PAC Generalization via Invariant Representations

Advait Parulekar [1] **Karthikeyan Shanmugam** [2] **Sanjay Shakkottai** [1]

Abstract

Invariant representations are transformations of the covariates such that the best model on top of the representation is invariant across training environments. In the context of linear Structural Equation Models (SEMs), invariant representations might allow us to learn models with out-of-distribution guarantees, i.e., models that are robust to interventions in the SEM. To address the invariant representation problem in a *finite sample* setting, we consider the notion of ϵ-approximate invariance. We study the following question: If a representation is approximately invariant with respect to a given number of training interventions, will it continue to be approximately invariant on a larger collection of unseen intervened SEMs? Inspired by PAC learning, we obtain finite-sample out-of-distribution generalization guarantees for approximate invariance that holds *probabilistically* over a family of linear SEMs without faithfulness assumptions. We also show how to extend our results to a linear indirect observation model that incorporates latent variables.

1. Introduction

A common failure of empirical risk minimization in machine learning is that it is beneficial to exploit spurious correlations to minimize training loss. A classic example of this comes from (Beery et al., 2018) where we have a dataset with pictures of cows and camels. In most cases, the cows are in pastures and the camels are in deserts, and a classifier that works well on such a dataset generalizes poorly to classifying cows in deserts. How do we prevent a classifier tasked with distinguishing pictures of cows from pictures of camels from using the color of the images? Certainly doing so might actually help with the classification - take an extreme example wherein rather than just having an indicative background, the species of the animal is annotated in the image itself. In this case, simply learning to interpret what is annotated will lead to an excellent classifier. However, such a classifier will perform poorly on a future dataset with incorrect or missing annotations.

In asking for such "out-of-distribution" (OOD) generalization (generalization to different distributions rather than different samples from the same distribution), we are asking for more than what is guaranteed by traditional PAC learning. For instance, we might have a *covariate shift*, meaning the marginal on the covariates of the joint distributions of our training and test data is different. We would like a classifier that is trained on training environments to generalize to the test environments.

Recently, in the OOD generalization literature, the data for the training and test environments have been modeled to arise from a set of *causal models* that are intervened versions of each other. The true causal relationship in causal models between a target variable and its causal parents remains invariant while other relationships could change (Peters et al., 2016; Bareinboim et al., 2012). Building on this observation, authors in (Arjovsky et al., 2019) proposed to learn a representation across the training environments such that the optimal classifier on top of it remains the same. We refer to this as the *invariant representation* in this work. Will this representation and the predictor generalize (remain invariant) in unseen environments? Does such a formulation result in improvements over other methods such as empirical risk minimization or distributionally robust optimization?

There has been some debate about this matter. Some theoretical works (Rosenfeld et al., 2020), (Kamath et al., 2021) describe regimes and settings in which either the convex penalty suggested by (Arjovsky et al., 2019) or the invariance formulation in general are not sufficient to generalize to new environments. On the other hand, there are also settings described in (Arjovsky et al., 2019), (Ahuja et al., 2020), (Kamath et al., 2021) in which IRM does provably lead to OOD generalization guarantees that are not possible using other methods. However, these prior studies require strong conditions such as faithfulness (Yang et al., 2018), (Pearl, 2009), that come from the structure learning community. Indeed, as noted in (Uhler et al., 2013), faithfulness (that observed invariances imply structural constraints on the

[1]Department of Electrical and Computer Engineering, University of Texas at Austin [2]Google Research India. Correspondence to: Advait Parulekar <advaitp@utexas.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

causal graph) is a very strong assumption in the finite sample setting. This is because the volume of the set of linear SEMs that are "close" to ones with faithfulness violations is a large constant fraction of all linear SEMs. This makes it difficult to resolve the question of whether an observed conditional independence truly holds due to the underlying causal structure with finite samples. Invariance based results almost always study the infinite sample setting and assume general position conditions on the training environments; similar to the faithfulness assumption, these become much stronger assumptions in the finite sample setting.

Thus in the worst case, perhaps we cannot expect generalization guarantees without an exponentially many interventions/samples. Instead of studying the worst case setting, suppose we consider the *PAC* setting – can we now derive finite sample generalization guarantees without the need of strong faithfulness or general position assumptions, yet with a polynomial scaling in interventions/samples?

