A primer on optimal transport for causal inference with observational data

Florian Gunsilius (1)

Abstract. The theory of optimal transportation has developed into a powerful and elegant framework for comparing probability distributions, with wide-ranging applications in all areas of science. The fundamental idea of analyzing probabilities by comparing their underlying state space naturally aligns with the core idea of causal inference, where understanding and quantifying counterfactual states is paramount. Despite this intuitive connection, explicit research at the intersection of optimal transport and causal inference is only beginning to develop. Yet, many foundational models in causal inference have implicitly relied on optimal transport principles for decades, without recognizing the underlying connection. Therefore, the goal of this review is to offer an introduction to the surprisingly deep existing connections between optimal transport and the identification of causal effects with observational data—where optimal transport is not just a set of potential tools, but actually builds the foundation of model assumptions. As a result, this review is intended to unify the language and notation between different areas of statistics, mathematics, and econometrics, by pointing out these existing connections, and to explore novel problems and directions for future work in both areas derived from this realization.

Key words and phrases: causal inference, difference-in-differences, identification, instrumental variables, monotone rearrangement, observational data, optimal transport, synthetic controls, treatment heterogeneity.

1. INTRODUCTION

At the heart of many scientific and practical problems lies the challenge of understanding causal relationships of interventions and modifications: will I feel better if I take this medication? What are the health effects on the population of opening up a factory near this neighborhood? These questions all revolve around a fundamental principle: identifying and analyzing the effect an intervention—taking the medication, building the factory in this neighborhood—has on the system under consideration, that is, my body or the neighborhood. Such questions are difficult to answer with data since they require the knowledge of two mutually exclusive states of the world, the counterfactual states. Either the factory is built at this location or it is not, but it is impossible to ever observe both states at the same time. This is the "fundamental problem of causal inference" [75].

Department of Economics, Emory University (e-mail: fgunsil@emory.edu). Zheng Fang, Rex Hsieh, Esfandiar Maasoumi, Guido Romero, and Alejandro Sanchez-Becerra provided greatly helpful feedback. All errors are the author's.

This problem is amplified in settings where the statistician does not have the agency to devise the intervention directly—as is the case in randomized controlled trials for instance—but only has access to observational data. In such cases the treatment variable is endogenous to the system, meaning that it is not independent of the unobservables that affect the outcome of interest. A classic example is self-selection into treatment [7, 82], where treatment is not randomized, but unobservable criteria affect both the uptake of the treatment as well as the outcome. For instance, motivation to find a new job after losing it recently will make individuals more likely to sign up for potential job training programs (the treatment) but also makes them more likely to send more applications and land a new job (the outcome), something documented by the classic Ashenfelter Dip [71]. Therefore, simply comparing the outcome in both groups—the people in the job training program versus the rest of the population—does not provide the causal effect of attending the job training program. One essentially compares two different groups based on characteristics and would falsely attribute the difference exclusively to the treatment effect of attending the job training program.

Such problems have been at the center of interest in classical econometrics since the 1920's [65, 132, 142], starting with the analysis of supply and demand models. Fundamentally, in such settings, identification, estimation, and inference of the effect of the intervention on the outcome of interest requires additional assumptions tailored to the specific case. Over the decades, statisticians and econometricians have developed a plethora of such methods, including instrumental variable estimation, matching, difference-in-differences, and synthetic controls.

Traditionally, the focus in this literature has been on identifying average effects, essentially estimating an expected effect of the intervention on the outcome of interest in the system. While this is often an important measure, modern research in this area has emphasized the importance of accounting for *treatment heterogeneity* [69].

A classical and influential example is the literature on the effect of minimum wage increases on employment [29, 107]. To analyze the effect a minimum wage increase had on the employment growth rate at fast food chains in the state of New Jersey in 1992, David Card and Alan Krueger [29] used a difference-in-differences approach focusing on average effects; they compared the average change in the employment rate at the surveyed New Jersey fast-food restaurants to the average change of employment at fast-food restaurants in the neighboring Pennsylvania, where the minimum wage remained constant. Somewhat counterintuitively, they found that an increase in the minimum wage increased employment. In stark contrast, David Neumark and William Wascher [107], using a different dataset and different methodology, found a negative average effect.

While the debate of the effects of the minimum wage on employment is still ongoing, an interesting contribution [119] attempts to reconcile these two opposite results by considering the entire distribution of the size of fast-food restaurants in each state—using the changes-in-changes estimator [10], which is an extension of the difference-in-differences idea to account for individual heterogeneity, implicitly built using optimal transportation as shown below. The author finds that the positive effect seems to persist for smaller fast-food restaurants while a negative effect can be found for larger fast-food restaurants. This is a striking example of how accounting for heterogeneity in the causal framework can uncover important insights into the problem that an average effect cannot.

The goal of this review is to show how optimal transportation can be used to resolve such and similar problems by allowing to capture *heterogeneity* within the respective setting [68, 79]. Importantly, optimal transport allows to do this in two ways, the second being arguably more novel: the first way fundamentally focuses on identifying causal effects for different individuals in the outcome distribution. In contrast to classical approaches that

focus on *average* causal effects, this framework explicitly focuses on the heterogeneity of effects, away from the average, often by considering the quantiles in the potential outcome distributions. The second way is based on a "systems view". Here, the goal is not to identify the different individual effects within the system, but to analyze the different counterfactual states of the system *as a whole*.

This review therefore has several goals.

- 1. One goal is to illustrate that not just the methods, but actually the models and assumptions in the literature on identifying causal effects with observational data are implicitly based on optimal transportation. Moreover, I aim to provide a common ground to use optimal transport as a paradigm to connect different areas from econometrics, statistics, machine learning, and artificial intelligence that—knowingly or unknowingly—use optimal transportation in the causal argumentation. This seems useful, because in recent years, there has been an increase in articles explicitly using optimal transportation methods to do causal inference, and many of these approaches [e.g. 94] do not reference existing related results [e.g. 10, 79].
- 2. This review first and foremost focuses on *identification* in causal inference with observational data. Identification arguments are of paramount importance in causal inference and precede estimation and inference: researchers need to show that they are actually able to obtain the correct effect in the population before focusing on estimating it. This makes causal inference significantly more complicated than prediction, because the observable distributions in general do not provide correct estimates—unobservable confounders bias estimation results, and one needs special tools to circumvent the resulting endogeneity issues, some of which I outline in this review.
- 3. A third goal is to show how optimal transport can be used to enhance, augment, and generalize many of the existing linear regression methods, in particular when it comes to the identification of heterogeneous effects [69], or identifying effects on entire systems [63]. The key is to generalize the assumptions and models from the literature to more general settings.
- 4. Finally, I attempt to point out some open problems and potentially new connections in this quickly expanding area. For instance, I show how realizing that the influential control variable approach [81] is based on optimal transportation allows to straightforwardly generalize it to make it vastly more applicable in practical settings. As another example, I provide a connection between partial identification in instrumental variable models and optimal transport problems on path space.

Of course, this overview is necessarily biased towards my own area of research, but I made every effort to provide a general overview, trying to connect existing results from diverse areas.

2. A QUICK REFRESHER ON OPTIMAL TRANSPORT

The basic problem of optimal transportation is the following. Suppose we are given two probability distributions P and Q on some underlying sets (supports) \mathcal{X} and \mathcal{Y} . The original problem posed by Gaspard Monge [104] is to find an optimal map $T: \mathcal{X} \to \mathcal{Y}$ that preserves the mass between P and Q and minimizes the overall cost of transporting P onto Q measured in terms of the cost function $c: \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_0^+$, that is

(2.1)
$$\min_{\substack{T: \mathcal{X} \to \mathcal{Y} \\ T_x P = O}} \int c(x, T(x)) dP(x),$$

where $T_\#P(A) = P\left(\{x \in \mathcal{X}: T(x) \in A\}\right)$ is the pushforward measure from P via T. This problem need not have a solution as there need not exist a function T that accomplishes this, for instance when P has fewer points in its support than Q. The cost function c(x,y) is given and encapsulates all the important information on the problem. In many settings, it is simply a distance function on the underlying space on which the probability measures are supported. One of the arguably most influential contributions to theory was the following convex relaxation of the problem by Leonid Kantorovich, which instead of asking for a map T, only requires a coupling between P and Q, that is, a joint distribution γ on $\mathcal{X} \times \mathcal{Y}$ that solves the linear program

(2.2)
$$\min_{\gamma \in \Gamma(P,Q)} \int_{\mathcal{X} \times \mathcal{Y}} c(x,y) d\gamma(x,y),$$

where $\Gamma(P,Q)=\{\gamma\in P(\mathcal{X}\times\mathcal{Y}):\pi_1\gamma=P,\pi_2\gamma=Q\}$ is the set of all couplings of P and Q [85] and π_1 (respectively: π_2) denotes the projection onto the first (respectively: second) marginal distribution. (2.2) is the Kantorovich problem. It is a convex relaxation of the Monge problem (2.1) in the sense that if (2.1) has a solution T then the optimal coupling solving (2.2) is supported on the graph of T in general. Both the minimum and the optimizer are of interest, particularly in economics [54], because there one is often interested in the optimal allocation and matching of resources.

In the case where the cost function is the squared Euclidean distance $c(x,y)=|x-y|^2,\,x,y\in\mathbb{R}^d$, the square root of the value function, i.e., the square root of the minimum achieved via (2.2), can be shown to possess all the properties of a metric. In this case, the value function is denoted by $W_2^2(P,Q)$ and is called the (square of the) 2-Wasserstein distance. Technically, it would be more accurate to at least call it the "Monge-Kantorovich-Wasserstein" distance, but this nomenclature has stuck.

Finally, in the case of the squared Euclidean distance as a cost function and if P is absolutely continuous with respect to Lebesgue measure, then Brenier's theorem [e.g. 140, Theorem 2.12] implies that the optimal transport map T in (2.1) takes the form of the gradient of a convex function, i.e., $T(x) = \nabla \varphi(x)$ for some convex φ . This property has been a main contributor to a surge of optimal transport methods in statistics and econometrics [e.g. 30, 43, 60], because the gradient of a convex function is a natural generalization of a monotone function.

Since then the contributions to theory have exploded to dynamic [18], weak [12, 57], multimarginal [6, 31], regularized [40], unbalanced [37, 93, 129], and geometric representations [83], just to name a few. In the following, I introduce the respective mathematical theory whenever it is needed. For general overviews of the basic concepts, consider [113, 115, 128, 139, 140].

Fundamentally, the fact that optimal transportation induces a coupling between probability measures by map-ping between the underlying state spaces $\mathcal X$ and $\mathcal Y$ provides a natural connection to causal inference if one considers P and Q probability measures of outcomes of interest and unobservable confounders. To showcase the utility of optimal transportation, I now introduce a classical setting of structural equation models and show how optimal transportation has implicitly provided the underlying structure for the fundamental identification and estimation approaches in this area.

3. THE FOUNDATION: STRUCTURAL MODELS, COUNTERFACTUALS, AND THE MONOTONE REARRANGEMENT

This section analyzes how optimal transportation allows for the identification of causal effects while accounting for individual heterogeneity. In fact, I show how ideas from optimal transportation, in particular the *monotone rearrangement*, have built the foundation for models and assumptions in this area and how this insight can be used for generalizations and to develop new methods.

3.1 Structural models

A way to formalize the setup from the introduction is via *structural models* of the form

$$(3.1) Y = g(X, U),$$

where $Y:\Omega\to\mathbb{R}$ is the observed outcome of interest and $X:\Omega\to\mathbb{R}^d$ is the observed treatment variable, that is, the variable in the system whose effect on the outcome we want to isolate. $U:\Omega\to\mathcal{U}$ accounts for all other unobserved influences in the system, in particular the *treatment heterogeneity*: different values of U introduce different causal mechanisms between X and Y. \mathcal{U} can potentially be infinite dimensional All variables are defined on a common probability space (Ω,\mathcal{A},P) . In many classical

causal inference settings X is a univariate binary variable taking values in $\{0,1\}$, indicating the treatment status, but it can of course be more general. The unobserved parts in this model are the variable U and the causal mechanism g(X,U).

The generality in the assumption of the causal mechanism serves to not introduce additional structural assumptions into the model besides a postulated relationship between the variables Y, X, as part of the larger system. U captures all the unobservable factors in the system that affect the outcome Y and potentially the treatment X. This means that U and X are *not* assumed to be independent, as would be the case in a randomized controlled trial. This captures the fundamental problem of causal inference in this setting, as X is now itself affected by a change in Y through a back door path via U, depicted as a directed acyclic graph (DAG) [110] in Figure 1.

This induced "feedback" loop is the classical endogeneity problem, i.e., X is endogenous to the system so that the effect $X \to Y$ cannot be extracted directly by simple prediction. Therefore, for identification in these settings one needs more information and stronger assumptions to close the "back-door path" $U \to X$, as I will discuss below in examples such as instrumental variable models and difference-in-differences estimation. Before doing this, we first need to introduce the foundational ingredients for identifying causal effects in this setting by outlining how identification of treatment effects has been based on the monotone rearrangement, a solution to specific types of optimal transport problems.

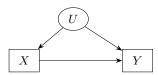


FIG 1. The DAG corresponding to (3.1) illustrating the backdoor path through the unobservable U.

The model (3.1) is deliberately general and in particular encompasses additively separable structures (g(X, U) =f(X) + U) and even more restrictive linear models $(g(X,U)=X^{\top}\beta+U)$. The focus is also on identifying the mechanism g(X, u) directly, i.e., the relationship $X \to Y$ while controlling for the unobservable U = u. The heterogeneity is captured by the unobservable U, which contains all characteristics that are unobserved to the statistician, but are of importance to the causal question of interest. Allowing for g to depend nonseparably on U allows for this heterogeneity to be of the most general form, without the need to introduce artificial structural assumptions like additive separability. It has been argued that additive separability is an artificial assumption in most settings of interest when dealing with human behavior [e.g. 74].

