Transportable Representations for Domain Generalization

Kasra Jalaldoust and **Elias Bareinboim**

Causal Artificial Intelligence Laboratory
Columbia University
{kasra,eb}@cs.columbia.edu

Abstract

One key assumption in machine learning literature is that the testing and training data come from the same distribution, which is often violated in practice. The anchors that allow generalizations to take place are causal, and provenient in terms of the stability and modularity of the mechanisms underlying the system of variables. Building on the theory of causal transportability, we define the notion of "transportable representations", and show that these representations are suitable candidates for the domain generalization task. Specifically, considering that the graphical assumptions about the underlying system are provided, the transportable representations can be characterized accordingly, and the distribution of label conditioned on the representation can be computed in terms of the source distributions. Finally, we relax the assumption of having access to the underlying graph by proving a graphical-invariance duality theorem, which delineates certain probabilistic invariances present in the source data as a sound and complete criterion for generalizable classification. Our findings provide a unifying theoretical perspective over several existing approaches to the domain generalization problem.

1 Introduction

Generalizing findings across settings is central throughout human experience. The discussion about the conditions under which induction can be formally justified can be traced back at least to Scottish philosopher David Hume circa the 18th century. Hume acknowledged that humans perform inferences from observed and particular experiences to more general and unobserved situations, but disputed its rational basis (Hume 1739). This challenge is called the *problem of induction* (Henderson 2018), and have puzzled generations of philosophers and mathematicians, from Kant to Popper, Goodman to Russell (Popper 1971, 1953; Russell 1912; Watkins et al. 2005).

The generalization problem plays a fundamental role in artificial intelligence and machine learning as well (Mitchell 1997; Russell and Norvig 2010), where it appears in different forms. For instance, one of the most well-studied tasks in the field is classification, where one tries to predict the label and generalize from something observed and specific (e.g., finite samples) to something unobserved and general (e.g., a probability distribution, a classifier). Tom Mitchell, one of the precursors of the field, noted (Mitchell 1997, p. 44): "a fundamental property of inductive inference: a learner that makes no a priori assumptions regarding the identity of the target concept has no rational basis for classifying any unseen instances.", refining Hume's observation. The question then becomes how to link the data collected from the distribution to the distribution itself. One of the fundamental approaches in the field is known as *empirical risk minimization*, due to Vapnik, which tied the risk between hypothetical and empirical distributions under some very general conditions (Vapnik 1991, 1998).

Despite the power of these ensuing results, we note that, in practice, the domains where the data is collected (called sources) are related to, but not necessarily the same as the one where the predictions are intended (target), violating a key assumption underlying many of the prior results. In fact, if the target domain is arbitrary, or drastically different from the source domains, no learning could take place (David et al. 2010; Bareinboim et al. 2022). However, the fact that we generalize and adapt relatively well to a new domain suggest that certain domains share common characteristics and that, owing to these commonalities, statistical claims can be generalized even to domains where no data is available (Pearl 2000; Spirtes, Glymour, and Scheines 2000; Bareinboim and Pearl 2016). How could one described the shared features across domain that allow this inferential leap? The anchors of knowledge that allow generalization to take place are eminently causal, following from the stability of the mechanisms shared across settings (Aldrich 1989). [1] The systematic analysis of these mechanisms and the conditions under which generalizations could be formally justified has been studied in the causal inference literature under the rubric of *transportability theory* (Bareinboim and Pearl 2014, 2016; Bareinboim et al. 2013; Pearl and Bareinboim 2011; Correa and Bareinboim 2020, 2019; Lee, Correa, and Bareinboim 2020).

In modern machine learning literature, the challenge of predicting in an unseen target domain is acknowledged and

[1] While arguing in response to Hume's skepticism, Kant noted that some *a priori* knowledge of concepts such as causation could be available before the inductive step (Kant 1781); for further discussion on this point, refer to (De Pierris and Friedman 2018).

broadly referred to as domain generalization problem. In this task, we have access to labeled data from the source domains, while no data in the target domain is available (Gulrajani and Lopez-Paz 2020). The theoretical proposals in this area rely on assumptions to define the target domains compatible with the source data such as the covariate shift assumption (Sugiyama and Müller 2005; Subbaswamy, Schulam, and Saria 2019; Subbaswamy and Saria 2020; Xu et al. 2021), or use of distance measures to relate the source and target distributions (Ben-David et al. 2006; Hanneke and Kpotufe 2019). Even under restrictive assumptions tying the source and target distributions, generalizing to the target domain might still be impossible (David et al. 2010). Another line of work takes into account the fact that the source and target domains are linked through the shared causal mechanisms, as alluded to earlier, which might entail probabilistic criteria that relates aspects of the source and target distributions. The invariance-based approaches then view the probabilistic invariances across the source data as proxies to the causal invariances across the source and target domains (Magliacane et al. 2018; Rojas-Carulla et al. 2018; Arjovsky et al. 2019; Rothenhäusler et al. 2021; Wald et al. 2021; Chen and Bühlmann 2021; Lu et al. 2021). Theoretical guarantees provided for these methods are contingent on assumptions such as linearity, additivity, markovianity (i.e., no confounders), yet there exists subtleties that limit the effectiveness and practicality of these methods (Rosenfeld, Ravikumar, and Risteski 2021). Another important ingredient present in modern machine learning methods is the use of representations. Those methods extract useful information to feed into the learning algorithm, which is particularly useful in high-dimensional and unstructured tasks (Bengio, Courville, and Vincent 2013). It has been noted both theoretically and empirically that enforcing certain restrictions to the representation learning stage yields performance boost for the downstream prediction tasks (Ben-David et al. 2006; Ganin et al. 2016; Long et al. 2018; Li et al. 2018; Zhang, Gong, and Schoelkopf 2015; Zemel et al. 2013). In some work, causal features have been used in constructing representations while filtering out the spurious correlations that might be unstable across domains (Wang and Jordan 2021; Schölkopf et al. 2021; Mao et al. 2022; Krueger et al. 2021). Considering this background, we note that solving the domain generalization problem can be seen as a two-step process:

1. **Evaluation:** for a fixed a representation, approximate the distribution of the label conditional on the representation in the target domain.
2. **Search:** find a representation that achieves maximal accuracy by using an evaluation method as a subroutine to assess the accuracy.

The above breakdown is natural, and is followed in several theoretical works on domain generalization. For instance, in the work by Ben-David et al. (2006), Theorem 1 provides an upper-bound to the risk in the target domain (step 1), and next, the authors treat this upper-bound as a proxy for the actual risk. They then propose an optimization procedure for finding the representation that minimizes it (step 2).

In this paper, we study the evaluation step through transportability lenses. In particular, we analyze the fundamental interplay between causal knowledge and the representation. For instance, we refute through formal analysis the belief that causal features are always desirable while spurious ones should be discarded. Moreover, throughout this paper we treat the true distributions as proxies for having access to samples drawn from them, thus, the challenges tied with finite/small sample size are considered outside the scope of this work. Our contributions are as follows:

- (Section 2) We formalize the domain generalization task by introducing the notion of transportable representations (Def. 5), and we develop a procedure to decide whether a representation is transportable given the structural assumptions encoded by a graph (Thm. 1). We demonstrate finite-sample performance of a transportability-based classifier via synthetic experiments, and show its superiority to vanilla ERM and the invariance-based method.
- (Section 3) We prove that the so-called invariance property, i.e., when the distribution of label conditioned on the representation is invariant across the source domains, is in fact a sound and complete criterion for transportability once we relax the assumption of having access to the graphs (Thm. 2). Also, this result provides a dual view on the graphical-invariance dichotomy, which highlights under what set of assumptions they coincide, and what are the limitations of operating graph-free.

Preliminaries. We use upper-case letters (e.g. $\mathbf{X}$ or Z) to denote random variables; The regular letter is used for univariate random variables, bold letter is used for multivariate ones. Support of random variables $\mathbf{Z}$ is denoted as $\text{supp}(\mathbf{Z})$, and values in the support are denoted by the corresponding lowercase letter, e.g., $\mathbf{z} \in \text{supp}(\mathbf{Z})$. To denote $P(\mathbf{A} = \mathbf{a} \mid \mathbf{B} = \mathbf{b})$, we use the shorthand $P(\mathbf{a} \mid \mathbf{b})$. The notion $\perp\!\!\!\perp_d$ denotes d-separation in graphs.

