TECHNICAL REPORT
R-99
First version: May, 2023
Last version: Feb, 2024

TRANSPORTABLE REPRESENTATIONS FOR DOMAIN GENERALIZATION

A PREPRINT

Kasra Jalaldoust and **Elias Bareinboim**
Causal Artificial Intelligence Laboratory
Columbia University
{kasra,eb}@cs.columbia.edu

ABSTRACT

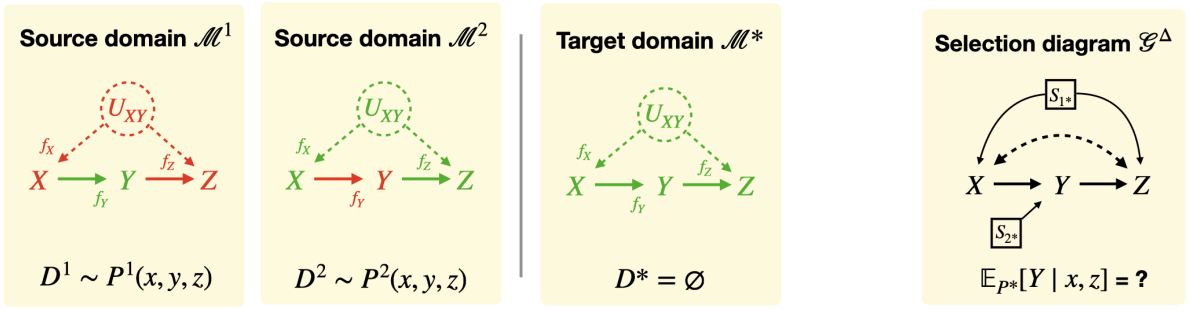
Generalizing across settings and changing conditions is one of the fundamental problems of AI. One critical assumption in this context is that the testing and training data come from the same distribution, which, despite its popularity in the literature, is often violated in practice. The anchors that allow generalizations to take place are causal and stem from the stability of the mechanisms underlying the system under investigation. Building on the theory of causal transportability introduced by Bareinboim & Pearl, we define the notion of "transportable representations" to provide data structures for allowing a formal analysis of the domain generalization task. We then develop an algorithm to decide whether the distribution of the label, conditioned on the representation, can be computed in terms of the source distributions and assumptions about the commonalities and disparities across source and target domains. Moreover, we relax the assumption of having the graph as the task's input and prove a graphical-invariance duality theorem, delineating the conditions under which certain invariances in the source data can be used as a sound and complete criterion for assessing the generalizability of a classifier. We review the prior literature and show how 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 [22]. This challenge is called the *problem of induction* [21], and have puzzled generations of philosophers and mathematicians throughout history, from Kant to Popper, Goodman to Russell [42, 43, 47, 59].

The generalization problem plays a fundamental role in artificial intelligence and machine learning as well [36, 48], 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 [36, 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.", which can be thought as a refinement of 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 Vladimir Vapnik, which tied the risk between hypothetical and empirical distributions under some very general conditions [55, 56].

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 most of the prior results. In fact, if the target domain is arbitrary, or drastically different from the source domains, no learning could take place [15, 6]. However, the fact that we generalize and adapt



(a) (b)

Figure 1: (a) The causal structure of the observed domains M^1, M^2, and the unseen domain M^*. For each of the variable the mechanisms are in the same color if they are shared across the domains. (b) The selection diagram $\mathcal{G}^\Delta$ that aggregates the causal diagrams and domain discrepancies.

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 [37, 50, 4]. 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 [1]. [1] Figure 1 illustrates this idea. In all domains, the value of each variable $V \in \{X, Y, Z\}$ is decided through a function f_V and based on the value of the variables pointing to V. Some of these mechanisms might be similar across the domains, and are shown with the same color; for instance, the mechanism of Y is the same between $\mathcal{M}^2, \mathcal{M}^*$. On the other hand, some mechanisms might be dissimilar across the domains, e.g., the mechanism for Z might differ between $\mathcal{M}^2, \mathcal{M}^*$. The goal is estimating $\mathbb{E}_{P^*}[Y \mid x, z]$ in the target domain, and the main challenge is that we do not observe any data from it. Ideally, we will leverage our knowledge about domain discrepancies to use the data collected from the source domains, and estimate the target quantity in the target domain. 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* [3, 4, 5, 39, 13, 14, 30].

In modern machine learning literature, the challenge of predicting in an unseen target domain has also been acknowledged and is broadly referred to as the domain generalization problem. In this task, one has access to labeled data from the source domains, while no data in the target domain is available [19]. The theoretical proposals in this area rely on,

(a) assumptions to define the target domains compatible with the source data, such as the covariate shift assumption [53, 52, 51, 60], or,

(b) (b) distance measures to relate the source and target distributions (e.g., [7, 20]).

Even under restrictive assumptions about the relation between the source and target distributions, generalization might still be impossible [15]. 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 [34, 44, 2, 46, 57, 11, 33]. Theoretical guarantees provided for these methods are contingent on assumptions such as linearity, additivity, markovianity (i.e., no unobserved common causes), yet there are subtleties that limit the effectiveness and practicality of these methods as we will elaborate it further [45]. 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 when data is high-dimensional and unstructured [9]. 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 [7, 17, 32, 31, 62, 61]. In some work, causal features have been used in constructing representations while filtering out the spurious correlations that might be unstable across domains [58, 49, 35, 29].

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, i.e., estimate $P^*(y \mid \phi(\mathbf{X}))$ or $\mathbb{E}_{P^i}[Y \mid \phi(\mathbf{X})]$.

[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 [28]; for further discussion on this point, refer to [16].

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. [7], 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 in depth through transportability lenses. In particular, we analyze the fundamental interplay between causal knowledge and the generalizability of the representation. These results have various implications to the practice in the field, and for example, we refute the belief that causal features are always desirable while spurious ones should be discarded when generalizing across the domains. Our contributions are as follows:

- (Section 2) We formalize the domain generalization task by introducing the notion of transportable representations (Def. 5), and then we develop a procedure (Thm. 1) to decide whether a representation is transportable given the structural assumptions encoded by a causal graph. We demonstrate finite-sample performance of a transportability-based classification via synthetic experiments, and show its superiority to both vanilla ERM and invariance-based method.
- (Section 3) We prove that when the distribution of label conditioned on the representation is the same across the source domains, i.e., the so called invariance property holds, the representation is suitable for domain generalization. In fact, the invariance property is 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 and trade-offs of operating graph-free.

Preliminaries. Here, we introduce the basic notation used throughout the paper. 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 a causal graphs.

We use semantics of Structural Causal Models [37], 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 $\mathcal{G}$. SCM $\mathcal{M}$ entails a probability distribution $P^{\mathcal{M}}(\mathbf{v})$ over the set of observed variables $\mathbf{V}$ such that*

$$P^{\mathcal{M}}(\mathbf{v}) = \int_{\text{supp}(\mathbf{U})} \prod_{V \in \mathbf{V}} P^{\mathcal{M}}(v \mid \mathbf{pa}_V, \mathbf{u}_V) \cdot P(\mathbf{u}) \cdot d\mathbf{u}, \tag{1}$$

where the 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 the positivity assumption, 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 $\mathbf{V} = \mathbf{X} \cup \{Y\}$, where Y is a binary label, and $\mathbf{X}$ is a set of features. 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 $\mathbb{P} = \{P^1, P^2, \ldots, P^T\}$ via Eq. 1, 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 it induces the causal diagram $\mathcal{G}^*$. The following example illustrates these concepts.

Example 1 (Covariate shift) Let $\mathbf{X} := \langle X_1, X_2, \ldots, X_N \rangle$ be a vector of binary covariates (or features), and Y be a binary label. Assume data from two source domains $\mathcal{M}^1, \mathcal{M}^2$ is available, and the task is to predict the label Y based on the features $\mathbf{X}$ in the target domain $\mathcal{M}^*$. More specifically, the two source domains and the target domain are specified as follows:

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

Implied by this specification, the label Y in all domains $\mathcal{M}^1, \mathcal{M}^2, \mathcal{M}^*$ follows the conditional distribution,

$$Y \mid \mathbf{X} \sim \text{Bernoulli}(\sigma(\alpha^\top \cdot \mathbf{X})), \tag{3}$$

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 the features $\mathbf{X}$ differs across the source and target domains, while the likelihood of $Y = 1$ is determined by $\mathbf{X}$. Consider now the distribution of the features in each domain:

1. In $\mathcal{M}^1$, the variables $\mathbf{X}$ are drawn from the uniform distribution, i.e., $P^1(\mathbf{x}) = \frac{1}{2^N}$ for all $\mathbf{x} \in \text{supp}(\mathbf{X})$.
2. In $\mathcal{M}^2$, $P^2(\mathbf{x}) \propto \|\mathbf{x}\|$, where $\|\cdot\|$ is the L2 norm, and in this case, it counts the number of ones in $\mathbf{x}$.
3. In the target domain, however, $P^*(\mathbf{x}) \propto \frac{1}{\|\mathbf{x}\|+1}$.

Further, the causal diagrams if each SCM, $\mathcal{G}^1, \mathcal{G}^2, \mathcal{G}^*$ coincide, as shown in Figure 2a. This setup is commonly referred to in the literature as the covariate shift [53, 52, 51], where it is assumed 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 [30].

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:*

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, if an endogenous variable V is not in Δ_{ij} , this means 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 [30] 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. The following example illustrates selection diagrams.

(a) (b)

Figure 2: (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, while the label is determined by the covariates and is not confounded with any 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.

Example 2 (Ex. 1 continued; covariate shift) Following Definition 2, the domain discrepancy sets are $\Delta_{12} = \Delta_{1*} = \Delta_{2*} = \mathbf{X}$, i.e., the variables in $\mathbf{X}$ are not structurally invariant across domains. Further, and more importantly, the set of domain discrepancies $\mathbf{\Delta}$ implies that the label Y is structurally invariant across domains. Putting these observations together, the induced selection diagram is shown Figure 2b, where the selection nodes S_{12}, S_{*1}, S_{*2} are pointing to $\mathbf{X}$ nodes. □

We next define the notion of representations and their score function.

Definition 4 (Representation and score function) *The random 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 $\mathbf{R} = \phi(\mathbf{X})$. Further, the corresponding score function is defined as*

$$l_\phi(\mathbf{r}) := \mathbb{E}_{P^*}[Y \mid \mathbf{R} = \mathbf{r}]. \tag{4}$$

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. □

A representation (a.k.a., featurizer) is an aggregation of the information contained 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 by the next example.

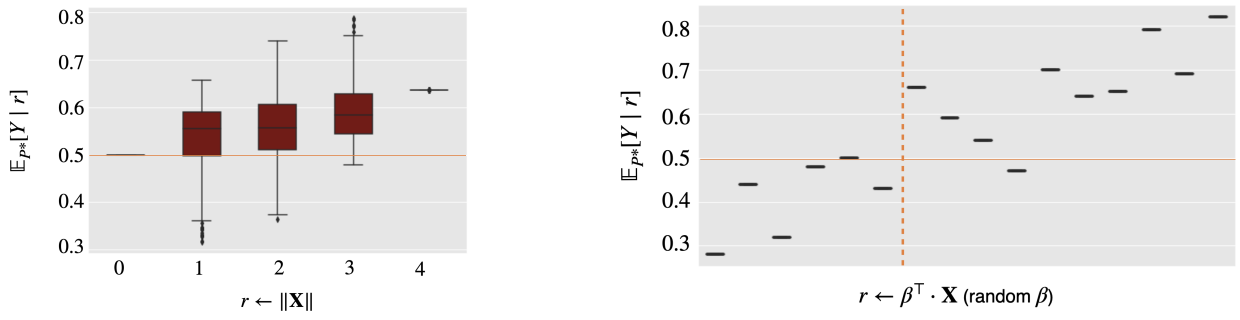
Example 3 (Ex. 1 continued; representations) The goal of the domain generalization task is to predict Y by observing $\mathbf{X}$ in the target domain $\mathcal{M}^*$. Now we consider our first representation that aggregates the information of $\mathbf{X}$, for instance:

$$R_{\text{sum}} = \phi_{\text{sum}}(\mathbf{X}) := \sum_{i=1}^{N} X_i, \tag{5}$$

We then may consider predicting Y based on the value of R. To do so, a natural route is to compute the quantity $\mathbb{E}_{P^*}[Y \mid R_{\text{sum}} = r]$ for every $r \in \{0, 1, \ldots, N\}$ using the data collected from source domains $\mathcal{M}^1, \mathcal{M}^2$. This is well-defined only if this quantity is unique under all the 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. 5. □

Further, 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 common in the causal inference literature. Traditionally, the expression on the r.h.s. of a the conditioning bar in a query denotes probabilistic conditioning on a set of certain values for a subset of variables, e.g., $P(y \mid X = x, Z = z)$. However, by conditioning on the value of a representation $\mathbf{R} = \phi(\mathbf{X})$, we ar able to express an extended family of queries, e.g., $P(y \mid \phi(\mathbf{X}) = \mathbf{r})$, where the mapping $\mathbf{R} = \phi(\mathbf{X})$ is arbitrary but known. In special case, ϕ can be a subset selection (i.e., feature selection map), which coincides with the traditional approach to the queries. 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 $\mathbf{r} \in \text{supp}(\mathbf{R})$. Note that this is a property of the representation ϕ which 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 data-driven (Kasra: instead of "algebraic"



(a) (b)

Figure 3: (a) The score function of ϕ_{sum} (Example 4) for a randomly generated example; the bounds indicate how much variation one might see in the score function for different plausible target SCMs $\mathcal{M}^*$. (b) The score function of ϕ_{rand} (Example 5) for the same instance of the problem; as seen here the bounds are collapsed, meaning that the score function is not unique across all plausible target SCMs, which allows us to have generalizable prediction based on this representation; for instance, the threshold indicated by the dotted line yields a very good classification rule.

which you suggested before) (Sec. 3) assumptions. To this end, we extend the notion of transportability [5] to study queries involving representations.

Definition 5 (Transportable representation) *The representation* $\mathbf{R} = \phi(\mathbf{X})$ *is said to be transportable from the collection of distributions* $\mathbb{P}$ *given* $\langle \mathcal{G}^{\mathbf{\Delta}}, \phi \rangle$ *if for every pair system of SCMs such as* $\langle \mathcal{M}_a^1, \mathcal{M}_a^2, \dots, \mathcal{M}_a^T, \mathcal{M}_a^* \rangle$ *and* $\langle \mathcal{M}_b^1, \mathcal{M}_b^2, \dots, \mathcal{M}_b^T, \mathcal{M}_b^* \rangle$ *that induce* $\mathcal{G}^{\mathbf{\Delta}}$ *and entail* $\mathbb{P}$*,*

$$\mathbb{E}_{P^{\mathcal{M}_a^*}}[Y \mid \phi(\mathbf{X}) = \mathbf{r}] = \mathbb{E}_{P^{\mathcal{M}_b^*}}[Y \mid \phi(\mathbf{X}) = \mathbf{r}], . \tag{6}$$

for all $\mathbf{r} \in \text{supp}(\mathbf{R})$. □

In words, ϕ is transportable if its 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 $\mathbb{P}$, considering:

1. The assumption encoded in the selection diagrams $\mathcal{G}^{\mathbf{\Delta}}$, and
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 what is called in the literature the notion of statistical transportability (e.g., see [14]). 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 evaluate it. 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 developing a method to compute an expression for the score function in terms of the available source distributions. This expression can be thought of as a blue-print for estimation. Below, we discuss a well-attended instance of the domain generalization problem.