1.1. Contributions

We answer the above affirmatively, and provide the *first generalization and associated finite-sample guarantees for invariant representations without faithfulness assumptions.* We study a setting in which we have access to a family of linear SEMs related to each other by hard and soft interventions on arbitrary subsets of variables. We assume that the training environments arise by random sampling from a distribution over this family. We derive a PAC bound for the number of interventions needed to get probabilistic generalization on the family of SEMs. A central result is that we need only $O(\frac{n^4}{\delta'})$ training environments, such that approximately invariant representations on the training set generalizes with probability $1-\delta'$ on the family of unseen linear SEMs. Further our theory also works for general linear representations beyond feature selection for indirect and direct observation models (e.g., could have latent variables) with a scaling of $O(\frac{n^8}{\delta'})$; please see Appendix 9.4 for details. We show tighter interventional complexity independent of n for SEM families that have simultaneous atomic interventions on any fixed set of k nodes and soft interventions on k nodes in a degree bounded setting.

Secondly, we characterize the number of samples in each training environment that is needed for approximate invariance to hold with high probability. This gives an end-to-end sample complexity guarantee along with the interventional complexity required in the above setting without any faithfulness or general position assumptions. Furthermore, we show extensions of our results to a linear indirect observation model that incorporates latent variables.

Finally, we empirically demonstrate that in the setting described above using a notion of generalization that we describe, *most* approximately invariant representations generalize to *most* new distributions.

1.2. Related Work

Causal Structure Learning: One approach, in the presence of data sampled from structural models, is simply to try and learn the DAG structure explicitly. Algorithms that use conditional independence (CI) tests and/or penalized likelihood scoring are used to determine the DAG (Spirtes et al., 2000; Solus et al., 2021; Chickering, 2020; Brouillard et al., 2020). These methods can only learn DAGs up to what is called the Markov Equivalence Class (MEC) (Verma & Pearl, 1990), or $\mathcal{I}$-MECs (Interventional MECs) when given access to general interventions (Hauser & Bühlmann, 2012; Squires et al., 2020; Brouillard et al., 2020) (Yang et al., 2018) under some form of faithfulness assumptions on the underlying causal model. It has been shown that some forms of faithfulness assumption are actually stronger than one might expect; linear SEMs that are close to a faithful violation form a much larger set (Uhler et al., 2013) and are difficult to distinguish in finite samples from faithful ones. However, in light of the fact that we only want generalization guarantees for invariance and not structure recovery, one could argue that learning the structure, or even feature selection is stronger than what we actually need.

Domain Adaptation Methods: Domain adaptation literature (Ganin et al., 2016; Muandet et al., 2013; Ben-David et al., 2007) learns representations whose distribution does not change across training and an unlabeled test distribution. Another line of work searches for the best mixture of training distributions to train on or mixture of pretrained models for generalization to unlabeled test data or test data with limited test labels (Mansour et al., 2021; 2008). Distributionally robust optimization approaches that optimize the worst risk over an uncertainty set of distributions have been proposed (Duchi et al., 2021; Sagawa et al., 2019). Please see Appendix A of (Gulrajani & Lopez-Paz, 2020) for a more exhaustive list of works. Robust optimization and multi source domain adaptation methods search in an uncertainty ball around the training distributions while domain adaptation fail even with a shift in marginal distributions of the labels (Zhao et al., 2019). Such methods fail when the test distribution is outside the convex hull of training distributions. (Magliacane et al., 2017) use invariance tests, which require faithfulness assumptions. (Chen et al., 2023) applies a multi-objective optimization perspective to the OOD problem to resolve some of the optimization difficulties of OOD generalization. Finally, the interventional complexity of learning invariant representations is also considered in (Chen et al., 2022) and (Wang et al., 2022). Under the data model of (Rosenfeld et al., 2020), (Chen et al., 2022) show that a smoothed analysis (with distributional assumptions) can get logarithmic interventional complexity, while (Wang

et al., 2022) use second-order moments to get just O(1) training environments in the infinite sample limit.