We use semantics of Structural Causal Models (Pearl 2000), 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 $\mathcal{M}$ 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 that are jointly independent; $\mathbf{V}$ is a set of endogenous (observed) variables; $\mathcal{F}$ represents a collection of functions $\mathcal{F} = \{f_V\}$ such that each endogenous variable $V \in \mathbf{V}$ is determined by a function $f_V \in \mathcal{F}$, where $f_V : \text{supp}(\mathbf{U}_V) \times \text{supp}(\mathbf{Pa}_V) \to \text{supp}(V)$ with $\mathbf{U}_V \subseteq \mathbf{U}$, and $\mathbf{Pa}_V \subseteq \mathbf{V} \setminus \{V\}$; The uncertainty is encoded through a distribution over the exogenous variables, $P(\mathbf{u})$.*

Every SCM $\mathcal{M}$ induces a causal diagram, which is a directed acyclic graph where any variable $V \in \mathbf{V}$ is a vertex, and there exists a directed edge from every variable in Pa_V to V. Also, for every pair $V, V' \in \mathbf{V}$ such that $\mathbf{U}_V \cap \mathbf{U}_{V'} \neq \emptyset$, there exists a bidirected edge between V and V'. We denote this causal diagram with the letter $\mathcal{G}$, and we say $\mathcal{M}$ is compatible with $\mathcal{G}$ if $\mathcal{M}$ induces $\mathcal{G}$. A SCM $\mathcal{M}$ entails a probability distribution $P^{\mathcal{M}}(\mathbf{v})$ over the set of observed

variables $\mathbf{V}$ such that $P^{\mathcal{M}}(\mathbf{v}) = \int_{\text{supp}(\mathbf{U})} \prod_{V \in \mathbf{V}} P^{\mathcal{M}}(v \mid \mathbf{pa}_V, \mathbf{u}_V) \cdot P(\mathbf{u}) \cdot d\mathbf{u}$,, where each term $P(v \mid \mathbf{pa}_V, \mathbf{u}_V)$ corresponds to the function $f_V \in \mathcal{F}$ in the underlying structural causal model $\mathcal{M}$. Throughout this paper, we assume the observational distributions entailed by the SCMs satisfy positivity, that is, $P^{\mathcal{M}}(\mathbf{v}) > 0$, for every $\mathbf{v}$. We will also operate non-parametrically, i.e., making no assumption about the particular functional form or the distribution of the unobserved variables.

2 Transportability of Representations

We study a system of endogenous variables $\mathbf{X} \cup \{Y\}$, where Y is a binary label. SCMs $\mathcal{M}^1, \mathcal{M}^2, \ldots, \mathcal{M}^T$ defined over $\mathbf{X} \cup \{Y\}$ denote the source domains, and entail the distributions $\mathbb{P} = \{P^1, P^2, \ldots, P^T\}$, while they induce the causal diagrams $\mathcal{G}^1, \mathcal{G}^2, \ldots, \mathcal{G}^T$. Also, an unknown SCM $\mathcal{M}^*$ represents the target domain, which entails the distribution P^*, and induces the causal diagram $\mathcal{G}^*$. The following example elaborates the concepts through a well-attended instance of the problem.

Example 1 (Covariate shift) Let $\mathbf{X} := \langle X_1, X_2, \ldots, X_N \rangle$ be a vector of binary variables, and Y be a binary label. We observe data from two source domains $\mathcal{M}^1, \mathcal{M}^2$, and the task is predicting Y based on $\mathbf{X}$ in the target domain $\mathcal{M}^*$. What follows describes the SCM $\mathcal{M}^i$ ($i \in \{1, 2, *\}$):

$$U_{\mathbf{X}} \sim P^i(u_{\mathbf{X}}) \tag{1}$$
$$U_Y \sim \text{Unif}([0, 1]) \tag{2}$$
$$\mathbf{X} \leftarrow U_{\mathbf{X}} \tag{3}$$
$$Y \leftarrow \begin{cases} 1 & \text{if } U_Y \geq \sigma(\alpha^\top \cdot \mathbf{X}) \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

As seen above, the label Y in all domains $\mathcal{M}^1, \mathcal{M}^2, \mathcal{M}^*$ follows the conditional distribution,

$$Y \mid \mathbf{X} \sim \text{Bernoulli}(\sigma(\alpha^\top \cdot \mathbf{X})). \tag{5}$$

Where α is a N-vector of coefficients, and σ is the sigmoid function defined as $\sigma(x) = \frac{1}{1+e^{-x}}$. In words, the distribution of $\mathbf{X}$ changes across the source and target domains, while the odds of Y is determined by $\mathbf{X}$, and is equal to a linear combination of entries in $\mathbf{X}$.

Based on this definition, the causal diagrams $\mathcal{G}^1, \mathcal{G}^2, \mathcal{G}^*$ coincide, and are depicted in Figure 1a. This setup is commonly referred to in the literature as the covariate shift (Sugiyama and Müller 2005; Subbaswamy, Schulam, and Saria 2019; Subbaswamy and Saria 2020), under which ones assumes that $\mathbb{E}[Y \mid \mathbf{x}]$ is invariant across the source and target domains, while the distribution of the covariates $P(\mathbf{x})$ might vary. □

To describe the mismatch of mechanisms between two SCMs, we adapt the following notion introduced in (Lee, Correa, and Bareinboim 2020).

Definition 2 (Domain discrepancy) *For every pair of SCMs M^i, M^j ($i, j \in \{*, 1, 2, \ldots, T\}$) defined over $\mathbf{X} \cup \{Y\}$, the domain discrepancy set $\Delta_{ij} \subseteq \mathbf{V}$ is defined such that for every $V \in \Delta_{ij}$ there might exist:*

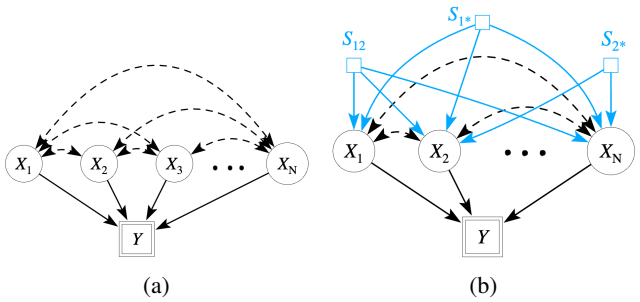

Figure 1: (a) The causal diagram induced by the source and targets SCMs in Example 1. All covariates might be confounded, which is indicated by a bidirected arrow between every pair of them. (b) Selection diagram of Example 1; the mechanism determining the covariates might vary across the source and target domains, so the selection nodes are connected to all the covariates, while the mechanism determining the label is the same across all the domains, thus no selection node is connected to the label.

1. *a discrepancy between $f_V^{M^i} \neq f_V^{M^j}$, or,*
2. $P^{M^i}(\mathbf{u}_V) \neq P^{M^j}(\mathbf{u}_V)$. □

In other words, $V \notin \Delta_{ij}$ is equivalent to assuming that the mechanisms for V across M^i, M^j are structurally invariant, i.e., $f_V^{M^i} = f_V^{M^j}$ and $P^{M^i}(\mathbf{u}_V) = P^{M^j}(\mathbf{u}_V)$. We introduce next a version of selection diagrams (Lee, Correa, and Bareinboim 2020) to graphically represent the system that includes multiple SCMs relative to the collection of domains.

Definition 3 (Selection diagram) *The selection diagram $\mathcal{G}^{\Delta_{ij}}$ is constructed from $\mathcal{G}^i$ ($i \in \{*, 1, 2, \ldots, T\}$) by adding the selection node S_{ij} to the vertex set, and adding the edge $S_{ij} \rightarrow V$ for every $V \in \Delta_{ij}$. The collection $\mathcal{G}^{\boldsymbol{\Delta}} = \{\mathcal{G}^{\Delta_{ij}}\}_{i,j \in \{*,1,2,\ldots,T\}}$ encodes the graphical assumptions. Whenever the causal diagram is shared across the domains, a single diagram can be utilized to depict $\mathcal{G}^{\boldsymbol{\Delta}}$.* □

In words, selection diagrams are parsimonious graphical expressions of the commonalities and disparities across the domains, which can be seen as grounding Kant's observation alluded to earlier.

Example 1 (Covariate shift: continued) Following Definition 2, the domain discrepancy sets are $\Delta_{12} = \Delta_{1*} = \Delta_{2*} = \mathbf{X}$. Thus, the in the induced selection diagram shown in Figure 1b where the selection nodes S_{12}, S_{*1}, S_{*2} are pointing to $\mathbf{X}$ nodes.