3.2 Relation to counterfactuals and other causal questions

Structural models also induce the classical counterfactuals Y(x) [108, 126, 76]. Here, Y(x) is the counterfactual outcome if the treatment X takes the value x. In the classical binary setting $X \in \{0,1\}$, Y(1) is the counterfactual under treatment while Y(0) is the counterfactual under no treatment. The fundamental problem of causal inference for an independent and identically distributed sample $\{Y_i, X_i\}_{i=1}^n$ in the simple binary case $X \in \{0,1\}$ can be stated simply as

$$(3.2) Y_i = (1 - X_i)Y_i(0) + X_iY_i(1),$$

that is, one can only ever observe one potential outcome in practice. Structural equations are sometimes argued to be more general than the counterfactual notation because they explicitly note the importance of the unobserved heterogeneity U in the model [111]. In particular, equivalent to structual equation models are the *counterfactual processes* $Y_x(u)$, which denote the counterfactual outcome under treatment x if the system is in the state U=u [15, 110]. Below I show how those processes induce new and interesting questions that relate to optimal transportation—analyzing those can open connections between optimal transport, statistical physics, and causal inference.

The model (3.1) is completely general in that it does not impose any structural assumptions on the system, and in particular U, a priori. It is also different in focus and in a sense more general than the models in other areas of causal inference for machine learning [e.g. 112]. The difference is that the goal in our setting is not to analyze the system and to understand the relationships between all the variables in the system relegated to the error term U, but to simply isolate the causal effect $X \to Y$ among all the other moving parts (captured by U) in the system. Intuitively, for our questions of causal inference with observational data, we simply want to control the unobservable variable U, while in other settings it might be interesting to explicitly model the relationships among all the unobservable variables and their effects on the outcome via directed acyclic graphs [110, 112]. Optimal transport is also being used in such settings [33, 135], based on the idea of optimal transport between different processes [13], exploiting the Markovian structure of DAGs.

The difference between those two viewpoints is important from an applied researcher's perspective, as one often wants to be as agnostic as possible about the other relationships in the system and only focus on the one relationship between X and Y. Consider [80] for further reading about this distinction. Throughout, this review concentrates on the latter, i.e., extracting the effect $X \to Y$, not modeling the latent space. I now show how optimal transport has built the foundation for this, via the monotone rearrangement.

3.3 Setup: identification and monotone rearrangement

The key question is to identify, that is, isolate, the causal effect $X \to Y$ from the system. "Identification" is defined as injectivity of the observable law $P_{Y,X}$ induced by the model (3.1) in unobservable quantities g, P_U , where the latter is the law of U [99, 110]. The focus for identification is the function g, as it encodes the effect $X \to Y$. Of course, without additional assumptions, it is impossible to identify the mechanism g; in such cases one can only obtain bounds on specific functionals, as I show below. I now quickly outline the predominant assumption to identification in this setting and show how it is fundamentally based on optimal transportation.

Let G be the set of all functions g we consider and Γ a set of laws P_U of U in question. Formally, the definition of identification is the following.

DEFINITION 3.1 ([99]). The pair (g, P_U) is identified in the set $(G \times \Gamma)$ if (i) $(g, P_U) \in (G \times \Gamma)$, and (ii) for all (g', P'_U) , in $(G \times \Gamma)$,

$$[P_{Y,X}(\cdot,\cdot;g,P_U) = P_{Y,X}(\cdot,\cdot;g',P'_U)]$$

$$\Rightarrow (g,P_U) = (g',P'_U).$$

This means that the map from the observable joint distribution $P_{Y,X}$ of outcome and treatment is injective in the model: two different models generate different observable outcomes. This would lead to point-identification, i.e., obtaining a unique function g(x,u).

3.3.1 The basic idea in exogenous settings

Before stating approaches for identification when X is truly endogenous in the model, I focus on the case where X is actually exogenous, that is, where X and U are independent (denoted in the following by $X \perp U$), so that the arrow $U \to X$ in Figure 1 is absent. This setting is of course significantly simpler, actually circumventing the main causal inference problem, but it serves to illustrate the first fundamental connection between optimal transport and such nonseparable representations.

A simple sufficient condition for identification in the case where all variables Y, X, and U are univariate, F_U , the cumulative distribution function (CDF) of the law of U is absolutely continuous and $X \perp U$ is monotonicity of the causal mechanism. In fact, the main assumption that guarantees identification of the causal mechanism up to some equivalence class is *continuity and strict monotonicity* of g in U for all x [45]. This follows from the simple string of equalities for almost every x:

(3.3)
$$F_{Y|X=x}(y) = P(Y \le y | X = x)$$
$$= P(g(X, U) \le y | X = x)$$
$$= P(U \le g^{-1}(X, y) | X = x)$$
$$= F_U(g^{-1}(x, y)),$$

where $F_{Y|X}$ is the conditional CDF of the observables Y and X. The invertibility of g follows because it is assumed to be continuous and strictly increasing.

This simple argument in the univariate case implies that the function g is identified up to observational equivalence in the exogenous setting $X \perp U$; in this sense two functions $g, g' \in G$ are observationally equivalent if there exist $F_U, F_U' \in \Gamma$ such that they all induce the same observational distribution, i.e.,

$$F_{Y,X}(\cdot,\cdot;g,F_U) = F_{Y,X}(\cdot,\cdot;g',F_U').$$

This implies that the entire mechanism g is identified up to a monotone transformation, since any monotone transformation will provide an observationally equivalent mechanism.

While this argument seems straightforward, its impact on the literature on identification of effects cannot be overstated. In fact, this idea is used in many fundamental approaches to identify causal effects, such as instrumental variable models [81, 72, 133] and extensions of the difference-in-difference method [10], which I revisit below. It is thus interesting to analyze the limits and potential extensions of this idea. This is facilitated by the connection to optimal transportation.

3.3.2 The monotone rearrangement

The connection to optimal transport comes from the fact that the identification argument (3.3) implies that $g(x,\cdot)$ is the *monotone rearrangement* between the unobservable F_U and the observable $F_{Y|X=x}$ for P_X -almost every x. That is, g is the solution to the Monge problem (2.1) transporting P_U to $P_{Y|X=x}$ for any symmetric convex cost function c(|u-y|), which exists if P_U does not have atoms (that is, gives positive mass to single points) [140].

The monotone rearrangement $g(x,\cdot)$ takes the form

(3.4)
$$g(x,u) = F_{Y|X=x}^{-1} (F_U(u)),$$

where $F_Z^{-1}(q)=\inf\{z\in\mathbb{R}:F_Z(z)\geq q\}$ is the quantile function of the random variable Z. It is an optimal transport map in the sense of Monge (2.1), because (for fixed x) it maps every point u in the support of P_U to a point y=T(u)=g(x,u) in the support of $P_{Y|X=x}$. Moreover, it is measure preserving, meaning that the preimage of a measurable subset B in the support of Y|X=x gets mapped to a subset $A=T^{-1}(B)$ in the support of U that has the same measure. Importantly, the monotonicity requirement makes this map the unique measure- and order preserving transformation, as depicted in Figure 2.

In this case, the monotonicity assumption for identification can be rephrased as: suppose U is a univariate index accounting for the heterogeneity of the system in question and assume that the causal mechanism captured by g(x,u) is such that for each hypothetical state u of the

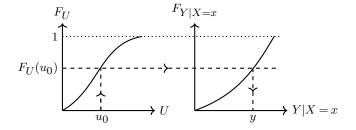


FIG 2. Depiction of the monotone rearrangement $y=g(x,u_0)=F_{Y|X=x}^{-1}(F_U(u_0))$

system, the counterfactual outcome Y(x) is created as the optimal transport coupling induced by a convex cost function. Then if the treatment is exogenous (i.e., $X \perp U$), the causal mechanism is the solution to the optimal transport problem with convex symmetric cost.

This simple connection allows to analyze and extend (3.3) in many different ways. For one, it connects this argument to other areas, in particular to *structural nested distribution models* [117, 116, 138], as pointed out in [14]. Structural nested distribution models are generalization of structural nested mean models, where the relationship between counterfactual outcomes Y(x) and Y(x') for different realizations x, x' is modeled by some general *known* function

$$\alpha(y, x, x', \beta) = F_{Y(x)|X=x'}^{-1} \left(F_{Y(x')|X=x'}(y) \right),$$

which is parametrized by β . In short, for structural nested distribution models [136], one assumes a parametric form for the monotone rearrangement as the solution to the optimal transport map between the counterfactual distributions, thus constructing a joint distribution over the joint potential outcomes $\{Y(x)\}_{x\in\mathbb{R}}$.

Moreover, the connection of (3.3) to the monotone rearrangement links the identification argument to the classical Fréchet-Hoeffding bounds [115]. The coupling γ induced by the monotone rearrangement g(x,u), which solves the Kantorovich problem (2.2) has the CDF

$$H(u, y; x) = \min\{F_U(u), F_{Y|X=x}(y)\}\$$

for P_X -almost every x. This is the upper Fréchet-Hoeffding copula, the copula that makes F_U and $F_{Y|X=x}$ comonotone, maximizing their dependence.

This rephrasing clarifies the obvious restrictiveness of the monotonicity assumption in (3.3). In many settings, comonotonicity is reasonable. For instance when analyzing a hypothetical setting of the causal effect of the gradepoint average in high-school (X) on future earnings (Y) of students, where the unobservable U is understood as an "index of ability" of the student. In this case, it is reasonable to assume that students with higher ability will also earn more conditional on their GPA. Of course, it is not

difficult to conceive of other settings where co- or countermonotonicity in an unobservable index are not reasonable; especially settings where extreme realizations in U lead to bad outcomes. These extensions are easy to model by picking different cost functions, for instance concave cost functions in the Monge problem (2.1), which provide optimal maps g(x,u) that have exactly this property [100].

3.3.3 Connection to the counterfactual notation

The connection to the Fréchet-Hoeffding copulas of course also allows to make a connection to Rubin's counterfacutal model in the actually interesting case where $X \not\perp U$. The fundamental problem of causal inference in this setting can also be stated from a coupling perspective that is more amenable to tools from optimal transportation. Consider the binary case $X \in \{0,1\}$ for simplicity. Then (3.2) implies that one can estimate the marginal distributions Y(1) and Y(0) by considering the two groups separately. The important question, however, is to identify the *joint* distribution of the potential outcomes (Y(0),Y(1)), which captures the causal mechanism. Without further assumptions or information, this joint distribution cannot be identified if $X \not\perp U$.

One can obtain bounds on the causal mechanism of interest by the standard Fréchet-Hoeffding bounds. In the binary setting, this has been done in [9, 50, 70]. The Fréchet-Hoeffding bounds for the joint distribution of (Y(0), Y(1)) are [e.g. 115, Theorem 3.1.1]

(3.5)
$$F_{Y(0)}(y) + F_{Y(1)}(y') - 1$$

 $\leq F_{Y(0),Y(1)}(y,y')$
 $\leq \min \left\{ F_{Y(0)}(y), F_{Y(1)}(y') \right\}.$

Based on this, one can bound other expressions such as the covariance Cov(Y(0), Y(1)) or any expression of the form $\mathbb{E}[c(Y(0), Y(1))]$ via

(3.6)
$$\int_{0}^{1} c(F_{Y(0)}^{-1}(u), F_{Y(1)}^{-1}(u)) du$$

$$\leq \mathbb{E}[c(Y(0), Y(1)]]$$

$$\leq \int_{0}^{1} c(F_{Y(0)}^{-1}(u), F_{Y(1)}^{-1}(1-u)) du$$

for any quasi-antitone cost function c, that is,

$$c(x', y') + c(x, y) \le c(x', y) + c(x, y')$$

for any $x' \ge x$ and $y' \ge y$, which is a classic inequality proved in [28]. These bounds are wide in general, often too wide to be of use, but they do provide a minimal restriction on the respective data-generating process.

I would be amiss to not mention that recently, [14] analyzed lower bounds on functionals of the joint distribution of all potential outcomes Y(x) in a setting where the

treatment X is continuous. By solving a general optimal transport problem, i.e., minimizing functionals such as the "quadratic effect"

$$\mathbb{E}\left[\int\int |Y(x)-Y(x')|^2 dP_X(x) dP_X(x')\right],$$

where P_X is the marginal distribution of the treatment Xand the expectation is over all joint laws of the counterfactual outcomes $\{Y(x)\}_{x\in\mathbb{R}}$ they obtain lower bounds on these effects. It can hence be seen as a generalization of the Fréchet-Hoeffding approach to continuous treatments. The same issue as above holds here: the bounds are usually wide without additional assumptions. Moreover, it is not always clear when a quadratic effect like this is interesting—recall that the assumption of the monotone rearrangement as the causal mechanism allows to identify the entire mechanism, not just functionals of it. Below in section 4.3 I make a connection to the problem of bounding treatment effects in instrumental variable models [15, 16], where more information is introduced into the above problem via linear constraints, leading to tighter bounds and a novel optimal transport adjacent optimization problem on path spaces.

3.3.4 Monotone rearrangement as the foundation of causal inference with observational data

The monotone rearrangement has been the workhorse for the identification of causal effects in a variety of different areas. Most likely, the main reason for this is its interpretability, it being an order preserving map. We have also seen, however, that it is in general not enough to identify the causal effects without stronger assumptions. First, in the structural setting with Y = q(X, U), we had to assume that $X \perp U$, which is the simple setting of a randomized controlled trial, circumventing the causal identification problem entirely—and then we could only identify the mechanism up to a monotone transformation. Second, in the case of the counterfactual notation setting (in which case $X \not\perp U$), we have seen that it only provides bounds on functionals of the causal mechanism in general. On the other hand, the monotone rearrangement is the fundamental connection to works on the intersection of optimal transport and causal inference [42] that argue that the fundamental model between potential outcomes should be induced by optimal transport: in a way, this has always been the case, but as a connection between unobservables and outcome, not between the counterfactual distributions directly.

In the following, I therefore have two goals. First, I want to show how the realization that the monotone rearrangement is the solution to certain one-dimensional optimal transport problems can be used to generalize causal inference to higher dimensions and more realistic settings. Second, I want to introduce and generalize

classical approaches such as instrumental variable estimation, difference-in-differences, and synthetic controls to *uniquely identify* either the entire causal mechanism g(x,u)—or causal effects based on it—in the case where X is endogenous, i.e., $X \not\perp U$, by exploiting additional structure and assumptions in the problem.

3.3.5 Problems with extension to more general settings Realizing the connection between (3.3) and optimal transportation, there is now a straightforward way to generalize the setting to higher dimensions. That is, instead of requiring that U and Y be univariate and q(x, u) be the monotone rearrangement between U and Y|X=x, we can assume that U and Y are multivariate of the same dimension and that g(x, u) is the optimal transport map for the quadratic cost function $c(u, y) = |u - y|^2$. Considering x fixed, by Brenier's theorem [25, 140, Theorem 2.12], this map is uniquely defined P_U -almost everywhere and takes the form of the gradient of a convex function, that is $g(x, u) \equiv T_x(u) = \nabla \varphi_x(u)$ for some convex function $\varphi_x: \mathbb{R}^d \to \mathbb{R}$. Identification would then come from the fact that the Brenier map is the unique mapping with this monotonicity property, analogous to the monotone rearrangement.

