From Statistical Transportability to Estimating the Effect of Stochastic Interventions

Juan D. Correa and **Elias Bareinboim**
Department of Computer Science, Purdue University, IN, USA
{correagr, eb}@purdue.edu

Abstract

Learning systems often face a critical challenge when applied to settings that differ from those under which they were initially trained. In particular, the assumption that both the source/training and the target/deployment domains follow the same causal mechanisms and observed distributions is commonly violated. This implies that the robustness and convergence guarantees usually expected from these methods are no longer attainable. In this paper, we study these violations through causal lens using the formalism of *statistical transportability* [Pearl and Bareinboim, 2011] (PB, for short). We start by proving sufficient and necessary graphical conditions under which a probability distribution observed in the source domain can be extrapolated to the target one, where strictly less data is available. We develop the first sound and complete procedure for statistical transportability, which formally closes the problem introduced by PB. Further, we tackle the general challenge of identification of stochastic interventions from observational data [Sec. 4.4, Pearl, 2000]. This problem has been solved in the context of atomic interventions using Pearl's do-calculus, which lacks complete treatment in the stochastic case. We prove completeness of stochastic identification by constructing a reduction of any instance of this problem to an instance of statistical transportability, closing the problem.

1 Introduction

Generalizing causal and statistical findings across settings is central in scientific inferences as well as in many applications throughout artificial intelligence and machine learning. The environment where the data is collected (source) is related to, but almost never the same as, the one where the inferences are intended (target). If the target environment is arbitrary, or drastically different from the training (source) environment, no learning could take place. However, the fact that we learn and perform relatively well in a new environment suggest that certain environments share common characteristics and that, owing to these commonalities, causal and statistical claims would be valid and robust even in settings where no or very little data is available [Pearl, 2000; Spirtes *et al.*, 2001; Bareinboim and Pearl, 2016; Pearl and Mackenzie, 2018].

Remarkably, the anchors of knowledge that allow extrapolations to take place are eminently causal, following from the stability of the mechanisms shared across settings [Aldrich, 1989]. The systematic analysis of these mechanisms and the conditions under which extrapolations could be formally justified has been studied in the literature under the rubric of *transportability theory* [Pearl and Bareinboim, 2011]. A number of results showed the robustness and efficiency of transportability under a wide range of conditions [Bareinboim and Pearl, 2012; Lee and Honavar, 2013a; Lee and Honavar, 2013b; Bareinboim and Pearl, 2013; Bareinboim and Pearl, 2014]; for a survey, refer to [Bareinboim and Pearl, 2016].

Despite all the progress achieved so far, most of these results were focused on the conditions under which *causal distributions* could be extrapolated, leaving a general class of extrapolation problems without solution, namely, non-causal distributions. In practice, on the other hand, many problems in AI and ML today, including classification and clustering, entail the learning of a (non-causal) probability distributions of the form $P(\mathbf{y}|\mathbf{x})$. In these settings, it's also the case that the training environment does not always match the one where the classifier, for example, is intended to be deployed. Depending on the differences between environments, the distribution $P(\mathbf{y}|\mathbf{x})$ may not be a good predictor in the target domain. The mismatches between the source and target environments are due to the differences in the underlying causal mechanisms and data-collection method.

Through the lens of causal reasoning, this setting has been called *statistical transportability* in [Pearl and Bareinboim, 2011], *dataset shift*, or *domain adaptation* in [Quiñonero-Candela *et al.*, 2009; Zhang *et al.*, 2013; Zhang *et al.*, 2015; Magliacane *et al.*, 2018; Rojas-Carulla *et al.*, 2018]. The statistical transportability problem has been formalized but has not been solved systematically nor in its non-parametric version. For concreteness, consider the following example.

Example 1. An internet company stores records on the type of advertisement displayed to its customers (X) and whether the corresponding product was sold (Y) in its website Π. Each user's age (W) is also measured, which affects both the ad format (X) as well as her/his propensity for buying the product (Y). The distribution $P(x, w, y)$ can be estimated from this dataset. The company plans to expand to a different country Π^* and call the data science team to help to predict Y given X, $P^*(y|x)$, in the new market. In transportability theory, the differences in the causal mechanisms across settings are encoded through square nodes (□), which will be formally defined later on. The causal diagrams shown in Fig. 1 entail different statistical patterns in the data when comparing against the source, when

C1 the age distribution ($P^*(W)$) is significantly different,
C2 the strategy to select the ad format (X) is different, and
C3 the buying behavior (Y) differs, for instance, since users are less wealthy in population Π^*.

For case (C1), the causal analyst in the team suggests that they should use data from the Census in population Π^*, and estimate the new age distribution, $P^*(w)$. The team goes on to say that this will allow the target query $P^*(y|x)$ to be written in a convenient form, namely,

$$P^*(y|x) = \frac{\sum_w P(y|x,w)P(x|w)P^*(w)}{\sum_w P(x|w)P^*(w)}, \qquad (1)$$

The r.h.s. of the expression is estimable by combining data from the source ($P(\mathbf{v})$) and the smaller dataset from the target ($P^*(w)$). For the second case, (C2), the new strategy to select ads (i.e., $P^*(x|w)$) is needed, the team suggests, so that the effect of the new policy in Π^* can be assessed. It further says that, for case (C3), $P^*(y|x)$ needs be obtained from scratch in Π^*.

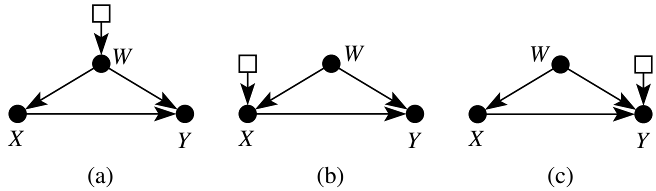
The main goal of this paper is to explicate the rationale behind this analysis and, more broadly, to provide a systematic way of deciding statistical transportability in arbitrary settings. Specifically, the contributions of the paper are as follows. In Sec. 3, we develop a novel graphical decomposition of the observational distribution that takes the latent structure into account, which generalizes the celebrated *C-component* strategy. In Sec. 4, we develop a sound, complete, and efficient algorithm to decide whether a query $P^*(\mathbf{y}|\mathbf{x})$ can be transported from a combination of data in the source ($P(\mathbf{v})$) and target domains ($P^*(\mathbf{w}), \mathbf{W} \subseteq \mathbf{V}$). In Sec. 5, we connect the problem of identification of dynamic plans (stochastic interventions) with statistical transportability, and show a reduction that proves the completeness of the former. After all, the algorithmic treatment and completeness results proved in Sections 4 and 5 close two long-standing problems in causal inference, respectively, statistical transportability [Pearl and Bareinboim, 2011] and identification of dynamic plans [Pearl and Robins, 1995; Pearl, 2000, Ch. 4.4; Dawid *et al.*, 2010].