Example 4 (Ex. 1 continued: lack of transportability) In the context of Example 1, we argue that for the representation $R_{\text{sum}} = \phi_{\text{sum}}(\mathbf{X})$ given in Eq. 5, 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,

$$l_{\phi_{\text{sum}}}(1) = \mathbb{E}_{P^*}[Y \mid \sum_{i=1}^{N} X_i = 1] \tag{7}$$

$$= \sum_{i=1}^{N} \sigma(\alpha_i) \cdot P^*(\mathbf{X} = \text{one hot at } i \mid \sum_{i=1}^{N} X_i = 1). \tag{8}$$

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 $P^*(\mathbf{X} = \text{one hot at } i \mid \sum_{i=1}^{N} X_i = 1)$ in Eq. 8 can be any set 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 $\sigma(\alpha_i)$ or just above the minimum of them. Precisely speaking, for every

$$c \in (\min_{1\leq i\leq N} \sigma(\alpha_i), \max_{1\leq i\leq N} \sigma(\alpha_i)), \tag{9}$$

there exists a plausible target SCM $\mathcal{M}_c^*$ that is compatible with the selection diagram in Figure 2b and the source distributions P^1, P^2, such that $\mathbb{E}_{P^{\mathcal{M}_c^*}}[Y \mid R_{\text{sum}} = 1] = c$. Thus, as long as the coefficients α_i are not all equal, the interval in Eq. 9 contains more than one value, and therefore, we can not assure that the score function is unique across all compatible target domains. We conclude that ϕ_{sum} is not transportable in this example.

Figure 3a shows the bounds above for all values of $r \in \{0, 1, \ldots, N\}$, obtained with randomly generated instance of the problem; as seen in this figure, the only value of R_{sum} for which a confident prediction can be made is $R_{\text{sum}} = 4$, as in all other cases the score function can be both over and below 0.5, making it impossible to decide Y. □

Despite its simplicity, this example carries an important message. In settings that involve representations, evaluating 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). This can be written more formally as follows.

Corollary 1 *Even under the covariate shift assumption, it is not possible to evaluate the score function of all representation, i.e.,*

$$\textit{Covariate shift assumption} \not\Longrightarrow \textit{transportability of representations}.$$

Moreover, there exist instances of the problem where the score function of certain representations can be evaluated using the source data, but the covariate shift assumption does not hold (e.g., Example 8), i.e.,

$$\textit{Transportability of representations} \not\Longrightarrow \textit{covariate shift assumption}.$$

□

We introduce a graphical criterion for in the selection diagrams that is useful to evaluate the probabilistic invariances in the distribution motivated by [39].

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 conditioned on $\mathbf{Z}$ w.r.t. the domains $\mathcal{M}^i, \mathcal{M}^j$ whenever $\mathbf{A}$ is d-separated from S_{*i} given $\mathbf{Z}$ in $\mathcal{G}^{\Delta_{ij}}$. The conditional distribution of $\mathbf{A}$ given $\mathbf{Z}$ is invariant across domain, that is,*

$$\mathbf{A} \perp\!\!\!\perp_d S_{ij} \mid \mathbf{Z} \textit{ in } \mathcal{G}^{\Delta_{ij}} \implies P^i(\mathbf{a} \mid \mathbf{z}) = P^j(\mathbf{a} \mid \mathbf{z}). \tag{10}$$

□

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 5 (Covariate shift: transportable representation) In the context of Example 1, consider the representation

$$R_{\text{rand}} = \phi_{\text{rand}}(\mathbf{X}) := \beta^\top \cdot \mathbf{x}, \tag{11}$$

where $\beta \sim \mathcal{N}(0, I)$ is drawn independently. Considering the expression above, we can almost surely compute an inverse function $\mathbf{X} = \phi_{\text{rand}}^{-1}(R_{\text{rand}})$ (proof in Appendix A). Intuitively, the reason is that $\text{supp}(R) = \mathbb{R}$ is larger than $\text{supp}(\mathbf{X})$ which has 2^N elements in it, and with a random β no two values of $\mathbf{X}$ get mapped to the same value in $\mathbb{R}$. Thus, the score function $l_{\phi_{\text{rand}}}(r)$ can be written as,

$$\mathbb{E}_{P^*}[Y \mid R_{\text{rand}} = r] = \mathbb{E}_{P^*}[Y \mid \mathbf{X} = \phi_{\text{rand}}^{-1}(r)]. \tag{12}$$

Licensed by the s-admissibility relation $Y \perp\!\!\!\perp_d S_{ij} \mid \mathbf{X}$ as readable from the selection diagram in Fig. 2b, the latter can be directly obtained from either of the source distributions, namely,

$$\mathbb{E}_{P^*}[Y \mid \mathbf{X} = \phi_{\text{rand}}^{-1}(r)] = \mathbb{E}_{P^{1,2}}[Y \mid \mathbf{X} = \phi_{\text{rand}}^{-1}(r)]. \tag{13}$$

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

$$\begin{aligned}
\mathbb{E}_{P^*}[Y \mid \phi_{\text{rand}}(\mathbf{X}) = r] &= \mathbb{E}_{P^*}[Y \mid \mathbf{X} = \phi_{\text{rand}}^{-1}(r)] && \text{(Eq. 12)} && (14)\\
&= \mathbb{E}_{P^{1,2}}[Y \mid \mathbf{X} = \phi_{\text{rand}}^{-1}(r)] && \text{(Eq. 13)} && (15)\\
&= \mathbb{E}_{P^{1,2}}[Y \mid \phi_{\text{rand}}(\mathbf{X}) = r] && \text{(Eq. 12).} && (16)
\end{aligned}$$

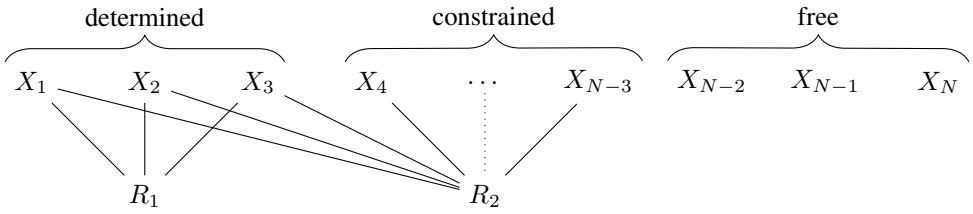


Figure 4: The relationship between the representation and the features.

Figure 3b shows the values of the score function for a randomly generated instance of the problem with four covariates. A classifier defined based on the representation ϕ_{rand},

$$\hat{h}(\phi_{\text{rand}}(\mathbf{X})) := \begin{cases} 1 \text{ if } \beta^\top \cdot \mathbf{X} > 0.2 \\ 0 \text{ otherwise} , \end{cases} \tag{17}$$

matches the optimal Bayes classifier of the target distribution, except for $\phi(\mathbf{X}) = 0.62$. This means that prediction based on the representation $R = \beta^\top \cdot \mathbf{X}$ is not only generalizable, but also efficient. □

As seen in Examples 4 and 5, 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 regarding 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 [18] regarding deterministic relations in SCMs.

Definition 7 (Determined, constrained, 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. The following example elaborates on this definition.

Example 6 (Definition 7 illustrated) The representation ϕ_{sum} in Example 4 does not determine any of the features, because the equation,

$$r = \phi_{\text{sum}}(\mathbf{X}) = \sum_{i=1}^{N} X_i, \tag{18}$$

does not have a unique solution for any X_i in terms of r, and therefore, $\det(\phi_{\text{sum}}) = \emptyset$. However, one can limit the possible values of all variables in $\mathbf{X}$ based on some values of $\mathbf{r}$; for instance, for $\mathbf{r} = N$, it is certain that all $X_i = 1$. Thus, $\text{cons}(\phi_{\text{sum}}) = \mathbf{X}$, and consequently, $\text{free}(\phi_{\text{sum}}) = \emptyset$.

On the other hand, the representation ϕ_{rand} in Example 5 determines all variables in $\mathbf{X}$, because the mapping ϕ_{rand} is almost surely invertible, as discussed in Example 5. Therefore, $\det(\phi_{\text{rand}}) = \mathbf{X}$, and $\text{cons}(\phi_{\text{rand}})\text{free}(\phi_{\text{rand}}) = \emptyset$.

Further, we consider a more complex situation. Suppose $\mathbf{X} = \langle X_1, X_2, \ldots, X_N \rangle$ is a binary vector of random variables ($N > 7$). Consider the representation,

$$\mathbf{R} = \phi(\mathbf{X}) := \langle \underbrace{\beta^\top \cdot \mathbf{X}_{1:3}}_{R_1}, \underbrace{\sum_{i=1}^{N-3} X_i}_{R_2} \rangle, \tag{19}$$

where β is drawn from $\mathcal{N}(0, I_3)$. In words, the first entry of the representation, R_1, is a random linear mapping of X_1, X_2, X_3, and the second entry, R_2, is the summation of the entries of $X_1, X_2, \ldots, X_{N-3}$; Fig. 4 shows the correspondence between the entries of $\mathbf{X}$ and the entries of $\mathbf{R}$.

Consider a specific value of the representation, such as $\mathbf{r} = \langle r_1, r_2 \rangle = \phi(\mathbf{X})$. As seen in Example 5, we can almost surely recover the values of X_1, X_2, X_3 from r_1, thus $\det(\phi) = \{X_1, X_2, X_3\}$. On the other hand, by plugging in the determined values into the expression for R_2, we have,

$$r_2 = x_1 + x_2 + x_3 + \sum_{i=4}^{N-3} X_i. \tag{20}$$

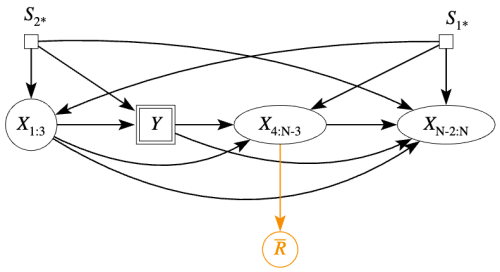


Figure 5: Augmented selection diagram corresponding to Example 7. The node S_{12} is removed to avoid clutter, and is connected to all variables $\mathbf{X}$.

Let $\tilde{\phi}(\mathbf{X}_{4:N-3}) = \sum_{i=4}^{N-3} X_i$. We can rewrite the above as,

$$\bar{\phi}(\mathbf{X}_{4:N-3}) = r_2 - x_1 - x_2 - x_3, \tag{21}$$

which specifies a constraint over the variables $X_4, X_5, \ldots, X_{N-3}$. However, these variables cannot be determined uniquely, thus, $\text{cons}(\phi) = \{X_4, X_5, \ldots, X_{N-3}\}$. The following can be stated:

$$\phi(\mathbf{X}) = \langle r_1, r_2 \rangle \iff \begin{cases} \langle X_1, X_2, X_3 \rangle = \langle x_1, x_2, x_3 \rangle \\ \bar{\phi}(\mathbf{X}_{4:N-3}) = r_2 - x_1 - x_2 - x_3 \end{cases} \tag{22}$$

Finally, we see that the value of variables X_{N-2}, X_{N-1}, X_N cannot be constrained by $\mathbf{r}$, and therefore $\text{free}(\phi) = \{X_{N-2}, X_{N-1}, X_N\}$. □

As seen in the example above, constraints over the certain variables can be derived such as the one given by Eq. 21. This equation is obtained by plugging-in the value of determined variables into the expression for the representation, and then massaging it to have only non-determined variables on the r.h.s. of the equation. Motivated by this example, we incorporate the knowledge of the representations into the selection diagram.

Definition 8 (Augmented selection diagrams) *Let $\mathcal{G}^{\boldsymbol{\Delta}}$ be a selection diagram over the variables $\mathbf{X}, Y$, and $\mathbb{P}$ be the set of 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. Also, let the equation $\bar{R} = \bar{\phi}(\mathbf{Z})$ obtained from $\mathbf{R} = \phi(\mathbf{X})$ specify the constraints. We construct a new selection diagram and distributions by adding the variable $\bar{\mathbf{R}}$ to $\mathcal{G}^{\boldsymbol{\Delta}}, \mathbb{P}$ following two steps:*

$$\mathcal{G}^{\boldsymbol{\Delta}}_{\text{aug}}: \textit{ add node } \bar{\mathbf{R}} \textit{ to } \mathcal{G}^{\boldsymbol{\Delta}} \textit{ with arrows from } \bar{\mathbf{Z}} \textit{ nodes to } \bar{\mathbf{R}} \tag{23}$$

$$\mathbb{P}_{\text{aug}}: \{P^i_{\text{aug}} := P^i(\mathbf{x}, y) \cdot 1_{\{\bar{\mathbf{r}} = \bar{\phi}(\bar{\mathbf{z}})\}}, \textit{ for } P^i \in \mathbb{P}\}. \tag{24}$$

The corresponding data structure is called the augmented selection diagram. □

The motivation behind constructing the augmented selection diagram is to extract as much information as possible from the condition $\mathbf{r} = \phi(\mathbf{X})$ and incorporate it into the selection diagram. This enables us to use the existing algorithms such as gTR [13] that operate on selection diagrams for transportability. For consistency, we need to augment the distribution with the added node; this is possible as we know the mechanism generating $\bar{\mathbf{R}}$ based on $\bar{\mathbf{Z}}$ across all domains.

Example 7 (Ex. 6 continued) Consider a system of variables that induces the selection diagram depicted in Figure 5 (in black), and the representation ϕ in Eq. 19. As seen in Example 6, $\det(\phi) = \mathbf{X}_{1:3}$, $\text{cons}(\phi) = \mathbf{X}_{4:N-3}$ and $\text{free}(\phi) = \mathbf{X}_{N-2:N}$. The constraint over $\text{cons}(\phi)$ is indicated by,

$$\bar{R} = \bar{\phi}(\mathbf{X}_{4:N-3}) = \sum_{i=4}^{N-3} X_i. \tag{25}$$

The augmented selection diagram $\mathcal{G}^{\boldsymbol{\Delta}}_{\text{aug}}$ is obtained by adding the variable $\bar{R}$ to the selection diagram and adding arrows from $\mathbf{X}_{4:N-3}$ nodes to it.

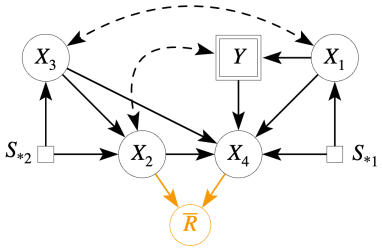


Figure 6: selection diagram corresponding to Example 8 (black). The node S_{12} is removed to avoid clutter, and is connected to all variables $\mathbf{X}$. The naively augmented (with orange) and augmented (with blue) selection diagrams.

Notice that in this example the covariate shift assumption does not hold, as Y conditioned on $\mathbf{X}$ changes across the domains. Still, one may wonder if ϕ is a transportable representation. In order to answer this question, we consider the following derivation that justifies transportability of ϕ:

$$
\begin{aligned}
l_\phi(\mathbf{r}) &= \mathbb{E}_{P^*}[Y \mid r_1, r_2] && \text{(Def. 4)} && (26)\\
&= P^*(Y=1 \mid \mathbf{x}_{1:3}, \bar{r}) && \text{(Eq. 22)} && (27)\\
&= P^*(Y=1 \mid \mathbf{x}_{1:3}) \cdot P^*(\bar{r} \mid Y=1, \mathbf{x}_{1:3}) && \text{(factorization)} && (28)\\
&= P^1(Y=1 \mid \mathbf{x}_{1:3}) \cdot P^*(\bar{r} \mid Y=1, \mathbf{x}_{1:3}) && (S_{1*} \perp\!\!\!\perp Y \mid \mathbf{X}_{1:3} \text{ in } \mathcal{G}^{\boldsymbol{\Delta}}_{\text{aug}}) && (29)\\
&= P^1(Y=1 \mid \mathbf{x}_{1:3}) \cdot P^2(\bar{r} \mid Y=1, \mathbf{x}_{1:3}) && (S_{2*} \perp\!\!\!\perp \bar{R} \mid \mathbf{X}_{1:3}, Y \text{ in } \mathcal{G}^{\boldsymbol{\Delta}}_{\text{aug}}). && (30)
\end{aligned}
$$

□

The following result generalizes the derivation in the previous example, and allows one to incorporate the knowledge of the representation ϕ into the symbolic inference pipeline.

Theorem 1 (Graphical transportability) *Consider a representation $\mathbf{R} = \phi(\mathbf{X})$, and a 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,*

$$l_\phi(\mathbf{r}) = \mathbb{E}_{P^*}[Y \mid \mathbf{r}] = P^*(Y=1 \mid \mathbf{Z}=\mathbf{z}, \bar{\mathbf{R}}=\bar{\mathbf{r}}). \tag{31}$$

The representation ϕ is a transportable (i.e., l_ϕ can be computed in terms of $\mathbb{P}$) given $\langle \mathcal{G}^{\boldsymbol{\Delta}}, \phi \rangle$ 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 $\mathcal{G}^{\boldsymbol{\Delta}}_{\text{aug}}$ via the gTR algorithm by Lee et al. [30].