Brenier's theorem has been the starting point for a recent explosion in interest in applications of static optimal transportation in statistics [e.g. 30, 34, 35, 43, 49]. The reason being that the gradient of a convex function is a "natural" generalization of a monotone function. In fact, it is not only monotone as a map $\mathbb{R}^d \to \mathbb{R}^d$, in the sense that

$$\langle T_x(u) - T_x(u'), u - u' \rangle \ge 0$$
 for all $u, u' \in \mathbb{R}^d$,

where $\langle \cdot, \cdot \rangle$ denotes the inner product on $\mathbb{R}^d \times \mathbb{R}^d$, but it actually possesses a *cyclically monotone* support [118, Theorem 24.8]. A map $T_x : \mathbb{R}^d \to \mathbb{R}^d$ is said to be cyclically monotone if for any positive integer m and any cycle $u_1, \ldots, u_m, u_{m+1} \equiv u_1$ in its domain, it holds

$$\sum_{i=1}^{m} \langle u_i, T_x(u_i) - T_x(u_{i+1}) \rangle \ge 0.$$

In the univariate setting, monotonicity and cyclic monotonicity coincide, but cyclic monotonicity provides more structure in multivariate settings, as the inequality has to hold for all finite m-cycles, while monotonicity only requires this for m=2. Below, I show how cyclic monotonicity is useful in the identification of individual effects.

While this extension is straightforward and useful in many areas of statistics, it is less useful in our setting. The main issue is the interpretation of the requirement that U and Y have to be of the same dimension. First, in causal inference settings, it is less common to have multivariate outcomes. Second, the interpretation of the unobservable in a multivariate setting is difficult. U is designed to

contain any covariate that can theoretically affect the outcome of interest and is not observable. Therefore, either assuming U is infinite dimensional or that U is univariate seem reasonable. In the first case, it is a general and allencompassing error term. In the second case it is an index that contains the information of all relevant unobservables (for instance an "ability"-index in the GPA example). The choice that U needs to be multivariate and of the same dimension as Y is therefore artificial in most applications.

In this sense, it is more realistic to assume Y is univariate or low-dimensional while U is higher-dimensional. Such generalizations have been analyzed in [36, 101]. The issue here is that unlike the case where the dimensions of Y and U align, the existence (and uniqueness) of an optimal transport map g(x, u) between $u \in \mathbb{R}^d$ and $y \in \mathbb{R}$ relies on the shapes of P_U and P_Y as well as the cost function c(u,y) in the optimal transport problem [36, Theorem 4.(b)]. When such a map exists in the setting $Y \in \mathbb{R}$, the authors call the corresponding model nested. The fact that the existence of g(x, u) depends on the geometries of $P_{Y|X=x}$ and P_U is interesting in general causal inference problems. It suggests that there are important differences in models that allow to identify causal models in these settings. Also, generalizing the results in [36, 101] to allow for infinite dimensional *U* is interesting.

While the standard generalization of the monotone rearrangements to Brenier maps in multiple dimensions is often not a realistic model for the relation between the outcome Y and the unobservables U in the system, it has proven to be very helpful and illuminating in actual approaches to causal inference when $X \not\perp U$; and there multivariate extensions are in fact important, as they allow to consider multiple endogenous treatments. I now review three of these approaches. The first is the *method of instrumental variables*, the second is difference-indifferences, and the third is synthetic controls.

4. INSTRUMENTAL VARIABLE MODELS: IDENTIFICATION, BOUNDS, ROBUSTNESS

Now is the first instance where we actually start dealing with the true causal inference problem depicted in Figure 1, i.e., where $X \not\perp U$. As indicated above, it is not possible to point-identify the structure of g(x,u) without additional assumptions, even if g is assumed to be the monotone rearrangement. The reason is the dependence between X and U, the backdoor channel (Figure 1). It implies that the observed conditional measure $P_{Y|X=x}$ is not the counterfactual measure of interest, which is denoted by $P_{Y|X^*}$, indicating that this is the conditional measure for the hypothetical scenario where $X \sim X^*$ but $X^* \perp U$. The observed joint measure $P_{Y,X}$ is expressed as

$$P_{Y,X} = \int P_{Y|X,U=u} P_{X|U=u} P_U(du).$$

If X were actually exogenous, that is, $X \perp U$, we would have $P_{X|U} = P_X$, so that we could identify $P_{Y|X^*}$ by

$$P_{Y|X} = \frac{P_{Y,X}}{P_X} = \frac{P_X \int P_{Y|X,U=u} P_U(du)}{P_X} = P_{Y|X^*}.$$

Since X^* is unobservable, one idea is to find a random variable Z that has its properties. In short, we would want a variable Z that has the same distribution as X but is itself independent of the unobservable U. This is too strong a restriction, so we require Z to be an *instrumental variable* with the following properties:

DEFINITION 4.1. A random variable $Z: \Omega \to \mathbb{R}$ is an instrument for the endogenous variable X in the model (3.1) if

- (i) (Relevance) $Z \not\perp X$,
- (ii) (Independence) $Z \perp U$,
- (iii) (Exclusion) The only influence Z has on the outcome Y is via X.

If Z satisfies (i) - (iii), we call it *valid* [82, 110]. Incorporating an instrument into our model, we can extend (3.1) to the following structural model.

(4.1)
$$Y = g(X, V)$$
$$X = h(Z, W), \quad Z \perp (W, V),$$

where we have split the unobservable error term U into an error term W of the *first stage* and an error term V of the second stage. This model is equivalent to a model where U appears in both the first- and second-stage, but is more convenient for the derivations below.

This model implicitly contains all the information we require from an instrumental variable model. Relevance is given if h is not constant in Z for every u. Independence is enforced by $Z \perp U$, and exclusion is captured by the fact that g is not a function of Z. One can depict (4.1) more elegantly as a DAG [110] as in Figure 3.

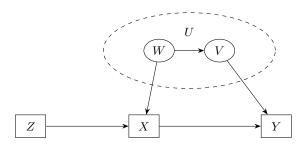


FIG 3. The DAG corresponding to (4.1).

To understand the intuitive idea, in a simple linear setting, i.e., where $g(X,V)=\alpha X+V$ and $h(Z,W)=\beta Z+W$, one can identify the average causal effect $X\to$

Y by the coefficient β^{IV} , which can be obtained via the Wald estimator $\beta^{IV} = \frac{\text{CoV}(Y,Z)}{\text{CoV}(X,Z)}$. This estimator provides an intuitive explanation in how instruments can be used to identify the causal effect even in more general settings. Since Z does not have a direct influence on Y and is itself independent of U, any correlation between Y and Z must go through (or in causal inference speak: "is mediated by") X. Therefore, to identify the causal effect in this setting, we only have to normalize this relationship by the relation between X and Z to identify the (linear) average causal effect $X \to Y$.

While similar, this argument is significantly more complex in the general nonseparable setting (4.1), which I focus on now. I briefly start with the basic idea of control variables, which captures the same information as the unobservable U. This allows to identify average effects [81]. I show how optimal transportation can be used to explore the limitations of this approach and provide a novel generalization. Then I show how arguments using fixed-point iterations [133, 44] allow to identify heterogeneous causal effects. I then see how far these can be extended to general diemnsions and what the limitations are [64]. In the process I introduce some interesting dynamic mathematical properties of optimal transport maps that warrant further investigation. Then I turn to the partial identification case [15, 16], where the goal is to obtain bounds on an average treatment effect. I argue that there are also some so far unexplored connections to optimal transportation in this setting that can be valuable to analyze. Finally, I want to mention the connection of distributionally robust optimization approaches [19] to instrumental variables estimation, which have recently been explored in linear models [114].

4.1 Control variables

From the classical backdoor criterion [110], we know that conditioning on W breaks the backdoor and allows us to extract the correct causal effect $X \to Y$: fixing W by conditioning on a realization W = w makes X = h(Z, w) just a function of the exogenous Z, so that by varying Z one can obtain the exogenous effect of X on Y. The issue is that W is unobservable, so that conditioning on it is impossible. This is where the control variable approach comes in [81].

DEFINITION 4.2. A random variable $R: \Omega \to \mathbb{R}^d$ is a control variable in the model (4.1) if $X \perp V | R$.

A control variable mimics the conditioning on W. Therefore, it needs to be constructed such that conditioning on W is the same as conditioning on R, i.e., their induced σ -algebras need to coincide. Doob's functional representation implies that the σ -algebras coincide if there exist maps ξ, ρ such that $R = \xi(W)$ and $W = \rho(R)$. In

particular, these maps may not depend on any other variables Y, Z, V [81], which is expected at this point, require W and X to be univariate and h(z, W) to be the monotone rearrangement in W for all z. Then they construct a control variable via

(4.2)
$$R = F_{X|Z}(X) = F_W(h^{-1}(Z,X)) = F_W(W).$$

They then assume F_W to be itself strictly increasing and continuous, that is, invertible, which makes R a valid control variable with the fixed function F_W .

Using this control variable approach, one can now identify average structural effects $\Lambda(X \to Y)$ of the form

$$\Lambda(X \to Y) = \int \Lambda(g(x, v)) dP_V(v),$$

which are unobservable because P_V is unobservable. For this, one needs the important *large support assumption*, which requires that for all realizations x of X, the support $\mathcal{R}(X)$ of the conditional measure $P_{R|X}$ is the same as that of the marginal measure R. Then, one can compute

$$\begin{split} E(\Lambda(Y)|X &= x, R = r) \\ &= \int \Lambda(g(x,v)) dF_{V|X=x,R=r}(v) \\ &= \int \Lambda(g(x,v)) dF_{V|R=r}(v). \end{split}$$

where the second equality follows from the fact that R is a control variable. Then by the large support assumption on R one can integrate over the marginal distribution of R to get

$$\int E(\Lambda(Y)|X=x,R=r)dF_R(r)$$

$$= \iint \Lambda(g(x,v))dF_{V|X=x,R=r}(v)dF_R(r)$$

$$= \iint \Lambda(g(x,v))dF_{V|R=r}(v)dF_R(r)$$

$$= \int \Lambda(g(x,v))dF_V(v).$$

As can be seen from the above argument, if the large support assumption does not hold, one can only identify the effect on the set where the supports of $P_{R|X=x}$ and P_R overlap for all x, which can be significantly smaller.

The large support assumption is testable in practice and is usually very badly violated [67], not least since many instruments have finite support [133]. In particular, the fact that the control variable is by construction uniform is a restricting factor.

The knowledge that this identification result is based on optimal transportation allows us to alleviate this issue slightly by replacing the map $F_{X|Z}(X)$ by a more general function $m^{-1}(X,Z)$ to generate an R

which need not have a uniform distribution. The important requirement on m^{-1} is that it can be written as $m^{-1}(X,Z) = T(h^{-1}(Z,X))$ for some measure-preserving isomorphism T, which must not depend on z. A measure-preserving isomorphism T is a map that is measurable and preserves the measure whose inverse exists and is also measurable and measure-preserving. In this case, the two σ -algebras induced by R and R coincide, so that R is a valid control variable. Using this idea, we have the following simple generalization of the control variable approach.

PROPOSITION 4.1 (Generalized control variables). Let F_W be absolutely continuous and strictly increasing and let h(z,w) be the monotone rearrangement between F_W and $F_{X|Z=z}$ for all z. If $F_{X|Z=z}$ is continuous in x for all z, then any univariate random variable R independent of Z for which there exists a measure-preserving isomorphism $T:[0,1]\to\mathbb{R}$ mapping the uniformly distributed $\tilde{R}=F_{X|Z}(X)$ to R can be made a control variable by setting

$$m^{-1}(X,Z) = T(F_{X|Z}(X)).$$

PROOF. The first part is the same as the proof of Theorem 1 in [81]. Let $h^{-1}(x,z)$ denote the inverse function of h(z,w). Then, by (3.3), it holds

$$F_{X|Z=z}(x) = F_W(h^{-1}(x,z)).$$

Plugging in the random variables X,Z into this expression gives

$$\tilde{R} = F_{X|Z}(X) = F_W(h^{-1}(Z,X)) = F_W(W),$$

where \tilde{R} is now a uniformly distributed random variable, i.e., $\tilde{R} \sim U[0,1]$. Now fix a random variable R for which there exists a measure-preserving isomorphism $\tilde{R} \mapsto T(\tilde{R}) = R$. Set

$$m^{-1}(X,Z) = T(F_{X|Z}(X)),$$

which implies

$$\begin{split} R = m^{-1}(X,Z) = & T\left(F_W(h^{-1}(Z,X))\right) \\ = & T\left(F_W(W)\right). \end{split}$$

Since F_W is strictly increasing and continuous, it is the unique monotone rearrangement between W and \bar{R} , and is measurable with measurable inverse since both \bar{R} and W do not give mass to points. In particular, F_W does not depend on X or Z. But the composition $T \circ F_W(W)$ is then a measure-preserving isomorphism that does not depend on X or Z. Hence, the σ -algebra induced by W is equal to the σ -algebra induced by R, so that conditional expectations given W are identical to those given R. Also, for any bounded function a(X), by $Z \perp (V, W)$

$$E[a(X)|V,W] = \int a(h(z,W))dF_Z(z) = E[a(X)|W].$$

Therefore, for any bounded function b(V), I have

$$\begin{split} E[a(X)b(V)|R] = & E[a(X)b(V)|W] \\ = & E[b(V)E[a(X)|V,W]|W] \\ = & E[b(V)E[a(X)|W]|W] \\ = & E[b(V)|W]E[a(X)|W] \\ = & E[b(V)|R]E[a(X)|R] \end{split}$$

which shows that R is a control variable.

Proposition 4.1 is a straightforward extension of the control variable approach once one realizes that the identification idea is again based on the monotone rearrangement. It is, however, very valuable in practical settings, as it can be used to achieve a much better coverage for the large support assumption.