2 Preliminaries

We use as basic semantical framework of our analysis Structural Causal Models [Pearl, 2000, pp. 204-207], which will allow the formal articulation of the invariances needed to extrapolate findings across settings, as defined next:

Definition 1 (Structural Causal Model (SCM))**.** A structural causal model M is a 4-tuple $\langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{u}) \rangle$, where

- $\mathbf{U}$ is a set of exogenous (unobserved) variables;
- $\mathbf{V}$ is a set of endogenous (observed) variables;
- $\mathcal{F}$ represents a collection of functions $\mathcal{F} = \{f_i\}$ such that each endogenous variable $V_i \in \mathbf{V}$ is determined by a function $f_i \in \mathcal{F}$, where f_i is a mapping from the respective domain of $U_i \cup Pa_i$ to V_i, with $U_i \subseteq \mathbf{U}$, $Pa_i \subseteq \mathbf{V} \setminus \{V_i\}$;
- The uncertainty is encoded through a probability distribution over the exogenous variables, $P(\mathbf{u})$.



(a) (b) (c)

Figure 1: Models where the differences between source and target domain affect different variables.

Every SCM M is associated with one causal diagram represented as a directed acyclic graph where any variable in $V_i \in \mathbf{V}$ is a vertex, and there exists a directed edge from every variable in Pa_i to V_i. Also, for every pair $V_i, V_j \in \mathbf{V}$ such that $U_i \cap U_j \neq \emptyset$, there exists a bidirected edge between V_i and V_j. We denote this causal diagram with the letter $\mathcal{G}$.

A SCM M induces a probability distribution $P(\mathbf{v})$ over the set of observed variables $\mathbf{V}$ such that

$$P(\mathbf{v}) = \sum\nolimits_{\mathbf{u}} \prod\nolimits_{\{i|V_i \in \mathbf{V}\}} P(v_i \mid pa_i, u_i) P(\mathbf{u}), \qquad (2)$$

where each term $P(v_i|pa_i, u_i)$ corresponds to the function $f_i \in \mathcal{F}$ in the underlying structural causal model M. These functions represent autonomous mechanisms affecting only its corresponding V_i, locally [Aldrich, 1989]. In this paper, we operate non-parametrically, i.e., making no assumption about the particular functional form or the distribution of the unobserved variables [Pearl, 2000]. In this case, the only assumption is that the arguments of the functions are known as encoded through the causal diagram $\mathcal{G}$.

If each exogenous $U_i \in \mathbf{U}$ affects only one observed variable $V_i \in \mathbf{V}$, the causal model is called *Markovian.* In this class of models, the joint distribution over observables can be factorized as the product of the conditional probabilities of each variable given its parents, i.e., $P(\mathbf{v}) = \prod_i P(v_i|pa_i)$. Those conditional distributions provide a canonical way to parametrize a model, since once each of them is given, the whole joint distribution is characterized. In contrast, in most real-world scenarios, Markovianity rarely holds since latent variables commonly affect more than one observable. In this paper, we focus on models where Markovianity does not hold, which we call non-Markovian.

Random variables are denoted with uppercase letters (e.g, C) while their instantiations to particular values are written in lowercase (e.g, c). Similarly, sets of variables are written in bold (e.g, $\mathbf{C}$) and a vector of a value assignment to them in lowercase-bold letters (e.g., $\mathbf{c}$). Following standard notation, we denote by $\mathcal{G}_{\overline{\mathbf{W}}\underline{\mathbf{X}}}$ the graph that is the same as $\mathcal{G}$ except that the edges incoming to variables in $\mathbf{W}$ and the edges going out from variables in $\mathbf{X}$ are removed. Let $\mathcal{G}_{[\mathbf{C}]}$ be the subgraph of $\mathcal{G}$ made only of nodes in $\mathbf{C} \subset \mathbf{V}$ and the edges between them.

When considering any topological order $V_1, \ldots, V_k$ consistent with $\mathcal{G}$, let $\mathbf{D}^{\leq i} = \{V_{d_1}, \ldots, V_{d_i}\}$ be the set of variables in $\mathbf{D} \subset \mathbf{V}$ ordered before V_{d_i} (including V_{d_i}), and $\mathbf{D}^{>i} = \mathbf{D} \setminus \mathbf{D}^{\leq i}$ for $i = 1, \ldots, k$, and $\mathbf{D}^{\leq 0} = \emptyset$.

We define $Pa(\mathbf{C})$ and $An(\mathbf{C})$, as the union of $\mathbf{C} \subset \mathbf{V}$ with its parents and ancestors, respectively.

3 Factorization of Non-Markovian Models

In order to find a complete method for obtaining $P^*(\mathbf{y}|\mathbf{x})$ from $P(\mathbf{v})$ and $P^*(\mathbf{w})$, we will decompose both input and output distributions into small factors identifiable from data.

We introduce a strategy for decomposing distributions induced by Non-Markovian models, and derive a number of corresponding graphical and algorithmic properties. Overall, once both input and output distributions are canonically decomposed into *factors*, we solve the task if every factor of the output can be obtained from the available inputs.

Our factorization builds on constructs known as *c-factors* and *c-components* developed by Tian and Pearl in (2002a; 2002b). We start by augmenting these definitions to explicitly account for marginalization. First, let $\mathbf{C}, \mathbf{H} \subseteq \mathbf{V}$ be disjoint subsets, define the quantity $Q[\mathbf{C} \parallel \mathbf{H}]$, called *c*-factor*, to denote the following function

$$Q[\mathbf{C} \parallel \mathbf{H}](pa(\mathbf{c} \cup \mathbf{h}) \setminus \mathbf{h}) = \sum_{u(\mathbf{C}\cup\mathbf{H}),\mathbf{h}} \prod_{\{i|V_i \in \mathbf{C}\cup\mathbf{H}\}} P(v_i \mid pa_i, u_i) P(u(\mathbf{C}\cup\mathbf{H})), \quad (3)$$

where pa_i and u_i are the sets of observable and unobservable parents of V_i, respectively; and for any set $\mathbf{B} \subseteq \mathbf{V}$, define $U(\mathbf{B}) = \bigcup_{V_i \in \mathbf{B}} U_i$. For consistency, we define $Q[\emptyset \parallel .] = 1$.