All proofs are provided in Appendix A. We elaborate on the procedure described above through an example.

Example 8 (Deciding transportability) Consider the selection diagram $\mathcal{G}^{\boldsymbol{\Delta}}$ shown in Figure 6 over the variables Y and $\mathbf{X} = \{X_1, X_2, X_3, X_4\}$ with $\text{supp}(X_i) \subset \mathbb{N}$. We have data coming from two source domains, and our goal is to compute the score function $Q : l_\phi(\mathbf{r}) = \mathbb{E}_{P^*}[Y \mid \mathbf{r}]$ for the representation,

$$\mathbf{R} = \phi(\mathbf{X}) := \langle \underbrace{\frac{X_1 \cdot X_2 \cdot X_3}{X_4}}_{R_1}, \underbrace{\frac{X_1 \cdot X_2}{X_3 \cdot X_4}}_{R_2}, \underbrace{\frac{X_2 \cdot X_3}{X_1 \cdot X_4}}_{R_3} \rangle. \tag{32}$$

As seen in the expression above, every entry of this representation depends on every feature, so it is not immediately clear which entries of $\mathbf{X}$ are determined, constrained, or free. However, Eq. 32 can be viewed as a system of equations that allows us to write $\mathbf{X}$ in terms of $\mathbf{R}$:

$$X_1 = \sqrt{\frac{R_1}{R_3}}, \tag{33}$$

$$X_3 = \sqrt{\frac{R_1}{R_2}}, \tag{34}$$

$$\frac{X_2}{X_4}\sqrt{R_2 \cdot R_3}. \tag{35}$$

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 Eqs. 33, 34. On the other hand, X_2, X_4 are constrained by the value of $\mathbf{R}$ through Eq. 35, but cannot be uniquely computed. Thus, by definition 7, $\text{det}(\phi) = \{X_1, X_3\}$ and $\text{cons}(\phi) = \{X_2, X_4\}$.

Now, we attempt to construct the augmented selection diagram. 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 given in Def. 8 adds a new variable that is shown in orange in Figure 6. For a fixed value $\mathbf{r} \in \text{supp}(\mathbf{R})$, let

$$x_1^{\mathbf{r}} := \sqrt{\frac{r_1}{r_3}}, x_3^{\mathbf{r}} = \sqrt{\frac{r_1}{r_2}}, \bar{r}^{\mathbf{r}} = \sqrt{r_2 \cdot r_3} \tag{36}$$

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

$$Q : l_\phi(\mathbf{r}) = \mathbb{E}_{P^*}[Y \mid \mathbf{R} = \mathbf{r}] = \underbrace{P^*(Y = 1 \mid x_1^{\mathbf{r}}, x_3^{\mathbf{r}}, \bar{r}^{\mathbf{r}})}_{Q'}. \tag{37}$$

Now, we attempt to transport the query Q' from the source distributions given the augmented selection diagram 6. We follow the gTR algorithm [13], and first use the definition of conditional probability:

$$Q' = P^*(y \mid \bar{r}^{\mathbf{r}}, x_1^{\mathbf{r}}, x_3^{\mathbf{r}}) \tag{38}$$

$$= \frac{\sum_{x_2, x_4} P^*(y, \bar{r}^{\mathbf{r}}, x_2, x_4 \mid x_1^{\mathbf{r}}, x_3^{\mathbf{r}})}{\sum_{y, x_2, x_4} \underbrace{P^*(y, \bar{r}^{\mathbf{r}}, x_2, x_4 \mid x_1^{\mathbf{r}}, x_3^{\mathbf{r}})}_{Q''}}. \tag{39}$$

Now, we factorize the numerator as,

$$\begin{aligned} Q'' &= P^*(\bar{r}^{\mathbf{r}} \mid y, x_1^{\mathbf{r}}, x_2, x_3^{\mathbf{r}}, x_4) \cdot P^*(y, x_2, x_4 \mid x_1^{\mathbf{r}}, x_3^{\mathbf{r}}) \\ &= P^*(\bar{r}^{\mathbf{r}} \mid x_2, x_4) \cdot P^*(y, x_2, x_4 \mid x_1^{\mathbf{r}}, x_3^{\mathbf{r}}) && (\bar{R} \perp\!\!\!\perp_d X_1, X_3, Y \mid X_2, X_4 \text{ in Fig. 6}) && (40) \\ &= 1_{\{\bar{r}^{\mathbf{r}} = \frac{x_2}{x_4}\}} \cdot \underbrace{P^*(y, x_2, x_4 \mid x_1^{\mathbf{r}}, x_3^{\mathbf{r}})}_{Q'''}, && (\text{Eq.24}) && (41) \end{aligned}$$

Finally, we transport Q''':

$$Q''' = P^*(y, x_2 \mid x_1^{\mathbf{r}}, x_3^{\mathbf{r}}) \cdot P^*(x_4 \mid y, x_2, x_1^{\mathbf{r}}, x_3^{\mathbf{r}}) \tag{42}$$

$$= P^1(y, x_2 \mid x_1^{\mathbf{r}}, x_3^{\mathbf{r}}) \cdot P^2(x_4 \mid y, x_2, x_1^{\mathbf{r}}, x_3^{\mathbf{r}}), \tag{43}$$

where Eq. 43 is licensed by the S-admissibility relations,

$$Y, X_2 \perp\!\!\!\perp_d S_{1*} \mid X_1, X_3 \tag{44}$$

$$X_4 \perp\!\!\!\perp_d S_{2*} \mid Y, X_1, X_2, X_3. \tag{45}$$

The derivation above allows expressing the score function solely in terms of the source distributions, which means that the representation ϕ specified in Eq. 32 is indeed transportable. □

Remark. The covariate shift assumption does not hold in Example 8; the reason is that $S_{*1} \rightarrow X_4 \leftrightarrow Y$ and $S_{*2} \rightarrow X_2 \leftarrow Y$ are d-connecting paths between the s-nodes and Y given $\mathbf{X}$, and therefore, $\mathbb{E}[Y \mid \mathbf{r}]$ varies across the domains in generic instances that admit this model (Figure 6). One might limit the scope of covariate shift assumption to a subset of features such as 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 strictly larger than the one generated by X_1, because X_1 is determined by $\mathbf{R}$ (Eq. 33). 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 "spurious". This observation shows 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.

Estimation

Transportability formulae such as the one derived in Example 8 provide a strategy for estimation using finite samples drawn from the source distributions. Extensive work in semiparametric statistics concerns robust estimation of the

real-valued mappings defined over the space of probability distributions, such as the transportability formulae [12, 10, 54, 24, 25]. In particular, the work by Jung et al. [26] studies estimation of causal effect using finite samples drawn from multiple source distributions, and is a natural choice for estimating of the transportability formula that are computed by our method. For demonstration purposes, we describe one strategy for obtaining an approximation of the score function $l_\phi(\mathbf{r})$ in the context of Example 8.

Consider fixed $\mathbf{r} \in \text{supp}(\mathbf{R})$, and compute $x_1^{\mathbf{r}}, x_3^{\mathbf{r}}, \bar{r}^{\mathbf{r}}$ as in Eq. 36; this is not a statistical computation, and is possible using merely ϕ. By this change of variables, it suffices to evaluate $Q' : P^*(y \mid x_1^{\mathbf{r}}, x_3^{\mathbf{r}}, \bar{r}^{\mathbf{r}})$ (as in Eq. 37). To evaluate Q' following Eq. 39, it suffices to evaluate the numerator $\tilde{Q} : \sum_{x_2,x_4} P^*(y, x_2, x_4 \mid x_1^{\mathbf{r}}, x_3^{\mathbf{r}}, \bar{r}^{\mathbf{r}})$, and then normalize it for $y \in \{0, 1\}$. Due to Eq. 41, we can express $\tilde{Q}$ as,

$$\begin{aligned}
\tilde{Q} &= \sum_{x_2,x_4} P^*(y, x_2, x_4 \mid x_1^{\mathbf{r}}, x_3^{\mathbf{r}}) && (46)\\
&= \sum_{x_2,x_4} P^1(y, x_2 \mid x_1^{\mathbf{r}}, x_3^{\mathbf{r}}) \cdot P^2(x_4 \mid y, x_1^{\mathbf{r}}, x_2, x_3^{\mathbf{r}}) \\
&= \sum_{x_2,x_4} \overbrace{P^1(x_2 \mid x_1^{\mathbf{r}}, x_3^{\mathbf{r}})}^{\lambda^1} \cdot \overbrace{P^1(y \mid x_1^{\mathbf{r}}, x_2, x_3^{\mathbf{r}})}^{\rho^1} \cdot \underbrace{P^2(x_4 \mid x_1^{\mathbf{r}}, x_2, x_3^{\mathbf{r}})}_{\lambda^2} \cdot \frac{\overbrace{P^2(y \mid x_1^{\mathbf{r}}, x_2, x_3^{\mathbf{r}}, x_4)}^{\rho_a^2}}{\underbrace{P^2(y \mid x_1^{\mathbf{r}}, x_2, x_3^{\mathbf{r}})}_{\rho_b^2}}. && (47)
\end{aligned}$$

Next, we train the generative models λ^1, λ^2 to approximate sampling from $P^1(x_2 \mid x_1, x_3), P^2(x_4 \mid x_1, x_2, x_3)$, respectively. Also, we train the likelihood models $\rho^1, \rho_a^2, \rho_b^2$ to approximate the conditional distributions $P^1(Y = 1 \mid x_1^{\mathbf{r}}, x_2, x_3^{\mathbf{r}}), P^2(Y = 1 \mid x_1^{\mathbf{r}}, x_2, x_3^{\mathbf{r}}, x_4), P^2(Y = 1 \mid x_1^{\mathbf{r}}, x_2, x_3^{\mathbf{r}})$, respectively. Finally, we sample from the generative models and plug them into the likelihood models to approximate $\tilde{Q}$ for label $Y = 1$ according to the following expression for $\tilde{Q}$ in form of expectations:

$$\tilde{Q} = \mathbb{E}_{x_2 \sim \lambda^1(x_1^{\mathbf{r}}, x_3^{\mathbf{r}})}\big[\frac{\rho^1(x_1^{\mathbf{r}}, x_2, x_3^{\mathbf{r}})}{\rho_b^2(x_1^{\mathbf{r}}, x_2, x_3^{\mathbf{r}})} \cdot \mathbb{E}_{x_4 \sim \lambda^2(x_1^{\mathbf{r}}, x_2 x_3^{\mathbf{r}})}[\rho_a^2(x_1^{\mathbf{r}}, x_2, x_3^{\mathbf{r}}, x_4)]\big]. \quad (48)$$

Similarly, we can estimate $\tilde{Q}$ for $y = 0$, and by normalizing the two values we estimate the score function $Q' = l_\phi(\mathbf{r})$.

We implemented the above estimator for randomly generated SCMs $\mathcal{M}^1, \mathcal{M}^2, \mathcal{M}^*$ and the results are shown in Figure 7. 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 data pooled from the source domains. Despite the popularity of this method, we see that due to the mismatch between the target and the source domains, 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, transportable (directly transportable from either of the source domains), and is suggested by the works on invariance-based domain generalization. INV has a better performance compared to ERM, however, the transportability-based classifier (green) performed significantly better as the source data grows. In implementation of the TR method, rejection sampling is used for the generative models, and random forests are used for the likelihood models.

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 a transportability problem based on data.

3 Data-Driven Transportability

The discussions so far assume that the true selection diagram $\mathcal{G}^{\mathbf{\Delta}}$ was available, which is hard to obtain in some settings. 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. Noting that there is no free lunch, and just a trade-off, we begin by imposing a a restriction over the structure of the [unknown] 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 access to the selection diagrams, we need to establish a structural correspondence between the target domain and the source domains.

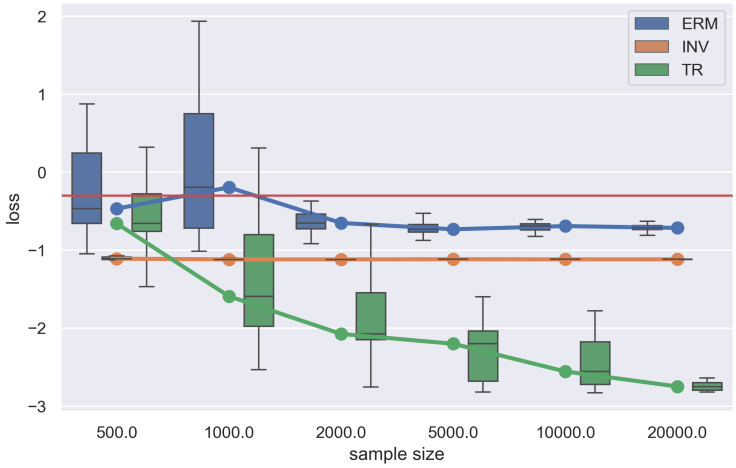


Figure 7: The metric is cross-entropy under the target distribution P^* (the lower, the better). The horizontal axis is the sample size from both source domains.

Assumption 1: Causal Mechanistic Stability (CMS). *The causal diagram of the source and target domains are shared, and for all variables $V \in \mathbf{X} \cup \{Y\}$,*

$$V \notin \bigcup_{i,j=1}^{T} \Delta_{ij} \implies V \notin \bigcup_{k=1}^{T} \Delta_{*k}. \tag{49}$$

In words, if a variable is not present in any of the cross-source discrepancy sets, then it is not present in any of the source-target discrepancy sets. The following example elaborates more on this assumption.

Example 9 (CMS 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, CMS imposes a structure to the selection diagrams as a whole, and is not explicit about the different variables/mechanisms in the SCMs. Next, we need an adjusted definition for transportability according to causal mechanistic stability assumption.

Definition 9 (Transportable representation: 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 CMS 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 $\mathcal{G}^{\boldsymbol{\Delta}}$ with CMS, 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 8 (Figure 6) satisfies CMS, it is not the only selection diagram over these variables that is valid under CMS. 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 CMS. 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. 32 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 CMS for describing the knowledge about the domains. Some readers might find it helpful to note that assuming CMS is weaker than assuming the selection diagrams, while CMS implies the covariate shift assumption. Thus, CMS lies between covariate shift and selection diagrams in terms of expressivity.

An important consequence of CMS 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 causal mechanistic stability 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., CMS 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 CMS 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, CMS 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 9, we have no reason to believe that the pictures taken in the same room during fall come from an SCM that is dissimilar to that of spring and summer domains. Therefore, the domains might be interchangeable in reality.

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 2 (necessity of invariance) *Under causal mechanistic stability 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.,*

$$\mathbb{E}_{P^1}[Y \mid \mathbf{r}] = \mathbb{E}_{P^2}[Y \mid \mathbf{r}] = \cdots = \mathbb{E}_{P^T}[Y \mid \mathbf{r}] = \mathbb{E}_{P^*}[Y \mid \mathbf{r}]. \tag{50}$$

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,*

$$\text{INV}_{ij}[\phi] : \mathbb{E}_{P^i}[Y \mid \mathbf{r}] = \mathbb{E}_{P^j}[Y \mid \mathbf{r}], \quad \forall \mathbf{r} \in \text{supp}(\mathbf{R}). \tag{51}$$

We define the source invariance property as $\bigwedge_{i,j=1}^{T} \text{INV}_{ij}[\phi]$. 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., [44, 2, 46, 34, 11]). In summary, Corollary 2 states that under the causal mechanistic stability assumption, the source invariant property is a necessary condition for transportability of representations.