A generalization of Proposition 4.1 to multivariate settings is essentially impossible since it is no longer the case that the half-open rectangles $(-\infty,w]$ and $(-\infty,w']$ necessarily have different probabilities if $w\neq w'$ for strictly increasing F_W , due to the lack of a natural complete order on \mathbb{R}^d . For instance, under the assumption that h(z,w) is the gradient of some convex function φ_z , that is, $h(z,w) := \nabla \varphi_z(w)$, which is the standard generalization of the monotone rearrangement to higher dimensions using Brenier's theorem as above [25, 140, Theorem 2.12], the important requirement

$$P_W(h^{-1}(z,A)) = P_{X|Z=z}(A) = P_R(m^{-1}(z,A))$$

does not hold for arbitrary Borel sets $A \subset \mathbb{R}^d$ and all z unless R = W, in which case the control variable approach is obsolete. The proof for this claim is omitted. Hence, one of the most influential approaches to identify causal effects in general instrumental variables seems to be restricted to univariate first stages; it would be interesting to investigate this further. Note, however, that there is no restriction on the second stage Y = g(X, V), the stage of interest.

4.2 Instruments with small support and dynamics of Brenier maps

As mentioned, the large support assumption is by definition violated when the instrument Z is not continuous [133]. In the following, I consider the case where Z is binary and can only take two values z, z'; the result can be straightforwardly extended to finitely many realizations. Such settings are ubiquitous in practice [79, 133]. Moreover, it would be nice to allow for several endogenous variables X and not just one, meaning that it is useful to generalize the first stage in (4.1) to a multivariate setting. Finally, the control variable approach only provided identification for average structural effects $\Lambda(X \to Y)$, but not the entire mechanism g(X, V), which would give

us full identification of the heterogeneous treatment effects.

To incorporate all of these modifications, we again start with the case where all variables are univariate. This has been introduced in [133] and [44]. The univariate model is the one where all variables, observable and unobservable, are univariate and where both g and h are assumed to be monotone rearrangements in V and W, respectively. Using the counterfactual notation [126], we can write $F_{Y(X)}$ as the counterfactual distribution of the effect of X on Y for an exogenous change in X (i.e., the counterfactual distribution). Due to the backdoor channel via V as depicted in Figure 3, this is not observable and the observable distribution $F_{Y|X}$ does not coincide with the counterfactual distribution.

The idea for identification of g is as before: vary Z in such a way that X varies but V stays constant. The fact that this is possible if Z is only binary was first shown in [133] and [44]. The argument of the former is more amenable to our ideas and rests on the fact that using the binary instrument Z with realizations z and z', there are two maps that do not change the distribution F_V of V, but change values of X. Using only those two maps therefore captures the exogenous effect of X on Y. These two maps are depicted in Figure 4, which is taken from [64].

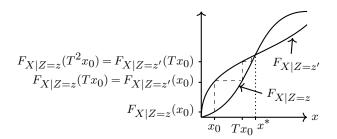


FIG 4. Fixed-point iteration in a univariate framework for identifying causal effects in (4.1).

The first map changes $z\mapsto z'$ for a fixed x, i.e. switches the distributions $F_{X|Z=z}(x)$ and $F_{X|Z=z'}(x)$ (the "vertical" map in Figure 4). The fact that $F_{V|X,Z}$ is not affected by this follows from the control variable approach introduced above [81] and because Z has no effect on g due to the exclusion restriction of Z. A control variable R can be constructed via $R=F_{X|Z}(X)$. Hence, by conditioning on R and the fact that Z is independent of (V,W), the vertical shift changes Z but does not change X, which therefore does not affect the function g(x,v) we want to identify [133].

The second map is the change of quantiles (the "horizontal" map in Figure 4), which follows by the fact that the change $(x,z') \to (Tx,z)$ is performed in such a way that $F_{X|Z=z'}(x) = F_{X|Z=z}(Tx)$. This map is of course the monotone rearrangement if it exists. This is

again achieved via the control variable approach by defining $R = F_{X|Z}(X)$ and conditioning on R. This implies that the horizontal map does not affect the distribution of V and hence the function $g(\cdot,V)$ we want to identify. If we keep alternating between these two maps for a given starting value x_0 , this sequence $\lim_{m \to +\infty} T^m(x_0)$, that is, $x_0, T(x_0), T(T(x_0)), T(T(T(x_0))) \dots$ either converges or diverges.

To make it converge, one now assumes the existence of a fixed point x^* , for instance by requiring that the conditional CDFs $F_{X|Z=z}$ and $F_{X|Z=z'}$ are continuous and that they intersect at a point, as depicted in Figure 4 [133]. In this case, the dynamics will converge to x^* where $F_{X|Z=z}$ and $F_{X|Z=z'}$ intersect. For points $x \le x^*$ we need to iterate $z \mapsto z'$ and for points $x \ge x^*$ we need to iterate $z' \mapsto z$. This allows us to identify the function g(x,V) by comparing different points x in this iterative approach to the point x^* , because the "vertical" and "horizontal" maps do not change the distribution of V and hence keep $g(\cdot,V)$ fixed. Now in combination with the standard assumption that g is the monotone rearrangement in V and some other regularity assumptions, this argument allows to identify the true causal mechanism g, but now in the setting $X \not\perp U$.

The mathematically interesting part of this argument is the fixed-point iteration. In particular, it is an interesting question if this argument can be extended to a multivariate first stage and what the dynamics look like in this case. This has been done in [64] in a special case of optimal transport maps, but there are many other open questions. The question then becomes: is it possible to generalize the fixed-point idea from the univariate setting when considering dynamics of Brenier maps $\nabla \varphi_z$? The first key is to generalize the "horizontal" map in Figure 4 to the multivariate setting (the "vertical" map trivially is the same as in the univariate setting).

The main requirement pointed out in [64] is that $h(z,\cdot)$ be a measure-preserving isomorphism for every z, i.e., a map that preserves the measure whose inverse also preserves the measure in the sense that $P_{X|Z=z}(A) = P_W(h^{-1}(z,A))$ for any Borel set A, and where $h^{-1}(z,A)$ denotes the preimage of A. This is because for any measure-preserving isomorphism, it holds (4.3)

$$P_{V|X=x,Z=z} = P_{V|W=h^{-1}(z,x),Z=z} = P_{V|W=h^{-1}(z,x)},$$

where the second equality follows since Z is an instrument. The first equality holds by defining the map ϕ : $(X,Z)\mapsto (h^{-1}(Z,X),Z)$, which is a measure-preserving isomorphism. The same reasoning holds for the map

$$(v, x, z) \mapsto (v, h^{-1}(z, x), z).$$

Thus $P_{V|X,Z}(A_v) = P_{V|W,Z}(A_v)$. This shows that V does not change when applying the map ϕ , which makes the latter the natural extension of the "horizontal map" in the univariate setting. This argument shows that a measure

preserving isomorphism is the required generalization of the "horizontal map".

As a result, this shows that if we make some standard regularity assumptions, such as continuity of $h(z,\cdot)$, then W must be of the same dimension as X. This necessary condition has been well-understood and has also been found in [73]. In particular, this implies that identification in general instrumental variable models of the form (4.1) puts a strong limit on the dimension on W in comparison to X—in short, the dimension of the unobservable in the first stage needs to be restricted, while the second stage can be much more general.

Generalizing the "horizontal" maps from the univariate case to the multivariate setting is challenging. In [64] this is done under some strong assumptions. The first is that both measures $P_{X|Z=z}$ and $P_{X|Z=z'}$ have the same support. The idea is then to work with the multivariate CDFs $F_{X|Z=z}$ and $F_{X|Z=z'}$. Instead of requiring a fixed point for these dynamics, one can allow for a more general fixed set. This set is the set of intersection of the two multivariate CDFs $F_{X|Z=z}$ and $F_{X|Z=z'}$.

This intersection requirement is captured in the following assumption.

ASSUMPTION 4.1. The distributions $F_{X|Z=z}$ and $F_{X|Z=z'}$ are continuously differentiable everywhere and strictly quasi-concave with supports that coincide and are convex. Moreover, one of the following two settings holds:

- (i) \mathcal{X} is bounded above or bounded below.
- (ii) \mathcal{X} is allowed to be unbounded, but F and G intersect in the sense that there exist some quantilevalues $\alpha, \beta \in (0,1)$ such that G(x) > F(x) for all $x \in \mathcal{X}$ with $\alpha \leq G(x) < 1$ and G(x') < F(x') for all $x' \in \mathcal{X}$ with $0 < G(x') \leq \beta$.

The two requirements (i) and (ii) are not exclusive. In general, the intersection is satisfied when one of the measures has thicker tails than the other, but it is satisfied in many more settings. The assumptions that the supports coincide is strong but required to make the dynamics work. It is also required in the univariate setting [44, 133]. The assumption that the CDFs are strictly quasi-concave is also strong. Strict quasi-concavity of the CDFs implies that the corresponding isoquants (or level-sets) are strictly convex. An *isoquant* or *level set* of a CDF F at a point x_0 is the set of all x which have the same value F(x) as $F(x_0)$. While many standard CDFs are quasi-concave—in particular all unimodal distributions—most distributions are not.

The idea is based on the straightforward insight that the optimal transport map, that is, the Brenier map $\nabla \varphi(x)$ solving (2.1) for $c(x,y) = |x-y|^2$ between two strictly quasi-concave cumulative distribution functions F and G

whose measures have the same support, is the metric projection of each point x onto the epigraph of the corresponding isoquant in the other measure. The metric projection T(x) of x onto a closed convex set A maps x onto the point $y \in A$ which is closest to x in the sense that

$$T(x) = y = \underset{s \in A}{\operatorname{arg\,min}} |x - s|^2.$$

While this result seems simple, it does not seem to have been mentioned in the literature before. A proof can be found in [64]. It is based on the fact that the metric projection onto a closed convex set in a Hilbert space is the gradient of a convex function [77, 105].

CLAIM 4.1. Let T be the Brenier map transporting a quasi-concave and continuous CDF G onto another quasi-concave and continuous CDF F with the same support. Then, for each x_0 in the support of both measures, $T(x_0)$ is either the metric projection of x_0 onto the epigraph

$$I_F^{\uparrow}(x_0) = \{x : F(x) \ge F(x_0)\}$$

of the corresponding isoquant

$$I_F(x_0) = \{x : F(x) = F(x_0)\}\$$

of x_0 or the inverse.

Using this result, one can analyze the dynamics and show that the corresponding analogues of the "vertical" and "horizontal" maps in this multivariate setting always converge to some point on the intersection set defined in Assumption 4.1. Since this set of lower-dimension under these assumptions and hence of Lebesgue-measure zero, it shows that under some structural assumptions on the structural functions of the second stage g(x, v), it is identified up to this set of measure zero [64].

The interesting part of this identification result lies in the connection between optimal transport, fixed-point approaches and overall dynamics. In particular, analyzing the "fixed-set dynamics" of different optimal transport maps between two measures $P_{X|Z=z}$ and $P_{X|Z=z'}$ seems to be an open question that is not only interesting for causal inference in instrumental variable models: the fixed-point iteration in the univariate setting is ubiquitous in many areas of mathematics, for instance in the analysis of convergence to equilibria [e.g. 98]. Extensions of this idea to general supports that are potentially not identical and to measures that are not quasi-concave seem fruitful in this regard.

4.3 Bounds on average outcomes in instrumental variable models

The above results and connections between optimal transport and causal inference were all quite straightforward so far in that they were all based on the monotone rearrangement. While the monotone rearrangement will make another appearance later, for now we explore a less straightforward connection between optimal transport and causal inference. It is still situated in the instrumental variable framework, but instead of trying to identify the full causal mechanism g(X,V), we now only care about average structural effects, similar to the control variable approach above. As we have seen, the control variable approach requires strong structural assumptions on the causal mechanism g and the first stage h in order to identify those average structural effects.

In this section, we therefore go the opposite way: instead of making strong structural assumptions on the mechanisms, we ask how tight the bounds on the object of interest will be if we make *no assumptions* on the structure. This mirrors [14], but in the setting of instrumental variables, where there is more structure. Also, as will become clear, this setup is not directly related to the classical optimal transport problem we have considered so far, but more general—it is a version of an optimal transport problem on path spaces. I want to cover it in the hope that this connection will bring mathematically novel insights into not just causal inference, but also optimal transport.

As in the control variables approach, our goal is to identify average structural effects. Before defining them, I want to recall the connection between structural equations, the counterfactual notation [126], and "counterfactual processes" [15, 110]: the structural equation Y=g(X,V) is equivalent to the counterfactual process $Y_x(v)$, and analogously for X=h(Z,W) and $X_z(w)$. This allows us to obtain the bounds by optimizing over the *counterfactual path space*, which is a convenient representation in many settings.

The original idea for this approach was introduced in [15, 16] in the case where all observable variables Y, X, Z are binary (see also the contributions in [95]). These results were generalized to continuous Y and binary X, and Z in [88] and to discrete Y, X, and Z in [127]. These partial identification results are part of a vast literature in econometrics and causal inference, see [e.g. 96, 97]. In the following we work in the general framework of Y, X, and Z having potentially continuous laws [58].

Our goal here is to obtain bounds on average structural effects of the form

$$\Lambda(X \to Y) = \mathbb{E}[\Lambda(g(x, V))]$$

as in the case of control variables above. The seminal idea of [15, 16] is to consider the counterfactual distributions $P_{Y(x)}$ and $P_{X(z)}$ as the laws of corresponding counterfactual processes $Y_x(v)$ and $X_z(w)$ of the first and second stage of the IV model (4.1). Each element $v \in \mathcal{V}$ indexes one path $Y_x(v)$ and each $w \in \mathcal{W}$ indexes one path $X_z(w)$ of the processes, respectively. Mathematically, this is possible if the spaces of paths are small enough, so that they

can be put in a one-to-one relation to the unit interval, for instance the space of all continuous paths [e.g. 21, Theorem 9.2.2]. The average structural effect can then be written as

$$\Lambda(X \to Y) = \mathbb{E}[\Lambda(Y_x)],$$

where the expectation is taken with respect to the joint distribution (P_V, P_W) . In this setting, it is often easier to consider the unobservable U instead of (V, W), but both representations are equivalent.

The idea for obtaining bounds on $\Lambda(X \to Y)$ is the following. As mentioned in the introduction, because of the endogeneity problem $X \not\perp U$, the observable distribution $P_{Y|X=x}$ is not the correct counterfactual law $P_{Y(x)}$. However, if Z is a valid instrument, then the observable conditional distribution $P_{Y,X|Z=z}$ does provide correct information on the causal system. So the key is to back out bounds on $P_{Y(x)}$, which is required for $\Lambda(X \to Y)$, from the observed distribution $P_{Y,X|Z=z}$. One can achieve this via a linear program, which is also the connection to optimal transport—and the reason for why I include it in this review.

