TECHNICAL REPORT
R-99
May, 2023

Transportable Representations for Out-of-distribution Generalization

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

Abstract

Generalizing across settings and changing conditions is one of the fundamental problem of machine learning and AI. One common assumption pervasive in the literature is that the testing and training data come from the same distribution, which is often violated in practice. The anchors that allow generalizations to take place are causal, and provenient from the stability and modularity of the mechanisms underlying the system under investigation. Building on the theory of causal transportability (Bareinboim & Pearl), we define in this paper the notion of "transportable representations," and show that the classifiers defined based on these representations exhibit desirable out-of-distribution properties. Specifically, we develop an algorithm to decide whether a specific representation satisfies the transportability property. We then study the problems of domain generalization (DG) and unsupervised domain adaptation (UDA) regarding these transportability properties. We show that the risk of classifiers defined based on transportable representations can be computed exactly in the UDA case, and can be bounded in the DG case, considering that graphical assumptions about the underlying system are provided. Finally, we study a data-driven approach for transportability through a collection of invariance tests in the source domain. We prove a graphical-invariance duality theorem, which delineates under which assumptions the data-driven approach and the graphical approaches coincide.

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 [16]. This challenge is called the *problem of induction*, and have puzzled generations of philosophers and mathematicians, from Kant to Popper, Goodman to Russell [29, 30, 34, 45].

The generalization problem plays a fundamental role in artificial intelligence and machine learning as well [24, 35], where it appears in different forms. For instance, one of the most well-studied tasks in the field is known as classification, where one tries to 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 [24, p. 44]: "a fundamental property of inductive inference: a learner that makes no a priori assumptions regarding the identity of the target concept has no rational basis for classifying any unseen instances.", refining Hume's observation. The question becomes how to link the data collected from the distribution to the distribution itself. One of the fundamental results in the field is known as *empirical risk minimization*, due to Vapnik, tied the risk between hypothetical and empirical distributions under some very general conditions [41, 42].

Query	Data		Rep ϕ	Assumptions			Method	Result
	Src.	Target		$\mathcal{G}^{\mathbf{\Delta}}$	A1	A2		
$P^*(y \mid \mathbf{z})$	$\mathbb{P}$			✓			do-calc.	g-TR [19, 11]
$P^*(y \mid \phi(\mathbf{x}))$	$\mathbb{P}$		✓	✓			g-TR	Alg. 1, Thm 1
$R_{P^*}(\tilde{h} \circ \phi)$	$\mathbb{P}$	$[P^*(\mathbf{x})]$	✓	✓			Alg 1	Thms 2, 3
$P^*(y \mid \phi(\mathbf{x}))$	$\mathbb{P}$		✓		✓	✓	$I[\phi; \mathbb{P}]$	Thm. 4
$R_{P^*}(\tilde{h} \circ \phi)$	$\mathbb{P}$	$[P^*(\mathbf{x})]$	✓		✓	✓	$I[\phi; \mathbb{P}]$	Thm 4

Table 1: Tasks discussed in this paper: The query is either a probabilistic term (blue), or the risk expression (green). Data is labeled in the source domains, but might be missing or unlabeled in the target. The tasks we study involve a representation ϕ. $\mathcal{G}^{\mathbf{\Delta}}$ denotes the selection diagram. Assumption $A1$ is a variation of faithfulness [37], and Assumption A2 asserts certain properties to the selection diagrams $\mathcal{G}^{\mathbf{\Delta}}$. The method column indicates the subroutines used in each solution, e.g., $I[\phi; \mathbb{P}]$ denotes the invariance property (Def. 8). The last column is a reference to the corresponding results.

Despite the power of these ensuing results, we note that, in practice, the domains where the data is collected (called sources) are related to, but not necessarily the same as the one where the predictions are intended (target), violating a key assumption underlying many of the prior results. In fact, if the target domain is arbitrary, or drastically different from the source domains, no learning could take place [12, 6]. However, the fact that we generalize and adapt relatively well to a new domain suggest that certain domains share common characteristics and that, owing to these commonalities, statistical claims can be generalized even to domains where no or partial data is available [26, 37, 4]. How could one described the shared features across environments that allow this inferential leap? The anchors of knowledge that allow generalization to take place are eminently causal, following from the stability and invariance of the mechanisms shared across settings [1]. [1] The systematic analysis of these mechanisms and the conditions under which generalizations could be formally justified has been studied in the literature under the rubric of *transportability theory* [3–5, 27, 10, 11, 19].

In modern machine learning literature, the challenge of predicting in an unseen target domain is acknowledged and broadly referred to as the out-of-distribution (OOD) generalization. The theoretical proposals in this area rely on assumptions to define the target domains compatible with the source data, e.g., the covariate shift assumption [40, 39, 38], or use of distance measures to relate the source and target distributions [7, 15]. Even under restrictive assumptions tying the source and target distributions, adapting to the target domain might still be impossible [12]. 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, and 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 and target data as proxies to the causal invariances across the source and target domains [22, 31, 2, 33, 43, 9]. These methods are contingent on assumptions such as linearity, additivity, markovianity, yet there exists subtleties that limit the effectiveness and practicality of these methods [32]. Another important ingredient present in modern machine learning methods is the use of representations. Those methods extract useful information to feed into the learning algorithm, which is particularly useful in high-dimensional and unstructured domains [8]. 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, 14, 21, 20, 47, 46]. Also, causal features has been used in representations to help predictions across domain, while filtering out the spurious correlations that might be unstable across domains [44, 36, 23, 18].

