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 \(X \cup \{Y\}\), where \(Y\) is a binary label. SCMs \(M_1, M_2, \ldots, M_T\) defined over \(X \cup \{Y\}\) denote the source domains, and entail the distributions \(P = \{P^1, P^2, \ldots, P^T\}\), while they induce the causal diagrams \(G^1, G^2, \ldots, G^T\). There exists an unknown SCM \(M^*\) representing the target domain, which entails the distribution \(P^*\), while it induces the causal diagram \(G^*\). 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 $X \cup \{Y\}$, the domain discrepancy set $\Delta_{ab} \subseteq V$ is defined such that for every $V \in \Delta_{ab}$ there might exist a discrepancy $f_{M^a}^V \neq f_{M^b}^V$ or $P^{M^a}(uv) \neq P^{M^b}(uv)$. \hfill $\Box$

In other words, $V \notin \Delta_{ab}$ is equivalent assuming the same mechanisms for $V$ across $M^a, M^b$, i.e., $f_{M^a}^V = f_{M^b}^V$ and $P^{M^a}(uv) = P^{M^b}(uv)$. 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 $G^{\Delta_{ij}}$ is constructed from $G^i \ (i \in \{*, 1, 2, \ldots, T\})$ 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 $G^A = \{G^i\} \cup \{G^{\Delta_{ij}}\}_{ij=1}^2$ encodes the graphical assumptions. If the causal diagram is shared across the domains, we can use a single graph to depict $G^A$. \hfill $\Box$

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 $C, W \subset X \cup \{Y\}$ in the SCM, the query $P^*(c \mid w)$ is transportable if for every pair of SCMs $M^a, M^b$ compatible with the selection diagrams $G^A$, and the distributions $P$ over $X \cup \{Y\}$, $P^{M^a}(c \mid w) = P^{M^b}(c \mid w)$. \hfill $\Box$

The joint distribution $P^*(x, y)$ 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 $P$ and qualitative assumptions encoded by the selection diagrams $G^A$. The notion of transportability describes such a property.

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

Definition 2.4 (Representations for classification). The variable $R = \phi(X)$ is called a representation for every mapping $\phi : \text{supp}(X) \rightarrow \text{supp}(R)$. Furthermore, a representation is said to satisfy the coverage property w.r.t. the distribution $P(x)$ if $P(X \in \{x : \phi(x) = r\}) > 0$ for every $r \in \text{supp}(R)$. A mapping $h : \text{supp}(X) \rightarrow \{0, 1\}$ is said to be a classifier defined based on the representation $R = \phi(X)$ if it can be expressed as composition with $\phi$, i.e., $h = h \circ \phi$. \hfill $\Box$

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

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 $X = \{Z, W\}$ 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 $M^1, M^2$, respectively. The patients from $M^1$ 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 $M^*$ are at risk of HBP based on the same features $X$. The relationships across domains are summarized through the selection diagrams $G^A$ shown in Figure 1. In the domain $M^1$, patients are genetically prone to HBP, similar to $M^2$, thus, the mechanisms deciding blood pressure ($Y$) in $M^*$ is the same as $M^1$, while differing from $M^2$. However, in $M^*$, the government is not running the exercising TV ads, and the mechanism determining exercise is the same as in $M^2$, while differing from $M^1$. Further, the mechanism determining anxiety ($W$) is invariant across sources and target domains. All these invariances can be written as $\Delta_{s1} = \{Z\}$ and $\Delta_{s2} = \{Y\}$, 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,$$
\hat{h}(z, w) := \mathbb{I}_{\{\phi(z, w) \leq c\}} = \mathbb{I}_{\{r \leq c\}} = \mathbb{I}_{\{z - w \leq c\}}.
$$

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

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 $M^i, M^j \ (i, j \in \{*, 1, 2, \ldots, T\})$, and sets of variables $Z, A \subset X \cup \{Y\}$. $A$ is said to be $S$-admissible given $Z$ w.r.t. the domains $M^i, M^j$ whenever $A$ is separated from $S_{ij}$ given $Z \in G^{\Delta_{ij}}$. Furthermore, if $S$-admissibility holds, then the conditional distribution of $A$ given $Z$ is invariant.
across $\mathcal{M}^2$ and $\mathcal{M}^4$. In summary,

$$A \perp_{\mathcal{D}} S_{ij} \mid Z \in \mathcal{G}^{\Delta_{ij}} \implies P_i^j(a \mid z) = P^j(a \mid z). \quad (1)$$