The linear program to obtain bounds on $\Lambda(X \to Y)$ is

$$(4.4) \quad \begin{aligned} & \min/\max_{P_U \in \mathcal{P}^*} \quad \mathbb{E}_{P_U}[\Lambda(Y_x)] \\ \text{s.t.} \quad & F_{Y,X|Z=z}(y,x) = P_U(Y_{X_z} \leq y, X_z \leq x), \end{aligned}$$

where \mathcal{P}^* is a set of probability measures on the joint path space of the processes $Y_x(u)$ and $X_z(u)$, and $Y_{X_z(u)}(u)$ is the process based on a realization of the process $X_z(u)$ for given u. In words, the optimization problem tries to find a corresponding measure P_U on path space which maximizes (for an upper bound) or minimizes (for a lower bound) the average structural effect $\Lambda(X \to Y)$ under the restriction that P_U induces counterfactual processes Y_x and X_z whose induced joint law

$$F_{[Y,X]_z} = P_U(Y_{X_z} \le y, X_z \le x)$$

coincides with the *observable* marginal $F_{Y,X|Z=z}$.

The problem (4.4) is a generalized optimal transport problem on path spaces. To see the connection, suppose that Z can only take two values, z and z'. In this case, (4.4) requires to find a joint measure for $([Y,X]_z,[Y,X]_{z'})$ such that the marginals coincide with the observable marginal measures $P_{Y,X|Z=z}$ and $P_{Y,X|Z=z'}$. For more general Z, it becomes the problem on path measures. It is very closely related to—but more general than—problems from statistical physics, in particular large particle dynamics [e.g. 41, 53, 32]. The considered problem is more general, because the objective expression only depends on Y and X, while the marginals are for Y, X, and Z.

So far, only a rather inefficient sampling method has been proposed to solve this problem in practice [58], for a

very general set of processes Y_x, X_z . [86] provide a clever efficient estimator by simplifying the problem: instead of requiring a replication of the entire marginal distribution in the constraint, they essentially replicate moments of the data, turning the problem into a generalized method of moments problem. While this provides a robust and efficient estimator, it does not solve the original causal inference problem. Furthermore, this setup seems like a useful generalization of classical optimal transport problems to measures on path spaces, whose analysis is likely to provide new insights and connections in the mathematical theory of optimal transportation. In the finite setting, i.e., where X and Z are supported on finitely many points, the problem reduces to a simple finite-dimensional linear program, which has been analyzed recently in statistics [89] and econometrics [51].

4.4 Distributionally robust IV methods

Before moving on to other general approaches for identification of causal effects in settings with observable data, I want to mention another area where optimal transportation arguments are paramount, and which has recently gained attention in the literature on causal inference: distributionally robust optimization (DRO) [19, 20, 56].

Here, the object of interest is $\mathbb{E}[l(D,\beta)]$ for some general loss function $l(D,\beta)$, where the expectation is taken with respect to the law of the data D in an i.i.d. setting. β is the parameter one wants to optimize over. The idea of DRO is to make the estimator robust to different heterogeneous environments, as encoded by the law from the data D is drawn. This heterogeneity is captured by defining a region of possible distributions the problem can take around the data-distribution. The utility of DRO comes from the fact that when one chooses a Wasserstein ball for the regions of possible distributions, the primal DRO problem

$$\min_{\beta} \max_{Q \in B_{\rho}(P)} \mathbb{E}_{Q}[l(D,\beta)],$$

where $B_{\rho}(P) = \{Q : W_2(Q, P) \leq \rho\}$ is a ball in the Wasserstein space of radius ρ , admits a dual problem that often takes the form of a standard constrained prediction problem [20, 56].

Recently, a very interesting contribution [114] considered the DRO approach for *linear* instrumental variable models, that is where

$$g(X, V) = X^{\top} \beta_0 + V$$
 and $h(Z, W) = Z^{\top} \gamma + W$.

The optimization problem in this case is

$$\min_{\beta} \max_{Q \in B_{\rho}(\tilde{P}_n)} \mathbb{E}_Q[(Y - X^{\top}\beta_0)^2],$$

where \tilde{P}_n is not the empirical measure for the observations $\{(Y_i, X_i)\}_{i=1}^n$, which are considered i.i.d. draws

from P, but the empirical measure of the observations projected onto the instrument Z, that is,

$$\{(\tilde{Y}_i, \tilde{X}_i)\}_{i=1}^n = \{(\Pi_Z Y_i, \Pi_Z X_i)\}_{i=1}^n,$$

where $\Pi_Z = (Z^\top Z)^{-1} Z^\top$ is the projection onto the columns space of Z. The ingenuity of this approach lies in the dual problem, which takes a regularized regression form:

$$\min_{\beta} \sqrt{\frac{1}{n} \|\Pi_{\mathbf{Z}} \mathbf{Y} - \Pi_{\mathbf{Z}} \mathbf{X} \boldsymbol{\beta}\|^2} + \sqrt{\rho(\|\boldsymbol{\beta}\|^2 + 1)},$$

where the interesting part is the novel regularizer, dubbed the "square root ridge" regularizer [114]. In particular, the authors show that the empirical estimator $\hat{\beta}_{n,\rho}$ of β , is consistent for small enough $\rho > 0$, even if ρ does not vanish as the number of data points increases, amid some other nice properties with respect to the weak instrument problem [24, 130] and invalid instruments. Generalizing this idea to the semi-parametric or fully nonparametric setting to understand the corresponding properties and how it compares to the other approaches in instrumental variable models we have considered so far could be useful.

5. USING TIME VARIATION FOR IDENTIFICATION: DIFFERENCE-IN-DIFFERENCES AND PARALLEL TRENDS

While instruments are a "silver bullet" for the general endogeneity problem introduced above, finding a valid and relevant instrument is the challenge in practice. This is amplified by the fact that instrument validity is difficult to test, which turns out to be an impossibility when the instrument Z is distributed on a continuum [62, 109].

Moreover, in many settings, other information is available on the problem, often a time domain. The simplest setting is the one with two time periods—a pre-treatment period t=0 and a post-treatment period t=1—and two groups of interest—usually a treated group g=1 and an untreated control group g=0. This is the setting in the minimum wage analysis by Card and Krueger [29] for instance.

It is natural to exploit the given structure. The idea is to make the *parallel trends assumption*. This assumption essentially states that the change over time of the outcome in the control group, that is, the group that did not receive the treatment between t=0 and t=1, captures all the unobservable influences that cause the outcome to change over time *except for the influence of the actual treatment*. Under this assumption, one can then extract the effect of receiving treatment between t=0 and t=1 by comparing the change in the outcome over time for the treatment group with the change in the outcome over time for the control group: a difference-in-differences.

This method of difference-in-difference has become one of the main pillars for reduced-form causal inference in applied research, extended to many time periods and several different techniques. The models there are parametric linear, that is, $g(X,V) = X^{\top}\beta + V$ and $h(Z,W) = Z^{\top}\gamma + W$. This allows for simple linear regression methods to identify average treatment effects, as in [29]. For a recent overview of this applied literature, consider [124]. A formal treatment of the semiparametric case, still focusing on average effects, can be found in [1]. The methods focusing on average effects are fundamental, but can miss important details as laid out in the minimum wage example in the introduction.

5.1 Nonlinear difference-in-differences and the changes-in-changes estimator

A nonparametric method to identify general heterogeneous effects for univariate outcomes was introduced in [10]. It is based, of course, on the monotone rearrangement, and considers the entire distribution of outcomes and unobservables. The abstract relations between all measures are depicted in Figure 5, adapted from [134].

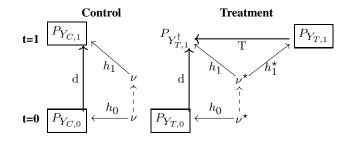


FIG 5. Illustration of various maps in the "nonlinear difference-indifferences" setup. An arrow indicates a pushforward map between two measures; for example $P_{Y_{C,1}} = \mathrm{d}_\# P_{Y_{C,0}}$. The maps h_j are the "production functions" linking the unobservable measures ν and ν^* to the potential outcomes. A dashed arrow indicates a map from a measure to itself. $P_{Y_{T,1}^\dagger}$ is the counterfactual outcome measure of the treated units had they not received treatment. d is the natural trend map and T is the map from an observed outcome to its counterfactual. The observable data is drawn from the four boxed measures.

Each unit i in an empirical setting is sampled from a larger population. Adapting the notation of [10], a random variable G_i denotes a unit's treatment: C for no treatment, T for treatment. Let the random vectors $Y_{i;C,0}$ and $Y_{i;C,1}$ to model unit i's observable potential outcomes in the control group in the pre- and post-intervention periods, respectively; $Y_{i;T,0}$ and $Y_{i;T,1}$ are the unit's observable potential outcomes when $G_i = T$. Each unit has indicator random variables $T_{i,0}$ and $T_{i,1}$ denoting whether an outcome was observed in each study period.

The main assumptions needed for the model are as follows.

- 1. Each potential outcome in the absence of treatment is generated by a deterministic production function $h(t,\cdot)$. That is, $Y_{i;C,0}=h(0,U_i)$, $Y_{i;T,0}=h(0,U_i^*)$, and $Y_{i:C,1}=h(1,U_i)$.
- 2. Moreover, the laws ν and ν^* of the random variables U_i and U_i^* describing the unobservable characteristics of the individual do not change over time within a treatment group. In Figure 5 this is captured by the fact that ν and ν^* —which can be arbitrarily different—stay the same over time.

The basic idea is as in the linear case: the "parallel trends" assumption implies that the natural trend of the control group (the map "d" in Figure 5) is the change in the outcome distribution of the treatment group had it not received treatment. By definition, for the control group, in both time periods, one observes the counterfactual outcome of no treatment (the two boxed measures $P_{Y_{C,0}}$ and $P_{Y_{T,0}}$ in the control group). For the treated group, one observes the counterfactual outcome of no treatment at t=0 and the potential outcome of being treated at t=1 (the two boxed probability measures $P_{Y_{T,0}}$ and $P_{Y_{T,1}}$ in the treatment group in Figure 5).

To isolate the effect of receiving the treatment between t=0 and t=1, we want to "net the change" over time in the observed outcomes in the treatment group by the change in the control group. In order to identify the counterfactual outcome of the treatment group had it not received treatment, we want to transplant the "natural trend" d—that is the change in the outcome in the case where no treatment is administered—to the outcome $P_{Y_{T,0}}$ of the treatment group before treatment.

Of course, there are infinite potential ways to (i) define the natural trend and (ii) translate it to $P_{Y_{T,0}}$. A consistent way in the univariate setting is to work with the order structure, which leads directly into the use of the monotone rearrangement. The additional assumptions needed for this are as follows [10].

- 3. The production functions h(t, U) are the monotone rearrangement between the corresponding unobservables U, U^* and the corresponding potential outcomes.
- 4. Since the idea is to "transplant" the natural drift d, one needs to assume that the support of U^* is contained in the support of U.

Under these assumptions—and if U is either continuously or discretely distributed—the counterfactual CDF $F_{Y_{\tau}^{+}}$ is identified as [10, Theorem 3.1]

$$F_{Y_{T,1}^{\dagger}}(y) = F_{Y_{T,0}}\left(F_{Y_{C,0}}^{-1}\left(F_{Y_{C,1}}(y)\right)\right),$$

and the difference between the observed $F_{Y_{T,1}(y)}$ and the induced $F_{Y_{T,1}^{\dagger}}(y)$ is the *changes-in-changes estimator*.

Note that the monotone rearrangement in this identification result goes in *the opposite direction* as one would expect, that is, from the post-treatment period to the pretreatment period in the control group. The intuition is the same, however: we want to translate the natural trend from the control onto the treatment group to induce the counterfactual of what would have happened to the treatment group without treatment.

The statistical estimator based on this argument was used in [119] to analyze the minimum wage debate as outlined in the introduction. By being able to estimate the entire counterfactual CDF and not just an average, [119] showed that there is a positive employment effect for smaller fast-food restaurants and a negative employment effect for larger ones. This already sheds some more light on the problem, but there are many different dimensions that one should consider. It is therefore useful to extend this approach to allow for multivariate outcomes.

Finally, I want to point out that in the univariate setting, other approaches have been proposed which generalize the "parallel trends" assumption to allow for general heterogeneity. These models in general make stronger or less interpretable assumptions than monotonicity. [27] directly assume that the copulas between $P_{Y_C,0}$ and $P_{Y_T,0}$ as well as $P_{Y_T,0}$ and $P_{Y_T,1}^{\dagger}$ are identical. [123] analyze what happens when pointwise differences between the corresponding cumulative distribution functions are equal: $F_1(x) - F_0(x) = F_1^{\dagger}(x) - F_0^{\star}(x)$ for all $x \in \mathbb{R}$. [22] restrict the heterogeneity of the model to be additively separable and assume that the pointwise differences between the corresponding logarithms of the characteristic functions are equal.

5.2 Extensions via (cyclic) comonotonicity

One would think that an extension of the changes-inchanges idea to multivariate outcomes now follows the straightforward path: replace the monotone rearrangement by Brenier maps. While this is true to an extent, it is not quite so simple, surprisingly. The reason is that in the multivariate setting compositions of Brenier maps are not necesarily Brenier maps; unlike in the univariate setting, where this is true. The multivariate setting has recently been analyzed in detail in [134]

One key insight is that the monotonicity assumptions on the production functions in the univariate setting of [10] allow the latent variable to be entirely abstracted away! This follows from the weaker notion of *comonotonicity*. We say $h_0, h_1 : \mathbb{R}^d \to \mathbb{R}^d$ are *comonotone* [47] if

$$\langle h_0(x) - h_0(y), h_1(x) - h_1(y) \rangle \ge 0$$

for all $x,y \in \mathbb{R}^d$. Note that with an identity production function, h(t,U)=U, comonotonicity reduces to classical monotonicity. In the univariate case comonotonicity implies locally that the derivatives $h'_0 \cdot h'_1 \geq 0$ have the

same sign; in addition to further global constraints. As a concrete example, if h_0 is a polynomial and h_1 its pointwise scaling by some $\gamma>0$, then this pair of functions is comonotone because they have the same sign between all zeros and hence the same signed difference between any pair of points. This example emphasizes that h_0 and h_1 need not be individually monotone themselves for them to be comonotone. Cyclical comonotonicity is hence an important weakening in this setting.

