Transportable Representations for Out-of-distribution Generalization

Kasra Jalaldoust   Elias Bareinboim

Abstract

Building on the theory of causal transportability (Bareinboim & Pearl), we define in this paper the notion of "transportable representations," and show that the out-of-distribution generalization risk of classifiers defined based on these representations can be bounded, considering that graphical assumptions about the underlying system are provided.

1. Introduction

Generalizing findings across settings is central throughout human experience. 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). 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 [21; 27; 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 of the mechanisms shared across settings [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; 4; 5; 22; 10; 11; 16].

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 [30; 29; 28], or use of distance measures to relate the source and target distributions [7; 14]. 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 [19; 23; 2; 25; 31; 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 [24]. 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; 13; 18; 17; 34; 33]. Also, causal features have been used in representations to help predictions across domain, while filtering out the spurious correlations that might be unstable across domains [32; 26; 20; 15].

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 a setting where labeled data from source domains is available, however, no data from the target domain is available. We also analyze in this setting 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 always desirable while spurious should be discarded. The preliminaries are provided in Appendix A.

2. Examples & Results

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 (Lee et al., 2020) to describe mismatch of mechanisms between two SCMs.

Definition 2.1 (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 (Lee et al., 2020) to graphically represent the system that includes multiple SCMs relative to the collection of source and target domains.

Definition 2.2 (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 $\mathcal{G}^{\boldsymbol{\Delta}} = \{\mathcal{G}^*\} \cup \{\mathcal{G}^{\Delta_{ij}}\}_{i,j=1}^T$ 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 2.3 (Transportability)**.** For subsets of variables $\mathbf{C}, \mathbf{W} \subset \mathbf{X} \cup \{Y\}$ in the SCM, the query $P^*(\mathbf{c} \mid \mathbf{w})$ is transportable if for every pair of SCMs $\mathcal{M}_a^*, \mathcal{M}_b^*$ compatible with the selection diagrams $\mathcal{G}^{\boldsymbol{\Delta}}$, and the distributions $\mathbb{P}$ over $\mathbf{X} \cup \{Y\}$, $P^{\mathcal{M}_a^*}(\mathbf{c} \mid \mathbf{w}) = P^{\mathcal{M}_b^*}(\mathbf{c} \mid \mathbf{w})$. □

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 OOD generalization task comprises the labeled data drawn from each $P^i \in \mathbb{P}$. Next, we formally define classifiers which use a representation of the input.

Definition 2.4 (Representations for classification)**.** The variable $\mathbf{R} = \phi(\mathbf{X})$ is called a representation for every mapping $\phi : \text{supp}(\mathbf{X}) \rightarrow \text{supp}(\mathbf{R})$. 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 $h : \text{supp}(\mathbf{X}) \rightarrow \{0, 1\}$ 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, $R_{P^*}(\hat{h}) := P^*(Y \neq h(\mathbf{X}))$.

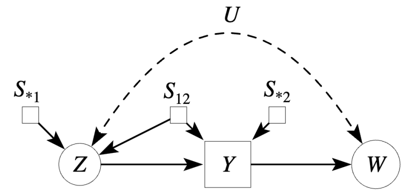

Figure 1: Selection diagram corresponding to Examples 2.5 & 2.7

Example 2.5 (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 are 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. 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 $R = \phi(Z, W) := Z - W$. One can construct a classifier based on the value of this representation, namely,

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

In words, $\hat{h}$ suggests that the person is in high risk if their wellness index R is below threshold c. □

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

Definition 2.6 (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,

$$\mathbf{A} \perp\!\!\!\perp_d S_{ij} \mid \mathbf{Z} \text{ in } \mathcal{G}^{\Delta_{ij}} \implies P^i(\mathbf{a} \mid \mathbf{z}) = P^j(\mathbf{a} \mid \mathbf{z}). \quad (1)$$

