\#}P$ and $T_{2\#}P$ as first and second marginal respectively. Thus, the deterministic coupling π between P and Q characterized by a push-forward operator T satisfying $T_{\#}P = Q$ can be written as $\pi = (I \times T)_{\#}P$ 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. We write $\mathcal{T}(P, Q)$ for the set of measurable mappings pushing forward P to Q .

2.2.2 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.

Arbitrary cost

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

$$\min_{T \in \mathcal{T}(P, Q)} \int_{\mathbb{R}^d} c(x, T(x)) dP(x). \quad (2.1)$$

We refer to solutions to [\(2.1\)](#) 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 ,

$$\min_{\pi \in \Pi(P, Q)} \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x, x') d\pi(x, x'). \quad (2.2)$$

Solutions to [\(2.2\)](#) are *optimal transport plans* (possibly non deterministic) between P and Q with respect to c . In contrast to optimal transport maps, they exist under very mild assumptions, like the non negativity of the cost. Notice that, since a push-forward operator can be identified to a coupling, the set of feasible solutions to [\(2.1\)](#) is included in the set of feasible solutions to [\(2.2\)](#).

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 [\(2.2\)](#), whose form is $(I \times T)_\# P$ where $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$ solves the corresponding squared Monge problem,

$$\min_{T: T_\# P = Q} \int_{\mathbb{R}^d} \|x - T(x)\|^2 dP(x). \quad (2.3)$$

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 [\(2.3\)](#). 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 $\nabla\phi_{\#}P = Q$. We combine Brenier's and McCann's theorems into the following lemma, which simplifies the search for the solutions to (2.3).

Lemma 2.2.1: “Brenier + McCann”

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 $T : \text{supp}(P) \rightarrow \text{supp}(Q)$ is a solution to (2.3) if and only if it satisfies the two following conditions:

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

This result will play a key role in Section 2.4.2 to prove a link between optimal transport and causality.

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 (2.2) between their empirical counterparts P^n and Q^m . By identifying a discrete coupling to a matrix, we write this problem as,

$$\min_{\pi \in \Sigma(n, m)} \sum_{i=1}^n \sum_{j=1}^m c(x_i, x'_j) \pi(i, j), \quad (2.4)$$

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 (2.2) admits a unique solution π , then a sequence $\{\pi^{n, m}\}_{n, m \in \mathbb{N}}$ of solutions to Problem (2.2) 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 (2.4) is necessarily nondeterministic as soon as $n \neq m$, the corresponding solution to Problem (2.2) 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 (2.4) 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 2.7) 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 nonzero coefficients, which precludes memory-efficient implementations. This is why we address only standard optimal transport in our numerical experiments.

2.3 Counterfactual models

We now have all the tools to focus on the main subject of this chapter: counterfactual reasoning. As mentioned in the introduction, both causality and transport techniques have been used for this purpose. However, a yet nonappreciated 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.

2.3.1 Problem setup

Set $d \geq 1$, and define the random vector $V := (X, S) \in \mathbb{R}^{d+1}$, where the variables $X : \Omega \rightarrow \mathcal{X} \subseteq \mathbb{R}^d$ represent some observed features, while the variable $S : \Omega \rightarrow \mathcal{S} \subset \mathbb{R}$ 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 values 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.

2.3.2 Structural counterfactuals

Answering the counterfactual question from Section 2.3.1 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

$$V_k \stackrel{\mathbb{P}\text{-a.s.}}{=} G_k(V_{\text{Endo}(k)}, U_{\text{Exo}(k)}),$$

where G_k is a real-valued measurable function, U is a vector of *exogenous* variables, while $V_{\text{Endo}(k)}$ and $U_{\text{Exo}(k)}$ 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 $\text{do}(S = s)$, 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.

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 $x' \in \text{supp}(\mu_{\langle s'|s \rangle})$. 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 $\text{do}(S = s')$ 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 2.3.1: Structural counterfactuals

Let \mathcal{M} satisfy **(A)**. For an observed evidence $\{X = x, S = s\}$ and an intervention $\text{do}(S = s')$, 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.

Mass-transportation viewpoint

While the mainstream literature on causality generally operates with the definition of structural counterfactuals given by the three-step procedure (Kusner et al., 2017; Barocas

et al., 2019), we focus in this chapter on a *mass-transportation viewpoint* of counterfactuals, formalized by the following definition.

Definition 2.3.2: Structural counterfactual model

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

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

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 $d\pi_{\langle s'|s \rangle}^*(x, x')$ 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 2.3.1 and 2.3.2 characterize the exact same counterfactual statements, the formal link being $d\pi_{\langle s'|s \rangle}^*(x, x') = \mu_{\langle s'|s \rangle}(x'|x)d\mu_s(x)$. 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 2.4.1. 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).

2.3.3 Transport-based counterfactuals

The main issue of structural counterfactuals, which will be widely discussed in Section 2.3.5, 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.

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, $t : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^d \times \mathbb{R}^d, (x, x') \mapsto (x', x)$ denotes the permutation function.

Definition 2.3.3: Transport-based counterfactual model

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

- (i) $\pi_{\langle s'|s} \in \Pi(\mu_s, \mu_{s'})$;
- (ii) $\pi_{\langle s|s} = (I \times I)_{\#} \mu_s$;
- (iii) $\pi_{\langle s|s'} = t_{\#} \pi_{\langle s'|s}$.

An element of Π is called a *counterfactual coupling*. We say that Π is a *random counterfactual model* if at least one coupling for $s \neq s'$ is not deterministic. Otherwise, we say that Π is a *deterministic counterfactual model*. In the deterministic case, Π can be identified almost everywhere to a collection $\mathcal{T} := \{T_{\langle s'|s} \}_{s, s' \in \mathcal{S}}$ of measurable mappings from \mathcal{X} to \mathcal{X} satisfying for every $s, s' \in \mathcal{S}$,

- (i) $T_{\langle s'|s} \# \mu_s = \mu_{s'}$;
- (ii) $T_{\langle s|s} = I$;
- (iii) $T_{\langle s'|s}$ is invertible μ_s -almost everywhere such that $T_{\langle s|s'} = T_{\langle s'|s}^{-1}$.

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 2.3.3 translates the intuition that a realistic counterfactual operation on S should morph the nonintervened 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 counterfactual counterparts. Remark that in the case of discrete measures, the operation $t_{\#}$ in condition (iii) simply amounts to transposing the associated coupling matrices. Lastly, note that this definition replaces the unobservable, SCM-dependent distributions $\{\mu_{\langle s'|s} \}_{s, s' \in \mathcal{S}}$ of structural counterfactual models by the observational $\{\mu_{s'} \}_{s' \in \mathcal{S}}$ for feasibility reasons. In Section 2.4.1, we will see that this approximation makes sense in typical fairness settings where $\mu_{\langle s'|s} = \mu_{s'}$ 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 $x' = T_{\langle s'|s}(x) \in \mathcal{X}_{s'}$. In

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

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 nondeterministic 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 noncontinuous 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 2.6.

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 2.3.3, 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 (2.1) 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 2.2.2) has remarkable properties. According to Lemma 2.2.1, solutions to Problem (2.3) are gradients of convex functions, which extends the notion of nondecreasing 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 one marginal to the other is an intuitive definition of the counterfactual counterpart. For the sake of illustration, Section 2.3.4 below provides several examples of optimal transport applied to counterfactual reasoning.

In Section 2.4.2 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 2.3.3 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 chapter. Additionally, this generalization permits the use of regularized optimal transport (Cuturi, 2013), which deviates from the original formulation of Problem (2.2), 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 2.3.3 taking into account random couplings. Lastly, we will see in Section 2.4.1 that structural counterfactual models are transport-based counterfactual models—but not necessarily optimal-transport-based—under some assumptions.

2.3.4 Examples

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

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 2.3.2. The coupling from $S = 0$ to $S = 1$ is given by

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

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

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

Figures 2.1a and 2.1b 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 2.4.1 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 $t_{\#}\pi_{\langle 1|0 \rangle}^* = \pi_{\langle 0|1 \rangle}^*$. 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 2.4.1.

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 (2.4) between samples provides a matching that cannot generalize to new, out-of-sample observations. This is why we employ POT’s in-built nonregularized 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 2.1a. The converse counterfactual operation $T_{\langle 0|1 \rangle}$ represented in Figure 2.1d is produced by inversion.

We emphasize that all the couplings in Figure 2.1, 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 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 2.1c 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 2.4.2.

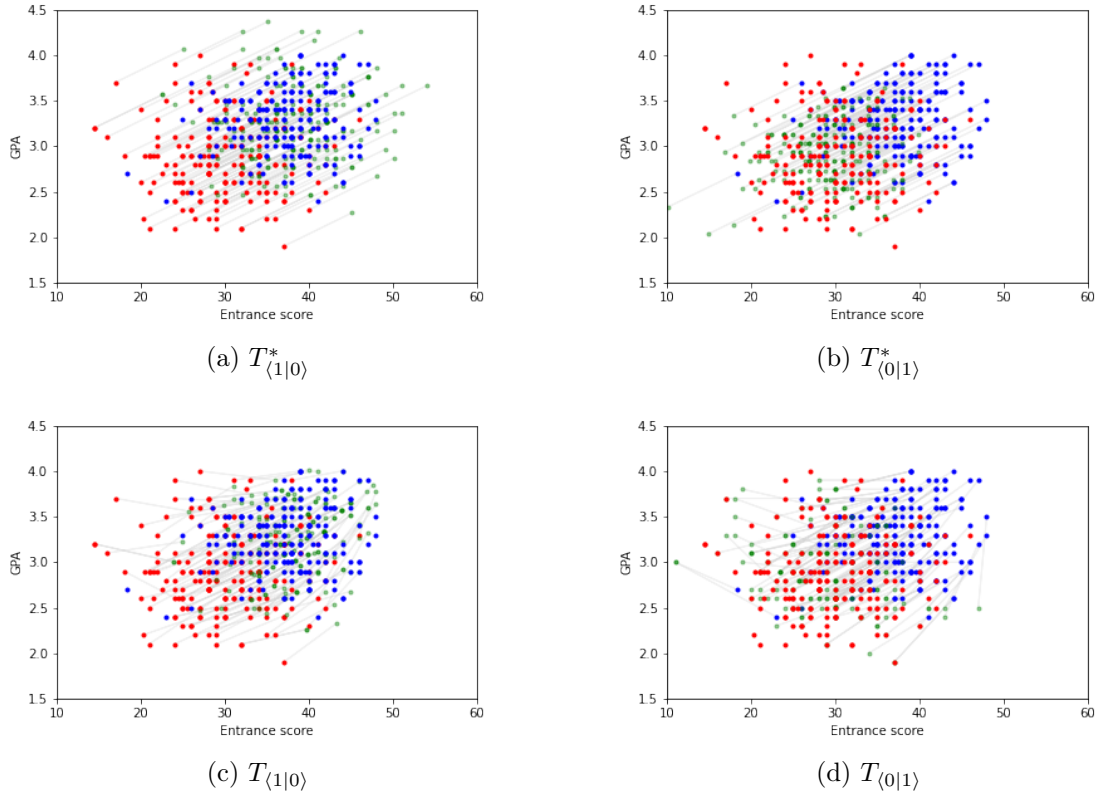


Figure 2.1: Counterfactual models for the Law dataset. The red sample represents 200 factual black students while the blue sample represents 200 factual white students. The green sample depicts counterfactual instances: the first column (Figures 2.1a and 2.1c) has white counterfactual students; the second column (Figures 2.1b and 2.1d) has black counterfactual students. The lightgray lines describe the coupling between factual and counterfactual instances.

Body-measurement dataset

We now further illustrate the properties of optimal-transport counterfactuals on a dataset of body measurements from $n_0 = 260$ females and $n_1 = 247$ males. The features of interest are the weight X_1 and the height X_2 , while S encodes the sex. Suppose now that Bob is a 80kg and 190cm male. What would have been Bob’s height and weight had he been a female? 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 female, she would have been 59kg and 177cm.

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

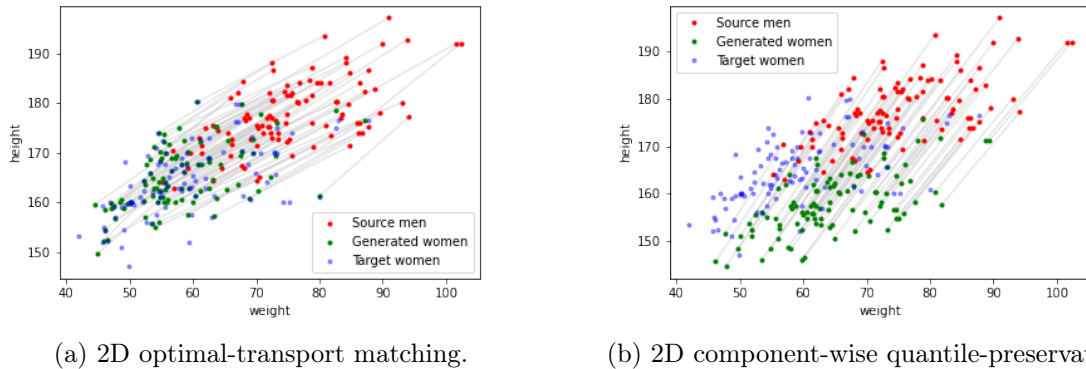


Figure 2.2: Body dataset. The red dots represent a data sample of men, while the blue dots represent a data sample of women. The green dots are the estimated counterfactual counterparts of the male sample.

due to the optimal map being the unique gradient of a convex function between distributions as previously explained. We underscore in Figure 2.2 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 sex’s distribution.

2.3.5 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 2.6 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 2.3 by discussing important drawbacks of the causal account to counterfactual reasoning as well as the limitations of the transport approach.

Shortcomings of the causal approach

The main limitation of *structural* counterfactual reasoning, as for any causal-based framework, is its feasibility. Notably, counterfactual inference requires a *fully specified* SCM. What follows is a recap of the concluding discussion from Section 1.3, which explained the obstacles of every specification step.

First of all, fully specifying the causal model from scratch is a too strong assumption in practice. It would demand to agree on (1) the causal graph, (2) the structural equations, (3) the distribution of the input exogenous variables, and (4) to check whether the model fits the observations. This is not a realistic scenario, especially for high dimensional data with

nonlinear correlations. Besides, this is likely not workable for businesses since this tedious procedure would have to be repeated for each new prediction task.

Secondly, techniques to learn the causal graph from observational data are computationally challenging (Cooper, 1990; Chickering et al., 2004; Scutari et al., 2019) and do not furnish the structural equations, which are necessary for counterfactual inference. Estimating these equations amounts to a complex multivariate regression problem, hence why the literature only handles simple, low-dimensional models (Shimizu et al., 2006; Hoyer et al., 2008; Kusner et al., 2017; Russell et al., 2017). Moreover, the approximation error implied by the choice of simple functional forms for the structural equations can incur unrealistic, out-of-distribution counterfactuals, as exemplified in Figure 2.1 above. To our knowledge, the literature on causal counterfactuals has not pointed out this flaw to date.

Thirdly, an inferred causal model will always suffer from causal uncertainty: there exist several causal models corresponding to a same data distribution (see Bongers et al., 2021, Example 4.2), which leads in particular to possibly different counterfactual models. It cannot be tested whether the adjusted model is the “true” one, making the modeling inherently uncertain. Moreover, for nondeterministic 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.

Lastly, counterfactual quantities are sometimes nonexistent in Pearl’s causal framework if Assumption (A) does not hold. We emphasize that observational data can be generated through an acyclical mechanism. Nevertheless, (solvable) acyclic models do not always admit solutions under do-interventions (Bongers et al., 2021, Example 2.17), implying that $X_{S=s}$ and all the related counterfactual quantities may not be defined in such situations.

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 $|\mathcal{S}|(|\mathcal{S}| - 1)/2$ 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 sex 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 2.1. 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.

2.4 Theoretical results

Until now, we have recalled the basics of causality and transport in Sections 1.3 and 2.2 and introduced counterfactual models, either causal-based or transport-based, in Section 2.3. In what follows, we demonstrate connections between both approaches. Concretely, we firstly explore in Section 2.4.1 the relationship between an SCM and the counterfactual model it induces, providing justifications to what we observed in Section 2.3.4. 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 2.3, we demonstrate in Section 2.4.2 that optimal transport recovers structural counterfactuals in specific cases.

2.4.1 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 2.4.1 involves any specific knowledge on optimal transport theory, only on causal modeling and (general) mass transportation.

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 2.4.1: Solution-map expression of the features

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

This is a direct consequence of Proposition 1.3.1 by partitioning V into X and S . For clarity, we wrote the proof of this specific case in the dedicated section. It 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 $s \in S$ 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, \bar{E} denotes the closure of any $E \subseteq \mathbb{R}^d$.

Proposition 2.4.1: Set of counterfactual counterparts

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

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

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): Invertibility of the causal generative process

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 **1.3.1** and Section **2.3.4**.

Example 2.4.1: Additive noise model

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

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

where $\phi_k : \mathcal{V}_{\text{Endo}(k)} \rightarrow \mathcal{V}_k$. Note that the random seed U is fully determined by the value of V , meaning that for any $v \in V$ the posterior distribution $\mathcal{L}(U | V = v)$ 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

$$X \stackrel{\mathbb{P}\text{-a.s.}}{=} \varphi(S, X) + U_X,$$

where $\varphi : \mathcal{S} \times \mathcal{X} \rightarrow \mathcal{X}$ is a deterministic measurable function; the equation on S does not matter. Assumption **(A)** entails through unique solvability that $X \stackrel{\mathbb{P}\text{-a.s.}}{=} (I - \varphi(S, \cdot))^{-1}(U_X)$. 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 2.4.2, 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 2.4.2: Deterministic structural counterfactuals

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. $\mu_{\langle s'|s \rangle} = T_{\langle s'|s \rangle}^* \# \mu_s$;
3. $\pi_{\langle s'|s \rangle}^* = (I \times T_{\langle s'|s \rangle}^*) \# \mu_s$.

We say that $T_{\langle s'|s \rangle}^*$ is a *structural counterfactual operator*, and identify $\mathcal{T}^* := \{T_{\langle s'|s \rangle}^*\}_{s, s' \in \mathcal{S}}$ to the deterministic structural counterfactual model $\Pi^* = \{(I \times T_{\langle s'|s \rangle}^*) \# \mu_s\}_{s, s' \in \mathcal{S}}$.

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 counterfactual instances for *any* changes on S are deterministic, leading to a fully-deterministic counterfactual model. However, according to Proposition 2.4.1, 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.

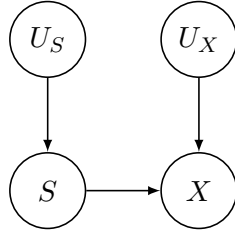


Figure 2.3: DAG of a structural causal model satisfying **(RE)**. The nodes U_S , U_X and X possibly represent several variables.

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 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): Relative exogeneity

$$U_S \perp U_X \text{ and } X_{\text{Endo}(S)} = \emptyset.^a$$

^aThese conditions resemble Assumption **(O)** introduced for the identification of potential outcomes. Note that **(RE)** interprets *all* nonintervened endogenous variables as outcomes of S , whereas **(O)** divides the nonintervened variables into two groups: covariates and outcomes.

The first item, $U_S \perp U_X$, 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 $(U_{\text{Exo}(i)})_{i \in \mathcal{I}}$ are independent (Shimizu et al., 2006; Karimi et al., 2021; Bongers et al., 2021; Dominguez-Olmedo et al., 2022). The second item, $X_{\text{Endo}(S)} = \emptyset$, means that S is *ancestrally closed*: no variable in X is a direct cause (or parent) of S (see Figure 2.3). This holds typically in fairness problems, such as in Section 2.3.4, where the variable S to alter generally encodes someone’s sex, race or age, which do not have any observable causes. As pointed out by Fawkes et al. (2022), 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 $S \perp U_X$, which can be interpreted as blinding S to the latent features (similarly to randomization). Then, it is easy to see that at the distributional level, intervening on S amounts to conditioning X by a value of S .

Proposition 2.4.3: Ignorability of structural counterfactuals

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* nonobservable counterfactual distribution $\mu_{\langle s'|s \rangle}$. According to Proposition 2.4.3, **(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 2.3.4. 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. (2022), a practical consequence of **(RE)** is that it enables to link observational and causal notions of fairness. In Section 2.5, 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 2.4.4: Exogenously-induced structural counterfactuals

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) $t_{\#} \pi_{\langle s'|s \rangle}^* = \pi_{\langle s|s' \rangle}^*$.

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

- (iii) $T_{\langle s'|s \rangle, \#}^* \mu_s = \mu_{s'}$;
- (iv) The operator $T_{\langle s'|s \rangle}^*$ is invertible μ_s -almost everywhere, such that $\mu_{s'}$ -almost everywhere $T_{\langle s'|s \rangle}^{*-1} = T_{\langle s|s' \rangle}^*$.

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 2.4.4 is that items (ii) and (iv) 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.

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 2.4.2: Linear additive causal model

Under **(RE)** and **(A)**, a *linear additive* SCM is characterized by the structural equations

$$\begin{aligned} X &\stackrel{\mathbb{P}\text{-a.s.}}{=} MX + wS + b + U_X, \\ S &\stackrel{\mathbb{P}\text{-a.s.}}{=} U_S, \end{aligned}$$

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 $U_S \perp U_X$. Solving the equations we get $X \stackrel{\mathbb{P}\text{-a.s.}}{=} (I - M)^{-1}(wS + b + U_X) =: F(S, U_X)$. 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 **2.3.4**.

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.

2.4.2 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 **(2.3)** between all pairs of factual distributions. That is, for every $s, s' \in \mathcal{S}$,

$$T_{\langle s'|s \rangle} := \arg \min_{T \in \mathcal{T}(\mu_s, \mu_{s'})} \int_{\mathcal{X}_s} \|x - T(x)\|^2 d\mu_s(x). \quad (2.5)$$

As explained before in Section **2.3**, 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 2.4.1: Optimal-transport identification of structural counterfactuals

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 **(2.5)**.

The mass-transportation formalism of Pearl’s counterfactual reasoning introduced in Section 2.3.2 and developed in Section 2.4.1 renders the proof of this theorem straightforward. The nontriviality 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 2.2.1. Thus, for the sake of illustration and clarity, we reproduce it directly below.

Proof According to (I) and Proposition 2.4.2, 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 2.4.3 that $\mu_{\langle s'|s \rangle} = \mu_{s'}$. Therefore, $T_{\langle s'|s \rangle}^* \mu_s = \mu_{s'}$. 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 2.2.1 it is the solution to Problem (2.3) between μ_s and $\mu_{s'}$, that the solution to Problem (2.5). ■

Understanding the strengths and limitations of Theorem 2.4.1 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 2.4.1. For instance, as the structural counterfactual operator from Example 2.4.2 is the gradient of a convex function, we obtain the following corollary.

Corollary 2.4.1: The case of linear models

Consider a linear additive SCM satisfying (RE) (see Example 2.4.2). 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 (2.3) 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 2.3.4. Besides, the scope of Theorem 2.4.1 goes beyond linear additive SCMs, as shown in the following nonlinear nonadditive example.

Example 2.4.3: Nonlinear nonadditive model

Consider the following SCM,

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

where α, β_1, β_2 are \mathbb{R} -valued functions such that $\alpha > 0$, $U_1 > 0$, and $U_S \perp (U_1, U_2)$. It satisfies **(A)**, **(I)** and **(RE)**, such that for any $s, s' \in \mathcal{S}$, the associated structural counterfactual operator is given by,

$$T_{\langle s'|s \rangle}^*(x) = \frac{\alpha(s')}{\alpha(s)}x + [\beta(s') - \beta(s)],$$

where $\beta = (\beta_1, \beta_2)$ is \mathbb{R}^2 -valued. This is the gradient of the convex function $x \mapsto \frac{\alpha(s')}{2\alpha(s)}\|x\|^2 + [\beta(s') - \beta(s)]^T x$. 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 **(2.3)** between μ_s and $\mu_{s'}$.

Note that the converse of the implication in Theorem **2.4.1** 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 2.4.4: Counterexample

Consider the following SCM,

$$\begin{cases} X_1 \stackrel{\mathbb{P}\text{-a.s.}}{=} U_1, \\ X_2 \stackrel{\mathbb{P}\text{-a.s.}}{=} SX_1^2 + U_2, \\ S \stackrel{\mathbb{P}\text{-a.s.}}{=} U_S, \end{cases}$$

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

$$T_{\langle s'|s \rangle}^*(x_1, x_2) = (x_1, x_2 + (s' - s)x_1^2).$$

It cannot be written as the gradient of a function. Consequently, it is not a solution to **(2.3)**.

Through Section **2.4**, we aimed notably at justifying the pertinence of optimal transport in counterfactual frameworks on top of the insights and illustrations given in Section **2.3**. 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 **2.4.1**, 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 **2.4.1** and Corollary **2.4.1** 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.

2.5 Transport-based counterfactual fairness

The strength of the unified mass-transportation viewpoint of counterfactual reasoning we proposed in Section 2.3 and further studied in Section 2.4 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 sex) 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 2.3 to reformulate the accepted *counterfactual fairness* condition (Kusner et al., 2017). On the basis of the reformulation, we then propose new fairness criteria derived from transport-based counterfactual models.

2.5.1 Causal counterfactual fairness from a mass-transportation viewpoint

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

Definition 2.5.1: Counterfactual fairness

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 2.5.1: Mass-transportation viewpoint of counterfactual fairness

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 2.5.1 are variations of the first item under the implications of **(RE)** and **(I)** through respectively Propositions 2.4.4 and 2.4.2. 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 2.5.2: Counterfactual fairness entails statistical parity

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.

2.5.2 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 2.3.

Definition 2.5.2: Transport-based counterfactual fairness

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 2.3.3) that only half of the couplings are truly required in the above conditions. Besides, because the proof of Proposition 2.5.2 only relies on the assumption that the couplings have factual distributions for marginals, the following proposition automatically holds.

Proposition 2.5.3: Π -counterfactual fairness entails statistical parity

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 2.5.1 and Proposition 2.5.1, Definition 2.5.2 relies on computationally feasible counterfactual models that obviate any assumptions about the data-generation process. In addition, as Definition 2.5.1 amounts to Π^* -counterfactual fairness (when (RE) holds), one can as well think of Definition 2.5.2 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 2.6.

2.5.3 Counterfactual fairness as nonarbitrary statistical parity

We conclude this section by discussing in more details the interest of counterfactual fairness compared to statistical parity. While it was taking into account causal links rather than correlations that mainly motivated the original definition of counterfactual fairness, a less appreciated interest—shared with transport-based counterfactual fairness—is that it amounts to statistical parity at the individual level.

Recall that statistical parity in binary classification consists in attributing the same proportion of positive outputs across protected groups. However, there is no further constraint on the allocation: statistical parity concerns only the *relative proportions* of individuals

receiving a positive answer, not the *total number* of individuals nor *which* individuals. This critically means that this fairness notion is arbitrary, thereby unfair—who would call fair an allocation rule that can interchangeably favor or unfavor a same person? Naturally, we should also keep in mind that “fair” predictors are built in practice by also maximizing accuracy (fairness being either a constraint or a penalty), which restricts the set of admissible allocations. But the issue persists even when taking accuracy into account, as recently evidenced by Krco et al. (2023). After conducting a comparative study of state-of-the-art bias-mitigation strategies, they observed that predictors with equal performances in accuracy and group fairness targeted different individuals. More precisely, the impacted individuals change according to the debiasing procedure or even the random seed.

Transport-based counterfactual fairness can remedy to this issue. To understand this, note that the behaviour of a fair machine-learning model can be divided into two distinct features: how the predictions are *relatively* distributed across groups or counterfactual counterparts—which is a matter of fairness; what are the actual *values* of the predictions at each input—which is driven by accuracy. In particular, if \hat{Y} is a fair binary classifier (according to statistical parity or Π -counterfactual fairness) then $1 - \hat{Y}$ is also fair in the same sense but with a radically different behaviour. This means that, as such, the considered notions of fairness cannot *fully* rule out the overall arbitrariness of the decisions. However, counterfactual fairness revokes all arbitrariness *on the fairness side* by fixing the relative allocation of outputs at the individual scale. Let us provide a more mathematical interpretation of this statement through the proposition below:

Proposition 2.5.4: Arbitrariness of statistical parity

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.

Recall that Π -counterfactual fairness implies statistical parity according to Proposition 2.5.3. Proposition 2.5.4 states that (somehow) conversely, statistical parity in classification can always be associated to some transport-based counterfactual model Π . Therefore, enforcing Π -counterfactual fairness—for a chosen Π —can be viewed as *specifying* the counterfactual model underlying statistical parity. This is precisely this specification that makes this notion stronger than group fairness, and restrains the arbitrariness of fair rules. To sum-up, Π -counterfactually fair predictors—for a same Π —can make different decisions at given inputs, but they all make invariant decisions across the same counterfactual counterparts listed in Π .

Beware that Proposition 2.5.4 also raises an ethical issue, as practitioners could argue that their group-fair predictors are counterfactually fair according to some Π . However, recall that not all transport-based counterfactual models define legitimate notions of counterparts, leading to potentially unfair decisions at the subgroup or individual level. Avoiding fair washing hence requires practitioners to always be able to justify the counterfactual models they employ when not imposed by legal experts of the prediction task.

2.6 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.

2.6.1 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 $Y : \Omega \rightarrow \mathcal{Y} \subseteq \mathbb{R}$ 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 $\Pi := \{\pi_{\langle s'|s} \rangle\}_{s, s' \in \mathcal{S}}$ and a given weight $\lambda > 0$, we define the following *expected* risk on the predictors,

$$\mathcal{R}_{\mathcal{D}, \Pi, \lambda}(\theta) := \mathbb{E}[\ell(h_\theta(X, S), Y)] + \lambda \sum_{s \in \mathcal{S}} \mathbb{P}(S = s) \sum_{s' \neq s} \mathbb{E}[r_\theta(X_s, s, X_{s'}, s')], \quad (2.6)$$

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 $r_\theta(x, s, x', s') := \max\{0, |h_\theta(x, s) - h_\theta(x', s')| - \epsilon\}$ for some $\epsilon > 0$. In this chapter, 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 $\{\pi_{\langle s'|s}^n(i, j)\}_{i \in \mathcal{I}_s, j \in \mathcal{I}_{s'}}$, typically by solving Problem (2.4) as explained in Section 2.2.2. Then, we define the following *empirical*

risk,

$$\mathcal{R}_n(\theta) := \frac{1}{n} \sum_{i=1}^n \ell(h_\theta(x_i, s_i), y_i) + \lambda n_{s_i} \sum_{s' \neq s_i} \sum_{j \in \mathcal{I}_{s'}^n} \pi_{(s'|s_i)}^n(i, j) r_\theta(x_i, s_i, x_j, s'). \quad (2.7)$$

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.

2.6.2 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 2.6.1: Consistency of the regularized predictor

Suppose that for every pair of factual distributions, the Kantorovich problem (2.2) with cost $c(x, x') := \|x - x'\|^2$ admits a unique solution. Thus, we can define the counterfactual model $\Pi = \{\pi_{(s'|s)}\}_{s, s' \in \mathcal{S}}$ and its empirical counterpart $\Pi^n = \{\pi_{(s'|s)}^n\}_{s, s' \in \mathcal{S}}$ as, for every $s, s' \in \mathcal{S}$,

$$\pi_{(s'|s)} := \arg \min_{\pi \in \Pi(\mu_s, \mu_{s'})} \int_{\mathcal{X}_s \times \mathcal{X}_{s'}} \|x - x'\|^2 d\pi(x, x'), \quad (2.8)$$

$$\pi_{(s'|s)}^n \in \arg \min_{\pi \in \Sigma(n_s, n_{s'})} \sum_{i \in \mathcal{I}_s} \sum_{j \in \mathcal{I}_{s'}} \|x_i - x_j\|^2 \pi(i, j). \quad (2.9)$$

Now, let $\Phi : \mathcal{X} \times \mathcal{S} \rightarrow \mathbb{R}^p$ be a feature map such that for every $s \in \mathcal{S}$ 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 $\mathcal{X} \subseteq \{x \in \mathbb{R}^d \mid \|x\| \leq R\}$,
- (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,

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

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 2.6.1: On the assumptions

1. Uniqueness of the solution (2.8) 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 nonlinear 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 $\theta_n \xrightarrow[n \rightarrow +\infty]{a.s.} \theta^*$ (this additional step is detailed in the proof of Theorem 2.6.1).

2.7 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/lucadelara/PI-Fair>.

2.7.1 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 model 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 (nonregularized) 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.

Evaluation metrics

In what follows, $h : \mathcal{X} \times \mathcal{S} \rightarrow \mathcal{Y}$ 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

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

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

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

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

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

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

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

Note that this extends the parity gap to the continuous case. The purpose of these two group-fair indicators is to illustrate Proposition 2.5.3, 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

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

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 $s' \neq s$,

$$\text{CFT}_\epsilon(h, x, s, s', \Pi) \geq 1 - \delta. \quad (2.10)$$

Dataset	Adult	COMPAS	Law	Crimes
Task	Classification	Classification	Regression	Regression
$S : 0/1$	Woman/Man	Black/White	Black/White	Black/Nonblack
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 2.1: 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 nondeterministic—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),

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

This corresponds to the proportion of points satisfying Condition [2.10](#). In the classification setting we set $\epsilon = 0$ while in the regression setting we work with $\epsilon = \frac{1}{2} \mathbb{E} [|Y - Y'|]$ where Y' is an independent copy of Y .

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. These 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 $\Pi\text{-Fair}(\lambda)$ 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.

2.7.2 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 [2.1](#) summarizes information about each dataset after preprocessing.

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 *sex*: $S = 0$ 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 $n_{test} = 16,118$.

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.² 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 $n_{test} = 2,030$.

Law School

This is the dataset used in Section 2.3.4, 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.

²<https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>

Communities and crimes

The Communities and Crimes dataset can also be found in the UCI Machine Learning Repository (Dua and Graff, 2019). It contains socioeconomics, 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.

2.7.3 Results

The regularization weight λ takes successively all the values in a grid $\{10^{-4}, 10^{-3.5}, \dots, 10^1\}$. 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.

Trade-off between accuracy and fairness

Figures 2.4 to 2.7 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.

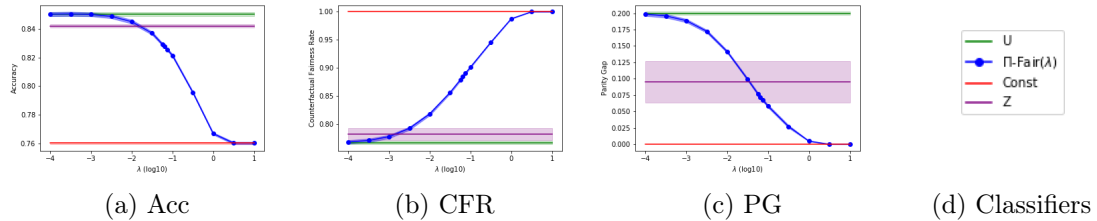


Figure 2.4: Evaluation metrics on the Adult dataset.

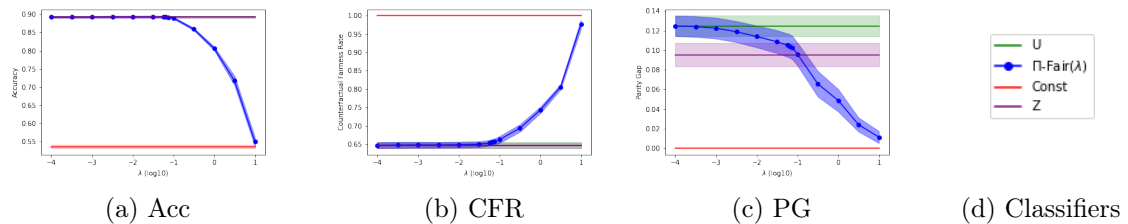


Figure 2.5: Evaluation metrics on the COMPAS dataset.

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

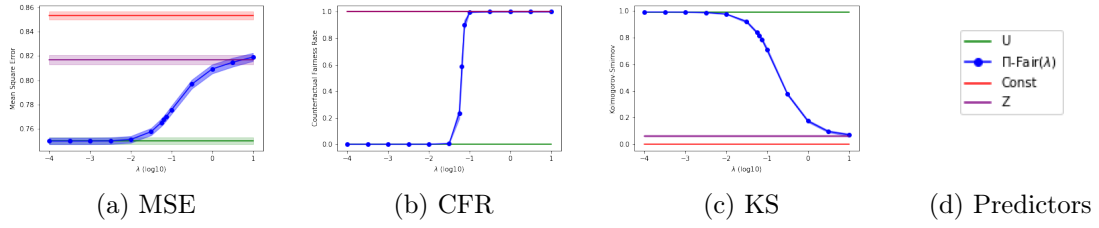


Figure 2.6: Evaluation metrics on the Law dataset.

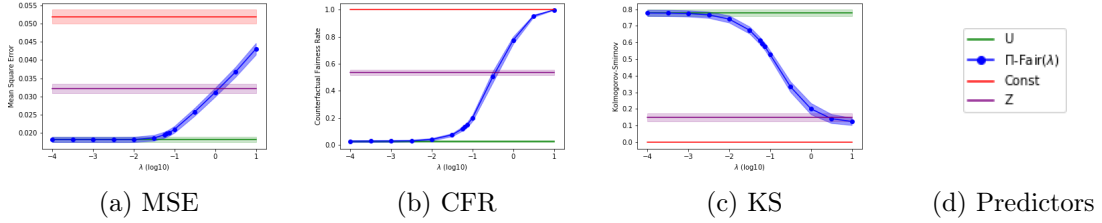


Figure 2.7: Evaluation metrics on the Crimes dataset.

evaluation metrics remain stable across the different repeats. As anticipated from Proposition 2.5.3, 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 2.6) 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.

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 2.3.4) and satisfies the assumptions of Corollary 2.4.1. Figure 2.8b 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 2.8a serves as a sanity check: it plots the normalized-variance indicator $\sqrt{\frac{\mathbb{E}[(\hat{Y} - \mathbb{E}[\hat{Y}])^2]}{\mathbb{E}[(Y - \mathbb{E}[Y])^2]}}$ to control how close a predictor is to being constant.

As anticipated by theory, the training process does promote causal counterfactual fairness: the two curves in Figure 2.8b are almost identical. Crucially, this is not a consequence of the predictors merely becoming constant, since the sequence of predictions in Figure 2.8a have variations that remain significantly higher than the best constant predictor.

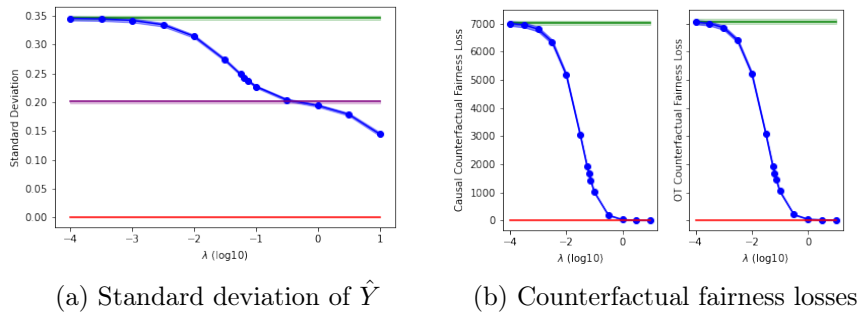


Figure 2.8: Promotion of causal counterfactual fairness on the Law dataset.

2.7.4 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), or neural-network-based approximations (see (Leygonie et al., 2019; Black et al., 2020; Makuva et al., 2020; Korotin et al., 2021; Huang et al., 2021) and Chapter 3) 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.

2.8 Perspectives of extensions

This concluding section summarizes our contributions and discusses the limitations and potential improvements of transport-based counterfactual models and more generally of any counterfactual approaches to fairness. The goal is to provide a better understanding of the potential merits and drawbacks of counterfactual fairness, and open new research tracks. It ends with a transition towards the second part of the manuscript.