The goal of the domain generalization task is to predict Y by observing $\mathbf{X}$ in the target domain $\mathcal{M}^*$. We consider an aggregation of the information in $\mathbf{X}$ that we call a representation, for instance:

$$R_{\text{sum}} = \phi_{\text{sum}}(\mathbf{X}) := \sum_{i=1}^{N} X_i, \tag{6}$$

We then may try to predict Y based on the value of R. To do so, a natural objective is to compute the quantity

$\mathbb{E}_{P^*}[Y \mid R_{\text{sum}} = r]$ for every $r \in \{0, 1, \ldots, N\}$ using the source data collected from source domains $\mathcal{M}^1, \mathcal{M}^2$. This objective is well-defined only if this quantity is unique under all distributions P^* that can be entailed by some SCM that can possibly govern the target domain. In causal inference literature, this concept is known as transportability. □

As in the example above, we are interested in quantities that involve a variable, such as R computed according to Eq. 6. What follows is a formal definition of representations and their score function.

Definition 4 (Representation and scores) *The variable* $\mathbf{R}$ *with support* $\text{supp}(\mathbf{R})$ *is said to be a representation (of* $\mathbf{X}$*) if there exists a mapping* $\phi : \text{supp}(\mathbf{X}) \to \text{supp}(\mathbf{R})$ *such that* $\mathbf{R} = \phi(\mathbf{X})$*. The corresponding score function is defined as*

$$l_\phi(\mathbf{r}) := \mathbb{E}_{P^*}[Y \mid \mathbf{R} = \mathbf{r}]. \tag{7}$$

For source distributions $P^i \in \mathbb{P}$*, the quantity* $\mathbb{E}_{P^i}[Y \mid \mathbf{R} = \mathbf{r}]$ *is called an empirical score function.* □

A representation is an aggregation of the information in $\mathbf{X}$. For example, when $\mathbf{X}$ is a binary vector, the representation can be the number of ones in this vector, as illustrated in Example 1. In a special case, one might discard certain entries of $\mathbf{X}$ while keeping the rest; for instance $\phi(X_1, X_2, X_3) = \langle X_1, X_3 \rangle$. The latter is well-attended in the causal inference literature, as the causal queries are usually denoted by $P(y \mid \mathbf{z})$ where $\mathbf{Z} \subseteq \mathbf{X}$. By using a representation $\mathbf{R} = \phi(\mathbf{X})$, we can express a much larger class of queries of the form $P(y \mid \phi(\mathbf{X}) = \mathbf{r})$, where the mapping $\mathbf{R} = \phi(\mathbf{X})$ is arbitrary but known. Throughout this work, we consider representations that satisfy the coverage of property, that is, $P^i(\mathbf{r}) > 0$ for every $P^i \in \mathbb{P}$ and $\mathbf{r} \in \text{supp}(\mathbf{R})$. Inherently, this property is testable using the data.

Motivated by Example 1, the main objective of this paper is to compute the score function of a given representation using the source data, and guarantee its validity given graphical (sec. 2) or algebraic (sec. 3) assumptions. To this end, we extend the notion of transportability (Bareinboim et al. 2013) to study queries involving representations.

Definition 5 (Transportable representation) *The representation* $\mathbf{R} = \phi(\mathbf{X})$ *is called transportable if its score function* $l_\phi(\mathbf{r}) = \mathbb{E}_{P^*}[Y \mid \mathbf{R} = \mathbf{r}]$ *can be uniquely computed from the source distributions* $\mathbb{P}$*, considering:*

1. *The assumption encoded in the selection diagrams* $\mathcal{G}^{\mathbf{\Delta}}$*,*
2. *The arithmetic expression for* ϕ*.* □

For the representations that are feature selection, such as $\phi(X_1, X_2, X_3) = \langle X_1, X_3 \rangle$, the definition above coincides with the notion of statistical transportability (Correa and Bareinboim 2019). If a representation ϕ is not transportable, then its score function l_ϕ is not unique across possible target domains, and therefore, it is not possible to compute it regardless of statistical challenges of estimation. Thus, the task of computing the score function of a given representation is only well-defined if that representation is transportable. In the rest of this paper, we focus on characterizing transportable representations, and develop a method to compute an expression for the score functions in terms of the available source distributions. This expression is a blueprint for estimation, but some of the subtleties of using finite samples is outside the scope of our paper. Below, we discuss a well-attended instance of the domain generalization problem.

Example 2 (Covariate shift: not TR case) In the context of Example 1, we argue that for the representation $R_{\text{sum}} = \phi_{\text{sum}}(\mathbf{X})$ in Eq. 6, the score function $l_{\phi_{\text{sum}}}(r) = \mathbb{E}_{P^*}[Y \mid R_{\text{sum}} = r]$ is not unique for all compatible target domains. Consider,

$$l_{\phi_{\text{sum}}}(1) = \mathbb{E}_{P^*}[Y \mid \sum_{i=1}^{N} X_i = 1] \tag{8}$$

$$= \sum_{i=1}^{N} \sigma(\alpha_i) \cdot P^*(\mathbf{X} \text{ one hot } i \mid R_{\text{sum}} = 1). \tag{9}$$

The last expression indicates that at $r = 1$ the score function might vary for different choices of $P^*(\mathbf{x})$. In case of the covariate shift example, the distribution of covariates in the target domain, namely $P^*(\mathbf{x})$, can be any arbitrary positive distribution. Thus, the terms $P^*(\mathbf{X} = \text{one hot at } i \mid R_{\text{sum}} = 1)$ in Eq. 9 can be any vector of positive values that sum to one. This fact indicates that, in extreme cases, $l_{\phi_{\text{sum}}}(1)$ can lie just below the maximum of $\{\sigma(\alpha_i)\}_i$ or just above the minimum of them. Precisely speaking, for every

$$c \in (\min_{1 \le i \le N} \sigma(\alpha_i), \max_{1 \le i \le N} \sigma(\alpha_i)), \tag{10}$$

there exists a plausible target SCM $\mathcal{M}_c^*$ that is compatible with the selection diagram in Figure 1b and the source distributions P^1, P^2, such that $\mathbb{E}_{P^{\mathcal{M}_c^*}}[Y \mid R_{\text{sum}} = 1] = c$. As long as the coefficients α_i are not all equal, the interval in Eq. 10 contains more than one value, and therefore, we can not assure that the score function is unique across all compatible target domains, let alone compute it. We conclude that ϕ_{sum} is not transportable in this example. □

The above example carries an important message despite its simplicity; in settings that involve representations, computing or estimating the score function might be impossible, even when the covariate shift assumption can be ascertained and access to true distributions is given (instead of finite samples).

We introduce a graphical criterion that is useful to evaluate the probabilistic invariances in the distribution motivated by (Pearl and Bareinboim 2011).

Definition 6 (S-Admissibility) *Consider the domains* $\mathcal{M}^i, \mathcal{M}^j$ *(*$i, j \in \{*, 1, 2, \ldots, T\}$*), and sets of variables* $\mathbf{Z}, \mathbf{A} \subset \mathbf{X} \cup \{Y\}$*.* $\mathbf{A}$ *is said to be* S*-admissible given* $\mathbf{Z}$ *w.r.t. the domains* $\mathcal{M}^i, \mathcal{M}^j$ *whenever* $\mathbf{A}$ *is d-separated from* S_{*i} *given* $\mathbf{Z}$ *in* $\mathcal{G}^{\Delta_{ij}}$*. In that case,*

$$\mathbf{A} \perp\!\!\!\perp_d S_{ij} \mid \mathbf{Z} \text{ in } \mathcal{G}^{\Delta_{ij}} \implies P^i(\mathbf{a} \mid \mathbf{z}) = P^j(\mathbf{a} \mid \mathbf{z}). \tag{11}$$

In words, the s-admissibility criterion enables us to read the probabilistic invariances across domains by evaluating the d-separation relations in the selection diagram. Next, we elaborate through an example the use of the S-admissibility criterion for deciding if a representation is transportable by computing its score function.