Note that S-admissibility connects the assumptions encoded in the graphical model about the underlying mechanisms, as formalized in Def. 2.2, and the mechanisms represented by the underlying and unobserved generating SCMs, to elicit invariances at the probabilistic level (r.h.s. of Eq. 1). 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 2.7 (Risk evaluation through joint transportability)**.** Considering the classifier $\hat{h}(z, w)$ of Ex. 2.5, we attempt to transport the joint distribution of Z, Y, W as,

$$\begin{aligned} P^*(z, y, w) &= P^*(z) \cdot P^*(y \mid z) \cdot P^*(w \mid y, z) \\ &= P^2(z) \cdot P^1(y \mid z) \cdot P^2(w \mid y, z) \end{aligned}$$

The last line 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 $P^*(r \mid y, z, w) = \mathbb{1}_{\{z-w=r\}}$, we can derive,

$$\begin{aligned} &P^*(y, r) \\ &= \int P^*(z, y, w, r) \cdot dz \cdot dw \\ &= \int P^*(z, y, w) \cdot P^*(r \mid z, y, w) \cdot dz \cdot dw \\ &= \int P^2(z) \cdot P^1(y \mid z) \cdot P^2(w \mid y, z) \cdot \mathbb{1}_{\{z-w=r\}} \cdot dz \cdot dw \end{aligned}$$

Having this joint distribution allows us to compute the risk $R_{P^*}(\hat{h}) = P^*(Y \neq \hat{h}(Z, W)) = P^*(Y \neq R)$. Thus, the first step of the procedure discussed in Sec 1 (Evaluation) can be executed, i.e., the risk can be evaluated via the source data drawn from $P^1(z, w, y), P^2(z, w, y)$. Tuning the parameters of the classifier and the representation to minimize this quantity (Search) would asymptotically yield a min-max optimal solution under the graphical assumptions encoded in the selection diagrams. □

The derivation in Example 2.7 leads to a more general decision problem that asks whether certain distributions can be computed from the available data considering a given representation. The next example shows that the strategy used in Ex. 2.7 is not always applicable for deciding r-transportability, but it's neither necessary.

Example 2.8 (Complex representation)**.** 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 only one source domain $\mathcal{M}^1$. Further, consider the

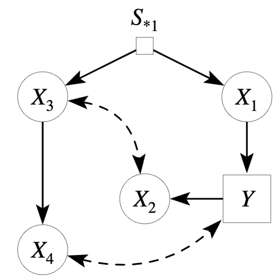

Figure 2: selection diagram corresponding to Example 2.8

representation

$$\begin{aligned} R_1 &= -\log(X_1) + 2 \cdot \sqrt{X_3} + 3 \cdot \lfloor 10 \cdot X_4 \rfloor \\ R_2 &= -3\log(X_1) + 1 \cdot \sqrt{X_3} + 2 \cdot \lfloor 10 \cdot X_4 \rfloor \\ R_3 &= -2\log(X_1) + 3 \cdot \sqrt{X_3} + \lfloor 10 \cdot X_4 \rfloor \end{aligned}$$

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

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

The matrix above 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. 2, we can derive the following conditions on $\mathbf{X}$ equivalent to the condition $\mathbf{R} = \mathbf{r}$;

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

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,

$$\begin{aligned} &P^*(y \mid r) \\ &= P^*(y \mid x_1, x_3, X_4 \in [x_4^a, x_4^b)) \\ &= \frac{P^*(y, X_4 \in [x_4^a, x_4^b) \mid x_1, x_3)}{\sum_{y=0}^1 P^*(y, X_4 \in [x_4^a, x_4^b) \mid x_1, x_3)} \\ &= \frac{P^1(y, X_4 \in [x_4^a, x_4^b) \mid x_1, x_3)}{\sum_{y=0}^1 P^1(y, X_4 \in [x_4^a, x_4^b) \mid x_1, x_3)} \qquad (S\text{-adm.}) \\ &= P^1(y \mid x_1, x_3, X_4 \in [x_4^a, x_4^b)) = P^1(y \mid \mathbf{r}) \qquad (3) \end{aligned}$$