We have now seen that that the monotone production functions assumption in [10] in the univariate setting implies that the natural trend d is cyclically monotone. Furthermore, the production functions are comonotone. To extend the changes-in-changes estimator to higher dimensions, [134] exploit the idea of comonotonicity. They show that the assumption of cyclically comonotone production functions has the desired properties to generalize the changes-in-changes estimator.

DEFINITION 5.1 (Cyclic comonotonicity, [134]). Two production functions h_0 and h_1 are cyclically comonotone if for any positive integer m and any cycle

$$u_1, \ldots, u_m, u_{m+1} = u_1$$

in their common domain, it holds

$$\sum_{i=1}^{m} \langle h_0(u_i), h_1(u_i) - h_1(u_{i+1}) \rangle \ge 0.$$

Just as cyclical monotonicity collapses to monotonicity when m=2, cyclical comonotonicity collapses to comonotonicity in that case. Whenever h_0 has an inverse h_0^{-1} , cyclical comonotonicity implies that the natural trend $d=h_1\circ h_0^{-1}$ is also cyclically monotone [134, Theorem 2]. By the uniqueness of Brenier maps [e.g. 139, Theorem 2.12], d is then the unique Brenier map such that $P_{Y_C,1}=d\#P_{Y_C,0}$. For the identification result, one needs some structural assumptions on the supports, mirroring the univariate setting: the observable measures $P_{Y_C,t}$, $P_{Y_T,t}$, t=0,1, and the counterfactual measure $P_{Y_T,1}^{\dagger}$ are supported on proper convex subsets K_t , K_t^{\star} , and K_1^{\dagger} of \mathbb{R}^d and are absolutely continuous with respect to Lebesgue measure. Moreover, $K_0^{\star} \subset K_0$, which is required to transport the map entirely to the new domain.

THEOREM 5.1 (Multivariate extension of the changes-in-changes estimator, [134]). Consider the causal model depicted in Figure 5. Let the above regularity assumptions hold. Moreover, assume that the production function h_0 has a well-defined inverse and that h_0 and h_1 are cyclically comonotone in the sense of (6). Then there exists a unique map $d: K_0^{\star} \to K_1^{\dagger}$. It is the Brenier map from $P_{Y_C,0}$ to $P_{Y_C,1}$. The counterfactual distribution $P_{Y_T,1}^{\dagger}$ of

the treated unit had it not received treatment is then identified via

$$P_{Y_T,1}^{\dagger} = d \# P_{Y_T,0}.$$

As in the univariate case, when the measures are not absolutely continuous with respect to Lebesgue measure, one can only obtain bounds on the counterfactual measures. Moreover, the optimal transport approach allows us to identify the actual counterfactual random variable, a generalization of [10].

COROLLARY 5.1 ([134]). Consider the setting and assumptions from the previous theorem. If the production function h_1 has a well-defined inverse and h_1^* and h_1 are cyclically comonotone, then there exists a unique map $T: K_1^* \to K_1^{\dagger}$ such that

$$P_{Y_T,1}^{\dagger} = T_{\#} P_{Y_T,1}.$$

T is the Brenier map from $P_{Y_T,1}$ to $P_{Y_T,1}^{\dagger}$. $Y_{T,1}^{\dagger}$ is then identified via

$$Y_{T,1}^{\dagger} = T(Y_{T,1}).$$

To show that the multivariate extension is useful, consider again the minimum wage setting. A multivariate outcome can now for instance also distinguish between fulltime and part-time employees—two groups that a priori should be affected very differently by a minimum wage increase—while accounting for the correlation structure between those sets of employees. [134] reproduce the same result of [119] about the size of the restaurants in question: positive effects on employment for smaller restaurants and negative effects for larger ones. Now, one can also split the problem into full-time and part-time employees. They find a strong (average) positive effect for full-time employees and a strong (average) negative effect for part-time employees, an additional dimension to further understand the effects of increasing the minimum wage.

5.3 Individual heterogeneity vs. the entire system

The above changes-in-changes estimator and its extension are focused on identification of individual heterogeneity. This is obvious in the univariate setting, where the monotone rearrangement implies that quantiles of U are preserved when mapped to the potential outcomes. In the next section, when introducing a generalization of the synthetic control method, I focus on the systems view. There, the level of interest are not the individual quantiles, but the distribution as a whole. This also means that the standard monotone rearrangement will not underlie the identification strategy. Of course, a similar view would be interesting in the difference-in-differences setting. This

can have important applications in many areas of science; a canonical example is single-cell RNA-seq data: measuring the gene expression levels of a cell is a destructive process, and as a result a given cell may be only measured once, fitting into the standard causal inference setting we consider [26].

6. RECREATING TRENDS: SYNTHETIC CONTROLS

The existence of parallel trends is the fundamental assumption that allows identification in difference-in-difference settings. If one has access to more pretreatment time periods than just one, this assumption can be tested by checking if the pre-trends between treatment and control are sufficiently parallel. Very often, the parallel trends assumption is violated and should not be upheld. In such settings, it is still possible to obtain estimates of causal effects if one has access to several units that never receive treatment. The seminal idea, introduced in [3, 5] is to find a convex combination of such control units that can replicate the pre-treatment trend of the target unit as closely as possible.

6.1 Classic synthetic controls

The classic method of synthetic controls is designed for comparative case studies [2, 4], where an aggregate system or unit, such as a state or industry sector, is exposed to a treatment at one point in time $t=t^*$ and stays treated thereafter. This setting is a case study in that there are only few units to consider: in the most extreme case, only one treated unit and a few control units. Moreover, each unit is observed over several time periods $t=T_0,\ldots,T$, with $T_0 < t^* < T$. I call $t \le t^*$ the pre-treatment- or pre-treatment periods and $t > t^*$ the post-intervention- or post-treatment periods.

It is consensus that in such settings, a combination of untreated aggregate units can provide a better control group than a single one [2]. The original method of synthetic controls then provides a canonical approach to obtain a close replication by projecting the outcome of the target (the treated unit) onto the convex hull of the potential control units to generate the synthetic control.

Let $\{Y_{0t}\}_{t\in[T_0,T]}$ be the observed time series of the treated unit, and $\{Y_{jt}\}_{t\in[T_0,T]}$ for $j=1,\ldots,J$ be the observed time series of aggregate units that never receive treatment—the potential controls. The key quantity to estimate is $Y_{0t,N}$, the outcome of the treatment unit had it not received the treatment in the post-intervention periods. Based on this, one defines the effect $\alpha_{jt}=Y_{jt,I}-Y_{jt,N}$ of the intervention for unit j at time t, so that one can write the observable outcome in terms of the counterfactual notation as

$$Y_{jt} = Y_{jt,N} + \alpha_{jt} D_{jt},$$

where $D_{jt} \in \{0,1\}$ is the treatment indicator.

The goal in the suggested setting is to estimate the treatment effect on the treated group in the post-treatment period, i.e.

$$\alpha_{0t} = Y_{0t,I} - Y_{0t,N} = Y_{0t} - Y_{0t,N}$$
 for $t > t^*$,

which requires a model for the unobservable $Y_{0t,N}$. [3] introduce a linear factor model

$$Y_{jt,N} = \delta_t + \theta_t X_j + w_t \mu_j + \varepsilon_{jt},$$

where δ_t is a univariate factor, w_t is a vector containing factors whose loadings are captured in μ_j , and where X_j are observed covariates of the respective units. The error terms ε_{jt} are zero mean transitory shocks. The idea for the synthetic controls method is that $Y_{jt,N} = Y_{jt}$ for $t > t^*$ and $j = 1, \ldots, J$, so that the treatment effect on the treatment group can be obtained by a weighted average

$$\hat{\alpha}_{0t} = Y_{0t} - \sum_{j=1}^{J} \lambda_j^{\star} Y_{jt},$$

where $\{\lambda_i^{\star}\}_{j=1,\dots,J}$ is an optimal set of weights.

The classical synthetic controls estimator in this setting then proceeds in two stages. In the first stage, one obtains the optimal weights $\lambda^\star := \{\lambda_j^\star\}_{j=1,\dots,J}$ which lie in the J-dimensional probability simplex Δ^J and are chosen such that they minimize a weighted Euclidean distance

$$\left(\sum_{k=1}^{K} v_k (X_{k0} - \lambda_1 X_{k1} - \dots - \lambda_J X_{kJ})^2\right)^{1/2},\,$$

where $v \in \Delta^K$ is another set of weights which needs to be chosen by the researcher. [2, 3, 4, 8] provide possible choices for v. In the second stage, the obtained optimal weights $\{\lambda_j^\star\}_{j=1,\dots,J}$ from this minimization are used to create \hat{Y}_{0t}^N in the post-treatment periods as

$$\hat{Y}_{0t}^N = \sum_{j=1}^J \lambda_j^{\star} Y_{jt}, \quad \text{for } t > t^*,$$

based on which one can estimate $\hat{\alpha}_{0t} = Y_{0t} - \hat{Y}_{0t}^{N}$. Since the introduction of the original method, there have been a myriad of extensions and generalizations, see the overview [2].

6.2 Distributional synthetic controls

So far, there is no direct connection between the synthetic controls method and optimal transport. Recently, however, [63] introduced an extension to the synthetic control method in settings where more information about the aggregate unit in question is available. This extension is based on optimal transportation, in particular barycenters in Wasserstein space [6].

The setting is one where the researcher observes j = 0, ..., J aggregate systems/units over time periods t, but

also has access to data Y_{ijt} within the system. A canonical example is a state j, for instance New Jersey, where the researcher observes fast-food restaurants i within the state j at time t. Then one can consider the system "fast-food restaurants within the state of New Jersey". The difference to classic longitudinal or panel data settings is that the level of interest of the causal effect is the state level j, not the individual level i for each restaurant within the state. In particular, this means that the method is applicable in settings where the individuals i within the system j cannot be traced over time. In the above example this can happen when restaurants close or new ones open for instance. Another setting is employees i in a company j and the treatment is administered at the company level. Such a setting has recently been analyzed in [137].

The method is designed for univariate outcomes, but the concept can be generalized to multivariate outcomes [59], as I briefly show below. The quantity of interest is the quantile function $F_{Y_{j_t}}^{-1}(q)$. The goal is to estimate the counterfactual quantile function $F_{Y_{0t,N}}^{-1}(q)$ of the treated unit had it not received treatment by an optimally weighted average of the control quantile functions $F_{Y_{j_t}}^{-1}(q)$ for all $j=1,\ldots,J,\,t>t^*$:

$$F_{Y_{0t,N}}^{-1}(q) = \sum_{j=1}^J \lambda_j^{\star} F_{Y_{jt}}^{-1}(q) \quad \text{for all } q \in (0,1).$$

The weights, as in the classical method, are obtained by trying to replicate the treatment quantile function $F_{Y_{0t}}^{-1}(q)$ as closely as possible by a convex combination of the control quantile functions.

Since the 2-Wasserstein space—the space of all probability measures equipped with the 2-Wasserstein distance—for measures supported on the line is flat [90], this can be done by exploiting the linear structure on the space. The optimization for the replication becomes:

$$\vec{\lambda}_t^* = \arg\min_{\vec{\lambda} \in \Delta^J} \int_0^1 \left| \sum_{j=1}^J \lambda_j F_{Y_{jt}}^{-1}(q) - F_{Y_{0t}}^{-1}(q) \right|^2 dq$$

for all $t < t^*$.

In some settings, for instance, when it is known that the distributions are mixtures, it is useful to work with distribution functions instead of quantiles. In this case, using the 1-Wasserstein distance in CDF form is useful:

$$\vec{\lambda}_t^* = \arg\min_{\vec{\lambda} \in \Delta^J} \int_{\mathbb{R}} \left| \sum_{j=1}^J \lambda_j F_{Y_{jt}}(y) - F_{Y_{0t}}(y) \right| dy.$$

To obtain one set of weights λ^* over all pre-treatment time periods t, one usually forms another weighted average over time; [8] provide useful approaches. In most practical settings the quantile functions are not given directly but have to be estimated from observations $\{Y_{ijt}\}$, $i=1,\ldots,n_j, j=0,\ldots,J, t\in [T_0,T]$.

In addition to the proposed estimator, it is interesting to analyze the causal model in this setting. While the classic method is based on a linear factor model as seen above, the question is how general one can be for a causal model of the form

$$P_{Y_{it,N}} = h_t \# P_{U_{it}}$$
 for $P_{U_{it}} = g_t \# P_{U_{i(t-1)}}$.

Since the synthetic control method replicates across groups, but extrapolates over time—the key identification assumption is that the weights $\vec{\lambda}^*$ obtained in the pre-treatment periods $t < t^*$ stay optimal in the posttreatment periods $t \ge t^*$ in order to identify the correct counterfactual—it turns out that the maps h and g must be isometries between the respective measures [63, Appendix]. In the 2-Wasserstein space on the line the set of such isometries is larger than the standard isometries on \mathbb{R} [90]. The exotic isometries in this setting are difficult to describe, so [63] focuses on linear maps, i.e. $h(t, U_{jt}) = \alpha_t + \beta_t U_{jt}$. The argument shows that linearity is close to necessary and that this also extends to the classic method of [3].

While the univariate case is the most useful in applied settings, the multivariate setting is also interesting, not least since the 2-Wasserstein space over \mathbb{R}^d is non-negatively curved [90]. One hence needs to replace the classic weighted average based on the linear structure from the univariate setting by a metric analogue: the barycenter [6]. A direct approach for obtaining the optimal weights in the univariate setting would be

$$\vec{\lambda}_t^* = \underset{\lambda \in \Lambda^J}{\operatorname{argmin}} W_2^2(P_{Y_{0t}}, P(\lambda))$$

where

$$P(\vec{\lambda}) = \underset{P \in \mathcal{P}_2(\mathbb{R}^d)}{\operatorname{argmin}} \sum_{j=1}^J \frac{\lambda_j}{2} W_2^2(P, P_{Y_{jt}})$$

is the barycenter in the 2-Wasserstein space for the weights $\vec{\lambda} = (\lambda_1, \dots, \lambda_J) \in \Delta^J$. This expression has been used in different settings, for instance in applications in computer vision [23]. Solving this problem directly is challenging, because it is a computationally complex bilevel optimization problem which usually has several local optima. While this is not an issue for vision application, it is often useful to have unique weights in causal inference.