Finite Samples and Invariant Risk Minimization: One framework proposed to bypass structure learning difficulties and directly look for invariant representations is to use *Invariant Risk Minimization* (IRM) (Arjovsky et al., 2019). (Arjovsky et al., 2019) show that under some general position assumptions on the population covariances of the training distributions the exact invariant representation recovers the true causal parents. The work of (Ahuja et al., 2020) makes similar assumptions to derive sample complexity guarantees. (Kamath et al., 2021) also consider the question of getting generalization guarantees from IRM for general distributions, and address the number of training intervention needed for IRM; however, they work in the infinite sample limit with exact invariance and require an analytical mapping between intervention index set and training distributions. All these works assume regularity conditions and provide deterministic guarantees for a test distribution.

Invariant Predictors on DAGs: Recent works (Subbaswamy et al., 2019) have assumed the knowledge of the causal DAG behind the unseen test and/or train distributions along with information about the nodes that are intervened on in the test. It is then possible to characterize all possible invariant representations as a function of the graph. In our study, we do not assume any such side information about the SEM underlying our distributions. Finally, in (Rothenhäusler et al., 2021), there is a known set of variables (anchors) whose effect is considered to be spurious (in our setting, we do not know these). They study a weighted projection that penalizes estimates that are aligned with the anchor observations. In contrast, we study the IRM objective: rather than being a min-max of residual error over some set of distributions, our objective encourages representations that lead to invariant least squares solutions.

2. Model

2.1. Structural Equation Model

We consider a causal generative model specified by a linear Structural Equation Model (SEM) - data that is generated by linear equations on a directed acyclic graph (DAG) of variables. Specifically, we work with a random $n+1-$dimensional vector $X = (X_1, X_2, \cdots, X_n, X_t)$. We assume that there is some unknown directed acyclic graph G with nodes $(v_1, v_2, \cdots, v_n, v_t)$ which specifies the relationships between the variables in X, so that the node v_i corresponds to the random variable X_i. The variables are related to each other by structural equations of the form:

$$X_i^\theta = \left(\beta_i^\theta\right)^\top \mathtt{Pa}_i^\theta + \eta_i^\theta = \sum_{j \in \mathtt{Pa}_i^\theta} \beta_{j,i}^\theta X_j^\theta + \eta_i^\theta, \quad (1)$$

$$X_t = \beta_t^\top \mathtt{Pa}_t + \eta_t = \sum_{j \in \mathtt{Pa}} \beta_{j,t} X_j + \eta_t. \quad (2)$$

Here η_i are the exogenous variables in our system and X_i denotes the endogenous variables in the system. $\mathtt{Pa}_i$ denote the set of random variables (stacked as a column vector) corresponding to the parents of node v_i in G. The weights of the linear SEM, when there is a directed edge between node v_j and v_i are represented by $\beta_{j,i}$. There is some target node X_t that is special in that we would like to predict its value given the rest of the variables. We will denote the vector consisting of the rest of the variables as X_{-t}. We will denote by barred quantities the zero-padded versions of the above vectors, so $\overline{\beta}_t$, is such that $\beta_t^\top \mathtt{Pa}_t = \overline{\beta}_t^\top X_{-t}$.

2.2. Interventions

An important feature of Equation (1) is the superscript θ for $\theta \in \Theta$, an element of some *intervention* index set Θ. Θ specifies a parameterized intervention family, i.e. a family of linear SEMs related to each other by interventions. Conceptually various environments correspond to different mechanisms by which causal parents influence children. The changes in mechanisms are reflected in the weights $\beta_{j,i}^\theta$ parameterized by θ. Note that the conditional distribution of the *target* variable X_t given its parents $\mathtt{Pa}_t$ is assumed to be the same even for different $\theta \in \Theta$. This is a property that we would like to exploit so that our task of predicting X_t is robust to different interventions in Θ, i.e. there exists at least one robust predictor for all environments. We begin with an observational distribution with weights given by $\beta^{\theta_\circ}$. There are two common types of interventions studied:

Atomic or Hard Interventions: In hard interventions θ at some node v_i, we assign a value to X_i. This corresponds to setting $\beta_{j,i} = 0$ for all $X_j \in \mathtt{Pa}_i$, and setting $\eta_i^\theta = a_i^\theta$.