The transformation $\mathbb{W}$ in Eq. 2 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:

$$\begin{aligned}&P^*(y, \tilde{r}_3 \mid \tilde{r}_1, \tilde{r}_2)\\&= P^*(y, X_4 \in [x_4^a, x_4^b) \mid x_1, x_3)\\&= P^1(y, X_4 \in [x_4^a, x_4^b) \mid x_1, x_3) \qquad \text{(S-adm.)}\\&= P^1(y, \tilde{r}_3 \mid \tilde{r}_1, \tilde{r}_2).\end{aligned}$$

This enables us to derive the following to bound the risk as,

$$\begin{aligned}&R_{P^*}(\hat{h})\\&= P^*(Y \neq \tilde{h} \circ \phi(\mathbf{x}))\\&= P^*(Y \neq \tilde{h}(\mathbb{W} \cdot \tilde{\mathbf{R}}))\\&= \sum_{\tilde{r}_1, \tilde{r}_2} P^*(Y \neq \tilde{h}(\mathbb{W} \cdot \langle \tilde{r}_1, \tilde{r}_2, \tilde{R}_3 \rangle^T) \mid \tilde{r}_1, \tilde{r}_2) \cdot P^*(\tilde{r}_1, \tilde{r}_2)\\&= \sum_{\tilde{r}_1, \tilde{r}_2} P^*(\tilde{r}_1, \tilde{r}_2) \cdot P^1(Y \neq \tilde{h}(\mathbb{W} \cdot \langle \tilde{r}_1, \tilde{r}_2, \tilde{R}_3 \rangle^T) \mid \tilde{r}_1, \tilde{r}_2)\\&\leq \max_{\tilde{r}_1, \tilde{r}_2 \in \mathrm{supp}(\tilde{R}_1, \tilde{R}_2)} P^1(Y \neq \tilde{h}(\mathbb{W} \cdot \langle \tilde{r}_1, \tilde{r}_2, \tilde{R}_3 \rangle^T) \mid \tilde{r}_1, \tilde{r}_2)\end{aligned} \tag{4}$$

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

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. Motivated by Example 2.8, we define the following concepts.

Definition 2.9 (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**.

As seen in Example 2.8, the key to blocking the path $S \to X_1 \to Y$ is through discovering the fact that the condition $\mathbf{R} = \mathbf{r}$ in this special case of the expression ϕ determines the value of X_1. The following definition is introduced accordingly.

Definition 2.10. (Determined and constrained variables) The variables $\mathbf{Z} \subseteq \mathbf{X}$ are determined by the system of equations $\mathbf{R} = \phi(\mathbf{X})$ if for some mapping ψ the equation $\mathbf{Z} = \psi(\mathbf{R})$ can be derived algebraically. A variable is unconstrained by $\mathbf{R} = \phi(\mathbf{X})$ if it can be algebraically removed from the expression ϕ. Variables $\bar{\mathbf{Z}} \subseteq \mathbf{X}$ are constrained by $\mathbf{R} = \phi(\mathbf{Z})$ if they are neither unconstrained nor determined.

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

1: $\langle \mathbf{Z} = \psi(\mathbf{R}), \bar{\mathbf{R}} = \bar{\phi}(\bar{\mathbf{Z}}) \rangle \leftarrow \mathrm{solve}(\mathbf{R} = \phi(\mathbf{X}))$
2: $\mathcal{G}^{\mathbf{\Delta}}_{\mathrm{aux}}$: Add to every graph in $\mathcal{G}^{\mathbf{\Delta}}$ the variable $\bar{\mathbf{R}}$ & arrows from $\bar{\mathbf{Z}}$ to $\bar{\mathbf{R}}$
3: $\mathbb{P}_{\mathrm{aux}} := \{P^i_{\mathrm{aux}}(\mathbf{x}, y, \bar{\mathbf{r}}) := P^i(\mathbf{x}, y) \cdot \mathbb{1}_{\{\bar{\mathbf{r}} = \bar{\phi}(\bar{\mathbf{z}})\}}\}_{P^i \in \mathbb{P}}$
4: **return** gTR(query: $P^*(y \mid \mathbf{z}, \bar{\mathbf{r}}); \mathcal{G}^{\mathbf{\Delta}}_{\mathrm{aux}}, \mathbb{P}_{\mathrm{aux}}$) [Lee et al. (2020)]