For the three diagrams in Fig. 1, the factor $Q[X, W, Y \parallel \emptyset]$ is equal to $P(w)P(x|w)P(y|w,x){=}P(\mathbf{v})$. Similarly, the c*-factor $Q[X, Y \parallel W]{=}\sum_w P(w)P(x|w)P(y|w,x){=}P(x,y)$. On the other hand, for the non-Markovian model in Fig. 2(a),

$$Q[Z \parallel C, D] = \sum_{u',c,d} P(z|c,u_1)P(c|u_1,u_2)P(d|u_3)P(u'), \quad (4)$$

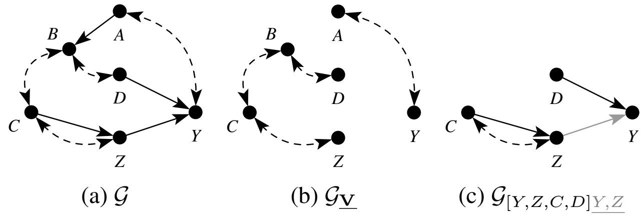
where $U' = \{U_1, U_2, U_3\}$ is the corresponding set of unobserved variables represented by the bidirected edges between the pairs of observables $(C, Z), (B, C), (B, D)$.

To simplify the notation, whenever clear from the context, we will write $Q[\mathbf{C} \parallel \mathbf{H}](pa(\mathbf{c}{\cup}\mathbf{h}){\setminus}\mathbf{h})$ as $Q[\mathbf{C} \parallel \mathbf{H}]$, $Q[\mathbf{C} \parallel \emptyset]$ as $Q[\mathbf{C}]$, and whenever $\mathbf{C} = \{V_i\}$, $\mathbf{H} = \{V_j\}$, we will write $Q[V_i \parallel V_j]$ instead of $Q[\{V_i\} \parallel \{V_j\}]$.

Decomposing Distributions based on the Causal Graph. As functions of probability distributions, c*-factors can, sometimes, be "factorized" depending on the variables that are marginalized out. The following definition characterize such decomposability property.

Definition 2 (C*-Component)**.** Let $\mathcal{G}$ be a graph over variables $\mathbf{V}$ and let $\mathbf{C}, \mathbf{H} \subseteq \mathbf{V}$ be two disjoint subsets such that $\mathbf{C} \cup \mathbf{H} = \mathbf{V}$. Then, $\mathbf{C} \cup \mathbf{H}$ can be partitioned, relative to $\mathbf{C}$, into sets $\{(\mathbf{C}_1\|\mathbf{H}_1), (\mathbf{C}_2\|\mathbf{H}_2), \dots (\mathbf{C}_l\|\mathbf{H}_l)\}$, such that $\{C_1, \dots, C_l\}$ partition $\mathbf{C}$ and $\{H_1, \dots, H_l\}$ partition $\mathbf{H}$. Two variables $V_j, V_k \in \mathbf{V}$ belong to the same $(\mathbf{C}_i\|\mathbf{H}_i)$ if there exists any path (regardless of the directionality of the arrows) between them in $\mathcal{G}_{\underline{\mathbf{C}}}$.

For $\mathcal{G}$ in Fig. 2(a), with $\mathbf{V} = \{A, B, Y, Z, C, D\}$ there are two c*-components relative to $\mathbf{V}$, i.e.: $(Z, C, B, D\|\emptyset)$ and $(A, Y\|\emptyset)$, which are the two disconnected subgraphs in $\mathcal{G}_{\underline{\mathbf{V}}}$ in (Fig. 2(b)). Meanwhile, for $\mathcal{G}_{[Y,Z,C,D]}$ in Fig. 2(c), where $\mathbf{V} = \{Y, Z, C, D\}$, the c*-components relative to $\{Z, Y\}$ are $(Y\|D)$ and $(Z\|C)$. This is clear when we look at the graph $\mathcal{G}_{[Y,Z,C,D]\underline{Y,Z}}$ where the outgoing arrows from Z and Y are removed (in this case, the arrow is shown in grey).



(a) $\mathcal{G}$ (b) $\mathcal{G}_{\underline{\mathbf{V}}}$ (c) $\mathcal{G}_{[Y,Z,C,D]\underline{Y,Z}}$

Figure 2: (a) Causal graph $\mathcal{G}$, (b) subgraph $\mathcal{G}_{\underline{\mathbf{V}}}$, and (c) subgraph $\mathcal{G}_{[Y,Z,C,D]}$. Further, $\mathcal{G}_{[Y,Z,C,D]\underline{Y,Z}}$ can be obtained from the subgraph (c) by removing the arrow $Z \to Y$ (shown in gray).

The importance of the c*-components stems from the fact that they lead to a natural decomposition of the corresponding c*-factors, which will disentangle the latents in a fundamental way. For example, we note that the factor $Q[Y, Z \parallel C, D] = \sum_{\mathbf{u},c,d} P(z|c,u_1,u_2)P(c|u_1,u_2)P(d|u_3)P(y|z,d,u_4)P(\mathbf{u})$ can be decomposed as the following independent factors,

$$\Big(\sum_{u_1,u_2,c} P(z|c,u_1,u_2)P(c|u_1,u_2)P(u_1,u_2)\Big)$$
$$\Big(\sum_{u_3,u_4,d} P(d|u_3)P(y|z,d,u_4)P(u_3,u_4)\Big), \quad (5)$$

which is equal to $Q[Z \parallel C]Q[Y \parallel D]$ (i.e., product of $(Y\|D)$ and $(Z\|C)$). The lemma below generalizes this property.

Lemma 1 (C*-Decomposition)**.** *Let $\mathbf{D} \subseteq \mathbf{V}\setminus\mathbf{H}$, and assume that $\mathbf{D}$ is partitioned into c*-components $(\mathbf{D}_i\|\mathbf{H}_i)_{i=1,\dots,l}$ in the subgraph $\mathcal{G}_{[\mathbf{D}\cup\mathbf{H}]}$, relative to $\mathbf{D}$. Then we have*

(i) $Q[\mathbf{D} \parallel \mathbf{H}]$ decomposes as

$$Q[\mathbf{D} \parallel \mathbf{H}] = \prod_i Q[\mathbf{D}_i \parallel \mathbf{H}_i]. \quad (6)$$

(ii) Let a topological order of the variables in $\mathbf{D}$ be $V_{d_1} < \cdots < V_{d_k}$ in $\mathcal{G}_{[\mathbf{D}\cup\mathbf{H}]}$.

$$Q[\mathbf{D}_j \parallel \mathbf{H}_j] = \prod_{\{i|V_{d_i} \in \mathbf{D}_j\}} \frac{Q[\mathbf{D}^{\leq i} \parallel \mathbf{H}]}{Q[\mathbf{D}^{\leq i-1} \parallel \mathbf{H}]}, \quad (7)$$

where each $Q[\mathbf{D}^{\leq i} \parallel \mathbf{H}]$, $i = 0, 1, \dots, k$, is given by

$$Q[\mathbf{D}^{\leq i} \parallel \mathbf{H}] = \sum_{\mathbf{d}^{>i}} Q[\mathbf{D} \parallel \mathbf{H}]. \quad (8)$$

To illustrate this lemma, suppose we are given $Q[Y, Z \parallel C, D]$, and note that according to Eqs. (7)-(8),

$$Q[Y \parallel D] = Q[Z, Y \parallel C, D] / \sum_y Q[Z, Y \parallel C, D], \quad (9)$$

$$Q[Z \parallel C] = \sum_y Q[Z, Y \parallel C, D] / \sum_{y,z} Q[Z, Y \parallel C, D]. \quad (10)$$

As it turns out, it is not always possible to uniquely identify a particular c*-factor from the available data. The following definition formalizes this notion.