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 $h(z, w)$ of Ex. 2.5, we attempt to transport the joint distribution of $Z, Y, W$ as,

$$P^*(z, y, w) = P^*(z) \cdot P^*(y \mid z) \cdot P^*(w \mid y, z)$$

$$= P^2(z) \cdot P^1(y \mid z) \cdot P^2(w \mid y, z)$$

The last line follows since $Z$ is (marginally) $S$-admissible in $\mathcal{M}^2$, $\mathcal{M}^4$, $Y$ 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,

$$P^*(y, r) = \int P^*(z, y, w, r) \cdot dz \cdot dw$$

$$= \int P^*(z, y, w) \cdot P^*(r \mid z, y, w) \cdot dz \cdot dw$$

$$= \int P^2(z) \cdot P^1(y \mid z) \cdot P^2(w \mid y, z) \cdot \mathbb{1}_{\{z = w = r\}} \cdot dz \cdot dw$$

Having this joint distribution allows us to compute the risk $R_P(h) = P^*(Y \neq h(Z, W)) = P^*(Y \neq R)$. 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 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 decision transportability, but it’s neither necessary.

**Example 2.8** (Complex representation). Consider the selection diagram $\mathcal{G}^\Delta$ in Figure 2, over the variables $Y$ and $X = \{X_1, X_2, X_3, X_4\}$ with $\text{supp}(X_i) = (0, 1)$. There exists only one source domain $\mathcal{M}^1$. Further, consider the representation

$$R_1 = -\log(X_1) + 2 \cdot \sqrt{X_3} + 3 \cdot [10 \cdot X_4]$$

$$R_2 = -3\log(X_1) + 1 \cdot \sqrt{X_3} + 2 \cdot [10 \cdot X_4]$$

$$R_3 = -2\log(X_1) + 3 \cdot \sqrt{X_3} + [10 \cdot X_4]$$

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

$$R = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \\ 2 & 3 & 1 \end{bmatrix} \cdot \langle -\log(X_1), \sqrt{X_3}, [10 \cdot X_4] \rangle^T. \quad (2)$$

The matrix above is full-rank, which means it is invertible; it will be called $\mathbb{W}$. For every value of $R$ such as $r = \langle r_1, r_2, r_3 \rangle$, let $\tilde{r} := \mathbb{W}^{-1} \cdot r$. From Eq. 2, we can derive the following conditions on $X$ equivalent to the condition $R = r$:

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

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

$$P^*(y \mid r) = P^*(y \mid x_1, x_3, X_4 \in [x_3^a, x_4^b])$$

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

$$= \sum_{y=0}^1 \frac{P^1(y, X_4 \in [x_3^a, x_4^b] \mid x_1, x_3)}{\sum_{y=0}^1 \sum_{y=0}^1 P^1(y, X_4 \in [x_3^a, x_4^b] \mid x_1, x_3)} \quad (S\text{-adm.})$$

$$= \sum_{y=0}^1 \frac{P^1(y \mid x_1, x_3, X_4 \in [x_3^a, x_4^b])}{P^1(y \mid x_1, x_3)} = P^1(y \mid r) \quad (3)$$

The transformation $\mathbb{W}$ in Eq. 2 can be used to rewrite $\hat{h} = (\hat{h} \circ \mathbb{W}) \circ (\mathbb{W}^{-1} \circ \phi)$, so that the classification component $\hat{h} \circ W$ takes transformed representation $\tilde{R} = (\mathbb{W}^{-1} \circ \phi)(x)$.\]
as the input. We can then write:
\[ P^*(y, r_3 \mid \tilde{r}_1, \tilde{r}_2) = P^*(y, X_4 \in [x_{14}^1, x_{14}^2] \mid x_1, x_3) = P^1(y, X_4 \in [x_{14}^1, x_{14}^2] \mid x_1, x_3) \quad \text{(S-adm.)} \]
\[ = P^1(y, r_3 \mid \tilde{r}_1, \tilde{r}_2). \]