2.8.1 Summary of the contributions

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 work, 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.

2.8.2 Improving transport-based counterfactual models

Definition 2.3.3 does not fully capture what could be expected of a counterfactual model. To illustrate this, let us further study the properties of *structural* counterfactual models.

Recall that items (ii) and (iv) of Proposition 2.4.4 ensure the reciprocity of structural counterfactual counterparts under (RE), which holds by construction in transport-based counterfactual models. This signifies concretely that the male counterfactual counterpart of Bob's female counterpart is Bob himself. But we could ask for a more general and equally desirable property: the commutativity of interventions, which we write as follows:

Definition 2.8.1: Commutativity of interventions

Let $\Pi := \{\pi_{\langle s'|s \rangle}\}_{s,s' \in \mathcal{S}}$ be a transport-based counterfactual model^a on X with respect to S . We say that Π is *commutative* if there exists a tuple of variables $(X_s)_{s \in \mathcal{S}}$ such that for any $s, s' \in \mathcal{S}$, $\mathcal{L}((X_s, X_{s'})) = \pi_{\langle s'|s \rangle}$. If Π is deterministic, identifiable to a collection of operators $\{T_{\langle s'|s \rangle}\}_{s,s' \in \mathcal{S}}$, this means that $T_{\langle s''|s \rangle}^* = T_{\langle s''|s' \rangle} \circ T_{\langle s'|s \rangle}$ μ_s -almost everywhere for $s, s', s'' \in \mathcal{S}$.

^aRecall that under (RE) a structural counterfactual model is a transport-based counterfactual model.

Under a commutative counterfactual model, the European counterfactual counterpart of an American person is the European counterfactual counterpart of the Asian counterfactual counterpart of this American person. Interestingly, structural counterfactual models are commutative under (RE), since the tuple $(X_{S=s})_{s \in \mathcal{S}}$ satisfies Definition 2.8.1 when $S \perp U_X$. However, the definition of transport-based counterfactuals (Definition 2.3.3) does not include this property. Notably, as soon as the cardinality of \mathcal{S} exceeds 2, a family of couplings $\{\pi_{\langle s'|s \rangle}\}_{s,s' \in \mathcal{S}}$ does not necessarily verify the above definition. Moreover, typical mass-transportation frameworks will not yield commutative couplings. In particular, the composition of two gradients of convex functions does not always yields the gradient of a

convex function, implying that the optimal-transport counterfactual model defined by (2.5) does not satisfy Definition 2.8.1 in general. This means that our construction of counterfactual models fails to capture an essential characteristic of structural ones.

This flaw motivates a new recipe for *commutative* transport-based counterfactuals. The narrative thread leading to the definition of transport-based counterfactual model rested on the observation that reasoning counterfactually with SCMs amounted to manipulating couplings between distributions. While noteworthy, this interpretation neglects a fundamental aspect of structural counterfactuals. In contrast to transport-based couplings, structural counterfactual couplings are not independently generated; under (RE), they stem from the same source U_X . Using our notations, this can be written as $\pi_{\langle s'|s \rangle}^* = (f_s \times f_{s'})_{\#} \mathcal{L}(U_X)$ for any $s, s' \in \mathcal{S}$.³ This is precisely the back and forth through the common distribution of exogenous variables that ensures the commutativity of the induced counterfactuals. Thus, we can improve Definition 2.3.3 by constructing transport-based counterfactual models through the following procedure:

1. Set a distribution of reference ν on \mathbb{R}^d , absolutely continuous with respect to the Lebesgue measure;
2. Using mass-transportation techniques, compute for any $s \in \mathcal{S}$ a function \tilde{f}_s such that $\tilde{f}_{s\#} \nu = \mu_s$;
3. Define the counterfactual couplings as $\pi_{\langle s'|s \rangle} = (\tilde{f}_s \times \tilde{f}_{s'})_{\#} \nu$ for any $s, s' \in \mathcal{S}$.

This is a mathematically well-posed construction since one can always define a (deterministic) mapping from a Lebesgue-absolutely-continuous measure towards another. Notably, the continuity of ν does not restrain generality; the input measures $\{\mu_s\}_{s \in \mathcal{S}}$ can be continuous, discrete, or neither of both. Whenever the mappings $\{\tilde{f}_s\}_{s \in \mathcal{S}}$ are invertible, the obtained couplings are deterministic. Commutativity holds since $(X_s)_{s \in \mathcal{S}} := (\tilde{f}_s(\tilde{U}_X))_{s \in \mathcal{S}}$ where $\mathcal{L}(\tilde{U}_X) = \nu$ satisfies Definition 2.8.1. All in all, we end up with an arguably more adequate model, fitting every requirement from Definition 2.3.3 plus commutativity. Moreover, any counterfactual model constructed as such can be generated by an SCM satisfying (RE), as stated below:

Proposition 2.8.1: Causal representation of transport

Let $\Pi := \{\pi_{\langle s'|s \rangle}\}_{s, s' \in \mathcal{S}}$ be a transport-based counterfactual in the sense of Definition 2.3.3 with \mathcal{S} finite and satisfying Definition 2.8.1. Then, there exists an SCM $\mathcal{M}^b := \langle U^b, G^b \rangle$ satisfying (A), (RE), and inducing a structural counterfactual model Π^b such that $\Pi^b = \Pi$.

Interestingly, this result provides together with Proposition 2.4.4 a characterization of the counterfactual models induced by SCMs verifying (RE), and a representation-like equivalence between structural counterfactual reasoning and the augmented version of transport-based counterfactual reasoning.

³We recall that $\pi_{\langle s'|s \rangle}^* = (f_s \times f_{s'})_{\#} \mathcal{L}(U_X | S = s)$ when (RE) does not hold.

From a computational viewpoint, note that this new recipe for transport-based counterfactuals requires solving only $|\mathcal{S}|$ transportation problems instead of $|\mathcal{S}|(|\mathcal{S}| - 1)/2$. Nonetheless, this does not necessarily entails a lower cost overall, since obtaining accurate solutions in this setting may demand a large sample from ν , increasing the computation time of every mapping. From an interpretability viewpoint, although opinions may diverge, the construction through an intermediate distribution of reference diminishes the understanding of the produced counterfactuals. In particular, optimal transport explained the pairing of instances between μ_s and $\mu_{s'}$ by the minimization of a global effort; now, pairs simply share an abstract ν value. Nevertheless, we can obtain meaningful interpretations by properly choosing ν and the $\{\tilde{f}_s\}_{s \in \mathcal{S}}$, giving them grounded significations. For example, setting ν as the spherical uniform distribution over the unit hypersphere and \tilde{f}_s as the optimal transport map for the quadratic cost between ν and μ_s coincides with the construction of multivariate center-outward quantiles proposed in (Hallin et al., 2021). This furnishes a concrete, understandable characterization of counterfactual counterparts: they live on the same quantile of their respective distributions.

To conclude, we underline that if $|\mathcal{S}| = 2$, then transport-based counterfactual models automatically satisfy Definition 2.8.1. Therefore, Definition 2.3.3 is adequate in many classical fairness problems addressing binary groups such as racialized/white or female/male. However, the new procedure to construct counterfactual models can still be useful, if for example one would prefer an interpretation of counterparts through quantiles rather than a global cost minimization.

2.8.3 The shape of counterfactual fairness constraints

We now turn to a detailed discussion on the *mathematical* (not conceptual) interest of achieving counterfactual fairness—be it causal-based or transport-based—rather than statistical parity. This question is motivated by our empirical observations in Section 2.7: the group-fair predictor had similar performances to the counterfactually fair predictor on the Law dataset. This subsection provides a rigorous analysis of this phenomenon by unpacking fairness constraints for simple predictors.

Predictor setup

Investigating the interest of achieving stronger fairness than statistical parity necessitates a class of predictive models that is: (1) simple enough to enable clear mathematical characterizations of fairness, (2) large enough for the evidences to shed light on more general cases. As in Section 2.7, we propose the class of predictors that are linear in their parameter. Let $\Phi : \mathcal{X} \times \mathcal{S} \rightarrow \mathbb{R}^p$ be a given mapping, and define

$$\mathcal{H} := \{h_{\theta_0, \theta} : (x, s) \mapsto \theta^T \Phi(x, s) + \theta_0 \mid \theta_0 \in \mathbb{R}, \theta \in \mathbb{R}^p\}.$$

The feature map Φ allows for modelling various decision rules, and typically corresponds to a kernel function. Of particular interest for our work are the cases of linear predictors, when $\Phi(x, s) := (x, s)$, and of unaware predictor, when $\Phi(x, s) := \bar{\Phi}(x)$ for some $\bar{\Phi} : \mathbb{R}^d \rightarrow \mathbb{R}^p$. The parameter θ represents the weight vector of the predictor, while θ_0 models an offset. Note that the offset could have been included in the feature map Φ , but we prefer this

formulation to emphasize its role. This setting is tailored to both regression and binary classification. In the case of classification, $h_{\theta_0, \theta}$ represents the decision boundary of the classifier $g_{\theta_0, \theta}(x, s) := \mathbf{1}_{\{h_{\theta_0, \theta}(x, s) \geq 0\}}$. Importantly, be it for statistical parity or counterfactual fairness, the fairness of the decision rule implies the one of the classifier. Note that the converse is not true in general.

Fairness constraints

While in Section 1.6 we wrote the fairness conditions for general predictors, we now specify them for the class \mathcal{H} . We consider three nested fairness constraints: no correlation, statistical parity, and counterfactual fairness. In what follows, $\text{Span}(E)$ denotes the *linear span* of a set of vectors $E \subseteq \mathbb{R}^p$ while F^\perp refers to the *orthogonal complement* of a linear subspace $F \subseteq \mathbb{R}^p$.

Recall that a predictor $h(X, S)$ satisfies statistical parity if $h(X, S) \perp S$. This requires in particular no correlation between the predictor and the protected status. For weight-linear predictions, the covariance between these variables can be expressed as $\text{Cov}(h_{\theta_0, \theta}(X, S), S) = \theta^T \mathbb{E}[\Phi(X, S)(S - \mathbb{E}[S])]$. Therefore, by defining

$$v_{NC} := \mathbb{E}[\Phi(X, S)(S - \mathbb{E}[S])], \quad (2.11)$$

we can characterize the set of weights θ ensuring no correlation as $\Theta_{NC} := \text{Span}(\{v_{NC}\})^\perp$, which describes a hyperplane of \mathbb{R}^p if $v_{NC} \neq 0$. Notice that this condition does not depend on the offset θ_0 . The next proposition provides an alternative formulation of v_{NC} , allowing for better interpretation of this orthogonality condition.

Proposition 2.8.2: No-correlation constraint

Let $\Phi : \mathcal{X} \times \mathcal{S} \rightarrow \mathbb{R}^p$ be a mapping with \mathcal{S} finite, and define v_{NC} as in (2.11). Then,

$$v_{NC} = \sum_{s > s'} \mathbb{P}(S = s)\mathbb{P}(S = s')(s - s') (\mathbb{E}[\Phi(X, S)|S = s] - \mathbb{E}[\Phi(X, S)|S = s']).$$

For simplicity, consider a case where S is binary. For instance, $S = 0$ stands for *female* while $S = 1$ stands for *male*. Then, it follows from Proposition 2.8.2 that

$$v_{NC} = \mathbb{P}(S = 1)\mathbb{P}(S = 0) (\mathbb{E}[\Phi(X, S)|S = 1] - \mathbb{E}[\Phi(X, S)|S = 0]).$$

Up to a multiplicative constant, v_{NC} is the mean translation vector from the female features to the male features. Hence, being fair in the sense of no correlation means having a weight that is orthogonal to the mean translation between protected groups. Figure 2.9 provides an illustration in linear classification.

In the case where approximated fairness is tolerated, one can leverage the absolute value of the covariance between the predictor and the protected attribute to measure unfairness. This corresponds to the following score:

$$\text{NC}(\theta) := |\theta^T v_{NC}|.$$

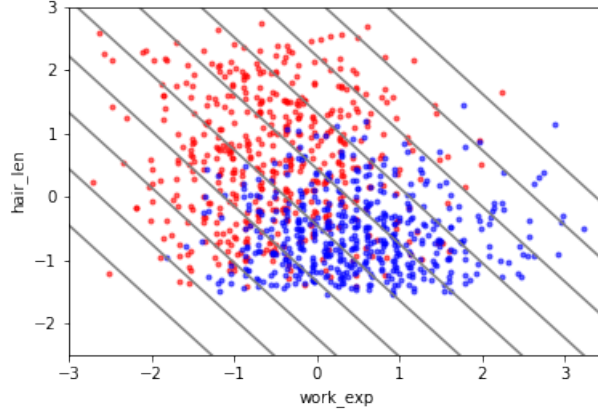


Figure 2.9: Fair decisions boundaries for the synthetic dataset from [Lipton et al. \(2018\)](#). We set $\Phi(x, s) := x$ so that the predictors are linear and unaware of the protected attribute. The female population is in red while the male population is in blue. The dark lines corresponds the boundaries of to linear decision rules satisfying the orthogonality condition. They all satisfy no correlation but have different accuracy rates.

For the counterfactual-fairness constraint, we consider a counterfactual model $\Pi := \{\pi_{(s'|s)}\}_{s,s' \in \mathcal{S}}$ (either structural or transport-based) and define

$$\Pi\Phi := \{\Phi(x, s) - \Phi(x', s') \mid (x, x') \in \text{supp}(\pi_{(s'|s)}^*), (s, s') \in \mathcal{S}^2\}, \quad (2.12)$$

which is the set of all possible counterfactual changes in the features when intervening on the protected attribute. Then, we prove that the set of counterfactually fair weights θ is $\Theta_{CF} := \text{Span}(\Pi\Phi)^\perp$.

Proposition 2.8.3: Counterfactual-fairness constraint

Let $\Phi : \mathcal{X} \times \mathcal{S} \rightarrow \mathbb{R}^p$ be a mapping such that $\Phi(\cdot, s)$ is continuous for every $s \in \mathcal{S}$, $(\theta, \theta_0) \in \mathbb{R}^p \times \mathbb{R}$ a parameter, and $\Pi := \{\pi_{(s'|s)}\}_{s,s' \in \mathcal{S}}$ a transport-based counterfactual model. The predictor $\hat{Y} := \theta^T \Phi(X, S) + \theta_0$ is Π -counterfactually fair if and only if,

$$\theta \in \text{Span}(\Pi\Phi)^\perp,$$

where $\Pi\Phi$ is defined as in [\(2.12\)](#).

Observe that this condition does not involve θ_0 either.

The statistical parity constraint is more difficult to write. Nonetheless, it is easy to see that the set of weights θ satisfying statistical parity Θ_{SP} is a linear subspace of \mathbb{R}^p . In addition, because counterfactual fairness implies statistical parity and statistical parity implies no correlation, we have $\Theta_{CF} \subseteq \Theta_{SP} \subseteq \Theta_{NC}$. Next, we use these inclusions along with the introduced formulations of the fairness constraints to discuss the tensions that exist between them.

Consequences

For this analysis, we exclude the trivial case where $v_{NC} = 0$. Remark that in contrast to no correlation, counterfactual fairness requires orthogonality to a linear space of dimension possibly higher than 1. The richer $\Pi\Phi$, the smaller Θ_{CF} . In particular, if the dimension of $\text{Span}(\Pi\Phi)$ reaches p , then $\Theta_{CF} = \{0\}$, meaning that the only possible fair predictors in \mathcal{H} narrow down to the constant functions. Therefore, if the counterfactual model Π and the features Φ are not well adapted to each others, counterfactual fairness ends up undesirable. More specifically, this questions the choice of the feature map Φ : should it be chosen with accuracy in mind and no care for the underlying counterfactual model Π , it would render being counterfactually fair to being merely constant.

Moreover, we know from $\Theta_{CF} \subseteq \Theta_{NC}$ that $v_{NC} \in \text{Span}(\Pi\Phi)$. Consequently, if $\dim(\text{Span}(\Pi\Phi)) = 1$, then $\Theta_{CF} = \Theta_{NC}$, and therefore, $\Theta_{CF} = \Theta_{SP}$. In this case, counterfactual fairness and statistical parity are equivalent, implying that enforcing counterfactual fairness is useless. Actually, because the codimension of Θ_{SP} is in general larger than one, this equivalence could be attained even with $\dim(\text{Span}(\Pi\Phi)) > 1$. However, except for specific but standard choices of Π and Φ that we exemplify below, the dimension of $\text{Span}(\Pi\Phi)$ is likely to largely exceed 1.

Example

To illustrate the above tension, we must figure out a scenario where $\Pi\Phi$ can be explicitly computed. We set $\Phi(x, s) := (x, s)$ (linear predictions) and assume that $\mathcal{S} = \{0, 1\}$. If the features are continuous and generated by a linear causal model $\mathcal{M} := \langle U, G \rangle$ (see Example 2.4.2) satisfying **(RE)**, then the counterfactual counterparts of the population $S = 0$ in population $S = 1$ are given by a mapping of the form $T_{\langle 1|0 \rangle}(x) := x + \Delta_{\langle 1|0 \rangle}$ where $\Delta_{\langle 1|0 \rangle}$ is a constant depending on the parameters of G .

Then, imposing the predictor to be counterfactually fair means that for any μ_0 -almost every $x \in \mathcal{X}_0$,

$$\theta^T [(x, 0) - (T_{\langle 1|0 \rangle}(x), 1)] = 0.$$

As a set of probability one is necessarily nonempty, this leads to

$$\theta^T (\Delta_{\langle 1|0 \rangle}, 1) = 0.$$

The above equation characterizes the space of counterfactually fair parameters θ . Typically, we would look within this space for the parameters that maximizes accuracy. But more importantly, remark that $\Delta_{\langle 1|0 \rangle}$ is simply the mean translation between μ_0 and μ_1 by definition of $T_{\langle 1|0 \rangle}$. Besides,

$$\begin{aligned} v_{NC} &:= \mathbb{P}(S = 1)\mathbb{P}(S = 0) (\mathbb{E}[(X, 1) | S = 1] - \mathbb{E}[(X, 0) | S = 0]) \\ &= \mathbb{P}(S = 1)\mathbb{P}(S = 0) (\mathbb{E}[X | S = 1] - \mathbb{E}[X | S = 0], 1 - 0) \\ &= \mathbb{P}(S = 1)\mathbb{P}(S = 0)(\Delta_{\langle 1|0 \rangle}, 1) \end{aligned}$$

according to Proposition 2.8.2. Therefore, $v_{CF} := (\Delta_{\langle 1|0 \rangle}, 1)$ is colinear to v_{NC} . This critically implies that counterfactual fairness and statistical parity are equivalent in this scenario. We

have just showed in this particular setting what we explained above: as the set $\text{Span}(\Pi\Phi)$ has a dimension equal to 1, $\Theta_{SP} = \Theta_{CF}$.

The same phenomenon occurs when thinking in terms of approximate fairness. In a similar fashion to no correlation, counterfactual fairness in this setting can be quantified using,

$$\text{CF}(\theta) := |\theta^T v_{CF}|.$$

Remark that $\text{NC}(\theta) = \mathbb{P}(S = 0)\mathbb{P}(S = 1)\text{CF}(\theta)$ in this example. This means that, up to a multiplicative constant, the notions are equivalent in terms of approximate fairness. In particular, increasing the fairness with respect to one criterion necessarily increases the fairness with respect to the other.

To sum-up—in the specific situation we considered—working with statistical parity or counterfactual fairness is absolutely equivalent, rendering pointless the use of the causal model. This is exactly what happened for the Law dataset: as the data fits a linear additive SCM, the counterfactual model generated by optimal transport for the quadratic cost produces translation mapping according to Corollary 2.4.1. In addition, we considered linear predictors, making the perfectly group-fair predictor also counterfactually fair. Nevertheless, this does not question the utility of counterfactual fairness in general: obviously, most of the real-world datasets do not fit linear generative processes, and practitioners often implement nonlinear predictive models. But this provides a better understanding of the shapes and boundaries of counterfactual fairness.

2.8.4 Interpretation and applicability of counterfactuals for fairness

Lastly, we must mention a neglected critical limitation of structural counterfactuals recently pointed out by Fawkes et al. (2022), which deserves a special attention from the fairness community. As we explain below, this also concerns transport-based counterfactuals.

Consider the Law dataset for the sake of illustration. To address questions of the form “Had they been black, would have they been predicted a lower score?” we computed the black counterfactual counterparts of a white student. But beware: we computed the counterparts *within the dataset*. This implicitly assumes that had the white student been born black, they would have pursued law studies, getting a GPA and taking the LSAT. It is obviously false in most scenarios, as inequalities across races induce disparate opportunities and incitations. So, what is the problem in our model? Mathematically, we justified that counterfactual counterparts (be them causal-based or transport-based) were observable, therefore ended up in the dataset, by Proposition 2.4.3 which requires Assumption (RE)—namely independent noises plus ancestral closure of the protected attribute S . There is nothing wrong in considering race to be ancestrally closed; the issue is assuming the noises independent. More precisely, $S \perp\!\!\!\perp U_X$ would mean that race is equally distributed over units. While this seems valid over the whole population, this is unlikely to hold on a specific population of law students. As formalized by Fawkes et al. (2022), the general idea is that accessing a dataset produces a selection bias: the sensitive attribute is not guaranteed to be independent of the random seed. In addition, because fairness problems cannot be randomized as controlled experiments, we cannot even produce studies for which $S \perp\!\!\!\perp U_X$ by design.

This signifies that the *true* SCM generating the dataset—not the whole population—includes a dependence between U_X and U_S , which violates (RE). Computing the “correct”

counterfactuals is still conceptually possible but requires to apply the three-step procedure to this true causal model (should we know it). This raises two issues in practice: a mere but consequential numerical challenge, and a more profound obstacle. First, many causal-discovery techniques work under the independent-noise assumption, thereby are unreliable for inferring the true SCM. This complicates even more the learning of the causal model, making even less feasible the causal approach to counterfactual reasoning. Second, even if the authentic generative process of the dataset is accessible, the produced counterfactuals may be useless for the prediction task. The fact that counterparts end up outside of the dataset not only means that $\mu_{(s'|s)} \neq \mu_s$ for every $(s, s') \in \mathcal{S}^2$, but also that some variables could be undefined at these instances. As mentioned, had one been born with a different race, there is no guarantee that they would have passed the LSAT. This definition problem could perhaps be handled by adding **None** to the attainable values. Nevertheless, the obtained counterfactuals would still be inoperative in most machine-learning tasks: what would be the meaning of predicting the grades of someone who does not attend law school? All in all, requiring equal treatment between an instance and its true causal counterfactual counterparts seems to be an intrinsically ill-posed query. Regarding the transport-based approach, should we care to mimic the correct behaviour of structural counterfactual reasoning, where “had they been black” means “had they been *born* black”, we would have to define counterfactual couplings as joint distributions between observable marginals of factual instances and unobservable marginals of counterfactual instances. This obviously cannot be achieved by a purely data-based approach; it requires additional knowledge.

Whether this issue renders within-dataset counterfactuals, hence counterfactual reasoning as typically applied in fairness, illegitimate is a question that must be addressed by the community. One could still argue that doing comparisons within the dataset makes sense, but we must crucially keep in mind that this transcribes a different principle to “had they been *born* different”. Note also that if **(RE)** does not hold, then structural counterfactual fairness does not imply statistical parity. Therefore, in real situations, statistical parity and causal-based counterfactual fairness are incompatible. This means in particular that “true” counterfactual fairness tolerates unbalanced acceptance rates across protected groups in binary classification tasks. On the contrary, transport-based counterfactual fairness always entails statistical parity, as it compares observable distributions. As we argued in Section 2.5.3, having stronger notions than statistical parity is crucial to limit the arbitrariness of group fairness; this is why we believe transport-based counterfactuals (which are necessarily within the dataset) to still be a valuable tool for implementing fairness.

2.8.5 Transport-based counterfactuals: from theory to practice

Transport-based counterfactual models, as introduced in Definition 2.3.3, are generally theoretical objects. In practice, one does not work with the true couplings but estimations from data, as in Sections 2.6 and 2.7. Such estimations can take various forms. In the case of optimal transport for instance, there exist natural empirical formulations of Monge and Kantorovich problems which admit exact solutions, expressed as matrices (see Problem 2.4). This is notably what we used in our numerical experiments.

However, note that these approximations of the true couplings cannot generalize to new out-of-sample inputs; they can only match the points on which there were constructed. This

can be limiting in practice. In particular, applying the fairness-auditing technique of [Black et al. \(2020\)](#) on different samplings from a same dataset or implementing mini-batch learning in [Section 2.7](#) would require to recompute the couplings at each batch of observations. This issue also concerns other optimal-transport applications such as domain adaptation and transfer learning, which led many researchers to develop approximations of optimal transport plans (especially maps) defined on the whole sample space. Notably, we used a solution based on [\(Ferradans et al., 2014\)](#) to compute the mappings in [Section 2.3.4](#).

Nevertheless, these approximations do not always come with statistical guarantees: on the contrary to the empirical solutions, it is not certified that they converge to the true couplings as the sample size tends to infinity. As explained in more details in the beginning of [Chapter 3](#), the literature has mostly addressed either theoretically grounded statistical estimators of optimal transport maps, but unsuitable for large-scale implementations, or efficient heuristic approximations, at the cost of statistical proofs. As statisticians, we believe such guarantees to be vital.

Ideally, we would like counterfactual models that can be computed easily and efficiently, able to generalize to unseen observations, all the while being statistically consistent. This is precisely what motivated the work presented in the next part of this manuscript. We made three attempts to design such models over the last three years: two dealing with optimal transport maps ([\(De Lara et al., 2021b\)](#); [\(González-Sanz et al., 2022\)](#)), one with diffeomorphic registration ([\(De Lara et al., 2023\)](#)). The third and fourth chapter focus on these works. For transparency, we must say that, even though we believe that they are valuable contributions, none of them happened to be truly satisfactory for counterfactual computations. Consequently, for our approach to counterfactual reasoning to be versatile, we need to carry out further research in the statistical estimation and practical implementation of transport models.

Appendix 2.A Proofs of Section 2.4

Proof of Lemma 2.4.1 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 $\text{Endo}(k) = \emptyset$ or $\text{Endo}(k) = \{S\}$, we can recursively replace the terms $X_{\text{Endo}(k)}$ in the formulas $G_k(X_{\text{Endo}(k)}, S_{\text{Endo}(k)}, U_{\text{Exo}(k)})$ by expressions depending only on U_X and S . This yields a measurable function F such that $X \stackrel{\mathbb{P}\text{-a.s.}}{=} F(S, U_X)$. The same computation but changing S to s for some $s \in \mathcal{S}$ leads to $X_{S=s} \stackrel{\mathbb{P}\text{-a.s.}}{=} F(s, U_X)$. ■

Proof of Proposition 2.4.1 Recall that $X \stackrel{\mathbb{P}\text{-a.s.}}{=} F(S, U_X)$. This implies that, \mathbb{P} -almost surely, $(X = x, S = s) \implies U_X \in f_s^{-1}(\{x\})$. Besides, $X_{S=s'} \stackrel{\mathbb{P}\text{-a.s.}}{=} f_{s'}(U_X)$ according to Lemma 2.4.1. Then, let $E \subseteq \mathcal{X}$ be an arbitrary Borel set and compute:

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

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

Proof of Proposition 2.4.2 Set $s, s' \in \mathcal{S}$ and $x \in \mathcal{X}_s$. Note that, according to (I), $U_X \stackrel{\mathbb{P}\text{-a.s.}}{=} f_S^{-1}(X)$. Let us address each item of the proposition separately.

• **Item 1.** Proposition 2.4.1 states that $\text{supp}(\mu_{\langle s'|s \rangle}(\cdot|x)) \subseteq \overline{f_{s'} \circ f_s^{-1}(\{x\})}$ for μ_s -almost every $x \in \mathcal{X}_s$. This means according to (I) that $\text{supp}(\mu_{\langle s'|s \rangle}(\cdot|x)) \subseteq \{f_{s'} \circ f_s^{-1}(x)\}$. 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

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

This proves the second item.

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

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

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

Proof of Proposition 2.4.3 Set $s \in \mathcal{S}$ and recall that $X \stackrel{\mathbb{P}-a.s.}{=} F(S, U_X)$ while $X_{S=s} \stackrel{\mathbb{P}-a.s.}{=} F(s, U_X)$. Thanks to Assumption **(RE)**, we have that $S \perp U_X$. Therefore,

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

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

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

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

Proof of Proposition 2.4.4 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 **2.4.3**.

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

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

- **Item (iii).** It is a direct consequence of $T_{\langle s'|s \rangle}^* \mu_s = \mu_{\langle s'|s \rangle}$ from Proposition **2.4.2** and $\mu_{\langle s'|s \rangle} = \mu_{s'}$ from Proposition **2.4.3**.

- **Item (iv).** We know according to Lemma **2.4.1** that $X_{S=s} \stackrel{\mathbb{P}\text{-a.s.}}{=} f_s(U_X)$ and $X_{S=s'} \stackrel{\mathbb{P}\text{-a.s.}}{=} f_{s'}(U_X)$. Furthermore, it follows from **(RE)** and Proposition **2.4.3** that $\mu_s = \mathcal{L}(X_{S=s})$ and $\mu_{s'} = \mathcal{L}(X_{S=s'})$. Wrapping this up, there exists a measurable set $\Omega^* \subseteq \Omega$ with $\mathbb{P}(\Omega^*) = 1$ such that for every $\omega \in \Omega^*$,

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

In the rest of the proof we implicitly work on an $\omega \in \Omega^*$. Assumption **(I)** ensures that $U_X = f_s^{-1}(X_{S=s})$ so that $X_{S=s'} = (f_{s'} \circ f_s^{-1})(X_{S=s})$. Noting that $X_{S=s} \in \mathcal{X}_s$, we obtain $X_{S=s'} = (f_{s'} \circ f_s^{-1}|_{\mathcal{X}_s})(X_{S=s}) = T_{\langle s'|s \rangle}^*(X_{S=s})$. Following the same computation after switching s and s' , we additionally get that $X_{S=s} = (f_s \circ f_{s'}^{-1}|_{\mathcal{X}_{s'}})(X_{S=s'}) = T_{\langle s|s' \rangle}^*(X_{S=s'})$.

Therefore, $T_{\langle s'|s \rangle}^*$ is invertible on $X_{S=s'}(\Omega^*)$ such that $T_{\langle s'|s \rangle}^{*-1} = T_{\langle s|s' \rangle}^*$ on $X_{S=s}(\Omega^*)$. Since $\mu_s(X_{S=s}(\Omega^*)) = \mathbb{P}(\Omega^*) = 1$ and $\mu_{s'}(X_{S=s'}(\Omega^*)) = \mathbb{P}(\Omega^*) = 1$, this means that $T_{\langle s'|s \rangle}^*$ is invertible μ_s -almost everywhere such that $T_{\langle s'|s \rangle}^{*-1} = T_{\langle s|s' \rangle}^*$ $\mu_{s'}$ -almost everywhere. This completes the proof. ■

Proof of Corollary **2.4.1** We address the structural equations

$$\begin{aligned}
X &\stackrel{\mathbb{P}\text{-a.s.}}{=} MX + wS + b + U_X, \\
S &\stackrel{\mathbb{P}\text{-a.s.}}{=} U_S,
\end{aligned}$$

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 $x \mapsto \frac{1}{2}\|x\|^2 + [(I - M)^{-1}w(s' - s)]^T x$. As **(RE)** holds and μ_s is Lebesgue-absolutely continuous with finite second order moment, it follows from Theorem **2.4.1** that $T_{\langle s'|s \rangle}^*$ is the solution to **(2.3)** between μ_s and $\mu_{s'}$.

■

Appendix 2.B Proofs of Section 2.5

Proof of Proposition 2.5.1 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 Borel set $E_{12} = E_{12}(s, s') \subseteq \mathcal{X} \times \mathcal{X}$ satisfying $\pi_{\langle s'|s \rangle}^*(E_{12}) = 1$ such that for every $(x, x') \in E_{12}$

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

Note that a direct reformulation of the original counterfactual fairness condition is

(CF) For every $s, s' \in \mathcal{S}$, there exists a Borel set $E_1 = E_1(s)$ satisfying $\mu_s(E_1) = 1$, such that for every $x \in E_1$ and every Borel set $E \subseteq \mathbb{R}$

$$\mathbb{P}(\hat{Y}_{S=s} \in E \mid X = x, S = s) = \mathbb{P}(\hat{Y}_{S=s'} \in E \mid X = x, S = s). \quad (2.13)$$

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 Borel set $E_1 = E_1(s)$ satisfying $\mu_s(E_1) = 1$, such that for every $x \in E_1$ and every Borel set $E \subseteq \mathbb{R}$ there exists a Borel set $E_2 = E_2(s, s', x, E)$ satisfying $\mu_{\langle s'|s \rangle}(E_2|x) = 1$ and such that for every $x' \in E_2$,

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

► **Proof that **(CF)** \iff **(IF)**.** Set $s, s' \in \mathcal{S}$, $x \in E_1$ and $E \subseteq \mathbb{R}$ measurable. 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 **(2.13)** as

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

Then, using Definition **2.3.1**, we reframe the right term of **(2.13)** as

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

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

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

is equivalent to $\mathbf{1}_{\{h(x,s) \in E\}} = \mathbf{1}_{\{h(x',s') \in E\}}$ for $\mu_{\langle s'|s \rangle}(\cdot|x)$ -almost every x' . This means that there exists a Borel set $E_2 = E_2(s, s', x, E)$ such that $\mu_{\langle s'|s \rangle}(E_2|x) = 1$ and for every $x' \in E_2$,

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

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

► **Proof that (IF) \implies (Goal).** As **(IF)** is true for any arbitrary Borel set $E \subseteq \mathbb{R}$, we can apply this result with $E = \{h(x, s)\}$ to obtain a Borel set $E_2 = E_2(s, s', x)$ such that $\mu_{\langle s'|s \rangle}(E_2|x) = 1$ and for every $x' \in E_2$, $h(x', s') = h(x, s)$. To sum-up, for every $s, s' \in \mathcal{S}$, there exists a Borel set $E_1 = E_1(s)$ satisfying $\mu_s(E_1) = 1$ such that for every $x \in E_1$, there exists a Borel set $E_2 = E_2(s, s', x)$ satisfying $\mu_{\langle s'|s \rangle}(E_2|x) = 1$, such that for every $x' \in E_2$, $h(x', s') = h(x, s)$. Now, we must show that the latter equality holds for $\pi_{\langle s'|s \rangle}^*$ -almost every (x, x') .

To this end, set $E_{12} = E_{12}(s, s') = \{(x, x') \in \mathcal{X} \times \mathcal{X} | x \in E_1(s), x' \in E_2(s, s', x)\}$. Remark that by definition of E_1 and E_2 , for every $(x, x') \in E_{12}$, $h(x, s) = h(x', s')$. To conclude, let us prove that $\pi_{\langle s'|s \rangle}^*(E_{12}) = 1$.

$$\begin{aligned} \pi_{\langle s'|s \rangle}^*(E_{12}) &= \int_{E_1} \mathbb{P}(X_{S=s'} \in E_2 | X = x, S = s) d\mu_s(x) \\ &= \int_{E_1} \mu_{\langle s'|s \rangle}(E_2|x) d\mu_s(x) \\ &= \int_{E_1} 1 d\mu_s(x) \\ &= \mu_s(E_1) \\ &= 1. \end{aligned}$$

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

► **Proof that (Goal) \implies (IF).** Using **(Goal)**, consider a Borel set $E_{12} = E_{12}(s, s')$ satisfying $\pi_{\langle s'|s \rangle}^*(E_{12}) = 1$ and such that for every $(x, x') \in E$, $h(x, s) = h(x', s')$. Then, define for any $x \in \mathcal{X}$, the Borel set $E_2(s, s', x) = \{x' \in \mathcal{X} | (x, x') \in E_{12}\}$. We use disintegrated formula of $\pi_{\langle s'|s \rangle}^*$ to write

$$1 = \int \mu_{\langle s'|s \rangle}(E_2|x) d\mu_s(x).$$

Since $0 \leq \mu_{\langle s'|s \rangle}(E_2|x) \leq 1$, this implies that for μ_s -almost every x , $\mu_{\langle s'|s \rangle}(E_2|x) = 1$. Said differently, there exists a Borel set $E_1 = E_1(s)$ satisfying $\mu_s(E_1) = 1$ such that for every $x \in E_1$, the Borel set $E_2(s, s', x)$ satisfies $\mu_{\langle s'|s \rangle}(E_2|x) = 1$. By construction of E_2 and by definition of E , for every $x \in E_1$ and every $x' \in E_2$, $h(x, s) = h(x', s')$. To obtain **(IF)**, it suffices to take any measurable $E \in \mathbb{R}$ and to note that the latter equality implies that $\mathbf{1}_{\{h(x,s) \in E\}} = \mathbf{1}_{\{h(x',s') \in E\}}$.

• **Item 2.** Recall that $\pi_{\langle s|s \rangle}^* = (I \times I)_{\#} \mu_s$. 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')

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

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

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

Moreover, **(RE)** implies through Proposition 2.4.4 that $\pi_{\langle s|s' \rangle}^* = t_{\#} \pi_{\langle s'|s \rangle}^*$. Therefore, the second condition above can be written as: for $\pi_{\langle s'|s \rangle}^*$ -almost every (x, x')

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

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 $s, s' \in \mathcal{S}$, μ_s -almost every x and every measurable $E \subseteq \mathbb{R}$ the left term of (2.13) is $\mathbf{1}_{\{h(x, s) \in E\}}$. Let us now reframe the right-term of (2.13). If **(I)** holds, using that $U_X \stackrel{\mathbb{P}\text{-a.s.}}{=} f_S^{-1}(X)$ we obtain

$$\begin{aligned} \mathbb{P}(\hat{Y}_{S=s'} \in E \mid X = x, S = s) &= \mathbb{P}(h(X_{S=s'}, s') \in E \mid X = x, S = s) \\ &= \mathbb{P}(h(F(s', U_X), s') \in E \mid X = x, S = s) \\ &= \mathbb{P}(h(f_{s'}(f_S^{-1}(X)), s') \in E \mid X = x, S = s) \\ &= \mathbb{P}(h(f_{s'} \circ f_S^{-1}(x), s') \in E) \\ &= \mathbb{P}(h(T_{\langle s'|s \rangle}^*(x), s') \in E) \\ &= \mathbf{1}_{\{h(T_{\langle s'|s \rangle}^*(x), s') \in E\}}. \end{aligned}$$

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

$$\mathbf{1}_{\{h(x, s) \in E\}} = \mathbf{1}_{\{h(T_{\langle s'|s \rangle}^*(x), s') \in E\}}.$$

Using the same reasoning as before, we take $E = \{h(x, s)\}$ 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 2.4.3, it follows that counterfactual fairness can be written as: for every $s, s' \in \mathcal{S}$ such that $s' < s$, for μ_s -almost every x

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

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

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

Set $s, s' \in \mathcal{S}$ such that $s' < s$. To prove the fourth item, we show as for item 2 that the two above conditions are equivalent. Set E_1 a measurable subset of \mathcal{X}_s such that $\mu_s(E_1) = 1$, and $h(x, s) = h(T_{\langle s'|s \rangle}^*(x), s')$ for any $x \in E_1$. Then, make the change of variable $x' = T_{\langle s'|s \rangle}^*(x)$ so that $h(T_{\langle s'|s \rangle}^{*-1}(x'), s') = h(x', s')$ for every $x' \in T_{\langle s'|s \rangle}^*(E_1)$. By Propositions 2.4.2 and 2.4.3, $T_{\langle s'|s \rangle}^* \mu_s = \mu_{s'}$, which implies that $\mu_{s'}(T_{\langle s'|s \rangle}^*(E_1)) = 1$. Therefore, the equality $h(T_{\langle s'|s \rangle}^{*-1}(x'), s) = h(x', s')$ holds for $\mu_{s'}$ -almost every x' . Finally, recall that according to Proposition 2.4.4, $T_{\langle s'|s \rangle}^{*-1} = T_{\langle s|s' \rangle}^*$ $\mu_{s'}$ -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 2.5.2 According to Proposition 2.5.1, 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 $\mathbf{1}_{\{h(x,s) \in E\}} = \mathbf{1}_{\{h(x',s') \in E\}}$ for every measurable $E \in \mathbb{R}$. 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 $E \subseteq \mathbb{R}$

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

This can be written as,

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

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,

$$X \stackrel{\mathbb{P}\text{-a.s.}}{=} S \cdot U_X + (1 - S) \cdot (1 - U_X).$$

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,

$$h(T_{(1|0)}^*(x)) = -h(x) \neq h(x).$$

■

Proof of Proposition 2.5.4 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 $y \in \mathcal{Y}$,

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

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

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

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

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

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

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

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

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

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

■

Appendix 2.C Proofs of Section 2.6

Proof of Theorem 2.6.1 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,

$$\mathcal{R}(\theta_n) - \min_{\theta \in \Theta} \mathcal{R}(\theta) \leq 2 \sup_{\theta \in \Theta} |\mathcal{R}_n(\theta) - \mathcal{R}(\theta)|. \quad (2.14)$$

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

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

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

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 $1 - \delta$,

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

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 almost every $\omega \in \Omega$, there exists a threshold $N(\omega)$ such that for any $n \geq N(\omega)$,

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

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

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

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,

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

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

$$\begin{aligned} & \sup_{\theta \in \Theta} \left| \left(\frac{n_s}{n} \pi_{\langle s'|s \rangle}^n - \mathbb{P}(S = s) \pi_{\langle s'|s \rangle} \right) (r_\theta(\cdot, s, \cdot, s')) \right| \\ & \leq \left| \frac{n_s}{n} - \mathbb{P}(S = s) \right| \sup_{\theta \in \Theta} \int r_\theta(x, s, x', s') d\pi_{\langle s'|s \rangle}(x, x') \end{aligned} \quad (2.15)$$

$$+ \mathbb{P}(S = s) \sup_{\theta \in \Theta} \left| \int r_\theta(x, s, x', s') \left(d\pi_{\langle s'|s \rangle}^n(x, x') - d\pi_{\langle s'|s \rangle}(x, x') \right) \right|. \quad (2.16)$$

The terms (2.15) tends to zero almost surely as n increases to infinity. We now turn to the convergence of the term (2.16).