Example 3 (Covariate shift: TR case) In the context of Example 1, consider the representation

$$R_{\text{rand}} = \phi_{\text{rand}}(\mathbf{X}) := \beta^\top \cdot \mathbf{x}, \tag{12}$$

where $\beta \sim \mathcal{N}(0, I)$ is drawn independently. Considering the expression above, we can almost surely compute an inverse function $\mathbf{X} = \phi^{-1}_{\text{rand}}(R_{\text{rand}})$ (proof in Appendix A); Thus, we can rewrite the score function $l_{\phi_{\text{rand}}}(r)$ as,

$$\mathbb{E}_{P^*}[Y \mid R_{\text{rand}} = r] = \mathbb{E}_{P^*}[Y \mid \mathbf{X} = \phi^{-1}_{\text{rand}}(r)]. \tag{13}$$

Licensed by the s-admissibility relation $Y \perp\!\!\!\perp_d S_{ij} \mid \mathbf{X}$ readable from the selection diagram in Figure 1b, the latter can be directly obtained from either of the source distributions, namely,

$$\mathbb{E}_{P^*}[Y \mid \mathbf{X} = \phi^{-1}_{\text{rand}}(r)] = \mathbb{E}_{P^{1,2}}[Y \mid \mathbf{X} = \phi^{-1}_{\text{rand}}(r)]. \tag{14}$$

Thus, we conclude that ϕ_{sum} is transportable in this example, because,

$$\begin{aligned}\mathbb{E}_{P^*}[Y \mid \phi_{\text{rand}}(\mathbf{X}) = r] &= \mathbb{E}_{P^*}[Y \mid \mathbf{X} = \phi^{-1}_{\text{rand}}(r)] &(15)\\ &= \mathbb{E}_{P^{1,2}}[Y \mid \mathbf{X} = \phi^{-1}_{\text{rand}}(r)] &(16)\\ &= \mathbb{E}_{P^{1,2}}[Y \mid \phi_{\text{rand}}(\mathbf{X}) = r]. &(17)\end{aligned}$$

As seen in Examples 2 and 3, transportability of representations depends on the expression of the representation; due to determinism of the mapping ϕ_{rand}, the condition $\phi(\mathbf{X}) = \mathbf{r}$ could be translated to a condition about the variables $\mathbf{X}$, which enabled us to use s-admissibility relations effectively in transporting the score function. The next definition extends the results by Geiger (1990) regarding deterministic relations in SCMs.

Definition 7 (Det., cons., and free) *Consider a representation $\mathbf{R} = \phi(\mathbf{X})$. Variables $\det(\phi) = \mathbf{Z} \subseteq \mathbf{X}$ are determined by ϕ if for every value $\mathbf{r} \in \text{supp}(\mathbf{R})$ a single value for $\mathbf{Z}$ can be derived from $\phi(\mathbf{X}) = \mathbf{r}$. The variables $\text{cons}(\phi) = \bar{\mathbf{Z}} \subseteq \mathbf{X} \setminus \mathbf{Z}$ are constrained by ϕ if they are not determined by ϕ, and for at least one value $\mathbf{r} \in \text{supp}(\mathbf{R})$ and at least one value $\bar{\mathbf{z}} \in \text{supp}(\bar{\mathbf{Z}})$, the system of equations $\phi(\mathbf{X} \setminus \bar{\mathbf{Z}}, \bar{\mathbf{z}}) = \mathbf{r}$ is inconsistent. The variables $\text{free}(\phi) = \mathbf{X} \setminus (\mathbf{Z} \cup \bar{\mathbf{Z}})$ do not depend on the representation, and are called free from ϕ.* □

The notions introduced in this definition are properties of the mapping ϕ, and in principle, they can be decided given an arithmetic expression for ϕ. An example in Appendix E elaborates on this definition. Next, we augment the selection diagrams to incorporate the knowledge of the representations into the symbolic inference pipeline.

Definition 8 (Augmented selection diagrams) *Let $\mathcal{G}^{\mathbf{\Delta}}$ be a selection diagram over the variables $\mathbf{X}, Y$, the set of distributions $\mathbb{P}$ be the source distributions, and $\phi(\mathbf{X})$ be a representation. Let $\mathbf{Z} = \det(\phi)$ denote the variables determined by ϕ, and $\bar{\mathbf{Z}} = \text{cons}(\phi)$ denote the variables constrained by ϕ, and let the equation $\bar{R} = \bar{\phi}(\mathbf{Z})$ obtained from $\mathbf{R} = \phi(\mathbf{X})$ specify the constraints. We construct the augmented selection diagrams and corresponding distributions*

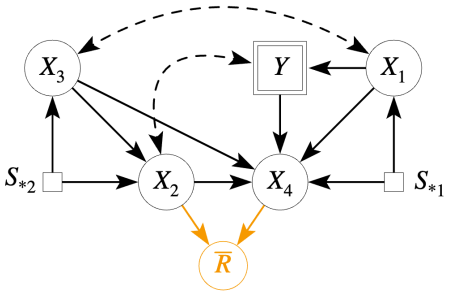

Figure 2: selection diagram corresponding to Example 4 (black). The node S_{12} is removed to avoid clutter, and is connected to all variables $\mathbf{X}$. The naively augmented (with orange) and augmented (with blue) selection diagrams.

by adding the variable $\bar{\mathbf{R}}$ to $\mathcal{G}^{\mathbf{\Delta}}, \mathbb{P}$ as follows:

$$\mathcal{G}^{\mathbf{\Delta}}_{\text{aug}}: \textit{ add node } \bar{\mathbf{R}} \textit{ to all graphs in } \mathcal{G}^{\mathbf{\Delta}} \textit{ with arrows from } \bar{\mathbf{Z}} \textit{ nodes to } \bar{\mathbf{R}} \tag{18}$$

$$\mathbb{P}_{\text{aug}}: \{P^i_{\text{aug}} := P^i(\mathbf{x}, y) \cdot 1_{\{\bar{\mathbf{r}} = \bar{\phi}(\bar{\mathbf{z}})\}}, \textit{ for } P^i \in \mathbb{P}\}. \tag{19}$$

An example in Appendix E elaborates on the Definition above. The following result allows us to evaluate queries involving representations, an in particular, to decide transportability of representation.

Theorem 1 (Graphical transportability) *Consider a representation $\mathbf{R} = \phi(\mathbf{X})$, and the value $\mathbf{r}$ for it. Let $\mathbf{Z} = \det(\phi)$ where $\mathbf{z}$ is its value, and $\bar{\mathbf{Z}} = \text{cons}(\phi)$ denotes the constrained variables where the equation $\bar{\mathbf{r}} = \bar{\phi}(\bar{\mathbf{Z}})$ specifies the constraints on them. The score function can be expressed as,*

$$l_\phi(\mathbf{r}) = \mathbb{E}_{P^*}[Y \mid \mathbf{r}] = P^*(Y = 1 \mid \mathbf{Z} = \mathbf{z}, \bar{\mathbf{R}} = \bar{\mathbf{r}}). \tag{20}$$

The representation ϕ is a transportable (i.e., l_ϕ can be computed in terms of $\mathbb{P}$) given $\langle \mathcal{G}^{\mathbf{\Delta}}, \phi \rangle$ if the equivalent query $P^*(y \mid \mathbf{z}, \bar{\mathbf{r}})$ is transportable from the augmented source distributions $\mathbb{P}_{\text{aug}}$ given the augmented selection diagrams $\mathcal{G}^{\mathbf{\Delta}}_{\text{aug}}$ via the gTR algorithm by Lee, Correa, and Bareinboim (2020). □

All proofs are provided in Appendix A. We elaborate on Theorem 1 through the following example.

Example 4 (Theorem 1 illustrated) Consider the selection diagram $\mathcal{G}^{\mathbf{\Delta}}$ in Figure 2, over the variables Y and $\mathbf{X} = \{X_1, X_2, X_3, X_4\}$ with $\text{supp}(X_i) \subset \mathbb{N}$. There exists two source domains, and the goal is to compute the score function $Q: l_\phi(\mathbf{r}) = \mathbb{E}_{P^*}[Y \mid \mathbf{r}]$ for the representation,

$$\mathbf{R} = \phi(\mathbf{X}) := \langle \underbrace{\frac{X_1 \cdot X_2 \cdot X_3}{X_4}}_{R_1}, \underbrace{\frac{X_1 \cdot X_2}{X_3 \cdot X_4}}_{R_2}, \underbrace{\frac{X_2 \cdot X_3}{X_1 \cdot X_4}}_{R_3} \rangle. \tag{21}$$

The expression for the representation ϕ can be viewed as a system of equations which allows us to express $\mathbf{X}$ in terms of $\mathbf{R}$;

$$X_1 = \sqrt{\frac{R_1}{R_3}}, \quad X_3 = \sqrt{\frac{R_1}{R_2}}, \quad \frac{X_2}{X_4} = \sqrt{R_2 \cdot R_3}. \tag{22}$$

More elaboration on the equation solving procedure is provided in Appendix B. In words, the variables X_1, X_3 are determined by $\mathbf{R}$, as they attain a unique value for every realization $\mathbf{R} = \mathbf{r}$ through the first two equation above. On the other hand, X_2, X_4 cannot be uniquely determined by $\mathbf{R}$, even though they are constrained by the value of $\mathbf{R}$ through the last equation. Thus, by definition 7, $\det(\phi) = \{X_1, X_3\}$ and $\text{cons}(\phi) = \{X_2, X_4\}$.