Soft Interventions: In *soft* interventions, the weights $\beta_i^\theta \neq \beta_i^{\theta_\circ}$ are altered while keeping the noise the same $\eta_i^\theta = \eta_i^{\theta_\circ}$.

The SEM parameters θ determine a joint distribution over X, which we denote by Δ_θ.

2.3. Data

Distributions over Interventions: In our model, we consider a distribution $\mathcal{D}$ over Θ from which we assume that our training and test data-sets are drawn by sampling the intervention index θ independently and randomly from $\mathcal{D}$. We provide generalization bounds that work with high probability over the randomness of the choice of test distribution, rather than deterministically as has been done in prior works (as we will explore in more detail, providing deterministic bounds is not possible without additional assumptions). Formally, Θ_{train} is an index set of interventions drawn i.i.d. from $\mathcal{D}$ over Θ. The test intervention index θ_{test} is also drawn from $\mathcal{D}$ independently. The learner does not see any details

about the intervention (which nodes are affected, etc), but sees the distinct datasets from the distinct interventions.

Data generation: In this sense, the training data is generated by first sampling Θ_{train} from $\mathcal{D}$, and then sampling $\mid \Theta_{\text{train}} \mid$ training datasets $\mathbf{X}^\theta \sim \Delta_\theta$ for $\theta \in \Theta_{\text{train}}$.

3. PAC-Invariant Representations

Motivated by Invariant Risk Minimization (IRM), we consider the search for a model that performs well under a variety of distributions in the hopes of attaining generalization guarantees. Consider a prediction function f with loss $\mathcal{R}^\theta(f)$. Here the superscript θ represents the intervention index. In this paper, we consider least squares regression, $f \in \mathbb{R}^n$ is a vector, and the loss is given by

$$\mathcal{R}^\theta(f) = \mathbb{E}_\theta[\left(f^\top X_{-t} - X_t\right)^2] \tag{3}$$

where the expectation is taken with respect to Δ_θ. We are interested in generalization guarantees for representations Φ that satisfy the property that

$$f = \arg\min_f \mathcal{R}^\theta(f \circ \Phi) \qquad \forall \theta \in \Theta_{\text{train}}, \tag{4}$$

is invariant across environments. That is, the least squares solution on top of the representation is the same for every intervention. We will denote the set of all invariant representations over a class of SEMs Θ' by $\mathcal{I}(\Theta')$.

$$\mathcal{I}(\Theta') = \{\Phi : \exists\, f \text{ s.t. } f = \arg\min_f \mathcal{R}^\theta(f \circ \Phi)\ \forall \theta \in \Theta'\}$$

We refer to the full model $f \circ \Phi$ when Φ is invariant as an *invariant solution*. We use $\mathcal{R}^\theta_\Phi(f)$ to denote the loss for f on top of representation Φ, so $\mathcal{R}^\theta_\Phi = \mathcal{R}^\theta(f \circ \Phi)$. We refer to f as the *head* of the model.

Feature selection, a motivation for the study of invariant representations, corresponds to diagonal representations; henceforth, we focus on the class of representations that are feature selectors, given by diagonal matrices.

Invariance via gradients: Rather than working with the loss itself, we follow (Ahuja et al., 2020), (Arjovsky et al., 2019) and use the *gradient* of the loss instead (see Lemma 8.5).

$$\Phi \in \mathcal{I}(\Theta_{\text{train}}) \iff \exists f \text{ s.t. } \nabla_f \mathcal{R}^\theta_\Phi(f) = 0 \ \ \forall \theta \in \Theta_{\text{train}}.$$

Unlike previous work, we do not further minimize the weighted empirical loss over training environments. We will instead consider the generalizability of *any* approximately invariant representation.