Is the source invariance property 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 $\text{INV}_{ij}[\phi]$ necessarily corresponds to an s-admissibility condition in the underlying $\mathcal{G}^{\Delta_{ij}}$. This is analogous to c-faithfulness Jaber et al. [23], which is an extension of faithfulness assumption [38] 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 collection of source distributions $\mathbb{P}$ is r-faithful to the underlying selection diagrams $\mathcal{G}^{\boldsymbol{\Delta}}$ if for all representations $\mathbf{R} = \phi(\mathbf{X})$ and for every $i, j \in \{1, 2, \ldots, T\}$,*

$$\text{INV}_{ij}[\phi] \implies S_{ij} \perp\!\!\!\perp_d Y \mid \mathbf{Z}, \bar{\mathbf{R}} \text{ in } \mathcal{G}^{\Delta_{ij}}_{\text{aug}}, \tag{52}$$

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

Under r-faithfulness assumption, we can use the structure that CMS 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 causal mechanistic stability 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. Below, we discuss some of the related work in depth and see them through the lens of data-driven transportability.

3.1 On Magliacane et al. [34]

The authors study the unsupervised domain adaptation (UDA) task, where unlabeled data from the target domain is available as well as labeled data from the source domain. Next, we restate some of the assumptions made by this paper.

- **Assumption 2 (i).** The underlying SCMs are Markovian, i.e., there exists no unobserved common cause. Moreover, the distributions entailed by the graphs are faithful to the causal diagrams, meaning that d=separations in the causal graph and conditional independence relations in the distribution correspond to one another.
- **Assumption 2 (ii).** The conditional independence relations involving Y among the source distributions must hold in the target distribution as well.

On Assumption 2 (i). Assuming Markovianity simplifies the task, to the extent that the discussion about causal identification/transportability in presence of Markovianity becomes almost trivial. This assumption, however, is quite restrictive, and numerous cases lie outside the scope of the problem studied by Magliacane et al. [34]. The faithfulness assumption is analogous to r-faithfulness in this paper; as discussed in Section 3, asserting an instance of faithfulness is necessary to avoid coincidental invariances that do not correspond to the causal graphs, and therefore, are unreliable.

On Assumption 2 (ii). This assumption, akin to the covariate shift assumption, is stated at the distribution level. This is undesirable, because statements at the distributional level can be derived from structural assumptions (such as r-faithfulness and CMS in this paper), so they are essentially the properties implied by structural stability rather than the core assumption. In principle, asserting such distributional properties is not drastically different from assuming that the proposed algorithm generalizes, because such statement is also a distributional property. In contrast with this paper, we assert Assumptions CMS & r-faithfulness corresponding to the SCMs entailing the distribution, which in turn implies the property assumed by the authors (via Theorem 2) for a much more general instance of the problem (Corollary 2).

On representation of X. It is worth mentioning that the work by Magliacane et al. [34] considered more traditional queries, where the representation of $\mathbf{X}$ is only a feature selection, e.g., $\phi(X_1, X_2, X_3) = \langle X_1, X_3 \rangle$. However, our analysis allows for more generic mappings such as Eq. 32 in Example 8.

In conclusion, our results in Section 3 (data-driven transportability) subsumes the result by Magliacane et al. [34] in the following sense:

1. We impose weaker assumptions about the underlying system of SCMs.
2. We consider prediction based on representations of $\mathbf{X}$, as opposed to subsets of it considered by Magliacane et al. [34].
3. We prove that invariance of $Y \mid \phi(\mathbf{X})$ across the source domains implies its generalizability, which is also in-line with Magliacane et al. [34].
4. We also prove that any generalizable (transportable) prediction corresponds to an invariant representation.

□

3.2 On Arjovsky et al. [2].

The authors study a constrained optimization problem called invariant risk minimization (IRM) in the context of domain generalization. In the notation of our paper, IRM is as follows:

$$\begin{aligned} \min_{\phi,h} \quad & \sum_{i=1}^{T} \mathbb{E}_{P^i}[Y \neq h \circ \phi(\mathbf{X})] \\ \text{s.t.} \quad & h \in \underset{\tilde{h}:\text{supp}(\mathbf{R})\to\{0,1\}}{\arg\min} \quad \mathbb{E}_{P^i}[Y \neq \tilde{h} \circ \phi(\mathbf{X})] \quad \forall i \end{aligned} \tag{53}$$

In words, a classifier h composed with a representation $\phi(\mathbf{x})$ satisfy the invariant risk minimization property if $h \circ \phi$ attains the minimum risk across all classifiers defined based on ϕ, across all source domains; we call this the IRM constraint, formally stated in Eq. 53. The search procedure suggests choosing the classifier that satisfies the mentioned constraint, and achieves minimum risk on the pooled source data.

The constrained optimization program above is highly non-convex and difficult to solve. To approximate the solution, the authors consider the Langrangian form below:

$$\min_{\phi,h_\theta} \quad \sum_{i=1}^{T} \mathbb{E}_{P^i}[Y \neq h_\theta \circ \phi(\mathbf{X})] + \lambda \cdot \|\nabla_\theta \mathbb{E}_{P^i}[Y \neq h_\theta \circ \phi(\mathbf{X})]\|^2. \tag{54}$$

In this program, θ parametrizes the classifier h, and the penalization with λ accounts for how restrictive we want to enforce the IRM constraint; for $\lambda = 0$ the objective is the vanilla ERM, and for $\lambda \to \infty$, the solution is guaranteed to satisfy the IRM constraint.

Now, let us consider the representations that satisfy the IRM constraint in Eq. 53. For every distribution P, the optimal classifier defined over a representation ϕ is the Bayes classifier, defined as,

$$h(\mathbf{x}) = \tilde{h}(\phi(\mathbf{x})) = \begin{cases} 1 & \text{if } \mathbb{E}_P[Y \mid \phi(\mathbf{x})] > \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases} \tag{55}$$

In words, the Bayes classifier under the distribution P gives the label $Y = 1$ to the point $\mathbf{x}$ if the empirical score function under P exceeds the threshold $\frac{1}{2}$. This threshold is due to the symmetric form of the loss function used in Eq. 53. Thus, the requirement by the IRM constraint is that the $\frac{1}{2}$ level-sets of the empirical score functions under the source distributions $P^i \in \mathbb{P}$ coincide. In contrast, the source invariance property (Definition 10) requires full equality between the empirical score functions under the source distributions, i.e., equality of all level-sets. In this sense, the IRM constraint is relevant to domain generalization and data-driven transportability, although it is a much weaker constraint compared to the source invariance property.

Criticism

The above is our attempt to find the relevance of IRM to the domain generalization problem by considering its objective in our framework of data-driven transportability. The IRM paper itself justifies no theoretical claims beyond asymptotic properties of the IRM solution for a few specific systems. The paper lacks clear definition of environment/domain, the objective of domain generalization problem is not explicitly stated, and therefore, the assumptions that guarantee the validity of the IRM solution for domain generalization are missing.

One might speculate that a weaker instance of the source invariance property, such as the original IRM in contrast with α-IRM, might yield a slightly weaker solution to the domain generalization problem. The work by Rosenfeld et al. [45] rejects this hypothesis through careful analysis of assumptions underlying the validity of the IRM solution:

- Theorem 5.1 indicates that in the linear setting, the number of domains that need to be observed in the source data must exceed the dimensionality of the domain index to guarantee proper generalization to unseen domains. Otherwise, an predictor that relies on unstable mechanisms (which makes it invalid for domain generalization) would achieve a better value in the IRM optimization problem.
- for arbitrarily small ϵ, Theorem 6.1. constructs a classifier that matches with the invariant risk minimizer on $1 - \epsilon$ fraction of the source distributions, and matches with bayes optimal classifier for the remaining ϵ. This allows this classifier to satisfy the IRM constraint approximately (for large number of source domains), and also achieve a small empirical risk on the source data. The latter justifies this classifier as a plausible solution to the IRM optimization problem, however, the target risk of this classifier would be lower-bounded by a large quantity, making it a poor classifier for domain generalization.

Moreover, the work by Kamath et al. [27] also concerns theoretical validity IRM for domain generalization problem, as summarized below:

1. In section 4, the authors elaborate through a simple example that even the perfect solution to the IRM objective is not guaranteed to have minimum worst-case loss for the target domain; they prove this point by constructing an alternative predictor that violates the IRM objective and yet has superior worst-case target domain risk.
2. In section 5, the authors characterize a setting where satisfying the IRM objective guarantees a stable score function for the target domain; to this end, the authors need to assume that the domain index is continuous, and the mapping between the domain index and joint distribution over $\mathbf{X}, Y$ is an analytic mapping. An analytic mapping has a converging Taylor expansion at every point in the domain. In this sense, the result is intuitive, as local information about this mapping at reveals the entire mapping through a Taylor expansion, thus, in principle, we expect local stability in the source domains to be extrapolated to all possible target domains. In this sense, the assumption of analytic mapping seems to be too strong.

In conclusion, even though the IRM criterion can be viewed as a weaker version of the source invariance property, it would not necessarily guarantee desirable generalization guarantees. In fact, the solution to the IRM optimization fails to generalize in various instances of the problem, though it works when further parametric assumption as imposed. In contrast with IRM, we prove that satisfying the source invariance property, despite its practical challenges, would provably guarantee domain generalization without restrictive parametric assumptions. Next, we propose an enhanced version of the IRM criterion that coincides with the source invariance property, and thus, the implications due to Theorem 2 give us a generalization guarantee.

Fixing IRM

We extend the IRM criterion such that perfectly satisfying it becomes equivalent to the source invariance property, and next, we turn the criterion, akin to the original IRMv1, into a penalized likelihood objective.

Definition 11 (α-loss) *We define a loss function that is class-sensitive:*

$$\mathcal{L}^{\alpha}(y, \hat{y}) := \frac{1}{\alpha} \cdot 1[y = 0, \hat{y} = 1] + \frac{1}{1-\alpha} \cdot 1[y = 1, \hat{y} = 0]. \tag{56}$$

For a classifier $h : \text{supp}(\mathbf{X}) \to \{0, 1\}$, the expected loss under $P(\mathbf{x}, y)$ is called α-risk, and is denoted by,

$$\mathcal{R}^{\alpha}_{P}(h) = \mathbb{E}_{P}[\mathcal{L}^{\alpha}(Y, h(\mathbf{X}))]. \tag{57}$$

Notice that $\frac{1}{2}$-loss is the symmetric 0-1 loss used in the IRM formulation in Eq. 53. As discussed earlier, IRM criterion requires existence of a classifier defined based on ϕ that minimizes $\frac{1}{2}$-risks across all source domains. Next, in IRMv1 the authors write the Lagrangian of the IRM program by adding a term to the main objective that penalizes sub-optimality of the classifier at hand in each of the source domains. Taking a similar approach, we propose the following program.

Definition 12 (α-IRM) *We define an extension of IRM criterion by enforcing coincidence of optimal classifiers of all α-risks;*

$$\begin{aligned} \min_{\phi, \{h^{\alpha}\}_{\alpha \in [0,1]}} \quad & \sum_{i=1}^{T} \mathcal{R}^{\frac{1}{2}}_{P^i}(h^{\frac{1}{2}} \circ \phi) \\ \text{s.t.} \quad & \forall \alpha \in [0,1] \quad \forall i \in \{1, 2, \ldots, T\} : h^{\alpha} \in \underset{\tilde{h}:\text{supp}(\mathbf{R}) \to \{0,1\}}{\arg\min} \mathcal{R}^{\alpha}_{P^i}(\tilde{h} \circ \phi) \end{aligned} \tag{58}$$

In this program, $\{h^{\alpha}\}$ is a family of classifiers. Moreover, the following penalized risk minimization program,

$$\min_{\phi, \{h^{\alpha}_{\theta}\}_{\alpha \in [0,1]}} \quad \sum_{i=1}^{T} \mathcal{R}^{\frac{1}{2}}_{P^i}(h^{\frac{1}{2}}_{\theta} \circ \phi) + \lambda \cdot \int_{0}^{1} \|\nabla_{\theta} \mathcal{R}^{\alpha}_{P^i}(h^{\alpha}_{\theta} \circ \phi)\|^2 \cdot d\alpha \tag{59}$$

where θ parametrizes the class of classifier, is a relaxation of α-IRM in the sense that for $\lambda \to \infty$ the solution of Eq. 59 coincides with that of Eq. 58.

Notice that the IRM constraint is a special case of the α-IRM constraints; precisely, for $\alpha = \frac{1}{2}$, the constraint in Eq.58 coincides with the IRM constraint in Eq. 53. In words, the α-IRM constraints require that for every $0 \leq alpha \leq 1$, there exist a classifier which minimizes α-risk across all source domains. Next, we justify that the α-IRM constraints are equivalent to the source invariance property.

Proposition 1 (Source invariance & α-IRM) *A representation ϕ satisfies the source invariance property (Def. 10) if and only if there exists a class of classifiers $\{h^\alpha\}_{\alpha\in[0,1]}$ which satisfies the α-IRM constraints in Eq. 58. Thus, due to Theorem 2, any solution to the α-IRM program generalizes to the target domain, under r-faithfulness and CMS assumptions.*

As shown by the result above, the extension of IRM that we called α-IRM is a theoretically justified proposal for domain generalization. It remains an open question to be explored in future work to verify whether minimizing the penalized risk in Eq. 58 in practice with finite samples is viable.

On Rothenhäusler et al. [46]

This work is a thorough study of domain generalization in the linear setting, i.e., when the causal mechanisms f_V are linear in terms of observed and unobserved variables. The authors study linear regression problem in the context of domain generalization. They propose the following procedure to obtain a linear estimator of Y for domain generalization.

- For fixed coefficients $\mathbf{b} \in \mathbb{R}^N$, define,

$$\mathbf{a}^{\mathbf{b}} = \langle a_1^{\mathbf{b}}, a_2^{\mathbf{b}}, \dots, a_T^{\mathbf{b}} \rangle := \arg\min_{\mathbf{a}\in\mathbb{R}^T} \sum_{i=1}^{T} \mathbb{E}_{P^i}[(Y - \mathbf{X}^\top \cdot \mathbf{b} - a_i)^2]. \tag{60}$$

  This is the projection of the residual $(Y - \mathbf{X}^\top \cdot \mathbf{b})$ on the span of the T one-hot indicator vectors for each of the source domains.

- Anchor regression coefficient vector $\mathbf{b}^\gamma$ is then obtained as,

$$\mathbf{b}^\gamma := \arg\min_{\mathbf{b}} \sum_{i=1}^{T} \mathbb{E}_{P^i}[(Y - \mathbf{X}^\top \cdot \mathbf{b})^2] + \gamma \cdot \|\mathbf{a}^{\mathbf{b}}\|^2, \tag{61}$$

  where γ indicates the penalization parameter.

In words, the above is a penalized squared-loss minimization, where the penalization term can be interpreted as the statistical dependence of the residuals $Y - \mathbf{X}^\top \cdot \mathbf{b}$ with the domain indicator vectors. In the work by [46], the domain index can be continuous and multi-dimensional, but here we only stated a form of the objective that matches our framework, where the domain indicators are one-hot vectors in $\mathbb{R}^T$.

For $\gamma \to \infty$, the solution to the anchor regression needs to minimize the penalty term $\gamma \cdot \|\mathbf{a}^{\mathbf{b}}\|^2$. This quantity can be no less than zero, which means the residual $Y - \mathbf{X}^\top \cdot \mathbf{b}$ would be orthogonal to (i.e., uncorrelated with) the domain indicator vectors. This objective is closely related to source invariance property, as stated below.

Proposition 2 (Source invariance & Anchor regression) *Let $R = \mathbf{X}^\top \cdot \mathbf{b}$ be a representation. If $\mathbf{b}$ is a solution to the anchor regression objective for $\gamma \to \infty$, then $R = \mathbf{X}^\top \cdot \mathbf{b}$ satisfies the source invariance property, i.e.,*

$$\mathbf{a}^{\mathbf{b}} = \mathbf{0} \implies \mathbb{E}_{P^i}[Y \mid \mathbf{X}^\top \cdot \mathbf{b} = r] = \mathbb{E}_{P^j}[Y \mid \mathbf{X}^\top \cdot \mathbf{b} = r], \forall r \in \mathbb{R}, P^i, P^j \in \mathbb{P}. \tag{62}$$

In words, Proposition 2 shows that any solution to the anchor regression optimization for the extreme case $\gamma \to \infty$ corresponds to a representation that satisfies the source invariance property. Notice, we can not immediately apply Theorem 2 to provide a guarantee for the generalization properties of anchor regression, as Y can be real-valued while our analysis is limited to the classification problem (i.e., binary Y). Extending Theorem 2 to the case of continuous Y is straightforward, but we leave it for future work. Also, in this paper, we assume that the unobserved variables in $\mathbf{U}$ are mutually independent. However, in the framing of Anchor regression, there might exist hidden variables denoted by $\mathbf{H}$ that can cause or be caused by the variables $\mathbf{X}, Y$ as well as the other hidden variables. These disparities makes it challenging to compare the two approaches fully. Thus, we can consider a special case of linear SCMs that can be considered in both frameworks.