Define $\bar{R} \leftarrow \frac{X_2}{X_4}$ as a new variable in the SCM with $\mathbf{Pa}_{\bar{R}} = \{X_2, X_4\}$; the augmentation of the selection diagram with this new variable is depicted in orange in Figure 2. For a fixed value $\mathbf{r} \in \text{supp}(\mathbf{R})$ let,

$$x_1^{\mathbf{r}} := \sqrt{\frac{r_1}{r_3}}, x_3^{\mathbf{r}} = \sqrt{\frac{r_1}{r_2}}, \bar{r}^{\mathbf{r}} = \sqrt{r_2 \cdot r_3} \tag{23}$$

be the values for $X_1, X_3, \bar{R}$, respectively. Next, we can use this change of variables to rewrite the score function Q as follows:

$$l_\phi(\mathbf{r}) = \mathbb{E}_{P^*}[Y \mid \mathbf{R} = \mathbf{r}] = \underbrace{P^*(Y = 1 \mid x_1^{\mathbf{r}}, x_3^{\mathbf{r}}, \bar{r}^{\mathbf{r}})}_{Q'}. \tag{24}$$

Now, we attempt to transport the query Q' from the source distributions given the augmented selection diagram 2. We follow the gTR algorithm (Correa and Bareinboim 2020):

$$Q' = P^*(y \mid \bar{r}^{\mathbf{r}}, x_1^{\mathbf{r}}, x_3^{\mathbf{r}}) \tag{25}$$

$$= \frac{\sum_{x_2,x_4} P^*(y, \bar{r}^{\mathbf{r}}, x_2, x_4 \mid, x_1^{\mathbf{r}}, x_3^{\mathbf{r}})}{\sum_{y,x_2,x_4} \underbrace{P^*(y, \bar{r}^{\mathbf{r}}, x_2, x_4 \mid, x_1^{\mathbf{r}}, x_3^{\mathbf{r}})}_{Q''}}. \tag{26}$$

We factorize the query Q'' as,

$$Q'' = P^*(\bar{r}^{\mathbf{r}} \mid y, x_1^{\mathbf{r}}, x_2, x_3^{\mathbf{r}}, x_4) \cdot P^*(y, x_2, x_4 \mid x_1^{\mathbf{r}}, x_3^{\mathbf{r}})$$

$$= P^*(\bar{r}^{\mathbf{r}} \mid x_2, x_4) \cdot P^*(y, x_2, x_4 \mid x_1^{\mathbf{r}}, x_3^{\mathbf{r}}) \tag{27}$$

$$= 1_{\{\bar{r}^{\mathbf{r}} = \frac{x_2}{x_4}\}} \cdot \underbrace{P^*(y, x_2, x_4 \mid x_1^{\mathbf{r}}, x_3^{\mathbf{r}})}_{Q'''}, \tag{28}$$

where Eq. 27 is due to $\bar{R} \perp\!\!\!\perp_d X_1, X_3, Y \mid X_2, X_4$ that is readable from the selection diagram in Figure 2, and Eq. 28 is due to the construction (Eq.19). Finally, we transport Q''' as,

$$Q''' = P^*(y, x_2 \mid x_1^{\mathbf{r}}, x_3^{\mathbf{r}}) \cdot P^*(x_4 \mid y, x_2, x_1^{\mathbf{r}}, x_3^{\mathbf{r}}) \tag{29}$$

$$= P^1(y, x_2 \mid x_1^{\mathbf{r}}, x_3^{\mathbf{r}}) \cdot P^2(x_4 \mid y, x_2, x_1^{\mathbf{r}}, x_3^{\mathbf{r}}), \tag{30}$$

where Eq. 30 is licensed by the S-admissibility relations,

$$Y, X_2 \perp\!\!\!\perp_d S_{1*} \mid X_1, X_3 \tag{31}$$

$$X_4 \perp\!\!\!\perp_d S_{2*} \mid Y, X_1, X_2, X_3. \tag{32}$$

The derivation above allows us to express the score function solely in terms of the source distributions, which means that the representation ϕ specified in Eq. 21 is indeed transportable (Def. 5). □

Note that the covariate shift assumption does not hold in Example 4; for instance, $S_{*1} \rightarrow X_4 \leftrightarrow Y$ and $S_{*2} \rightarrow X_2 \leftarrow Y$ are d-connecting paths between the s-nodes and Y conditional on $\mathbf{X}$. Therefore, $\mathbb{E}[Y \mid \mathbf{r}]$ varies across the domains in generic instances that admit this model (Figure 2). One might limit the scope of covariate shift assumption to X_1 and argue that $\mathbb{E}[Y \mid x_1]$ is invariant across the source and target domains. However, notice that the representation $\mathbf{R}$ is richer than the covariate X_1 alone, i.e., the σ-algebra generated by $\mathbf{R}$ is larger than the one generated by X_1, because X_1 is determined by $\mathbf{R}$ (Eq. 22). Therefore, $\mathbf{R}$ has higher predictive power (for prediction of Y) compared to X_1, even though X_1 is "causal" to Y and the rest of the covariates are not. This observation indicates that predictions based on causal features are not necessarily superior to the predictions based on non-causal features, as the transportability machinery might license the use of some non-causal features for better classification accuracy.

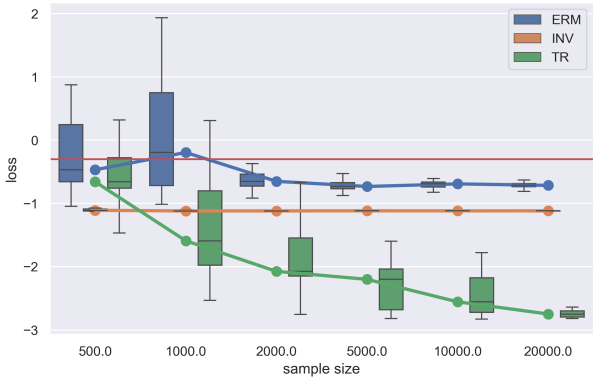

Figure 3: The metric is cross-entropy under the target distribution P^* (the lower, the better). The horizontal axis is the sample size from both source domains.

2.1 Experiments

Appendix D contains a detailed discussion on estimating the score function in Example 4. We also implemented the described estimator, and the results are depicted in Figure 3 for a set of randomly generated SCMs $\mathcal{M}^1, \mathcal{M}^2, \mathcal{M}^*$. The red line indicates the loss for a random guess. ERM stands for empirical risk minimization (cobalt), and it simply regresses Y on $\mathbf{R}$ using the pooled data. Despite the popularity of ERM, we see that due to mismatch between the target domain and the sources ERM performs only slightly better than a random guess, and the performance does not improve for larger data. INV (orange) regresses Y on the best invariant representation (that is X_1) using pooled data; this classifier is in fact what existing work on invariance-based domain generalization suggests. INV has a better performance compared to ERM, which is consistent with our theoretical guarantees, however, the transportability-based classifier (green) performed significantly better for larger data. In implementation of TR, we used rejection sampling for the generative models, and for the likelihood models we used random forests, and it

This experiment shows that transportability might allow us to make a generalizable prediction superior to the so-called *causal prediction*. However, having access to the expert knowledge encoded as selection diagrams is required.

In the next section, we remove this restriction and study an instance of the transportability problem that operates based on no explicit graphical expert knowledge.

3 Data-Driven Transportability

The discussions so far assume that the true selection diagram is available to us, however, it might be hard to obtain. In this section, we provide an alternative route for articulating assumptions and obtaining results on transportability. In particular, we will express the structural assumptions about the underlying mechanisms in terms of the data itself, in the absence of the selection diagrams $\mathcal{G}^{\mathbf{\Delta}}$. We impose a restriction to the structure of the selection diagrams $\mathcal{G}^{\mathbf{\Delta}}$ (Assumptions 1), and assert a correspondence between source data $\mathbb{P}$ and the selection diagram $\mathcal{G}^{\mathbf{\Delta}}$ (Assumption 2).

The means of inference in graphical setting are S-admissibility relations, so as we remove the assumption of having access to the selection diagrams, we need to establish a structural correspondence between the target domain and the source domains.

Assumption 1: Stability of Mechanisms (SoM). *The causal diagram is common across source and target domains, and for all variables $V \in \mathbf{X} \cup \{Y\}$,*

$$V \notin \bigcup_{i,j=1}^{T} \Delta_{ij} \implies V \notin \bigcup_{k=1}^{T} \Delta_{*k}. \tag{33}$$

What follows elaborates more on this assumption.