One way to achieve this is via the tangent structure. [59] introduced a notion of tangential projection in the 2-Wasserstein space that is efficient to implement via linear regression. The idea is to lift the problem to the tangent space centered at the target measure $P_{Y_{0t}}$, which linearizes the problem. This approach is related to—but in its implementation significantly more efficient and general in that it can be applied for general measures $P_{Y_{it}}$

that need not be absolutely continuous with respect to Lebesgue measure—than other approach in computer vision and machine learning, in particular the interesting works [141], [103], and [48]. This multivariate setting hence connects computer vision, machine learning, and causal inference through the general challenge of trying to find efficient methods of projections in nonlinear spaces. Of course, many statistically interesting questions are still open, while some, like large sample distributions of the distributional synthetic controls estimator in the univariate setting are currently been analyzed [137, 144]. Overall, exploring further connections between optimal transport and synthetic controls, also with respect to more general objects than just distributions—using the Gromov-Wasserstein distance [102]—might be fruitful.

7. BUT WHAT ABOUT MATCHING?

Of course, on the outset, the closest connection between optimal transport and causal inference is via matching [131]. In the most basic setup, we have two treatment groups $T \in \{0,1\}$, the corresponding potential outcomes (Y_0,Y_1) , and a (potentially high-dimensional) set of observable covariates $X \in \mathbb{R}^d$. The additional information supplied by the covariates in each group can be used in many different ways [131]. One is *balancing* [17], which essentially is a pre-processing method for downstream estimation of causal effects. In this step, one selects a sample where the treatment and control samples are more similar than in the original sample [82]. Another main way in which the additional information of the covariates is used is by matching on them.

Optimal transport is potentially useful for both. Of course, optimal transportation is predestined to be used as a pre-processing step to balance covariates. Many related methods have been influential in this area [e.g. 66, 78], and classic optimal transportation is beginning to be used in related methods [e.g. 46, 143]. However, classic optimal transport is *not* well-suited to be used in the second problem, that is, as a direct estimator for obtaining causal effects via matching. The issue here is the mass-preserving constraint of classic optimal transport, which requires every individual in treatment- and control group to be matched, as I now argue.

The idea of using matching for obtaining estimates of causal effects is based on the *unconfoundedness assumption*. Formally, it reads $(Y_0, Y_1) \perp T | X$. Intuitively, it implies that the researcher is able to observe all important covariates that can influence the potential outcomes and treatment reception, meaning that all unobserved variates U are not correlated with T. To identify the correct causal effects, one hence wants to compare two individuals with the same observable covariates, one in the treatment group and one in the control group. Averaging

over the differences in outcomes over all of these matched pairs should identify the average treatment effect.

Due to random sampling, perfect matches between the groups are rare in practice. The question is then how to find good matches. There is a vast literature on matching approaches [131], including propensity score matching [11, 122], and other methods of direct matching [e.g. 84, 106, 120, 121, and references therein]. The most widely used idea is to specify a distance, e.g. a weighted Euclidean distance or some discrepancy measure like the Kullback-Leibler divergence, and proceed via some form of k-nearest neighbor matching [125, 131]. In the most basic form, every individual in the treatment group gets matched with an individual in the control group. This will lead to biased estimates if the overlap of the supports of the covariate distributions of the two groups is not perfect in the population. In this case, individuals not in the intersection of the two supports do not have a good match in the other group and should be matched for an unbiased estimate of the treatment effect.

While the assumption of perfect overlap in the population is standard in this literature, it is often violated. This implies that matching via classic optimal transportation is not the right approach: the measure-preserving constraint in the Monge-Kantorovich problem implies that all individuals in both groups have to be matched. Especially in finite samples this introduces excessive bias into the estimator.

The solution to this problem is to use unbalanced optimal transportation [37, 93, 129], which relaxes the measure-preservation constraint, hence allowing for individuals to remain unmatched. The unbalanced optimal transport problem is

(7.1)
$$\inf_{\gamma \in \mathcal{M}^{+}(\mathcal{X} \times \mathcal{Y})} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\gamma(x, y) + \varepsilon K L(\gamma || P_{X_0} \otimes P_{X_1}) + \rho D_{\phi} (\pi_0 \gamma || P_{X_0}) + \rho D_{\phi} (\pi_1 \gamma || P_{X_1}),$$

where $KL(\gamma||P_{X_0}\otimes P_{X_1})$ is the Kullback-Leibler divergence [92, 91] between the optimal Kantorovich coupling γ and the independence coupling of the two marginals P_{X_0} and P_{X_1} , $\pi_j\gamma$ denotes the projection onto the j-th marginal of the coupling γ , D_φ denotes the ϕ -divergence (or Csiszàr-divergence) [38, 39], and $\mathcal{M}^+(\mathcal{X})$ denotes the set of all non-negative finite measures on $\mathbb{R}^d\times\mathbb{R}^d$ whose marginals are P_{X_0} and P_{X_1} . The first penalty term using the Kullback-Leibler divergence is only used to improve computational properties, analogous to the classic Sinkhorn regularization [40, 55], and can be dropped by setting $\varepsilon=0$.

The main difference to the classical optimal transport problem are (i) that the optimal coupling γ does not need to be a probability measure and (ii) the addition of the ϕ -divergence terms, which allow for the creation and destruction of mass and do not enforce the constraint from

classical optimal transport that the corresponding measures P_{X_0} and P_{X_1} need to have the same mass. In practice, this means that the optimal couplings obtained will provide partial optimal matches [52, 87] where individuals that do not have a close enough match are automatically discarded.

This automatic and disciplined way of optimal transportation to match two groups based on the relative distance should provide finite sample improvements in mean-squared error over existing methods, particularly when the overlap condition is violated. Recently, [61] analyzed the statistical properties of unbalanced optimal transport problems as the covariate measures P_{X_0} and P_{X_1} are approximated by empirical measures, showing that for fixed penalties $\varepsilon, \rho > 0$, as the number of observations increase, the limit element of the empirical process is Gaussian. Moreover, they show that as $\rho \to 0$ after $\varepsilon \to 0$ in the population problem (7.1) for regular Csiszàr divergences, the optimal coupling γ will only put weight on perfect matches from both groups, i.e., matches where the covariates coincide perfectly. This implies that, at least in the population, matching via unbalanced optimal transport can automatically provide unbiased estimates of the average treatment effect on the intersection of the supports of the two covariate distributions as the balancing penalty ρ vanishes.

This is backed by simulations in [61], which show that an unbalanced approach can beat existing benchmarks in such settings. To fully introduce unbalanced optimal transportation into the toolbox of applied causal inference researchers, *proof* of its superiority over other methods is needed. One way to do it is to analyze the statistical properties of (7.1) in the setting where the penalty terms ε and ρ are *data-dependent* and vanish at specific data-dependent rates. A conjecture is that by choosing appropriate divergence terms as penalties and optimal rates of convergence for ρ_n , it will be possible to show that unbalanced optimal transportation is able to beat essentially all other existing methods in terms of mean-squared error when it comes to estimating treatment effects in this setting.

8. CONCLUSION

This review introduced a selective overview of the uses of optimal transport theory in classic causal inference for observational data. The goal is to unify nomenclature and notation and to introduce new and potentially fruitful avenues for future research in this area. There are potentially many other uses of optimal transport in causal inference. However, the reviewed areas are particularly close since optimal transportation has built the foundation for most of the existing results in this area. It is my hope that this review simplifies the exchange between the two areas of research, and uncovers the connections and potential gains for both fields, on which future collaborations can be built.

REFERENCES

- [1] ABADIE, A. (2005). Semiparametric difference-in-differences estimators. *The review of economic studies* **72** 1–19.
- [2] ABADIE, A. (2021). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of economic literature* 59 391–425.
- [3] ABADIE, A., DIAMOND, A. and HAINMUELLER, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American statistical Association* **105** 493–505.
- [4] ABADIE, A., DIAMOND, A. and HAINMUELLER, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science* 59 495–510.
- [5] ABADIE, A. and GARDEAZABAL, J. (2003). The economic costs of conflict: A case study of the Basque Country. *Ameri*can economic review 93 113–132.
- [6] AGUEH, M. and CARLIER, G. (2011). Barycenters in the Wasserstein space. SIAM Journal on Mathematical Analysis 43 904–924.
- [7] ANGRIST, J. D. and PISCHKE, J.-S. (2009). Mostly harmless econometrics: An empiricist's companion. Princeton university press.
- [8] ARKHANGELSKY, D., ATHEY, S., HIRSHBERG, D. A., IMBENS, G. W. and WAGER, S. (2021). Synthetic difference-in-differences. *American Economic Review* 111 4088–4118.
- [9] ARONOW, P. M., GREEN, D. P. and LEE, D. K. (2014). Sharp bounds on the variance in randomized experiments. *The Annals* of Statistics 850–871.
- [10] ATHEY, S. and IMBENS, G. W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica* 74 431–497.
- [11] AUSTIN, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research* 46 399–424.
- [12] BACKHOFF-VERAGUAS, J. and PAMMER, G. (2022). Applications of weak transport theory. *Bernoulli* 28 370–394.
- [13] BACKHOFF-VERAGUAS, J. and PAMMER, G. (2022). Stability of martingale optimal transport and weak optimal transport. *The Annals of Applied Probability* 32 721–752.
- [14] BALAKRISHNAN, S., KENNEDY, E. and WASSERMAN, L. (2023). Conservative inference for counterfactuals. *arXiv* preprint arXiv:2310.12757.
- [15] BALKE, A. and PEARL, J. (1994). Counterfactual probabilities: Computational methods, bounds and applications. In *Un*certainty in artificial intelligence 46–54. Elsevier.
- [16] BALKE, A. and PEARL, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American statistical Association* **92** 1171–1176.
- [17] BEN-MICHAEL, E., FELLER, A., HIRSHBERG, D. A. and ZU-BIZARRETA, J. R. (2021). The balancing act in causal inference. *arXiv preprint arXiv:2110.14831*.
- [18] BENAMOU, J.-D. and BRENIER, Y. (2000). A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik* 84 375–393.
- [19] BLANCHET, J., KANG, Y. and MURTHY, K. (2019). Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability* 56 830–857.
- [20] BLANCHET, J. and MURTHY, K. (2019). Quantifying distributional model risk via optimal transport. *Mathematics of Opera*tions Research 44 565–600.
- [21] BOGACHEV, V. I. (2007). Measure theory 2. Springer.
- [22] BONHOMME, S. and SAUDER, U. (2011). Recovering distributions in difference-in-differences models: A comparison of

- selective and comprehensive schooling. Review of Economics and Statistics 93 479-494.
- [23] BONNEEL, N., PEYRÉ, G. and CUTURI, M. (2016). Wasserstein barycentric coordinates: histogram regression using optimal transport. *ACM Trans. Graph.* **35** 71–1.
- [24] BOUND, J., JAEGER, D. A. and BAKER, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American statistical association* 90 443–450.
- [25] BRENIER, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics* 44 375–417.
- [26] BUNNE, C., SCHIEBINGER, G., KRAUSE, A., REGEV, A. and CUTURI, M. (2024). Optimal transport for single-cell and spatial omics. *Nature Reviews Methods Primers* 4 58.
- [27] CALLAWAY, B. and LI, T. (2019). Quantile treatment effects in difference in differences models with panel data. *Quantitative Economics* 10 1579–1618.
- [28] CAMBANIS, S., SIMONS, G. and STOUT, W. (1976). Inequalities for Ek(x,y) when the marginals are fixed. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete **36** 285–294.
- [29] CARD, D. and KRUEGER, A. B. (1994). Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania. *The American Economic Review* 84 772–793.
- [30] CARLIER, G., CHERNOZHUKOV, V. and GALICHON, A. (2016). Vector Quantile Regression: An Optimal Transport Approach. *Annals of Statistics* 44 1165–1192.
- [31] CARLIER, G. and EKELAND, I. (2010). Matching for teams. *Economic theory* **42** 397–418.
- [32] CATTIAUX, P. and LEONARD, C. (1995). Large deviations and Nelson processes. In *Forum Math* 7 95–115.
- [33] CHERIDITO, P. and ECKSTEIN, S. (2023). Optimal transport and Wasserstein distances for causal models. *arXiv preprint* arXiv:2303.14085.
- [34] CHERNOZHUKOV, V., GALICHON, A., HALLIN, M. and HENRY, M. (2017). Monge-kantorovich depth, quantiles, ranks and signs. *Annals of Statistics* **45** 223–256.
- [35] CHERNOZHUKOV, V., GALICHON, A., HENRY, M. and PASS, B. (2021). Identification of hedonic equilibrium and non-separable simultaneous equations. *Journal of Political Economy* **129** 842–870.
- [36] CHIAPPORI, P.-A., MCCANN, R. J. and PASS, B. (2017). Multi-to One-Dimensional Optimal Transport. Communications on Pure and Applied Mathematics 70 2405–2444.
- [37] CHIZAT, L., PEYRÉ, G., SCHMITZER, B. and VIALARD, F.-X. (2018). Unbalanced optimal transport: Dynamic and Kantorovich formulations. *Journal of Functional Analysis* 274 3090–3123.
- [38] CSISZÁR, I. (1967). Information-type measures of difference of probability distributions and indirect observation. *Studia Sci*entiarum Mathematicarum Hungarica 2 229–318.
- [39] CSISZÁR, I. (1975). I-divergence geometry of probability distributions and minimization problems. *The annals of probability* 146–158.
- [40] CUTURI, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems 26.
- [41] DAWSON, D. A. and GÄRTNER, J. (1987). Large deviations from the McKean-Vlasov limit for weakly interacting diffusions. Stochastics: An International Journal of Probability and Stochastic Processes 20 247–308.