Definition 13 (Linear SCMs with shared noise) *In linear SCMs, the set of unobserved variables $\mathbf{U}$ contains K real-valued variables, with a full-support distribution $P(\mathbf{u})$ under which variables in $\mathbf{U}$ are jointly independent. For every SCM $\mathcal{M}^i$ ($i \in \{1, 2, \dots, T, *\}$), the distribution of the noise in the corresponding domain is,*

$$P^i(U_k = u) = P^0(U_k - M_k^i = u), \tag{63}$$

where P^0 is some fixed but unknown distribution, and $\mathbf{M}^i \in \mathbb{R}^K$ is a domain-specific vector of constants. For each variable $V \in \mathbf{X} \cup \{Y\}$, the mechanism determining its value is,

$$V \leftarrow \sum_{V' \in \mathbf{X} \cup \{Y\}}^{N} \beta_{V',V} \cdot V' + \sum_{k=1}^{K} \tau_{k,V} \cdot U_k. \tag{64}$$

We further assume that the sparsity pattern of matrixes β, τ correspond to an acyclic (i.e., non-recursive) model.

In words, in this system of linear SCMs, the noise variables are still jointly independent, however, they may be shared between the causal mechanisms, which might induce confounding effects. Also, the distribution of noise in each domain, namely $P^i(\mathbf{u})$, is assumed to be a shift $\mathbf{M}^i$ applied to a baseline distribution P^0.

Suppose we consider one-hot vectors $\mathbf{A}^i$ ($i \in \{1, 2, \ldots, T\}$) appended to each data point that indicates its domain, i.e., $\mathbf{A}^i_j = 1[i = j]$. Also, let $\mathbf{U}^0$ denote $\mathbf{U}$ drawn from the baseline distribution $P^0(\mathbf{u})$. Then, we can rewrite Eq. 64 as,

$$\mathbf{V} \leftarrow \beta^\top \cdot \mathbf{V} + \tau^\top \cdot (\mathbf{U}^0 + \mathbf{M}^\top \cdot \mathbf{A}), \tag{65}$$

where $\mathbf{M} := [\mathbf{M}^1, \mathbf{M}^2, \ldots, \mathbf{M}^T]$. The above matches Eq. (8) (and the subsequent equation) by Rothenhäusler et al. [46] which describes the data-generation process of their setting, and the matrix $\mathbf{M}$ is called the "shift matrix". For simplicity, we assume that the shift matrix is full-rank. What follows characterizes the selection diagram of this linear system in terms of the structural matrices $\gamma, \tau, \mathbf{M}$.

Proposition 3 (Selection diagram of Anchor regression) *The causal diagram of the linear SCMs defined in Def. 13 is shared across the domains. The selection diagram of the whole system of SCMs can be characterized as follows:*

1. *The directed edge $V' \to V$ is present in $\mathcal{G}^{\mathbf{\Delta}}$ if and only if $\beta_{V',V} \neq 0$.*
2. *The bidirected edge $V' \dashleftarrow\dashrightarrow V$ is present in $\mathcal{G}^{\mathbf{\Delta}}$ if and only if,*
$$\{k : \tau_{k,V} \neq 0\} \cap \{k : \tau_{k,V'} \neq 0\} \neq \emptyset. \tag{66}$$
3. *The cross-source domain discrepancies that are pointed at by the selection node S_{ij} are characterized as,*
$$V \in \Delta_{ij} \iff \tau_{\cdot,V}^\top \cdot \mathbf{M}^i \neq \tau_{\cdot,V}^\top \cdot \mathbf{M}^j. \tag{67}$$

Notice that Prop. 3 does not make any claims about S_{i*} nodes. For that, we need to consider Theorem 1 by [46] that characterizes a collection of plausible target distributions $P^*(\mathbf{x}, y)$ denoted as C^γ. By this result, the solution of the anchor regression problem in Eq. 61, namely b^γ, is min-max optimal under C^γ; this set of distributions is produced by taking different instances of $\mathbf{A}^*$ *close* to $\langle \frac{1}{T}, \frac{1}{T}, \ldots, \frac{1}{T} \rangle$ vector. The parameter γ can be interpreted as a measure of closeness of $\mathbf{A}^*$ to $\langle \frac{1}{T}, \frac{1}{T}, \ldots, \frac{1}{T} \rangle$. Thus, the extreme $\gamma \to \infty$, C^∞ is analogous to having arbitrary $\mathbf{A}^*$. Due to this above property, we can determine the source-target domain discrepancies Δ_{i*}; what follows is a formal statement.

Proposition 4 (Anchor regression and CMS) *The selection diagram of the system for every $\gamma > 1$ satisfies causal mechanistic stability assumption, which means that the source-target discrepancy sets are contained in cross-source discrepancy sets, i.e.,*

$$\Delta_{l*} \subseteq \bigcup_{1 \leq i,j \leq T} \Delta_{ij}, \quad \forall 1 \leq l \leq T. \tag{68}$$

Proof: *Consider a variable $V \in \mathbf{X} \cup \{Y\}$ whose mechanism has changed (i.e., unstable) across the sources, i.e., $V \in \bigcup_{1 \leq i,j \leq T} \Delta_{ij}$. By definition of domain discrepancy sets, this variable might lies in Δ_{l*} if there exists a plausible target domain for which the mechanism for V is different between $\mathcal{M}^l$ and $\mathcal{M}^*$. Next, we construct such target domain.*

As we assumed $V \in \bigcup_{1 \leq i,j \leq T} \Delta_{ij}$, it is convenient to assume that $V \in \Delta_{i^0,j^0}$ for some indices i^0, j^0. Due to line (3) of Proposition 3,

$$\underbrace{\tau_{\cdot,V}^\top \cdot \mathbf{M}^{i^0}}_{e} \neq \underbrace{\tau_{\cdot,V}^\top \cdot \mathbf{M}^{j^0}}_{e'}. \tag{69}$$

Now consider the quantity $\tau_{\cdot,V}^\top \cdot \mathbf{M}^l$; it can be equal to at most one of e, e', and therefore, it is necessarily unequal to one of them. Without loss of generality, suppose,

$$\underbrace{\tau_{\cdot,V}^\top \cdot \mathbf{M}^l}_{\tilde{e}} \neq \underbrace{\tau_{\cdot,V}^\top \cdot \mathbf{M}^{j^0}}_{e'}. \tag{70}$$

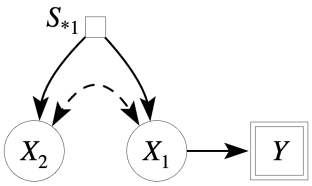


Figure 8: The selection diagram corresponding to Example 10

Due to Eq. 65, plausible target domains can be obtained by choices of $\mathbf{A}^*$*. These choices, however, are restricted to be close to the average of all one-hot* $\mathbf{A}^i$ *vectors. In particular for* $\gamma > 1$*, Rothenhäusler et al. [46] shows that the choice of* $\mathbf{A}^i$ *are valid for* $\mathbf{A}^*$ *for every* $1 \leq i \leq T$*, which means that the target domain can possibly be identical to each of the source domain. Lemma 1 (interchangeable domains) indicates that this property is the key to utility of invariance-based approaches for generalization. This approves that the choice of* $\tilde{\mathbf{A}^*} = \mathbf{A}^{j^0}$*, yields a plausible target domains, however,*

$$\tau_{\cdot,V}^{\top} \cdot (\mathbf{M}^{\top} \cdot \tilde{\mathbf{A}^*}) = \tau_{\cdot,V}^{\top} \cdot (\mathbf{M}^{\top} \cdot \mathbf{A}^{j^0}) \tag{71}$$

$$= \tau_{\cdot,V}^{\top} \cdot \mathbf{M}^{j^0} \tag{72}$$

$$\neq \tau_{\cdot,V}^{\top} \cdot (\mathbf{M}^{l}) \qquad \textit{(due to Equation 70).} \tag{73}$$

Therefore, $V \in \Delta_{l*}$*. The choice of* l *was arbitrary, which proves the objective.* □

In conclusion, our results in section 3 validate the approach taken by the authors. We emphasized the connections between the two approaches, and showed that source invariance property, as exploited by the authors in the case of linear models, can indeed be applied for generalization in a broader set of models with non-parametric mechanisms. □

3.3 On Ben-David et al. [7] & Ganin et al. [17]

The work by Ben-David et al. [7] focuses on learning representations for the unsupervised domain adaptation task, where unlabeled samples drawn from the target distribution $P^*(\mathbf{x})$ are available as well as labeled data drawn from the source distributions $\mathbb{P}$. Under the assumption that the target distribution is close to the source distribution in some distance measure $d_{\mathcal{A}}$, the authors prove a bound for the difference of classification risk in the target domain and the source domain that involves the term $d_{\mathcal{A}}(P(\phi(\mathbf{x})), P^*(\phi(\mathbf{x})))$ (Theorem 2 in Ben-David et al. [7]). Motivated by this observation, the authors suggest that an effective representation for domain generalization is characterized by its ability to prevent an algorithm from determining the original domain based on the input $\mathbf{X}$. Precisely, a representation is preferred in this proposal if,

- It yield small classification error on source distribution;
- It yield *similar* distributions $P(\phi(\mathbf{x})), P^*(\phi(\mathbf{x}))$ in the source and target domains.

The minimum possible value for the distance measure $d_{\mathcal{A}}(P(\phi(\mathbf{x})), P^*(\phi(\mathbf{x})))$ would be equal to zero, which is the case for choices to representations such as $\mathbf{R} = \phi(\mathbf{X})$ where $P(\mathbf{r}) = P^*(\mathbf{r})$.

Inspired by the above proposal, Ganin et al. [17] implements a method called domain-adversarial neural network (DANN) for learning such representations using high-dimensional data. In this subsection, we aim to understand the original proposal theoretically, and refer to the main idea as DANN for convenience. Next, we consider DANN in our framework, and we focus on a special instance of the domain adaptation task.

Example 10 (A failure case of DANN) Consider the following SCMs $\mathcal{M}^1$ (left) and $\mathcal{M}^*$ (right):

$$\begin{aligned} U &\sim \text{unif}(\{1, 2, \ldots, 100\}) \\ U_Y &\sim \mathcal{N}(0, 1) \\ X_1 &\leftarrow \alpha_1 \cdot U \\ X_2 &\leftarrow \alpha_2 \cdot U \\ Y &\leftarrow \begin{cases} 1 \text{ if } X + U_Y \geq c \\ 0 \text{ otherwise.} \end{cases} \end{aligned} \qquad \begin{aligned} U &\sim \text{unif}(\{1, 2, \ldots, 100\}) \\ U_Y &\sim \mathcal{N}(0, 1) \\ X_1 &\leftarrow \alpha_1 \cdot U + \delta_1 \\ X_2 &\leftarrow \alpha_2 \cdot U + \delta_2 \\ Y &\leftarrow \begin{cases} 1 \text{ if } X + U_Y \geq c \\ 0 \text{ otherwise.} \end{cases} \end{aligned} \tag{74}$$

As seen above, the distribution of X_1, X_2 is shifted between the source and target domains, while the mechanism for Y remains unchanged. The selection diagram corresponding to this system is shown in Figure 8. The ideal representation, according to DANN's objective, is the one that minimizes the difference between distribution of $\phi(X_1, X_2)$ in the source and target domains, as well as maximizing the prediction accuracy in the source domain. Consider the representation,

$$R = \phi(X_1, X_2) := X_1 - \frac{\delta_1}{\delta_2} \cdot X_2. \tag{75}$$

We derive the distribution of this representation in both domains:

$$\begin{aligned}
P^1(R = r) &= P^1(X_1 - \frac{\delta_1}{\delta_2} \cdot X_2 = r) && (76)\\
&= P^1(\alpha_1 \cdot U - \frac{\delta_1}{\delta_2} \cdot (\alpha_2 \cdot U) = r) && (77)\\
&= P^1((\alpha_1 - \frac{\delta_1 \cdot \alpha_2}{\delta_2}) \cdot U = r) && (78)\\
&= P^1(U = \frac{r}{\alpha_1 - \frac{\delta_1 \cdot \alpha_2}{\delta_2}}) && (79)\\
&= \begin{cases} \frac{1}{100} \text{ if } \frac{r}{\alpha_1 - \frac{\delta_1 \cdot \alpha_2}{\delta_2}} \in \{1, 2, \ldots, 100\} \\ 0 \text{ otherwise.} \end{cases} && (80)
\end{aligned}$$

$$\begin{aligned}
P^1 * (R = r) &= P^*(X_1 - \frac{\delta_1}{\delta_2} \cdot X_2 = r) && (81)\\
&= P^*(\alpha_1 \cdot U + \cancel{\delta_1} - \frac{\delta_1 \cdot \alpha_2}{\delta_2} \cdot U - \cancel{\frac{\delta_1}{\delta_2} \cdot \delta_2} = r) && (82)\\
&= P^*((\alpha_1 - \frac{\delta_1 \cdot \alpha_2}{\delta_2}) \cdot U = r) && (83)\\
&= P^*(U = \frac{r}{\alpha_1 - \frac{\delta_1 \cdot \alpha_2}{\delta_2}}) && (84)\\
&= \begin{cases} \frac{1}{100} \text{ if } \frac{r}{\alpha_1 - \frac{\delta_1 \cdot \alpha_2}{\delta_2}} \in \{1, 2, \ldots, 100\} \\ 0 \text{ otherwise.} \end{cases} && (85)
\end{aligned}$$

The above derivation holds as long as $\alpha_1 \neq \frac{\delta_1 \cdot \alpha_2}{\delta_2}$, which is true in non-degenerate cases. As seen above, the distribution of this representation matches perfectly between source and target domains; this means $d_{\mathcal{A}}(P(\phi(x_1, x_2)), P^*(\phi(x_1, x_2))) = 0$, which is its minimum. Moreover, the optimal Bayes classifier in the source domain can be obtained using this representation;

$$\begin{aligned}
h^1_{\text{bayes}}(X_1, X_2) &:= 1[X_1 \geq c] && (86)\\
&= 1[U \geq c] && (87)\\
&= 1[\frac{r}{\alpha_1 - \frac{\delta_1 \cdot \alpha_2}{\delta_2}} \geq c] && (88)\\
&= 1[r \geq c \cdot (\alpha_1 - \frac{\delta_1 \cdot \alpha_2}{\delta_2})] && (89)\\
&= 1[\phi(X_1, X_2) \geq c \cdot (\alpha_1 - \frac{\delta_1 \cdot \alpha_2}{\delta_2})] =: \hat{h}_\phi(X_1, X_2). && (90)
\end{aligned}$$

The latter is a function of the representation $\phi(X_1, X_2) = X_1 - \frac{\delta_1}{\delta_2} \cdot X_2$, which means that one can, in principle, learn the optimal Bayes classifier of the source domain by thresholding this representation using large enough data. In summary, this representation has both of the desired properties of DANN, however, in what follows we demonstrate that this approach performs poorly for the task of domain adaptation.

The optimal Bayes classifier in the target domain can still be expressed as a threshold over $\phi(X_1, X_2)$;

$$h^*_{\text{bayes}}(X_1, X_2) := 1[X_1 \geq c] \tag{91}$$
$$= 1[U + \delta_1 \geq c] \tag{92}$$
$$= 1[\frac{r}{\alpha_1 - \frac{\delta_1 \cdot \alpha_2}{\delta_2}} + \delta_1 \geq c] \tag{93}$$
$$= 1[r \geq (c - \delta_1) \cdot (\alpha_1 - \frac{\delta_1 \cdot \alpha_2}{\delta_2})] \tag{94}$$
$$= 1[\phi(X_1, X_2) \geq (c - \delta_1) \cdot (\alpha_1 - \frac{\delta_1 \cdot \alpha_2}{\delta_2})]. \tag{95}$$

This threshold, however, is $|\delta_1 \cdot (\alpha_1 - \frac{\delta_1 \cdot \alpha_2}{\delta_2})|$ units away from the threshold used in $\hat{h}_\phi(X_1, X_2)$ (Eq. 90). Notice that this difference grows linearly with δ_1, which is the difference of the mean of X_1 across the source and target domain. Thus, despite the fact that ϕ is absolutely ideal w.r.t. the proposal made by Ben-David et al. [7] (DANN), it has a possibly significant bias when deployed in the target domain, even if we observe unlimited data from the source and unlabeled data from the target. Finally, note that discarding X_2 in the a representation, such as in $\tilde{\phi}(X_1, X_2) = X_1$, would yield optimal classification in both source and target domains.