By and large, we note that solving an OOD generalization problem can be seen as a two-step process – step 1 (evaluation). given a classifier, compute/bound its worst-case risk; step 2 (search). find a classifier that minimizes the quantity obtained by an evaluation method. In this paper, we study the evaluation step through transportability lenses in the context of two canonical tasks within OOD generalization, namely, **domain generalization (DG)** and **unsupervised domain adaptation (UDA)**. In the former, only labeled data from source domains is available, while in the latter, unlabeled data from the target domain is available as well as the labeled data from the source domains. We also analyze in these settings the fundamental interplay between causal knowledge and the complexity of a representation. For instance, we refute through our analysis the belief that causal features are

[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 [17]; for further discussion on this point, refer to [13].

always desirable while spurious should be discard. Table 1 provides a summary of the setup studied in this paper, and more specifically, our contributions are as follows:

- (Section 2) We introduce the notion of transportable representations (Def. 7), motivate their usefulness for risk evaluation, and develop an algorithm (Alg. 1 & Thm. 1) to decide whether a representation is transportable from the source distributions and structural assumptions.
- (Section 3) We show the worst-case risk of classifiers defined based on transportable representations can be computed for the UDA task (Thm 2) and can be bounded for the DG task (Thm 3).
- (Section 4) We prove that invariance of certain distributions across the source domains is a sound and complete data-driven criterion to find transportable representations (Thm. 4), and use their properties for risk evaluation. In particular, this provides a dual view on the graphical-invariance dichotomy, which highlights under what set of assumptions they coincide.

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

We use semantics of Structural Causal Models [26], 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; $\mathbf{V}$ 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 $\mathcal{G}$. A SCM $\mathcal{M}$ induces a probability distribution $P^M(\mathbf{v})$ over the set of observed variables $\mathbf{V}$ such that $P^{\mathcal{M}}(\mathbf{v}) = \int_{\text{supp}(\mathbf{U})} \prod_{V \in \mathbf{V}} P^M(v \mid \mathbf{pa}_V, \mathbf{u}_V) \cdot P(\mathbf{u}) \cdot d\mathbf{u}$, where each term $P(v \mid \mathbf{pa}_V, \mathbf{u}_V)$ corresponds to the function $f_V \in \mathcal{F}$ in the underlying structural causal model $\mathcal{M}$. Throughout this paper, we assume the observational distributions entailed by the SCMs we study satisfy positivity, that is, $P^{\mathcal{M}}(\mathbf{v}) > 0$, for every $\mathbf{v}$. We will also operate non-parametrically, i.e., making no assumption about the particular functional form or the distribution of the unobserved variables. In this case, the only assumption is that the arguments of the functions are known as encoded through the causal diagram $\mathcal{G}$.

2 OOD Generalizations through Transportable Representations

We study a system of variables $\mathbf{X} \cup \{Y\}$, where Y is a binary label. SCMs $\mathcal{M}^1, \mathcal{M}^2, \ldots, \mathcal{M}^T$ defined over $\mathbf{X} \cup \{Y\}$ denote the source domains, and entail the distributions $\mathbb{P} = \{P^1, P^2, \ldots, P^T\}$, while they induce the causal diagrams $\mathcal{G}^1, \mathcal{G}^2, \ldots, \mathcal{G}^T$. There exists an unknown SCM $\mathcal{M}^*$ representing the target domain, which entails the distribution P^*, while it induces the causal diagram $\mathcal{G}^*$. We adapt the following notion introduced in [19] to describe mismatch of mechanisms between two SCMs.

Definition 2 (Domain discrepancy) *For every pair of SCMs M^a, M^b ($a, b \in \{*, 1, 2, \ldots, T\}$) defined over $\mathbf{X} \cup \{Y\}$, the domain discrepancy set $\Delta_{ab} \subseteq \mathbf{V}$ is defined such that for every $V \in \Delta_{ab}$ there might exist a discrepancy $f_V^{M^a} \neq f_V^{M^b}$ or $P^{M^a}(\mathbf{u}_V) \neq P^{M^b}(\mathbf{u}_V)$.* □

In other words, $V \notin \Delta_{ab}$ is equivalent assuming the same mechanisms for V across M^a, M^b, i.e., $f_V^{M^a} = f_V^{M^b}$ and $P^{M^a}(\mathbf{u}_V) = P^{M^b}(\mathbf{u}_V)$. We introduce next a version of selection diagrams [19] to graphically represent the system that includes multiple SCMs relative to the collection of source and target 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 $S_{ij} \to V$ for every $V \in \Delta_{ij}$. The collection $\mathcal{G}^{\boldsymbol{\Delta}} = \{G^*\} \cup \{\mathcal{G}^{\Delta_{ij}}\}_{i,j=1}^{T}$ encodes the graphical assumptions. If the causal diagram is shared across the domains, we can use a single graph to depict $\mathcal{G}^{\boldsymbol{\Delta}}$.* □

In words, a selection diagram is a parsimonious graphical representation of the commonalities and disparities across domains, which can be seen as grounding Kant's observation alluded to earlier.

Definition 4 (Transportability) *For subsets of variables $\mathbf{A}, \mathbf{B} \subset \mathbf{X} \cup \{Y\}$ in the SCM, the query $P^*(\mathbf{a} \mid \mathbf{b})$ is transportable if for every pair of SCMs $\mathcal{M}_1^*, \mathcal{M}_2^*$ compatible with the selection diagrams $\mathcal{G}^{\boldsymbol{\Delta}}$, and the distributions $\mathbb{P}$ over $\mathbf{X} \cup \{Y\}$, $P^{\mathcal{M}_1^*}(\mathbf{a} \mid \mathbf{b}) = P^{\mathcal{M}_2^*}(\mathbf{a} \mid \mathbf{b})$.* □

In both DG and UDA tasks, the joint distribution $P^*(\mathbf{x}, y)$ is unknown, yet we might be able to infer certain aspects of it (e.g., the conditional distributions, the risk of a classifier) from the source distributions $\mathbb{P}$ and qualitative assumptions encoded by the selection diagrams $\mathcal{G}^{\boldsymbol{\Delta}}$. The notion of transportability describes such a property.