- [42] DE LARA, L., GONZÁLEZ-SANZ, A., ASHER, N., RISSER, L. and LOUBES, J.-M. (2024). Transport-based counterfactual models. *Journal of Machine Learning Research* 25 1–59.
- [43] DEL BARRIO, E., SANZ, A. G. and HALLIN, M. (2024). Non-parametric multiple-output center-outward quantile regression. *Journal of the American Statistical Association* 1–15.
- [44] D'HAULTFŒUILLE, X. and FÉVRIER, P. (2015). Identification of nonseparable triangular models with discrete instruments. *Econometrica* **83** 1199–1210.
- [45] DOKSUM, K. (1974). Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *The annals of statistics* 267–277.
- [46] DONG, M., WANG, B., WEI, J., DE O. FONSECA, A. H., PERRY, C. J., FREY, A., OUERGHI, F., FOXMAN, E. F., ISHIZUKA, J. J. and DHODAPKAR, R. M. (2023). Causal identification of single-cell experimental perturbation effects with CINEMA-OT. *Nature methods* **20** 1769–1779.
- [47] EKELAND, I., GALICHON, A. and HENRY, M. (2012). Comonotonic measures of multivariate risks. *Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics* **22** 109–132.
- [48] FAN, J. and ALVAREZ-MELIS, D. (2023). Generating synthetic datasets by interpolating along generalized geodesics. In *Uncertainty in Artificial Intelligence* 571–581. PMLR.
- [49] FAN, Y., HENRY, M., PASS, B. and RIVERO, J. A. (2022). Lorenz map, inequality ordering and curves based on multidimensional rearrangements. arXiv preprint arXiv:2203.09000.
- [50] FAN, Y. and PARK, S. S. (2010). Sharp bounds on the distribution of treatment effects and their statistical inference. *Econometric Theory* 26 931–951.
- [51] FANG, Z., SANTOS, A., SHAIKH, A. M. and TORGOVIT-SKY, A. (2023). Inference for Large-Scale Linear Systems With Known Coefficients. *Econometrica* **91** 299–327.
- [52] FIGALLI, A. (2010). The optimal partial transport problem. *Archive for rational mechanics and analysis* **195** 533–560.
- [53] FÖLLMER, H. (1988). Random fields and diffusion processes. Lect. Notes Math 1362 101–204.
- [54] GALICHON, A. (2018). Optimal transport methods in economics. Princeton University Press.
- [55] GALICHON, A. and SALANIÉ, B. (2010). Matching with trade-offs: Revealed preferences over competing characteristics. CEPR Discussion Paper No. DP7858.
- [56] GAO, R. and KLEYWEGT, A. (2023). Distributionally robust stochastic optimization with Wasserstein distance. *Mathematics* of Operations Research 48 603–655.
- [57] GOZLAN, N., ROBERTO, C., SAMSON, P.-M. and TETALI, P. (2017). Kantorovich duality for general transport costs and applications. *Journal of Functional Analysis* **273** 3327–3405.
- [58] GUNSILIUS, F. (2019). A path-sampling method to partially identify causal effects in instrumental variable models. arXiv preprint arXiv:1910.09502.
- [59] GUNSILIUS, F., HSIEH, M. H. and LEE, M. J. (2024). Tangential wasserstein projections. *Journal of Machine Learning Research* **25** 1–41.
- [60] GUNSILIUS, F. and SCHENNACH, S. (2023). Independent nonlinear component analysis. *Journal of the American Statistical* Association 118 1305–1318.
- [61] GUNSILIUS, F. and XU, Y. (2021). Matching for causal effects via multimarginal unbalanced optimal transport. arXiv preprint arXiv:2112.04398.
- [62] GUNSILIUS, F. F. (2021). Nontestability of instrument validity under continuous treatments. *Biometrika* 108 989–995.
- [63] GUNSILIUS, F. F. (2023). Distributional synthetic controls. Econometrica 91 1105–1117.

- [64] GUNSILIUS, F. F. (2023). A condition for the identification of multivariate models with binary instruments. *Journal of Econometrics* 235 220–238.
- [65] HAAVELMO, T. (1943). The Statistical Implications of a System of Simultaneous Equations. *Econometrica* 11 1–12.
- [66] HAINMUELLER, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political analysis* 20 25–46.
- [67] HECKMAN, J. (1990). Varieties of selection bias. The American Economic Review 80 313–318.
- [68] HECKMAN, J. J. (2001). Micro data, heterogeneity, and the evaluation of public policy: Nobel lecture. *Journal of political Economy* 109 673–748.
- [69] HECKMAN, J. J., SMITH, J. and CLEMENTS, N. (1997). Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. The Review of Economic Studies 64 487–535.
- [70] HECKMAN, J. J. and SMITH, J. A. (1995). Assessing the case for social experiments. *Journal of economic perspectives* 9 85– 110.
- [71] HECKMAN, J. J. and SMITH, J. A. (1999). The pre-programme earnings dip and the determinants of participation in a social programme. Implications for simple programme evaluation strategies. *The Economic Journal* 109 313–348.
- [72] HECKMAN, J. J. and VYTLACIL, E. (2005). Structural equations, treatment effects, and econometric policy evaluation 1. Econometrica 73 669–738.
- [73] HODERLEIN, S., HOLZMANN, H., KASY, M. and MEISTER, A. (2017). Corrigendum: Instrumental variables with unrestricted heterogeneity and continuous treatment. *The Review of Economic Studies* 84 964–968.
- [74] HODERLEIN, S. and MAMMEN, E. (2007). Identification of marginal effects in nonseparable models without monotonicity. *Econometrica* 75 1513–1518.
- [75] HOLLAND, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association* 81 945–960.
- [76] HOLLAND, P. W. (1988). Causal inference, path analysis and recursive structural equations models. *ETS Research Report Series* **1988** i–50.
- [77] HOLMES, R. B. (1973). Smoothness of certain metric projections on Hilbert space. *Transactions of the American Mathematical Society* **184** 87–100.
- [78] IMAI, K. and RATKOVIC, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **76** 243–263.
- [79] IMBENS, G. W. (2007). Nonadditive models with endogenous regressors. Econometric Society Monographs 43 17.
- [80] IMBENS, G. W. (2020). Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics. *Journal of Economic Literature* 58 1129–79.
- [81] IMBENS, G. W. and NEWEY, W. K. (2009). Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica* 77 1481–1512.
- [82] IMBENS, G. W. and RUBIN, D. B. (2015). Causal inference in statistics, social, and biomedical sciences. Cambridge university press.
- [83] JORDAN, R., KINDERLEHRER, D. and OTTO, F. (1998). The variational formulation of the Fokker–Planck equation. SIAM journal on mathematical analysis 29 1–17.
- [84] KALLUS, N. (2020). Generalized optimal matching methods for causal inference. *Journal of Machine Learning Research* 21 1–54.

- [85] KANTOROVICH, L. (1942). On the translocation of masses, CR Dokl. Acad. Sci. URSS 37 191–201.
- [86] KILBERTUS, N., KUSNER, M. J. and SILVA, R. (2020). A class of algorithms for general instrumental variable models. Advances in Neural Information Processing Systems 33 20108– 20119.
- [87] KITAGAWA, J. and PASS, B. (2015). The multi-marginal optimal partial transport problem. In *Forum of Mathematics, Sigma* 3. Cambridge University Press.
- [88] KITAGAWA, T. (2021). The identification region of the potential outcome distributions under instrument independence. *Journal* of Econometrics 225 231–253.
- [89] KLATT, M., MUNK, A. and ZEMEL, Y. (2022). Limit laws for empirical optimal solutions in random linear programs. *Annals* of Operations Research 315 251–278.
- [90] KLOECKNER, B. (2010). A geometric study of Wasserstein spaces: Euclidean spaces. Annali della Scuola Normale Superiore di Pisa-Classe di Scienze 9 297–323.
- [91] KULLBACK, S. (1959). Information theory and statistics. Wiley.
- [92] KULLBACK, S. and LEIBLER, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics* 22 79– 86.
- [93] LIERO, M., MIELKE, A. and SAVARÉ, G. (2018). Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Inventiones mathematicae* 211 969–1117.
- [94] LIN, Z., KONG, D. and WANG, L. (2023). Causal inference on distribution functions. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **85** 378–398.
- [95] MANSKI, C. F. (1990). Nonparametric bounds on treatment effects. *The American Economic Review* 80 319–323.
- [96] MANSKI, C. F. (1999). *Identification problems in the social sciences*. Harvard University Press.
- [97] MANSKI, C. F. (2003). Partial identification of probability distributions. Springer Science & Business Media.
- [98] MAS-COLELL, A., WHINSTON, M. D. and GREEN, J. R. (1995). Microeconomic Theory. Oxford University Press.
- [99] MATZKIN, R. L. (2003). Nonparametric estimation of nonadditive random functions. *Econometrica* 71 1339–1375.
- [100] MCCANN, R. J. (1999). Exact solutions to the transportation problem on the line. Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences 455 1341–1380.
- [101] MCCANN, R. J. and PASS, B. (2020). Optimal transportation between unequal dimensions. Archive for Rational Mechanics and Analysis 238 1475–1520.
- [102] MÉMOLI, F. (2014). The Gromov–Wasserstein distance: A brief overview. Axioms 3 335–341.
- [103] MÉRIGOT, Q., DELALANDE, A. and CHAZAL, F. (2020). Quantitative stability of optimal transport maps and linearization of the 2-Wasserstein space. In *International Conference on Artificial Intelligence and Statistics* 3186–3196. PMLR.
- [104] MONGE, G. (1781). Mémoire sur la théorie des déblais et des remblais. Mem. Math. Phys. Acad. Royale Sci. 666–704.
- [105] MOREAU, J.-J. (1965). Proximité et dualité dans un espace hilbertien. Bulletin de la Société mathématique de France 93 273–299.
- [106] MORUCCI, M., ORLANDI, V., ROY, S., RUDIN, C. and VOL-FOVSKY, A. (2020). Adaptive Hyper-box Matching for Interpretable Individualized Treatment Effect Estimation. In Conference on Uncertainty in Artificial Intelligence 1089–1098. PMLR.

- [107] NEUMARK, D. and WASCHER, W. (2000). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania: Comment. *American Economic Review* 90 1362–1396.
- [108] NEYMAN, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Statistical Science 465–472.
- [109] PEARL, J. (1995). On the testability of causal models with latent and instrumental variables. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence* 435–443.
- [110] PEARL, J. (2009). Causality. Cambridge university press.
- [111] PEARL, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys* **3** 96 146.
- [112] PETERS, J., JANZING, D. and SCHÖLKOPF, B. (2017). *Elements of causal inference: foundations and learning algorithms*. The MIT Press.
- [113] PEYRÉ, G., CUTURI, M. et al. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning* **11** 355–607.
- [114] QU, Z. and KWON, Y. (2024). Distributionally Robust Instrumental Variables Estimation. *arXiv preprint arXiv:2410.15634*.
- [115] RACHEV, S. T. and RÜSCHENDORF, L. (2006). Mass Transportation Problems: Volume 1: Theory. Springer Science & Business Media.
- [116] ROBINS, J. (1992). Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika* 79 321–334.
- [117] ROBINS, J. M. (1989). The analysis of randomized and nonrandomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. *Health service research* methodology: a focus on AIDS 113–159.
- [118] ROCKAFELLAR, R. T. (1970). Convex analysis. Princeton Mathematical Series. Princeton University Press.
- [119] ROPPONEN, O. (2011). Reconciling the evidence of Card and Krueger (1994) and Neumark and Wascher (2000). *Journal of Applied Econometrics* 26 1051–1057.
- [120] ROSENBAUM, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association* 84 1024–1032.
- [121] ROSENBAUM, P. R. (2020). Modern algorithms for matching in observational studies. *Annual Review of Statistics and Its Appli*cation 7 143–176.
- [122] ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70 41–55.
- [123] ROTH, J. and SANT'ANNA, P. H. (2023). When is parallel trends sensitive to functional form? *Econometrica* **91** 737–747.
- [124] ROTH, J., SANT'ANNA, P. H., BILINSKI, A. and POE, J. (2023). What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics* 235 2218–2244.
- [125] RUBIN, D. B. (1973). Matching to remove bias in observational studies. *Biometrics* 29 159–183.
- [126] RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* **66** 688.
- [127] RUSSELL, T. M. (2021). Sharp bounds on functionals of the joint distribution in the analysis of treatment effects. *Journal of Business & Economic Statistics* 39 532–546.
- [128] SANTAMBROGIO, F. (2015). Optimal transport for applied mathematicians. *Birkäuser*, NY 55 94.
- [129] SÉJOURNÉ, T., PEYRÉ, G. and VIALARD, F.-X. (2023). Unbalanced optimal transport, from theory to numerics. *Handbook* of *Numerical Analysis* 24 407–471.

- [130] STOCK, J. H., WRIGHT, J. H. and YOGO, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics* 20 518–529.
- [131] STUART, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics* **25** 1.
- [132] TINBERGEN, J. (1930). Determination and interpretation of supply curves: an example. Zeitschrift für Nationalökonomie 1 669–679.
- [133] TORGOVITSKY, A. (2015). Identification of nonseparable models using instruments with small support. *Econometrica* 83 1185–1197.
- [134] TOROUS, W., GUNSILIUS, F. and RIGOLLET, P. (2024). An optimal transport approach to estimating causal effects via nonlinear difference-in-differences. *Journal of Causal Inference* 12 20230004.
- [135] TU, R., ZHANG, K., KJELLSTRÖM, H. and ZHANG, C. (2022). Optimal transport for causal discovery. In *ICLR 2022-The Tenth International Conference on Learning Representa*tions (Virtual), Apr 25th-29th, 2022. International Conference on Learning Representations, ICLR.
- [136] VAN DER LAAN, M. J. and ROBINS, J. M. (2003). Unified methods for censored longitudinal data and causality. Springer.
- [137] VAN DIJCKE, D., GUNSILIUS, F. and WRIGHT, A. (2024). Return to office and the tenure distribution. *arXiv* preprint *arXiv*:2405.04352.
- [138] VANSTEELANDT, S. and JOFFE, M. (2014). Structural Nested Models and G-estimation: The Partially Realized Promise. *Statistical Science* **29** 707 731.
- [139] VILLANI, C. (2009). *Optimal transport: old and new* **338**. Springer Science & Business Media.
- [140] VILLANI, C. (2021). Topics in optimal transportation 58. American Mathematical Society.
- [141] WERENSKI, M., JIANG, R., TASISSA, A., AERON, S. and MURPHY, J. M. (2022). Measure estimation in the barycentric coding model. In *International Conference on Machine Learn*ing 23781–23803. PMLR.
- [142] WRIGHT, P. G. (1928). The tariff on animal and vegetable oils **26**. Macmillan.
- [143] YAN, Y., ZHOU, H., YANG, Z., CHEN, W., CAI, R. and HAO, Z. (2024). Reducing balancing error for causal inference via optimal transport. In *Proceedings of the 41st International Conference on Machine Learning* 55913–55927.
- [144] ZHANG, L., ZHANG, X. and ZHANG, X. (2024). Asymptotic Properties of the Distributional Synthetic Controls. arXiv preprint arXiv:2405.00953.