This enables us to derive the following to bound the risk as,
\[ R_{P^*}(\hat{h}) = P^*(Y \neq \hat{h} \circ \phi(x)) \]
\[ = \sum_{\tilde{r}_1, \tilde{r}_2} P^*(Y \neq \tilde{h}(\mathbb{W} \cdot \tilde{R})) \cdot P^*(\tilde{r}_1, \tilde{r}_2) \]
\[ = \sum_{\tilde{r}_1, \tilde{r}_2} P^*(\tilde{r}_1, \tilde{r}_2) \cdot P^1(Y \neq \tilde{h}(\mathbb{W} \cdot \tilde{r}_1, \tilde{r}_2, \tilde{R}_3) \mid \tilde{r}_1, \tilde{r}_2) \]
\[ \leq \max_{\tilde{r}_1, \tilde{r}_2 \in \text{supp}(\tilde{r}_1, \tilde{r}_2)} P^1(Y \neq \tilde{h}(\mathbb{W} \cdot \tilde{r}_1, \tilde{r}_2, \tilde{R}_3) \mid \tilde{r}_1, \tilde{r}_2) \quad (4) \]

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

Noticably, the features \(X_3, X_4\) are non-causal to the label \(Y\), as there exists no direct path from them to \(Y\) in \(G^\Delta\). 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 \(R = \phi(X)\) be a representation. The query \(P^*(y \mid r)\) is r-transportable given (1) the set of distributions \(P\), (2) the selection diagrams \(G^\Delta\), and (3) the arithmetic expression \(\phi\), if for every two SCMs \(M^*_1, M^*_2\) compatible with \(P\) and \(G^\Delta\), \(P^M_\phi(y \mid r) = P^{M_2}_\phi(y \mid r)\). If so, \(\phi\) will be called a transportable representation.

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

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

**Algorithm 1** rTR: r-transport \(P^*(y \mid r)\) from \(P, G^\Delta, \phi\).

1: \(Z = \psi(R), \tilde{R} = \phi(Z) \leftarrow \text{solve}(R = \phi(X))\)
2: \(G^\Delta_{aux}\): Add to every graph in \(G^\Delta\) the variable \(\tilde{R}\) & arrows from \(Z\) to \(\tilde{R}\)
3: \(P_{aux} := \{P_{aux}^i(x, y, \tilde{r}) := P^i(x, y) \cdot I_{\tilde{r} = \phi(Z)}\} P_{aux} \in P\)
4: return \(\text{gTR}(query: P^*(y \mid z, \tilde{r}); G^\Delta_{aux}, P_{aux})\) [Lee et al. (2020)]

In Example 2.8, the variables \(X_1, X_3\) are determined by \(R\), \(X_2\) is unconstrained by it, and \(X_4\) is constrained 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 \(\phi\) 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 \(R = r\). 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 \(R = \phi(X)\), and let \(Z, Z, R, G^\Delta_{aux}, P_{aux}\) denote the objects obtained by solving the system of equations \(R = \phi(X)\) (Def. 2.10). Suppose the query \(P^*(z \mid x)\) is transportable given \(P\) and \(G^\Delta\) (e.g., via gTR [16]). Then, the query \(P^*(y \mid z, \tilde{r})\) is transportable from \(G^\Delta_{aux}, P_{aux}\). Moreover, we can construct a mapping \(\phi^*(Z, R) = R\), which enables us to compute a bound to the risk of \(\hat{h} = h \circ \phi\) via,
\[ R_{P^*}(\hat{h}) \leq \max_{z \in \text{supp}(Z)} P^\Delta(Y \neq h \circ \phi^*(z, R) \mid z). \quad (5) \]

Theorem 2.12 offer a systematic method for bounding the worst-case risk, under the assumption that \(P^*(z \mid x)\) 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


A. Preliminaries

We use upper-case letters (e.g., 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 Z is denoted as \( \text{supp}(Z) \), and values in the support are denoted by the corresponding lowercase letter, e.g., \( z \in \text{supp}(Z) \). To denote \( P(A = a \mid B = b) \), we use the shorthand \( P(a \mid b) \). The notion \( \triangleq \) 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 \( \langle U, V, \mathcal{F}, P(u) \rangle \), where \( U \) is a set of exogenous (unobserved) variables; \( V \) is a set of endogenous (observed) variables; \( \mathcal{F} \) represents a collection of functions \( f = \{f_v\} \) such that each endogenous variable \( V \in V \) is determined by a function \( f_v \in \mathcal{F} \), where \( f_v : \text{supp}(U_V) \times \text{supp}(P_{a_V}) \rightarrow \text{supp}(V) \) with \( U_V \subseteq U \), and \( P_{a_V} \subseteq V \setminus \{V\} \); The uncertainty is encoded through a distribution over the exogenous variables, \( P(u) \).