The input for the DG task comprises the labeled data drawn from each $P^i \in \mathbb{P}$. In the UDA task, unlabeled data drawn from $P^*(\mathbf{x})$ is also available. Next, we formally define classifiers which use a representation of the input.

Definition 5 (Representations for classification) *The variable $\mathbf{R} = \phi(\mathbf{X})$ is called a representation for every mapping $\phi : \text{supp}(\mathbf{X}) \to \text{supp}(\mathbf{R})$. Furthermore, a representation is said to satisfy the coverage property w.r.t. the distribution $P(\mathbf{x})$ if $P(\mathbf{X} \in \{\mathbf{x} : \phi(\mathbf{x}) = \mathbf{r}\}) > 0$ for every $\mathbf{r} \in \text{supp}(\mathbf{R})$. A mapping $h : \text{supp}(\mathbf{X}) \to \{0, 1\}$ is said to be a classifier defined based on the representation $\mathbf{R} = \phi(\mathbf{X})$ if it can be expressed as composition with ϕ, i.e., $h = \tilde{h} \circ \phi$.* □

Throughout this work, we consider representations that satisfy the coverage of property w.r.t. all $P^i \in \mathbb{P}$. Our performance measure for the classifier $\hat{h}$ is called *risk*, a.k.a., classification error, defined as, $R_{P^*}(\hat{h}) := P^*(Y \neq h(\mathbf{X}))$.

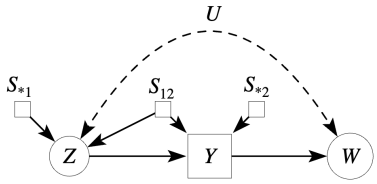

Figure 1: Selection diagram corresponding to Examples 1.

Example 1 (High blood pressure (HBP)) Let Y be a binary variable indicating whether a patient has HBP. For each patient, a set of features $\mathbf{X} = \{Z, W\}$ is measured, which denotes the level of exercise and anxiety, respectively. The unobserved confounders U is the patient's wealth. In this population, wealth directly affects the patients' exercise and anxiety levels. Data is drawn from P^1, P^2 entailed by domains $\mathcal{M}^1, \mathcal{M}^2$, respectively. The patients from $\mathcal{M}^1$ are genetically prone to HBP, which leads the government to run TV ads to promote exercising.

We are asked to classify whether patients in another domain $\mathcal{M}^*$ are at risk of HBP based on the same features $\mathbf{X}$. The relationships across domains is summarized through the selection diagrams $\mathcal{G}^{\boldsymbol{\Delta}}$ shown in Figure 1. In the domain $\mathcal{M}^*$, patients are genetically prone to HBP, similar to $\mathcal{M}^1$, thus, the mechanisms deciding blood pressure (Y) in $\mathcal{M}^*$ is the same as $\mathcal{M}^1$, while differing from $\mathcal{M}^2$. However, in $\mathcal{M}^*$, the government is not running the exercising TV ads, and the mechanism determining exercise is the same as in $\mathcal{M}^2$, while differing from $\mathcal{M}^1$. Further, the mechanism determining anxiety (W) is invariant across sources and target domains. In formal notation, all these invariances can be written as $\Delta_{*1} = \{Z\}$ and $\Delta_{*2} = \{Y\}$, and $\Delta_{12} = \{Z, Y\}$.

As a representation of Z, W, consider a mind & body wellness R that is decreasing in anxiety (W) and increasing in exercise (Z), defined as

$$R = \phi(Z, W) := Z - W. \tag{1}$$

One can construct a classifier based on the value of this representation, namely,

$$\hat{h}(z, w) := \mathbb{1}_{\{\phi(z,w) \leq c\}} = \mathbb{1}_{\{r \leq c\}} = \mathbb{1}_{\{z - w \leq c\}}. \tag{2}$$

In words, $\hat{h}$ suggests that the person is in high risk if their wellness index R is below threshold c. To verify whether R satisfies the coverage property, one needs to evaluate if $P^1(Z - W = r) > 0$ for every $-1 \leq r \leq 1$, which can be done using the data from the source domains. □

We next introduce a criterion useful to judge certain invariances about the underlying mechanisms that will imply probabilistic invariances in the distribution.

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

$$\mathbf{A} \perp\!\!\!\perp_d S_{ij} \mid \mathbf{Z} \text{ in } \mathcal{G}^{\Delta_{ij}} \implies P^i(\mathbf{a} \mid \mathbf{z}) = P^j(\mathbf{a} \mid \mathbf{z}). \tag{3}$$

Note that S-admissibility connects the assumptions encoded in the graphical model about the underlying mechanisms, as formalized in Def. 3, and the mechanisms represented by the underlying and unobserved generating SCMs, to elicit invariances at the probabilistic level (r.h.s. of Eq. 3). Next, we elaborate on whether (and how) the risk of a classifier can be transported (i.e., uniquely computed) given the source data through the S-admissibility criterion.