Example 5 (SoM illustrated) Suppose the source data contains pictures of cats and dogs in a room, and the task is to distinguish them. Two source datasets are collected in spring and summer. The target domain is on pictures of cats and dogs in the same room, but during fall season. Suppose, for simplicity, that the lighting condition in the room depends solely on the natural light coming from outside, e.g., $\mathrm{lightOn} \leftarrow \neg\mathrm{isSunny} \oplus U$. In the context of lighting, SoM assumption states that if the mechanism determining the lighting has been the same for both spring and summer datasets (sources), we assume that it is going to remain unchanged in fall (target). Notice that unchanged lighting mechanism does not translate to the same amount of total lighting, because that depends on the distribution of the parents of $\mathrm{lightOn}$ as well, including $\mathrm{isSunny}$, which most probably varies across the domains. SoM, on the other hand, asserts that if this mechanism has changed, e.g., the light was less likely to be on due to energy preservation during summer, then we shall not rely on stability of this mechanism anymore, and the lighting mechanism might change arbitrarily in the fall (target domain). We emphasize that the variables such as lighting are not explicitly available to the learner; by assuming SoM we impose such stability properties to all mechanisms (such as lighting) that generate the images through the unknown underlying SCMs. In words, SoM requires that all stables mechanisms that have generated the source data to remain stable in the target as well. □

As indicated through the example above, SoM imposes a structure to the selection diagrams as a whole, while it is not explicit about the different variables/mechanisms in the SCMs. Next, we define an adjusted notion of transportability according to stability of mechanisms assumption.

Definition 9 (Transportable represent.: Data-driven) *In a data-driven setup, the representation $\mathbf{R} = \phi(\mathbf{X})$ is called transportable if the score function $l_\phi(\mathbf{r}) = \mathbb{E}_{P^*}[Y \mid \mathbf{R} = \mathbf{r}]$ can be uniquely computed in terms of the source distributions $\mathbb{P}$, under the SoM assumption.* □

According to the definition above, when considering transportability in the data-driven setup, we consider the most conservative scenario. In other words, when we replace the graph $\mathcal{G}^{\mathbf{\Delta}}$ with SoM, we can only leverage it to reject the possibility of scenarios that are inconsistent with it. This conservative approach is similar to how we treat the selection diagrams in graphical transportability; in that setup, once an edge exists in a selection diagram it indicates the *possibility* of a cause-effect relationship, and our inference strategy must be correct for all possibilities that conform to the selection diagram.

For instance, even though the selection diagram $\mathcal{G}^{\mathbf{\Delta}}$ in Example 4 (Figure 2) satisfies SoM, it is not the only selection diagram over these variables that is valid under SoM. We can construct $\mathcal{G}_0^{\mathbf{\Delta}}$ by connecting all s-nodes to all covariates $\mathbf{X}$. Notice, $\mathcal{G}_0^{\mathbf{\Delta}}$ is still compatible with the true SCMs while it also satisfies SoM. The selection diagram $\mathcal{G}_0^{\mathbf{\Delta}}$ contains the edges present in $\mathcal{G}^{\mathbf{\Delta}}$, and therefore, is a weaker assumption. This is a critical point, because the representation in Eq. 21 is no longer transportable given $\mathcal{G}_0^{\mathbf{\Delta}}$. This observation highlights a limitation of the data-driven approach in comparison to the graphical approach discussed in section 2. The above example indicates that selection diagrams offer a more expressive language compared to SoM for describing the knowledge about the domains. Some readers might find it helpful to note that assuming SoM is weaker than assuming the selection diagrams, while SoM implies the covariate shift assumption. Thus, SoM lies between covariate shift and selection diagrams in terms of expressivity.

An important consequence of SoM is that each of the source domains is compatible as a possible target domain; below, this property is stated formally.

Lemma 1 (Interchangeable domains) *Under the stability of mechanisms assumption, an SCM that is identical to one of the source SCM $\mathcal{M}^T$ ($1 \leq t \leq T$) can be a compatible target domain, i.e., SoM does not preclude the possibility of $\mathcal{M}^*$ being identical to $\mathcal{M}^T$.* □

Lemma 1 uncovers a key limitation of the data-driven approach, as it indicates that SoM can not express a family of possible target domains while rejecting some of the source domains as a possible target domain. For instance, suppose some economical data is collected from the US and China, and the target domain is India. Even though there might be similarities between individual mechanism across these domains, it is unlikely that India's economy is totally identical to either of the US and China, i.e., domains are likely non-interchangeable. However, SoM does not rule out such possibility according to Lemma 1. In contrast, the graphical assumptions encoded in the selection diagrams might help

us express the experts' knowledge in finer granularity by allowing us to reject more possibilities for the target domain. On the other hand, in the context of Example 5, we have no reason to believe that the pictures taken in the same room during fall season (target) are generated by an SCM necessarily dissimilar to that of the spring and summer seasons. Therefore, the domains might be interchangeable in the reality of Example 5.

Due to Lemma 1, $\mathcal{M}^k$ $(1 \leq k \leq T)$ can be the target SCM, so the quantity $\mathbb{E}_{P^k}(Y \mid \mathbf{r})$ is a possible value that the score function $l_\phi(\mathbf{r})$ might attain. If the representation ϕ is transportable (Def. 9), i.e., $l_\phi(\mathbf{r})$ attains a unique value across all compatible target SCMs, then $l_\phi(\mathbf{r})$ must be identical to $\mathbb{E}_{P^k}[Y \mid \mathbf{r}]$ for all $1 \leq k \leq T$. What follows is a formal statement.

Corollary 1 (necessity of invariance) *Under stability of mechanisms assumption, if a representation $\mathbf{R} = \phi(\mathbf{X})$ is transportable (Def. 9), i.e., $\mathbb{E}_{P^*}[Y \mid \mathbf{r}]$ attains a unique value for all compatible target SCMs, then $\mathbb{E}[Y \mid \mathbf{r}]$ is invariant across the source and target domains, i.e.,*

$$\mathbb{E}_{P^1}[Y \mid \mathbf{r}] = \cdots = \mathbb{E}_{P^T}[Y \mid \mathbf{r}] = \mathbb{E}_{P^*}[Y \mid \mathbf{r}]. \tag{34}$$

This motivates the following definition.

Definition 10 (Invariance Property) *A representation $\mathbf{R} = \phi(\mathbf{X})$ is said to satisfy the invariance property w.r.t. the distributions P^i, P^j $(i, j \in \{*, 1, \ldots, T\})$ if,*

$$\mathrm{INV}_{ij}[\phi] : \mathbb{E}_{P^i}[Y \mid \mathbf{r}] = \mathbb{E}_{P^j}[Y \mid \mathbf{r}], \quad \forall \mathbf{r} \in \mathrm{supp}(\mathbf{R}). \tag{35}$$

We define the source invariance property as $\bigwedge_{i,j=1}^{T} \mathrm{INV}_{ij}[\phi]$. Such a representation is then called an invariant representation. □

The source invariance property is statistically testable given sufficiently large data collected from all the source domains. This activity is acknowledged in the literature, and representations that satisfy the source invariance property w.r.t. the source domains $\mathbb{P}$ are proposed for domain generalization in numerous existing work (e.g., (Rojas-Carulla et al. 2018; Arjovsky et al. 2019; Rothenhäusler et al. 2021; Magliacane et al. 2018; Chen and Bühlmann 2021)). In summary, Corollary 1 states that under the stability of mechanisms assumption, the source invariant property is a necessary condition for transportability of representations.

Is the source invariance property also a sufficient criterion for transportability? We need to assure that the probabilistic invariances present within the source data are not coincidental, i.e., the invariance property $\mathrm{INV}_{ij}[\phi]$ must necessarily corresponds to an s-admissibility condition in the underlying $\mathcal{G}^{\Delta_{ij}}$. This is analogous to c-faithfulness Jaber et al. (2020), which is an extension of faithfulness assumption (Pearl 2009) for the setting where we have access to multiple datasets obtained from controlled soft interventions. What follows is our proposed variation of faithfulness assumption tailored to the problem at hand.