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,

$$\begin{aligned} |r_\theta(x_1, s, x'_1, s') - r_\theta(x_2, s, x'_2, s')| & \leq |\theta^T (\Phi(x_1, s) - \Phi(x'_1, s') - \Phi(x_2, s) + \Phi(x'_2, s'))|^2 \\ & \leq |\theta^T (\Phi(x_1, s) - \Phi(x_2, s))|^2 \\ & \quad + |\theta^T (\Phi(x'_1, s') - \Phi(x'_2, s'))|^2, \\ & \leq \|\theta\|^2 \|\Phi(x_1, s) - \Phi(x_2, s)\|^2 \\ & \quad + \|\theta\|^2 \|\Phi(x'_1, s') - \Phi(x'_2, s')\|^2, \\ & \leq D^2 \left\{ L_s^2 \|x_1 - x_2\|^2 + L_{s'}^2 \|x'_1 - x'_2\|^2 \right\}, \\ & \leq D^2 \max_{s \in \mathcal{S}} L_s^2 \|(x_1, x'_1) - (x_2, x'_2)\|^2, \\ & \leq 4D^2 \max_{s \in \mathcal{S}} L_s^2 R^2 \|(x_1, x'_1) - (x_2, x'_2)\|. \end{aligned}$$

Let us set $\Lambda := 4D^2(\max_{s \in \mathcal{S}} L_s)^2 R^2$, 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 $\pi_{\langle s'|s \rangle}^n$ converges almost-surely weakly to $\pi_{\langle s'|s \rangle}$ as $n_s, n_{s'} \rightarrow +\infty$. Moreover, for any $s \in \mathcal{S}$, we have $\frac{n_s}{n} \xrightarrow{n \rightarrow +\infty} \mathbb{P}(S = s) > 0$, hence $n_s \xrightarrow{n \rightarrow +\infty} +\infty$. As a consequence, $\pi_{\langle s'|s \rangle}^n$ converges almost-surely weakly to $\pi_{\langle s'|s \rangle}$ as $n \rightarrow +\infty$. 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 $W_1(\pi_{\langle s'|s \rangle}^n, \pi_{\langle s'|s \rangle}) \xrightarrow{n \rightarrow +\infty} 0$, where W_1 denotes the Wasserstein-1 distance. Using the dual formulation of the Wasserstein distance, this convergence can be written as,

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

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

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

This entails that the term (2.16) converges almost surely to 0, and completes the proof.

• **Step 2. Consistency of the minimum.** 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 (2.14) 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,

$$\theta_n \xrightarrow{n \rightarrow +\infty} \theta^*.$$

■

Appendix 2.D Proofs of Section 2.8

Proof of Proposition 2.8.1 Without loss of generality, we can suppose that $\mathcal{S} = \{1, \dots, N\}$ for some $N \geq 1$. The proof follows two steps. Firstly, we postulate an SCM $\mathcal{M}^b := \langle U^b, G^b \rangle$ satisfying (A) and (RE) such that its solution (X^b, S^b) generates the same distribution as (X, S) . Secondly, we show that the counterfactual model induced by \mathcal{M}^b is equal to $\Pi := \{\pi_{\langle s'|s \rangle}\}_{s, s' \in \mathcal{S}}$.

Since by assumption the counterfactual model verifies Definition 2.8.1, there exists a collection of random vectors $(X_1, \dots, X_N) \in (\mathbb{R}^d)^N$ such that for any $s, s' \in \mathcal{S}$:

1. $\mathcal{L}((X_s, X_{s'})) = \pi_{\langle s'|s \rangle}$,
2. $\mathcal{L}(X_s) = \mu_s$.

On this basis, we create the structural causal model using the so-called *Rosenblatt transform*: according to [Rosenblatt \(1952\)](#), there always exists a measurable function $\text{RT} : (\mathbb{R}^d)^N \rightarrow (\mathbb{R}^d)^N$ and a random vector $U_X^b \in (\mathbb{R}^d)^N$ (following the uniform distribution on the hypercube) such that $(X_1, \dots, X_N) = \text{RT}(U_X^b)$. Therefore, by dividing RT into N groups of d components we can write $\text{RT} = (\text{RT}_1, \dots, \text{RT}_N)$ so that $X_s = \text{RT}_s(U_X^b)$ for every $s \in \mathcal{S}$. Lastly, we define U_S^b such that $\mathcal{L}(U_S^b) = \mathcal{L}(S)$ and $U_S^b \perp U_X^b$.

We now have all the ingredients to postulate an adequate structural causal model. Let \mathcal{M}^b be the SCM with random seed $U^b := (U_X^b, U_S^b)$ and structural equations given by:

$$\begin{cases} X^b \stackrel{\mathbb{P}\text{-a.s.}}{=} G_X^b(S^b, U_X^b) := \text{RT}_{S^b}(U_X^b), \\ S^b \stackrel{\mathbb{P}\text{-a.s.}}{=} G_S^b(U_S^b) := U_S^b. \end{cases}$$

Note that \mathcal{M}^b satisfies [\(A\)](#) and [\(RE\)](#).

Next, let us verify that \mathcal{M}^b fits the data, that is $\mathcal{L}((X^b, S^b)) = \mathcal{L}((X, S))$. Let E be a Borel set of \mathbb{R}^{d+1} and compute,

$$\begin{aligned} \mathbb{P}((\text{RT}_{S^b}(U_X^b), U_S^b) \in E) &= \sum_{s \in \mathcal{S}} \mathbb{P}(U_S^b = s) \mathbb{P}((\text{RT}_s(U_X^b), s) \in E \mid U_S^b = s) \\ &= \sum_{s \in \mathcal{S}} \mathbb{P}(U_S^b = s) \mathbb{P}((X_s, s) \in E) \\ &= \sum_{s \in \mathcal{S}} \mathbb{P}(S = s) \mathbb{P}((X_s, s) \in E) \\ &= \sum_{s \in \mathcal{S}} \mathbb{P}(S = s) \mathbb{P}((X, S) \in E \mid S = s) \\ &= \mathbb{P}((X, S) \in E), \end{aligned}$$

using that $U_X^b \perp U_S^b$, $\mathcal{L}(U_S^b) = S$ and $\mathcal{L}(X_s) = \mathcal{L}(X \mid S)$.

To finish, we show that $\mathcal{L}((X^b, X_{S^b=s'}^b) \mid S^b = s) = \pi_{\langle s'|s \rangle}$. Let us compute,

$$\begin{aligned} \mathcal{L}((X^b, X_{S^b=s'}^b) \mid S^b = s) &= \mathcal{L}((\text{RT}_{S^b}(U_X^b), \text{RT}_{s'}(U_X^b)) \mid S^b = s) \\ &= \mathcal{L}((\text{RT}_s(U_X^b), \text{RT}_{s'}(U_X^b)) \mid S^b = s) \\ &= \mathcal{L}((\text{RT}_s(U_X^b), \text{RT}_{s'}(U_X^b))) \\ &= \mathcal{L}((X_s, X_{s'})) \\ &= \pi_{\langle s'|s \rangle}. \end{aligned}$$

This completes the proof. ■

Proof of Proposition [2.8.2](#) We first condition over each level of S to obtain,

$$v_{NC} = \sum_{s \in \mathcal{S}} \mathbb{P}(S = s) \mathbb{E}[\Phi(X, S)(s - \mathbb{E}[S]) \mid S = s].$$

Then, we write $\mathbb{E}[S] = \sum_{s' \in \mathcal{S}} \mathbb{P}(S = s')s'$ so that by linearity of the expectation

$$v_{NC} = \sum_{s \in \mathcal{S}} \mathbb{P}(S = s) \sum_{s' \in \mathcal{S}} \mathbb{P}(S = s') \mathbb{E}[\Phi(X, S)(s - s') \mid S = s].$$

This can be expressed as,

$$v_{NC} = \sum_{s \in \mathcal{S}} \left\{ \sum_{s' < s} \mathbb{P}(S = s') \mathbb{P}(S = s) \mathbb{E}[\Phi(X, S)(s - s') \mid S = s] \right. \\ \left. + \sum_{s' > s} \mathbb{P}(S = s') \mathbb{P}(S = s) \mathbb{E}[\Phi(X, S)(s - s') \mid S = s] \right\}.$$

Then, by switching s and s' in the second sum we have

$$v_{NC} = \sum_{s \in \mathcal{S}} \sum_{s' : s > s'} \mathbb{P}(S = s) \mathbb{P}(S = s') \mathbb{E}[\Phi(X, S)(s - s') \mid S = s] \\ - \sum_{s' \in \mathcal{S}} \sum_{s : s > s'} \mathbb{P}(S = s) \mathbb{P}(S = s') \mathbb{E}[\Phi(X, S)(s - s') \mid S = s'].$$

Therefore,

$$v_{NC} = \sum_{s > s'} \mathbb{P}(S = s) \mathbb{P}(S = s') (s - s') (\mathbb{E}[\Phi(X, S) \mid S = s] - \mathbb{E}[\Phi(X, S) \mid S = s']).$$

■

Proof of Proposition 2.8.3 First of all, let us unpack the definition of counterfactual fairness for a predictor of the form $\hat{Y} := \theta^T \Phi(X, S) + \theta_0$. For every $(s, s') \in \mathcal{S}^2$ and $\pi_{\langle s' | s \rangle}$ -almost every (x, x') this requires,

$$\theta^T \Phi(x, s) + \theta_0 = \theta^T \Phi(x', s') + \theta_0,$$

that is,

$$\theta \perp \Phi(x', s') - \Phi(x, s).$$

Reformulating, counterfactual fairness demands that for every $(s, s') \in \mathcal{S}^2$ there exists a set $E_{\langle s' | s \rangle}$ such that $\pi_{\langle s' | s \rangle}(E_{\langle s' | s \rangle}) = 1$ and for every $(x, x') \in E_{\langle s' | s \rangle}$,

$$\theta \perp \Phi(x', s') - \Phi(x, s).$$

Now, we show that $\theta \in \text{Span}(\Pi\Phi)^\perp$ entails that \hat{Y} is counterfactually fair. If $\theta \in \text{Span}(\Pi\Phi)^\perp$, then θ is orthogonal to all vectors in $\Pi\Phi$. This means that for every $(s, s') \in \mathcal{S}^2$ and every $(x, x') \in \text{supp}(\pi_{\langle s' | s \rangle})$,

$$\theta \perp \Phi(x', s') - \Phi(x, s).$$

This implies counterfactual fairness since the set $E_{\langle s' | s \rangle} := \text{supp}(\pi_{\langle s' | s \rangle})$ is such that $\pi_{\langle s' | s \rangle}(E_{\langle s' | s \rangle}) = 1$ for every $(s, s') \in \mathcal{S}$.

Next, we turn to proving the converse implication. Let $(s, s') \in \mathcal{S}^2$. and note that it follows from $\pi_{\langle s'|s \rangle}(E_{\langle s'|s \rangle}) = 1$ that $\overline{E_{\langle s'|s \rangle} \cap \text{supp}(\pi_{\langle s'|s \rangle})} = \text{supp}(\pi_{\langle s'|s \rangle})$. Additionally, the mapping $(x, x') \mapsto \Phi(x', s') - \Phi(x, s)$ is continuous by continuity of $\Phi(\cdot, s)$ and $\Phi(\cdot, s')$. Therefore, the assumption that for every $(x, x') \in E_{\langle s'|s \rangle} \cap \text{supp}(\pi_{\langle s'|s \rangle})$,

$$\theta \perp \Phi(x', s') - \Phi(x, s),$$

extends by continuity of the scalar product to the closure $\overline{E_{\langle s'|s \rangle} \cap \text{supp}(\pi_{\langle s'|s \rangle})}$, that is $\text{supp}(\pi_{\langle s'|s \rangle})$. Taking the linear span of this set, which preserves the orthogonality condition, completes the proof. ■

Part II

Statistical estimation of transport models

Chapter 3

GAN estimation of Lipschitz optimal transport maps

This chapter introduces the first statistically consistent estimator of the optimal transport map between two probability distributions, based on neural networks. Building on theoretical and practical advances in the field of Lipschitz neural networks, we define a Lipschitz-constrained generative adversarial network penalized by the quadratic transportation cost. Then, we demonstrate that, under regularity assumptions, the obtained generator converges uniformly to the optimal transport map as the sample size increases to infinity. Furthermore, we show through a number of numerical experiments that the learnt mapping has promising performances. In contrast to previous work tackling either statistical guarantees or practicality, we provide an expressive and feasible estimator which paves way for optimal transport applications where the asymptotic behaviour must be certified.

3.1 Introduction

An *optimal transport map* is the fundamental object of Monge’s seminal formulation of optimal transport (Monge, 1781). It transforms one distribution into another with minimal effort. Formally, given two probability distributions P and Q on $\Omega \subseteq \mathbb{R}^d$, an optimal transport map from P to Q is a solution to,

$$\min_{T \in \mathcal{T}(P, Q)} \int_{\Omega} \|x - T(x)\|^2 dP(x), \quad (3.1)$$

where $\mathcal{T}(P, Q)$ is the set of measurable maps $T : \Omega \rightarrow \Omega$ pushing forward P to Q , that is $Q(E) = P(T^{-1}(E))$ for every Borel set $E \subseteq \Omega$.¹ This property, denoted by $T_{\#}P = Q$, means that if a random variable X follows the distribution P then its image $T(X)$ follows the distribution Q . According to (Villani, 2003, Theorem 2.12), originally demonstrated in (Cuesta and Matrán, 1989; Brenier, 1991), when P and Q admit densities with respect to the Lebesgue measure and have finite second-order moments, then there exists a unique (up to P -negligible sets) solution to Problem (3.1), which we denote by T_0 .

¹In general, an optimal transport map refers to a solution of Monge’s problem for any cost function. In this chapter, we use this term for the quadratic cost only.

Due to their transparent mathematical formulation and well-established theory, optimal transport maps became popular in many applications from statistics-related fields, where one aims at modeling shifts between distributions. This includes multivariate-quantile analysis (Beirlant et al., 2020; Hallin et al., 2021), signal analysis (Kolouri et al., 2017), domain adaptation (Courty et al., 2014; Seguy et al., 2018; Redko et al., 2019), transfer learning (Gayraud et al., 2017), fairness in machine learning (Gordaliza et al., 2019; Black et al., 2020), and counterfactual reasoning (De Lara et al., 2021a; Berk et al., 2021). However, in such practical frameworks, one typically does not have access to the true distributions P and Q but to independent samples $x_1, \dots, x_n \sim P$ and $y_1, \dots, y_n \sim Q$. This raises the question of constructing a tractable approximation of the solution T_0 on the basis of these empirical observations. The simplest way to compute an empirical optimal transport map from data points is to solve Problem (3.1) between the empirical measures $P_n := n^{-1} \sum_{i=1}^n \delta_{x_i}$ and $Q_n := n^{-1} \sum_{i=1}^n \delta_{y_i}$ instead of P and Q . Implementing this solution suffers from three main drawbacks. The first one is the *computational cost*, since it requires at least $O(n^3)$ operations to compute the empirical optimal transport map (Peyré and Cuturi, 2019). The second is the *memory cost*, since this map is typically stored as an $n \times n$ matrix. As a consequence of these two issues, this approach does not scale well with the size of the dataset. The third limitation of the empirical map is its *inability to generalize* to new out-of-sample observations: by construction it is only matching the set $\{x_1, \dots, x_n\}$ to $\{y_1, \dots, y_n\}$.

These practical drawbacks triggered a vast literature on continuous approximations of optimal transport maps. The proposed mappings all come with different practical limitations, theoretical guarantees, and experimental performances. On the one hand, a wide range of these constructions provably converge in some sense to the true map T_0 as n increases to infinity, making them consistent estimators. The so-called plug-in estimators, such as the ones proposed in (Beirlant et al., 2020; Hallin et al., 2021; Manole et al., 2021), extend the empirical solution to the whole domain Ω by leveraging regularity assumptions. However, they still bear the burdens of computing and storing the empirical transport map. The smooth estimator introduced by Hütter and Rigollet (2021) reaches near-optimal minimax convergence rate, but fails to be computationally tractable. In contrast, Seguy et al. (2018) and Pooladian and Niles-Weed (2021) employed entropic regularization, a numerical scheme based on Sinkhorn’s algorithm (Cuturi, 2013), to build an implementable and scalable estimator. On the other hand, several papers proposed learning the optimal transport map through neural networks, leading to expressive approximations with high generalization power. Specifically, Leygonie et al. (2019) and Black et al. (2020) developed approximations based on a generative-adversarial-network (GAN) objective (Goodfellow et al., 2014; Arjovsky et al., 2017). More recently, the use of input convex neural networks, building on the convexity of the optimal transport potential, has received a growing attention (Makkuva et al., 2020; Korotin et al., 2021; Huang et al., 2021). However, while these neural-based mappings display strong experimental performances, they generally lack theoretical guarantees, in particular the statistical convergence.

To sum-up, the literature has mostly addressed either theoretically grounded statistical estimators of optimal transport maps, but unsuitable for large-scale implementations, or efficient heuristic approximations, at the cost of statistical guarantees. In this chapter, we propose a novel GAN-based estimator G_n of T_0 which, under some assumptions, converges

uniformly:

$$\|G_n - T_0\|_\infty \xrightarrow[n \rightarrow +\infty]{a.s.} 0.$$

Our construction takes root in the approximation from (Black et al., 2020), defined as the generator of a penalized Wasserstein-GAN (WGAN) training problem (Arjovsky et al., 2017), and improve it by assuming a setting where the optimal transport map is Lipschitz and by leveraging recent theoretical and practical advances on Lipschitz neural networks (Anil et al., 2019; Tanielian and Biau, 2021; Béthune et al., 2022). Formally, G_n solves the following adversarial training:

$$\inf_{G \in \mathcal{G}_n} \left\{ \|I - G\|_{L^2(P_n)}^2 + \lambda_n \sup_{D \in \mathcal{D}_n} \int D(d(G_\# P_n) - dQ_n) \right\},$$

where \mathcal{D}_n is a class of 1-Lipschitz discriminators providing a proxy for the Wasserstein-1 distance, and \mathcal{G}_n is a class of Lipschitz generators parametrizing the space of feasible mappings. The positive parameter λ_n governs the trade-off between minimizing the quadratic transportation cost, promoting the objective of the Monge problem (3.1), and minimizing the distance between the generated and the target distributions, enforcing the push-forward constraint.

The most similar papers to ours are the ones of Seguy et al. (2018) and Pooladian and Niles-Weed (2021), as they propose feasible estimators with statistical guarantees. We note two main differences. First, we do not rely on entropic regularization while still ensuring scalability to large datasets. Second, our estimator innovates by being defined as a neural network. In particular, Seguy et al. (2018) relies on a neural network in practice, but the statistical convergence holds for a theoretical estimator. Regarding theoretical guarantees, we lack the convergence rates provided in (Pooladian and Niles-Weed, 2021), but we prove a stronger result than Seguy et al. (2018) by ensuring the uniform convergence of the estimator.

Outline. The rest of the chapter is organized as follows:

1. Section 3.2 introduces the necessary background on so-called *GroupSort* neural networks, which became the gold standard to parametrize Lipschitz feed-forward neural networks. By studying the multivariate setting, we provide generalizations of the main approximation theorem from (Tanielian and Biau, 2021).
2. Section 3.3 presents the technical assumptions of our framework, in particular the regularity of the optimal transport map, details construction of our GAN estimator, and states the statistical consistency theorem.
3. Section 3.4 focuses on the practical implementation of the estimator, and study its performance through a number of numerical experiments.

Notations. The absolute value of real numbers and the Euclidean norm of vectors are respectively given by $|\cdot|$ and $\|\cdot\|$. The notation B_r refers to the centered Euclidean ball of \mathbb{R}^d with radius $r > 0$. We denote by $\text{diam}(\Omega)$ the diameter of a set $\Omega \subseteq \mathbb{R}^d$. If Ω is a closed convex set, then Proj_Ω stands for the projection onto Ω . The support of a

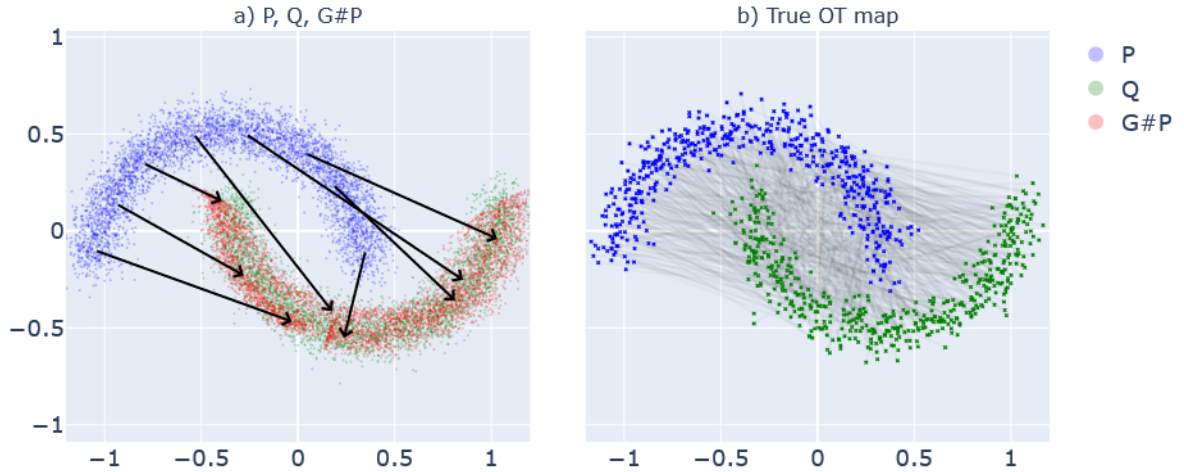


Figure 3.1: Estimation of the optimal transport map on the TwoMoons dataset. (a) GAN estimator G after 800 gradient steps on the generator, on the basis of 4,000 points from each distribution. The black arrows represent the transport of specific points. (b) Empirical optimal transport map (discrete matching) between samples of size 500.

probability measure is given by $\text{supp}(\cdot)$. In the following $\Omega_1 \subseteq \mathbb{R}^{d_1}$ and $\Omega_2 \subseteq \mathbb{R}^{d_2}$ denote two arbitrary subsets. For a function $F : \Omega_1 \rightarrow \Omega_2$ and μ a probability measure on Ω_1 , we write $\|F\|_{L^2(\mu)} := \sqrt{\int_{\Omega_1} \|F(x)\|^2 d\mu(x)}$. The supremum norm of function is given by $\|\cdot\|_{\infty}$. For some $L > 0$, we write $\text{Lip}_L(\Omega_1, \Omega_2)$ the set of L -Lipschitz functions from Ω_1 to Ω_2 . For some $\alpha > 0$, we call $\mathcal{C}^\alpha(\Omega_1, \Omega_2)$ the set of α -Hölder functions from Ω_1 to Ω_2 and write $\|\cdot\|_{\alpha, \infty}$ for the α -Hölder norm of functions. For a differentiable function $F : \Omega_1 \rightarrow \Omega_2$, we call F' its derivative, where for any $x \in \Omega_1$ the quantity $F'(x)$ is a $d_1 \times d_2$ matrix. For a real symmetric matrix S and a real number γ , the relation $\gamma \preceq S$ indicates that all the eigenvalues of S are greater than γ . The relation \succeq is defined similarly.

3.2 Lipschitz neural networks

The GAN estimator defined by (3.2) and further described in Section 3.3 requires generators and discriminators that are both Lipschitz. The question of imposing sharp Lipschitz constraints on neural networks has attracted much attention from the field of machine learning, especially with the popularization of WGANs which rely on 1-Lipschitz discriminators. In particular, gradient penalization (Gulrajani et al., 2017) has proven to be more efficient than the parameter-clipping approach originally proposed by Arjovsky et al. (2017). In this chapter, we focus on the recently introduced *GroupSort* activation function to impose the Lipschitz constraint, which have proven to yield tighter estimates of 1-Lipschitz functions (Anil et al., 2019; Tanielian and Biau, 2021). We recall the necessary background on GroupSort-based networks, and show that their ability to approximate any bounded classes of Lipschitz functions holds for arbitrary output dimension.

3.2.1 Multivariate GroupSort neural networks

We introduce GroupSort neural networks in a similar fashion to (Tanielian and Biau, 2021). In contrast, we consider a more general setting where the output dimension $p \geq 1$ is arbitrary. This difference is motivated by the optimal transport map being a multivariate function.

We write σ_k for the GroupSort activation function of grouping size $k \geq 2$. By definition, it splits the pre-activation input into groups of size k , and then sorts each group by decreasing order. This operation is 1-Lipschitz, gradient-norm preserving and homogeneous (Anil et al., 2019). In this chapter, we only address the grouping size 2. We call a GroupSort feed-forward neural network (with grouping size 2) any function $N_\theta : \mathbb{R}^d \rightarrow \mathbb{R}^p$ of the form

$$N_\theta = h_l \circ h_{l-1} \circ \dots \circ h_1, \quad (3.2)$$

where

$$\begin{aligned} h_1(x) &:= W_1 x + b_1 \text{ with } W_1 \in \mathbb{R}^{w_1 \times d}, b_1 \in \mathbb{R}^{w_1}; \\ h_2(x) &:= W_2 \sigma_2(x) + b_2 \text{ with } W_2 \in \mathbb{R}^{w_2 \times w_1}, b_2 \in \mathbb{R}^{w_2}; \\ &\dots \\ h_l(x) &:= W_l \sigma_2(x) + b_l \text{ with } W_l \in \mathbb{R}^{p \times w_{l-1}}, b_l \in \mathbb{R}^p. \end{aligned}$$

The integer $l \geq 1$ denotes the *depth* of the network while the integers $\{w_1, \dots, w_{l-1}\}$ refer to the *widths* of the hidden layers $\{h_1, \dots, h_{l-1}\}$. The widths are assumed to be divisible by 2 (the grouping size). Additionally, we define $s := \sum_{i=1}^{l-1} w_i$ the *size* of the network. The parameter $\theta := (W_1, \dots, W_l, b_1, \dots, b_l) \in \Theta$ represents the *weights* matrices and *offset* vectors of N_θ .

For a matrix W , let $\|W\|_\infty := \sup_{\|x\|_\infty=1} \|Wx\|_\infty$ and $\|W\|_{2,\infty} := \sup_{\|x\|=1} \|Wx\|_\infty$, where $\|x\|_\infty$ denotes the maximum norm of vectors. Consider the following assumption on the parameters:

Assumption (C): Compactness

There exists a constant $C > 0$ such that for all $(W_1, \dots, W_l, b_1, \dots, b_l) \in \Theta$,

$$\begin{aligned} \|W_1\|_{2,\infty} &\leq 1, \\ \max(\|W_2\|_\infty, \dots, \|W_l\|_\infty) &\leq 1, \\ \max(\|b_1\|_\infty, \dots, \|b_l\|_\infty) &\leq C. \end{aligned}$$

In the following, we denote by $\mathcal{N}_C^p(l, s)$ the class of GroupSort feed-forward neural networks with depth l , size s , output dimension p , satisfying Assumption (C) for the constant $C > 0$. When the depth and size are arbitrary, we simply write \mathcal{N}_C^p . The following result is a trivial extension to the multivariate case of (Tanielian and Biau, 2021, Lemma 1), stating that GroupSort neural networks satisfying Assumption (C) are 1-Lipschitz.

Lemma 3.2.1: GroupSort neural networks are 1-Lipschitz

For any $C > 0$, $\mathcal{N}_C^p \subset \text{Lip}_1(\mathbb{R}^d, \mathbb{R}^p)$.

Next, we study their ability to approximate Lipschitz continuous functions.

3.2.2 Approximating Lipschitz continuous functions

We now restrict the input domain to a *compact* subset of \mathbb{R}^d denoted by Ω . The following theorem states that for a well-chosen C the class \mathcal{N}_C^1 approximates with given precision any bounded subclass of $\text{Lip}_1(\Omega, \mathbb{R})$. It generalizes Theorem 2 in (Tanielian and Biau, 2021) by providing the universal constant for which Assumption (C) is satisfied, and extending the result to any compact domain Ω while it was restricted to $[0, 1]^d$.

Theorem 3.2.1: Approximation of univariate Lipschitz functions

Let $\mathcal{F} \subseteq \text{Lip}_1(\Omega, \mathbb{R})$ be a class of functions such that $\sup_{f \in \mathcal{F}} \|f\|_\infty \leq K_{\mathcal{F}}$ for some $K_{\mathcal{F}} > 0$. Set $\epsilon > 0$ and $C := K_{\mathcal{F}} + \sqrt{d}(\sup_{x \in \Omega} \|x\| + 1) + \epsilon$. Then, for any $f \in \mathcal{F}$, there exists a neural network $N \in \mathcal{N}_C^1(l, s)$ where

$$l = O\left(d^2 \log_2\left(\frac{2\sqrt{d}}{\epsilon}\right)\right) \text{ and } s = O\left(\left(\frac{2\sqrt{d}}{\epsilon}\right)^{d^2}\right),$$

such that $\|N - f\|_\infty \leq \epsilon$.

The proof essentially follows that of (Tanielian and Biau, 2021). It generalizes some parts by tracking the bound on the offset vectors of the approximating network. Interestingly, Theorem 3.2.1 can be extended to the case where the output is of dimension p .

Theorem 3.2.2: Approximation of multivariate Lipschitz functions

Let $\mathcal{G} \subseteq \text{Lip}_1(\Omega, \mathbb{R}^p)$ be a class of functions such that $\sup_{g \in \mathcal{G}} \|g\|_\infty \leq K_{\mathcal{G}}$ for some $K_{\mathcal{G}} > 0$. Set $\epsilon > 0$ and $C = K_{\mathcal{G}} + \sqrt{d}(\sup_{x \in \Omega} \|x\| + 1) + \epsilon$. Then, for any $g \in \mathcal{G}$ there exists a neural network $N \in \mathcal{N}_C^p(l, s)$ where

$$l = O\left(d^2 \log_2\left(\frac{2\sqrt{d}\sqrt{p}}{\epsilon}\right)\right), \text{ and } s = O\left(p \left(\frac{2\sqrt{d}\sqrt{p}}{\epsilon}\right)^{d^2}\right),$$

such that $\|N - g\|_\infty \leq \epsilon$.

The proof amounts to applying Theorem 3.2.1 to the univariate function along each dimension. Note that Theorems 3.2.1 and 3.2.2 can be extended to approximate L -Lipschitz

functions, for an arbitrary $L > 0$, by multiplying by L the output later of 1-Lipschitz neural networks. This remark will be useful to approximate the optimal transport map, assumed to be L -Lipschitz.

3.3 GAN estimator

In this section, we address the construction of an estimator of the optimal transport map, and show its uniform convergence as the sample size increases to infinity.

3.3.1 Optimal transport setup

Set P and Q two measures on \mathbb{R}^d admitting densities with respect to the Lebesgue measure and with finite second-order moments. We aim at estimating with a GroupSort neural network the unique optimal transport map T_0 between P and Q through the knowledge of the empirical distributions P_n and Q_n . As mentioned in the introduction, we consider a setting where the optimal transport map T_0 is Lipschitz.

As in the previous section, $\Omega \subset \mathbb{R}$ is a compact set, and we denote by $\Omega_P := \text{supp}(P)$ the *source* domain and $\Omega_Q := \text{supp}(Q)$ the *target* domain. Then, we let $L \geq 2$ and make the following assumptions:

Assumption (S1): Smoothness of the source domain

The source domain $\Omega_P \subseteq B_L$ is a bounded and connected Lipschitz domain. The measure P admits a density ρ with respect to the Lebesgue measure such that $L^{-1} \leq \rho(x) \leq L$ for almost every $x \in \Omega_P$.

Assumption (S2): Smoothness of the optimal transport map

Let $\tilde{\Omega}_P$ denote a convex set with Lipschitz boundary such that $\Omega_P + B_{L^{-1}} \subseteq \tilde{\Omega}_P \subseteq B_L$. The optimal transport map T_0 is a differentiable function from $\tilde{\Omega}_P$ to \mathbb{R}^d such that $T_0 = \nabla f_0$ where $f_0 : \tilde{\Omega}_P \rightarrow \mathbb{R}^d$ is a differentiable convex function. Additionally it satisfies:

- (i) $T_0 \in C^2(\tilde{\Omega}_P, \mathbb{R}^d)$ such that $\|T_0\|_{2,\infty} \leq L$;
- (ii) $L^{-1} \preceq T'_0(x) \preceq L$ for all $x \in \tilde{\Omega}_P$.

These are the same hypothesis as in (Hütter and Rigollet, 2021, Section 5), specified with a Hölder regularity α equal to 2. This makes our setting milder, as we do not require the optimal transport map to be highly regular. Assumptions (S1) and (S2) ensure the existence of a near-optimal minimax estimator of T_0 , which play a key role in the proof of our estimator's consistency. Note that, without loss of generality, we can consider that P and Q are measures on a compact set $\Omega \subset \mathbb{R}^d$ sufficiently large to contain B_L . Then, Assumption (S2) implies that $T_0 \in \text{Lip}_L(\Omega, B_L)$.

Now that the optimal transport problem is properly specified, we turn to the GAN architecture through which our estimator is defined.

3.3.2 GAN setup

The optimal transport map T_0 satisfies two objectives: it is constrained to pushing forward P to Q , that is $T_{0\#}P = Q$; it minimizes the quadratic transportation cost $\|I - T_0\|_{L^2(P)}^2$. Due to the push-forward condition, T_0 can be regarded as a generative model. This observation is the foundation of the approximation of [Black et al. \(2020\)](#). They proposed to regularize the WGAN objective function, promoting only the push-forward condition, with an optimal transport penalty on the generator. We proceed similarly, with three critical differences. First, we penalize the quadratic transportation cost with the push-forward condition instead of the converse. Second, we employ GroupSort neural networks to implement the discriminator and generator. Third, because we aim at proving the statistical convergence of the generator, we emphasize for all the objects involved in the GAN their dependence to the sample size n , including the penalty weight.

Discriminator

In the WGAN framework, the *discriminator* $D : \mathbb{R}^d \rightarrow \mathbb{R}$ is a neural network defining a proxy for the Wasserstein-1 distance, while the *generator* $G : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is a neural network minimizing this proxy between $G_{\#}P_n$ and Q_n , thereby aiming at generating Q from P .

We recall that the Wasserstein-1 distance between two measures μ and ν on Ω is defined as,

$$\mathcal{W}(\mu, \nu) := \inf_{\pi \in \Pi(\mu, \nu)} \int_{\Omega \times \Omega} \|x - y\| d\pi(x, y),$$

where $\Pi(\mu, \nu)$ is the set of couplings with μ as first marginal and ν as second marginal. Interestingly, this distance enjoys the following dual formulation, known as the Kantorovich-Rubinstein formula ([Kantorovich and Rubinshtein, 1958](#)). According to ([Villani, 2008, Particular Case 5.15](#)), this can be written as:

$$\mathcal{W}(\mu, \nu) = \sup_{f \in \text{Lip}_1(\Omega, \mathbb{R})} \int f(d\mu - d\nu). \quad (3.3)$$

The key idea of WGAN is to approximate this distance by computing the supremum over a class of neural networks included in $\text{Lip}_1(\Omega, \mathbb{R})$. The larger the class, the better the approximation. Originally, this was done by clipping, thresholding the weights of the network, leading to a coarse approximation of the Wasserstein distance. Later, several papers showed that using GroupSort neural networks led to sharper approximations ([Anil et al., 2019](#); [Biau et al., 2021](#)).

Actually, note that if f is an optimal function in Problem [\(3.3\)](#), then the function $f + b$ for any constant b is also an optimal solution. As a consequence, we can without loss of generality restrict the set of feasible potentials to 1-Lipschitz functions taking the value zero at a given arbitrary anchor point $x_0 \in \Omega$. Formally, let's define

$$\mathcal{F} := \{f \in \text{Lip}_1(\Omega, \mathbb{R}) \mid f(x_0) = 0\}. \quad (3.4)$$

Then we can write,

$$\mathcal{W}(\mu, \nu) = \sup_{f \in \mathcal{F}} \int f(d\mu - d\nu).$$

The interest of this formulation is that the feasible potentials now belongs to a bounded subclass of Lipschitz functions.

Lemma 3.3.1: Control on the discriminators

Let \mathcal{F} be defined as in Equation (3.4). Then,

$$K_{\mathcal{F}} := \sup_{f \in \mathcal{F}} \|f\|_{\infty} \leq \text{diam}(\Omega).$$

Thus, Theorem 3.2.1 entails that they can be approximated by GroupSort neural networks with specific depth and size. Following this remark, we define for each sample size n the class of feasible discriminators \mathcal{D}_n as well-chosen GroupSort neural networks. Specifically, the discriminators are defined as in the next assumption.

Assumption (G1): Construction of the discriminators

Set a sequence of positive numbers $\{\epsilon_n\}_{n \in \mathbb{N}}$ such that $\lim_{n \rightarrow +\infty} \epsilon_n = 0$, and a sequence of constants $\{C_n\}_{n \in \mathbb{N}}$ defined as

$$C_n := \text{diam}(\Omega) + \sqrt{d}(\sup_{x \in \Omega} \|x\| + 1) + \epsilon_n.$$

For every $n \in \mathbb{N}$, define $\mathcal{D}_n := \mathcal{N}_{C_n}^1(l_n, s_n)$ where,

$$l_n = O\left(d^2 \log_2\left(\frac{2\sqrt{d}}{\epsilon_n}\right)\right), \quad \text{and} \quad s_n = O\left(\left(\frac{2\sqrt{d}}{\epsilon_n}\right)^{d^2}\right).$$

Then, we approximate the Wasserstein-1 distance through the following integral probability metric:

$$\mathcal{W}_n(\mu, \nu) := \sup_{D \in \mathcal{D}_n} \int D(d\mu - d\nu). \quad (3.5)$$

An important consequence of Assumption (G1) through Lemma 3.3.1 and Theorem 3.2.1 is that $\bigcup_{n \in \mathbb{N}} \mathcal{D}_n$ is dense in \mathcal{F} , rendering \mathcal{W}_n asymptotically close to \mathcal{W} as n increases to infinity. Note that the sequence $\{\epsilon_n\}_{n \in \mathbb{N}}$ characterizes the rate at which the class \mathcal{D}_n approximates \mathcal{F} . Now that we have properly defined the discriminators, we focus on the generators.