Finite samples and $\epsilon-$approximately Invariant Representations: Since we are working with finite samples, we can no longer hope for *exact* invariance across environments. We slightly change the definition of invariance to be about the gradient being close to 0.

$$\Phi \in \mathcal{I}^\epsilon(\Theta_{\text{train}}) \iff \exists f \text{ s.t. } \|\nabla \mathcal{R}^\theta_\Phi(f)\|_2 \le \epsilon\ \forall \theta \in \Theta_{\text{train}}. \tag{5}$$

We refer to approximate versions of quantities using a superscript $\cdot^\epsilon$, and, given specific datasets, we refer to finite sample versions of these quantities using hats, so for instance the set of representations that are invariant for some set Θ' is denoted $\mathcal{I}(\Theta')$, the set of $\epsilon-$approximately invariant representations is given by $\mathcal{I}^\epsilon(\Theta')$, finally, given a particular dataset, the set of $\epsilon-$approximately invariant representations would be denoted $\hat{\mathcal{I}}^\epsilon(\Theta')$. Note that the finite sample quantities are random (from the randomness of the samples drawn from Δ_θ), while the $\epsilon-$approximate quantities are deterministic. The dataset of N points is denoted in matrix form as $\mathbf{X}$ with $\mathbf{X}_{-t}$ and $\mathbf{X}_t$ denoting the non-target matrix and target vector respectively.

Significance of linear representations: A linear representation Φ affects invariance non-trivially by selecting an "effective" column space for the regression, that is, in determining the column space of $\Phi^\top \mathbf{X}_{-t}$. Because we are looking for a fixed (across environments) least-squares head on top of the representation, Φ can be further composed with any invertible linear map, only to be inverted at the head to obtain another invariant solution $(\Phi A, A^{-1} f)$ from an invariant solution (Φ, f). Because of this freedom, we can actually choose to work with some fixed head $f_\circ$. When the space of Φ is all matrices $f_\circ$ is chosen to be $(1, 0 \ldots 0)$. When the space of Φ is only diagonal matrices, then $f_\circ$ is taken to be $(1, 1 \ldots 1)$. These are formalized in Lemma 8.3 and Lemma 8.4. As we will see, a key construction is that of $\mathcal{I}^\epsilon_{f_\circ}(\theta)$, defined as those Φ such that the gradient is near-zero at $f_\circ$, that is,

$$\Phi \in \mathcal{I}^\epsilon_{f_\circ}(\theta) \iff \|\nabla \mathcal{R}^\theta_\Phi(f_\circ)\| \le \epsilon,$$

for some specific $f_\circ$. Similarly, for some set of interventions Θ', we define $\mathcal{I}^\epsilon_{f_\circ}(\Theta') = \bigcap_{\theta \in \Theta'} \mathcal{I}^\epsilon_{f_\circ}(\theta)$. This allows us to look at invariance as a property of a representation and a *single* intervention, rather than that over collection of interventions as in Equation 4. This is discussed further in Remark 3.6.

3.1. Motivation for Probabilistic Guarantees

Consider a distribution $\mathcal{D}$ over interventions. During training, we see samples from specific interventions drawn from this distribution Θ_{train}, and compute invariant representations for these interventions. Ideally, we would like to be able to say that an invariant solution computed in this way will generalize to all future interventions on the same DAG. However, this is not possible without further assumptions, as the following examples illustrate.

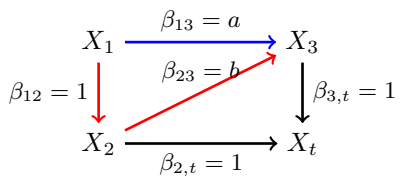

Figure 1. Figure for Example 3.1 showing that there could be more independencies in the data than apparent from the DAG alone.

Example 3.1 (Faithfulness)**.** *Consider a linear SEM on four variables X_1, X_2, X_3, X_t given by*

$$X_1 = \mathcal{N}(0,1), \qquad X_2 = X_1 + \mathcal{N}(0,1),$$
$$X_3 = aX_1 + bX_2 + \mathcal{N}(0,1), \quad X_t = X_2 + X_3 + \mathcal{N}(0,1).$$

depicted in Figure 1. This system has at least the following two invariant representations of $X = (X_1, X_2, X_3)$ if $a = -b$ for every intervention: (X_2, X_3) and (X_2). However, only the first continues to be invariant when $a \neq -b$. In other words, given an arbitrary number of training environments, each of which satisfies $a = -b$, we might decide that (X_2) is an invariant representation. However, this fails to be invariant once we include a test intervention for which $a \neq -b$. This happens because the joint distribution is not faithful to the DAG provided. In short, this means that the conditional independencies indicated by the DAG are not the only ones found in the data. In our case, the effect of X_1 on X_3 through the blue path and the red path exactly cancel in all training interventions.