Every SCM \( \mathcal{M} \) induces a causal diagram, which is a directed acyclic graph where any variable \( V \in V \) is a vertex, and there exists a directed edge from every variable in \( P_{a_V} \) to \( V \). Also, for every pair \( V, V' \in V \) such that \( U_V \cap U_{V'} \neq \emptyset \), there exists a bidirected edge between \( V \) and \( V' \). We denote this causal diagram with the letter \( G \). A SCM \( \mathcal{M} \) induces a probability distribution \( P^M(v) \) over the set of observed variables \( V \) such that \( P^M(v) = \int_{\text{supp}(U)} \prod_{v \in V} P^M(v \mid \text{pa}_V, u_v) \cdot P(u) \cdot du \), where each term \( P(v \mid \text{pa}_V, u_v) \) 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, \( P^M(v) > 0 \), for every \( v \). 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 \( G \).

B. Proofs

B.1. Proof of Theorem 2.11

The condition \( \mathbf{R} = \phi(X) \) is equivalent to \( Z = z, \bar{R} = \bar{\phi}(\bar{Z}) \), and the latter is obtained by solving the system of equations \( \mathbf{R} = \phi(X) \) (more in Appendix C.1). Therefore, \( P^*(y \mid \bar{r}) = P^*(y \mid z, \bar{r}) \).

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

\[
Q^*[C](c, \text{pa}_C) := P^*(c \mid do(\text{pa}_C \setminus C)),
\]

where \( \text{pa}_C := \bigcup_{C \subseteq C} \text{pa}_C \). By Theorem 2 from Lee et al. (2020),

\[
P^*(y \mid z, \bar{r}) = \frac{\sum_{\omega(y) \in \text{W}_V} Q^*[\mathcal{A}]}{\sum_{\omega(y) \in \text{W}_V} Q^*[\mathcal{A}]},
\]

where,

\[
(\mathcal{G}^*_\text{aux})_{Z \cup \bar{R}} = \text{Take } \mathcal{G}^*_\text{aux} \in \mathcal{G}^\Delta \text{ and cut the outgoing arrows of } Z \cup \bar{R}
\]

\[
\text{W}_Y = \{ V \in Z \cup \bar{R} \text{ connected to } Y \text{ by any path in } (\mathcal{G}^*_\text{aux})_{Z \cup \bar{R}} \}
\]

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

The gTR algorithm decomposes \( Q[A] \) according to

\[
Q^*[A] = Q^*[A^1] \cdot Q^*[A^2] \cdot \cdots \cdot Q^*[A^K] \cdot Q^*[\bar{R}] =: Q^*[A_0] \cdot Q^*[\bar{R}]
\]

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 \( Q^*[\bar{R}] \), the algorithm can transport it from any source distribution, i.e., \( Q^*[\bar{R}] = Q^i[\bar{R}] \) for every \( 1 \leq i \leq T \). In \( P \) notation,

\[
Q^*[\bar{R}] = Q^i[\bar{R}] \quad (c\text{-factor rules})
\]

\[
= P^i(\bar{r} \mid \bar{z}) \quad (computable from } P^i_{\text{aux}})
\]

\[
= \mathbb{1}_{(\bar{r} = \phi(\bar{z}))}
\]
Suppose the gTR algorithm returns an expression for the c-factor $Q^*[A]$. We can apply Lemma 4 by Lee et al. (2020) in a topological order to deduce $P^*(y \mid z, r)$ is transportable if and only if $\sum_{A \subseteq \{(y)\cup yz\}} Q^*[A]$ is transportable. In case $Q^*[A]$ is transported by gTR, the algorithm returns the expression in Equation 7 which is a valid transportation formula for $P^*(y \mid z, r)$ and is equal to the target query $P^*(y \mid r)$.