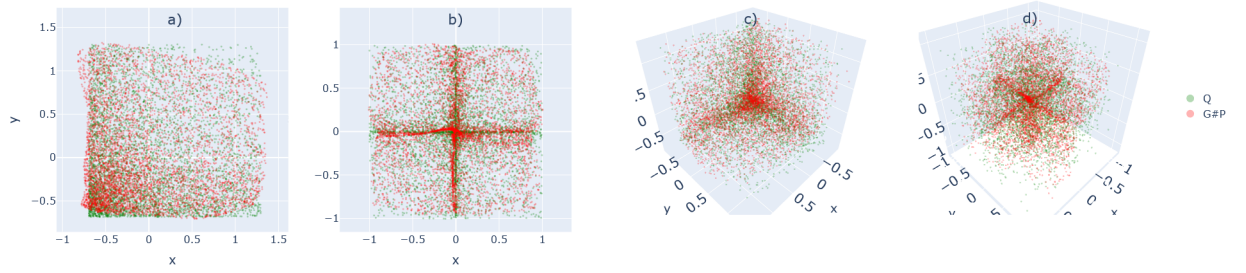


Figure 3.2: Visualisation of $G_{\#}P$ and $Q := T_0\#P$ with 10,000 points. P is the uniform distribution on $[-1, 1]^d$. The generator is trained for 120 gradient steps. The Figures (a)-(b) corresponds to $d = 2$. The Figures (c)-(d) corresponds to $d = 3$. In Figures (a)-(c), we defined T_0 by coordinate-wise application of $x \mapsto \frac{1}{1.18}(\exp x - 1.18)$. In Figures (b)-(d), we defined T_0 by coordinate-wise application of $x \mapsto x^2 \text{sign}(x)$.

Generator

On the contrary to a standard WGAN, the generator must additionally minimize the quadratic transportation cost in order to approach the optimal transport map T_0 . Let us denote by \mathcal{G}_n the class of feasible generators, which will be specified later. A naive formulation for our estimator $G_n \in \mathcal{G}_n$ would be,

$$G_n \in \arg \min_{G \in \mathcal{G}_n \text{ s.t. } G_{\#}P_n = Q_n} \|I - G\|_{L^2(P_n)}^2.$$

However, since the push-forward condition is intractable as such, we replace it by a penalty term based on the neural proxy of the Wasserstein-1 distance. Formally, we set $\lambda_n > 0$ a regularization weight and we define the GAN estimator G_n as an optimal solution to Problem (3.2), that is

$$G_n \in \arg \min_{G \in \mathcal{G}_n} \mathcal{L}_n(G),$$

where

$$\mathcal{L}_n(G) := \|I - G\|_{L^2(P_n)}^2 + \lambda_n \mathcal{W}_n(G_{\#}P_n, Q_n).$$

We note that Problem (3.2) is well-posed under mild conditions.

Proposition 3.3.1: Existence of an optimal generator

If $\mathcal{D}_n \subseteq \text{Lip}_1(\Omega, \mathbb{R})$ and \mathcal{G}_n is compact, then Problem (3.2) admits solutions.

This result is a direct consequence of the Lipschitz continuity of the loss function \mathcal{L}_n , which we demonstrate in the proof.

At this stage, we should make further assumptions on \mathcal{G}_n to exploit the smoothness of the optimal transport problem. Let us define

$$\mathcal{G} := \text{Lip}_L(\Omega, B_L), \quad (3.6)$$

which is a class of bounded Lipschitz functions.

Lemma 3.3.2: Control on the generators

Let \mathcal{G} be defined as in Equation (3.6). Then,

$$K_{\mathcal{G}} := \sup_{g \in \mathcal{G}} \|g\|_{\infty} \leq L \operatorname{diam}(\Omega) + \sup_{x \in \Omega} \|x\|.$$

Critically, under Assumption (S2), the solution T_0 belongs to \mathcal{G} , and as such can be approximated by GroupSort neural networks according to Theorem 3.2.2. This motivates the following conditions on the set of feasible generators \mathcal{G}_n :

Assumption (G2): Construction of the generators

Set $\{\varepsilon_n\}_{n \in \mathbb{N}}$ a sequence of positive numbers such that $\lim_{n \rightarrow +\infty} \varepsilon_n = 0$, and a sequence of constants $\{C_n\}_{n \in \mathbb{N}}$ defined as

$$C_n := L \operatorname{diam}(\Omega) + (\sqrt{d} + 1) \sup_{x \in \Omega} \|x\| + \sqrt{d} + \varepsilon_n.$$

For every $n \in \mathbb{N}$, we define \mathcal{G}_n as

$$\{x \in \Omega \mapsto \operatorname{Proj}_{B_L}(L \times N(x)), N \in \mathcal{N}_{C_n}^d(l_n, s_n)\}$$

where,

$$l_n = O\left(d \log_2\left(\frac{2d}{\varepsilon_n}\right)\right), \quad \text{and} \quad s_n = O\left(d \left(\frac{2d}{\varepsilon_n}\right)^{d^2}\right).$$

Defined as such, \mathcal{G}_n is included in \mathcal{G} . The idea behind Assumption (G2) is similar to that of Assumption (G1). In particular, the condition on the depth and size of the networks guarantees through Theorem 3.2.2 that \mathcal{G}_n asymptotically fills \mathcal{G} at speed ε_n , allowing to recover T_0 at the limit.

3.3.3 Main theorem

The convergence of $\{G_n\}_{n \in \mathbb{N}}$ towards T_0 revolves around two antagonistic conditions. Instinctively, the sequence of regularization weights $\{\lambda_n\}_{n \in \mathbb{N}}$ must tend to infinity in order to impose the push-forward condition at the limit. Concurrently, the sequence of feasible generators $\{\mathcal{G}_n\}_{n \in \mathbb{N}}$ must fill \mathcal{G} sufficiently fast. This corresponds to the following assumptions:

Assumption (G3): Speed of the regularization

The sequence $\{\lambda_n\}$ is such that $\lim_{n \rightarrow +\infty} \lambda_n = +\infty$ and

$$\lambda_n = \begin{cases} o\left(n^{\frac{1}{d}}\right) & \text{if } d > 2, \\ o\left(n^{\frac{1}{2}}/\log n\right) & \text{if } d = 2, \\ o\left(n^{\frac{1}{2}}/\sqrt{\log n}\right) & \text{if } d = 1. \end{cases}$$

Assumption (G4): Speed of the covering

The sequence $\{\varepsilon_n\}_{n \in \mathbb{N}}$ from Assumption **(G2)** is such that, $\varepsilon_n = o\left(\frac{1}{\lambda_n}\right)$.

We are now ready to state our main theorem.

Theorem 3.3.1: Uniform convergence of the optimal generators

Let P and Q be such that the smoothness assumptions **(S1)** and **(S2)** on the optimal transport problem hold, and denote by T_0 the (almost everywhere) unique optimal transport map between P and Q . Suppose that the GAN problem satisfies Assumptions **(G1)**, **(G2)**, **(G3)** and **(G4)**. Then, for G_n defined as a solution to Problem **(3.2)** we have

$$\|G_n - T_0\|_{\infty} \xrightarrow[n \rightarrow +\infty]{a.s.} 0.$$

To the best of our knowledge, this is the first statistical consistency result for a neural-network-based optimal transport map. We leave the analysis of consistency rates for future work. In particular, we could obtain sharper results by imposing conditions on the parameter ε_n which characterizes the rate at which the discriminators \mathcal{D}_n approximate the 1-Lipschitz potentials, and by leveraging stronger regularity assumptions on T_0 . The proof is quite technical; the convergence of λ_n to infinity prevents from using classical empirical process techniques. Instead, we rely on more analytical arguments based on the relative compactness properties of Lipschitz functions. Moreover, we note that the proof still holds for more general classes of generators as long as they maintain certain universality properties and have a Lipschitz constant that can be controlled. This is one of the main strengths of GroupSort neural networks: they can sharply approximate any classes of bounded Lipschitz functions with the same Lipschitz constant.

3.4 Numerical experiments

The rest of the chapter addresses the implementation of our method, and showcases experimental results. Specifically, we do not try to illustrate the convergence rate of the estimator,

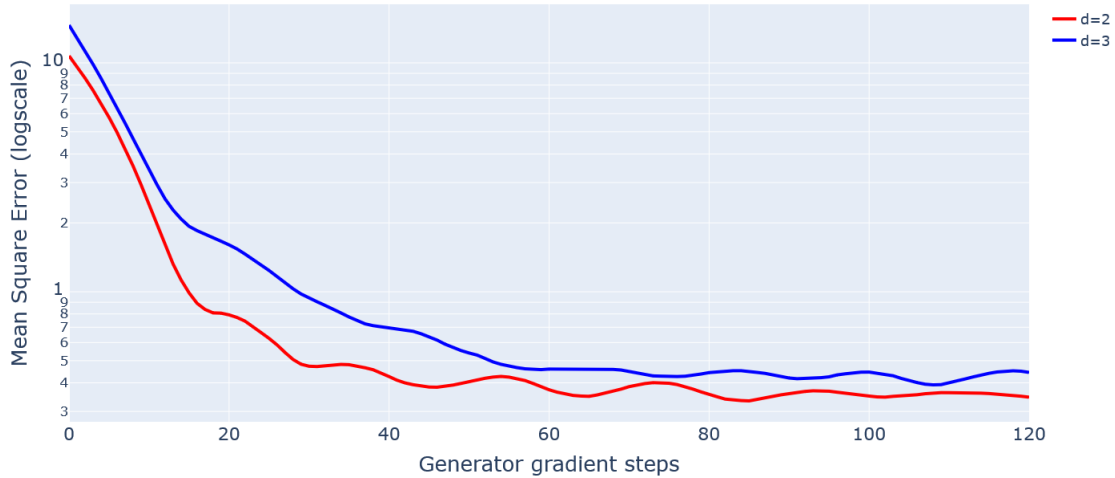


Figure 3.3: Evolution of the mean square error $\|T_0 - G\|_2^2$ during the learning process as function of the number of gradient steps on generator with batch size 512, for $x \mapsto \frac{1}{1.18}(\exp x - 1.18)$. The number of samples used is proportional to the number of steps.

which is yet to be found, but instead focus on the efficiency and practicality of our GAN-based optimal transport map.

We would also like to emphasize in passing that while we left this short section in the thesis for the sake of illustration and completeness, the code and figures are the work of my colleague Louis Béthune.

3.4.1 Implementation

In the following experiments, we use $(\cdot \rightarrow 80 \rightarrow 80 \rightarrow 80 \rightarrow \cdot)$ densely connected neural networks with GroupSort activation functions for both the generator and the discriminator. We implement GroupSort using Deel-Lip library^[2]. The 1-Lipschitz constraint is enforced through projections onto a parameter space satisfying Assumption **(C)**. The output layer of the generator is multiplied by L to be made L -Lipschitz. Critically, since this constant is unknown in practice, we must rely on a large-enough user-defined upper bound. We use Adam with default parameters for the optimization. All experiments have been run on personal workstation with 32GB RAM and NVIDIA Quadro RTX 8000 48GB GPU.

The learning procedure is detailed in Algorithm **1**. In contrast to a WGAN, the generator loss includes the quadratic transportation cost. It also differs from the procedure proposed in (Black et al., 2020) by implementing a sharper weight projection than clipping.

Algorithm 1: GAN learning of the optimal transport map

Input: source distribution P , target distribution Q , regularization parameter λ , discriminator $\{D_\psi\}_{\psi \in \Psi}$, generator $\{G_\phi\}_{\phi \in \Phi}$, respective learning rates η_D and η_G , minibatch size m

repeat

repeat

 Sample minibatches: $\{x_i\}_{i=1}^m \sim P$, $\{y_i\}_{i=1}^m \sim Q$

 Define cost function:

$$\mathcal{W}_D(\psi) := \frac{1}{m} \sum_{i=1}^m D_\psi(G_\phi(x_i)) - \frac{1}{m} \sum_{i=1}^m D_\psi(y_i)$$

 Projected gradient ascent step on discriminator:

$$\psi \leftarrow \text{Proj}_\Psi(\psi + \eta_D \nabla_\psi \mathcal{W}_D(\psi))$$

until convergence of D_ψ

 Sample minibatch: $\{x'_i\}_{i=1}^m \sim P$

 Define cost functions:

$$\mathcal{W}_G(\phi) := \frac{1}{m} \sum_{i=1}^m D_\psi(G_\phi(x'_i))$$

$$\mathcal{C}(\phi) := \frac{1}{m} \sum_{i=1}^m \|x'_i - G_\phi(x'_i)\|^2$$

 Projected gradient descent step on generator:

$$\phi \leftarrow \text{Proj}_\Phi(\phi - \eta_G \nabla_\phi (\mathcal{C}(\phi) + \lambda \mathcal{W}_G(\phi)))$$

until convergence of G_ϕ

3.4.2 Experimental results

We evaluate how close the trained generator G is to the optimal transport map T_0 . Recall that our construction, as in (Hütter and Rigollet, 2021; Pooladian and Niles-Weed, 2021), is tailored to settings where the optimal transport map is at least Lipschitz, hence continuous. This excludes in particular target distributions with disconnected supports. Firstly, we address a setting where the true optimal transport map T_0 is unknown. Figure 3.1 benchmarks the GAN estimator against the empirical optimal transport map on the TwoMoons dataset. We used the POT library to compute the discrete matching (Flamary et al., 2021). It shows that the generator faithfully matches the two moons with respect to the quadratic transportation cost.

²<https://dee1-lip.readthedocs.io>

Secondly, we consider synthetic examples for which T_0 has an explicit formula. We follow the protocol adopted in the aforementioned papers by defining P as the uniform distribution on the hypercube $[-1, 1]^d$ and setting $Q := T_0\#P$, where $T_0 : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is obtained by applying a monotone scalar function coordinate-wise. The combination of McCann’s theorem (McCann, 1995), stating that there exists a unique gradient of a convex function achieving the push-forward between two Lebesgue-absolutely-continuous distributions, and (Villani, 2003, Theorem 2.12), stating that an optimal transport map coincide almost-everywhere with the gradient of a convex function, ensures that T_0 constructed as such is the (almost everywhere unique) optimal transport map between P and Q . Note that for practical reasons, we choose T_0 such that Q is a distribution with zero mean and width less than 2: normalizing the input and output distributions of a neural network ensures faster convergence. The result are illustrated in Figure 3.2.

Additionally, we investigate in Figure 3.3 the evolution of the mean square error between the generator G and the optimal transport map T_0 as the learning process goes on. It confirms that the optimization scheme has the expected behaviour. Furthermore, since the mean square error is evaluated on an independent sample to the training set, it illustrates the generalization ability of the learnt map.

3.5 Conclusion

We conclude this chapter by discussing the significance of its contributions.

3.5.1 Summary of contributions

The method we propose has the advantage of providing a theoretically sound and feasible estimation of the optimal transport map whose statistical convergence can be mathematically certified. Theorem 4.5.1 proves its consistency, while Section 4.6.2 illustrates its ability to learn the underlying map. This renders this estimator potentially suitable for many applications where guarantees of convergence are required such as FlipTest (De Lara et al., 2021b).

Additionally, we extended in Section 3.2 the established theory on approximating Lipschitz continuous functions by GroupSort neural networks to the multivariate case. This also opens new lines of inquiry for further applications of these networks, such as imposing regularity properties on generative models. Finally, our statistical framework and mathematical proofs addressed several interesting problems at the frontier between neural network modeling and statistics. We hope this effort will contribute to bridge the gap between deep learning and statistical theory.

3.5.2 Limitations and further research

We believe that some contributions of this chapter deserve improvements. Notably, the optimization step is currently too cumbersome for our estimator to be widely deployed. As for any neural optimal-transport approximation, it depends on numerous hyper-parameters, and additionally requires to enter an unknown Lipschitz constant. This leads to several tries

and repeats even on low-dimensional data to reach convergence. Therefore, automatizing hyper-parameter tuning would be a significant advance.

However, practical limitations are more radical. Similarly to (Hütter and Rigollet, 2021), our framework demands the distributions to be highly regular to render the optimal transport map Lipschitz, which significantly restricts its validity domain. Moreover, Salmona et al. (2022) recently pointed out that Lipschitz generative models could hardly fit complex distributions. These are notably the reasons why we did not use the GAN estimator when implementing transport-based counterfactuals; it always felt easier to rely on true empirical solution or simple plug-in generalizations, as in Chapter 2. In this sense, it seems more promising to develop more flexible approaches. In particular, Uscidda and Cuturi (2023) introduces a regularizer to learn transport maps for general costs and arbitrary distributions.

Nevertheless, the specific Lipschitz scenario we considered enabled us to derive unique convergence results for a neural estimators. This is according to us the greatest merit of this work; it pioneers the study of the statistical convergence of neural-network-based optimal transport maps.

Appendix 3.A Proofs of Section 3.2

The demonstration of (Tanielian and Biau, 2021, Theorem 2) relies on (He et al., 2020, Theorem 5.1), according to which any 1-Lipschitz piecewise-affine function q defined on a compact set can be written as,

$$q(x) = \max_{1 \leq s \leq m} \min_{i \in I_s} (a_i \cdot x + c_i), \quad (3.7)$$

where for any $1 \leq s \leq m$, I_s is a subset of $\{1, \dots, m\}$ and $\|a_i\| \leq 1$. Our proof of Theorem 3.2.1 generalizes the one of (Tanielian and Biau, 2021, Theorem 2) by notably involving the next lemma:

Lemma 3.A.1: Approximation of 1-Lipschitz piecewise-linear functions

Let $f_1 \in \text{Lip}_1([0, 1]^d, \mathbb{R})$ and f_2 be a 1-Lipschitz piecewise-linear function on $[0, 1]^d$ such that $\|f_2 - f_1\|_\infty < \epsilon$. Then, f_2 can be expressed in the form (3.7) with

$$\max_{1 \leq i \leq m} |c_i| \leq \|f_1\|_\infty + \epsilon + \sqrt{d}.$$

Proof Note that we can suppose without loss of generality that for any $k \in \{1, \dots, m\}$ there exists a point $x_k \in \Omega$ such that $f_2(x_k) = a_k \cdot x_k + c_k$, otherwise this index is meaningless and we can eliminate it. Since $\|a_k\| \leq 1$, we have that $|c_k| \leq \|f_2\|_\infty + \sup_{x \in [0, 1]^d} \|x\|$. We conclude using the fact that $\|f_2\|_\infty \leq \|f_1\|_\infty + \epsilon$. ■

Let us now prove Theorem 3.2.1.

Proof of Theorem 3.2.1 Let $f \in \mathcal{F} \subset \text{Lip}_1(\Omega, \mathbb{R})$ such that $\sup_{f \in \mathcal{F}} \|f\|_\infty \leq K_{\mathcal{F}}$. The idea is to generalize (Tanielian and Biau, 2021, Theorem 2), restricted to 1-Lipschitz functions on the hypercube $[0, 1]^d$, to functions on the arbitrary compact set Ω . To this end, we first transform f into a 1-Lipschitz function on the hypercube $[0, 1]^d$.

Since Ω is compact then there exists some $R > 0$ such that $\Omega \subset [-R, R]^d$. Kirszbraun's theorem, see for instance (Heinonen, 2005, Theorem 2.5), implies that we can extend f on $[-R, R]^d$ while preserving the 1-Lipschitz property. Concretely, there exists a function $\tilde{f} \in \text{Lip}_1([-R, R]^d, \mathbb{R})$ such that $\tilde{f}(x) = f(x)$ for all $x \in \Omega$.

Now, we transform the extension \tilde{f} into a 1-Lipschitz function on the hypercube $[0, 1]^d$. This requires to translate and scale the inputs. Set $x_R = R \cdot \mathbf{1}$ where $\mathbf{1} := (1, \dots, 1) \in \mathbb{R}^d$, and define $f_R(x) := \frac{1}{2R} \tilde{f}(2Rx - x_R)$ as a function on $[0, 1]^d$. (Tanielian and Biau, 2021, Theorem 2) yields that, for every $\epsilon > 0$, there exists a neural network N of the form (3.2) satisfying Assumption (C) defined on $[0, 1]^d$ whose depth and size are respectively

$$l = O\left(d^2 \log_2\left(\frac{2\sqrt{d}}{\epsilon}\right)\right) \text{ and } s = O\left(\left(\frac{2\sqrt{d}}{\epsilon}\right)^{d^2}\right),$$

such that

$$\sup_{x \in [0, 1]^d} |f_R(x) - N(x)| < \epsilon. \quad (3.8)$$

However, [Tanielian and Biau \(2021\)](#) never clearly specified a universal bound C for which Assumption [\(C\)](#) was satisfied, which is necessary to conclude. To find such a bound, we detail how they constructed the GroupSort neural network N approximating f_R . First, note that according to [\(He et al., 2020, Theorem 5.1\)](#), any 1-Lipschitz piecewise-affine function q defined on a compact set can be written in [\(3.7\)](#). Second, following the proof of Theorem 2 in [\(Tanielian and Biau, 2021\)](#), one can find a 1-Lipschitz piecewise-affine function q such that $\|q - f_R\| \leq \epsilon$. Finally, [\(Tanielian and Biau, 2021, Theorem 1\)](#), states that q can be represented by a neural network N of the form [\(3.2\)](#) with depth l and size s . Critically, the representing N is built with weights $(W_1, \dots, W_l, b_1, \dots, b_l)$ such that the offset vectors of N are all equal to zero except b_1 . More precisely, the coefficients of b_1 are the constants c_1, \dots, c_m from the representation [\(3.7\)](#). This entails that $\max_{1 \leq i \leq l} \|b_i\|_\infty \leq \max_{1 \leq i \leq m} |c_i|$. Hence, bounding the constants in [\(3.7\)](#) will bound the offsets vectors in [\(3.2\)](#). To find a bound on the constants, we rely on Lemma [3.A.1](#), which implies that the ϵ -approximation q of f_R is such that $\max_{1 \leq i \leq m} |c_i| \leq K_{\mathcal{F}} + \epsilon + \sqrt{d}$. Consequently, the neural network N approximating f_R belongs to $\mathcal{N}_{C_0}^1(l, s)$ with $C_0 = K_{\mathcal{F}} + \epsilon + \sqrt{d}$.

Now, recall the the objective is to construct a neural network approximating f . Note that, after a change of variable, [\(3.8\)](#) can be written as

$$\sup_{x \in [-R, R]^d} \left| \tilde{f}(x) - 2RN \left(\frac{x + x_R}{2R} \right) \right| < 2R\epsilon.$$

Since the activation functions are GroupSort, hence homogeneous, we have that $2RN \left(\frac{x + x_R}{2R} \right) = N(x + x_R)$. This leads to

$$\|f - N_R\|_\infty \leq \sup_{x \in [-R, R]^d} \left| \tilde{f}(x) - N_R(x) \right| < 2R\epsilon,$$

Finally, remark that the neural network $N_R : x \mapsto N(x + x_R)$ belongs to $\mathcal{N}_C^1(l, s)$ with $C = \sqrt{d}R + C_0$ that is $\sqrt{d}(R + 1) + K_{\mathcal{F}} + \epsilon$. Setting $R = \sup_{x \in \Omega} \|x\|$ completes the proof. \blacksquare

Proof of Theorem [3.2.2](#) Let $g \in \mathcal{G} \subset \text{Lip}_1(\Omega, \mathbb{R}^p)$ such that $\sup_{g \in \mathcal{G}} \|g\|_\infty = K_{\mathcal{G}} > 0$. We generalize Theorem [3.2.1](#) to \mathbb{R}^p -valued output by approximating g along each dimension by a GroupSort neural network. The function g can be written as (g_1, \dots, g_p) where $g_i \in \text{Lip}_1(\Omega, \mathbb{R})$ and $\|g_i\|_\infty \leq K_{\mathcal{G}}$ for every $1 \leq i \leq p$. Then, we know from Theorem [3.2.1](#) that there exists a neural network $N^i \in \mathcal{N}_C^1$ where $C = K_{\mathcal{G}} + \sqrt{d}(\sup_{x \in \Omega} \|x\| + 1) + \epsilon$, whose depth and size are respectively

$$l = O \left(d^2 \log_2 \left(\frac{2\sqrt{d}}{\epsilon} \right) \right) \text{ and } s = O \left(\left(\frac{2\sqrt{d}}{\epsilon} \right)^{d^2} \right),$$

such that,

$$\|g_i - N^i\|_\infty \leq \epsilon.$$

We build the \mathbb{R}^p -valued neural network $N = (N^1, \dots, N^p)$. Then, for any $x \in \Omega$,

$$\|g(x) - N(x)\|^2 = \sum_{i=1}^p |g_i(x) - N^i(x)|^2 \leq p\epsilon^2.$$

As a consequence, $\|g - N\|_\infty \leq \sqrt{p}\varepsilon$. To conclude, note that N has depth l and size $p \times s$. Moreover, it satisfies Assumption **(C)** for the constant C , as the weight matrices and offset vectors of N are obtained by concatenation of the ones of the N^i , which preserves the upper-bound on the norms $\|\cdot\|_{2,\infty}$ and $\|\cdot\|_\infty$. Consequently, $N \in \mathcal{N}_C^p(l, p \times s)$. \blacksquare

Appendix 3.B Proofs of Section 3.3

Proof of Proposition 3.3.1 The proof amounts to showing that \mathcal{L}_n is continuous on the compact set \mathcal{G}_n .

Firstly, we note that the map $\mathcal{L}_n^{ot} : T \mapsto \|I - T\|_{L^2(P_n)}$ is continuous. Secondly, we prove that $\mathcal{L}_n^{gen} : T \mapsto \lambda_n \mathcal{W}_n(T_{\#}P_n, Q_n)$ is Lipschitz continuous. Let $T_1, T_2 \in \mathcal{C}(\Omega, \Omega)$ and compute,

$$\begin{aligned} |\mathcal{W}_n(T_{1\#}P_n, Q_n) - \mathcal{W}_n(T_{2\#}P_n, Q_n)| &\leq \left| \sup_{D \in \mathcal{D}_n} \left\{ \int D(T_1(x)) - D(T_2(x)) dP_n(x) \right\} \right| \\ &\leq \sup_{D \in \text{Lip}_1(\Omega, \mathbb{R})} \left| \int D(T_1(x)) - D(T_2(x)) dP_n(x) \right| \\ &\leq \int \|T_1(x) - T_2(x)\| dP_n(x) \\ &\leq \|T_1 - T_2\|_\infty. \end{aligned}$$

As a conclusion, $\mathcal{L}_n := \mathcal{L}_n^{ot} + \mathcal{L}_n^{gen}$ is continuous, and as such admits a minimizer on any compact set, in particular \mathcal{G}_n . \blacksquare

Appendix 3.C Proofs of Section 3.4

The proof of Theorem 3.3.1 relies on an intermediary result on the minimax estimator described in (Hütter and Rigollet, 2021, Section 5). Existence and statistical guarantees follow from the smoothness assumptions **(S1)** and **(S2)**.

Lemma 3.C.1: Properties of the minimax estimator

Assume that Assumptions **(S1)** and **(S2)** hold, and let T_n^{MM} be the minimax estimator from (Hütter and Rigollet, 2021) of the optimal transport map T_0 . It satisfies,

$$\|T_n^{\text{MM}} - I\|_{L^2(P_n)}^2 \xrightarrow[n \rightarrow +\infty]{a.s.} \|T_0 - I\|_{L^2(P)}^2. \quad (3.9)$$

Additionally, if Assumptions **(G1)**, **(G3)** and **(G4)** hold, then

$$\lambda_n \mathcal{W}_n(T_n^{\text{MM}} \# P_n, Q_n) \xrightarrow[n \rightarrow +\infty]{a.s.} 0, \quad (3.10)$$

hence,

$$\mathcal{L}_n(T_n^{\text{MM}}) \xrightarrow[n \rightarrow +\infty]{a.s.} \|T_0 - I\|_{L^2(P)}^2. \quad (3.11)$$

Proof Let's start by proving **(3.9)**. According to the triangle inequality,

$$\begin{aligned} \|T_n^{\text{MM}} - I\|_{L^2(P_n)} &\leq \|T_n^{\text{MM}} - T_0\|_{L^2(P_n)} + \|T_0 - I\|_{L^2(P_n)}, \\ &\leq \sqrt{\left| \int \|T_n^{\text{MM}} - T_0\|^2 (dP_n - dP) \right|} + \|T_n^{\text{MM}} - T_0\|_{L^2(P)} + \|T_0 - I\|_{L^2(P_n)}. \end{aligned}$$

We address each of the three terms of the upper bound in order. For the first term, recall that both T_n^{MM} and T are L -Lipschitz on Ω . Let's show that this entails that $x \mapsto \|T_n^{\text{MM}}(x) - T_0(x)\|^2$ is Lipschitz. For any $x, y \in \Omega$,

$$\begin{aligned} &\left| \|T_n^{\text{MM}}(x) - T_0(x)\|^2 - \|T_n^{\text{MM}}(y) - T_0(y)\|^2 \right| \\ &\leq 2 \text{diam}(\Omega) \left| \|T_n^{\text{MM}}(x) - T_0(x)\| - \|T_n^{\text{MM}}(y) - T_0(y)\| \right|, \\ &\leq 2 \text{diam}(\Omega) \|T_n^{\text{MM}}(x) - T_0(x) - T_n^{\text{MM}}(y) + T_0(y)\|, \\ &\leq 2 \text{diam}(\Omega) (\|T_n^{\text{MM}}(x) - T_n^{\text{MM}}(y)\| + \|T_0(x) - T_0(y)\|), \\ &\leq 2 \text{diam}(\Omega) (L\|x - y\| + L\|x - y\|), \\ &\leq 4L \text{diam}(\Omega) \|x - y\|. \end{aligned}$$

Denoting $L' = 4L \text{diam}(\Omega)$, we conclude that $x \mapsto \|T_n^{\text{MM}}(x) - T_0(x)\|^2$ belongs to $\text{Lip}_{L'}(\Omega, \mathbb{R})$. As a consequence,

$$\left| \int \|T_n^{\text{MM}} - T_0\|^2 (dP_n - dP) \right| \leq \sup_{f \in \text{Lip}_{L'}(\Omega, \mathbb{R})} \left| \int f (dP_n - dP) \right|.$$

The upper bound is a centered empirical process indexed by $\text{Lip}_{L'}(\Omega, \mathbb{R})$. According to [\(Van Der Vaart and Wellner, 1996, Corollary 2.7.2\)](#) and [\(Van Der Vaart and Wellner, 1996, Theorem 2.4.1\)](#), it tends to zero almost surely as n increases to infinity. This shows the convergence of the first term.

To control the second term we rely on [\(Hütter and Rigollet, 2021, Proposition 12\)](#). It states that with probability at least $1 - \delta$,

$$\|T_n^{\text{MM}} - T_0\|_{L^2(P)}^2 = \begin{cases} O\left(n^{-\frac{4}{2+d}} (\log n)^2 + \frac{\log \delta^{-1}}{n}\right) & \text{if } d > 2 \\ O\left(n^{-1} (\log n)^2 + \frac{\log \delta^{-1}}{n}\right) & \text{if } d = 2 \\ O\left(n^{-1} + \frac{\log \delta^{-1}}{n}\right) & \text{if } d = 1 \end{cases}$$

Hence,

$$\|T_n^{\text{MM}} - T_0\|_{L^2(P)} = \begin{cases} O\left(n^{-\frac{2}{2+d}}(\log n) + \sqrt{\frac{\log \delta^{-1}}{n}}\right) & \text{if } d > 2 \\ O\left(n^{-\frac{1}{2}}(\log n) + \sqrt{\frac{\log \delta^{-1}}{n}}\right) & \text{if } d = 2 \\ O\left(n^{-\frac{1}{2}} + \sqrt{\frac{\log \delta^{-1}}{n}}\right) & \text{if } d = 1 \end{cases} \quad (3.12)$$

Then, by setting $\delta_n = \frac{1}{n^2}$, it follows from Borel-Cantelli's theorem that $\|T_n^{\text{MM}} - T_0\|_{L^2(P)}$ tends almost-surely to zero. This shows the desired convergence of the second term. Moreover, as n increases to infinity, the third term of the upper bound tends almost surely to $\|T_0 - I\|_{L^2(P)}$, by weak convergence of P_n to P almost surely,

We now turn to the demonstration of (3.10). Let $D \in \mathcal{D}_n$ and write the following decomposition,

$$\begin{aligned} \int D \circ T_n^{\text{MM}} dP_n - \int D dQ_n &= \int D \circ T_n^{\text{MM}} d(P_n - P) + \int (D \circ T_n^{\text{MM}} - D \circ T_0) dP + \int D \circ T_0 dP \\ &\quad - \int D dQ_n \leq \left| \int D \circ T_n^{\text{MM}} d(P_n - P) \right| + \int \|T_n^{\text{MM}} - T_0\| dP + \left| \int D d(Q - Q_n) \right|, \end{aligned}$$

where we use that $\int D \circ T_0 dP = \int D dQ$ since $T_0 \# P = Q$. Noting that $\mathcal{D}_n \subseteq \text{Lip}_1(\Omega, \mathbb{R})$ we obtain,

$$\begin{aligned} \mathcal{W}_n(T_n^{\text{MM}} \# P_n, Q_n) &\leq \sup_{D \in \text{Lip}_1(\Omega, \mathbb{R})} \left| \int D \circ T_n^{\text{MM}} d(P_n - P) \right| + \int \|T_n^{\text{MM}} - T_0\| dP \\ &\quad + \sup_{D \in \text{Lip}_1(\Omega, \mathbb{R})} \left| \int D d(Q - Q_n) \right|. \end{aligned}$$

Recall now that T_n^{MM} is L -Lipschitz so that for any $D \in \mathcal{D}_n$ we have $D \circ T_n^{\text{MM}} \in \text{Lip}_L(\Omega, \mathbb{R})$. As a consequence,

$$\mathcal{W}_n(T_n^{\text{MM}} \# P_n, Q_n) \leq \sup_{g \in \text{Lip}_L(\Omega, \mathbb{R})} \left| \int g d(P_n - P) \right| + \int \|T_n^{\text{MM}} - T_0\| dP + \sup_{D \in \text{Lip}_1(\Omega, \mathbb{R})} \left| \int D d(Q - Q_n) \right|. \quad (3.13)$$

Next, we control each of the three terms of the upper bound in (3.13) with high probability.

Let us start with the first one, which is the supremum of a centered empirical process indexed by Lipschitz functions. Recall that P_n is supported by n independent variables $x_1, \dots, x_n \sim P$. Set $X \sim P$ and define

$$Z_n := \sup_{g \in \text{Lip}_L(\Omega, \mathbb{R})} \left| \frac{1}{n} \sum_{i=1}^n g(x_i) - \mathbb{E}g(X) \right| = \sup_{g \in \text{Lip}_L(\Omega, \mathbb{R})} \left| \int g d(P_n - P) \right|.$$

By L -Lipschitz continuity, changing x_i by an independent duplicate $x'_i \sim P$ changes Z_n of at most $\frac{1}{n}L \text{diam}(\Omega)$. Thus, it follows from MacDiarmid's inequality (Boucheron et al., 2013) that for any $t > 0$,

$$\mathbb{P}(Z_n \leq \mathbb{E}Z_n + t) \leq 1 - \exp\left(-\frac{2t^2}{\frac{1}{n}L^2 \text{diam}^2(\Omega)}\right).$$

After a change of variable, we get for every $0 < \delta < 1$,

$$\mathbb{P}(Z_n \leq \mathbb{E}Z_n + \frac{L \operatorname{diam}(\Omega)}{\sqrt{2n}} \sqrt{\log(\delta^{-1})}) \leq 1 - \delta.$$

(Schreuder, 2020, Theorem 4) provides an upper bound on $\mathbb{E}Z_n$. Up to logarithmic factors we have,

$$\mathbb{E}Z_n = \begin{cases} O\left(n^{-\frac{1}{d}}\right) & \text{if } d > 2 \\ O\left(n^{-\frac{1}{2}} \log n\right) & \text{if } d = 2 \\ O\left(n^{-\frac{1}{2}}\right) & \text{if } d = 1 \end{cases}$$

Hence, with probability at least $1 - \delta$,

$$Z_n = \begin{cases} O\left(n^{-\frac{1}{d}} + \sqrt{\frac{\log(\delta^{-1})}{n}}\right) & \text{if } d > 2 \\ O\left(n^{-\frac{1}{2}} \log n + \sqrt{\frac{\log(\delta^{-1})}{n}}\right) & \text{if } d = 2 \\ O\left(n^{-\frac{1}{2}} + \sqrt{\frac{\log(\delta^{-1})}{n}}\right) & \text{if } d = 1 \end{cases}$$

The third term of (3.13) can be bounded similarly, as the smoothness L only affects the hidden constant in the O . We now turn to the second term of (3.13). It follows from Cauchy-Schwarz inequality that

$$\int \|T_n^{\text{MM}} - T_0\| dP \leq \|T_n^{\text{MM}} - T_0\|_{L^2(P)}.$$

Recall that with probability at least $1 - \delta$, the right-term of this inequality is bounded as in (3.12).

By summing the bounds in probability holding for each of the three terms of (3.13), and after rescaling δ by 3, we obtain that with probability at least $1 - \delta$,

$$\mathcal{W}_n(T_n^{\text{MM}} \# P_n, Q_n) = \begin{cases} O\left(n^{-\frac{1}{d}} + n^{-\frac{4}{2+d}} (\log n) + \sqrt{\frac{\log \delta^{-1}}{n}}\right) & \text{if } d > 2 \\ O\left(n^{-\frac{1}{2}} (\log n) + \sqrt{\frac{\log \delta^{-1}}{n}}\right) & \text{if } d = 2 \\ O\left(n^{-\frac{1}{2}} + \sqrt{\frac{\log \delta^{-1}}{n}}\right) & \text{if } d = 1 \end{cases}$$

Now, we replace δ by $\frac{1}{n^2}$ and we multiply both sides of the inequality by λ_n so that with probability at least $1 - \frac{1}{n^2}$,

$$\lambda_n \mathcal{W}_n(T_n^{\text{MM}} \# P_n, Q_n) = \begin{cases} \lambda_n O\left(n^{-\frac{1}{d}} + n^{-\frac{4}{2+d}} \log n + \sqrt{\frac{\log(n)}{n}}\right) & \text{if } d > 2 \\ \lambda_n O\left(n^{-\frac{1}{2}} \log n + \sqrt{\frac{\log(n)}{n}}\right) & \text{if } d = 2 \\ \lambda_n O\left(n^{-\frac{1}{2}} + \sqrt{\frac{\log(n)}{n}}\right) & \text{if } d = 1 \end{cases}$$

Then, Assumption **(G3)** on λ_n implies that with probability at least $1 - \frac{1}{n^2}$,

$$\lambda_n \mathcal{W}_n(T_n^{\text{MM}} \# P_n, Q_n) = \begin{cases} o(1) + o\left(n^{-\frac{3d-2}{d(2+d)}} \log n\right) + o\left(n^{-\frac{d-2}{2d}} \sqrt{\log(n)}\right) & \text{if } d > 2 \\ o(1) + o\left(\frac{1}{\sqrt{\log n}}\right) & \text{if } d = 2 \\ o\left(\frac{1}{\sqrt{\log(n)}}\right) + o(1) & \text{if } d = 1 \end{cases}$$

We conclude, using Borel-Cantelli's theorem, that $\lim_{n \rightarrow +\infty} \lambda_n \mathcal{W}_n(T_n^{\text{MM}} \# P_n, Q_n) = 0$ almost surely. ■

We now turn to the proof of Theorem **3.3.1**, which will be divided in three steps.