Example 3.2 (Degenerate Interventions)**.** *Consider the degenerate situation in which each of the interventions is the same. The Empirical Risk Minimization (ERM) solution itself is invariant, however, the ERM solution comes with no generalization guarantees to other interventions. It is clear that some "diversity" condition among the training environments is necessary to get generalization guarantees.*

The key idea is that in such situations one might say that these issues (faithfulness violations as in Example 3.1 or degeneracies as in 3.2) will likely continue to manifest in future distributions. In particular, if we assume that both training and test intervention come from a single distribution over interventions, can we at least say that with high probability over a test intervention also drawn from $\mathcal{D}$, we will generalize to that intervention? More formally, we ask:

How big should Θ_{train} be so that for $\theta \sim \mathcal{D}$, $\mathcal{I}^{\epsilon}_{f_\circ}(\Theta_{train}) \in \mathcal{I}^{\epsilon}_{f_\circ}(\theta)$ with high probability?

In other words, we would like invariance on our training interventions to *certify* invariance over a future interventions with high probability. To further clarify the type of guarantee considered, we look at the following example.

Example 3.3. *Consider the DAG given in Figure (2). Consider the following distribution over interventions: We set $X_1 = 1$. Independently and with probability $1/2$ for each node that is not X_1 or X_t, we assign its value to be 0. Else we set its value equal to its parent. The only invariant representations in this case are (X_n) and (X_1, X_n). Now consider any training set of m interventions. If $m \ll 2^n$, we expect to never see the intervention in which none of the edges are disconnected (i.e., an environment with no intervention on any of $X_2, X_3, .., , X_{n-1}$), in which case (X_1) also appears to be invariant over every observed intervention. However, it fails to be invariant on the disconnected intervention described above, which occurs with an extremely low probability of $1/2^{n-1}$.*

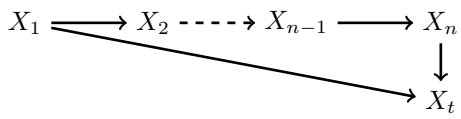

Figure 2. DAG for intervention in Example 3.3

Our result instead gives a probabilistic guarantee that says that with a certain probability over the randomness from which we are drawing interventions, any representation that appears to be invariant over the training set will also be invariant when the set is expanded to include another i.i.d. intervention (the test intervention). As demonstrated in this example, we may need $\Theta(2^n)$ interventions to get deterministic guarantees while our poly(n) result highlights the benefit of getting a probabilistic guarantee in this setting.

3.2. PAC Formulation

In this section, we show that the question of generalization of $\epsilon-$invariant representations introduced in the previous section can be rewritten as a question of PAC generalization.

3.2.1. VC DIMENSION AND PAC GENERALIZATION

Let $X \subset \mathbb{R}^d$ be a subset of points. A function class $\mathcal{G}$ is said to *shatter* X if every binary assignment to X is realized by some element of $\mathcal{G}$, that is, for any $\sigma : X \to \{\pm 1\}$, we have that there is some $f \in \mathcal{G}$ such that $f(x) = \sigma(x)$ for all $x \in X$. The VC dimension of a function class is defined as the size of the largest set that is shattered by it. The following result from PAC learning theory allows us to determine how many interventions are needed to generalize.