In Example 2.8, the variables X_1, X_3 are determined by $\mathbf{R}$, X_2 is unconstrained by it, and X_4 is constraint by it.

We propose algorithm 1 to decide r-transportability, and show the following.

Theorem 2.11. *Algorithm 1 is sound for r-transportability.*

All proofs are provided in Appendix B.

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 2.8) or constrained (e.g., X_4 in Example 2.8) 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. (2020)). Detailed explanation of the Algorithm 1 is provided in Appendix C. The next result provides a bound for the risk.

Theorem 2.12 (Risk Evaluation)**.** *Consider a transportable representation $\mathbf{R} = \phi(\mathbf{X})$, and let $\mathbf{Z}, \bar{\mathbf{Z}}, \bar{\mathbf{R}}, \mathcal{G}^{\mathbf{\Delta}}_{\mathrm{aux}}, \mathbb{P}_{\mathrm{aux}}$ denote the objects obtained by solving the system of equations $\mathbf{R} = \phi(\mathbf{X})$ (Def. 2.10). Suppose the query $P^*(\bar{\mathbf{z}} \mid \mathbf{z})$ is transportable given $\mathbb{P}$ and $\mathcal{G}^{\mathbf{\Delta}}$ (e.g., via gTR [16]). Then, the query $P^*(y, \bar{\mathbf{r}} \mid \mathbf{z})$ is transportable from $\mathcal{G}^{\mathbf{\Delta}}_{\mathrm{aux}}, \mathbb{P}_{\mathrm{aux}}$. 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,*

$$R_{P^*}(\hat{h}) \leq \max_{\mathbf{z} \in \mathrm{supp}(\mathbf{Z})} P^{\mathrm{tr}}(Y \neq \tilde{h} \circ \phi^*(\mathbf{z}, \bar{\mathbf{R}}) \mid \mathbf{z}). \tag{5}$$

Theorem 2.12 offer a systematic method for bounding the worst-case risk, under the assumption that $P^*(\bar{\mathbf{z}} \mid \mathbf{z})$ is transportable, which can be verified graphically. Further discussions on the nuances of computing risks are provided in Appendix D.

3. Conclusion

Our findings suggest study of transportable representation as promising choices for the OOD generalization task. We characterize these representations graphically via Algorithm 1 (Theorem 2.11), and propose a risk evaluation method by computing a bound for the risk of classifiers defined based on them through Theorem 2.12. This bound can be further used for the search procedure to find an optimal classifier.

References

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

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

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

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

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

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

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. Analysis of representations for domain adaptation. In Schölkopf, B., Platt, J., and Hoffman, T. (eds.), *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.

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

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

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

Correa, J. D. and Bareinboim, E. From statistical transportability to estimating the effect of stochastic interventions. In *IJCAI*, pp. 1661–1667, 2019.

David, S. B., Lu, T., Luu, T., and Pal, D. Impossibility theorems for domain adaptation. In Teh, Y. W. and Titterington, M. (eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 129–136, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.

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

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

Krueger, D., Caballero, E., Jacobsen, J.-H., Zhang, A., Binas, J., Zhang, D., Priol, R. L., and Courville, A. Out-of-distribution generalization via risk extrapolation (rex). In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5815–5826. PMLR, 18–24 Jul 2021.