Example 2 (Risk evaluation through joint transportability) Considering the classifier $\hat{h}(z, w)$ of Ex. 1, we attempt to transport the joint distribution of Y, Z, W in the target domain;

$$\begin{aligned} P^*(z, y, w) &= P^*(z) \cdot P^*(y \mid z) \cdot P^*(w \mid y, z) && \text{(factorization)} && (4)\\ &= P^2(z) \cdot P^1(y \mid z) \cdot P^2(w \mid y, z) && (S\text{-admissibility}) && (5) \end{aligned}$$

Specifically, Eq. 5 follows since Z is (marginally) S-admissible in $\mathcal{M}^2, \mathcal{M}^*$, Y is S-admissible conditional on Z in $\mathcal{M}^1, \mathcal{M}^*$, and W is S-admissible conditioned on $\{Y, Z\}$ w.r.t. $\mathcal{M}^2, \mathcal{M}^*$.

Considering the representation, $R = Z - W$ implies $P^*(r \mid y, z, w) = \mathbb{1}_{\{z-w=r\}}$, so,

$$\begin{aligned} P^*(z, y, w, r) &= P^*(z, y, w) \cdot P^*(r \mid z, y, w) && \text{(factorization)} && (6)\\ &= P^2(z) \cdot P^1(y \mid z) \cdot P^2(w \mid y, z) \cdot \mathbb{1}_{\{z-w=r\}} && \text{Eq. 5} && (7) \end{aligned}$$

Having this joint distribution allows us to derive the following expression for $\hat{h}$'s risk,

$$R_{P^*}(\hat{h}) = P^*(Y \neq \hat{h}(Z, W)) = \int_{-1}^{1} P^*(Y \neq \mathbb{1}_{\{R \leq c\}}) \cdot dr, \tag{8}$$

where the joint distribution $P^*(y, r)$ is computed in the target via,

$$\int P^*(y, z, w, r) \cdot dz \cdot dw \overset{\text{Eq. 7}}{=} \int P^2(z) \cdot P^1(y \mid z) \cdot P^2(w \mid y, z) \cdot \mathbb{1}_{\{z-w=r\}} \cdot dz \cdot dw \tag{9}$$

The first step of the procedure discussed in Sec 1 (Evaluation) can be executed, i.e., the risk (as Eq. 8) can be evaluated via the source data drawn from $P^1(z, w, y), P^2(z, w, y)$ using various estimates. □

A few observations follow from this example. First, note that evaluating the risk from the source distributions allows us to compare the performance of the classifier $\hat{h}$ with other candidate classifiers, and execute the second step (search) to find optimal classification in both DG and UDA tasks. Second, note that the strategy undertaken in the derivation leveraged the fact that the joint distribution over all variables and the representation could be factorized and then transported through the S-admissibility criterion. Third, still, this computation leads to a more general decision problem that asks whether certain distributions can be computed from the available data considering a given representation.

Definition 7 (r-Transportability & transportable representations) *Let $\mathbf{R} = \phi(\mathbf{X})$ be a representation. The query $P^*(y \mid \mathbf{r})$ is r-transportable given (1) the set of distributions $\mathbb{P}$, (2) the selection diagrams $\mathcal{G}^{\mathbf{\Delta}}$, and (3) the arithmetic expression ϕ, if for every two SCMs $\mathcal{M}^*_a, \mathcal{M}^*_b$ compatible with $\mathbb{P}$ and $\mathcal{G}^{\mathbf{\Delta}}$, $P^{\mathcal{M}^*_a}(y \mid \mathbf{r}) = P^{\mathcal{M}^*_b}(y \mid \mathbf{r})$. If so, ϕ will be called a **transportable representation**.*

The next example shows that the strategy used in Ex. 2 is not always applicable for deciding r-transportability, but it's neither necessary.

Example 3 (Alternative for UDA & role of determinism) Consider the selection diagram $\mathcal{G}^{\mathbf{\Delta}}$ in Figure 2, over the variables Y and $\mathbf{X} = \{X_1, X_2, X_3, X_4\}$ with $\text{supp}(X_i) = (0, 1)$. There exists

Algorithm 1 rTR: r-transport $P^*(y \mid \mathbf{r})$ from $\mathbb{P}$, $\mathcal{G}^{\mathbf{\Delta}}$, ϕ.

1: $\langle \mathbf{Z} = \psi(\mathbf{R}), \bar{\mathbf{R}} = \bar{\phi}(\bar{\mathbf{Z}}) \rangle \leftarrow \text{solve}(\mathbf{R} = \phi(\mathbf{X}))$ (determined $\mathbf{Z}$ & constrained $\bar{\mathbf{Z}}$)
2: $\mathcal{G}^{\mathbf{\Delta}}_{\text{aux}}$: Add to every graph in $\mathcal{G}^{\mathbf{\Delta}}$ the variable $\bar{\mathbf{R}}$ & arrows from $\bar{\mathbf{Z}}$ to $\bar{\mathbf{R}}$
3: $\mathbb{P}_{\text{aux}} : \{P^i_{\text{aux}}(\mathbf{x}, y, \bar{\mathbf{r}}) := P^i(\mathbf{x}, y) \cdot \mathbb{1}_{\{\bar{\mathbf{r}} = \bar{\phi}(\bar{\mathbf{z}})\}},$ for every $P^i \in \mathbb{P}\}$
4: **return** $\text{gTR}\big(\text{query: } P^*(y \mid \mathbf{z}, \bar{\mathbf{r}}); \mathcal{G}^{\mathbf{\Delta}}_{\text{aux}}, \mathbb{P}_{\text{aux}}\big)$ [Lee et al. [19]]

only one source domain $\mathcal{M}^1$ and our goal is to solve the UDA task, where only $P^*(\mathbf{x})$ is available. Further, consider a representation $R = \phi(\mathbf{X}) := -\log(X_1)$, and a classifier $\hat{h} := \tilde{h} \circ \phi$ defined based on it. The goal is to compute the risk $R_{P^*}(\hat{h})$. In this case, unlike Example 2, the joint distribution of the label and the representation cannot be transported because the path $S_{*1} \to X_1 \to Y$ is open, and in general $P^*(\mathbf{x}, y) \neq P^1(\mathbf{x}, y)$.