Proof of Theorem 3.3.1

Recall that for any $n \in \mathbb{N}$, $G_n \in \mathcal{G}_n \subset \mathcal{G} := \text{Lip}_L(\Omega, B_L)$ according to Assumption **(G2)**. Since \mathcal{G} is a compact set, there exists a subsequence $\{G_{\varphi(n)}\}_{n \in \mathbb{N}}$ and some $G_\varphi \in \mathcal{G}$ such that $\|G_{\varphi(n)} - G_\varphi\|_\infty \xrightarrow[n \rightarrow +\infty]{a.s.} 0$. The goal of the proof is to show that $G_\varphi = T_0$ regardless of the extraction φ . For the sake of clarity, we will not track φ in the notations for the rest of the proof.

Moreover, note that since the minimax estimator T_n^{MM} belongs to \mathcal{G} , we know from Assumption **(G2)** and Theorem **3.2.2** that there exists a GroupSort neural network $G_n^{\text{MM}} \in \mathcal{G}_n$ such that $\|G_n^{\text{MM}} - T_n^{\text{MM}}\|_\infty \leq \varepsilon_n$. This neural network approximation of the minimax estimator will play a key role throughout the proof.

Step 1. In this first part, we aim at showing that $\lim_{n \rightarrow +\infty} \lambda_n \mathcal{W}_n(G_n \# P_n, Q_n) = 0$ almost surely when λ_n verifies Assumption **(G3)**. Let's assume ad absurdum that $\lambda_n \mathcal{W}_n(G_n \# P_n, Q_n)$ does not tend to zero. As $0 \in \mathcal{D}_n$, we have that $\mathcal{W}_n(G_n \# P_n, Q_n) > 0$ and consequently $\lim_{n \rightarrow +\infty} \lambda_n \mathcal{W}_n(G_n \# P_n, Q_n) = +\infty$. We will show a contradiction to this convergence.

Recall that $\|G_n^{\text{MM}} - T_n^{\text{MM}}\|_\infty \leq \varepsilon_n$, and that $G \mapsto \lambda_n \mathcal{W}_n(G \# P_n, Q_n)$ is λ_n -Lipschitz continuous. This leads to,

$$\begin{aligned} |\mathcal{L}_n(G_n^{\text{MM}}) - \mathcal{L}_n(T_n^{\text{MM}})| &\leq \lambda_n \left| \mathcal{W}_n(G_n^{\text{MM}} \# P_n, Q_n) - \mathcal{W}_n(T_n^{\text{MM}} \# P_n, Q_n) \right| + \|I - G_n^{\text{MM}}\|_{L^2(P_n)}^2 \\ + \|I - T_n^{\text{MM}}\|_{L^2(P_n)}^2 &\leq \lambda_n \|G_n^{\text{MM}} - T_n^{\text{MM}}\|_\infty + \text{diam}^2(\Omega) + \text{diam}^2(\Omega) \leq \lambda_n \varepsilon_n + 2 \text{diam}^2(\Omega). \end{aligned}$$

As G_n minimizes \mathcal{L}_n over \mathcal{G}_n , and since $G_n^{\text{MM}} \in \mathcal{G}_n$, we additionally have,

$$\mathcal{L}_n(G_n) \leq \mathcal{L}_n(G_n^{\text{MM}}) = \{\mathcal{L}_n(G_n^{\text{MM}}) - \mathcal{L}_n(T_n^{\text{MM}})\} + \mathcal{L}_n(T_n^{\text{MM}}).$$

Hence,

$$\begin{aligned} \lambda_n \mathcal{W}_n(G_n \# P_n, Q_n) + \|I - G_n\|_{L^2(P_n)} &\leq \{\lambda_n \varepsilon_n + 2 \text{diam}^2(\Omega)\} + \lambda_n \mathcal{W}_n(T_n^{\text{MM}} \# P_n, Q_n) \\ &\quad + \|I - T_n^{\text{MM}}\|_{L^2(P_n)}, \end{aligned}$$

leading to

$$0 \leq \lambda_n \mathcal{W}_n(G_n \# P_n, Q_n) \leq \lambda_n \varepsilon_n + 3 \text{diam}^2(\Omega) + \lambda_n \mathcal{W}_n(T_n^{\text{MM}} \# P_n, Q_n).$$

From Lemma 3.C.1, it follows that the right term is bounded, which contradicts the fact that $\lambda_n \mathcal{W}_n(G_{n\#}P_n, Q_n)$ tends to infinity. Consequently, $\mathcal{W}_n(G_{n\#}P_n, Q_n) \xrightarrow[n \rightarrow +\infty]{a.s.} 0$.

Step 2. Now, we prove that $G_{\#}P = Q$. Note that,

$$\begin{aligned}
& |\mathcal{W}_n(G_{n\#}P_n, Q_n) - \mathcal{W}(G_{\#}P, Q)| \\
& \leq \left| \sup_{D \in \mathcal{D}_n} \left(\int D \circ G_n dP_n - \int D dQ_n \right) - \left(\int D \circ G dP - \int D dQ \right) \right| \\
& + \left| \sup_{D \in \mathcal{D}_n} \left(\int D \circ G dP - \int D dQ \right) - \sup_{D \in \text{Lip}_1(\Omega, \mathbb{R})} \left(\int D \circ G dP - \int D dQ \right) \right|, \\
& \leq \left| \sup_{D \in \text{Lip}_1(\Omega, \mathbb{R})} \left(\int D \circ G_n dP_n - \int D dQ_n \right) - \left(\int D \circ G dP - \int D dQ \right) \right| \\
& + \left| \sup_{D \in \mathcal{D}_n} \left(\int D \circ G dP - \int D dQ \right) - \sup_{D \in \text{Lip}_1(\Omega, \mathbb{R})} \left(\int D \circ G dP - \int D dQ \right) \right|, \\
& \leq \left| \sup_{D \in \text{Lip}_1(\Omega, \mathbb{R})} \int D \circ G_n dP_n - \int D \circ G dP \right| + \left| \sup_{D \in \text{Lip}_1(\Omega, \mathbb{R})} \int D (dP_n - dP) \right| \\
& + \left| \sup_{D \in \mathcal{D}_n} \left(\int D \circ G dP - \int D dQ \right) - \sup_{D \in \text{Lip}_1(\Omega, \mathbb{R})} \left(\int D \circ G dP - \int D dQ \right) \right|.
\end{aligned}$$

The second term of the upper bound is the supremum of a centered empirical process indexed by the class of 1-Lipschitz functions, which tends to zero almost surely as n increases to infinity. The third term tends to zero according to Assumption **(G1)**. To address the first term, remark that for any $D \in \text{Lip}_1(\Omega, \mathbb{R})$,

$$D(G_n(x)) \leq \|G_n(x) - G(x)\| + D(G(x)).$$

Consequently,

$$\begin{aligned}
& \left| \sup_{D \in \text{Lip}_1(\Omega, \mathbb{R})} \int D \circ G_n dP_n - \int D \circ G dP \right| \leq \|G_n - G\|_{\infty} \\
& + \left| \sup_{D \in \text{Lip}_1(\Omega, \mathbb{R})} \int (D \circ G)(dP_n - dP) \right| \leq \|G_n - G\|_{\infty} + \left| \sup_{f \in \text{Lip}_L(\Omega, \mathbb{R})} \int f(dP_n - dP) \right|,
\end{aligned}$$

where we used the fact that $D \circ G \in \text{Lip}_L(\Omega, \mathbb{R})$, since $D \in \text{Lip}_1(\Omega, \mathbb{R})$ and $G \in \text{Lip}_L(\Omega, \Omega)$. By definition of G , we know that $\|G - G_n\|_{\infty} \xrightarrow[n \rightarrow +\infty]{a.s.} 0$. Moreover, the second term is here again the supremum of a centered empirical process indexed by Lipschitz functions, which tends to zero almost surely.

All in all, $|\mathcal{W}_n(G_{n\#}P_n, Q_n) - \mathcal{W}(G_{\#}P, Q)| \xrightarrow[n \rightarrow +\infty]{a.s.} 0$, and it follows from the first step that $\mathcal{W}(G_{\#}P, Q) = 0$, hence $G_{\#}P = Q$.

Step 3. We know that $G_{\#}P = Q$. To conclude that G is the unique optimal transport map T_0 between P and Q , we show that G minimizes the transportation cost. Firstly, we write,

$$\begin{aligned} & \left| \|I - G_n\|_{L^2(P_n)}^2 - \|I - G\|_{L^2(P)}^2 \right| \leq \left| \|I - G_n\|_{L^2(P_n)}^2 - \|I - G\|_{L^2(P_n)}^2 \right| \\ & \quad + \left| \|I - T_0\|_{L^2(P_n)}^2 - \|I - G\|_{L^2(P)}^2 \right|, \\ & \leq 2 \operatorname{diam}(\Omega) \|G_n - G\|_{\infty} + 2 \operatorname{diam}(\Omega) \left| \int \|T_0(x) - G(x)\|^2 (dP_n(x) - dP(x)) \right|. \end{aligned}$$

Hence,

$$\|I - G_n\|_{L^2(P_n)} \xrightarrow[n \rightarrow +\infty]{a.s.} \|I - G\|_{L^2(P)}. \quad (3.14)$$

Secondly, using that G_n minimizes \mathcal{L}_n on \mathcal{G}_n we have

$$\begin{aligned} \mathcal{L}_n(G_n) & \leq \mathcal{L}_n(G_n^{\text{MM}}), \\ & \leq \lambda_n \sup_{D \in \text{Lip}_1(\Omega, \mathbb{R})} \left\{ \int (D \circ G_n^{\text{MM}}) dP_n - \int D dQ_n \right\} + \|I - G_n^{\text{MM}}\|_{L^2(P_n)}^2, \\ & \leq \lambda_n \sup_{D \in \text{Lip}_1(\Omega, \mathbb{R})} \left\{ \int (D \circ G_n^{\text{MM}}) dP_n - \int (D \circ T_n^{\text{MM}}) dP_n \right\} \\ & \quad + \lambda_n \sup_{D \in \text{Lip}_1(\Omega, \mathbb{R})} \left\{ \int (D \circ T_n^{\text{MM}}) dP_n - \int D dQ_n \right\} + \|I - G_n^{\text{MM}}\|_{L^2(P_n)}^2, \\ & \leq \lambda_n \|T_n^{\text{MM}} - G_n^{\text{MM}}\|_{\infty} + \lambda_n \mathcal{W}_n(T_n^{\text{MM}}_{\#}P_n, Q_n) + \|I - G_n^{\text{MM}}\|_{L^2(P_n)}^2, \\ & \leq \lambda_n \varepsilon_n + \mathcal{L}_n(T_n^{\text{MM}}) + \|I - G_n^{\text{MM}}\|_{L^2(P_n)}^2 - \|I - T_n^{\text{MM}}\|_{L^2(P_n)}^2, \\ & \leq \lambda_n \varepsilon_n + \mathcal{L}_n(T_n^{\text{MM}}) + \|I - G_n^{\text{MM}}\|_{L^2(P_n)}^2 - \|I - T_n^{\text{MM}}\|_{L^2(P_n)}^2, \\ & \leq \lambda_n \varepsilon_n + \mathcal{L}_n(T_n^{\text{MM}}) + \left(\|I - G_n^{\text{MM}}\|_{L^2(P_n)} - \|I - T_n^{\text{MM}}\|_{L^2(P_n)} \right) \\ & \quad \times \left(\|I - G_n^{\text{MM}}\|_{L^2(P_n)} + \|I - T_n^{\text{MM}}\|_{L^2(P_n)} \right), \\ & \leq \lambda_n \varepsilon_n + \mathcal{L}_n(T_n^{\text{MM}}) + 2\varepsilon_n \operatorname{diam}(\Omega). \end{aligned}$$

This inequality can be written as,

$$\lambda_n \mathcal{W}_n(G_n_{\#}P_n, Q_n) + \|I - G_n\|_{L^2(P_n)}^2 \leq \mathcal{L}_n(T_n^{\text{MM}}) + \lambda_n \varepsilon_n + 2\varepsilon_n \operatorname{diam}(\Omega).$$

Then, according to the first step of the proof and the convergence (3.14), the left term tends almost surely to $\|I - G\|_{L^2(P)}^2$ as n increases to infinity. Besides, according to Lemma 3.C.1 and Assumptions (G4) and (G3), the right term tends to $\|I - T_0\|_{L^2(P)}^2$. Consequently,

$$\|I - G\|_{L^2(P)}^2 \leq \|I - T_0\|_{L^2(P)}^2.$$

This means that G minimizes the transportation cost. By uniqueness of the optimal transport map we conclude that $G = T_0$. This completes the proof. ■

Chapter 4

Diffeomorphic registration using Sinkhorn divergences

The diffeomorphic registration framework enables one to define an optimal matching function between two probability measures with respect to a data-fidelity loss function. The nonconvexity of the optimization problem renders the choice of this loss function crucial to avoid poor local minima. Recent work showed experimentally the efficiency of entropy-regularized optimal transportation costs, as they are computationally fast and differentiable while having few minima. Following this approach, we provide in this chapter a new framework based on Sinkhorn divergences, unbiased entropic optimal transportation costs, and prove the statistical consistency with rate of the empirical optimal deformations.

4.1 Introduction

Diffeomorphic deformations describe a large class of computational frameworks whose goal is to find optimal deformations of the ambient space, defined as a diffeomorphisms generated through flow equations (Joshi and Miller, 2000; Beg et al., 2005; Younes, 2019). They amount to solving an optimization problem involving two terms: an objective loss function characterizing in which sense the deformation should be optimal; a penalization over the kinetic energy spent by the transformation. The versatility of the problem formulation along with the appealing mathematical properties of diffeomorphisms made diffeomorphic deformations widely used in various application fields. In particular, they have been popularized for *diffeomorphic registration* in medical image analysis. This task consists of constructing diffeomorphic matching functions between shapes in order to establish spatial correspondences (Sotiras et al., 2013). More recently, Younes (2020) proposed to apply flows of diffeomorphisms in a machine-learning context, where the optimal deformation is designed to render the data classes linearly separable.

This chapter focuses on the diffeomorphic registration problem between two shapes. More specifically, we address the setting where the shapes are represented by probability measures: a formulation that has received a growing interest over the past few years to address unlabeled landmarks (Glaunes, 2005; Bauer et al., 2015; Feydy et al., 2017; Feydy and Trounev, 2018). In this case, the objective loss function, referred as the *data-fidelity loss*, is defined as a metric

between probability measures. Squares of *maximum mean discrepancies* (MMD), which are well-known kernel-based distances, became the canonical choice for such settings. In particular, their use for diffeomorphic registration enjoys a well-established theory (Glaunes et al., 2004; Glaunes, 2005; Younes, 2019). However, they also suffer from important practical drawbacks.

As pointed out by Feydy et al. (2017), the nonconvexity of the optimization problem on the diffeomorphic deformation renders the choice of the loss function crucial to avoid poor local minima, whereas an MMD possesses many. This is why they proposed to use optimal transport metrics as an alternative. More precisely, they define the data-fidelity loss as the entropy-regularized optimal transportation cost between unbalanced measures, which has two critical advantages. Firstly, it benefits from the nonlocality of optimal transport metrics, leading to few local minima. Secondly, entropic regularization alleviates the computational burden of standard optimal transport: it allows for fast computation and differentiation of the cost through the celebrated Sinkhorn’s algorithm (Cuturi, 2013). Nevertheless, while this alternative loss for diffeomorphic registration performs better experimentally, it lacks the statistical theory that was proven for squares of MMDs. Moreover, the entropic regularization induces a well-known bias making the loss not minimal between two identical measures. The latter issue motivates the employment of a *Sinkhorn divergence*: a symmetric unbiased version of the standard entropy-regularized optimal transportation cost. In Feydy et al. (2019), the authors showed that Sinkhorn divergences performed significantly better than their biased counterparts for registration purpose. However, they carried out their analysis using flows of gradients (an approach reviewed by Santambrogio (2017)) instead of flows of diffeomorphisms.

This chapter addresses *diffeomorphic* registration for Sinkhorn-divergence-based fidelity losses from both a theoretical and practical viewpoint. By leveraging some recent advances on these divergences (Feydy et al., 2019; Genevay et al., 2019), we show in a statistically-driven approach that the deformation obtained by solving the optimization problem between empirical measures converges with the parametric rate \sqrt{n} to its population counterpart, where n is the sample size. Additionally, we illustrate the practicality of our method through numerical experiments. This furnishes a new theoretically and practically grounded framework for diffeomorphic matching of probability measures.

Related work Several papers bear resemblances with our work as they combine entropic optimal transport with diffeomorphic registration at some point of their pipeline. Let us underline the major differences with our framework. The work of Croquet et al. (2021) leverages a Sinkhorn divergence as the data-fidelity loss of a regularized diffeomorphic-registration engine restricted to flows induced by *stationary velocity fields* (SVF), which are notoriously not tailored to match significantly different shapes (Arsigny et al., 2006). In contrast, our approach applies to the more flexible *large deformation diffeomorphic metric mapping* (LDDMM) framework where the flows are time dependent. In Shen et al. (2021), the authors also interface entropic optimal transport with large diffeomorphic deformations but for a different role: optimal transport computes a prior landmark alignment instead of acting as the data attachment term. The closest approach to ours is the one of Feydy et al. (2017) who first suggested to use entropic optimal transport as the data-fidelity loss for diffeomorphic registration. However, they relied on the *biased* transportation cost

between *unbalanced* measures whereas we tackle the *unbiased* divergence between probability distributions. Additionally, their work focuses on practical applications while we also provide theoretical background. Finally, one implementation of time-variant diffeomorphic registration driven by an unbiased Sinkhorn divergence can be found in a PhD manuscript (Feydy, 2020, Figure 4.6). In our work, we go further by filling the theoretical gap, as well as by proposing more comprehensive experiments illustrating the behaviour of these loss functions.

Outline The rest of the chapter is organized as follows. In Section 4.2, we specify the basic mathematical notations that will be used throughout the chapter. In Section 4.3, we set up the general problem we address by introducing the diffeomorphic registration framework for arbitrary data-fidelity losses. In Section 4.4, we present the necessary background on optimal transport and entropic regularization, in order to properly define Sinkhorn divergences. Additionally, we study some indispensable regularity properties of entropic optimal transport. In Section 4.5, we state our main results, that is the existence and statistical consistency of the optimal deformations. In Section 4.6, we recall the implementation of diffeomorphic registration, and present the numerical experiments where we benchmark Sinkhorn divergences with other losses. All the proofs are deferred to Appendix 4.B, while Appendix 4.A recalls key mathematical tools from empirical process theory and Frechet differentiability.

4.2 Preliminaries and notations

In this section, we introduce the definitions and notations that will be used throughout the chapter. The first part is dedicated to classes of smooth functions; the second one addresses probability measures.

4.2.1 Smooth functions

Let $d_1 \geq 1$ and \mathcal{X} be an arbitrary subset of \mathbb{R}^{d_1} with nonempty interior denoted by $\overset{\circ}{\mathcal{X}}$. For $p \geq 1$ and $d_2 \geq 1$, we define $\mathcal{C}^p(\mathcal{X}, \mathbb{R}^{d_2})$ as the set of p -continuously Frechet-differentiable functions from \mathcal{X} to \mathbb{R}^{d_2} . We also define $\mathcal{L}^p(\mathbb{R}^{d_1}, \mathbb{R}^{d_2})$ the set of symmetric p -multilinear operators from \mathbb{R}^{d_1} to \mathbb{R}^{d_2} . The p -th derivative of some $F \in \mathcal{C}^p(\mathcal{X}, \mathbb{R}^{d_2})$ is denoted by $F^{(p)}$. It maps any point $x \in \overset{\circ}{\mathcal{X}}$ to $F^{(p)}(x)[\cdot] \in \mathcal{L}^p(\mathbb{R}^{d_1}, \mathbb{R}^{d_2})$. By convention we set $F^{(0)} = F$. For any $L \in \mathcal{L}^p(\mathbb{R}^{d_1}, \mathbb{R}^{d_2})$, we define the operator norm as

$$\|L\|_{op} := \sup\{\|L[\delta_1, \dots, \delta_k]\| \mid \delta_i \in \mathbb{R}^{d_1}, \|\delta_i\| \leq 1\}$$

where $\|\cdot\|$ is the Euclidean norm. For example, if $F \in \mathcal{C}^1(\mathcal{X}, \mathbb{R})$, then $\|F'(x)\|_{op} = \|\nabla F(x)\|$ where ∇F is the gradient of F . This enables to define, for any $F \in \mathcal{C}^p(\mathcal{X}, \mathbb{R}^{d_2})$, the functional norm,

$$\|F\|_{p,\infty} := \max_{0 \leq k \leq p} \|F^{(k)}\|_{\infty},$$

where $\|F\|_\infty := \sup_{x \in \mathcal{X}} \|F(x)\|$, and $\|F^{(k)}\|_\infty := \sup_{x \in \mathcal{X}} \|F^{(k)}(x)\|_{op}$ for $k \geq 1$. In addition, for any $R > 0$ we denote by $\mathcal{C}_R^p(\mathcal{X}, \mathbb{R}^{d_2})$ the class of functions $F \in \mathcal{C}^p(\mathcal{X}, \mathbb{R}^{d_2})$ such that $\|F\|_{p,\infty} \leq R$, and write B_R for the centered Euclidean ball of radius R .

4.2.2 Actions on probability measures

We write $\mathbb{E}[X]$ for the expectation of any random variable X . The symbol \otimes denotes the product of measures. For two measures μ and ν on \mathbb{R}^d , the relation $\mu \ll \nu$ means that μ is absolutely continuous with respect to ν , that is $(\nu(E) = 0 \implies \mu(E) = 0)$ for every Borel set $E \subseteq \mathbb{R}^d$.

We define two kinds of actions involving probability measures. Let μ be a probability measure on \mathbb{R}^d and $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be a measurable function. The action of μ on f defines the real number:

$$\mu(f) := \int f d\mu = \mathbb{E}_{X \sim \mu}[f(X)].$$

Now, consider a measurable function $F : \mathbb{R}^d \rightarrow \mathbb{R}^d$. The action of F on μ defines a probability measure called the *push-forward* measure, defined as:

$$F_{\#}\mu := \mu \circ F^{-1}(\cdot).$$

If a random variable X follows the law μ , then the image variable $F(X)$ follows the law $F_{\#}\mu$. The push-forward operation enables to write changes of variables. Formally,

$$\int f d(F_{\#}\mu) = \int (f \circ F) d\mu.$$

4.3 Diffeomorphic measure transportation

In this section we present the necessary background on diffeomorphic registration of probability measures. We refer to (Younes, 2019) for a complete and precise treatment of this topic. Firstly, we recall how to define diffeomorphisms through flow equations. Secondly, we introduce the diffeomorphic measure transportation problem for arbitrary data-fidelity losses.

4.3.1 Generating diffeomorphic deformations

The diffeomorphic deformation framework can be framed as a fluid mechanics problem, where points in \mathbb{R}^d are transported by a vector field representing a stream varying across time in the ambient space. We begin by reviewing the corresponding formalism and theory.

For an integer $p \geq 1$ let \mathcal{B}_p be the space of functions in $\mathcal{C}^p(\mathbb{R}^d, \mathbb{R}^d)$ whose derivatives up to order p vanish to zero at infinity. This together with the norm $\|\cdot\|_{p,\infty}$ is a Banach space. Next, denote by V a Hilbert space with inner product $\langle \cdot, \cdot \rangle_V$ and norm $\|\cdot\|_V$, and assume that V is *continuously embedded* in \mathcal{B}_p . This corresponds to the hypothesis below.

¹Note that we allowed ourselves to clash with Part I, where V denoted a random variable, because this chapter is largely random-variable free.

Assumption (CE): Continuous embedding

The space V is included in \mathcal{B}_p , and there exists a constant $c_V > 0$ such that for any $v \in V$,

$$\|v\|_{p,\infty} \leq c_V \|v\|_V.$$

Physically, a function $v \in V$ represents a stationary vector field in the ambient space, specifying the speed vector $v(x) \in \mathbb{R}^d$ of the stream running at every position $x \in \mathbb{R}^d$. Then, define the class L_V^2 of vector fields $t \in [0, 1] \mapsto v_t \in V$ indexed by time and space satisfying $\int_0^1 \|v_t\|_V^2 dt < \infty$, which is a Hilbert space endowed with the inner product,

$$\langle v, u \rangle_{L_V^2} := \int_0^1 \langle v_t, u_t \rangle_V dt.$$

We recall that a sequence $\{v^n\}_{n \in \mathbb{N}}$ in L_V^2 converges weakly to v if for any $u \in L_V^2$,

$$\langle v^n, u \rangle_{L_V^2} \xrightarrow{n \rightarrow +\infty} \langle v, u \rangle_{L_V^2}. \quad (4.1)$$

The associated norm in L_V^2 is given by

$$\|v\|_{L_V^2} := \sqrt{\int_0^1 \|v_t\|_V^2 dt},$$

and we use the notation

$$L_{V,M}^2 := \{v \in L_V^2 \mid \|v\|_{L_V^2} \leq M\}$$

for the centered ball of radius $M > 0$ in L_V^2 .

We can now turn to the definition of diffeomorphic deformations. Any vector field $v \in L_V^2$ generates a deformation $\phi^v := (\phi_t^v)_{t \in [0,1]}$, function of both time and space variables, defined as the unique solution to the following *flow equation*,

$$\forall x \in \mathbb{R}^d, \forall t \in [0, 1], \quad \phi_t(x) = x + \int_0^t v_s(\phi_s(x)) ds. \quad (4.2)$$

The parametric curve $(\phi_t^v(x))_{t \in [0,1]}$ represents the trajectory across time of a point initially located at $\phi_0(x) = x \in \mathbb{R}^d$. Remarkably, for every $t \in [0, 1]$ the transformation ϕ_t^v is a p -continuously differentiable diffeomorphism. Moreover, as a direct consequence of (Glaunes, 2005, Theorem 5), these diffeomorphic transformations are smooth over compact sets.

Lemma 4.3.1: Smoothness of diffeomorphic deformations

Suppose that Assumption **(CE)** holds. Then for any radius $M > 0$ and any compact set $K \subset \mathbb{R}^d$, there exists a constant $R = R((K, d); (V, p); M) > 0$ such that for any vector field $v \in L_{V,M}^2$,

$$\max_{0 \leq k \leq p} \left\{ \sup_{t \in [0,1], x \in K} \left\| (\phi_t^v)^{(k)}(x) \right\|_{op} \right\} \leq R.$$

In particular, $\sup_{x \in K} \|x\| \leq R$.

In practice, the space of vector fields V is constructed through the choice of a kernel function. This is enabled by Assumption **(CE)** which entails that V is a reproducing kernel Hilbert space (RKHS), characterized by a unique nonnegative symmetric matrix-valued kernel function $\text{Ker} : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$. In particular, the choice of the kernel function sets the order of regularity p of the vector fields. For instance, the typical choice of a Gaussian kernel, that is

$$\text{Ker}(x, y) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right) I_d \quad (4.3)$$

where $\sigma > 0$ is the bandwidth parameter and I_d the identity matrix, leads to $p = +\infty$.

4.3.2 Diffeomorphic matching of distributions

In general, diffeomorphic deformation frameworks amount to finding solutions to Equation **(4.2)** that are optimal in some sense. In this work, we focus on the diffeomorphic measure transportation framework, which aims at matching two probability measures.

Formally, let Λ be a positive loss function between probability measures, and set α and β two distributions on the ambient space \mathbb{R}^d . For a given regularization weight $\lambda > 0$, an optimal matching function between α and β is a diffeomorphism ϕ^v solution to **(4.2)** where v minimizes

$$J_\lambda(v) := \Lambda(\phi_{1\#}^v \alpha, \beta) + \lambda \|v\|_{L_V^2}^2. \quad (4.4)$$

The first term of the objective function **(4.4)** is the *data-fidelity loss*, which tends to match $\phi_{1\#}^v \alpha$ with β , while the second term is the regularizer, which penalizes the kinetic energy spent by the trajectories $(\phi_t^v)_{t \in [0,1]}$, keeping them as close as possible to the identity function. The parameter λ governs the trade-off between the two contributions. The objective J_λ always admits minimizers provided that the term $v \in L_V^2 \mapsto \Lambda(\phi_{1\#}^v \alpha, \beta) \in \mathbb{R}^+$ is weakly continuous. For a minimizer v^* , the function $\phi_1^{v^*}$ is an optimal matching between α and β , and the family $(\phi_t^{v^*})_{t \in [0,1]}$ provides an approximated interpolation between the two measures.

In practical settings, one typically does not have access to the full probability measures α and β but to empirical observations. This naturally raises the question of estimating an optimal matching function between α and β on the basis of independent samples. Concretely, let $x_1, \dots, x_n \sim \alpha$ and $y_1, \dots, y_n \sim \beta$ be independent samples, and define the empirical probability measures $\alpha_n := n^{-1} \sum_{i=1}^n \delta_{x_i}$ and $\beta_n := n^{-1} \sum_{j=1}^n \delta_{y_j}$. Plugging these discrete measures in the original objective function **(4.4)** leads to the following empirical objective function:

$$J_{\lambda,n}(v) := \Lambda(\phi_{1\#}^v \alpha_n, \beta_n) + \lambda \|v\|_{L_V^2}^2. \quad (4.5)$$

In Theorem **4.5.1** we prove under some assumptions that if the data-fidelity loss Λ is a *Sinkhorn divergence*, a divergence derived from entropic optimal transport, then any sequence of minimizers $\{v^n\}_{n \in \mathbb{N}}$ of the empirical problem **(4.5)** converges up to the extraction of a subsequence to a minimizer of the population problem **(4.4)** as the sample size n increases to infinity.

4.4 Entropic optimal transport

In this section, we first briefly present the necessary background on optimal transport and entropic regularization, in order to properly define Sinkhorn divergences. We refer to (Villani, 2003, 2008; Peyré and Cuturi, 2019) for further insight on these topics. Then, we introduce some properties of these divergences, which will be useful to later demonstrate the main results of this chapter.

4.4.1 Transportation costs and Sinkhorn divergences

Let α and β be two probability measures on \mathcal{X} a subset of \mathbb{R}^d , and $c : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^+$ a positive ground cost function. Typically, $c(x, y) := \|x - y\|^2$. The optimal transportation cost with respect to c between α and β is defined as,

$$\mathcal{T}_c(\alpha, \beta) := \min_{\pi \in \Pi(\alpha, \beta)} \int_{\mathcal{X} \times \mathcal{X}} c(x, y) d\pi(x, y), \quad (4.6)$$

where $\Pi(\alpha, \beta)$ is the set of couplings admitting α as first marginal and β as second marginal. In particular, for an integer $k \geq 1$ and D a distance on \mathcal{X} , the quantity $(\mathcal{T}_{D^k})^{\frac{1}{k}}$ yields a distance between measures referred as the *Wasserstein distance* of order k . Transportation costs and optimal transport distances became popular in many machine-learning-related problems for their appealing geometric properties, but suffer from being computationally challenging in practice. This triggered a growing literature on fast approximations of (4.6), the most popular being entropy-regularized versions, which can be computed through the Sinkhorn algorithm (Cuturi, 2013). For $\varepsilon > 0$, the *entropy-regularized* transportation cost w.r.t. c is defined as

$$\mathcal{T}_{c, \varepsilon}(\alpha, \beta) := \min_{\pi \in \Pi(\alpha, \beta)} \int_{\mathcal{X} \times \mathcal{X}} c(x, y) d\pi(x, y) + \varepsilon \text{KL}(\pi | \alpha \otimes \beta), \quad (4.7)$$

where $\text{KL}(\mu | \nu)$ denotes the *Kullback-Leibler* divergence between probability measures μ and ν given by $\int \log\left(\frac{d\mu}{d\nu}(z)\right) d\mu(z)$ if $\mu \ll \nu$, and $+\infty$ otherwise.

Critically, the entropic transportation cost $\mathcal{T}_{c, \varepsilon}$ suffers from the so-called *entropic bias*, that is $\mathcal{T}_{c, \varepsilon}(\alpha, \alpha) \neq 0$ in general. As illustrated in (Feydy et al., 2019), this entails that the minimum of $\mathcal{T}_{c, \varepsilon}(\alpha, \cdot)$ is not reached at α but at a shrunken version of α with smaller support, making the entropic cost an unreliable loss function. The Sinkhorn divergence was originally introduced to fix this undesirable effect. It is formally defined as

$$S_{c, \varepsilon}(\alpha, \beta) := \mathcal{T}_{c, \varepsilon}(\alpha, \beta) - \frac{1}{2} \mathcal{T}_{c, \varepsilon}(\alpha, \alpha) - \frac{1}{2} \mathcal{T}_{c, \varepsilon}(\beta, \beta).$$

As aforementioned, using a nonlocal similarity measure such as an entropic-optimal-transport cost instead of a local similarity measure such as a squared MMD leads to fewer local solutions when minimizing (4.5). Moreover, it does not suffer from the computational burden of standard optimal transport. This is why Feydy et al. (2017) advocated the use of the entropy-regularized transportation cost (4.7) for diffeomorphic registration, providing empirical evidences of the benefits of this approach. However, they did not rely on the unbiased Sinkhorn divergences, for which little was known until (Feydy et al., 2019) that

demonstrated several key properties. In particular, if c is continuous, $e^{-\frac{c}{\varepsilon}}$ defines a positive universal kernel, and \mathcal{X} is compact, then $S_{c,\varepsilon}$ is symmetric positive definite, smooth and convex in each of its input distributions. Additionally, in contrast to the standard regularized transportation cost, it metrizes the convergence in law. In particular, these properties hold for the classical cost functions $c(x, y) := \|x - y\|$ and $c(x, y) := \|x - y\|^2$ defined on compact domains. The goal of this chapter is precisely to use a Sinkhorn divergence for the data-fidelity loss, while providing statistical guarantees. The demonstrations are based on the dual formulation of entropic optimal transport for which we derive some important results next.

4.4.2 Regularity of the dual formulation

The minimization problem (4.7) has the following dual formulation,

$$\mathcal{T}_{c,\varepsilon}(\alpha, \beta) = \sup_{f,g \in \mathcal{C}(\mathcal{X}, \mathbb{R})} \int_{\mathcal{X}} f(x) d\alpha(x) + \int_{\mathcal{X}} g(y) d\beta(y) - \varepsilon \int_{\mathcal{X}^2} e^{\frac{f(x)+g(y)-c(x,y)}{\varepsilon}} d\alpha(x) d\beta(y) + \varepsilon. \quad (4.8)$$

The functions f and g are referred as *potentials*. Note that Equation (4.8) can also be compactly written as,

$$\mathcal{T}_{c,\varepsilon}(\alpha, \beta) = \sup_{f,g \in \mathcal{C}(\mathcal{X}, \mathbb{R})} (\alpha \otimes \beta) \left(h_{c,\varepsilon}^{f,g} \right),$$

where

$$h_{c,\varepsilon}^{f,g}(x, y) := f(x) + g(y) - \varepsilon e^{\frac{f(x)+g(y)-c(x,y)}{\varepsilon}} + \varepsilon. \quad (4.9)$$

We call the function $h_{c,\varepsilon}^{f,g}$ the *global potential*. It will play a key role in the proofs.

A remarkable property of entropic optimal transport, investigated in (Genevay et al., 2019; Feydy et al., 2019), is that the potentials of the dual formulation inherit the regularity of the ground cost function c if the measures α and β are compactly supported. This setting will be useful to derive statistical guarantees. More specifically, it allows to restrict the set of feasible potentials to smooth functions regardless of the involved probability measures, as stated in the next lemma which readily follows from (Genevay et al., 2019, Proposition 1) (see also (del Barrio et al., 2022, Lemma 4.1) for the particular case of the quadratic ground cost).

Lemma 4.4.1: Smoothness of the optimal potentials

Let μ and ν be two measures on a compact set $K \subset \mathbb{R}^d$, and suppose that the ground cost function c belongs to $\mathcal{C}^q(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R}^+)$ with $q \geq 1$. Then, there exists a constant $m = m((K, d); (c, q); \varepsilon) > 0$ such that

$$\mathcal{T}_{c,\varepsilon}(\mu, \nu) = \sup_{f,g \in \mathcal{C}(K, \mathbb{R})} (\mu \otimes \nu) \left(h_{c,\varepsilon}^{f,g} \right) = \sup_{f,g \in \mathcal{C}_m^q(K, \mathbb{R})} (\mu \otimes \nu) \left(h_{c,\varepsilon}^{f,g} \right).$$

Naturally, the smoothness of f , g and c renders the global potential $h_{c,\varepsilon}^{f,g}$ smooth as well. Combining Lemma 4.4.1 with the following result ensures the smoothness of the optimal global potential under smooth data-processing transformations, such as diffeomorphic transformations.

Proposition 4.4.1: Smoothness of the optimal global potential

Let \mathcal{X} be a compact subset of \mathbb{R}^d , suppose that the ground cost function c belongs to $\mathcal{C}^q(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R}^+)$ with $q \geq 1$, set $p \geq 1$ and write $\kappa := \min\{p, q\}$. Then for any $m > 0$ and $R > 0$, there exists a constant $H = H(m; R; (c, q); \varepsilon; p) > 0$ such that for any $f, g \in \mathcal{C}_m^q(B_R, \mathbb{R})$ and $T_1, T_2 \in \mathcal{C}_R^p(\mathcal{X}, \mathbb{R}^d)$,

$$h_{c, \varepsilon}^{f, g} \circ (T_1, T_2) \in \mathcal{C}_H^\kappa(\mathcal{X} \times \mathcal{X}, \mathbb{R}).$$

We are now ready to state and prove our main results.

4.5 Main results

This section focuses on the main theoretical contributions of the chapter, namely the existence and statistical consistency of the empirical optimal matching function between α and β when using a Sinkhorn divergence.

Firstly, we show that the objective functions J_λ and $J_{\lambda, n}$ with $\Lambda = S_{c, \varepsilon}$ admit minimizers. We recall that a function $\Psi : L_V^2 \rightarrow \mathbb{R}$ is *weakly continuous* if for any sequence $\{v^n\}_{n \in \mathbb{N}}$ weakly converging to some $v \in L_V^2$ (see (4.1)), we have $\Psi(v^n) \xrightarrow[n \rightarrow +\infty]{} \Psi(v)$. (Glaunes, 2005, Theorem 7) states that J_λ admits a minimum if $v \in L_V^2 \mapsto \Lambda(\phi_{1\#}^v \alpha, \beta)$ is weakly continuous and nonnegative while (Feydy et al., 2019, Theorem 1) guarantees the nonnegativeness of Sinkhorn divergences when $e^{-\frac{c}{\varepsilon}}$ defines positive universal kernel. Therefore, existence of an optimal matching directly follows from the proposition below.

Proposition 4.5.1: Existence of the optimal vector fields

Let α and β be two probability measures on \mathcal{X} a compact subset of \mathbb{R}^d , suppose that the ground cost function c belongs to $\mathcal{C}^1(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R}^+)$, and assume that Assumption (CE) holds. Then the function $v \in L_V^2 \mapsto S_{c, \varepsilon}(\phi_{1\#}^v \alpha, \beta)$ is weakly continuous. If additionally $e^{-\frac{c}{\varepsilon}}$ defines a positive universal kernel, then J_λ for $\Lambda = S_{c, \varepsilon}$ admit minimizers.

The minimizer is not unique in general due to the nonconvexity of the data-fidelity loss with respect to v . Uniqueness could be artificially achieved by choosing λ very large, thereby rendering the objective function strictly convex, but this would make the purpose of the regularization meaningless.

We now turn to our main theorem, which is divided in two items. The first one ensures the convergences of the empirical solutions to their population counterparts; the second one specifies the speed of this convergence.