Assumption 2: r-faithfulness. *The source distributions $\mathbb{P}$ are r-faithful to the underlying selection diagrams $\mathcal{G}^{\boldsymbol{\Delta}}$ if for all representations $\mathbf{R} = \phi(\mathbf{X})$ and for every $i, j \in [T]$,*

$$\mathrm{INV}_{ij}[\phi] \implies S_{ij} \perp\!\!\!\perp_d Y \mid \mathbf{Z}, \bar{\mathbf{R}} \text{ in } \mathcal{G}^{\Delta_{ij}}_{\mathrm{aug}}, \tag{36}$$

where $\mathbf{Z} = \det(\phi)$, $\bar{\mathbf{R}} = \bar{\phi}(\bar{\mathbf{Z}})$ denotes the constraints on constrained variables $\bar{\mathbf{Z}} = \mathrm{cons}(\phi)$, and $\mathcal{G}^{\Delta_{ij}}_{\mathrm{aug}}$ is the augmented selection diagram (Def. 8) □

Under r-faithfulness assumption, we can use the structure that SoM assumption imposes to the underlying selection diagram, and prove the source invariance property as a sound and complete data-driven criterion for transportability; what follows is a formal statement.

Theorem 2 (Data-driven transportability) *For a representation $\mathbf{R} = \phi(\mathbf{X})$, the score function $l_\phi(\mathbf{r}) = \mathbb{E}_{P^*}[Y \mid \mathbf{r}]$ can be computed in terms of the source distributions under r-faithfulness and stability of mechanisms assumption, if and only if ϕ satisfies the source invariance property.* □

Theorem 2 unifies some of the existing approach to domain generalization, and shed lights on the weaknesses of some other proposals. In appendix C, we provided a thorough discussion on the related work. Below, we consider two special cases of invariant representations, namely *balanced-rate classifiers* and *multi-calibrated scores*.

3.1 Balanced-rate Classifiers

A well-attended family of criteria for training generalizable models is on balancing/equalizing different notions of prediction error across the source distributions, e.g., (Wald et al. 2021; Rothenhäusler et al. 2021; Krueger et al. 2021; Pfister et al. 2021; Arjovsky et al. 2019). Below, we propose our version of error balancing criterion that corresponds to data-driven transportability.

For a classifier $h : \mathrm{supp}(\mathbf{X}) \to \{0, 1\}$, false omission (FOR) rate and false discovery rate (FDR) w.r.t. the distribution $P(y, \mathbf{x})$ are denoted as follows;

$$\mathrm{FOR}_P(h) := P(Y = 1 \mid h(\mathbf{X}) = 0), \tag{37}$$

$$\mathrm{FDR}_P(h) := P(Y = 0 \mid e(\mathbf{X}) = 1). \tag{38}$$

We consider classifiers that attain equal/balanced false omission rate and false discovery rate across all source domains.

Definition 11 (Balanced-rate classification) *A balanced-rate classifier is a solution to the following penalized ERM;*

$$\min_h \gamma \cdot [\mathrm{Var}\big(\{\mathrm{FOR}_{P^i}(h)\}_{i=1}^T\big) \tag{39}$$

$$+ \mathrm{Var}\big(\{\mathrm{FDR}_{P^i}(h)\}_{i=1}^T\big)] + \sum_{i=1}^{T} \mathcal{R}_{P^i}(h). \tag{40}$$

In the above, γ penalizes variation among the FDR and FOR terms, and in the extreme case $\gamma \to \infty$, the solution coincide with the following constrained ERM;

$$\min_h \sum_{i=1}^{T} \mathcal{R}_{P^i}(h) \tag{41}$$

$$\text{s.t.} \quad \mathrm{FOR}_{P^i}(h) = \mathrm{FOR}_{P^j}(h), \quad \forall P^i \in \mathbb{P} \tag{42}$$

$$\mathrm{FDR}_{P^i}(h) = \mathrm{FDR}_{P^j}(h), \quad \forall P^i \in \mathbb{P}. \tag{43}$$

It remains a future work to assess the effectivity of this method in practice, but below we state the generalization guarantee of balanced-rate classification.

Proposition 1 (Balanced-rate generlization) *Under r-faithfulness and SoM assumptions, any solution to the extreme case of balanced-rate classification such as h_{ri}^{∞} is a generalizable classifier. Particularly,*

$$\text{FDR}_{P^*}(h_{\text{ri}}^{\infty}) = \text{FDR}_{P^i}(h_{\text{ri}}^{\infty}) \quad \forall P^i \in \mathbb{P}, \tag{44}$$

$$\text{FOR}_{P^*}(h_{\text{ri}}^{\infty}) = \text{FOR}_{P^i}(h_{\text{ri}}^{\infty}) \quad \forall P^i \in \mathbb{P}. \tag{45}$$

In conclusion, balancing notions of prediction error across the source distributions is indeed relevant to the domain generalization task (Krueger et al. 2021; Ben-Tal, Ghaoui, and Nemirovski 2009). We formulated an optimization scheme to seek a balance of false discovery rate and false omission rate across the sources, as well as minimizing the prediction error on the source data. Our findings on data-driven transportability of representations theoretically justifies this objective for the domain generalization task.

3.2 Multi-Calibrated Scores

In this subsection, we discuss the relation between multi-calibration and domain generalization.

Definition 12 (Multi-Calibrated (MC) score) *A representation $\psi(\mathbf{X})$ with the support $[0, 1]$ is called a score function. It is calibrated w.r.t. the distribution P if,*

$$\forall e \in [0, 1] : \mathbb{E}_P[Y \mid \psi(\mathbf{X}) = e] = e. \tag{46}$$

It is called multi-calibrated if it is calibrated w.r.t. all source distributions. □

Apparent from the above definition, the calibrated scores serve as unbiased estimation of the empirical score function. An MC score ψ can be viewed as a representation with support $[0, 1]$. Our results in Section 3 justifies MC scores for domain generalization, as stated below.

Corollary 2 (Multi-calibrated generalization) *An MC score qualifies as an invariant representation (Def.10). Due to Theorem 2, under r-faithfulness and SoM assumptions, MC score ψ is transportable, i.e.,*

$$\mathbb{E}_{P^*}[Y \mid \psi] = \mathbb{E}_{P^1}[Y \mid \psi] = \cdots = \mathbb{E}_{P^T}[Y \mid \psi]. \tag{47}$$

This observation validates use of MC for domain generalization. This finding is in-line with the claims made in Wald et al. (2021), however, the theoretical guarantees provided in that work are limited to two linear instances of the problem.

Lemma 1 by Wald et al. (2021) shows that, in fact, invariant scores (with support of $[0, 1]$) can be transformed into multi-calibrated scores; we extend this result.

Lemma 2 (Source invariance & multi-calibration) *If any representation $\mathbf{R} = \phi(\mathbf{X})$ satisfies the source invariance property, then the score $\psi(\mathbf{x}) := \mathbb{E}_{P^i}[Y \mid \mathbf{R} = \phi(\mathbf{x})]$ (for any of the source distributions $P^i \in \mathbb{P}$) is MC. Moreover, for every $P^i \in \mathbb{P}$,*

$$I_{P^i}(Y; \phi(\mathbf{X})) = I_{P^i}(Y; \psi(\mathbf{X})), \tag{48}$$

where I denotes the mutual information.

In words, Lemma 2 states a that for every representation that satisfies the source invariance property, there exists a multi-calibrated score with equivalent prediction power. This fact suggests that in search for invariant representations with high prediction power, one might limit the search space to MC scores only without a trade-off.

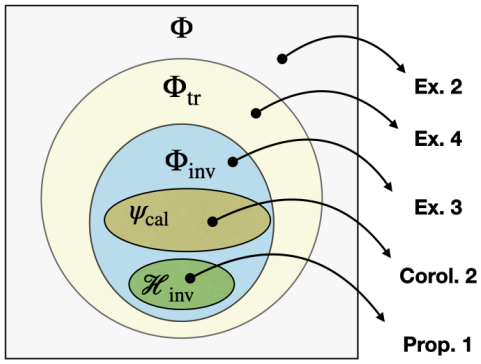

Figure 4: Φ denotes the set of all rep.; Φ_{tr} denotes the class of transportable rep.; Φ_{inv} denotes the class of invariant rep.; Ψ_{cal} denotes the class of calibrated score functions; $\mathcal{H}_{\text{inv}}$ denotes the class of balanced-rate classifiers

3.3 Taxonomy of Representations

In section 2 we elaborated through examples the relevance of transportable representations for domain generalization task. Some representations are non-transportable, e.g., Eq. 6 in Example 2, and some are transportable, e.g., Eq.12 in Example 3. The latter is not only transportable but also invariant, i.e., $\mathbb{E}[Y \mid \phi]$ matches across the source and target domains. Some representations are transportable but not invariant, e.g., Eq. 21 in Example 4. The classifiers that have balanced false negative and false discovery rates across the source and target domains constitute a subset of invariant representations that we call balanced-rate classifiers. The class of multi-calibrated scores is another subclass of invariant representation, which is also equivalent to it in terms of prediction power. Under r-faithfulness and SoM assumptions, the class of transportable representations collapses to invariant representations. In conclusion, our findings suggest the taxonomy in Figure 4 for the space of representations.