Note that $P^*(\mathbf{x})$ is also available, which means that ϕ can be applied to the unlabeled data, and the distribution $P^*(\phi(\mathbf{x})) = P^*(r)$ can be computed. Once we compute the query $P^*(y \mid r)$ is evaluated instead of the full joint, the risk can be computed by $R_{P^*}(\hat{h}) = \int_0^\infty P^*(Y \neq \tilde{h}(r) \mid r) \cdot P^*(r) \cdot dr (\ddagger)$.

The mapping $R = -\log(X_1)$ is a deterministic relation between X_1, R, which can be expressed as, $R = r \iff X_1 = \exp(-r)$ $(\star)$, alloowing us to rewrite the conditional distribution as,

$$P^*(y \mid R = r) \overset{(\star)}{=} P^*(y \mid X = \exp(-r)) \overset{\dagger}{=} P^1(y \mid X_1 = \exp(-r)) \overset{(\star)}{=} P^1(y \mid R = r), \quad (10)$$

where $(\dagger)$ is due to the S-admissibility of Y given X_1 w.r.t. domains $\mathcal{M}^1, \mathcal{M}^*$. The risk of $\hat{h}$ is computable by plugging Eq. 10 into Eq. $(\ddagger)$ using data from both $P^1(\mathbf{x}, y)$ and $P^*(\mathbf{x})$. □

One may surmise that the more fine-grained strategy pursued in Example 3 to evaluate transportability may always lead to a positive solution. The next example suggests that not all representations satisfy the requirements of r-transportability. Also, this impossibility may be the case even when the representation is more refined by including additional feaures. In fact, the property of transportability is non-monotonic, i.e., more refined representations are not necessarily uniquely transportable; this point is elaborated in the next two examples.

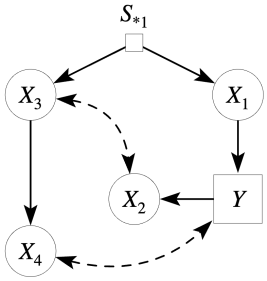

Figure 2: selection diagram of Ex. 3, 4, 5

Example 4 (Lack of transportability; sometimes less is more) In the context of Example 3, consider the representation $\langle R_1, R_2 \rangle := \langle -\log(X_1), 10 \cdot X_4 \rangle$, and a classifier $\hat{h} := \tilde{h} \circ \phi$ defined based on it. Notice, the entry R_1 of the representation is equal to R of Example 3, however, the risk of $\hat{h}$ is not computable using a similar strategy, first, $\langle R_1, R_2 \rangle = \langle r_1, r_2 \rangle \iff \langle X_1, X_4 \rangle = \langle \exp(-r_1), \frac{r_2}{10} \rangle$, which can be written as,

$$P^*(y \mid R_1 = r_1, R_2 = r_2) = P^*(y \mid X_1 = \exp(-r_1), X_4 = \frac{r_2}{10}). \quad (11)$$

However, Y is not S-admissible given X_1, X_4 w.r.t. $\mathcal{M}^1, \mathcal{M}^*$, which precludes transportability. □

While S-admissibility can be used to decide whether a certain quantity is transportable, deciding whether an arbitrary representation is transportable requires a more systematic method towards this goal. We introduce a procedure called rTR (Algorithm 1) to solve this task and show the following.

Theorem 1 (Graphical r-Transportability) *Consider a representation $\mathbf{R} = \phi(\mathbf{X})$. Algorithm 1 is sound for r-transportability of the query $P^*(y \mid \mathbf{r})$ given the distributions $\mathbb{P}$, the selection diagrams $\mathcal{G}^{\mathbf{\Delta}}$, and the arithmetic expression ϕ.* □

All proofs are provided in Appendix A.

Algorithm rTR uses the arithmetic expression for ϕ to solve a system of equation and decides the variables that are determined (e.g., X_1, X_3 in Example 5) or constrained (e.g., X_4 in Example 5) by the condition $\mathbf{R} = \mathbf{r}$. Next, it reduces the r-transportability task into an equivalent transportability task, and solves it by using the gTR algorithm (Lee et al. [19]). Detailed explanation of the Algorithm 1 is provided in Appendix B. In the next section, we build on this r-transportability machinery for evaluating the risk in both UDA and DG tasks.

3 Risk Evaluation for Transportable Representations

In this section, we study the OOD risk of classifiers defined based on transportable representations. In particular, we elaborate how r-transportation yields a method to evaluate the risk for UDA and DG tasks. We first consider an example.

Example 5 (Complex representation) In the context of Examples 3 & 4, consider the representation

$$R_1 = -\log(X_1) + 2 \cdot \sqrt{X_3} + 3 \cdot \lfloor 10 \cdot X_4 \rfloor \tag{12}$$

$$R_2 = -3\log(X_1) + 1 \cdot \sqrt{X_3} + 2 \cdot \lfloor 10 \cdot X_4 \rfloor \tag{13}$$

$$R_3 = -2\log(X_1) + 3 \cdot \sqrt{X_3} + \lfloor 10 \cdot X_4 \rfloor \tag{14}$$

In this case, the relation between $\mathbf{R} = \langle R_1, R_2, R_3 \rangle$ and the variables X_1, X_3, X_4 is not immediately clear, however, we can rewrite the above equations as

$$\mathbf{R} = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \\ 2 & 3 & 1 \end{bmatrix} \cdot \langle -\log(X_1), \sqrt{X_3}, \lfloor 10 \cdot x_4 \rfloor \rangle^T. \tag{15}$$