Definition 3 (C*-factor Identifiability)**.** A c*-factor $Q[\mathbf{C} \parallel \mathbf{H}]$ is said to be identifiable from a causal graph $\mathcal{G}$ and a c*-factor $Q[\mathbf{T} \parallel \mathbf{L}]$ if $Q[\mathbf{C} \parallel \mathbf{H}]$ can be computed uniquely from $Q[\mathbf{T} \parallel \mathbf{L}]$. That is, if $Q^{M_1}[\mathbf{C} \parallel \mathbf{H}]{=}Q^{M_2}[\mathbf{C} \parallel \mathbf{H}]$ for every pair of models M_1 and M_2 that induce the same $\mathcal{G}$ with $Q^{M_1}[\mathbf{T} \parallel \mathbf{L}]{=}Q^{M_2}[\mathbf{T} \parallel \mathbf{L}]$, $Q[\mathbf{C} \parallel \mathbf{H}]$ is identifiable.

For instance, the c*-factor $Q[Z \parallel \emptyset]$ in Fig. 2(a) cannot be identified from $Q[Z, C \parallel \emptyset]$. In words, there exists two different SCMs M_1 and M_2 with the same $\mathcal{G}$ and generating the same $Q[Z, C \parallel \emptyset] = \sum_{u_1,u_2} P(z|c, u_1)P(c|u_1, u_2)P(u_1, u_2)$, but with different $Q[Z \parallel \emptyset] = \sum_{u_1} P(z|c, u_1)P(u_1)$, for some values z, c.

In Alg. 1, we develop a procedure to identify c*-factors from other c*-factors, which we called *Identify**. This algorithm uses a modified version of *Identify* [Tian and Pearl, 2002a] as a subroutine (for details, see Appendix A).

Theorem 1. *$Q[\mathbf{C} \parallel \mathbf{H}]$ is identifiable from $Q[\mathbf{T} \parallel \mathbf{L}]$ in $\mathcal{G}$ if and only if Identify* returns an expression for it.*

In words, failure of *Identify** implies there exist SCMs M_1, M_2 compatible with $\mathcal{G}$ that induce the same $Q[\mathbf{T} \parallel \mathbf{L}]$ but $Q^{M_1}[\mathbf{C} \parallel \mathbf{H}](\mathbf{v}) \neq Q^{M_2}[\mathbf{C} \parallel \mathbf{H}](\mathbf{v})$, for some values $\mathbf{v}$.

Query Decomposition. Once the c*-factors and components are well-understood given a specific causal graph, we turn our attention to the query distribution, $P(\mathbf{y}|\mathbf{x})$, given by

$$P(\mathbf{y}|\mathbf{x}) = P(\mathbf{y}, \mathbf{x})/P(\mathbf{x}) = P(\mathbf{y}, \mathbf{x})/\sum\nolimits_{\mathbf{y}} P(\mathbf{y}, \mathbf{x}). \quad (11)$$

The distribution $P(\mathbf{y}, \mathbf{x})$ can be expressed in terms of c*-factors, as shown in the following example.

Example 2. Consider the query $P(y|z)$ in the context of the graph in Fig. 2(a), we have

$$P(y, z) = \sum\nolimits_{c,d} P(y, z, c, d) = Q[Y, Z \parallel C, D] \quad (12)$$

$$P(z) = Q[Z \parallel C, D, Y] = Q[Z \parallel C], \quad (13)$$

where the last equality follows from the fact that $D, Y \notin An(Z)_{\mathcal{G}_{[Z,C,D,Y]}}$. The graph $\mathcal{G}_{[An(Y,Z)]} = \mathcal{G}_{[Y,Z,C,D]}$ can be partitioned relative to $\{Y, Z\}$ into c*-components $(Y \| D)$ and $(Z \| C)$ (the connected components in $\mathcal{G}_{[Y,Z,C,D]\underline{Y,Z}}$, shown in Fig. 2(c)), and by Lemma 1, $Q[Y, Z \parallel C, D]$ decomposes as $Q[Z \parallel C]Q[Y \parallel D]$, and

$$P(y, z) = Q[Z \parallel C]Q[Y \parallel D] \quad (14)$$

The conditional distribution will be equal to

$$P(y|z) = \frac{P(y, z)}{P(z)} = \frac{Q[Z \parallel C]Q[Y \parallel D]}{Q[Z \parallel C]} = Q[Y \parallel D]. \quad (15)$$

Remarkably, while $P(y, z)$ depends on $Q[Z \parallel C] = \sum_c P(z|c)P(c)$ and $Q[Y \parallel D] = \sum_d P(y|z, d)P(d)$, the query $P(y|z)$ is independent of $Q[Z \parallel C]$.

Naturally, if the joint distribution $P(y, z, d, c, b, a)$ is available, $P(y, z)$ is trivially computable. But, if we consider a scenario with two domains (e.g., websites), Π and Π^*, with the goal of estimating $P^*(y|z)$ in the target domain, the previous analysis (example 2) shows that even if the distribution $P^*(z, c)$ and $P(z, c)$ are different, $P^*(y|z)$ can still be obtained from data in the source alone. A more surprising conclusion is that if there are differences across settings in the distribution of D, $P^*(d)$ needs to be measured *despite* D not appearing in the query expression.

This leads to a more fundamental issue about which factors of the model are really required to compute a certain query. The following result answers this question in generality.

Algorithm 1 Identify*$(\mathbf{C}, \mathbf{H}, \mathbf{T}, \mathbf{L}, Q, \mathcal{G})$

Input: $\mathbf{C} \subseteq \mathbf{T} \subseteq \mathbf{V}$, $\mathbf{H} \subseteq \mathbf{V} \setminus An(\mathbf{C})_{\mathcal{G}_{[\mathbf{C}\cup\mathbf{H}]}}$, $\mathbf{L} \subseteq \mathbf{V} \setminus An(\mathbf{T})_{\mathcal{G}_{[\mathbf{T}\cup\mathbf{L}]}}$, $Q = Q[\mathbf{T} \parallel \mathbf{L}]$, graph $\mathcal{G}$. Assuming $\mathcal{G}_{[\mathbf{C}\cup\mathbf{H}]}$ and $\mathcal{G}_{[\mathbf{T}\cup\mathbf{L}]}$ are composed of a single c*-component.
Output: Expression for $Q[\mathbf{C} \parallel \mathbf{H}]$ in terms of Q or Fail.

```
$\mathbf{B} = \mathbf{H} \setminus \mathbf{L}$.
if $\mathbf{B} \neq \emptyset$ then
    Let $(\mathbf{C}_1, \mathbf{H}_1), (\mathbf{C}_2, \mathbf{H}_2), \dots$ be the c*-components of
        $\mathcal{G}$ relative to $\mathbf{B}$ intersecting variables in $\mathbf{C}$.
    return $\sum_{\mathbf{b}} \prod_i$ Identify*$(\mathbf{C}_i, \mathbf{H}_i, \mathbf{T}, \mathbf{L}, Q, \mathcal{G})$
end if
return Identify$(\mathbf{C}, \mathbf{T}, \mathbf{H}, Q, \mathcal{G})$
```