B.2. Proof of Theorem 2.12

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

$$P^*(y, r \mid z) = P^*(y \mid r, z) \cdot P^*(r \mid z)$$

(r-transportable query) (19)

(15)

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

$$R_{P^{tr}}(h) = P^*(Y \neq \tilde{h} \circ \phi(X))$$

$$= P^*(Y \neq \tilde{h}(R))$$

$$= P^*(Y \neq (\tilde{h} \circ \phi^*)(Z, \tilde{R}))$$

$$= \int_{\text{supp}(Z)} P^*(Y \neq (\tilde{h} \circ \phi^*)(Z, \tilde{R}) \mid Z) \cdot P^*(Z) \cdot dZ$$

(Law of total prob.) (20)

(25)

(26)

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

A variable $X \in X$ is unconstrained by $R$, if for every value $r_0 \in \text{supp}(R)$, the set

$$\{x_0 : x_0 \in \text{supp}(X) \text{ s.t. } (r_0, x_0) \in T\}$$

(27)

In Example 2.8, $X_1, X_3$ are determined, $X_4$ is constrained, and $X_2$ is unconstrained by $R$. A solution to the system of equations $R = \phi(X)$ is a function $\psi : \text{supp}(R) \rightarrow \text{supp}(Z)$ for which the equation $Z = \psi(R)$ can be algebraically proved from $R = \phi(X)$. 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 $Z = \psi(R)$ in the expression for $\phi$ to obtain $R = \phi(Z, \psi(R), X \setminus (Z \cup Z))$. Next, we can massage this expression to rewrite it without the unconstrained variables $X \setminus (Z \cup Z)$. Without loss of generality, suppose $R = \phi(Z, \psi(R))$. Next, we massage the expression to move every term containing $R$ to the l.h.s., and call the expression $\bar{R}$. Then, the expression in terms of $\bar{Z}$ remained on the r.h.s. is denoted as $\phi$, i.e.,

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

once we fix the value of $\bar{r} = \bar{\phi}(\bar{Z})$ and $Z = z$, we can obtain $r = \phi(Z, Z)$ in the following way: As we have access to $\bar{r}$, we can revert the derivation in equation 28 to obtain $R = \phi(Z, \psi(R))$ only dependent on the unknown $\psi(R)$. Next, we can substitute the term $\psi(R)$ with its known value $z$ that is given to us, and then the whole expression for $R$ is determined, i.e., does not depend on any unknown variable. Let $\phi^* : \text{supp}(Z) \times \text{supp}(\bar{R})$ denote the described mapping that allows us to compute $r$ from $\bar{r}, z$. 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,

$$R_{P^*}(h) \leq \max_{\hat{r}_1, \hat{r}_2 \in \text{supp}(\hat{r}_1, \hat{r}_2)} P^1(Y \neq \hat{h}(\hat{W} \cdot (\hat{r}_1, \hat{r}_2, \hat{R}_3)^T) | \hat{r}_1, \hat{r}_2).$$

(29)

An important question is whether there exists a target SCM compatible with $G^\Delta, \mathcal{P}$ 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 $\tilde{M}$ from $M^1$ by modifying the assignments for $X_1, X_3$ into,

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

(30)

Notice, $\tilde{M}$ is compatible with the selection diagram $G^\Delta$ in Figure 2, as the domain discrepancy between $M, M^1$ matches $\Delta_{x_1} = \{X_1, X_3\}$. The risk under domain $\tilde{M}$ is,

$$R_{P,\tilde{M}}(h) = \int P^M(r) \cdot P^1(Y \neq \tilde{h}(\mathcal{W} \cdot (\tilde{r}_1, \tilde{r}_2, \tilde{R}_3)^T) | \tilde{r}_1, \tilde{r}_2) \cdot d\tilde{r}_1 \cdot d\tilde{r}_2$$

(31)

$$= P^1(Y \neq \tilde{h}(\mathcal{W} \cdot (\tilde{r}_1, \tilde{r}_2, \tilde{R}_3)^T) | r_1^*, r_2^*).$$

(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.