Theorem 4.5.1: Consistency of the optimal vector fields

Let α_n and β_n be empirical measures corresponding respectively to α and β , two probability measures on \mathcal{X} a compact subset of \mathbb{R}^d , suppose that the ground cost

function c belongs to $\mathcal{C}^q(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R}^+)$ with $q \geq 1$ and induces a positive universal kernel $e^{-\frac{c}{\varepsilon}}$. Finally, assume that Assumption **(CE)** holds. If, for any $n \in \mathbb{N}^*$, v^n denotes a minimizer of $J_{\lambda,n}$ for $\Lambda = S_{c,\varepsilon}$, then the following results hold.

- (i) There exists a minimizer of J_λ denoted by v^* such that up to the extraction of a subsequence

$$\|v^n - v^*\|_{L_V^2} \xrightarrow[n \rightarrow \infty]{a.s.} 0$$

and

$$\sup_{t \in [0,1]} \left\{ \left\| \phi_t^{v^n} - \phi_t^{v^*} \right\|_\infty + \left\| (\phi_t^{v^n})^{-1} - (\phi_t^{v^*})^{-1} \right\|_\infty \right\} \xrightarrow[n \rightarrow \infty]{a.s.} 0.$$

- (ii) If $\kappa := \min\{p, q\} > d$, then there exists a constant $A = A(\lambda; (\mathcal{X}, d); (c, q); \varepsilon; (V, p)) > 0$ such that

$$\mathbb{E} [|J_\lambda(v^n) - J_\lambda(v^*)|] \leq \frac{A}{\sqrt{n}}.$$

Note that [Glaunes et al. \(2004\)](#) proved a similar consistency result when the data-fidelity loss is the square of an MMD, but did not determine the speed of convergence as in (ii). The demonstration of (i) follows the steps of their proof (see [Glaunes, 2005](#), Theorem 16)). The idea is to show the convergence of $\sup_{v \in L_{V,M}^2} |J_{\lambda,n}(v) - J_\lambda(v)|$ as n increases to infinity, where $L_{V,M}^2$ contains all the minimizers independently of n . The main challenge when addressing an entropic optimal transport cost comes from the fact that it does not satisfy a triangle inequality, nor a data-processing inequality, and is hence harder to control. We remedy to this issue by proving and applying the following intermediary result:

Proposition 4.5.2: Consistency of entropic optimal transport

Let α_n and β_n be empirical measures corresponding respectively to α and β , two probability measures on \mathcal{X} a compact subset of \mathbb{R}^d , and suppose that the ground cost function c belongs to $\mathcal{C}^q(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R}^+)$ with $q \geq 1$. Set $p \geq 1$ and write $\kappa := \min\{p, q\}$. Then, the following results hold:

- (i) For any $R > 0$

$$\sup_{T_1, T_2 \in \mathcal{C}_R^p(\mathcal{X}, \mathbb{R}^d)} |\mathcal{T}_{c,\varepsilon}(T_{1\#}\alpha_n, T_{2\#}\beta_n) - \mathcal{T}_{c,\varepsilon}(T_{1\#}\alpha, T_{2\#}\beta)| \xrightarrow[n \rightarrow +\infty]{a.s.} 0.$$

- (ii) If $\kappa > d$, then for any $R > 0$ there exists a constant $A = A(R; (c, q); \varepsilon; (\mathcal{X}, d); p) > 0$ such that

$$\mathbb{E} \left[\sup_{T_1, T_2 \in \mathcal{C}_R^p(\mathcal{X}, \mathbb{R}^d)} |\mathcal{T}_{c,\varepsilon}(T_{1\#}\alpha_n, T_{2\#}\beta_n) - \mathcal{T}_{c,\varepsilon}(T_{1\#}\alpha, T_{2\#}\beta)| \right] \leq \frac{A}{\sqrt{n}}.$$

Notice that as a direct consequence of the triangle inequality, a similar result holds for $S_{c,\varepsilon}$. Hence, as diffeomorphisms are smooth on compact sets according to [Lemma 4.3.1](#), we can apply [Proposition 4.5.2](#) to control $\sup_{v \in L_{V,M}^2} |J_{\lambda,n}(v) - J_\lambda(v)|$.

Although Proposition 4.5.2 is motivated by diffeomorphic registration, we believe it has further interest. Remark in particular that the objective (4.4) shares similarities with generative modelling (Goodfellow et al., 2014); an input distribution α is passed through a parametric function ϕ_1^v meant to generate a target distribution β by minimizing a certain loss Λ . In particular, generative modelling using the Wasserstein-1 distance or a Sinkhorn divergence has proved to be efficient for diverse applications (Arjovsky et al., 2017; Genevay et al., 2018). The main difference in (4.4) comes from the parameter v being infinitely dimensional, and characterizing a diffeomorphism instead of a neural network. However, Proposition 4.5.2 is general enough to be applied in the context of generative modelling with Sinkhorn divergences, in order to derive statistical guarantees for smooth generators.

Remark 4.5.1: What about the biased cost?

Proposition 4.5.1 and Theorem 4.5.1 do not hold for $\mathcal{T}_{c,\varepsilon}$ instead of $S_{c,\varepsilon}$ because $v \mapsto \mathcal{T}_{c,\varepsilon}(\phi_{1\#}^v \alpha, \beta)$ is not lower bounded on L_V^2 . We also emphasize that it is preferable to use a Sinkhorn divergence in practice, since it does not suffer from the aforementioned entropic bias. In particular, the experiments from the next section illustrate that debiasing leads to more accurate registrations.

Remark 4.5.2: Similar results

Item (ii) in Proposition 4.5.2 resembles classical sampling complexity bounds of entropic optimal transport such as (Genevay et al., 2019, Theorem 3), (Séjourné et al., 2019, Theorem 7) and (Mena and Niles-Weed, 2019, Corollary 1). Our result differs critically by handling a supremum over a class of smooth push-forward maps within the expectation, which enables to prove item (ii) in Theorem 4.5.1.

4.6 Implementation

This section addresses the practical aspects of diffeomorphic registration through Sinkhorn divergence. Firstly, we briefly recall how to compute a minimizer of $J_{\lambda,n}$ for an arbitrary loss Λ . Then, we illustrate the procedure for Sinkhorn divergences on numerical experiments.

4.6.1 Resolution procedure

This subsection introduces the basic knowledge for solving a diffeomorphic registration problem. It is meant to keep the manuscript as self-contained as possible. Several minimization strategies coexist, corresponding to different parametrizations of the optimization problem 4.5. We refer to (Younes, 2019, Section 10.6) for a complete overview of the resolution procedures.

Gradient descent over the time-dependent momentum

To practically minimize $J_{\lambda,n}$, one must first write the optimal vector fields v in a finite parametric form, and then perform a gradient descent on the coefficients of this decomposition. Recall that Assumption **(CE)** implies that V is a RKHS, thereby characterized by a unique matrix-valued symmetric positive kernel function $\text{Ker} : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$. For simplicity, we address the case of the Gaussian kernel defined in **4.3**. Statistically, the bandwidth parameter σ represents the correlation between the morphed points; physically, it quantifies the fluid viscosity. When σ is small, the points have independent trajectories; when it is large, the points move as a whole.

The RKHS viewpoint enables to parametrize the optimal vector fields through a kernel trick. Firstly, note that the minimization of $J_{\lambda,n}$ can be formulated as an optimal control problem. It amounts to solving

$$\min_{v \in L_V^2} \Lambda(\alpha_n(1), \beta_n) + \lambda \|v\|_{L_V^2}^2; \text{ subject to } \alpha_n(t) = \phi_t^v \alpha_n \text{ for any } t \in [0, 1]. \quad (4.10)$$

Then, since the constraint involves a finite number n of trajectories, the so-called *reduction principle* (see **(Glaunes, 2005)**, Theorem 14)) entails that any solution to problem **4.10**, that is any minimizer of $J_{\lambda,n}$, can be written as,

$$v_t^n(x) = \sum_{i=1}^n \text{Ker}(x, z_i^a(t)) a_i(t),$$

where the *momentum* $a := (a_1, \dots, a_n)$ denotes n unspecified time functions of $L^2([0, 1], \mathbb{R}^d)$, and the *control trajectories* $z^a := (z_1^a, \dots, z_n^a)$ are defined by

$$z_i^a(t) = x_i + \int_0^t \sum_{j=1}^n \text{Ker}(z_i^a(s), z_j^a(s)) a_j(s) ds. \quad (4.11)$$

This enables to recast **(4.10)** as minimizing,

$$E_{\lambda,n}(a) := \Lambda\left(\frac{1}{n} \sum_{k=1}^n \delta_{z_k^a(1)}, \beta\right) + \lambda \int_0^1 \sum_{i,j=1}^n a_i(t) \cdot \text{Ker}(z_i^a(t), z_j^a(t)) a_j(t) dt, \quad (4.12)$$

where \cdot denotes the Euclidean inner product. The gradient of $E_{\lambda,n}$ was originally derived in **(Glaunes et al., 2004)** for the MMD case, and re-expressed in **(Glaunes, 2005; Younes, 2020)** for more general settings. It can be written as $\nabla E_{\lambda,n}(a) = 2\lambda a - p^a$ where $p^a := (p_1^a, \dots, p_n^a)$ denotes n functions of $L^2([0, 1], \mathbb{R}^d)$ satisfying for any $i \in \{1, \dots, n\}$ and $t \in [0, 1]$,

$$\begin{aligned} p_i^a(t) &:= \nabla_{z_i^a(1)} \Lambda\left(\frac{1}{n} \sum_{k=1}^n \delta_{z_k^a(1)}, \beta\right) \\ &- \frac{1}{\sigma^2} \int_t^1 \sum_{j=1}^n \text{Ker}(z_i^a(t), z_j^a(t)) [a_i(t) \cdot p_j^a(t) + a_j(t) \cdot p_i^a(t) - 2\lambda a_i(t) \cdot a_j(t)] (z_i^a(t) - z_j^a(t)). \end{aligned} \quad (4.13)$$

In order to practically track all the functions of the continuous time variable, one must discretize the time scale $[0, 1]$ into τ sub-intervals of equal sizes, which recasts a , z^a and p^a as $(\tau + 1) \times n \times d$ tensors. Then, equations (4.11) and (4.13) are successively solved at each iteration of the gradient descent by solving the associated discrete dynamical systems. By plugging the solutions z^a and p^a into the formula of $\nabla E_{\lambda, n}(a)$ one can update the variable a with $a \leftarrow a - \xi \times (2\lambda a - p^a)$ where ξ denotes the step size. The computational complexity of an iteration is in $O(n^2 d \tau)$. However, the dynamical systems can be parallelized in the number of points and the dimension. At the end of the process, we obtain the following deformation,

$$\phi_t^{a, \tau}(x) := x + \frac{1}{\tau} \sum_{s=0}^{t-1} \sum_{j=1}^n \text{Ker}(x, z_j^a(s)) a_j(s). \quad (4.14)$$

This approach handles any data-fidelity loss Λ as long as it is differentiable with respect to the data points of the discrete distributions. Both Sinkhorn divergences and squares of MMDs satisfy this property.

Geodesic shooting of the initial momentum

A widely used variant of the above approach is the *geodesic shooting of the initial momentum* which relies on the equations satisfied at the minimum to uniquely constrain the time-dependent solution $a(\cdot)$ by its initial value, allowing for optimizing solely over $a(0)$.

More specifically, as demonstrated in (Miller et al., 2006), the Hamiltonian viewpoint of the control problem yields the following joint dynamic of the optimal control trajectories and momentum:

$$\begin{aligned} z_i^a(t) &= x_i + \int_0^t \sum_{j=1}^n \text{Ker}(z_i^a(s), z_j^a(s)) a_j(s) ds, \\ a_i(t) &= a(0) - \frac{1}{2} \nabla_{z_i^a(t)} \int_0^t \left(\sum_{j=1}^n a_i(s) \cdot \text{Ker}(z_i^a(s), z_j^a(s)) a_j(s) \right) ds. \end{aligned} \quad (4.15)$$

This entails that both the control trajectories and the momentum at any instant t are fully characterized by $a(0)$. Slightly abusing notations we write $z^a = z^{a(0)}$.

Additionally, the kinetic energy remains constant along optimal solutions, implying that

$$\int_0^1 \sum_{i,j=1}^n a_i(t) \cdot \text{Ker}(z_i^{a(0)}(t), z_j^{a(0)}(t)) a_j(t) dt = \sum_{i,j=1}^n a_i(0) \cdot \text{Ker}(x_i, x_j) a_j(0). \quad (4.16)$$

Therefore, (4.16) together with (4.15) enable to recast the functional (4.12) to minimize as

$$E_{\lambda, n}^0(a(0)) := \Lambda \left(\frac{1}{n} \sum_{k=1}^n \delta_{z_k^{a(0)}(1)}, \beta \right) + \lambda \sum_{i,j=1}^n a_i(0) \cdot \text{Ker}(x_i, x_j) a_j(0), \quad (4.17)$$

which is a well-defined function of the time-invariant parameter $a(0) \in \mathbb{R}^{n \times d}$ only. After minimizing (4.17) using a gradient-descent-based method, one can *shoot* the obtained $a(0)$ along the discretized system (4.15) to generate the optimal control trajectories $z^a(\cdot)$ and

time-dependent momentum $a(\cdot)$. Then, the trajectory of any new point $x \in \mathbb{R}^d$ at any time $t \in [0, 1]$ can be computed by integrating the flow equation as in (4.14).

Naturally, for a nonconvex program such as (4.5) the quality of the output solution may heavily depend on the chosen resolution procedure. In the coming experiments, we compare the deformations obtained with both solving strategies.

4.6.2 Numerical experiments

We present a series of numerical experiments on synthetic and real 2-D and 3-D shapes. The objective is to illustrate the practical benefits of using a Sinkhorn divergence as the data-fidelity loss. Our Python code² operates with the GeomLoss package (Feydy et al., 2019) to compute the losses and their gradients by automatic differentiation, and the KeOps package (Charlier et al., 2021) to handle kernel-reduction operations. It is largely inspired by the example codes from these libraries' websites.³

2-D dataset

In (Feydy et al., 2019), the authors proposed an alternative measure registration framework based on the gradient flow of the data-fidelity loss. It amounts to updating the source distribution $\alpha_n := n^{-1} \sum_{i=1}^n \delta_{x_i}$ by carrying out a gradient descent on $\Lambda(\alpha_n, \beta_n)$ with respect to the positions x_1, \dots, x_n . This model-free method enables to faithfully match one distribution to another, even when the supports have irregularities such as holes. In this section, we firstly adapt their experiments, more precisely the ones from the example section of the GeomLoss package website, by using diffeomorphic deformations instead of gradient flows.

The objective is matching two blob-like point clouds in dimension 2. We proceed as follows. Firstly, we learn the optimal matching between two samples of size $n = 1,000$ using each of the two previously described procedures. Secondly, we display the obtained time interpolation between two new independent samples of size $m = 2,000$. In order to benchmark the influence of the data-fidelity loss, we consider a fixed setting where V is defined through a Gaussian kernel with bandwidth $\sigma = 0.175$, the regularization has weight $\lambda = 10^{-8}$, and the time scale is uniformly divided into $\tau = 16$ intervals. Then, we compare the results for different losses: (unbiased) Sinkhorn divergences, biased entropic transportation costs, and squared Gaussian maximum mean discrepancies. Recall that the squared Gaussian MMD with bandwidth parameter $\theta > 0$ is defined as,

$$\text{MMD}_\theta^2(\mu, \nu) := \int_{\mathbb{R}^d \times \mathbb{R}^d} \exp\left(-\frac{\|x - y\|^2}{2\theta^2}\right) d(\mu - \nu)(x) d(\mu - \nu)(y).$$

The ground cost function for the Sinkhorn divergences is always $c(x, y) := \|x - y\|^2$ throughout the experiments. Figures 4.1 to 4.3 compare the optimal matchings obtained with respectively the gradient descent on the momentum (GDM) and geodesic shooting (GS) for different values of the relevant parameters ε and θ . Note that whatever the minimization

²<https://github.com/lucaselara/lldmm-sinkhorn/>

³<https://www.kernel-operations.io/geomloss/> and <https://www.kernel-operations.io/keops/>

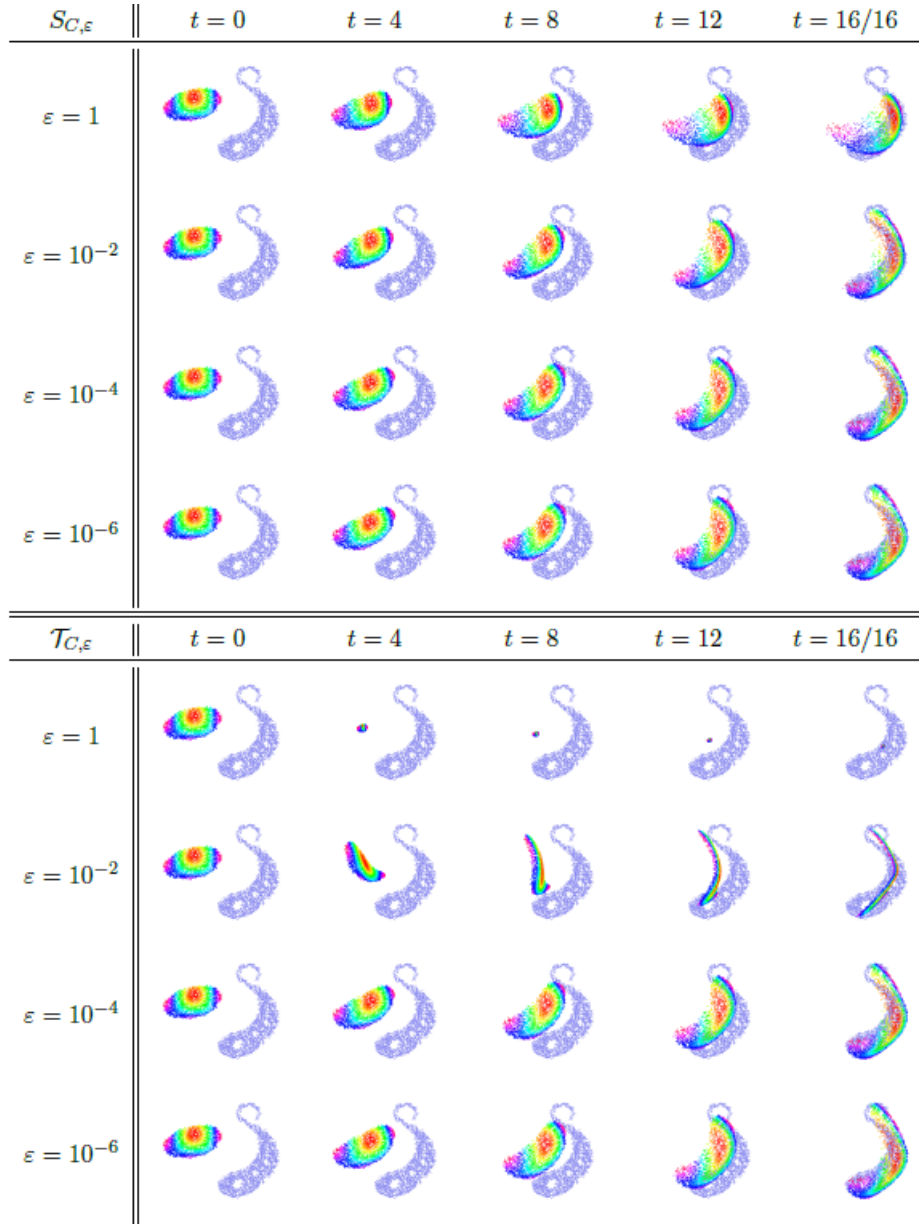


Figure 4.1: Optimal-transport-driven 2-D diffeomorphic registration optimized by GDM. The colored distribution is $\alpha_m(t)$, while the blue distribution is β_m .

strategy, we used a fixed number of iterations with a constant learning rate, and initialized the momentum with the zero tensor. Also, while we programmed a standard gradient descent for GDM, we relied on the PyTorch (Paszke et al., 2019) in-built L-BFGS solver for the geodesic shooting. The results are arranged as follows: Figure 4.1 shows the deformations for both Sinkhorn divergences and (biased) entropic transportation costs optimized with GDM; Figure 4.2 is the counterpart of Figure 4.1 for GS; Figure 4.3 displays the deformations generated by Gaussian maximum mean discrepancies for both resolution procedures.

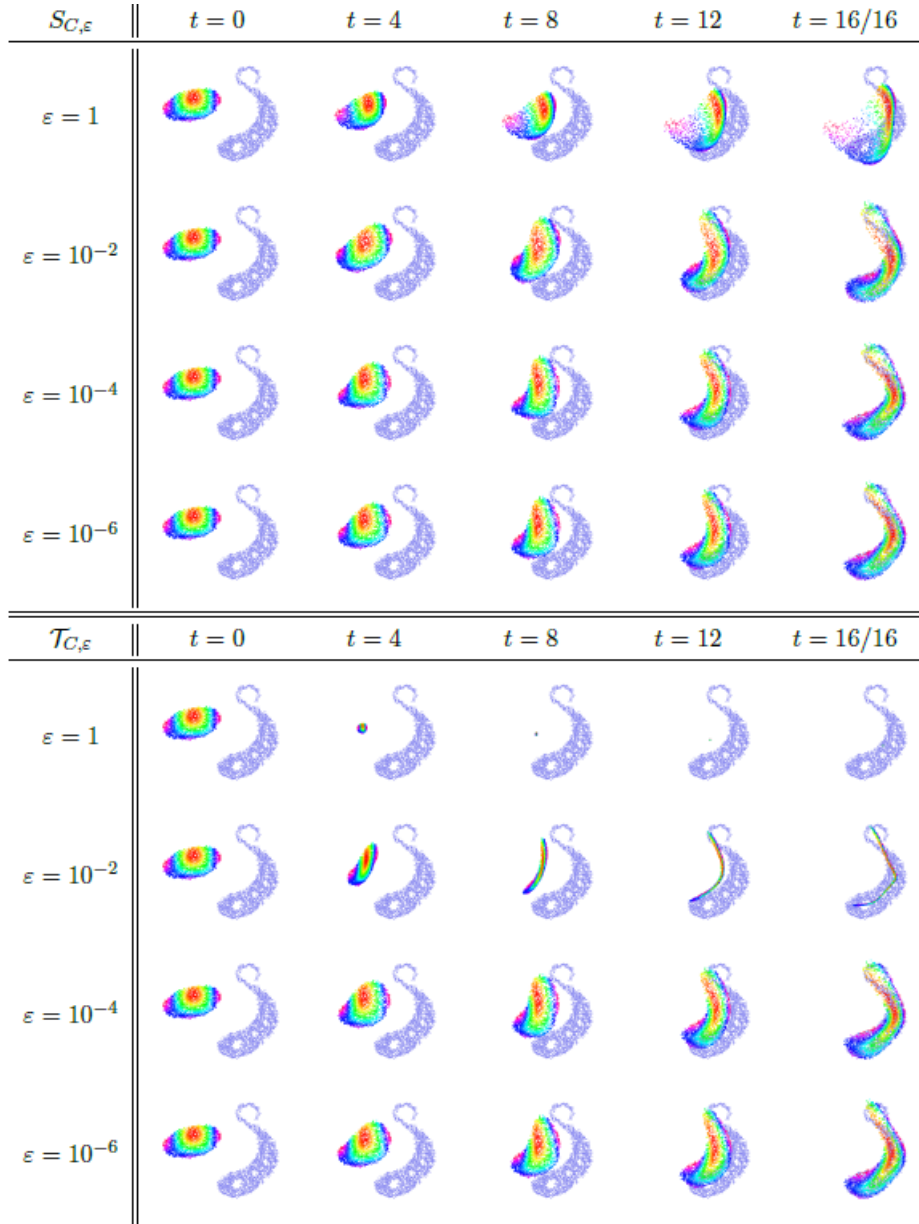


Figure 4.2: Optimal-transport-driven 2-D diffeomorphic registration optimized by GS. The colored distribution is $\alpha_m(t)$, while the blue distribution is β_m .

Firstly, we observe from Figures 4.1 and 4.2 that entropic optimal-transport metrics yield consistent results across minimization strategies. In contrast, the registration for maximum mean discrepancies depicted in Figure 4.3 varies with the chosen method. This instability of the optimization problem underlines that MMDs give more local minima.

Secondly, Figures 4.1 and 4.2 clearly exhibit the entropic bias: in contrast to Sinkhorn divergences, standard entropic transportation costs shrink the morphed distribution for large values of the regularization parameter ε , leading to unacceptable registrations. However,

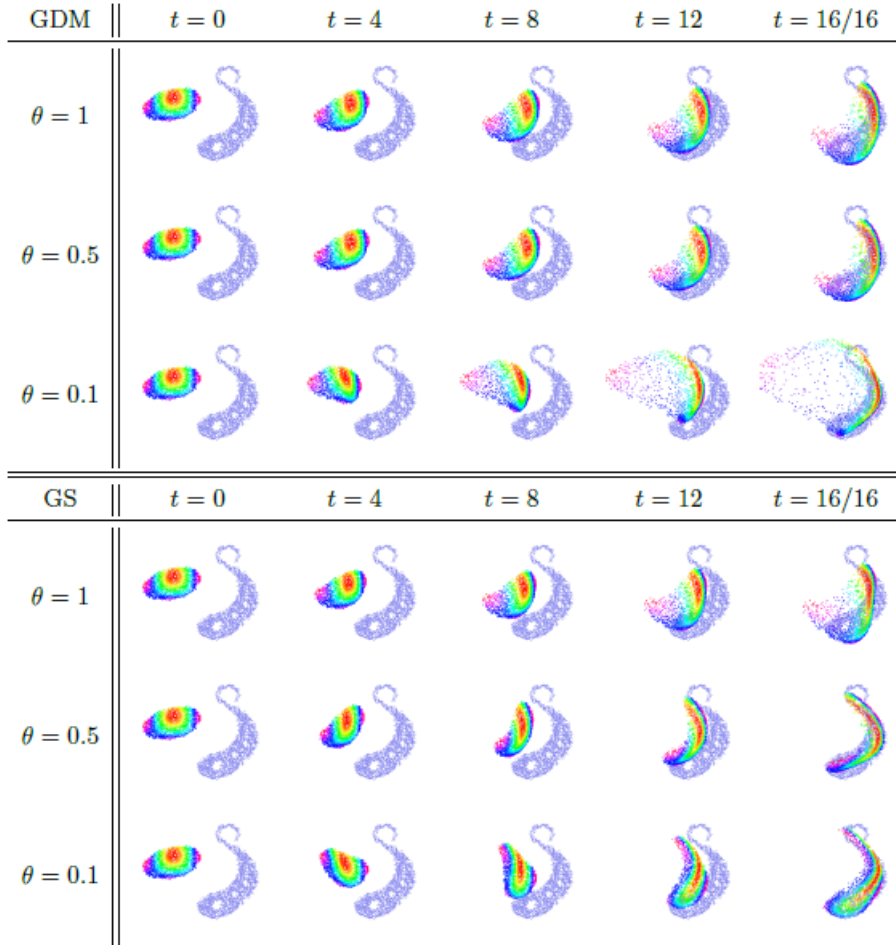


Figure 4.3: 2-D diffeomorphic registration driven by MMD_θ^2 . The colored distribution is $\alpha_m(t)$, while the blue distribution is β_m .

choosing a too large ε for the unbiased divergence yields a blurry, poorly accurate solution. As expected, debiasing becomes less critical as the regularization diminishes, and both entropic losses provide sharp matchings for small values of ε . Note also that there is no need to decrease ε below a certain threshold to ensure accurate deformations.

Finally, Figure 4.3 indicates that the consistency of the results between resolution procedures weakens as the bandwidth of the Gaussian kernel decreases. This is due to Gaussian maximum mean discrepancies ignoring disparities smaller than θ . As such, setting a large bandwidth facilitates the registration but degrades the quality of the matching. In contrast, a small bandwidth allows for sharper registration but induces more local minima. This aspect is epitomized for $\theta = 0.1$ in the experiments: with the gradient descent on the time-dependent momentum, the morphed points end up diverging, trapped into minimizing the auto-correlation contribution of the MMD, while geodesic shooting produces a fine matching.

All in all, our experimental observations about the role of the losses are similar to the ones made by Feydy et al. (2019) in the context of gradient flows. Critically, compared to their approach, we work with a transformation that is smooth at any time. This regularity constraint reduces the flexibility of the matching, which leads to a less accurate fitting than gradient flows. This affects particularly the anomalous parts of the targeted support, namely the holes and the tail. In contrast, regularity enables the deformation to generalize to any new out-of-sample observations. Additionally, it prevents from tearing the mass apart. The color map on the distribution $\alpha_m(t)$ enables to track the location of the moved points through time. Notice that, as a direct consequence of the smoothness, the chromatic continuity between morphed points is preserved throughout the process.

Before turning to more complex 3-D shapes, let us push further the quality analysis of local minima on this illustrative dataset. In the sequel, we consider the same setting as before, and focus on the optimal matchings obtained by geodesic shooting for Sinkhorn divergences and Gaussian maximum mean discrepancies with different parameter values. However, instead of initializing the optimized variable $a(0)$ to zero, we now study the stability and accuracy of the solutions over various initial values. More specifically, we rely on a warm-start strategy: solutions from the above experiments are reused as starting points in the solver. The results are gathered in Figure 4.4, which reports the final matchings obtained with different initializations along with their associated loss values.

Let us firstly analyze the results for the losses that previously gave the finest registrations: the Sinkhorn divergence with $\varepsilon = 10^{-4}$ (rows 1 and 5) and the Gaussian MMD with $\theta = 0.1$ (rows 3 and 7). As anticipated, the matchings vary with the initialization. Visually, this phenomenon is stronger for the MMD than for the Sinkhorn divergence and the quality of the final matchings remains quite accurate for the optimal-transport loss. By checking the loss values, we note that the warm start downgrades the solutions for both losses, except for the MMD using initialization via $S_{c,\varepsilon}$ with $\varepsilon = 10^{-4}$ which gets significantly closer to the global minimum. In sum, it seems that the entropic divergence induces fewer or better local minima. Regarding the Sinkhorn divergence with $\varepsilon = 1$ (rows 2 and 6) and the MMD with $\theta = 0.5$ (rows 4 and 8), which previously yielded imprecise matchings, they have analogous behaviours with respect to warm start. We observe that the results are less robust to initialization and can be significantly improved by using already accurate solutions as starting points, underlining that the registrations obtained with the initialization to zero corresponded to bad local minima.

3-D surfaces

Next, we implement the diffeomorphic matching of two shapes embedded in \mathbb{R}^3 : the source is the unit sphere while the target is the centered scaled Stanford bunny,⁴ both encoded through the associated uniform distributions. Similarly to the above experiments, we firstly learn the diffeomorphism on a training set of size $n = 5,000$ using geodesic shooting for various losses, and then display the final matching on a testing set of size $m = 10,000$. The setup is characterized by $\sigma = 0.05$, $\lambda = 10^{-8}$, and $\tau = 16$. The results can be found in Figure 4.5. We make comparable observations to before. Powering diffeomorphic registration with a Sinkhorn divergence instead of the biased regularized cost avoids the shrinkage effect

⁴<http://graphics.stanford.edu/data/3Dscanrep/>



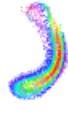
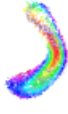










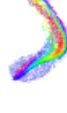
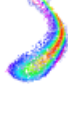
Loss \ Init	Zero	$S_{C,\varepsilon}$ $\varepsilon = 10^{-4}$	$S_{C,\varepsilon}$ $\varepsilon = 1$	MMD_θ^2 $\theta = 0.1$	MMD_θ^2 $\theta = 0.5$
$S_{C,\varepsilon}$ $\varepsilon = 10^{-4}$					
$S_{C,\varepsilon}$ $\varepsilon = 1$					
MMD_θ^2 $\theta = 0.1$					
MMD_θ^2 $\theta = 0.5$					
$S_{C,\varepsilon}$ $\varepsilon = 10^{-4}$	$7.35 \cdot 10^{-5}$		$1.01 \cdot 10^{-4}$ (+37, 4%)	$9.77 \cdot 10^{-5}$ (+32, 9%)	$7.72 \cdot 10^{-5}$ (+5, 0%)
$S_{C,\varepsilon}$ $\varepsilon = 1$	$1.61 \cdot 10^{-6}$	$3.06 \cdot 10^{-8}$ (-98, 1%)		$5.26 \cdot 10^{-8}$ (-96, 7%)	$3.60 \cdot 10^{-8}$ (-97, 8%)
MMD_θ^2 $\theta = 0.1$	$3.39 \cdot 10^{-4}$	$1.01 \cdot 10^{-4}$ (-70, 2%)	$4.06 \cdot 10^{-4}$ (+19, 8%)		$3.59 \cdot 10^{-4}$ (+5, 9%)
MMD_θ^2 $\theta = 0.5$	$1.55 \cdot 10^{-6}$	$6.30 \cdot 10^{-8}$ (-95, 9%)	$3.34 \cdot 10^{-6}$ (+115, 5%)	$1.23 \cdot 10^{-7}$ (-92, 1%)	

Figure 4.4: 2-D diffeomorphic matching optimized by GS with warm start. A row specifies the studied loss while each column refers to an initialization. Except for the first column which indicates the initialization to zero, the columns correspond to a solution from the previous experiments. The first four rows show the final matchings while the last four rows give the associated loss values with their growth rates compared to the initialization via zero; row-wise-minimal loss values are written in bold.

of the entropic bias for large values of ε , and the matchings are accurate for both losses when ε is small. The Gaussian MMD requires a small bandwidth θ to potentially fit the bunny, but the solution falls into a poor local minima where several morphed points are not attracted by the target. Note also that, due to their regularity, the deformations tend to smooth the sharpest edges of the target bunny.

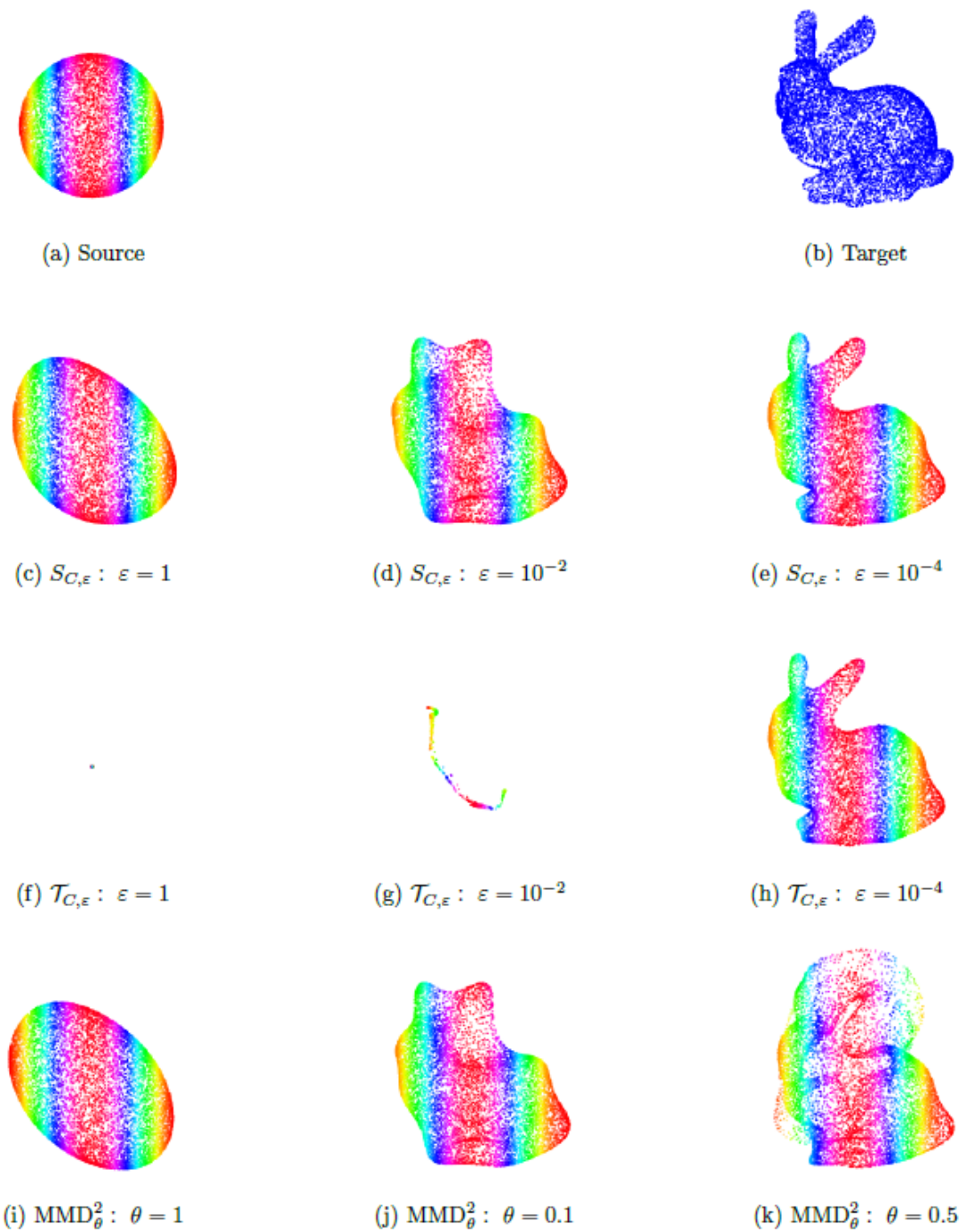


Figure 4.5: 3-D diffeomorphic matchings. Both shapes (a) and (b) are centered scaled.

4.7 Final remarks

Similarly to the previous chapters, we conclude by discussing the interests of our contributions and proposing lines of further research.

4.7.1 Application to counterfactual reasoning

As mentioned in Chapter 2, diffeomorphic registration could serve as a transport-based counterfactual model. Let us comment on this idea.

First of all, note that diffeomorphic registration can only define approximate counterfactual models in contrast to optimal transport, as the generated matching never perfectly fits the target distribution (i.e., $\phi_{1\#}^v \alpha \approx \beta$). Nevertheless, by relaxing the push-forward constraint, it can furnish deterministic matching functions whatever the probability distributions. Moreover, the optimization procedure has a $O(n^2)$ complexity instead of the $O(n^3 \log(n))$ of optimal transport.

Besides a computational comparison, diffeomorphic registration does not interpret counterfactual counterparts by the minimization of a cost, but through the choice of a space of vector fields—or equivalently a kernel function. The kernel function can be seen as a local similarity measure regulating the interactions between the counterfactual correspondences. This could open interesting debates on the semantics carried by such counterfactual models. However, the choice of the kernel function is what ensures an accurate registration in practice. Therefore, adding constraints of interpretation on this choice would make the matching process too complicated.

Another issue that we faced when trying to use diffeomorphisms as structural counterfactual operators comes from the impossibility to visualize the matching when the dimension of the data exceeds 3. Even though the data-fidelity loss gives an idea of how close the morphed and target distributions are, this only makes sense relatively to another value. Classical applications of diffeomorphic registration deals with 2-D or 3-D shapes, for which one can visually assess the quality of the matching. This cannot be achieved in typical fairness tasks. To sum-up, it did not feel reliable to use diffeomorphic registration as a matching process from one world to another. For all these reasons, we leave the application of diffeomorphic mappings to counterfactual reasoning, and more generally to algorithmic fairness and explainability, for further research.

4.7.2 Conclusion

We proposed to use Sinkhorn divergences as the fidelity loss in diffeomorphic registration problems. We derived the statistical theory, and illustrated the efficiency of this method compared to past approaches based on MMDs or *biased* entropic transportation costs. As such, this chapter paves way for accurate and smooth measure registration with certifiable asymptotic guarantees. Moreover, carrying out this work led us to further investigate the dual formulation of entropic optimal transport, complementing recent papers on the subject. A first avenue for extension could be to consider the registration of *unbalanced* measures using Sinkhorn divergences, which would align with the work of Feydy et al. (2017). A second one could be to derive sharper rates of convergences. Notably, (del Barrio et al., 2022) which demonstrates faster convergence rates for the empirical entropic transportation potentials

and (Chizat et al., 2020) which shows that debiasing decreases the approximation error of optimal transport induced by entropic regularization could serve as inspirations.