Lemma 2. *Let $\mathbf{Y}, \mathbf{X} \subseteq \mathbf{V}$ be disjoint sets of variables, then*

$$P(\mathbf{y} \mid \mathbf{x}) = Q[\mathbf{A} \parallel \mathbf{H}]/Q[\mathbf{A} \setminus \mathbf{Y} \parallel \mathbf{H} \cup \mathbf{Y}], \quad (16)$$

where $(\mathbf{A} \setminus \mathbf{Y} \| \mathbf{H} \cup \mathbf{Y})$ is the union of the c-components of $\mathcal{G}_{[An(\mathbf{Y}\cup\mathbf{X})]}$, relative to $\mathbf{X}$, intersecting the variables in $\mathbf{Y}$.*

In example 2, this maps to $\mathbf{A} = \{Y\}$ and $\mathbf{H} = \{D\}$ since $(\emptyset \| Y, D)$ is the only c*-component relative to $\mathbf{Z}$ in $\mathcal{G}_{[Z,Y,C,D]}$ that intersects variables in $\mathbf{Y}$.

4 Transporting Probabilistic Relations

In this section, we build on the machinery developed so far to decompose conditional distributions (e.g., classifiers) and extrapolate them from one domain to another, as illustrated in Example 1. This task is known as statistical transportability [Pearl and Bareinboim, 2011, Section 5].

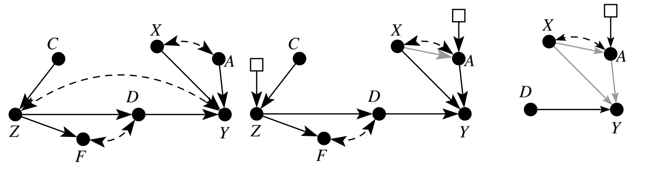
In any transportability instance, the source and target domains, Π and Π^*, are modelled explicitly and assumed to be governed by two corresponding SCMs, M and M^*. Each causal model is accompanied by its corresponding causal graph, in this case, $\mathcal{G}$ and $\mathcal{G}^*$. In the language of transportability, a special indicator variable T (drawn as squares in $\mathcal{G}^*$) is used to represent differences between the two domains. A T-node points to a variable affected by unobserved factors (causal mechanism or distribution of exogenous) that are distinct across domains. Formally, $\mathcal{G}^*$ contains an extra edge $T_i \to V_i$ whenever there *might* exist a discrepancy $f_i \neq f_i^*$ or $P(U_i) \neq P^*(U_i)$ between M and M^*. This also allows one to accommodate structural changes such as when V_i has a different parent set across domains.

Recall case 2 from Example 1. Note that a T-node pointing to X indicates that the underlying function that assign values to X is different in the source and target domains, which implies that $P^*(x|w) \neq P(x|w)$. The query $P^*(y|x)$ can be transported in this case if we have $P^*(x, w)$ or more specifically $P^*(x|w)$. To see how, write the query as

$$P^*(y|x) = \sum_w P^*(y|x, w) \frac{P^*(x|w)P^*(w)}{\sum_{w'} P^*(x|w')P^*(w')}. \quad (17)$$

Due to the absence of a square node pointing to W and Y in Fig. 1(b), it follow that $P^*(y|x, w) = P(y|x, w)$ and $P^*(w) = P(w)$, which leads to the following expression:

$$P^*(y \mid x) = \sum_w P(y \mid x, w) \frac{P^*(x \mid w)P(w)}{\sum_w P^*(x \mid w)P(w)}, \quad (18)$$



(a) $\mathcal{G}$ (b) $\mathcal{G}^*$ (c) $\mathcal{G}^*_{[An(Y,X,Z)]}$

Figure 3: Causal diagrams involved in transporting $P^*(y|x,z)$.

Eq. (18) is a mixture of factors from Π and Π^* within the available input distributions, $P^*(x|w)$ and $P(w,x,y)$.

On the other hand, the third case (Fig. 1(c)) requires one to measure $P^*(y|x,w)$ in the target domain, which usually means having the full distribution $P^*(y,x,w)$. In this case, the task is done from scratch, without any transportability.

Next, we formally state the statistical transportability task.

Definition 4 (Statistical Transportability [Pearl and Bareinboim, 2011])**.** Given two populations, Π and Π^*, characterized by probability distributions P and P^*, and causal diagrams $\mathcal{G}$ and $\mathcal{G}^*$, respectively; a statistical relation $R(P)$ is said to be statistically transportable from Π to Π^* over $\mathbf{W} \subset \mathbf{V}$ if $R(P^*)$ is identified from P, $P^*(\mathbf{W})$, $\mathcal{G}$, and $\mathcal{G}^*$.

We will now focus on the general case of transporting relationships R of the form $R(P^*) = P^*(\mathbf{y}|\mathbf{x})$ for any pair of disjoint sets of variables $\mathbf{X}, \mathbf{Y} \subset \mathbf{V}$, from inputs $P(\mathbf{v})$ and $P^*(\mathbf{w})$. Using Lemma 2 in the context of Π^*, we can write

$$P^*(\mathbf{y} \mid \mathbf{x}) = \frac{Q^*[\mathbf{A} \,\|\, \mathbf{H}]}{Q^*[\mathbf{A}\backslash\mathbf{Y} \,\|\, \mathbf{H}\cup\mathbf{Y}]} = \frac{Q^*[\mathbf{A} \,\|\, \mathbf{H}]}{\sum_{\mathbf{y}} Q^*[\mathbf{A} \,\|\, \mathbf{H}]}, \quad (19)$$

where $Q^*[.\,\|\,.]$ describes a c*-factor in the domain Π^* and consistent with the corresponding causal model $\mathcal{G}^*$.

Example 3. Figs. 3(a) and (b) represent the causal graphs in the source and target domains, Π and Π^*. The goal is to transport $R = P^*(y|x,z)$ to a domain Π^* where the behavior of the variable Z depends only on C, which contrasts with the Π's behavior where Z depends on an unobservable variable shared with Y (bidirected arrow). Note that the variable A in Π^* has an extra dependency on the value of X that is not present Π. Only data from $P^*(x,a)$ is available in Π^*.