4 Conclusions

We framed the domain generalization problem within causal transportability theory. We introduced representations into the transportability pipeline, and developed a method to decide transportability of queries involving representations given structural assumptions encoded in the form of selection diagrams. Finally, we relaxed the assumption of having access to the graphs, and showed that under r-faithfulness and stability of mechanisms assumption, invariance of the empirical score across the source distributions constitutes a sound and complete data-driven criterion for generalizability. Our findings unified some of the existing ideas on invariance-based domain generalization, and opens a new thread of research for the graphical analysis of representations and their properties through transportability lenses.

References

Aldrich, J. 1989. Autonomy. *Oxford Economic Papers*, 41(1): 15–34.

Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Bareinboim, E.; Correa, J. D.; Ibeling, D.; and Icard, T. 2022. On Pearl's Hierarchy and the Foundations of Causal Inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, 507–556. NY, USA: Association for Computing Machinery, 1st edition.

Bareinboim, E.; Lee, S.; Honavar, V.; and Pearl, J. 2013. Transportability from multiple environments with limited experiments. *Advances in Neural Information Processing Systems*, 26.

Bareinboim, E.; and Pearl, J. 2014. Transportability from multiple environments with limited experiments: Completeness results. *Advances in neural information processing systems*, 27.

Bareinboim, E.; and Pearl, J. 2016. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27): 7345–7352.

Ben-David, S.; Blitzer, J.; Crammer, K.; and Pereira, F. 2006. Analysis of Representations for Domain Adaptation. In Schölkopf, B.; Platt, J.; and Hoffman, T., eds., *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

Ben-Tal, A.; Ghaoui, L. E.; and Nemirovski, A. 2009. *Robust Optimization*, volume 28 of *Princeton Series in Applied Mathematics*. Princeton University Press. ISBN 978-1-4008-3105-0.

Bengio, Y.; Courville, A.; and Vincent, P. 2013. Representation Learning: A Review and New Perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35: 1798–1828.

Chen, Y.; and Bühlmann, P. 2021. Domain adaptation under structural causal models. *The Journal of Machine Learning Research*, 22(1): 11856–11935.

Correa, J.; and Bareinboim, E. 2020. General Transportability of Soft Interventions: Completeness Results. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 10902–10912. Vancouver, Canada: Curran Associates, Inc.

Correa, J. D.; and Bareinboim, E. 2019. From Statistical Transportability to Estimating the Effect of Stochastic Interventions. In *IJCAI*, 1661–1667.

David, S. B.; Lu, T.; Luu, T.; and Pal, D. 2010. Impossibility Theorems for Domain Adaptation. In Teh, Y. W.; and Titterington, M., eds., *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, 129–136. Chia Laguna Resort, Sardinia, Italy: PMLR.

De Pierris, G.; and Friedman, M. 2018. Kant and Hume on Causality. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2018 edition.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1): 2096–2030.

Geiger, D. 1990. *Graphoids: A Qualitative Framework for Probabilistic Inference*. Ph.D. thesis, USA. UMI Order No. GAX90-16109.

Gulrajani, I.; and Lopez-Paz, D. 2020. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*.

Hanneke, S.; and Kpotufe, S. 2019. On the Value of Target Data in Transfer Learning. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Henderson, L. 2018. The problem of induction.

Hume, D. 1739. *A Treatise of Human Nature*. Oxford: Oxford University Press.

Jaber, A.; Kocaoglu, M.; Shanmugam, K.; and Bareinboim, E. 2020. Causal Discovery from Soft Interventions with Unknown Targets: Characterization and Learning. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 9551–9561. Curran Associates, Inc.

Kant, I. 1781. *Critique of Pure Reason*. St. Martin's Press (NY).

Krueger, D.; Caballero, E.; Jacobsen, J.-H.; Zhang, A.; Binas, J.; Zhang, D.; Priol, R. L.; and Courville, A. 2021. Out-of-Distribution Generalization via Risk Extrapolation (REx). In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 5815–5826. PMLR.

Lee, S.; Correa, J.; and Bareinboim, E. 2020. Generalized Transportability: Synthesis of Experiments from Heterogeneous Domains. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*. New York, NY: AAAI Press.

Li, Y.; Gong, M.; Tian, X.; Liu, T.; and Tao, D. 2018. Domain generalization via conditional invariant representations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Long, M.; Cao, Z.; Wang, J.; and Jordan, M. I. 2018. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31.

Lu, C.; Wu, Y.; Hernández-Lobato, J. M.; and Schölkopf, B. 2021. Nonlinear invariant risk minimization: A causal approach. *arXiv preprint arXiv:2102.12353*.

Magliacane, S.; Van Ommen, T.; Claassen, T.; Bongers, S.; Versteeg, P.; and Mooij, J. M. 2018. Domain adaptation by using causal inference to predict invariant conditional distributions. *Advances in neural information processing systems*, 31.

Mao, C.; Xia, K.; Wang, J.; Wang, H.; Yang, J.; Bareinboim, E.; and Vondrick, C. 2022. Causal Transportability for Visual Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 7511–7521. IEEE.

Mitchell, T. M. 1997. *Machine learning*, volume 1. McGraw-hill New York.

Pearl, J. 2000. *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press.

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

Pearl, J.; and Bareinboim, E. 2011. Transportability of causal and statistical relations: A formal approach. In *Twenty-fifth AAAI conference on artificial intelligence*.

Pfister, N.; Williams, E. G.; Peters, J.; Aebersold, R.; and Bühlmann, P. 2021. Stabilizing variable selection and regression. *The Annals of Applied Statistics*, 15(3): 1220–1246.

Popper, K. 1953. The problem of induction.

Popper, K. R. 1971. Conjectural knowledge: my solution of the problem of induction. *Revue internationale de Philosophie*, 167–197.

Rojas-Carulla, M.; Schölkopf, B.; Turner, R.; and Peters, J. 2018. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1): 1309–1342.

Rosenfeld, E.; Ravikumar, P. K.; and Risteski, A. 2021. The Risks of Invariant Risk Minimization. In *International Conference on Learning Representations*.

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

Russell, B. 1912. On induction. *First published as*, 19–26.

Russell, S.; and Norvig, P. 2010. *Artificial Intelligence: A Modern Approach*. Prentice Hall.

Schölkopf, B.; Locatello, F.; Bauer, S.; Ke, N. R.; Kalchbrenner, N.; Goyal, A.; and Bengio, Y. 2021. Toward causal representation learning. *Proceedings of the IEEE*, 109(5): 612–634.

Spirtes, P.; Glymour, C. N.; and Scheines, R. 2000. *Causation, prediction, and search*. MIT press.

Subbaswamy, A.; and Saria, S. 2020. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics*, 21(2): 345–352.

Subbaswamy, A.; Schulam, P.; and Saria, S. 2019. Preventing failures due to dataset shift: Learning predictive models that transport. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 3118–3127. PMLR.

Sugiyama, M.; and Müller, K.-R. 2005. Input-dependent estimation of generalization error under covariate shift. 23(4): 249–279.

Vapnik, V. 1991. Principles of Risk Minimization for Learning Theory. In Moody, J.; Hanson, S.; and Lippmann, R., eds., *Advances in Neural Information Processing Systems*, volume 4. Morgan-Kaufmann.

Vapnik, V. 1998. Statistical learning theory Wiley. *New York*, 1(624): 2.

Wald, Y.; Feder, A.; Greenfeld, D.; and Shalit, U. 2021. On Calibration and Out-of-Domain Generalization. In Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*.

Wang, Y.; and Jordan, M. I. 2021. Desiderata for representation learning: A causal perspective. *arXiv preprint arXiv:2109.03795*.

Watkins, E.; et al. 2005. *Kant and the Metaphysics of Causality*. Cambridge University Press.

Xu, R.; Zhang, X.; Shen, Z.; Zhang, T.; and Cui, P. 2021. A Theoretical Analysis on Independence-driven Importance Weighting for Covariate-shift Generalization. In *International Conference on Machine Learning*.

Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning Fair Representations. In Dasgupta, S.; and McAllester, D., eds., *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, 325–333. PMLR.

Zhang, K.; Gong, M.; and Schoelkopf, B. 2015. Multi-Source Domain Adaptation: A Causal View. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1).