The example above indicates that the proposal made by Ben-David et al. [7] (and implemented using neural networks by Ganin et al. [17]) is suboptimal for finding representations suitable for unsupervised domain adaptation.

3.4 On balancing prediction error

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., [57, 46, 29, 41, 2]. Below, we discuss the work by Krueger et al. [29] as an example, and propose our version of error balancing criterion with regard to the data-driven transportability.

Risk Extrapolation (MM-REx)

The authors take an approach similar to distributionally robust optimization (DRO) by Ben-Tal et al. [8], where it is assumed that the target distributions lies inside the convex hull of the source distribution. Although, DRO was not proposed for domain generalization task initially, REx can be viewed as an extension of DRO that is suggested for domain generalization task by Krueger et al. [29]. In REx the authors assume that the distribution entailed by the target SCM must lie inside or *close to* the convex hull of the distributions entailed by the source SCMs. Precisely, they propose the following.

Definition 14 (MM-REx) *In the notation of this paper, the MM-REx predictor is the solution to the following optimization problem,*

$$h^{\text{MMREx}} \in \underset{h:\text{supp}(\mathbf{X})\to\{0,1\}}{\arg\min} \max_{\{\lambda_i\}_{i=1}^T \geq \lambda_{\min}} \sum_{i=1}^{T} \lambda_i \mathcal{R}_{P^i}(h), \tag{96}$$

where $-\infty < \lambda_{\min} \leq 0$ is the extrapolation parameter. This objective can be rewritten as,

$$h^{\text{MMREx}} \in \underset{h:\text{supp}(\mathbf{X})\to\{0,1\}}{\arg\min} (1 - T\lambda_{\min}) \cdot \max_{1\leq i\leq T} \mathcal{R}_{P^i}(h) + \lambda_{\min} \cdot \sum_{i=1}^{T} \mathcal{R}_{P^i}(h). \tag{97}$$

For the special case of $\lambda_{\min} = 0$, the above coincides with DRO. Smaller choices of $\lambda_{\min}$ consider larger geometric extrapolation in the space of distributions. According to [29], in the extreme case of $\lambda_{min} \to -\infty$, presence of the penalty term,

$$(1 - T\lambda_{\min}) \cdot \max_{1\leq i\leq T} \mathcal{R}_{P^i}(h), \tag{98}$$

enforces strict equality between risks across the source domains. Motivated by this observation, the authors propose a similar objective that is claimed to be more stable and efficient, as described below.

Definition 15 (Variance Risk Extrapolation (V-REx).) *In the notation of this paper, the V-REx predictor is the solution to the following optimization problem,*

$$h^{\text{VREx}} \in \underset{h:\text{supp}(\mathbf{X})\to\{0,1\}}{\arg\min} \beta \cdot \text{Var}(\{\mathcal{R}_{P^1}(h), \mathcal{R}_{P^2}(h) \ldots, \mathcal{R}_{P^T}(h)\}) + \sum_{i=1}^{T} \mathcal{R}_{P^i}(h), \tag{99}$$

where $0 < \beta \leq \infty$ penalizes the variation in the value of the risk across the source domains.

For the special case of $\beta = 0$, the above coincides with ERM. Larger choices of β consider the classifiers with less variation in their risk across the source domains. In the extreme case of $\beta \to \infty$, presence of the penalty term,

$$\text{Var}(\{\mathcal{R}_{P^1}(h), \mathcal{R}_{P^2}(h) \ldots, \mathcal{R}_{P^T}(h)\}), \tag{100}$$

enforces strict equality between risks across the source domains. The objective of both MM-REx and V-REx is to equalize/balance the classification error across the source domains. The authors study effectivity of these predictors using tools from causal inference. Below, we restate the main theoretical claims of the paper.

- **Theorem 1.** For convenience, suppose $Y : X_0$. Consider a linear SCMs, where $P(\mathbf{u})$ is a normal distribution, and $X_i \leftarrow \sum_{j \neq i} \beta_{j,i} \cdot X_j + U_i$. If the source domains contain hard do-intervention over each of the variables for at least 3 distinct values (total of at least $3 \cdot N$ domains), then equalizing the risk across the source domains recovers the true coefficients for $Y : X_0$.
- **Theorem 2.** In a more general case, assuming that we have access to all possible hard interventions $do(\mathbf{Z} \leftarrow \mathbf{z})$ where $\mathbf{Z} \subseteq \mathbf{X}$, equalizing the risks across the these distributions recovers the true causal mechanism f_Y.

The results above rely on three assumptions about the underlying SCMs:

1. The label Y is not confounded with any other observable variable, i.e., there exists no common cause between Y and some X_i.
2. Each of the domains (source or target) corresponds to a hard do-intervention.
3. Homoskedasticity: a slight generalization of additive noise as used by Peters et al. [40].

In both statements, it is assumed that we have access to vast class of hard interventions. In this setting, the fact that the parents of Y are recoverable is not surprising, because the only variables that are consistently dependent on Y (i.e., $Y \not\perp\!\!\!\perp X_i$) across all these interventions constitute the set $\mathbf{Pa}_Y$. Once this set is identified, the relation between Y and its parents can be recovered by choosing a proper regression method. It is, however, interesting to see that equalizing the risk (MM-REx & V-REx) automatically yields the desired outcome. As we show in the next example, REx looses its charm even in very simple examples where domains are not merely do-interventions.

Example 11 (REx failure case) Consider the SCMs $\mathcal{M}^1$ (left) and $\mathcal{M}^2$ (right);

$$\begin{aligned} U_X &\sim \text{Bern}(0.9) \\ U_{Y1} &\sim \text{Bern}(0.9) \\ U_{Y0} &\sim \text{Bern}(0.6) \\ X &\leftarrow U_X \\ Y &\leftarrow \begin{cases} U_{Y1} & \text{if } X = 1 \\ U_{Y0} & \text{otherwise.} \end{cases} \end{aligned} \qquad \begin{aligned} U_X &\sim \text{Bern}(0.1) \\ U_{Y1} &\sim \text{Bern}(0.9) \\ U_{Y0} &\sim \text{Bern}(0.6) \\ X &\leftarrow U_X \\ Y &\leftarrow \begin{cases} U_{Y1} & \text{if } X = 1 \\ U_{Y0} & \text{otherwise.} \end{cases} \end{aligned} \tag{101}$$

The selection diagram corresponding to this example is $S \to X \to Y$; this example is also known as the covariate shift problem where the distribution of X changes while the distribution of $Y \mid X$ remain unchanged. In this case, the constant classifier $h(x) = 1$ is Bayes optimal in both of the domains, and remains Bayes optimal under arbitrary changes on the mechanism of X. However,

$$\begin{aligned} \mathcal{R}_{P^i}(h) &= P^i(Y \neq \overbrace{h(X)}^{1}) && && (102) \\ &= P^i(Y = 0 \mid X = 1) \cdot P^i(X = 1) + P^i(Y = 1 \mid X = 0) \cdot P^i(X = 0) && \text{law of total prob.} && (103) \\ &= 0.1 \cdot P^i(X = 1) + 0.4 \cdot P(X = 0) && \text{from SCMs defn.} && (104) \\ & && && (105) \end{aligned}$$

Thus, the risks in the source domains are,

$$\mathcal{R}_{P^1}(h) = 0.1 \cdot 0.9 + 0.4 \cdot 0.1 = \mathbf{0.15}, \tag{106}$$

$$\mathcal{R}_{P^2}(h) = 0.1 \cdot 0.1 + 0.4 \cdot 0.9 = \mathbf{0.37}. \tag{107}$$

Despite optimality of h in source and target distributions, we observe that there exists a significant mismatch between its risks across the source domains.

Example 11 highlights that in some generic cases, any encouragement for equalizing/balancing the risk across the source domains would sacrifice optimality in exchange for no clear benefit. This example refutes the utility of MM-REx and V-REx criteria for domain generalization in the framework studied in this paper.

We contemplate the key idea behind REx; it is expected that the target distribution somehow has a geometric correspondence to the source distributions. Even though the authors attempt to draw connections with causality (e.g., Theorem 1 & 2 of the paper), this main presumption is stated at the distribution level, agnostic to the causal mechanisms generating the data. Although use of geometry of the probability distributions is widespread throughout the robustness literature [8], it is still unclear if generalizing to the unseen domains can be dealt similarly. In contrast with [29], our results (e.g., Theorem 2) rely on assumptions about the differences between the mechanisms which entail the distributions in the domains. We prove that a distributional property such as source invariance (Def. 10) emerges as a consequence of structural assumption, and turns out to be a sounds and complete for transportability, making it a solid proposal for domain generalization.

Witnessed by Example 11, equalizing the risk across the source distributions does not necessarily characterize a transportable representation. However, we discover that balancing more refined notions of prediction error yields generalizable predictors; we elaborate below.

Balanced-rate Classifiers

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

$$\text{FOR}_P(h) := P(Y = 1 \mid h(\mathbf{X}) = 0), \tag{108}$$
$$\text{FDR}_P(h) := P(Y = 0 \mid e(\mathbf{X}) = 1). \tag{109}$$

one can verify that predictor $h(X) = 1$ in Example 11 has equal FDR and FOR across the sources. Motivated by this observation, we consider equalizing FDR and FOR across the source domains.

Risk and FOR/FDR are related; it is notable that the classification risk can be expressed as a weighted average of FOR and FDR;

$$\begin{aligned}
\mathcal{R}_{P^i}(h) &= P^i(Y \neq h(\mathbf{X})) && (110)\\
&= P^i(Y = 0, h(\mathbf{X}) = 1) + P^i(Y = 1, h(\mathbf{X}) = 0) && (111)\\
&= P^i(Y = 0 \mid h(\mathbf{X}) = 1) \cdot \underbrace{P^i(h(\mathbf{X}) = 1)}_{q} + P^i(Y = 1 \mid h(\mathbf{X}) = 0) \cdot P^i(h(\mathbf{X}) = 0) && (112)\\
&= q \cdot \text{FOR}_{P^i}(h) + (1 - q) \cdot \text{FDR}_{P^i}(h). && (113)
\end{aligned}$$

Notably, despite the above relation between risk and FDR/FOR, equalizing risk across the domains is not particularly relevant to equalizing FDR and FOR. This is due to the coefficient $q : P^i(h(X))$ in Eq. 113 that might differ across the domains. While the proposal of equalizing risk by Krueger et al. [29] was shown to be ineffective in our setting, below we show that equalizing FDR and FOR across the source distributions implies the source invariance property, and in turn, yields a generalizable classifier.

Suppose a classifier $h^e(\mathbf{X})$ attains equal false omission rate and false discovery rate across all source domains, i.e.,

$$\text{FOR}_{P^1}(h^e) = \text{FOR}_{P^2}(h^e) = \cdots = \text{FOR}_{P^T}(h^e) \tag{114}$$
$$\text{FDR}_{P^1}(h^e) = \text{FDR}_{P^2}(h^e) = \cdots = \text{FDR}_{P^T}(h^e). \tag{115}$$

The function h^e can be viewed as a representation with a binary support $\{0, 1\}$. In that case, one can verify that by Definition 10, h^e satisfies the source invariance property. Thus, under r-faithfulness and CMS assumptions, due to Theorem 2, h^e is transportable (invariant), and therefore, its false omission rate and false discovery rate must match in the unseen target domain as well. This observation justifies the benefit of equalizing these specific notions of prediction error across the source distributions as the means to domain generalization. Motivated by this, we propose the solution of the following optimization problems for domain generalization.

Definition 16 (Balanced-rate classification) *A balanced-rate classifier is a solution to the following penalized ERM;*

$$h^\gamma_{\text{ri}} \in \min_{h:\text{supp}(\mathbf{X})\to\{0,1\}} \gamma \cdot [\text{Var}\big(\{\text{FOR}_{P^i}(h)\}_{i=1}^T\big) + \text{Var}\big(\{\text{FDR}_{P^i}(h)\}_{i=1}^T\big)] + \sum_{i=1}^T \mathcal{R}_{P^i}(h). \tag{116}$$

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;

$$h_{\text{ri}}^{\infty} \in \min_{h:\text{supp}(\mathbf{X})\to\{0,1\}} \sum_{i=1}^{T} \mathcal{R}_{P^i}(h) \tag{117}$$

$$\text{s.t.} \quad \text{FOR}_{P^i}(h) = \text{FOR}_{P^j}(h), \quad \forall P^i \in \mathbb{P} \tag{118}$$

$$\text{FDR}_{P^i}(h) = \text{FDR}_{P^j}(h), \quad \forall P^i \in \mathbb{P}. \tag{119}$$

In words, balanced-rate classification aims to find a classifier for which FDR and FOR are as close as possible across the source domains. Depending on the choice of γ, this objective is prioritized over the regular ERM. This form of penalization using variance is inspired by Krueger et al. [29]. It remains a future work to assess the effectivity of this method in practice. Below, we state the generalization guarantee of balanced-rate classification.

Corollary 3 (Balanced-rate classifiers generalize to the unseen domain.) *Under r-faithfulness and CMS assumptions, any solution to the extreme case of balanced-rate classification (Def. 16, namely h_{ri}^{∞}, is a generalizable classifier. Particularly,*

$$\text{FDR}_{P^*}(h_{\text{ri}}^{\infty}) = \text{FDR}_{P^1}(h_{\text{ri}}^{\infty}) = \cdots = \text{FDR}_{P^T}(h_{\text{ri}}^{\infty}) \tag{120}$$

$$\text{FOR}_{P^*}(h_{\text{ri}}^{\infty}) = \text{FOR}_{P^1}(h_{\text{ri}}^{\infty}) = \cdots = \text{FOR}_{P^T}(h_{\text{ri}}^{\infty}). \tag{121}$$

In conclusion, the idea of balancing prediction error across the source distributions is indeed relevant to the domain generalization task. We formulated a scheme to find balanced-rate classifiers by seeking 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.

3.5 On Wald et al. [57]

In this work, the authors study the relation between multi-calibration and domain generalization.

Definition 17 (Multi-Calibrated 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,*

$$\forall e \in [0, 1] : \mathbb{E}_P[Y \mid \psi(\mathbf{X}) = e] = e. \tag{122}$$

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. A multi-calibrated score ψ can be viewed as a representation with support $[0, 1]$. One can verify that, by definition, multi-calibrated scores satisfy the source invariance property (Def. 10). Due to Theorem 2, we deduce that under r-faithfulness and CMS assumptions, ψ is transportable, i.e.,

$$\mathbb{E}_{P^*}[Y \mid \psi] = \mathbb{E}_{P^1}[Y \mid \psi] = \cdots = \mathbb{E}_{P^T}[Y \mid \psi]. \tag{123}$$

This observation serves as a theoretical justification for use of multi-calibration as a criterion for domain generalization. This finding is in-line with the claims made in Wald et al. [57]; which we iterate below.