Appendix 4.A Preliminary results

This section recalls some useful results. Section [4.A.1](#) contains a brief reminder on entropy numbers of classes of functions, in order to derive an upper bound on empirical processes; Section [4.A.2](#) focuses on the chain rule for composite Frechet derivatives up to arbitrary high orders.

4.A.1 Empirical processes

In the proof of Proposition [4.5.2](#), we will bound the sampling error between the empirical entropic transportation cost and its population counterpart by a centered empirical process indexed by a class of smooth functions. Recalling the theory introduced in ([Van Der Vaart and Wellner, 1996](#); [Koltchinskii, 2011](#)), we present in this subsection intermediary results on such processes.

Let \mathcal{X} be a compact convex subset of \mathbb{R}^d . For any probability measure μ on \mathcal{X} and $r \geq 1$, we define the $L_r(\mu)$ -norm on $\mathcal{C}(\mathcal{X}, \mathbb{R})$ as $\|h\|_{r,\mu} := (\int |h|^r d\mu)^{1/r}$. In empirical process theory, the complexity of classes of functions is commonly evaluated through the so-called *covering* and *bracketing* numbers. Let \mathcal{H} be a class of function included in $\mathcal{C}(\mathcal{X}, \mathbb{R})$, and $\epsilon > 0$ a constant. The covering number $N(\epsilon, \mathcal{H}, L_r(\mu))$ is defined as the minimal number of $L_r(\mu)$ -balls of radius ϵ needed to cover the class of functions \mathcal{H} . The center of the balls need not belong to \mathcal{H} , but must have finite norm. Additionally, given two functions l and u with finite norm but not necessarily in \mathcal{H} , the *bracket* $[l, u]$ is the set of all functions h such that $l \leq h \leq u$. An $(\epsilon, L_r(\mu))$ -bracket is a bracket $[l, u]$ such that $\|l - u\|_{r,\mu} \leq \epsilon$. Then, the bracketing number $N_{[\cdot]}(\epsilon, \mathcal{H}, L_r(\mu))$ is the minimal number of $(\epsilon, L_r(\mu))$ -bracket needed to cover \mathcal{H} .

These numbers have essential applications in statistics. The supremum of a centered empirical process indexed by a class of functions with a finite bracketing number converges uniformly almost-surely to zero. Moreover, with a sharper control on the bracketing number, one can derive the following convergence rate:

Proposition 4.A.1: Empirical processes indexed by smooth functions

Let μ_n be an empirical version of a probability measure μ on a compact convex subset \mathcal{X} of \mathbb{R}^d , and set $H > 0$ a constant. Consider the class of functions $\mathcal{H} := \mathcal{C}_H^\kappa(\mathcal{X}, \mathbb{R})$ for some integer $\kappa \geq 0$. If $\kappa > d/2$, then there exists a constant $A = A((H, \kappa); (\mathcal{X}, d))$ such that,

$$\mathbb{E} \left[\sup_{h \in \mathcal{H}} |\mu_n(h) - \mu(h)| \right] \leq \frac{A}{\sqrt{n}}.$$

Proof Combining ([Koltchinskii, 2011](#), Theorem 2.1) with ([Koltchinskii, 2011](#), Theorem 3.11), we directly have that,

$$\mathbb{E} \left[\sup_{h \in \mathcal{H}} |\mu_n(h) - \mu(h)| \right] \leq 2 \times \frac{b}{\sqrt{n}} \mathbb{E} \int_0^{2\sigma_n} \sqrt{\log N(\epsilon, \mathcal{H}, L_2(\mu_n))} d\epsilon,$$

where $b > 0$ is some constant and $\sigma_n := \sup_{h \in \mathcal{H}} \mu_n(h^2)$. By definition of \mathcal{H} , it follows that $\sigma_n \leq H^2$. Besides, we can upper bound the covering number in the right term by the

bracketing number $N_{[\cdot]}(2\epsilon, \mathcal{H}, L_2(\mu_n))$ (see (Van Der Vaart and Wellner, 1996, page 84)). In addition, according to (Van Der Vaart and Wellner, 1996, Corollary 2.7.2), there exists a constant $\rho = \rho((H, \kappa); (\mathcal{X}, d)) > 0$ such that,

$$\log N_{[\cdot]}(2\epsilon, \mathcal{H}, L_2(\mu_n)) \leq \rho(2\epsilon)^{-d/\kappa}.$$

Note that the right term does not depend on μ_n . All in all,

$$\mathbb{E} \left[\sup_{h \in \mathcal{H}} |\mu_n(h) - \mu(h)| \right] \leq \frac{2c}{\sqrt{n}} \mathbb{E} \int_0^{2H^2} \sqrt{\rho(2\epsilon)^{-d/\kappa}} d\epsilon.$$

The integral is finite as $\kappa > d/2$. Consequently, the upper bound defines a constant $A = A((H, \kappa); (\mathcal{X}, d))$. This concludes the proof. ■

Remark that the convexity assumption on the compact domain \mathcal{X} is not restrictive, as it suffices to extend the probability measure μ on the convex hull of \mathcal{X} .

4.A.2 Frechet derivative

The proof of Proposition 4.4.1 requires bounding the Frechet derivatives of arbitrary high orders of composite functions. We rely on the generalization of Faà di Bruno's formula proposed by (Clark and Houssineau, 2013) to carry out the computation.

Let $F : \mathbb{R}^{d_2} \rightarrow \mathbb{R}^{d_3}$ and $G : \mathbb{R}^{d_1} \rightarrow \mathbb{R}^{d_2}$ be two differentiable functions up to order $k \geq 1$. Denote by $\Omega(k)$ the set of partitions of $\{1, \dots, k\}$, and write $|\cdot|$ for the cardinality of a set. For any $\delta := (\delta_1, \dots, \delta_k) \in (\mathbb{R}^{d_1})^k$, $x \in \mathbb{R}^{d_1}$, and $\omega := \{\omega_1, \dots, \omega_{|\omega|}\} \in \Omega(k)$, we define $\delta_{\omega_i}^G(x) := G^{(|\omega_i|)}(x) [(\delta_j)_{j \in \omega_i}]$ for every $1 \leq i \leq |\omega|$. Then, according to (Clark and Houssineau, 2013, Theorem 2.1),

$$(F \circ G)^{(k)}(x) [\delta_1, \dots, \delta_k] = \sum_{\omega \in \Omega(k)} F^{(|\omega|)}(G(x)) \left[\delta_{\omega_1}^G(x), \dots, \delta_{\omega_{|\omega|}}^G(x) \right]. \quad (4.18)$$

This results implies a chain rule on the operator norms of derivatives of composite functions, which will greatly simplify the computations of later proofs.

Proposition 4.A.2: Control of high-order composite derivatives

Let $F : \mathbb{R}^{d_2} \rightarrow \mathbb{R}^{d_3}$ and $G : \mathbb{R}^{d_1} \rightarrow \mathbb{R}^{d_2}$ be two differentiable functions up to order $k \geq 1$. Then, for any $x \in \mathbb{R}^{d_1}$,

$$\left\| (F \circ G)^{(k)}(x) \right\|_{op} \leq \sum_{\omega \in \Omega(k)} \left\| F^{(|\omega|)}(G(x)) \right\|_{op} \times \prod_{1 \leq i \leq |\omega|} \left\| G^{(|\omega_i|)}(x) \right\|_{op}.$$

Proof According to the triangle inequality and (4.18)

$$\left\| (F \circ G)^{(k)}(x) \right\|_{op} \leq \sum_{\omega \in \Omega(k)} \sup_{\|\delta_1\|, \dots, \|\delta_k\| \leq 1} \left\| F^{(|\omega|)}(G(x)) \left[\delta_{\omega_1}^G(x), \dots, \delta_{\omega_{|\omega|}}^G(x) \right] \right\|.$$

Then, we can bound the right term of this inequality by,

$$\sum_{\omega \in \Omega(k)} \left\| F^{(|\omega|)}(G(x)) \right\|_{op} \times \prod_{1 \leq i \leq |\omega|} \sup_{\|\delta_1\|, \dots, \|\delta_k\| \leq 1} \left\| \delta_{\omega_i}^G(x) \right\|_{op}.$$

In addition, note that for any $x \in \mathbb{R}^{d_1}$,

$$\sup_{\|\delta_1\|, \dots, \|\delta_k\| \leq 1} \left\| \delta_{\omega_i}^G(x) \right\| \leq \left\| G^{(|\omega_i|)}(x) \right\|_{op}.$$

Therefore,

$$\left\| (F \circ G)^{(k)}(x) \right\|_{op} \leq \sum_{\omega \in \Omega(k)} \left\| F^{(|\omega|)}(G(x)) \right\|_{op} \times \prod_{1 \leq i \leq |\omega|} \left\| G^{(|\omega_i|)}(x) \right\|_{op}.$$

■

Appendix 4.B Proofs of Chapter 4

This sections details all the mathematical proofs of the chapter.

Proof of Lemma 3.2 Let us start with a preliminary remark. For any $v \in L_V^2$, it follows from Assumption **(CE)** that $\int_0^1 \|v_t\|_{p,\infty} dt \leq c_V \int_0^1 \|v_t\|_V dt$. Besides, by Cauchy-Schwarz inequality $\int_0^1 \|v_t\|_V dt \leq \|v\|_{L_V^2}$, leading to $\int_0^1 \|v_t\|_{p,\infty} dt \leq c_V \|v\|_{L_V^2}$.

We now turn to the proof. Recall that by definition $\phi_t^v(x) = x + \int_0^t v_s \circ \phi_s^v(x) ds$. Consequently, by the triangle inequality we have for any compact set $K \subset \mathbb{R}^d$ that

$$\sup_{t \in [0,1], x \in K} \left\| \phi_t^v(x) \right\| \leq \sup_{x \in K} \|x\| + \int_0^1 \|v_s\|_\infty ds \leq \sup_{x \in K} \|x\| + c_V \|v\|_{L_V^2}.$$

Therefore,

$$\sup_{v \in L_{V,M}^2, t \in [0,1], x \in K} \left\| \phi_t^v(x) \right\| \leq \sup_{x \in K} \|x\| + c_V M.$$

Moreover, combining ([Glaunes, 2005](#), Theorem 5) with the preliminary remark, we know that for any $1 \leq k \leq p$, there exist two positive constants c_k and c'_k such that for any $v \in L_V^2$,

$$\sup_{t \in [0,1]} \left\| (\phi_t^v)^{(k)} \right\|_\infty \leq c_k \exp \left(c'_k \|v\|_{L_V^2} \right).$$

Hence,

$$\sup_{v \in L_{V,M}^2, t \in [0,1]} \left\| (\phi_t^v)^{(k)} \right\|_\infty \leq c_k \exp \left(c'_k M \right).$$

Then, setting

$$R((K, d); (V, p); M) := \max \left\{ \max_{1 \leq k \leq p} \{c_k \exp(c'_k M)\}, \sup_{x \in K} \|x\| + c_V M \right\}$$

concludes the proof. ■

Proof of Lemma 4.1 Let μ and ν be probability measures on a compact set $K \subset \mathbb{R}^d$. In a first time, let us show that optimal potentials $(f, g) \in \mathcal{C}(K, \mathbb{R}) \times \mathcal{C}(K, \mathbb{R})$ for $\mathcal{T}_{c, \varepsilon}(\mu, \nu)$ can be chosen as universally-bounded Lipschitz functions. The optimality condition on the potentials (see for instance (Genevay, 2019)) can be written as,

$$\exp\left(-\frac{f(x)}{\varepsilon}\right) = \int_K \exp\left(\frac{g(y) - c(x, y)}{\varepsilon}\right) d\nu(y).$$

Remark that since c is continuously differentiable, f is therefore continuously differentiable. Differentiating both sides of this expression leads to,

$$\nabla f(x) = \int_K \nabla_1 c(x, y) \exp\left(\frac{f(x) + g(y) - c(x, y)}{\varepsilon}\right) d\nu(y),$$

where ∇_1 denotes the gradient with respect to x , the first variable of c . Let us define $\Gamma_{c, \varepsilon}^{f, g}(x, y) := \exp\left(\frac{f(x) + g(y) - c(x, y)}{\varepsilon}\right)$. According to the primal-dual relationship (Genevay, 2019, Proposition 7), an optimal solution π to the primal problem has the expression,

$$d\pi(x, y) = \Gamma_{c, \varepsilon}^{f, g}(x, y) d\mu(x) d\nu(y).$$

Since by definition $\pi \in \Pi(\mu, \nu)$, we consequently obtain that $\int_K \Gamma_{c, \varepsilon}^{f, g}(x, y) d\nu(y) = 1$. Therefore,

$$\|\nabla f\|_\infty \leq \sup_{x, y \in K} \|\nabla_1 c(x, y)\|.$$

A similar argument can be made for g . This shows that f and g are ℓ -Lipschitz with $\ell = \ell((K, d); c) > 0$. Now, note that for any constant $b \in \mathbb{R}$, the pair $(f + b, g - b)$ is still a pair of optimal potentials. As a consequence, they can be chosen without loss of generality such that $f(x_0) = 0$ for a given $x_0 \in K$. Thus, using the Lipschitz property we get $f(x) \leq \ell \|x - x_0\|$, hence $\|f\|_\infty \leq \ell \text{diam}(K)$. To bound g , we use (Genevay et al., 2019, Proposition 1) which states that $\inf_{x \in K} \{f(x) - c(x, y)\} \leq g(y) \leq \sup_{x \in K} \{f(x) - c(x, y)\}$. This entails that $\|g\|_\infty \leq \|f\|_\infty + \sup_{x, y \in K} |c(x, y)| \leq \ell \text{diam}(K) + \sup_{x, y \in K} |c(x, y)|$. All in all, there exists a constant $\ell_1 = \ell_1((K, d); c)$ such that f and g are ℓ_1 -bounded and ℓ_1 -Lipschitz continuous.

Analogously, one can bound the successive derivatives of f and g up to order q , the maximum order of differentiability of c , using (Genevay et al., 2019, Proposition 1). In particular, this result ensures that for any $1 \leq k \leq q$, both $\|f^{(k)}\|_\infty$ and $\|g^{(k)}\|_\infty$ are bounded by a polynomial in ε^{-1} whose coefficients depend only on c and K . This implies that there exists a constant $m = m((K, d); (c, q); \varepsilon) > 0$ such that f and g belong to $\mathcal{C}_m^q(K, \mathbb{R})$. ■

Proof of Proposition 4.2

Let $m > 0$ and $R > 0$. Set $f, g \in \mathcal{C}_m^q(B_R, \mathbb{R})$. Note that the function $h_{c, \varepsilon}^{f, g}$ belongs to $\mathcal{C}^q(B_R \times B_R, \mathbb{R})$. In a first time, we do not focus on any data processing operations, and show that $h_{c, \varepsilon}^{f, g}$ and its derivatives up to order q are uniformly bounded. By definition,

$$h_{c,\varepsilon}^{f,g}(x,y) = f(x) + g(y) - \varepsilon \exp\left(\frac{f(x) + g(y) - c(x,y)}{\varepsilon}\right) + \varepsilon.$$

Before going further, we define the constant

$$C_\infty(R) := \max_{0 \leq k \leq q} \left\{ \sup_{(x,y) \in B_R \times B_R} \left\| c^{(k)}(x,y) \right\|_{op} \right\} \quad (4.19)$$

Then, using the triangle inequality and the bounds on f, g and c we obtain,

$$\left\| h_{c,\varepsilon}^{f,g} \right\|_\infty \leq \|f\|_\infty + \|g\|_\infty + \varepsilon \exp\left(\frac{\|f\|_\infty + \|g\|_\infty + C_\infty}{\varepsilon}\right) + \varepsilon \leq 2m + \varepsilon \exp\left(\frac{2m + C_\infty}{\varepsilon}\right) + \varepsilon.$$

Notice that the upper bound does not depend on the choice of f and g . We prove similar bounds for arbitrary high orders of derivatives using the chain rule. We divide the problem by studying the function,

$$\Gamma_{c,\varepsilon}^{f,g} : (x,y) \in B_R \times B_R \mapsto \exp\frac{f(x) + g(y) - c(x,y)}{\varepsilon},$$

which is κ -continuously differentiable. Using Proposition [4.A.2](#) with $F = \exp$, we obtain for any $1 \leq k \leq q$,

$$\left\| \left(\Gamma_{c,\varepsilon}^{f,g} \right)^{(k)}(x,y) \right\|_{op} \leq \left| \Gamma_{c,\varepsilon}^{f,g}(x,y) \right| \sum_{\omega \in \Omega(k)} \prod_{1 \leq i \leq |\omega|} \varepsilon^{-1} \left\| f^{(|\omega_i|)}(x) + g^{(|\omega_i|)}(y) - c^{(|\omega_i|)}(x,y) \right\|_{op},$$

Then,

$$\left\| \left(\Gamma_{c,\varepsilon}^{f,g} \right)^{(k)} \right\|_\infty \leq \exp\left(\frac{2m + C_\infty(R)}{\varepsilon}\right) \sum_{\omega \in \Omega(k)} \varepsilon^{-|\omega|} (2m + C_\infty(R))^{|\omega|}. \quad (4.20)$$

We now turn back to $h_{c,\varepsilon}^{f,g}$. Since $\left(h_{c,\varepsilon}^{f,g} \right)^{(k)} = f^{(k)} + g^{(k)} - \varepsilon \left(\Gamma_{c,\varepsilon}^{f,g} \right)^{(k)}$ we finally have

$$\left\| \left(h_{c,\varepsilon}^{f,g} \right)^{(k)} \right\|_\infty \leq 2m + \exp\left(\frac{2m + C_\infty(R)}{\varepsilon}\right) \sum_{\omega \in \Omega(k)} \varepsilon^{-|\omega|+1} (2m + C_\infty(R))^{|\omega|}.$$

By defining,

$$H_0(m; R; (c, q); \varepsilon) := (2m + \varepsilon) + \varepsilon \exp\left(\frac{2m + C_\infty(R)}{\varepsilon}\right) \times \max_{0 \leq k \leq q} \left\{ \sum_{\omega \in \Omega(k)} \varepsilon^{-|\omega|} (2m + C_\infty(R))^{|\omega|} \right\},$$

we conclude that

$$\left\| \left(h_{c,\varepsilon}^{f,g} \right)^{(k)} \right\|_{q,\infty} \leq H_0.$$

We now include data processing transformations. Set $T_1, T_2 \in \mathcal{C}_R^p(\mathcal{X}, \mathbb{R}^d)$. It follows from the regularity of $h_{c,\varepsilon}^{f,g}$ that $h_{c,\varepsilon}^{f,g} \circ (T_1, T_2) \in \mathcal{C}^\kappa(\mathcal{X} \times \mathcal{X}, \mathbb{R})$. Since $T_1(x), T_2(y) \in B_R$, and because $h_{c,\varepsilon}^{f,g}$ is bounded by H_0 on $B_R \times B_R$, the function $h_{c,\varepsilon}^{f,g} \circ (T_1, T_2)$ is bounded on

$\mathcal{X} \times \mathcal{X}$ regardless of the choice of f, g, T_1 and T_2 . Here again, we use the chain rule to build higher-order bounds. From Proposition 4.A.2 applied with $F = h_{c,\varepsilon}^{f,g}$ and $G = (T_1, T_2)$ it follows that for any $1 \leq k \leq \kappa$,

$$\left\| (h_{c,\varepsilon}^{f,g} \circ (T_1, T_2))^{(k)}(x, y) \right\|_{op} \leq \sum_{\omega \in \Omega(k)} \left\| (h_{c,\varepsilon}^{f,g})^{(|\omega|)} \circ (T_1, T_2)(x, y) \right\|_{op} \times \prod_{1 \leq i \leq |\omega|} \left\| (T_1, T_2)^{(\omega_i)}(x, y) \right\|_{op}. \quad (4.21)$$

Then, remark that for any $1 \leq k \leq \kappa$,

$$\begin{aligned} \left\| (T_1, T_2)^{(k)}(x, y) \right\|_{op}^2 &= \sup_{\|\delta_i\| \leq 1} \left\| (T_1, T_2)^{(k)}(x, y)(\delta_1, \dots, \delta_k) \right\|^2 \\ &\leq \sup_{\|\delta_i\| \leq 1} \left\| T_1^{(k)}(x)(\delta_1, \dots, \delta_k) \right\|^2 + \sup_{\|\delta_i\| \leq 1} \left\| T_2^{(k)}(y)(\delta_1, \dots, \delta_k) \right\|^2 \\ &= \left\| T_1^{(k)}(x) \right\|_{op}^2 + \left\| T_2^{(k)}(y) \right\|_{op}^2 \\ &\leq 2R^2. \end{aligned}$$

We can therefore bound the right term of (4.21), leading to

$$\left\| (h_{c,\varepsilon}^{f,g} \circ (T_1, T_2))^{(k)} \right\|_{\infty} \leq \sum_{\omega \in \Omega(k)} H_0 \times \prod_{1 \leq i \leq |\omega|} \sqrt{2}R = H_0 \sum_{\omega \in \Omega(k)} (\sqrt{2}R)^{|\omega|}.$$

We conclude by defining

$$H(m; R; (c, q); \varepsilon, p) := H_0(m; R; (c, q); \varepsilon) \times \max_{0 \leq k \leq \kappa} \left\{ \sum_{\omega \in \Omega(k)} (\sqrt{2}R)^{|\omega|} \right\},$$

which leads to,

$$\left\| h_{c,\varepsilon}^{f,g} \circ (T_1, T_2) \right\|_{\kappa, \infty} \leq H. \quad \blacksquare$$

Proof of Proposition 5.1 Let $\{v^n\}_{n \in \mathbb{N}}$ be a sequence of vector fields in L_V^2 weakly converging to some $v \in L_V^2$. (Glaunes, 2005, Proposition 4) implies that for every $x \in \mathcal{X}$,

$$|\phi_1^{v^n}(x) - \phi_1^v(x)| \xrightarrow{n \rightarrow +\infty} 0. \quad (4.22)$$

Next, we aim at showing that this entails $\phi_1^{v^n} \alpha \xrightarrow[n \rightarrow +\infty]{w} \phi_1^v \alpha$, where w denotes the weak* convergence of *probability measures*. Firstly, note that as a consequence of the uniform-boundedness principle (Rudin, 1991, Theorem 2.5), the weak convergence of $\{v^n\}_{n \in \mathbb{N}}$ to v implies that there exists $M > 0$ such that $\{v^n\}_{n \in \mathbb{N}} \cup \{v\} \subset L_{V,M}^2$. Hence, according Lemma 4.3.1, there exists some $R = R((\mathcal{X}, d); (V, p); M) > 0$ such that the measures $\{\phi_1^{v^n} \alpha\}_{n \in \mathbb{N}}$, $\phi_1^v \alpha$, and β are all probability distributions on B_R . Secondly, recall that showing the weak* convergence amounts to check that for any bounded test functions $h \in \mathcal{C}(B_R, \mathbb{R})$ we have that $\int hd(\phi_1^{v^n} \alpha) \xrightarrow[n \rightarrow +\infty]{} \int hd(\phi_1^v \alpha)$. Let $h \in \mathcal{C}(B_R, \mathbb{R})$ be a bounded function and

use the push-forward change-of-variable formula to write $\int h d(\phi_1^{v^n} \alpha) = \int (h \circ \phi_1^{v^n}) d\alpha$. By continuity of h and according to (4.22), the sequence of functions $\{h \circ \phi_1^{v^n}\}_{n \in \mathbb{N}}$ converges pointwise to $h \circ \phi_1^v$. In addition, as h is bounded, this sequence is dominated by a constant. We can therefore apply the dominated convergence theorem to obtain that $\phi_1^{v^n} \alpha \xrightarrow[n \rightarrow +\infty]{w} \phi_1^v \alpha$.

We conclude the proof using (Feydy et al., 2019, Proposition 13), which states that $\mathcal{T}_{c,\varepsilon}$ (and consequently $S_{c,\varepsilon}$) is weak* continuous w.r.t. each of its input measures, provided that the ground cost function c is Lipschitz on their compact domains. This condition readily follows from the continuity of the derivative of c on the compact set $B_R \times B_R$. Therefore, $v \mapsto S_{c,\varepsilon}(\phi_1^v \alpha, \beta)$ is weakly continuous on L_V^2 . If additionally $e^{-\frac{c}{\varepsilon}}$ defines a positive universal kernel, then $v \mapsto S_{c,\varepsilon}(\phi_1^v \alpha, \beta)$ is non negative according to (Feydy et al., 2019, Theorem 1), which implies through (Glaunes, 2005, Theorem 7) that J_λ for $\Lambda = S_{c,\varepsilon}$ admits minimizers. ■

Proof of Proposition 5.3

Let $R > 0$. In a first time, we demonstrate the following Glivenko-Cantelli theorem (i):

$$\sup_{T_1, T_2 \in \mathcal{C}_R^p(\mathcal{X}, \mathbb{R}^d)} |\mathcal{T}_{c,\varepsilon}(T_1 \# \alpha_n, T_2 \# \beta_n) - \mathcal{T}_{c,\varepsilon}(T_1 \# \alpha, T_2 \# \beta)| \xrightarrow[n \rightarrow +\infty]{a.s.} 0.$$

In a second time, when $\kappa = \min\{p, q\} \geq d$, we show the following rate of convergence (ii):

$$\mathbb{E} \sup_{T_1, T_2 \in \mathcal{C}_R^p(\mathcal{X}, \mathbb{R}^d)} |\mathcal{T}_{c,\varepsilon}(T_1 \# \alpha_n, T_2 \# \beta_n) - \mathcal{T}_{c,\varepsilon}(T_1 \# \alpha, T_2 \# \beta)| \leq \frac{A}{\sqrt{n}},$$

where $A > 0$ is a constant. In both cases, the key idea of the proof is to note that the quantity

$$\sup_{T_1, T_2 \in \mathcal{C}_R^p(\mathcal{X}, \mathbb{R}^d)} |\mathcal{T}_{c,\varepsilon}(T_1 \# \alpha_n, T_2 \# \beta_n) - \mathcal{T}_{c,\varepsilon}(T_1 \# \alpha, T_2 \# \beta)|$$

is the supremum of a centered empirical process indexed by a class of smooth functions, and as such can be controlled via classical results from empirical process theory (see Section 4.A.1).

Let T_1 and T_2 be two arbitrary functions in $\mathcal{C}_R^p(\mathcal{X}, \mathbb{R}^d)$. By definition, the image sets $T_1(\mathcal{X})$ and $T_2(\mathcal{X})$ are contained in B_R . Thus, using the dual formulation, the entropic transportation costs can be written as,

$$\begin{aligned} \mathcal{T}_{c,\varepsilon}(T_1 \# \alpha_n, T_2 \# \beta_n) &= \sup_{f,g \in \mathcal{C}(B_R, \mathbb{R})} (T_1 \# \alpha_n \otimes T_2 \# \beta_n)(h_{c,\varepsilon}^{f,g}), \\ \mathcal{T}_{c,\varepsilon}(T_1 \# \alpha, T_2 \# \beta) &= \sup_{f,g \in \mathcal{C}(B_R, \mathbb{R})} (T_1 \# \alpha \otimes T_2 \# \beta)(h_{c,\varepsilon}^{f,g}). \end{aligned}$$

We apply Lemma 4.4.1 with $\mu = T_1 \# \alpha$ and $\nu = T_2 \# \beta$ which are probability measures on B_R . This implies that there exists a constant $m = m(B_R; (c, q), \varepsilon) > 0$ such that

$$\begin{aligned} \mathcal{T}_{c,\varepsilon}(T_1 \# \alpha_n, T_2 \# \beta_n) &= \sup_{f,g \in \mathcal{C}_m^q(B_R, \mathbb{R})} (T_1 \# \alpha_n \otimes T_2 \# \beta_n)(h_{c,\varepsilon}^{f,g}) \\ &= \sup_{f,g \in \mathcal{C}_m^q(B_R, \mathbb{R})} (\alpha_n \otimes \beta_n)(h_{c,\varepsilon}^{f,g} \circ (T_1, T_2)), \end{aligned}$$

where we used the push-forward change-of-variable formula. Proceeding similarly with the empirical measures we get,

$$\begin{aligned}\mathcal{T}_{c,\varepsilon}(T_{1\#}\alpha, T_{2\#}\beta) &= \sup_{f,g \in \mathcal{C}_m^q(B_R, \mathbb{R})} (T_{1\#}\alpha \otimes T_{2\#}\beta)(h_{c,\varepsilon}^{f,g}) \\ &= \sup_{f,g \in \mathcal{C}_m^q(B_R, \mathbb{R})} (\alpha \otimes \beta)(h_{c,\varepsilon}^{f,g} \circ (T_1, T_2)),\end{aligned}$$

Then, by using a classical error decomposition, we can control the difference between these two terms as follows,

$$\begin{aligned}|\mathcal{T}_{c,\varepsilon}(T_{1\#}\alpha_n, T_{2\#}\beta_n) - \mathcal{T}_{c,\varepsilon}(T_{1\#}\alpha, T_{2\#}\beta)| &\leq \\ &\sup_{f,g \in \mathcal{C}_m^q(B_R, \mathbb{R})} |(\alpha_n \otimes \beta_n)(h_{c,\varepsilon}^{f,g} \circ (T_1, T_2)) - (\alpha \otimes \beta)(h_{c,\varepsilon}^{f,g} \circ (T_1, T_2))|.\end{aligned}$$

After taking the supremum in $\mathcal{C}_R^p(\mathcal{X}, \mathbb{R}^d)$ on both sides of this inequality we get,

$$\begin{aligned}\sup_{T_1, T_2 \in \mathcal{C}_R^p(\mathcal{X}, \mathbb{R}^d)} |\mathcal{T}_{c,\varepsilon}(T_{1\#}\alpha_n, T_{2\#}\beta_n) - \mathcal{T}_{c,\varepsilon}(T_{1\#}\alpha, T_{2\#}\beta)| &\leq \\ \sup_{T_1, T_2 \in \mathcal{C}_R^p(\mathcal{X}, \mathbb{R}^d); f, g \in \mathcal{C}_m^q(B_R, \mathbb{R})} |(\alpha_n \otimes \beta_n)(h_{c,\varepsilon}^{f,g} \circ (T_1, T_2)) - (\alpha \otimes \beta)(h_{c,\varepsilon}^{f,g} \circ (T_1, T_2))|.\end{aligned}$$

The right term of this inequality can be seen as a centered empirical process indexed by the class of functions $\{h_{c,\varepsilon}^{f,g} \circ (T_1, T_2) \mid T_1, T_2 \in \mathcal{C}_R^p(\mathcal{X}, \mathbb{R}^d); f, g \in \mathcal{C}_m^q(B_R, \mathbb{R})\}$. Empirical process theory provides convergence guarantees when the index class is regular enough. Besides, we know from Proposition 4.4.1 that there exists a constant $H := H(R; (c, q); \varepsilon, p) > 0$ such that this class is included in $\mathcal{C}_H^\kappa(\mathcal{X} \times \mathcal{X}, \mathbb{R})$. Therefore,

$$\sup_{T_1, T_2 \in \mathcal{C}_R^p(\mathcal{X}, \mathbb{R}^d)} |\mathcal{T}_{c,\varepsilon}(T_{1\#}\alpha_n, T_{2\#}\beta_n) - \mathcal{T}_{c,\varepsilon}(T_{1\#}\alpha, T_{2\#}\beta)| \leq \sup_{h \in \mathcal{C}_H^\kappa(\mathcal{X} \times \mathcal{X}, \mathbb{R})} |(\alpha_n \otimes \beta_n)(h) - (\alpha \otimes \beta)(h)|.$$

Let us set $\mathcal{H} := \mathcal{C}_H^\kappa(\mathcal{X} \times \mathcal{X}, \mathbb{R})$. According to (Van Der Vaart and Wellner, 1996, Corollary 2.7.2) and (Van Der Vaart and Wellner, 1996, Theorem 2.4.1), \mathcal{H} is a so-called $(\alpha \otimes \beta)$ -Glivenko-Cantelli class of functions, meaning that

$$\sup_{h \in \mathcal{H}} |(\alpha_n \otimes \beta_n)(h) - (\alpha \otimes \beta)(h)| \xrightarrow[n \rightarrow +\infty]{a.s.} 0.$$

This implies (i). In addition, by Proposition 4.A.1, if $\kappa \geq (2d)/2$ then there exists a positive constant $A := A(R; (c, q); \varepsilon; (\mathcal{X}, d); p)$ such that,

$$\mathbb{E} \left[\sup_{h \in \mathcal{H}} |(\alpha_n \otimes \beta_n)(h) - (\alpha \otimes \beta)(h)| \right] \leq \frac{A}{\sqrt{n}}.$$

This proves (ii). ■

Before proving Theorem 4.5.1, we need the next intermediary result:

Lemma 4.B.1: Universal boundedness of the minimizers

Under the assumptions of Theorem 4.5.1, there exists a positive constant $M = M(\lambda; (\mathcal{X}, d); (c, q); \varepsilon)$ such that

$$\left\{ \bigcup_{n \in \mathbb{N}} \arg \min_{v \in L_V^2} J_{\lambda, n}(v) \right\} \cup \arg \min_{v \in L_V^2} J_{\lambda}(v) \subseteq L_{V, M}^2.$$

Proof The proof generalizes an argument made for a squared MMD in (Glaunes, 2005, Theorem 16) to a Sinkhorn divergence. Let $n \in \mathbb{N}$ and set v^n a minimizer of $J_{\lambda, n}$. Notice that the vector flow uniformly equal to zero generates the identity function, that is $\phi_t^0 = I$ for any $t \in [0, 1]$. Thus, by definition of a minimizer and by non negativity of the Sinkhorn divergence, we readily have that

$$\lambda \|v^n\|_{L_V^2}^2 \leq J_{\lambda, n}(v^n) \leq J_{\lambda, n}(0) = S_{c, \varepsilon}(\alpha_n, \beta_n).$$

Therefore, $\|v^n\|_{L_V^2}^2 \leq \lambda^{-1} S_{c, \varepsilon}(\alpha_n, \beta_n)$. To conclude, let us bound uniformly the right-term of this inequality. According to Lemma 4.4.1 applied with α_n and β_n there exists a constant $m = m((\mathcal{X}, d); (c, q); \varepsilon)$ such that,

$$\mathcal{T}_{c, \varepsilon}(\alpha_n, \beta_n) = \sup_{f, g \in \mathcal{C}_m^q(\mathcal{X}, \mathbb{R})} (\alpha_n \otimes \beta_n) \left(h_{c, \varepsilon}^{f, g} \right).$$

Moreover, for any $x, y \in \mathcal{X}$,

$$h_{c, \varepsilon}^{f, g}(x, y) = f(x) + g(y) - \varepsilon e^{\frac{f(x) + g(y) - c(x, y)}{\varepsilon}} + \varepsilon \leq m + m + 0 + \varepsilon.$$

Thus,

$$\mathcal{T}_{c, \varepsilon}(\alpha_n, \beta_n) \leq 2m + \varepsilon.$$

The same bound holds for the two auto-correlation terms of the Sinkhorn divergence, namely $\mathcal{T}_{c, \varepsilon}(\alpha_n, \alpha_n)$ and $\mathcal{T}_{c, \varepsilon}(\beta_n, \beta_n)$. Therefore, the triangle inequality leads to

$$S_{c, \varepsilon}(\alpha_n, \beta_n) \leq 4m + 2\varepsilon.$$

Consequently,

$$\|v^n\|_{L_V^2}^2 \leq \frac{4m + 2\varepsilon}{\lambda}.$$

To conclude, we set $M(\lambda; (\mathcal{X}, d); (c, q); \varepsilon) := \sqrt{\frac{4m + 2\varepsilon}{\lambda}}$. Note that this bound does not depend on n . As such, the minima $\{v^n\}_{n \in \mathbb{N}}$ all belong to $L_{V, M}^2$. A similar reasoning for v^* a minimizer of J_{λ} shows that all the minimizers of J_{λ} also belong to $L_{V, M}^2$. ■

Proof of Theorem 5.2 Let $M > 0$ be arbitrary (for now). Set $v \in L_{V, M}^2$ and compute

$$|J_{\lambda, n}(v) - J_{\lambda}(v)| = |S_{c, \varepsilon}(\phi_{1\#}^v \alpha_n, \beta_n) - S_{c, \varepsilon}(\phi_{1\#}^v \alpha, \beta)|.$$

According to Lemma 4.3.1, there exists a constant $R = R((\mathcal{X}, d); (V, p); M)$ such that for any $\phi \in \{\phi_t^v \mid t \in [0, 1], v \in L_{V,M}^2\}$, the restriction $\phi|_{\mathcal{X}}$ and the identity function I both belong to $\mathcal{C}_R^p(\mathcal{X}, \mathbb{R}^d)$. This leads to

$$\sup_{v \in L_{V,M}^2} |J_{\lambda,n}(v) - J_{\lambda}(v)| \leq \sup_{T_1, T_2 \in \mathcal{C}_R^p(\mathcal{X}, \mathbb{R}^d)} |S_{c,\varepsilon}(T_{1\#}\alpha_n, T_{2\#}\beta_n) - S_{c,\varepsilon}(T_{1\#}\alpha, T_{2\#}\beta)|. \quad (4.23)$$

From here, let us demonstrate the convergence of the minima, that is item (i). According to Lemma 4.B.1, there exists $M = M(\lambda; (\mathcal{X}, d); (c, q); \varepsilon) > 0$ such that all the minimizers of $J_{\lambda,n}$ belong to $L_{V,M}^2$. Next, we show that any weakly-converging subsequences of $\{v^n\}_{n \in \mathbb{N}}$ tend to a minimizer of J_{λ} . Set v^* a minimizer of J_{λ} , and let $\{u^n\}_{n \in \mathbb{N}}$ be a subsequence with limit u . First, let's show that $\lim_{n \rightarrow +\infty} J_{\lambda,n}(u^n) = J_{\lambda}(u)$. By the triangle inequality, $|J_{\lambda,n}(u^n) - J_{\lambda}(u)| \leq |J_{\lambda,n}(u^n) - J_{\lambda}(u^n)| + |J_{\lambda}(u^n) - J_{\lambda}(u)|$. The first term tends to zero by Proposition 4.5.2 and (4.23) specified with $M(\lambda; (\mathcal{X}, d); (c, q); \varepsilon)$, while the second term tends to zero according to Proposition 4.5.1 which ensures the weak continuity of J_{λ} . Second, note that the optimality condition entails that $J_{\lambda,n}(u^n) \leq J_{\lambda,n}(v^*)$, and that $\lim_{n \rightarrow +\infty} J_{\lambda,n}(v^*) = J_{\lambda}(v^*)$. Then, at the limit $J_{\lambda}(u) \leq J_{\lambda}(v^*)$, meaning that u is a minimizer of J_{λ} . Therefore, any weakly-converging subsequence $\{u^n\}_{n \in \mathbb{N}}$ of $\{v^n\}_{n \in \mathbb{N}}$ tends to a minimizer u of J_{λ} .

To conclude on the convergence of the generated diffeomorphisms, we rely on (Glaunes, 2005, Remark 1), stating that

$$\sup_{t \in [0,1]} \{ \|\phi_t^{u^n} - \phi_t^u\|_{\infty} + \|(\phi_t^{u^n})^{-1} - (\phi_t^u)^{-1}\|_{\infty} \} \leq 2c_V \|u^n - u\|_{L_V^2} \exp(c_V \|u\|_{L_V^2}).$$

We showed that $\|u^n - u\|_{L_V^2} \xrightarrow{n \rightarrow \infty} 0$. Consequently, the upper bound tends to zero as n increases to infinity. This completes the proof of (i).