Due to their distinct behaviors, $\mathcal{G}^*$ has T-nodes pointing to Z and A (i.e., $\square \to Z, \square \to A$). Using Lemma 2 w.r.t. R, note that the c*-components relative to $\{X, Z\}$ in $\mathcal{G}^*_{[An(Y,Z,X)]}$ that contain Y is $(X \| A, Y, D)$, yielding:

$$P^*(y|x,z) = Q^*[Y,X \,\|\, A,D] \Big/ \sum\nolimits_y Q^*[Y,X \,\|\, A,D], \quad (20)$$

which corresponds to $\mathbf{A} = \{Y, X\}$ and $\mathbf{H} = \{A, D\}$.

If $An(\mathbf{A}) \subseteq \mathbf{W}$, then the given distribution $P^*(\mathbf{w})$ includes all variables that are ancestors of the set $\mathbf{A}$ in the graph $\mathcal{G}^*$, and $P^*(\mathbf{y} \mid \mathbf{x})$ is easily estimated from $P^*(\mathbf{w})$. If this is not the case, it is possible to leverage data from the source domain Π, namely, $P(\mathbf{v})$. The key observation is that both domains Π and Π^* share some commonalities. The following lemma formalizes this notion.

Lemma 3. *[C*-invariance] Let $\mathcal{G}, \mathcal{G}^*$ be a pair of graphs for $\langle M, M^* \rangle$, then $Q^*[\mathbf{C} \,\|\, \mathbf{H}] = Q[\mathbf{C} \,\|\, \mathbf{H}]$ if $\mathcal{G}^*$ does not contain a selection node T_i pointing to any $V_i \in \mathbf{C} \cup \mathbf{H}$.*

Algorithm 2 Transport*$(\mathcal{G}, \mathcal{G}^*, \mathbf{Y}, \mathbf{X}, \mathbf{W})$

```
Input: $\mathcal{G}, \mathcal{G}^*$ causal diagrams over a set of variables $\mathbf{V}$, $\mathbf{Y}, \mathbf{X} \subseteq \mathbf{V}$ disjoint subsets of variables, $\mathbf{W} \subseteq \mathbf{V}$.
Output: $P^*(\mathbf{y}|\mathbf{w})$ in terms of $P(\mathbf{v}), P^*(\mathbf{w})$ or Fail.
Let $(\mathbf{A}\|\mathbf{H})$ be defined as in Lemma 2 with $\mathcal{G}^*$.
Let $\mathbf{D} \subseteq \mathbf{A}\cup\mathbf{H}$ be the set of vars. pointed by $T$-nodes.
Let $I$ be the set of c*-components $(\mathbf{A}_i\|\mathbf{H}_i), i = 1, \dots, k$ of $\mathcal{G}^*_{[\mathbf{A}\cup\mathbf{H}]}$, relative to $\mathbf{A}\cup(\mathbf{W}\cap\mathbf{H})$, such that $(\mathbf{A}_i\cup\mathbf{H}_i)\cap \mathbf{D} = \emptyset$, and Identify*$(\mathbf{A}_i, \mathbf{H}_i, \mathbf{V}, \emptyset, P(\mathbf{v}), \mathcal{G}) \neq$ Fail; and let $N$ be the rest of them.
$\mathbf{N_A} \leftarrow \bigcup_{(\mathbf{A}_i\|\mathbf{H}_i)\in N} \mathbf{A}_i$ and $\mathbf{N_H} \leftarrow \bigcup_{(\mathbf{A}_i\|\mathbf{H}_i)\in N} \mathbf{H}_i$.
if $\mathbf{N_A} \not\subseteq \mathbf{W}$ then return Fail.
$Q^*[\mathbf{N_A} \,\|\, \mathbf{N_H}] \leftarrow \sum_{\mathbf{w}\backslash\mathbf{n_A}} P^*(\mathbf{w})$.
$Q^*[\mathbf{A}\|\mathbf{H}] \leftarrow \sum_{\mathbf{w}\cap\mathbf{h}} Q^*[\mathbf{N_A}\|\mathbf{N_H}] \prod_{(\mathbf{A}_i\|\mathbf{H}_i)\in I} Q[\mathbf{A}_i\|\mathbf{H}_i]$.
return $Q^*[\mathbf{A} \,\|\, \mathbf{H}] \Big/ \sum_{\mathbf{y}} Q^*[\mathbf{A} \,\|\, \mathbf{H}]$.
```

In words, if there is no T-node pointing to any variable in a set $\mathbf{C} \cup \mathbf{H}$ in $\mathcal{G}^*$, the c*-factor $Q^*[\mathbf{C} \,\|\, \mathbf{H}]$ is invariant and can, therefore, be transported from the source domain. In example 3, the c*-factor $Q^*[Y \,\|\, D]{=}Q[Y \,\|\, D]$, while $Q^*[X, A \,\|\, \emptyset]$ may be different than $Q[X, A \,\|\, \emptyset]$.

Let $\mathbf{D} \subseteq \mathbf{A} \cup \mathbf{H}$ be the subset of variables involved in Eq. (19) that are pointed by T-nodes in $\mathcal{G}^*$. Then, factors involving variables in $\mathbf{D}$ need to be obtained from $P^*(\mathbf{w})$, which is equivalent to $Q^*[\mathbf{W} \,\|\, An(\mathbf{W})\backslash\mathbf{W}]$.

Let N be the set of c*-components in the graph $\mathcal{G}^*_{[An(\mathbf{A}\cup\mathbf{H})]}$, relative to $\mathbf{A}\cup(\mathbf{W}\cap\mathbf{H})$, that intersect any variable in $\mathbf{D}$ or those that cannot be identified from $P(\mathbf{v})$. Let $\mathbf{N_A} = \bigcup_{(\mathbf{A}_i\|\mathbf{H}_i)\in N} (\mathbf{A}_i)$ and $\mathbf{N_H} = \bigcup_{(\mathbf{A}_i\|\mathbf{H}_i)\in N} (\mathbf{H}_i)$.

In Example 3, $\mathcal{G}^*_{[\mathbf{A}\cup\mathbf{H}]}$ (Fig. 3(c)) has, relative to $\{Y, X, A\}$, c*-components $(Y\|D)$ and $(X, A\|\emptyset)$. The only variable pointed by a T-node is A, so $(\mathbf{N_A}\|\mathbf{N_H}) = (X, A\|\emptyset)$. That need to be obtained from $P(x,a)$; the following lemma characterizes this operation.