The matrix in Equation 15 is full-rank, which means it is invertible; it will be called $\mathbb{W}$. For every value of $\mathbf{R}$ such as $\mathbf{r} = \langle r_1, r_2, r_3 \rangle$, let $\tilde{r} := \mathbb{W}^{-1} \cdot \mathbf{r}$. From Eq. 15, we can derive,

$$\mathbf{R} = \mathbf{r} \iff X_1 = \exp(-\tilde{r}_1), X_3 = (\tilde{r}_2)^2, \text{ and } \frac{\tilde{r}_3}{10} \leq X_4 < \frac{\tilde{r}_3 + 1}{10}. \tag{16}$$

Let $x_1 := \exp(-\tilde{r}_1)$, $x_3 := (\tilde{r}_2)^2$, $x_4^a := \frac{\tilde{r}_3}{10}$, and $x_4^b := \frac{\tilde{r}_3+1}{10}$. We can compute $P^*(y \mid \mathbf{r})$ as,

$$P^*(y \mid \mathbf{r}) = P^*(y \mid x_1, x_3, X_4 \in [x_4^a, x_4^b)) \qquad \text{(By Equation 16)} \tag{17}$$

$$= \frac{P^*(y, X_4 \in [x_4^a, x_4^b) \mid x_1, x_3)}{\sum_{y=0}^{1} P^*(y, X_4 \in [x_4^a, x_4^b) \mid x_1, x_3)} \qquad \text{(coditioning)} \tag{18}$$

$$= \frac{P^1(y, X_4 \in [x_4^a, x_4^b) \mid x_1, x_3)}{\sum_{y=0}^{1} P^1(y, X_4 \in [x_4^a, x_4^b) \mid x_1, x_3)} \qquad (\{Y, X_4\}S\text{-adm. given } X_1, X_3) \tag{19}$$

$$= P^1(y \mid x_1, x_3, X_4 \in [x_4^a, x_4^b)) = P^1(y \mid \mathbf{r}). \qquad \text{(conditioning \& Equation 16)} \tag{20}$$

This allows us to compute the risk $R_{P^*}(\tilde{h} \circ \phi) = \int_{\text{supp}(\mathbf{R})} P^1(Y \neq \tilde{h}(\mathbf{r}) \mid \mathbf{r}) \cdot P^*(\mathbf{r}) \cdot d\mathbf{r}$.

Noticeably, the features X_3, X_4 are non-causal to the label Y, as there exists no direct path from them to Y in $\mathcal{G}^{\mathbf{\Delta}}$. However, it is valid in this case to use them for classification. This subtle point carries an important message; "causal" prediction is not necessarily superior, or even desirable, as the transportability theory might license us to use non-causal features for better classification. More generally, we can show the following formal result.

Theorem 2 (Risk Evaluation: UDA) *Consider a transportable representation $\mathbf{R} = \phi(\mathbf{X})$, and any classifier $\hat{h} = \tilde{h} \circ \phi$ defined based on it. For the UDA task where we have access to unlabeled data drawn from $P^*(\mathbf{x})$, the risk of $\hat{h}$, is unique under all target domains $\mathcal{M}^*$ compatible with $\mathbb{P}$ (i..e, data) and $\mathcal{G}^{\mathbf{\Delta}}$ (i.e., assumptions). Moreover, the risk can be transported (i.e., computed) via*

$$R^{tr}_{P^*}(\hat{h}) = \int_{\text{supp}(\mathbf{R})} P^{\text{tr}}(Y \neq \tilde{h}(\mathbf{r}) \mid \mathbf{r}) \cdot P^*(\mathbf{r}) \cdot d\mathbf{r}, \tag{21}$$

where $P^{\text{tr}}(y \mid \mathbf{r})$ denotes the expression obtained by Algorithm 1. □

This result provides a systematic way for computing the risk of a classifier for the UDA task, contingent on the corresponding representation being transportable. As alluded to in the introduction, this evaluation step yields optimal outcome once it is used for the search procedure.

Corollary 1 (Min-max optimality: UDA) *Minimizing $R^{\text{tr}}_{P^*}$ (Eq. 21) within the family of classifiers defined based on transportable representations yields min-max optimality in UDA task, i.e.,*

$$\arg\min_{h \in \mathcal{H}_{\text{tr}}} R^{\text{tr}}_{P^*}(h) = \arg\min_{h \in \mathcal{H}_{\text{tr}}} \max_{\bar{\mathcal{M}} \text{ compatible target given } \mathcal{G}^{\mathbf{\Delta}}, \mathbb{P}, P^*(\mathbf{x})} R_{P^{\bar{\mathcal{M}}}}(h), \tag{22}$$

where $\mathcal{H}_{\text{tr}} = \{h : \text{supp}(\mathbf{X}) \to \{0, 1\}$ s.t. $h = \tilde{h} \circ \phi$ where $\mathbf{R} = \phi(\mathbf{X})$ is transportable$\}$. □

We now turn our attention to the DG task, where $P^*(\mathbf{x})$ is not available. The next example illustrates that the risk might not be transported in these cases, but a bound over it can still be computed.

Example 6 (Transportability & risk: DG) In the context of Example 5, consider the representation $\mathbf{R} = \phi(\mathbf{X})$, where ϕ is defined in Example 5 (Eq. 15), and the classifier defined based on it is $\hat{h} = \tilde{h} \circ \phi$. The goal now is to evaluate the risk of $\hat{h}$ for the DG task.