Item (ii) readily follows from Proposition 4.5.2 stating that if $\kappa \geq d$, then there exists for any $M > 0$ a constant $A = A(\lambda; (\mathcal{X}, d); (c, q); \varepsilon; (V, p); M) > 0$ such that

$$\mathbb{E} \left[\sup_{v \in L_{V,M}^2} |J_{\lambda,n}(v) - J_{\lambda}(v)| \right] \leq \frac{A}{\sqrt{n}}.$$

To conclude, recall that both $\{v^n\}_{n \in \mathbb{N}}$ and v^* belong to $L_{V,M}^2$ for the constant M from Lemma 4.B.1, and apply the classical deviation inequality

$$J_{\lambda}(v^n) - J_{\lambda}(v^*) \leq 2 \sup_{v \in L_{V,M}^2} |J_{\lambda,n}(v) - J_{\lambda}(v)|.$$

■

Conclusion

The work presented in this thesis spans **counterfactual reasoning**, by clarifying misconceptions in Chapter 1 and introducing a sound mass-transportation approach in Chapter 2; to **machine-learning fairness**, by implementing new individual notions in Chapter 2; to **optimal transport**, by designing a novel neural mapping estimator with pioneering statistical guarantees in Chapter 3 and further studying the consistency of entropic-optimal-transport losses in Chapter 4; to **diffeomorphic registration**, by furnishing a theoretical and computational framework for Sinkhorn-divergence-based diffeomorphic mass transportation in Chapter 4. Although they can be considered independent, these contributions transcribe a common objective, as the development of efficient mass-transportation estimators serves to the implementation of transport-based counterfactual models, therefore provides more tools to build trustworthy AI systems. They also raise several open questions and further research directions, as summarized below:

- **Counterfactual reasoning in fairness** deserves a more thorough investigation. Ethical obligations and future legal constraints unquestionably demands agreeable notions of individual fairness. Although counterfactual fairness as introduced by [Kusner et al. \(2017\)](#) is an intuitive proposition in theory, its dependence to a structural causal model makes it hardly viable for large-scale deployment. Transport-based counterfactual fairness from Chapter 2 builds upon the same counterfactual-invariance principle while deviating from causality to retrieve feasibility. As such, it does not replace causal inference by any mean, but still preserves the discrimination control at the individual level. Nevertheless, be it causal or transport-based, counterfactual fairness faces a critical challenge: it manipulates counterfactuals that generally do not have a natural interpretation since they are computed under the implicit assumption that they are eligible to the dataset, as explained by [Fawkes et al. \(2022\)](#). This calls for a philosophical and legal debate on the scope of such counterfactuals to uncover discriminations. We believe that the merit of within-dataset counterfactuals is that they furnish individual fairness conditions, stronger than common group-fairness ones—which is not the case of outside-dataset counterfactuals. So, as long as algorithmic decision rules must conform to statistical parity, these counterfactuals will be useful. Specifically, transport-based counterfactual fairness enables one to control how statistical parity is achieved at the individual scale, considerably narrowing the arbitrariness of mere group-fair decision rules. Moreover, the mathematical analysis of the counterfactual fairness constraints in learning problems has been neglected by the community and should therefore be further studied, as initiated in Section 2.8. Specifically, we need to understand what counterfactual fairness incurs on the price in accuracy and on the

number of admissible models in order to know when it is recommended or not to use it. Lastly, one must crucially keep in mind that transport models and causal models are always estimated from data in practice, bringing statistical uncertainty to the assessment of counterfactual queries. This crucially raises the question of bridging the gap between counterfactual reasoning and statistics, by exploring how the statistical consistency of the estimated counterfactual model reverberates on the confidence of counterfactual notions. Note that our work in (De Lara et al., 2021b) follows this track by specifying conditions of statistical convergence for counterfactual fairness, but lacks confidence intervals.

- **Estimation of optimal transport maps** is a field that advances at a very fast pace, with a strong focus on computational performances. In particular, the neural approximation of (Uscidda and Cuturi, 2023) seems to attain an unprecedented level of flexibility. However, little is known about its statistical convergence, as for the other most expressive approximations of optimal transport. In contrast, our estimator from Chapter 3 has practical limitations, but comes with unique theoretical guarantees. This underlines a trade-off between statistical consistency and utility for neural-network-based approximations of optimal transport. Because trustworthiness in machine learning—as in any scientific field—needs statistical significance, we believe that proving such guarantees for any constructions represents a critical line of research. Moreover, as a direct improvement of our work, we could also precise the convergence rate of our estimator; we point out that our proof does not naturally extend to such a result.
- **Diffeomorphic registration**, tailored to shape analysis, has demonstrated its potential for machine learning applications in (Younes, 2020). We mentioned in Chapter 2 that diffeomorphic matching could possibly be applied to counterfactual reasoning, which is what originally motivated Chapter 4. In this work, we introduced a functional and statistically sound diffeomorphic-mass-transportation engine. Although this contributes to diffeomorphic registration in general, we came to the conclusion that the application of such a tool to counterfactual reasoning deserves further thinking. Diffeomorphisms could definitely bring promising properties to counterfactual association: they generate continuous paths between paired instances, they preserve topology, they work on (almost) any kind of distributions. Nevertheless, they do not come with an explicit counterfactual interpretation on the contrary to causal models and optimal transport. More precisely, while the kernel similarity function could inspire a notion of closeness between worlds, it already bears the responsibility of ensuring the distribution fitting and consequently cannot be forged for a different purpose. As such, applying diffeomorphic registration to fairness remains an open question worth exploring.

Lastly, I would like to conclude this manuscript on a more holistic remark. On the 22nd of June 2020, more than twenty scholars from the fields of machine learning and social sciences sent an open letter to the Springer Editorial Committee,⁵ sharing their consternation on the forthcoming publication of a scientific paper claiming to predict criminality from images of human faces. Gathered under the name of the Coalition for Critical Technology, they urged

⁵<https://medium.com/@CoalitionForCriticalTechnology/abolish-the-techtoprisonpipeline-9b5b14366b16>

Springer to publicly retract the publication and to explain how this study was reviewed. More than targeting a specific paper, their call denounces the general academic trend of relying on justice statistics to predict criminality as well as the resurgence of phrenological research through data science. On the basis of numerous past studies, they recalled that such research is unfair by design: since predictions rely on racially biased data from the justice system, they reproduce and amplify biases across class and race. To this day, the public letter accounts more than 2,400 signatures from researchers, practitioners and students, and the announced paper has not been released. However, the fact that such research has even been conducted and seriously considered for publication by a notorious scientific journal epitomizes the lack of critical insight blighting the spread of artificial intelligence. Naturally, this justifies more than ever the need to polarize machine-learning research towards improving fairness and explainability rather than accuracy. But more importantly, this case stresses out the priority of better educating researchers, practitioners and students on the risks and limitations of machine learning. There cannot not be trustworthy artificial intelligence systems without a trustworthy community developing them. In my opinion, this takes precedence over any groundbreaking fairness definition or explainability framework.

Conclusion (en français)

Les travaux présentés dans cette thèse s'étendent du **raisonnement contrefactuel**, en clarifiant des idées reçues dans le Chapitre 1 et en introduisant une approche de transport de masse étayée dans le Chapitre 2 ; à **l'équité de l'apprentissage machine**, en mettant en œuvre de nouvelles notions individuelles dans le Chapitre 2 ; au **transport optimal**, en concevant un nouvel estimateur "neuronal" avec des garanties statistiques pionnières dans le Chapitre 3 et en étudiant davantage la convergence statistique des coûts de transport optimaux entropiques dans le Chapitre 4 ; à **l'appariement difféomorphique**, en fournissant un cadre théorique et numérique pour le transport de masse difféomorphique basé sur les divergences de Sinkhorn dans le Chapitre 4. Bien qu'elles puissent être considérées comme indépendantes, ces contributions transcrivent un objectif commun, puisque le développement d'estimateurs de transport de masse efficaces sert à la mise en œuvre de modèles contrefactuels basés sur le transport, et fournit donc plus d'outils pour construire des systèmes d'IA dignes de confiance. Elles soulèvent également plusieurs questions ouvertes et de nouvelles directions de recherche, résumées ci-dessous :

- **Le raisonnement contrefactuel pour l'équité** mérite une étude plus approfondie. Les obligations éthiques et les contraintes juridiques futures exigent incontestablement des notions consensuelles d'équité individuelle. Bien que l'équité contrefactuelle telle qu'introduite par [Kusner et al. \(2017\)](#) soit une proposition intuitive en théorie, sa dépendance à un modèle causal structurel la rend difficilement viable pour un déploiement à grande échelle. L'équité contrefactuelle basée sur le transport du Chapitre 2 repose sur le même principe d'invariance contrefactuelle tout en s'écartant de la causalité pour retrouver la faisabilité. En tant que telle, elle ne remplace en aucun cas l'inférence causale, mais préserve toujours le contrôle de la discrimination au niveau individuel. Néanmoins, qu'elle soit causale ou basée sur le transport, l'équité contrefactuelle est confrontée à un défi majeur : elle manipule des contrefactuels qui n'ont généralement pas d'interprétation naturelle puisqu'ils sont calculés en partant de l'hypothèse implicite qu'ils sont éligibles au jeu de données, comme l'explique [Fawkes et al. \(2022\)](#). Cela appelle un débat philosophique et juridique sur la portée de ces contrefactuels dans la découverte de discriminations. Nous pensons que le mérite des contrefactuels "dans les données" est qu'ils fournissent des conditions d'équité individuelles, plus fortes que les conditions courantes d'équité de groupe, ce qui n'est pas le cas des contrefactuels "hors données". Ainsi, tant que les règles de décision algorithmiques doivent se conformer à la parité statistique, ces contrefactuels seront utiles. Plus précisément, l'équité contrefactuelle basée sur le transport permet de contrôler la manière dont la parité statistique est atteinte à l'échelle individuelle, ce qui réduit considérablement l'arbitraire des

règles de décision équitables à l'échelle globale. De plus, l'analyse mathématique des contraintes d'équité contrefactuelle dans les problèmes d'apprentissage a été négligée par la communauté et devrait donc être plus étudiée, comme initié dans la Section 2.8. Plus précisément, nous devons comprendre ce que l'équité contrefactuelle implique pour le coût en précision et le nombre de modèles admissibles afin de savoir quand il est recommandé ou non de l'utiliser. Enfin, il est essentiel de garder à l'esprit que les modèles de transport et les modèles causaux sont toujours estimés à partir de données dans la pratique, ce qui introduit une incertitude statistique dans l'évaluation des requêtes contrefactuelles. Cela soulève la question cruciale de combler le fossé entre le raisonnement contrefactuel et les statistiques, en explorant comment la cohérence statistique du modèle contrefactuel estimé se répercute sur la confiance des notions contrefactuelles. Notez que notre travail dans (De Lara et al., 2021b) suit cette voie en spécifiant des conditions de convergence statistique pour l'équité contrefactuelle, mais manque d'intervalles de confiance.

- **L'estimation des applications de transport optimales** est un domaine qui progresse à un rythme très rapide, avec un fort accent sur les performances numériques. En particulier, l'approximation neuronale de (Uscidda and Cuturi (2023)) semble atteindre un niveau de flexibilité sans précédent. Cependant, on sait peu de choses sur sa convergence statistique, comme pour les autres approximations les plus expressives du transport optimal. En revanche, notre estimateur du Chapitre 3 présente des limites pratiques, mais s'accompagne de garanties théoriques uniques. Cela met en évidence un compromis entre la cohérence statistique et l'utilité des approximations du transport optimal basées sur des réseaux de neurones. Parce que la confiance dans l'apprentissage automatique - comme dans tout domaine scientifique - nécessite une signification statistique, nous pensons que prouver de telles garanties pour n'importe quelle construction représente une ligne de recherche critique. De plus, comme amélioration directe de notre travail, nous pourrions également préciser la vitesse de convergence de notre estimateur ; nous soulignons que notre preuve ne s'étend pas naturellement à un tel résultat.
- **L'appariement difféomorphique**, adapté à l'analyse de forme, a démontré son potentiel pour les applications d'apprentissage automatique dans (Younes, 2020). Nous avons mentionné dans le Chapitre 2 que l'appariement difféomorphique pourrait éventuellement être appliqué au raisonnement contrefactuel, ce qui a originellement motivé le Chapitre 4. Dans ce travail, nous avons introduit un mécanisme de transport de masse difféomorphique fonctionnel et statistiquement solide. Bien que cela contribue à l'appariement difféomorphique en général, nous sommes arrivés à la conclusion que l'application d'un tel outil au raisonnement contrefactuel mérite une réflexion plus approfondie. Les difféomorphismes pourraient certainement apporter des propriétés prometteuses à l'association contrefactuelle : ils génèrent des chemins continus entre les instances appariées, ils préservent la topologie, ils fonctionnent sur (presque) n'importe quel type de distribution. Néanmoins, ils ne s'accompagnent pas d'une interprétation contrefactuelle explicite, contrairement aux modèles causaux et au transport optimal. Plus précisément, alors que la fonction noyau de similarité pourrait inspirer une notion de proximité entre les mondes, elle porte déjà la responsabilité d'assurer l'appariement

des distributions et ne peut donc pas être utilisée à d'autres fins. Ainsi, l'application de l'appariement difféomorphe à l'équité reste une question ouverte qui mérite d'être explorée.

Enfin, j'aimerais conclure ce manuscrit par une remarque plus holistique. Le 22 juin 2020, plus de vingt chercheur·euse·s des domaines de l'apprentissage automatique et des sciences sociales ont envoyé une lettre ouverte au comité éditorial de Springer.⁶ pour lui faire part de leur consternation face à la publication prochaine d'un article scientifique prétendant prédire la criminalité à partir d'images de visages humains. Réuni·e·s sous le nom de Coalition for Critical Technology, iels ont exhorté Springer à le rétracter publiquement et à expliquer comment cette étude a été évaluée. Plus que de viser un article en particulier, leur appel dénonce la tendance académique générale à s'appuyer sur les statistiques judiciaires pour prédire la criminalité, ainsi que la résurgence de la recherche phrénologique à travers la science des données. Sur la base de nombreuses études antérieures, iels rappellent que ces recherches sont injustes de par leur conception : puisque les prédictions s'appuient sur des données racialement biaisées du système judiciaire, elles reproduisent et amplifient les préjugés de classe et de race. À ce jour, la lettre publique compte plus de 2400 signatures de chercheur·euse·s, de praticien·ne·s et d'étudiant·e·s, et l'article annoncé n'a pas été publié. Cependant, le fait qu'une telle recherche ait été menée et même sérieusement envisagée pour publication par une revue scientifique de renom illustre le manque de vision critique qui gangrène la diffusion de l'intelligence artificielle. Naturellement, cela justifie plus que jamais la nécessité de polariser la recherche sur l'apprentissage automatique vers l'amélioration de l'équité et de l'explicabilité plutôt que la précision. Mais surtout, cette affaire souligne la priorité d'assurer une meilleure formation des chercheur·euse·s, des praticien·ne·s et des étudiant·e·s sur les risques et les limites de l'apprentissage automatique. Il ne peut y avoir de systèmes d'intelligence artificielle dignes de confiance sans une communauté digne de confiance pour les développer. À mon avis, cela prime sur tout critère d'équité ou toute méthode d'explicabilité révolutionnaires.

⁶<https://medium.com/@CoalitionForCriticalTechnology/abolish-the-techtoprisonpipeline-9b5b14366b16>

Bibliography

- Alvarez, J. M. and Ruggieri, S. (2023). Counterfactual situation testing: Uncovering discrimination under fairness given the difference. *arXiv preprint arXiv:2302.11944*.
- Anil, C., Lucas, J., and Grosse, R. (2019). Sorting out Lipschitz function approximation. In *International Conference on Machine Learning*, pages 291–301. PMLR.
- Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223. PMLR.
- Arsigny, V., Commowick, O., Pennec, X., and Ayache, N. (2006). A log-Euclidean framework for statistics on diffeomorphisms. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 924–931. Springer.
- Asher, N., De Lara, L., Paul, S., and Russell, C. (2022). Counterfactual models for fair and adequate explanations. *Machine Learning and Knowledge Extraction*, 4(2):316–349.
- Asher, N., Paul, S., and Russell, C. (2021). Fair and adequate explanations. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 79–97. Springer.
- Bajaj, M., Chu, L., Xue, Z. Y., Pei, J., Wang, L., Lam, P. C.-H., and Zhang, Y. (2021). Robust counterfactual explanations on graph neural networks. *Advances in Neural Information Processing Systems*, 34.
- Balaji, Y., Chellappa, R., and Feizi, S. (2020). Robust optimal transport with applications in generative modeling and domain adaptation. *Advances in Neural Information Processing Systems*, 33:12934–12944.
- Barocas, S., Hardt, M., and Narayanan, A. (2019). *Fairness and Machine Learning*. fairml-book.org. <http://www.fairmlbook.org>.
- Bauer, M., Joshi, S., and Modin, K. (2015). Diffeomorphic density matching by optimal information transport. *SIAM Journal on Imaging Sciences*, 8(3):1718–1751.
- Beg, M. F., Miller, M. I., Trounev, A., and Younes, L. (2005). Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International Journal of Computer Vision*, 61(2):139–157.

- Beirlant, J., Buitendag, S., del Barrio, E., Hallin, M., and Kamper, F. (2020). Center-outward quantiles and the measurement of multivariate risk. *Insurance: Mathematics and Economics*, 95:79–100.
- Bentler, P. M. and Weeks, D. G. (1980). Linear structural equations with latent variables. *Psychometrika*, 45(3):289–308.
- Berk, R. A., Kuchibhotla, A. K., and Tchetgen, E. T. (2021). Improving fairness in criminal justice algorithmic risk assessments using optimal transport and conformal prediction sets. *arXiv preprint arXiv:2111.09211*.
- Besse, P., del Barrio, E., Gordaliza, P., Loubes, J.-M., and Risser, L. (2021). A survey of bias in machine learning through the prism of statistical parity. *The American Statistician*, pages 1–11.
- Béthune, L., Boissin, T., Serrurier, M., Mamalet, F., Friedrich, C., and Gonzalez Sanz, A. (2022). Pay attention to your loss: understanding misconceptions about lipschitz neural networks. *Advances in Neural Information Processing Systems*, 35:20077–20091.
- Biau, G., Sangnier, M., and Tanielian, U. (2021). Some theoretical insights into Wasserstein GANs. *The Journal of Machine Learning Research*, 22(1):5287–5331.
- Bickel, P. J., Hammel, E. A., and O’Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley: Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation. *Science*, 187(4175):398–404.
- Black, E., Yeom, S., and Fredrikson, M. (2020). Fliptest: Fairness testing via optimal transport. In *Conference on Fairness, Accountability, and Transparency*, pages 111–121. Association for Computing Machinery.
- Bongers, S., Forré, P., Peters, J., and Mooij, J. M. (2021). Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885–2915.
- Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press.
- Brenier, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417.
- Byrne, R. M. (2019). Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning. In *International Joint Conference on Artificial Intelligence*, pages 6276–6282.
- Charlier, B., Feydy, J., Glaunès, J. A., Collin, F.-D., and Durif, G. (2021). Kernel operations on the GPU, with autodiff, without memory overflows. *Journal of Machine Learning Research*, 22(74):1–6.
- Chernozhukov, V., Fernández-Val, I., and Melly, B. (2013). Inference on counterfactual distributions. *Econometrica*, 81(6):2205–2268.

- Chernozhukov, V., Galichon, A., Hallin, M., and Henry, M. (2017). Monge–Kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, 45(1):223–256.
- Chiappa, S. (2019). Path-specific counterfactual fairness. In *AAAI Conference on Artificial Intelligence*, volume 33, pages 7801–7808.
- Chiappa, S., Jiang, R., Stepleton, T., Pacchiano, A., Jiang, H., and Aslanides, J. (2020). A general approach to fairness with optimal transport. In *AAAI Conference on Artificial Intelligence*, pages 3633–3640.
- Chickering, D. M., Heckerman, D., and Meek, C. (2004). Large-sample learning of Bayesian networks is NP-hard. *Journal of Machine Learning Research*, 5:1287–1330.
- Chizat, L., Roussillon, P., Léger, F., Vialard, F.-X., and Peyré, G. (2020). Faster Wasserstein distance estimation with the Sinkhorn divergence. *Advances in Neural Information Processing Systems*, 33:2257–2269.
- Chzhen, E., Denis, C., Hebiri, M., Oneto, L., and Pontil, M. (2020). Fair regression with Wasserstein barycenters. In *Advances in Neural Information Processing Systems*, volume 33, pages 7321–7331. Curran Associates, Inc.
- Civil Rights Act (1964). Civil rights act of 1964. *Title VII, Equal Employment Opportunities*.
- Clark, D. E. and Houssineau, J. (2013). Faa di Bruno’s formula and spatial cluster modelling. *Spatial Statistics*, 6:109–117.
- Colnet, B., Mayer, I., Chen, G., Dieng, A., Li, R., Varoquaux, G., Vert, J.-P., Josse, J., and Yang, S. (2020). Causal inference methods for combining randomized trials and observational studies: a review. *arXiv preprint arXiv:2011.08047*.
- Cooper, G. F. (1990). The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, 42(2–3):393–405. Research Note.
- Courty, N., Flamary, R., Habrard, A., and Rakotomamonjy, A. (2017). Joint distribution optimal transportation for domain adaptation. *Advances in Neural Information Processing Systems*, 30.
- Courty, N., Flamary, R., and Tuia, D. (2014). Domain adaptation with regularized optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 274–289. Springer.
- Croquet, B., Christiaens, D., Weinberg, S. M., Bronstein, M., Vandermeulen, D., and Claes, P. (2021). Unsupervised diffeomorphic surface registration and non-linear modelling. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 118–128. Springer.
- Cuesta, J. A. and Matrán, C. (1989). Notes on the Wasserstein metric in Hilbert spaces. *The Annals of Probability*, pages 1264–1276.
- Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 26:2292–2300.

- De Lara, L., González-Sanz, A., Asher, N., Risser, L., and Loubes, J.-M. (2021a). Transport-based counterfactual models. *arXiv preprint arXiv:2108.13025*.
- De Lara, L., González-Sanz, A., and Loubes, J.-M. (2023). Diffeomorphic registration using Sinkhorn divergences. *SIAM Journal on Imaging Sciences*, 16(1):250–279.
- De Lara, L., González-Sanz, A., and Loubes, J.-M. (2021b). A consistent extension of discrete optimal transport maps for machine learning applications. *arXiv preprint arXiv:2102.08644*.
- del Barrio, E., Gonzalez-Sanz, A., Loubes, J.-M., and Niles-Weed, J. (2022). An improved central limit theorem and fast convergence rates for entropic transportation costs. *arXiv preprint arXiv:2204.09105*.
- Dominguez-Olmedo, R., Karimi, A. H., and Schölkopf, B. (2022). On the adversarial robustness of causal algorithmic recourse. In *International Conference on Machine Learning*, pages 5324–5342. PMLR.
- Dua, D. and Graff, C. (2019). UCI machine learning repository. <http://archive.ics.uci.edu/ml>.
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Innovations in Theoretical Computer Science Conference*, pages 214–226.
- Fawkes, J., Evans, R., and Sejdinovic, D. (2022). Selection, ignorability and challenges with causal fairness. In *Conference on Causal Learning and Reasoning*, pages 275–289. PMLR.
- Fel, T., Cadène, R., Chalvidal, M., Cord, M., Vigouroux, D., and Serre, T. (2021). Look at the variance! efficient black-box explanations with Sobol-based sensitivity analysis. *Advances in Neural Information Processing Systems*, 34.
- Ferradans, S., Papadakis, N., Peyré, G., and Aujol, J.-F. (2014). Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882.
- Feydy, J. (2020). *Geometric data analysis, beyond convolutions*. PhD thesis, Université Paris-Saclay Gif-sur-Yvette, France.
- Feydy, J., Charlier, B., Vialard, F.-X., and Peyré, G. (2017). Optimal transport for diffeomorphic registration. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 291–299. Springer.
- Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S.-i., Trouné, A., and Peyré, G. (2019). Interpolating between optimal transport and MMD using Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 2681–2690. PMLR.
- Feydy, J. and Trouné, A. (2018). Global divergences between measures: from Hausdorff distance to optimal transport. In *International Workshop on Shape in Medical Imaging*, pages 102–115. Springer.

- Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N. T., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D. J., Tavenard, R., Tong, A., and Vayer, T. (2021). Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8.
- Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., and Davidian, M. (2011). Doubly robust estimation of causal effects. *American Journal of Epidemiology*, 173(7):761–767.
- Galles, D. and Pearl, J. (1998). An axiomatic characterization of causal counterfactuals. *Foundations of Science*, 3(1):151–182.
- Gayraud, N. T., Rakotomamonjy, A., and Clerc, M. (2017). Optimal transport applied to transfer learning for p300 detection. In *Graz Brain-Computer Interface Conference*, volume 7, page 6.
- Genevay, A. (2019). *Entropy-Regularized Optimal Transport for Machine Learning. (Régularisation Entropique du Transport Optimal pour le Machine Learning)*. PhD thesis, PSL Research University, Paris, France.
- Genevay, A., Chizat, L., Bach, F., Cuturi, M., and Peyré, G. (2019). Sample complexity of Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1574–1583. PMLR.
- Genevay, A., Peyré, G., and Cuturi, M. (2018). Learning generative models with Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617. PMLR.
- Ghosal, P. and Sen, B. (2022). Multivariate ranks and quantiles using optimal transport: Consistency, rates and nonparametric testing. *The Annals of Statistics*, 50(2):1012–1037.
- Glaunes, J. (2005). *Transport par difféomorphismes de points, de mesures et de courants pour la comparaison de formes et l’anatomie numérique*. PhD thesis.
- Glaunes, J., Trouvé, A., and Younes, L. (2004). Diffeomorphic matching of distributions: A new approach for unlabelled point-sets and sub-manifolds matching. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages II–II. IEEE.
- González-Sanz, A., De Lara, L., Béthune, L., and Loubes, J.-M. (2022). GAN estimation of Lipschitz optimal transport maps. *arXiv preprint arXiv:2202.07965*.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27.
- Gordaliza, P., Del Barrio, E., Gamboa, F., and Loubes, J.-M. (2019). Obtaining fairness using optimal transport theory. In *International Conference on Machine Learning*, pages 2357–2365. PMLR.

- Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., and Lee, S. (2019). Counterfactual visual explanations. In *International Conference on Machine Learning*, pages 2376–2384. PMLR.
- Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. (2017). Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*, pages 5769–5779.
- Hallin, M., del Barrio, E., Cuesta-Albertos, J., and Matrán, C. (2021). Distribution and quantile functions, ranks and signs in dimension d : A measure transportation approach. *The Annals of Statistics*, 49(2):1139 – 1165.
- Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323.
- He, J., Li, L., Xu, J., and Zheng, C. (2020). ReLU deep neural networks and linear finite elements. *Journal of Computational Mathematics*, 38(3):502–527.
- Heinonen, J. (2005). *Lectures on Lipschitz analysis*. Number 100. University of Jyväskylä.
- Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.
- Hoyer, P., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. (2008). Nonlinear causal discovery with additive noise models. *Advances in Neural Information Processing Systems*, 21.
- Huang, C.-W., Chen, R. T. Q., Tsirigotis, C., and Courville, A. (2021). Convex potential flows: Universal probability distributions with optimal transport and convex optimization. In *International Conference on Learning Representations*.
- Huang, X., Kwiatkowska, M., Wang, S., and Wu, M. (2017). Safety verification of deep neural networks. In *International Conference on Computer Aided Verification*, pages 3–29. Springer.
- Hütter, J.-C. and Rigollet, P. (2021). Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 49(2):1166–1194.
- Hytinen, A., Eberhardt, F., and Hoyer, P. O. (2012). Learning linear cyclic causal models with latent variables. *Journal of Machine Learning Research*, 13(1):3387–3439.
- Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 76(1):243–263.
- Imbens, G. W. (2020). Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature*, 58(4):1129–79.
- Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

- Joshi, S., Koyejo, O., Vijitbenjaronk, W., Kim, B., and Ghosh, J. (2019). Towards realistic individual recourse and actionable explanations in black-box decision making systems. *arXiv preprint arXiv:1907.09615*.
- Joshi, S. C. and Miller, M. I. (2000). Landmark matching via large deformation diffeomorphisms. *IEEE Transactions on Image Processing*, 9(8):1357–1370.
- Kang, J. D. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):523–539.
- Kantorovich, L. V. and Rubinshtein, S. (1958). On a space of totally additive functions. *Vestnik of the St. Petersburg University: Mathematics*, 13(7):52–59.
- Karimi, A.-H., Schölkopf, B., and Valera, I. (2021). Algorithmic recourse: from counterfactual explanations to interventions. In *ACM Conference on Fairness, Accountability, and Transparency*, pages 353–362.
- Keane, M. T., Kenny, E. M., Delaney, E., and Smyth, B. (2021). If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual XAI techniques. In Zhou, Z.-H., editor, *International Joint Conference on Artificial Intelligence*, volume 30, pages 4466–4474. International Joint Conferences on Artificial Intelligence Organization. Survey Track.
- Kilbertus, N., Ball, P. J., Kusner, M. J., Weller, A., and Silva, R. (2020). The sensitivity of counterfactual fairness to unmeasured confounding. In *Conference on Uncertainty in Artificial Intelligence*, pages 616–626. PMLR.
- Kleinberg, J., Mullainathan, S., and Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In *Innovations in Theoretical Computer Science Conference*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.
- Kolouri, S., Park, S. R., Thorpe, M., Slepcev, D., and Rohde, G. K. (2017). Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34(4):43–59.
- Koltchinskii, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d’Eté de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer Science & Business Media.
- Korotin, A., Egiazarian, V., Asadulaev, A., Safin, A., and Burnaev, E. (2021). Wasserstein-2 generative networks. In *International Conference on Learning Representations*.
- Krco, N., Laugel, T., Loubes, J.-M., and Detyniecki, M. (2023). When mitigating bias is unfair: A comprehensive study on the impact of bias mitigation algorithms. *arXiv preprint arXiv:2302.07185*.
- Kuhl, U., Artelt, A., and Hammer, B. (2022). Keep your friends close and your counterfactuals closer: Improved learning from closest rather than plausible counterfactual

- explanations in an abstract setting. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2125–2137.
- Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems*, volume 30, pages 4066–4076. Curran Associates, Inc.
- Lewis, D. (1973a). Causation. *Journal of Philosophy*, 70(17):556–567.
- Lewis, D. (1973b). *Counterfactuals*. Blackwell.
- Leygonie, J., She, J., Almahairi, A., Rajeswar, S., and Courville, A. (2019). Adversarial computation of optimal transport maps. *arXiv preprint arXiv:1906.09691*.
- Lipton, Z., McAuley, J., and Chouldechova, A. (2018). Does mitigating ml’s impact disparity require treatment disparity? In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.
- Liu, S., Kailkhura, B., Loveland, D., and Han, Y. (2019). Generative counterfactual introspection for explainable deep learning. In *IEEE Global Conference on Signal and Information Processing*, pages 1–5. IEEE.
- Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pages 4765–4774. Curran Associates, Inc.
- Mahajan, D., Tan, C., and Sharma, A. (2020). Preserving causal constraints in counterfactual explanations for machine learning classifiers. *arXiv preprint arXiv:1912.03277*.
- Makhlouf, K., Zhioua, S., and Palamidessi, C. (2020). Survey on causal-based machine learning fairness notions. *arXiv preprint arXiv:2010.09553*.
- Makkuva, A., Taghvaei, A., Oh, S., and Lee, J. (2020). Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning*, pages 6672–6681. PMLR.
- Manole, T., Balakrishnan, S., Niles-Weed, J., and Wasserman, L. (2021). Plugin estimation of smooth optimal transport maps. *arXiv preprint arXiv:2107.12364*.
- McCann, R. J. (1995). Existence and uniqueness of monotone measure-preserving maps. *Duke Math. J.*, 80(2):309–323.
- McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4):115–133.
- Mena, G. and Niles-Weed, J. (2019). Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. *Advances in Neural Information Processing Systems*, 32.
- Miller, M. I., Trounev, A., and Younes, L. (2006). Geodesic shooting for computational anatomy. *Journal of Mathematical Imaging and Vision*, 24(2):209–228.

- Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*.
- Moosavi-Dezfooli, S.-M., Fawzi, A., and Frossard, P. (2016). Deepfool: a simple and accurate method to fool deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2574–2582.
- Nabi, R. and Shpitser, I. (2018). Fair inference on outcomes. In *AAAI Conference on Artificial Intelligence*, volume 32.
- Neal, B. (2020). *Introduction to causal inference from a machine learning perspective*. bradyneal.com. https://www.bradyneal.com/Introduction_to_Causal_Inference-Dec17_2020-Neal.pdf.
- Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., and Swami, A. (2017). Practical black-box attacks against machine learning. In *ACM ASIA Conference on Computer and Communications Security*, pages 506–519.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32.
- Pearl, J. (2009). *Causality*. Cambridge university press.
- Pearl, J. (2010a). Brief report: On the consistency rule in causal inference: "axiom, definition, assumption, or theorem?". *Epidemiology*, pages 872–875.
- Pearl, J. (2010b). The mathematics of causal relations. *Causality and Psychopathology: Finding the Determinants of Disorders and their Cures*, pages 47–65.
- Pearl, J., Glymour, M., and Jewell, N. P. (2016). *Causal Inference in Statistics: A Primer*. John Wiley & Sons.
- Peterson, V., Nieto, N., Wyser, D., Lambercy, O., Gassert, R., Milone, D. H., and Spies, R. D. (2021). Transfer learning based on optimal transport for motor imagery brain-computer interfaces. *IEEE Transactions on Biomedical Engineering*, 69(2):807–817.
- Petsiuk, V., Das, A., and Saenko, K. (2018). RISE: randomized input sampling for explanation of black-box models. In *British Machine Vision Conference*, page 151. BMVA Press.
- Peyré, G. and Cuturi, M. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.
- Plecko, D. and Meinshausen, N. (2020). Fair data adaptation with quantile preservation. *Journal of Machine Learning Research*, 21:1–44.
- Pooladian, A.-A. and Niles-Weed, J. (2021). Entropic estimation of optimal transport maps. *arXiv preprint arXiv:2109.12004*.

- Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., and Flach, P. (2020). FACE: feasible and actionable counterfactual explanations. In *AAAI/ACM Conference on AI, Ethics, and Society*, pages 344–350.
- Rakotoarison, H., Milijaona, L., Rasoanaivo, A., Sebag, M., and Schoenauer, M. (2022). Learning meta-features for AutoML. In *International Conference on Learning Representations*.
- Rasouli, P. and Chieh Yu, I. (2022). CARE: Coherent actionable recourse based on sound counterfactual explanations. *International Journal of Data Science and Analytics*, pages 1–26.
- Redko, I., Courty, N., Flamary, R., and Tuia, D. (2019). Optimal transport for multi-source domain adaptation under target shift. In *International Conference on Artificial Intelligence and Statistics*, pages 849–858. PMLR.
- Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why should i trust you?" explaining the predictions of any classifier. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 22, pages 1135–1144.
- Richardson, T. S. and Robins, J. M. (2013). Single world intervention graphs (SWIGs): a unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences*, 128(30):2013. Working Paper.
- Risser, L., Sanz, A. G., Vincenot, Q., and Loubes, J.-M. (2022). Tackling algorithmic bias in neural-network classifiers using Wasserstein-2 regularization. *Journal of Mathematical Imaging and Vision*, pages 1–18.
- Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.
- Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The Annals of Mathematical Statistics*, 23(3):470–472.
- Rothenhäusler, D., Meinshausen, N., Bühlmann, P., Peters, J., et al. (2021). Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 83(2):215–246.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.
- Rudin, W. (1991). *Functional Analysis*. International series in pure and applied mathematics. McGraw-Hill.
- Russell, C., Kusner, M. J., Loftus, J., and Silva, R. (2017). When worlds collide: Integrating different counterfactual assumptions in fairness. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Salmona, A., De Bortoli, V., Delon, J., and Desolneux, A. (2022). Can push-forward generative models fit multimodal distributions? *arXiv preprint arXiv:2206.14476*.

- Santambrogio, F. (2017). {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7(1):87–154.
- Schreuder, N. (2020). Bounding the expectation of the supremum of empirical processes indexed by Hölder classes. *Mathematical Methods of Statistics*, 29(1):76–86.
- Scutari, M., Vitolo, C., and Tucker, A. (2019). Learning bayesian networks from big data with greedy search: computational complexity and efficient implementation. *Statistics and Computing*, 29(5):1095–1108.
- Seguy, V., Damodaran, B. B., Flamary, R., Courty, N., Rolet, A., and Blondel, M. (2018). Large-scale optimal transport and mapping estimation. In *International Conference on Learning Representations*, pages 1–15.
- Séjourné, T., Feydy, J., Vialard, F.-X., Trounev, A., and Peyré, G. (2019). Sinkhorn divergences for unbalanced optimal transport. *arXiv preprint arXiv:1910.12958*.
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision*, pages 618–626.
- Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.
- Shen, Z., Feydy, J., Liu, P., Curiale, A. H., San Jose Estepar, R., San Jose Estepar, R., and Niethammer, M. (2021). Accurate point cloud registration with robust optimal transport. *Advances in Neural Information Processing Systems*, 34:5373–5389.
- Shimizu, S., Hoyer, P. O., Hyvärinen, A., Kerminen, A., and Jordan, M. (2006). A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10).
- Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pages 3145–3153. PMLR.
- Shrikumar, A., Greenside, P., Shcherbina, A., and Kundaje, A. (2016). Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*.
- Slack, D., Hilgard, S., Lakkaraju, H., and Singh, S. (2021). Counterfactual explanations can be manipulated. *Advances in Neural Information Processing Systems*, 34.
- Smith, J. A. and Todd, P. E. (2005). Does matching overcome LaLonde’s critique of nonexperimental estimators? *Journal of Econometrics*, 125(1-2):305–353.
- Sotiras, A., Davatzikos, C., and Paragios, N. (2013). Deformable medical image registration: A survey. *IEEE Transactions on Medical Imaging*, 32(7):1153–1190.
- Stalnaker, R. C. (1980). A defense of conditional excluded middle. In *Ifs*, pages 87–104. Springer.

- Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR.
- Tanielian, U. and Biau, G. (2021). Approximating Lipschitz continuous functions with group-sort neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 442–450. PMLR.
- Thibaut Le Gouic, Jean-Michel Loubes, and Philippe Rigollet (2020). Projection to fairness in statistical learning. *arXiv preprint arXiv:2005.11720*.
- Torous, W., Gunsilius, F., and Rigollet, P. (2021). An optimal transport approach to causal inference. *arXiv preprint arXiv:2108.05858*.
- Uscidda, T. and Cuturi, M. (2023). The Monge gap: A regularizer to learn all transport maps. *arXiv preprint arXiv:2302.04953*.
- Ustun, B., Spangher, A., and Liu, Y. (2019). Actionable recourse in linear classification. In *Conference on Fairness, Accountability, and Transparency*, pages 10–19.
- Van Der Vaart, A. W. and Wellner, J. A. (1996). Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer.
- Villani, C. (2003). *Topics in optimal transportation*. Number 58 in Graduate Studies in Mathematics. American Mathematical Soc.
- Villani, C. (2008). *Optimal Transport: Old and New*. Number 338 in Grundlehren der mathematischen Wissenschaften. Springer, Berlin.
- von Kügelgen, J., Karimi, A.-H., Bhatt, U., Valera, I., Weller, A., and Schölkopf, B. (2022). On the fairness of causal algorithmic recourse. In *AAAI Conference on Artificial Intelligence*, volume 36, pages 9584–9594.
- Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31:841.
- Younes, L. (2019). *Shapes and diffeomorphisms*. Springer. Second Edition.
- Younes, L. (2020). Diffeomorphic learning. *Journal of Machine Learning Research*, 21(220):1–28.
- Zafar, M. B., Valera, I., Ródriguez, M. G., and Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. In *International Conference on Artificial Intelligence and Statistics*, pages 962–970. PMLR.
- Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833. Springer.
- Zhang, K. and Chan, L.-W. (2006). Extensions of ICA for causality discovery in the Hong Kong stock market. In *International Conference on Neural Information Processing*, pages 400–409. Springer.
- Zhang, Y. and Zhao, Q. (2023). What is a randomization test? *Journal of the American Statistical Association*, (just-accepted):1–29.