Lemma 4. *Given a causal diagram $\mathcal{G}$, let $\mathbf{C}, \mathbf{H} \subset \mathbf{V}$ be two disjoint subsets and $\mathbf{W} \subset \mathbf{V}$ such that $\mathbf{H} \subseteq An(\mathbf{C})_{\mathcal{G}_{\underline{\mathbf{W}}}}$. If $\mathbf{C} \subseteq \mathbf{W}$, then*

$$Q[\mathbf{C} \,\|\, \mathbf{H}] = \sum\nolimits_{\mathbf{w}\backslash\mathbf{c}} P(\mathbf{w}). \quad (21)$$

Then, if $\mathbf{N_A}$ and $\mathbf{N_H}$ satisfy $\mathbf{N_A} \cup \mathbf{N_H} \subseteq An(\mathbf{W})_{\mathcal{G}^*}$ and $\mathbf{N_A} \subseteq \mathbf{W}$, Lemma 4 licenses

$$Q^*[\mathbf{N_A} \,\|\, \mathbf{N_H}] = \sum\nolimits_{\mathbf{w}\backslash\mathbf{n_A}} P^*(\mathbf{w}). \quad (22)$$

In our example, $Q[X, A \,\|\, \emptyset] = P^*(x,a)$.

Let I be the c*-components of $\mathcal{G}^*_{[\mathbf{A}\cup\mathbf{H}]}$ not in N, that by definition, are all identifiable from $P(\mathbf{v})$. Consequently, the quantity $Q^*[\mathbf{A} \,\|\, \mathbf{H}]$ is identifiable from $P(\mathbf{v})$ and $P^*(\mathbf{w})$ as:

$$Q^*[\mathbf{A} \,\|\, \mathbf{H}] = \sum_{\mathbf{w}\cap\mathbf{h}} Q^*[\mathbf{N_A} \,\|\, \mathbf{N_H}] \prod_{(\mathbf{A}_i\|\mathbf{H}_i)\in I} Q[\mathbf{A}_i \,\|\, \mathbf{H}_i]. \quad (23)$$

For Example 3, we can obtain $Q^*[Y \,\|\, D]{=}Q[Y \,\|\, D]$, using *Identify**$(Y, D, \mathbf{V}, \emptyset, P(\mathbf{v}), \mathcal{G})$ and in terms of $P(\mathbf{v})$ as

$$Q[Y \,\|\, D] = \sum\nolimits_d P(d|z) \sum\nolimits_{z'} P(y|x,z',d,a)P(z'). \quad (24)$$

And the final expression is

$$P^*(y|x,z)=\sum_a P^*(a|x)\sum_d P(d|z)\sum_{z'} P(y|x,z',d,a)P(z'). \quad (25)$$

We incorporate the discussion so far in the procedure *Transport** shown in Alg. 2, which is complete for this task as given by the following statement.

Theorem 2. *[Completeness] The relationship $R = P^*(\mathbf{y}|\mathbf{x})$ is transportable from $P(\mathbf{V})$, $P^*(\mathbf{W})$ and $\mathcal{G}, \mathcal{G}^*$ if and only if Transport*$(\mathcal{G}, \mathcal{G}^*, \mathbf{Y}, \mathbf{X}, \mathbf{W})$ does not fail.*

5 Identifying Dynamic Plans

In this section, we investigate the identifiability of dynamic plans [Pearl and Robins, 1995; Pearl, 2000; Didelez *et al.*, 2006] and show that it can be reduced to the problem of transportability of marginals. We further show that the procedure introduced in [Tian, 2008] is not only sound but complete. This closes the problem of identifiability of dynamic sequential plans since all sequential plans that are computable from observational data can be algorithmically determined.

In this setting, we assume that the observational data $P(\mathbf{V})$ is collected from a SCM that induces a causal diagram $\mathcal{G}$ with observed variables $\mathbf{V}$ in domain Π. The goal is to determine the behavior of $\mathbf{Y} \subset \mathbf{V}$ under a hypothetical intervention that changes the mechanisms of variables $\mathbf{X} \subset \mathbf{V}$, which we call domain Π^*. The probability distribution over $\mathbf{V}$ in the intervened system is $P^*(\mathbf{V})$ and the associated graph is $\mathcal{G}^*$. The identifiability task amounts to computing $P^*(\mathbf{Y})$ in that hypothetical world using $P(\mathbf{V})$, the assumptions encoded in $\mathcal{G}, \mathcal{G}^*$, and knowledge about the interventions.

There are different ways of performing an intervention and changing the value of the corresponding variable, say $X \in \mathbf{X}$.

Atomic or do-intervention. The variable X is set to a constant value $\mathbf{x}'$ in its domain, i.e., $P^*(x \mid pa_x^*, u_x^*) = 1_{X=x'}$.

Conditional Intervention. The value of X is determined as a function $g(pa_x^*)$, that is, $P^*(x \mid pa_x^*, u_x^*) = 1_{X=g(pa_x^*)}$.

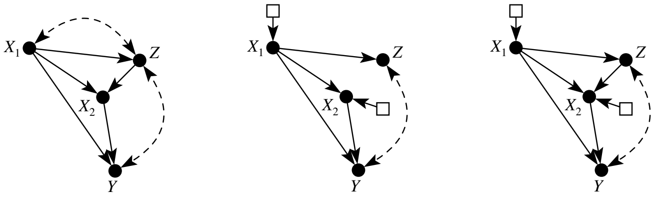
Stochastic Intervention. X will take value x in its domain according to an externally specified probability distribution $P^*(x \mid pa_x^*)$, namely, $P^*(x \mid pa_x^*, u_x^*) = P^*(x \mid pa_x^*)$.

We note that these different types of interventions can be encoded in a similar fashion using the formalism of transportability by adding T-nodes pointing to each variable in the interventional set, $\mathbf{X}$. This indicates that their functions are different that those in the original system Π. Since the new intervention is part of the input of the problem, $Q^*[\mathbf{X}] = \prod_{X \in \mathbf{X}} Q^*[X]$ is available in the target domain.

Example 4. Consider the the dynamic plan first studied in [Pearl and Robins, 1995] and shown in Fig. 4(a). With the new transportability mapping, the goal is to assess the distribution $P^*(\mathbf{y})$ in two hypothetical environments where:

C1 the values of (X_1, X_2) have been fixed to (x_1, x_2), in standard do-form, i.e., $P^*(x_1, x_2) = 1_{(X_1,X_2)=(x_1,x_2)}$. The corresponding $\mathcal{G}^*$ is shown in Fig. 4(b).

C2 the value of X_1 is fixed to x_1 and X_2 is set conditionally on Z based on a function $g(z)$. That is, $P^*(x_1) = I_{X_1=x_1}$ and $P^*(x_2 \mid z) = 1_{X_2=g(z)}$. The corresponding $\mathcal{G}^*$ is shown in Fig. 4(c).



(a) $\mathcal{G}$ (b) $\mathcal{G}^*$ (C1) (c) $\mathcal{G}^*$ (C2)

Figure 4: Causal diagrams associated with Example 4.