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

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

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

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

Mao, C., Xia, K., Wang, J., Wang, H., Yang, J., Bareinboim, E., and Vondrick, C. 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*, pp. 7511–7521. IEEE, 2022.

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

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

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

Rosenfeld, E., Ravikumar, P. K., and Risteski, A. The risks of invariant risk minimization. In *International Conference on Learning Representations*, 2021.

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

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

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

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

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

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

Wald, Y., Feder, A., Greenfeld, D., and Shalit, U. On calibration and out-of-domain generalization. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021.

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

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

Zhang, K., Gong, M., and Schoelkopf, B. Multi-source domain adaptation: A causal view. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), Feb. 2015.

A. 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 $\mathrm{supp}(\mathbf{Z})$, and values in the support are denoted by the corresponding lowercase letter, e.g., $\mathbf{z} \in \mathrm{supp}(\mathbf{Z})$. To denote $P(\mathbf{A} = \mathbf{a} \mid \mathbf{B} = \mathbf{b})$, we use the shorthand $P(\mathbf{a} \mid \mathbf{b})$. The notion $\perp\!\!\!\perp_d$ denotes d-separation in graphs.

We use semantics of Structural Causal Models (Pearl, 2000), which will allow the formal articulation of the invariances needed to extrapolate findings across settings, as defined next:

Definition A.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 $f_V : \mathrm{supp}(\mathbf{U}_V) \times \mathrm{supp}(\mathbf{Pa}_V) \to \mathrm{supp}(V)$ 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 $P^{\mathcal{M}}(\mathbf{v}) = \int_{\mathrm{supp}(\mathbf{U})} \prod_{V \in \mathbf{V}} P^M(v \mid \mathbf{pa}_V, \mathbf{u}_V) \cdot P(\mathbf{u}) \cdot d\mathbf{u}$, 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 $\mathcal{G}$.

B. Proofs

B.1. Proof of Theorem 2.11

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 C.1). Therefore, $P^*(y \mid \mathbf{r}) = P^*(y \mid \mathbf{z}, \bar{\mathbf{r}})$.

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

$$Q^*[\mathbf{C}](\mathbf{c}, \mathbf{pa_C}) := P^*(\mathbf{c} \mid do(\mathbf{pa_C} \setminus \mathbf{C})), \tag{6}$$

where $\mathbf{pa_C} := \bigcup_{C \in \mathbf{C}} \mathbf{pa}_C$. By Theorem 2 from Lee et al. (2020),

$$P^*(y \mid \mathbf{z}, \bar{\mathbf{r}}) = \frac{\sum_{\mathbf{a} \setminus (\{y\} \cup \mathbf{w}_Y)} Q^*[\mathbf{A}]}{\sum_{\mathbf{a} \setminus \mathbf{w}_Y} Q^*[\mathbf{A}]}, \tag{7}$$

where,

$$(\mathcal{G}^*_{\text{aux}})_{\underline{\mathbf{Z} \cup \bar{\mathbf{R}}}} = \text{Take } \mathcal{G}^*_{\text{aux}} \in \mathcal{G}^{\mathbf{\Delta}}_{\text{aux}} \text{ and cut the outgoing arrows of } \mathbf{Z} \cup \bar{\mathbf{R}} \tag{8}$$

$$\mathbf{W}_Y = \{V \in \mathbf{Z} \cup \bar{\mathbf{R}} \text{ connected to } Y \text{ by any path in } (\mathcal{G}^*_{\text{aux}})_{\underline{\mathbf{Z} \cup \bar{\mathbf{R}}}}\} \tag{9}$$

$$\mathbf{A} = \{V \in \mathbf{V} : \text{ there exists a directed path from } V \text{ to } Y \cup \mathbf{W}_Y \text{ in } (\mathcal{G}^*_{\text{aux}})_{\underline{\mathbf{Z} \cup \bar{\mathbf{R}}}} \} \tag{10}$$