- **Theorem 1 by Wald et al. [57]** The SCM for domain i is defined as follows;

$$U_Y \sim \text{Bernoulli}(\eta) \tag{124}$$

$$U_{\text{ac-ns}} \sim \mathcal{N}(\mathbf{0}, \Sigma_{\text{ns}}) \tag{125}$$

$$U_{\text{ac-sp}} \sim \mathcal{N}(0, \Sigma^{(i)}) \tag{126}$$

$$Y \leftarrow U_Y \tag{127}$$

$$X_{\text{ac-ns}} \leftarrow (y - 1/2) \cdot \mu_{\text{ns}} + U_{\text{ac-ns}} \tag{128}$$

$$X_{\text{ac-sp}} \leftarrow\sim (y - 1/2) \cdot \mu^{(i)} + U_{\text{ac-sp}}. \tag{129}$$

  The selection diagram corresponding to this system of SCMs is shown in Figure 9a. To predict Y based on $\mathbf{X} = \langle X_{\text{ac-ns}}, X_{\text{ac-sp}} \rangle$ - which are *anti-causal non-spurious* and *anti-causal spurious* features, respectively -

the authors consider a score $\psi(\mathbf{X}) = \sigma(\alpha^\top X + c) \in [0, 1]$, where σ is the sigmoid function. Since the mean of spurious features $X_{\text{ac-sp}}$, namely $\mu^{(i)}$, is dependent on Y, these features can help predict the label in some domains. Yet, these correlations do not carry to all domains, and $\psi(\mathbf{x})$ might rely on spurious correlations whenever the coefficients in α corresponding to X_{ac-sp} are non-zero. Such scores can carry an arbitrarily high risk in an unseen domain, because a new domain can alter the correlations observed in $\mathbb{P}$. Given these definitions, the result can be restated as below.

Statement. Given $T \geq 2 \cdot d_{\text{sp}}$ training domains (d_{sp} is the dimension of $X_{\text{ac-sp}}$) where data is generated according to the SCMs above, they are said to lie in general position if for all non-zero $x \in \mathbb{R}^{d_{sp}}$:

$$\dim\left(\text{span}\left\{\begin{bmatrix}\Sigma^{(i)}\cdot\mathbf{x}+\mu^{(i)}\\ 1\end{bmatrix}\right\}_{i=1}^{T}\right) = d_{\text{sp}} + 1. \tag{130}$$

If a score such as ψ is multi-calibrated on the T source distributions that lie in general position, then the coefficients of ψ for the features X_{ac-sp} are zero.

Comments. Theorem 1 states that for this specific instance of the problem, upon providing a set of *diverse* domains, any multi-calibrated linear score such as ψ automatically distinguishes the spurious/unstable features from the non-spurious/stable features.

- **Theorem 2 by Wald et al. [57]** Let $f(\mathbf{x}) = \alpha^\top \mathbf{x}$ be a linear score, and assume we have data collected from $k > \max\{d_c + 2, d_{sp}\}$ source domains, following the following specification for SCM $\mathcal{M}^i$:

$$U_c \sim \mathcal{N}(0, 1) \tag{131}$$
$$U_Y \sim \mathcal{N}(0, \sigma_Y^2) \tag{132}$$
$$U_{\text{ac-sp}} \sim \mathcal{N}(\mathbf{0}, \Sigma^{(i)}) \tag{133}$$
$$X_c \leftarrow \mu_c^{(i)} + \Sigma_c^{(i)} \cdot U_c \tag{134}$$
$$Y \leftarrow {\alpha_c^*}^\top \cdot X_c + U_Y \tag{135}$$
$$X_{\text{ac-sp}} \leftarrow Y \cdot \mu^{(i)} + U_Y. \tag{136}$$

The selection diagram corresponding to these SCMs is shown in Figure 9b. Under mild non-degeneracy conditions, if the regressor is multi-calibrated across all source domains, then the coefficients corresponding to $\mathbf{X}_c$ equal to α_c^* and those that correspond to $\mathbf{X}_{\text{ac-sp}}$ are zero.

Comment. Theorem 2 showcases another instance of the domain generalization task, this time involving a set of causal features and a set of effects of Y whose mechanism changes across the domains.

In conclusion, despite theoretical validity of multi-calibration as a criterion that ensures transportability (i.e., generalization), the theoretical guarantees provided in the work by Wald et al. [57] is limited to two linear instances of the problem.

As discussed earlier, multi-calibration is a special case of source invariance property. Lemma 1 by Wald et al. [57] shows that, in fact, invariant scores (with support of $[0, 1]$) can be transformed into multi-calibrated scores. We extend this result by showing a close connection between source invariance property of arbitrary representations and multi-calibration of the scores.

Lemma 2 (Source invariance & multi-calibration) *If any representation $\mathbf{R} = \phi(\mathbf{X})$ satisfies the source invariance property, which is,*

$$\text{INV}_{ij}(\phi) : \mathbb{E}_{P^i}[Y \mid \phi(\mathbf{X}) = \mathbf{r}] = \mathbb{E}_{P^j}[Y \mid \phi(\mathbf{X}) = \mathbf{r}], \quad \forall P^i, P^j \in \mathbb{P}, \tag{137}$$

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 multi-calibrated. Moreover, for every $P^i \in \mathbb{P}$,

$$I_{P^i}(Y; \phi(\mathbf{X})) = I_{P^i}(Y; \psi(\mathbf{X})), \tag{138}$$

meaning that ϕ and ψ have equivalent prediction power on all source domains.

In words, Proposition 2 states a that for every representation that which satisfies the source invariance property, there exists a multi-calibrated score with equivalent prediction power. It is also evident from Lemma 2 that there exists an invariant representation if and only if there exists a multi-calibrated score. Thus, to find an invariant representation for domain generalization purposes, one can limit the search space to the scores only and search for a calibrated score.

Figure 9: (a) The score function $P^*(Y = 1|\sum_i X_i = r)$ for a randomly generated example; the bounds indicate how much variation one might see in the score function for different plausible target SCMs $\mathcal{M}^*$. (b) The score function $P^*(Y = 1|\beta^\top \cdot \mathbf{X} = r)$ for the same instance of the problem; as seen here the bounds are collapsed, meaning that the score function is unique across all plausible target SCMs, which allows us to have generalizable prediction based on this representation.

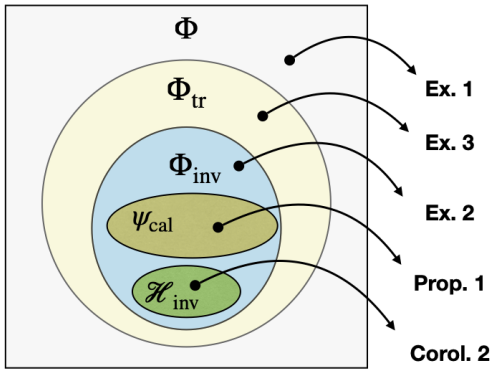


Figure 10: Φ denotes the set of all rep.; Φ_{tr} denotes the class of transportable rep.; Φ_{inv} denotes the class of invariant rep.; Ψ_{cal} denotes the class of calibrated score functions; $\mathcal{H}_{\text{inv}}$ denotes the class of invariant classifiers

3.6 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. 5 in Example 1, and some are transportable, e.g., Eq.11 in Example **??**. The latter is not only transportable but also invariant, i.e., $\mathbb{E}[Y \mid r]$ matches across the source and target domains. Some representations are transportable but not invariant, e.g., Eq. 32 in Example 8. The classifiers that have balanced false negative and false discovery rates across the source and target domains (such as h^e in subsection 3.4) constitute a subset of invariant representations that we call invariant classifiers. The class of multi-calibrated scores is another subclass of invariant representation that is equivalent to it for prediction. Under r-faithfulness and CMS assumptions, the class of transportable representations collapses to invariant representations. In conclusion, our findings suggest the taxonomy in Figure 10 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 of classifications made based on that representation. Our findings unified 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.

Acknowledgements

This manuscript benefits from the feedback and discussions with Bruno Ribeiro. This research was supported in part by the NSF, ONR, AFOSR, DARPA, DoE, Amazon, JP Morgan, and The Alfred P. Sloan Foundation.

A Proofs

A.1 Proof of invertibility in Example 1

We prove by induction on N.

Base: For $N = 1$, $R = \beta \cdot X_1$ has only one element in its support, and the inverse is trivial.

Step: Suppose we can compute $\phi^{-1}(\tilde{R})$ which is the inverse of the function $\tilde{R} = \phi(\mathbf{X}) = \beta^T \cdot \mathbf{X}$. Let

$$R = \beta^T \cdot \mathbf{X} + b \cdot X_{N+1}, \tag{139}$$

where $b \sim \mathcal{N}(0, 1)$. We are given the value $R = r$, and the goal is to recover the value of both $\mathbf{X}$ and X_{N+1} from r.

We can rewrite $R = \tilde{R} + b \cdot X_{N+1}$. Two cases are possible:

1. r lies in the domain of ϕ^{-1}, which means that $X_{N+1} = 0$, and $\mathbf{X} = \phi^{-1}(r)$.
2. $r - b$ lies in the domain of ϕ^{-1}, which means $X_{N+1} = 1$, and $\mathbf{X} = \phi^{-1}(r - b)$.

The probability of the event that both of the above conditions hold is zero, because there are only finitely many elements in the domain of the function ϕ^{-1} (as its range is $2^{\mathbf{X}}$), while b is drawn randomly and independently according to a univariate normal distribution. Thus, with probability one, exactly one of the above conditions is satisfied, which allows us to uniquely recover X_{N+1} and $\mathbf{X}$ from r.

A.2 Proof of Theorem 1

The condition $\mathbf{R} = \phi(\mathbf{X})$ is equivalent to $\mathbf{Z} = \mathbf{z}, \bar{\mathbf{R}} = \bar{\phi}(\bar{\mathbf{Z}})$, and the latter is obtained by solving the system of equations $\mathbf{R} = \phi(\mathbf{X})$ (more in Appendix B). Therefore, $\mathbb{E}_{P^*}[Y \mid \mathbf{r}] = P^*(Y = 1 \mid \mathbf{r}) = P^*(Y = 1 \mid \mathbf{z}, \bar{\mathbf{r}})$.

For convenience, let $\mathbf{V} := \mathbf{X} \cup \{Y\}$. A c-factor is defined as follows for every $\mathbf{C} \subseteq \mathbf{V}$:

$$Q^*[\mathbf{C}](\mathbf{c}, \mathbf{pa_C}) := P^*(\mathbf{c} \mid do(\mathbf{pa_C} \setminus \mathbf{C})), \tag{140}$$

where $\mathbf{pa_C} := \bigcup_{C \in \mathbf{C}} \mathbf{pa}_C$. By Theorem 2 from Lee et al. [30],

$$P^*(y \mid \mathbf{z}, \bar{\mathbf{r}}) = \frac{\sum_{\mathbf{a} \setminus (\{y\} \cup \mathbf{w}_Y)} Q^*[\mathbf{A}]}{\sum_{\mathbf{a} \setminus \mathbf{w}_Y} Q^*[\mathbf{A}]}, \tag{141}$$

where, $(\mathcal{G}^*_{\text{aug}})_{\underline{\mathbf{Z} \cup \bar{\mathbf{R}}}}$ is obtained by cutting the outgoing arrows of $\mathbf{Z} \cup \bar{\mathbf{R}}$ in $\mathcal{G}^*_{\text{aug}} \in \mathcal{G}^{\mathbf{\Delta}}_{\text{aug}}$, $\mathbf{W}_Y$ is the set of variable $V \in \mathbf{Z} \cup \bar{\mathbf{R}}$ that are connected to Y by any path in $(\mathcal{G}^*_{\text{aug}})_{\underline{\mathbf{Z} \cup \bar{\mathbf{R}}}}$, and $\mathbf{A}$ is the set of variables $V \in \mathbf{V}$ for which there exists a directed path from V to $Y \cup \mathbf{W}_Y$ in $(\mathcal{G}^*_{\text{aug}})_{\underline{\mathbf{Z} \cup \bar{\mathbf{R}}}}$. The gTR algorithm decomposes $Q[\mathbf{A}]$ according to

$$Q^*[\mathbf{A}^1] \cdot Q^*[\mathbf{A}^2] \cdot \cdots \cdot Q^*[\mathbf{A}^K] \cdot Q^*[\bar{\mathbf{R}}] =: Q^*[\mathbf{A}_0] \cdot Q^*[\bar{\mathbf{R}}] \tag{142}$$

Next, it attempts to identify each c-factor from some source domain using the sub-routine IDENTIFY [30]. For the last c-factor $Q^*[\bar{\mathbf{R}}]$, the algorithm can transport it from any source distribution, i.e., $Q^*[\bar{\mathbf{R}}] = Q^i[\bar{\mathbf{R}}]$ for every $1 \leq i \leq T$. In P notation,

$$\begin{aligned} Q^*[\bar{\mathbf{R}}] &= Q^i[\bar{\mathbf{R}}] && && (143)\\ &= P^i(\bar{\mathbf{r}} \mid \bar{\mathbf{z}}) && \text{(c-factor rules)} && (144)\\ &= 1_{\{\bar{\mathbf{r}} = \bar{\phi}(\bar{\mathbf{z}})\}} && \text{(computable from } P^i_{\text{aug}}\text{)} && (145) \end{aligned}$$

Suppose the gTR algorithm returns an expression for the c-factor $Q^*[\mathbf{A}]$. We can apply Lemma 4 by Lee et al. [30] in a topological order to deduce $P^*(y \mid \mathbf{z}, \bar{\mathbf{r}})$ is transportable if and only if $\sum_{\mathbf{a} \setminus (\{y\} \cup \mathbf{w}_Y)} Q^*[\mathbf{A}]$ is transportable. In case $Q^*[\mathbf{A}]$ is transported by gTR, the algorithm returns the expression in Equation 141 which is a valid transportation formula for $P^*(y \mid \mathbf{z}, \bar{\mathbf{r}})$ and is equal to the target query $P^*(y \mid \mathbf{r})$.

A.3 Proof of Proposition 2

For a fixed $e \in [0, 1]$ define,

$$\mathcal{D}^e = \{\mathbf{r} \in \text{supp}(\mathbf{R}) : \mathbb{E}_{P^i}[Y \mid \mathbf{r}] = e\}. \tag{146}$$

We can derive,

$$\mathbb{E}_{P^i}[Y \mid \psi(\mathbf{x})] = \mathbb{E}_{P^i}[Y \mid \mathbf{R} \in \mathcal{D}^{\psi(\mathbf{x})}] \tag{147}$$

$$= \int_{\mathcal{D}^{\psi(\mathbf{x})}} \mathbb{E}_{P^i}[Y \mid \mathbf{r}] \cdot P^i(\mathbf{r} \mid \mathbf{R} \in \mathcal{D}^{\psi(\mathbf{x})}) \cdot d\mathbf{r} \tag{148}$$

$$= \int_{\mathcal{D}^{\psi(\mathbf{x})}} \psi(\mathbf{x}) \cdot P^i(\mathbf{r} \mid \mathbf{R} \in \mathcal{D}^{\psi(\mathbf{x})}) \cdot d\mathbf{r} \tag{149}$$

$$= \psi(\mathbf{x}), \tag{150}$$

which proves multi-calibration of ψ.

On the mutual information, we have the following derivation:

$$I_{P^i}(Y; \phi(\mathbf{X})) = \sum_{y \in \{0,1\}} \int_{\text{supp}(\mathbf{R})} P^i(y, \mathbf{r}) \cdot \log \big(\frac{P^i(y, \mathbf{r})}{P^i(\mathbf{r}) \cdot P^i(y)} \big) \cdot d\mathbf{r} \tag{151}$$

$$= \sum_{y \in \{0,1\}} \int_0^1 \int_{\mathcal{D}^e} P^i(y, \mathbf{r}) \cdot \log \big(\frac{P^i(y, \mathbf{r})}{P^i(\mathbf{r}) \cdot P^i(y)} \big) \cdot d\mathbf{r} \cdot de \tag{152}$$

$$= \sum_{y \in \{0,1\}} \int_0^1 \int_{\mathcal{D}^e} \overbrace{P^i(y, \mid \mathbf{r})}^{P^i(y|e)} \cdot P^i(\mathbf{r}) \cdot \big(\log \overbrace{P^i(y \mid \mathbf{r})}^{P^i(y|e)} - \log P^i(y) \big) \cdot d\mathbf{r} \cdot de \tag{153}$$

$$= \sum_{y \in \{0,1\}} \int_0^1 P^i(y \mid e) \cdot \big(\log P^i(y \mid e) - \log P^i(y) \big) \int_{\mathcal{D}^e} \cdot P^i(\mathbf{r}) \cdot d\mathbf{r} \cdot de \tag{154}$$

$$= \sum_{y \in \{0,1\}} \int_0^1 P^i(y \mid e) \cdot P^i(e) \cdot \big(\log P^i(y \mid e) - \log P^i(y) \big) \cdot de \tag{155}$$

$$= I_{P^i}(Y; \psi(\mathbf{X})) \tag{156}$$

A.4 Proof of Lemma 1

In case $\mathcal{M}^* = \mathcal{M}^t$ ($t \in \{1, 2, \ldots, T\}$), we would have $\Delta_{*,k} = \Delta_{t,k}$ for all $k \in \{1, 2, \ldots, T\}$, and we can derive,

$$\bigcup_{k=1}^{T} \Delta_{*,k} = \bigcup_{k=1}^{T} \Delta_{t,k} \subseteq \bigcup_{i,j=1}^{T} \Delta_{i,j}, \tag{157}$$

which implies SoM assumption, i.e., for every $V \in \mathbf{X} \cup \{Y\}$,

$$V \notin \bigcup_{i,j=1}^{T} \Delta_{i,j} \implies \bigcup_{k=1}^{T} \Delta_{*,k}. \tag{158}$$

Therefore, the case $\mathcal{M}^* = \mathcal{M}^t$ is a compatible target domain under SoM assumption.