The transformation $\mathbb{W}$ in Eq. 15 can be used to rewrite $\hat{h} = (\tilde{h} \circ \mathbb{W}) \circ (\mathbb{W}^{-1} \circ \phi)$, so that the classification component $\tilde{h} \circ W$ takes transformed representation $\tilde{\mathbf{R}} = (\mathbb{W}^{-1} \circ \phi)(\mathbf{x})$ as the input. We can then write:

$$P^*(y, \tilde{r}_3 \mid \tilde{r}_1, \tilde{r}_2) = P^*(y, X_4 \in [x_4^a, x_4^b) \mid x_1, x_3) \quad \text{(Eq. 16)} \tag{23}$$

$$= P^1(y, X_4 \in [x_4^a, x_4^b) \mid x_1, x_3) = P^1(y, \tilde{r}_3 \mid \tilde{r}_1, \tilde{r}_2). \quad \text{(S-admissibility)} \tag{24}$$

This enables us to derive the following to bound $R_{P^*}(\hat{h})$,

$$P^*(Y \neq \tilde{h}(\mathbb{W} \cdot \tilde{\mathbf{R}})) \tag{25}$$

$$= \int P^*(Y \neq \tilde{h}(\mathbb{W} \cdot \langle \tilde{r}_1, \tilde{r}_2, \tilde{R}_3 \rangle^T) \mid \tilde{r}_1, \tilde{r}_2) \cdot P^*(\tilde{r}_1, \tilde{r}_2) \cdot d\tilde{r}_1 \cdot d\tilde{r}_2 \quad \text{(conditioning)} \tag{26}$$

$$= \int P^*(\tilde{r}_1, \tilde{r}_2) \cdot P^1(Y \neq \tilde{h}(\mathbb{W} \cdot \langle \tilde{r}_1, \tilde{r}_2, \tilde{R}_3 \rangle^T) \mid \tilde{r}_1, \tilde{r}_2) \cdot d\tilde{r}_1 \cdot d\tilde{r}_2 \quad \text{(By Eq. 24)} \tag{27}$$

$$\leq \max_{\tilde{r}_1, \tilde{r}_2 \in \text{supp}(\tilde{R}_1, \tilde{R}_2)} P^1(Y \neq \tilde{h}(\mathbb{W} \cdot \langle \tilde{r}_1, \tilde{r}_2, \tilde{R}_3 \rangle^T) \mid \tilde{r}_1, \tilde{r}_2) \quad \text{(avg} \leq \text{max)} \tag{28}$$

The bound provided in Eq. 28 is tight in this case, as the maximum is attained by a compatible target domain (see Appendix C). □

The following more general result regarding bounds on the risk can be shown.

Theorem 3 (Risk Evaluation: DG) *Consider a transportable representation $\mathbf{R} = \phi(\mathbf{X})$, and let $\mathbf{Z}, \bar{\mathbf{Z}}, \bar{\mathbf{R}}, \mathcal{G}^{\mathbf{\Delta}}_{\text{aux}}, \mathbb{P}_{\text{aux}}$ denote the objects obtained by running rTR (Alg. 1) to r-transport the query $P^*(y \mid \mathbf{r})$. Suppose the query $P^*(\bar{\mathbf{z}} \mid \mathbf{z})$ is transportable given $\mathbb{P}$ and $\mathcal{G}^{\mathbf{\Delta}}$ (e.g., via gTR [19]). Then, the query $P^*(y, \bar{\mathbf{r}} \mid \mathbf{z})$ is transportable from $\mathcal{G}^{\mathbf{\Delta}}_{\text{aux}}, \mathbb{P}_{\text{aux}}$. Moreover, we can construct a mapping $\phi^*(\mathbf{Z}, \bar{\mathbf{R}}) = \mathbf{R}$, which enables us to compute a bound to the risk of $\hat{h} = \tilde{h} \circ \phi$ via,*

$$R_{P^*}(\hat{h}) \leq \max_{\mathbf{z} \in \text{supp}(\mathbf{Z})} P^{\text{tr}}(Y \neq \tilde{h} \circ \phi^*(\mathbf{z}, \bar{\mathbf{R}}) \mid \mathbf{z}). \tag{29}$$

Theorem 3 uses components of Algorithm 1 to offer a systematic method for bounding the worst-case risk. Further discussions on the nuances of computing risks are provided in Appendix C.

Our findings in this section proposes transportable representation as promising choices for both UDA and DG tasks, as we developed methods to transport/bound the risk of classifiers defined based on them. Finding transportable representations and evaluating the risk is contingent on r-transportability of certain quantities, and Algorithm 1 provides the machinery necessary to evaluate it.

4 Data-Driven Transportability of Representations

In this section, our goal is to provide an alternative way of expressing assumptions about the underlying mechanisms whenever the selection diagrams $\mathcal{G}^{\mathbf{\Delta}}$ are not available. We replace this assumption by enforcing a correspondence between source data $\mathbb{P}$ and the selection diagram $\mathcal{G}^{\mathbf{\Delta}}$ (Assumption 1), and imposing a restriction to the selection diagrams $\mathcal{G}^{\mathbf{\Delta}}$ (Assumptions 2). We state below a probabilistic condition that corresponds to a certain notion of stability across source domains.