The gTR algorithm decomposes $Q[\mathbf{A}]$ according to

$$Q^*[\mathbf{A}] = Q^*[\mathbf{A}^1] \cdot Q^*[\mathbf{A}^2] \cdot \cdots \cdot Q^*[\mathbf{A}^K] \cdot Q^*[\bar{\mathbf{R}}] =: Q^*[\mathbf{A}_0] \cdot Q^*[\bar{\mathbf{R}}] \tag{11}$$

Next, it attempts to identify each c-factor from some source domain using the sub-routine IDENTIFY (Lee et al., 2020). 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,

$$Q^*[\bar{\mathbf{R}}] = Q^i[\bar{\mathbf{R}}] \tag{12}$$

$$= P^i(\bar{\mathbf{r}} \mid \bar{\mathbf{z}}) \qquad \text{(c-factor rules)} \tag{13}$$

$$= \mathbb{1}_{\{\bar{\mathbf{r}} = \bar{\phi}(\bar{\mathbf{z}})\}} \qquad \text{(computable from } P^i_{\text{aux}}) \tag{14}$$

Suppose the gTR algorithm returns an expression for the c-factor $Q^*[\mathbf{A}]$. We can apply Lemma 4 by Lee et al. (2020) 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 7 which is a valid transportation formula for $P^*(y \mid \mathbf{z}, \bar{\mathbf{r}})$ and is equal to the target query $P^*(y \mid \mathbf{r})$.

B.2. Proof of Theorem 2.12

Appendix C introduces concepts necessary for understanding the proof. First, we show that $P^*(y, \bar{\mathbf{r}} \mid \mathbf{z})$ is transportable.

$$
\begin{aligned}
P^*(y, \bar{\mathbf{r}} \mid \mathbf{z}) &= P^*(y \mid \bar{\mathbf{r}}, \mathbf{z}) \cdot P^*(\bar{\mathbf{r}} \mid \mathbf{z}) && (15)\\
&= P^*(y \mid \mathbf{R} = \phi^*(\mathbf{z}, \bar{\mathbf{r}})) \cdot P^*(\bar{\mathbf{r}} \mid \mathbf{z}) && (\phi^* \text{ from App. C.1}) \quad (16)\\
&= P^{\text{tr}}(y \mid \mathbf{r}) \cdot P^*(\bar{\mathbf{r}} \mid \mathbf{z}) && (\text{r-transportable query}) \quad (17)\\
&= P^{\text{tr}}(y \mid \mathbf{r}) \cdot \int_{\{\bar{\mathbf{z}}:\bar{\mathbf{r}}=\bar{\phi}(\bar{\mathbf{z}})\}} P^*(\bar{\mathbf{z}} \mid \mathbf{z}) \cdot d\bar{\mathbf{z}} && (\text{change of var.}) \quad (18)\\
&= P^{\text{tr}}(y \mid \mathbf{r}) \cdot \int_{\{\bar{\mathbf{z}}:\bar{\mathbf{r}}=\bar{\phi}(\bar{\mathbf{z}})\}} P^{\text{tr}}(\bar{\mathbf{z}} \mid \mathbf{z}) \cdot d\bar{\mathbf{z}} && (\text{transportable query}) \quad (19)
\end{aligned}
$$

Let the transportation formula in Equation 19 be denoted as $P^{\text{tr}}(y, \bar{\mathbf{r}} \mid \mathbf{z})$. Next, we derive the bound for the risk.