A.5 Proof of Theorem 2

If: Suppose $\mathbf{R} = \phi(\mathbf{X})$ satisfies source invariance property, i.e.,

$$s'(\mathbf{r}) := \mathbb{E}_{P^1}[Y \mid \mathbf{r}] = \mathbb{E}_{P^2}[Y \mid \mathbf{r}] = \cdots = \mathbb{E}_{P^T}[Y \mid \mathbf{r}]. \tag{159}$$

Assume the contrary; suppose $s'(\mathbf{r}) \neq l_\phi(\mathbf{r}) = \mathbb{E}_{P^*}[Y \mid \mathbf{r}]$. Therefore,

$$\forall k \in \{1, 2, \ldots, T\} : S_{*k} \not\perp\!\!\!\perp_d Y \mid \mathbf{Z}, \bar{\mathbf{R}} \text{ in } \mathcal{G}^{\Delta_{*k}}. \tag{160}$$

Due to SoM assumption, the causal diagram is shared across the source and target domains. Thus, the augmented selection diagram can be represented by a single graph $\mathcal{G}^{\mathbf{\Delta}}_{\text{aug}}$. Take $k \in \{1, 2, \ldots, T\}$ arbitrarily, and consider a d-connecting path in $\mathcal{G}^{\mathbf{\Delta}}_{\text{aug}}$,

$$Y = V_0, V_1, V_2, \ldots, V_d \leftarrow S_{*k}. \tag{161}$$

Due to SoM assumption, there must exist $i, j \in \{1, 2, \dots, T\}$ such that S_{ij} points to V_d. Now, we can see that same path would be d-connecting Y, S_{ij} in $\mathcal{G}^{\mathbf{\Delta}}_{\text{aug}}$, i.e.,

$$Y \not\perp\!\!\!\perp_d S_{ij} \mid \mathbf{Z}, \bar{\mathbf{R}} \text{ in } \mathcal{G}^{\mathbf{\Delta}}_{\text{aug}}. \tag{162}$$

Due to r-faithfulness, this implies that the invariant property $\text{INV}_{ij}[\phi]$ does not hold, which contradicts the assumption that ϕ satisfies the source invariance property.

Only if: By corollary 2.

B Solving the system of equations

Here, we elaborate more on the definition of determined, constrained, and free variables (Def. 7). Consider the system of equations $\mathbf{R} = \phi(\mathbf{X})$, and let $\mathbf{r} \in \text{supp}(\mathbf{R})$ be the value attained by the representation. Define,

$$\mathcal{T}^{\mathbf{r}} = \{\mathbf{x} \in \text{supp}(\mathbf{X}) \text{ s.t. } \mathbf{r} = \phi(\mathbf{x})\}. \tag{163}$$

A variable $Z \in \mathbf{X}$ is determined by ϕ, if for every value $\mathbf{r} \in \text{supp}(\mathbf{R})$, the set

$$\{z \in \text{supp}(Z) \text{ s.t. } \langle z, \mathbf{x} \setminus \{Z\}\rangle \in \mathcal{T}^{\mathbf{r}}\}, \tag{164}$$

contains only one element. A variable $W \in \mathbf{X}$ is free from ϕ, if for every $\mathbf{r} \in \text{supp}(\mathbf{R})$ and every pair of values for W such as $w_a, w_b \in \text{supp}(W)$,

$$\langle \mathbf{x} \setminus \{W\}, w_a\rangle \in \mathcal{T}^{\mathbf{r}} \implies \langle \mathbf{x} \setminus \{W\}, w_b\rangle \in \mathcal{T}^{\mathbf{r}} \tag{165}$$

A variable $\bar{Z} \in \mathbf{X}$ is constrained by $\mathbf{R}$, if it is neither free from ϕ nor determined by it. For example, in Example 8, X_1, X_3 are determined, and X_2, X_4 are constrained by ϕ.

Fix the determined variables. The determined variables can be computed from the value $\mathbf{R} = \mathbf{r}$ using the mapping $\psi : \text{supp}(\mathbf{R}) \to \text{supp}(\mathbf{Z})$ if the equation $\mathbf{Z} = \psi(\mathbf{R})$ can be algebraically proved given $\mathbf{R} = \phi(\mathbf{X})$. Automated proofs and algebraic derivations are well-studied subjects within theoretical computer science; here we treat the solving procedure as a black-box.

derive constraints. Let $\mathbf{Z} = \det(\phi), \bar{\mathbf{Z}} = \text{cons}(\phi)$. Once we have access to the mapping ψ, we can plug-in the value of $\mathbf{Z} = \psi(\mathbf{R})$ in the expression for ϕ to obtain a system of equations with less unknown variables:

$$\mathbf{R} = \phi(\bar{\mathbf{Z}}, \mathbf{Z} : \psi(\mathbf{r}), \mathbf{X} \setminus (\bar{\mathbf{Z}} \cup \mathbf{Z})). \tag{166}$$

Next, we can massage this expression to rewrite it in a form that does not contain any free variables $\mathbf{X} \setminus (\bar{\mathbf{Z}} \cup \mathbf{Z})$. Without loss of generality, suppose $\mathbf{R} = \phi(\bar{\mathbf{Z}}, \mathbf{Z} : \psi(\mathbf{R}))$ is the expression at hand. Next, we massage the expression to move every term containing $\mathbf{R}$ to the l.h.s., and call the resulting expression $\bar{\mathbf{R}}$. Then, the expression in terms of $\bar{\mathbf{Z}}$ remained on the r.h.s. is denoted as $\bar{\phi}$. In summary,

$$\mathbf{R} = \phi(\bar{\mathbf{Z}}, \psi(\mathbf{R})) \iff \bar{\mathbf{R}} = \bar{\phi}(\bar{\mathbf{Z}}). \tag{167}$$

Notice that the only requirement of the above procedure is to move all the expressions containing $\mathbf{R}$ to the l.h.s., and this does not interfere with entanglement of variables, as it is valid if the final expression contains $\bar{\mathbf{Z}}$ terms as well as $\mathbf{R}$ terms on the l.h.s., while it must contain only $\bar{\mathbf{Z}}$ terms (possibly none) on the r.h.s., i.e., no $\mathbf{R}$ terms shall remain on the r.h.s. of the equality. After this manipulation, the expression on the l.h.s. is renamed as $\bar{\mathbf{R}}$, and the r.h.s. would be an expression in terms of only $\bar{\mathbf{Z}}$, which is denoted by $\bar{\phi}(\bar{\mathbf{Z}})$.

once we fix the value of $\bar{\mathbf{r}} = \bar{\phi}(\bar{\mathbf{Z}})$ and $\mathbf{Z} = \mathbf{z}$, we can obtain $\mathbf{r} = \phi(\bar{\mathbf{Z}}, \mathbf{Z})$ in the following way: As we have access to $\bar{\mathbf{r}}$, we can revert the derivation in equation 167 to obtain $\mathbf{R} = \phi(\bar{\mathbf{Z}}, \psi(\mathbf{R}))$ only dependent on the unknown $\psi(\mathbf{R})$. Next, we can substitute the term $\psi(\mathbf{R})$ with its known value $\mathbf{z}$ that is at hand, and then there remains no unknown variables in the expression for $\mathbf{R}$. Let $\phi^* : \text{supp}(\mathbf{Z}) \times \text{supp}(\bar{\mathbf{R}}) \to \text{supp}(\mathbf{R})$ denote the described mapping that allows us to compute $\mathbf{r}$ from $\bar{\mathbf{r}}, \mathbf{z}$. This mapping and the derivations above allow us to translate back and forth between the equivalent queries $\mathbb{E}[Y \mid \mathbf{R} = \mathbf{r}]$ and $\mathbb{E}[Y \mid \bar{\mathbf{R}} = \bar{\mathbf{r}}, \mathbf{Z} = \mathbf{z}]$ (e.g., in Theorems 1 & 2)

References

[1] John Aldrich. Autonomy. *Oxford Economic Papers*, 41(1):15–34, 1989. ISSN 00307653, 14643812.

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[3] Elias Bareinboim and Judea Pearl. Transportability from multiple environments with limited experiments: Completeness results. *Advances in neural information processing systems*, 27, 2014.

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

[5] Elias Bareinboim, Sanghack Lee, Vasant Honavar, and Judea Pearl. Transportability from multiple environments with limited experiments. *Advances in Neural Information Processing Systems*, 26, 2013.

[6] Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. On pearl's hierarchy and the foundations of causal inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, page 507–556. Association for Computing Machinery, NY, USA, 1st edition, 2022.

[7] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.

[8] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*, volume 28 of *Princeton Series in Applied Mathematics*. Princeton University Press, 2009. ISBN 978-1-4008-3105-0.

[9] Y. Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35:1798–1828, 08 2013.

[10] Rohit Bhattacharya, Razieh Nabi, and Ilya Shpitser. Semiparametric inference for causal effects in graphical models with hidden variables. *J. Mach. Learn. Res.*, 23(1), jan 2022. ISSN 1532-4435.

[11] Yuansi Chen and Peter Bühlmann. Domain adaptation under structural causal models. *The Journal of Machine Learning Research*, 22(1):11856–11935, 2021.

[12] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018. ISSN 1368-4221. doi: 10.1111/ectj.12097. URL https://doi.org/10.1111/ectj.12097.

[13] J. Correa and E. Bareinboim. General transportability of soft interventions: Completeness results. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 10902–10912, Vancouver, Canada, Jun 2020. Curran Associates, Inc.

[14] Juan D Correa and Elias Bareinboim. From statistical transportability to estimating the effect of stochastic interventions. In *IJCAI*, pages 1661–1667, 2019.

[15] Shai Ben David, Tyler Lu, Teresa Luu, and David Pal. Impossibility theorems for domain adaptation. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 129–136, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.

[16] Graciela De Pierris and Michael Friedman. Kant and Hume on Causality. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2018 edition, 2018.

[17] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

[18] Dan Geiger. *Graphoids: A Qualitative Framework for Probabilistic Inference*. PhD thesis, USA, 1990. UMI Order No. GAX90-16109.

[19] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

[20] Steve Hanneke and Samory Kpotufe. On the value of target data in transfer learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[21] Leah Henderson. The problem of induction. 2018.

[22] D Hume. *A Treatise of Human Nature*. Oxford University Press, Oxford, 1739.

[23] Amin Jaber, Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Causal discovery from soft interventions with unknown targets: Characterization and learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9551–9561. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6cd9313ed34ef58bad3fdd504355e72c-Paper.pdf.

[24] Y. Jung, J. Tian, and E. Bareinboim. Estimating identifiable causal effects through double machine learning. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, number R-69, Vancouver, Canada, Feb 2021. AAAI Press.

[25] Y. Jung, J. Tian, and E. Bareinboim. Double machine learning density estimation for local treatment effects with instruments. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[26] Y. Jung, I. Diaz, J. Tian, and E. Bareinboim. Estimating causal effects identifiable from combination of observations and experiments. Technical Report R-97, Causal Artificial Intelligence Lab, Columbia University, May 2023.

[27] Pritish Kamath, Akilesh Tangella, Danica Sutherland, and Nathan Srebro. Does invariant risk minimization capture invariance? In *International Conference on Artificial Intelligence and Statistics*, pages 4069–4077. PMLR, 2021.

[28] Immanuel Kant. *Critique of Pure Reason*. St. Martin's Press (NY), 1781.

[29] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5815–5826. PMLR, 18–24 Jul 2021.

[30] S. Lee, J. Correa, and E. Bareinboim. Generalized transportability: Synthesis of experiments from heterogeneous domains. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, New York, NY, 2020. AAAI Press.

[31] Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[32] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31, 2018.

[33] Chaochao Lu, Yuhuai Wu, Jośe Miguel Hernández-Lobato, and Bernhard Schölkopf. Nonlinear invariant risk minimization: A causal approach. *arXiv preprint arXiv:2102.12353*, 2021.

[34] 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. *Advances in neural information processing systems*, 31, 2018.

[35] Chengzhi Mao, Kevin Xia, James Wang, Hao Wang, Junfeng Yang, Elias Bareinboim, and Carl Vondrick. 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*, pages 7511–7521. IEEE, 2022.

[36] Tom M Mitchell. *Machine learning*, volume 1. McGraw-hill New York, 1997.

[37] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000.

[38] Judea Pearl. *Causality*. Cambridge university press, 2009.

[39] Judea Pearl and Elias Bareinboim. Transportability of causal and statistical relations: A formal approach. In *Twenty-fifth AAAI conference on artificial intelligence*, 2011.

[40] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78 (5):947–1012, 2016.

[41] Niklas Pfister, Evan G Williams, Jonas Peters, Ruedi Aebersold, and Peter Bühlmann. Stabilizing variable selection and regression. *The Annals of Applied Statistics*, 15(3):1220–1246, 2021.

[42] Karl R Popper. Conjectural knowledge: my solution of the problem of induction. *Revue internationale de Philosophie*, pages 167–197, 1971.

[43] KR Popper. The problem of induction, 1953.

[44] 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.

[45] Elan Rosenfeld, Pradeep Kumar Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. In *International Conference on Learning Representations*, 2021.

[46] Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(2): 215–246, 2021.

[47] Bertrand Russell. On induction. *First published as*, pages 19–26, 1912.

[48] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2010.

[49] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[50] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

[51] Adarsh Subbaswamy and Suchi Saria. From development to deployment: dataset shift, causality, and shift-stable models in health ai. *Biostatistics*, 21(2):345–352, 2020.

[52] Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3118–3127. PMLR, 2019.

[53] Masashi Sugiyama and Klaus-Robert Müller. Input-dependent estimation of generalization error under covariate shift. 23(4):249–279, 2005.

[54] Mark J. van der Laan and Susan Gruber. Targeted minimum loss based estimation of causal effects of multiple time point interventions. *The International Journal of Biostatistics*, 8(1), 2012. doi: doi:10.1515/1557-4679.1370. URL https://doi.org/10.1515/1557-4679.1370.

[55] V. Vapnik. Principles of risk minimization for learning theory. In J. Moody, S. Hanson, and R.P. Lippmann, editors, *Advances in Neural Information Processing Systems*, volume 4. Morgan-Kaufmann, 1991.

[56] Vladimir Vapnik. Statistical learning theory wiley. *New York*, 1(624):2, 1998.

[57] Yoav Wald, Amir Feder, Daniel Greenfeld, and Uri Shalit. On calibration and out-of-domain generalization. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[58] Yixin Wang and Michael I Jordan. Desiderata for representation learning: A causal perspective. *arXiv preprint arXiv:2109.03795*, 2021.

[59] Eric Watkins et al. *Kant and the Metaphysics of Causality*. Cambridge University Press, 2005.

[60] Renzhe Xu, Xingxuan Zhang, Zheyan Shen, Tong Zhang, and Peng Cui. A theoretical analysis on independence-driven importance weighting for covariate-shift generalization. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:249848308.

[61] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 325–333. PMLR, 2013.

[62] Kun Zhang, Mingming Gong, and Bernhard Schoelkopf. Multi-source domain adaptation: A causal view. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), Feb. 2015.