In [Pearl and Robins, 1995], the impact of the plan in the outcome variable $\mathbf{Y}$ is denoted as $P(\mathbf{y} \mid \hat{x_1}, \ldots, \hat{x_n})$ (written as $P(\mathbf{y}; \sigma_{\mathbf{x}})$ in [Tian, 2008]). The following follows:

Definition 5 (Plan Identifiability)**.** A sequential plan is said to be identifiable if $P(\mathbf{y} \mid \hat{x_1}, \ldots, \hat{x_n})$ (equivalently $P(\mathbf{y}; \sigma_{\mathbf{x}})$) is uniquely computable from the joint distribution $P(\mathbf{v})$, for every assignment $(x_1, \ldots, x_n)$.

We show next that the reduction from plan identification to transportability is indeed valid and that the corresponding completeness follows.

Theorem 3. *The effect $P(\mathbf{y}|\hat{x_1}, \ldots, \hat{x_n})$ (or, $P(\mathbf{y}; \sigma_{\mathbf{x}})$) is equivalent to $P^*(\mathbf{y})$ where Π^* is related to $\mathcal{G}^*$ and $Q^*[\mathbf{X}]$ is determined by the corresponding intervention on $\mathbf{X}$.*

Corollary 1. *Transport* is complete for plan identification given $P(\mathbf{v})$ and $Q^*[\mathbf{X}]$.*

Corollary 2. *The condition in Thm. 1 from [Tian, 2008] is also necessary for dynamic plan identification.*

Using this result, one can immediately see that contrary to previous beliefs [Pearl, 2000, pp.120], and as hinted by [Tian, 2008], the plan in Example 4(C2) is not identifiable from $P(\mathbf{v})$ and $Q^*[\mathbf{X}]$. While case (C1), associated with Fig. 4(b), can be solved using do-calculus [Pearl, 1995], the extra edge $Z \to X_2$ in Fig. 4(c) (C2) makes X_1 and Y dependent conditional on X_2, hence the same derivation strategy does not work (see non-identifiability proof in Appendix C).

6 Conclusions

We developed a procedure (Alg. 2) that is complete (Thm. 2) for the task of estimating a distribution $P^*(\mathbf{y}|\mathbf{x})$ in a target domain Π^* using limited data from Π^* and a probability distribution collected in a different source domain Π. The algorithm uses a canonical factorization of marginal and conditional distributions, and a procedure to identify such factors (Alg. 1 & Thm. 1). Also, we showed that identifying the effect of plans can be reduced to a statistical transportability task for which our method is complete (Thm. 3). We hope that these results provide a better understanding of the assumptions and trade-offs involved in the construction of more robust and generalizable learning systems.

Acknowledgments

This research was supported by grants from IBM Research, Adobe Research, NSF IIS-1704352, and IIS-1750807 (CAREER). We thank Amin Jaber for useful discussion and feedback on a early version of this manuscript.

References

[Aldrich, 1989] John Aldrich. Autonomy. *Oxford Economic Papers*, 41:15–34, 1989.

[Bareinboim and Pearl, 2012] Elias Bareinboim and Judea Pearl. Transportability of causal effects: Completeness Results. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, CA, 2012. Department of Computer Science, University of California, Los Angeles.

[Bareinboim and Pearl, 2013] Elias Bareinboim and Judea Pearl. A general algorithm for deciding transportability of experimental results. *Journal of Causal Inference*, 1(1):107–134, 2013.

[Bareinboim and Pearl, 2014] Elias Bareinboim and Judea Pearl. Transportability from Multiple Environments with Limited Experiments: Completeness Results. In Z Ghahramani, M Welling, C Cortes, N D Lawrence, and K Q Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 280–288. Curran Associates, Inc., 2014.

[Bareinboim and Pearl, 2016] Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

[Dawid *et al.*, 2010] A Philip Dawid, Vanessa Didelez, and others. Identifying the consequences of dynamic treatment strategies: A decision-theoretic overview. *Statistics Surveys*, 4:184–231, 2010.

[Didelez *et al.*, 2006] Vanessa Didelez, A Philip Dawid, and Sara Geneletti. Direct and indirect effects of sequential treatments. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 138–146. AUAI Press, 2006.

[Lee and Honavar, 2013a] Sanghack Lee and Vasant Honavar. Causal Transportability of Experiments on Controllable Subsets of Variables: z-Transportability. In A Nicholson and P Smyth, editors, *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 361–370. AUAI Press, 2013.

[Lee and Honavar, 2013b] Sanghack Lee and Vasant Honavar. m-Transportability: Transportability of a Causal Effect from Multiple Environments. In M desJardins and M Littman, editors, *Proceedings of the Twenty-Seventh National Conference on Artificial Intelligence*, pages 583–590, Menlo Park, CA, 2013. AAAI Press.

[Magliacane *et al.*, 2018] Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain Adaptation by Using Causal Inference to Predict Invariant Conditional Distributions. In *Advances in Neural Information Processing Systems*, pages 10846–10856, 2018.

[Pearl and Bareinboim, 2011] Judea Pearl and Elias Bareinboim. Transportability of Causal and Statistical Relations: A Formal Approach. In *Proceedings of the Twenty-Fifth Conference on Artificial Intelligence (AAAI-11)*, pages 247–254, Menlo Park, CA, 8 2011.

[Pearl and Mackenzie, 2018] Judea Pearl and Dana Mackenzie. *The Book of Why*. Basic Books, New York, 2018.

[Pearl and Robins, 1995] Judea Pearl and J M Robins. Probabilistic evaluation of sequential plans from causal models with hidden variables. In P Besnard and S Hanks, editors, *Uncertainty in Artificial Intelligence 11*, pages 444–453. Morgan Kaufmann, San Francisco, 1995.

[Pearl, 1995] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

[Pearl, 2000] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition, 2000.

[Quiñonero-Candela *et al.*, 2009] Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009.

[Rojas-Carulla *et al.*, 2018] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

[Spirtes *et al.*, 2001] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2nd edition, 2001.

[Tian and Pearl, 2002a] Jin Tian and Judea Pearl. A General Identification Condition for Causal Effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI 2002)*, pages 567–573, Menlo Park, CA, 2002. AAAI Press/The MIT Press.

[Tian and Pearl, 2002b] Jin Tian and Judea Pearl. On the Testable Implications of Causal Models with Hidden Variables. *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI-02)*, pages 519–527, 2002.

[Tian, 2008] Jin Tian. Identifying Dynamic Sequential Plans. *Uncertainty in Artificial Intelligence*, 2008.

[Zhang *et al.*, 2013] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, pages III–819–III–827. JMLR.org, 2013.

[Zhang *et al.*, 2015] Kun Zhang, Mingming Gong, and Bernhard Scholkopf. Multi-source Domain Adaptation: A Causal View. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 3150–3157. AAAI Press, 2015.