$$
\begin{aligned}
R_{P^*}(h) &= P^*(Y \neq \tilde{h} \circ \phi(\mathbf{X})) && (h = \tilde{h} \circ \phi) \quad (20)\\
&= P^*(Y \neq \tilde{h}(\mathbf{R})) && (\phi(\mathbf{X}) = \mathbf{R}) \quad (21)\\
&= P^*(Y \neq (\tilde{h} \circ \phi^*)(\mathbf{Z}, \tilde{\mathbf{R}})) && (\phi^* \text{ in App. C.1}) \quad (22)\\
&= \int_{\text{supp}(\mathbf{Z})} P^*(Y \neq (\tilde{h} \circ \phi^*)(\mathbf{z}, \tilde{\mathbf{R}}) \mid \mathbf{z}) \cdot P^*(\mathbf{z}) \cdot d\mathbf{z} && (\text{Law of total prob.}) \quad (23)\\
&= \int_{\text{supp}(\mathbf{Z})} P^{\text{tr}}(Y \neq (\tilde{h} \circ \phi^*)(\mathbf{z}, \tilde{\mathbf{R}}) \mid \mathbf{z}) \cdot P^*(\mathbf{z}) \cdot d\mathbf{z} && (\text{Eq. 19}) \quad (24)\\
&\leq \max_{\mathbf{z} \in \text{supp}(\mathbf{Z})} P^{\text{tr}}(Y \neq (\tilde{h} \circ \phi^*)(\mathbf{z}, \tilde{\mathbf{R}}) \mid \mathbf{z}) && (\text{avg} \leq \text{max}) \quad (25)
\end{aligned}
$$

The latter is computable using the transportation formula for $P^{\text{tr}}(y, \bar{\mathbf{r}} \mid \mathbf{z})$ is Equation 19. Appendix D provides a discussion on tightness of the bounds achieved above.

C. Discussion on Algorithm 1

C.1. Solving the system of equations

For the system of equations $\mathbf{R} = \phi(\mathbf{X})$, Define

$$\mathcal{T} = \{\langle \mathbf{r}, \mathbf{x} \rangle \in \text{supp}(\mathbf{R}) \times \text{supp}(\mathbf{X}) \text{ s.t. } \mathbf{r} = \phi(\mathbf{x})\} \tag{26}$$

A variable $X \in \mathbf{X}$ is determined by $\mathbf{R}$, if for every value $\mathbf{r}_0 \in \text{supp}(\mathbf{R})$, the set

$$\{x_0 : \mathbf{x}_0 \in \text{supp}(\mathbf{X}) \text{ s.t. } \langle \mathbf{r}_0, \mathbf{x}_0 \rangle \in \mathcal{T}\} \tag{27}$$

has only one element. Let $\mathbf{Z} \subseteq \mathbf{X}$ denote the set of determined variables.

A variable $X \in \mathbf{X}$ is unconstrained by $\mathbf{R}$, if for every value $\mathbf{r}_0 \in \text{supp}(\mathbf{R})$, the set in Equation 27 is equal to $\text{supp}(X)$. Let $\bar{\mathbf{Z}} \subseteq \mathbf{X} \setminus \mathbf{Z}$ denoted the set of constrained variables, i.e., the variables that are neither determined nor unconstrained by $\mathbf{R}$.

In Example 2.8, X_1, X_3 are determined, X_4 is constrained, and X_2 is unconstrained by $\mathbf{R}$.

A solution to the system of equations $\mathbf{R} = \phi(\mathbf{X})$ is a function $\psi : \text{supp}(\mathbf{R}) \to \text{supp}(\mathbf{Z})$ for which the equation $\mathbf{Z} = \psi(\mathbf{R})$ can be algebraically proved from $\mathbf{R} = \phi(\mathbf{X})$. Solving systems of equations is a well-studied subject, and here we view the solving procedure as a black-box.

We can plug in the value of $\mathbf{Z} = \psi(\mathbf{R})$ in the expression for ϕ to obtain $\mathbf{R} = \phi(\bar{\mathbf{Z}}, \psi(\mathbf{R}), \mathbf{X} \setminus (\bar{\mathbf{Z}} \cup \mathbf{Z}))$. Next, we can massage this expression to rewrite it without the unconstrained variables $\mathbf{X} \setminus (\bar{\mathbf{Z}} \cup \mathbf{Z})$. Without loss of generality, suppose $\mathbf{R} = \phi(\bar{\mathbf{Z}}, \psi(\mathbf{R}))$. Next, we massage the expression to move every term containing $\mathbf{R}$ to the l.h.s., and call the expression $\bar{\mathbf{R}}$. Then, the expression in terms of $\bar{\mathbf{Z}}$ remained on the r.h.s. is denoted as $\bar{\phi}$, i.e.,

$$\mathbf{R} = \phi(\bar{\mathbf{Z}}, \psi(\mathbf{R})) \iff \bar{\mathbf{R}} = \bar{\phi}(\bar{\mathbf{Z}}) \tag{28}$$