Definition 8 (Invariance Property) *A representation $\mathbf{R} = \phi(\mathbf{X})$ is said to satisfy the invariance property w.r.t. a set of distributions $\mathbb{C} \subseteq \mathbb{P}$, when the following holds:*

$$I[\phi; \mathbb{C}] : \forall P^a, P^b \in \mathbb{C} \quad \forall \mathbf{r} \in \text{supp}(\mathbf{R}) \quad P^a(Y = 1 \mid \mathbf{r}) = P^b(Y = 1 \mid \mathbf{r}) \tag{30}$$

The invariance property is statistically testable assuming data from all the source domains is available, and is acknowledged in the literature (e.g., [31, 2, 33, 22, 9]). Representations that satisfy the invariance property w.r.t. the source domains $\mathbb{P}$ are proposed for OOD generalization purpose. This motivates us to examine this proposal theoretically. The next assumption enables us to read S-admissibility relations w.r.t. the pairs of source domains by testing the invariance property.

Assumption 1 (A1: r-Faithfulness) The set of distributions $\mathbb{P}$ is r-faithful to the selection diagrams $\mathcal{G}^{\boldsymbol{\Delta}}$ if for every representation $\mathbf{R} = \phi(\mathbf{X})$ the following holds:

$$I[\phi; \{P^i, P^j\}] \text{ (Def. 8)} \implies Y \perp\!\!\!\perp_d S_{i,j} \mid \mathbf{Z}, \bar{\mathbf{R}} \text{ in } \mathcal{G}^{\Delta_{i,j}}_{\text{aux}}, \tag{31}$$

Where $\mathbf{Z} = \mathbf{z}, \bar{\mathbf{R}} = \bar{\phi}(\bar{\mathbf{Z}})$ are determined & constrained variables obtained by solving the system of equation $\mathbf{R} = \phi(\mathbf{X})$, and $\mathcal{G}^{\boldsymbol{\Delta}}_{\text{aux}}$ is the auxiliary selection diagram (as constructed in Algorithm 1).

Asserting r-faithfulness guarantees that the invariances present within the source data are not coincidental, i.e., they correspond to separation statements in the selection diagrams. Under A1, the invariance properties within the source domains only inform about S-admissibility relations within the source domains, however, in the transportability activity, we use S-admissibility relations w.r.t. target domain and one of the source domains. To reason about source-target S-admissibility relations, we require that (1) the domain discrepancies between the target and the sources be constrained, and (2) the graphical representation of the target SCM corresponds to that of the source SCMs.

Assumption 2. (A2: Invariant Mechanisms & Shared Causal Structure)

$$\textbf{(a)} \quad V \notin \bigcup_{i,j=1}^{T} \Delta_{i,j} \implies V \notin \bigcup_{k=1}^{T} \Delta_{*,k}, \quad \forall V \in \mathbf{X} \cup \{Y\} \tag{32}$$

$$\textbf{(b)} \quad \mathcal{G}^1 = \mathcal{G}^2 = \cdots = \mathcal{G}^T = \mathcal{G}^* \tag{33}$$

In words, A2(a) states that if a mechanism is invariant across all source domains, this mechanism will match the mechanism in the target domain. For instance, the situation in Example 1 (HBP) satisfies this assumption, as the only unchanged mechanism across the sources is W (anxiety), and $\mathcal{G}^{\boldsymbol{\Delta}}$ suggests that it will remain unchanged in the target as well. On the other hand, if Δ_{*1} also contained W, i.e., $S_{*1} \to W$ existed in $\mathcal{G}^{\boldsymbol{\Delta}}$, then A2(a) would not hold. Imposing Assumption A2(b) is common in causal approaches to generalization (e.g., [28, 33, 25]), and rejects the possibility of cases where the underlying causal diagram of the domains is different. A more thorough discussion on assumptions 1&2 and their implications is provided in Appendix D & E. Asserting A1 & A2 above deems the source domain invariance property as a sound and complete criterion to decide r-transportability, as shown by the next result.

Theorem 4 (Data-driven r-transportability) *Under Assumptions 1,2, for a representation* $\mathbf{R} = \phi(\mathbf{X})$*, the query* $P^*(y \mid \mathbf{r})$ *is r-transportable from* $\mathcal{G}^{\boldsymbol{\Delta}}$ *and* $\mathbb{P}$ *if and only if the representation* ϕ *satisfies the invariance property* $I[\phi; \mathbb{P}]$*.* □

This result suggests that the invariance conditions can be used to decide r-transportability of queries using data, as an alternative to the assumptions encoded in the selection diagram. Under the non-parametric Assumption 1,2, source invariance properties (Def. 8) provide a sound and complete criterion to decide transportable representations, and then evaluate the risk of classifiers defined based on them; more discussion in Appendix E.

5 Conclusions

In this paper, we bridged the gap between transportability theory and the OOD generalization problems through formal causal language. We introduced and characterized transportable representations that exhibit desirable analytical properties, including the evaluation of the worst-case risk of a classifier for the domain generalization and the unsupervised domain adaptation tasks. We develop an algorithm for deciding whether a representation is transportable across domains given a collection of datasets and structural assumptions encoded in the form of selection diagrams. Finally, we develop an alternative approach for eliciting assumptions in graphical form, and showed that under r-faithfulness (A1), invariant mechanisms (A2(a)), and shared causal diagrams (A2(b)) assumptions, invariance tests across all source domains constitute a complete criterion for deciding transportability. Our findings provide a formal justification for the assumptions and conditions under which invariant representations can be evaluated in OOD tasks, which opens a new thread of research for the graphical analysis of representations and their properties through transportability lenses.

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] 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.

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

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

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

[18] 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.

[19] 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.

[20] 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.

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

[22] 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.

[23] 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.

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

[25] Joris M. Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *J. Mach. Learn. Res.*, 21(1), jun 2022. ISSN 1532-4435.

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

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

[28] 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.

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

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

[31] 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.

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

[33] 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.

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

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

[36] 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.

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

[38] 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.

[39] 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.

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

[41] 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.

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

[43] 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.

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

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

[46] 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.

[47] 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.