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 28 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 given to us, and then the whole expression for $\mathbf{R}$ is determined, i.e., does not depend on any unknown variable. Let $\phi^* : \text{supp}(\mathbf{Z}) \times \text{supp}(\bar{\mathbf{R}})$ denote the described mapping that allows us to compute $\mathbf{r}$ from $\bar{\mathbf{r}}, \mathbf{z}$. We use this mapping in other parts of the appendix.

D. Discussion on Risk Bounds for Domain Generalization

Theorem 2.12 provides us with a bound for the risk, however, minimizing this bound would not necessarily yield an optimal outcome, as the bound might be loose. Tightness of the bound can be an interesting discussion, however, here we only elaborate this in the context of Example 2.8.

D.1. Tightness of bounds: Example 2.8

In this example, we concluded with the bound,

$$R_{P^*}(h) \leq \max_{\tilde{r}_1, \tilde{r}_2 \in \text{supp}(\tilde{R}_1, \tilde{R}_2)} P^1(Y \neq \tilde{h}(\mathbb{W} \cdot \langle \tilde{r}_1, \tilde{r}_2, \tilde{R}_3 \rangle^T) \mid \tilde{r}_1, \tilde{r}_2). \tag{29}$$

An important question is whether there exists a target SCM compatible with $\mathcal{G}^{\mathbf{\Delta}}, \mathbb{P}$ such that it entails the risk equal to the upper-bound achieved above.

Suppose r_1^*, r_2^* denote the arguments achieving the maximum in Eq. 3. Construct the SCM $\bar{\mathcal{M}}$ from $\mathcal{M}^1$ by modifying the assignments for X_1, X_3 into,

$$X_1 \leftarrow \exp(\mathbb{W}^{-1} \cdot r_1^*), X_3 \leftarrow (\mathbb{W}^{-1} \cdot r_2^*)^2. \tag{30}$$

Notice, $\bar{M}$ is compatible with the selection diagram $\mathcal{G}^{\mathbf{\Delta}}$ in Figure 2, as the domain discrepancy between $\bar{\mathcal{M}}, \mathcal{M}^i$ matches $\Delta_{*1} = \{X_1, X_3\}$. The risk under domain $\bar{\mathcal{M}}$ is,

$$R_{P^{\bar{\mathcal{M}}}}(\hat{h}) = \int P^{\bar{\mathcal{M}}}(\mathbf{r}) \cdot P^1(Y \neq \tilde{h}(\mathbb{W} \cdot \langle \tilde{r}_1, \tilde{r}_2, \tilde{R}_3 \rangle^T \mid \tilde{r}_1, \tilde{r}_2) \cdot d\tilde{r}_1 \cdot d\tilde{r}_2 \tag{31}$$

$$= P^1(Y \neq \tilde{h}(\mathbb{W} \cdot \langle \tilde{r}_1, \tilde{r}_2, \tilde{R}_3 \rangle^T) \mid r_1^*, r_2^*). \tag{32}$$

Therefore, the bound for the risk is tight in this case, and minimizing it as an optimization objective yields min-max optimality.

We achieved the above tightness result because both determined variables X_1, X_3 were connected to the S-node, which makes it possible to construct a worst-case SCM so that they take their worst-case value. However, this approach fails once the determined variables are not directly connected to the S-nodes. In that case, the worst-case approach would yield a loose bound.