Lemma 3.1 ((Shalev-Shwartz & Ben-David, 2014))**.** *Consider a class $\mathcal{G}$ of functions from $\mathcal{X}$ to $\{\pm 1\}$ of VC dimension D, and a distribution $\tilde{\mathcal{D}}$ over $\mathcal{X} \times \{\pm 1\}$. In the realizable setting, when there exists $f \in \mathcal{G}$ such that $\mathbb{P}_{(X,Y)\sim\tilde{\mathcal{D}}}[f(X) \neq Y] = 0$, given $m = O(\frac{D + \log\frac{1}{\delta_1}}{\delta_2})$ samples $(x_i, y_i)_{i=1}^m$ and $\tilde{f}$ such that $\tilde{f}(x_i) = y_i$ for all i, with probability at least $1 - \delta_1$, $\tilde{f}$ satisfies $\mathbb{P}_{(X,Y)\sim\tilde{\mathcal{D}}}[\tilde{f}(X) \neq Y] < \delta_2$.*

We will need the following known VC dimension.

Lemma 3.2 ((Shalev-Shwartz & Ben-David, 2014)). *The VC dimension of halfspaces in $\mathbb{R}^D$, that is, the function class $\mathcal{G} = \left\{\mathbb{1}\{a^\top x < b\} : a \in \mathbb{R}^D, b \in \mathbb{R}\right\}$ is $D+1$.*

3.2.2. PAC Invariance

We will use PAC learning theory to study generalization for representations. For that we will rephrase the invariance problem to that of the generalization of binary classifiers. The proof is deferred to the Appendix, Lemma 8.7.

Lemma 3.3. *There is a function class $\mathcal{G}$ mapping interventions $\theta \in \Theta$ to $\{0,1\}$, such that*

$$\Phi \in \mathcal{I}^{\epsilon}_{f_\circ}(\Theta_{train}) \iff \exists g^{\epsilon}_{\Phi} \in \mathcal{G}$$
$$\text{such that } g^{\epsilon}_{\Phi}(\theta) = 1 \; \forall \theta \in \Theta_{train}.$$

In summary, if we can bound the VC dimension of $\mathcal{G}$, we can specify interventional complexity guarantees. The proof is deferred to the Appendix, Corollary 8.8

Corollary 3.4. *If the VC dimension of $\mathcal{G}$ is bounded by D, then given at least $O(\frac{D+\log\frac{1}{\delta}}{\delta'})$ training interventions, if $\mathcal{I}^{\epsilon}(\Theta) \neq \{\}$, we have with probability $1-\delta$ over the set of training interventions,*

$$\mathbb{P}_{\theta\sim\mathcal{D}}(\mathcal{I}^{\epsilon}_{f_\circ}(\Theta_{train}) \not\subset \mathcal{I}^{\epsilon}_{f_\circ}(\theta)) \leq \delta'.$$

Finally, there is at least one truly invariant representation, as demonstrated in the following Lemma.

Lemma 3.5. *There exists an invariant representation.*

Proof. The representation $\Phi = \text{diag}(\overline{\beta}_t)$ is invariant. That is, the diagonal matrix with β_t on the diagonal in indices corresponding to $\texttt{Pa}_t$. □

Remark 3.6. *(Fixing a head is important for PAC learning)* Considering a fixed head $f_\circ$ allows us to certify invariance locally, that is, looking at only a single intervention at a time. This is in contrast with the previous characterization, Equation (4), in which we simply ask for the best head on top of a representation to be the same for all interventions - a condition that we can only check for by considering all interventions simultaneously.

3.3. The VC dimension of certain classes of interventional distributions

Further, we show that for atomic interventions on some fixed set of k nodes, $O(k^4)$ training interventions suffices. For soft interventions on k nodes, assuming that each node has in-degree bounded by d, we show that $O(d^{4k})$ interventions suffices. This captures the intuition that if there is only some sparse, local change in the SEM, it should only take a number of interventions that scales in the sparsity to capture the invariances it leaves.

In the next section, we will show how to extend these results to the finite sample setting.

3.3.1. General Interventions

For this section we approach the task of bounding the VC dimension from the perspective of the complexity of the representation space. The class of interventions we consider are any interventions that lead to joint distributions over the variables that leave fixed the conditional distribution for the target variable given its parents.

Theorem 3.7. *Suppose that we are given $O(\frac{n^4+\log\frac{1}{\delta}}{\delta'})$ interventions drawn independently from distribution $\mathcal{D}$ over the intervention index set Θ. Then with probability at least $1-\delta$ over the randomness in Θ_{train}, the following statement holds:*

$$\mathbb{P}_{\theta\sim\mathcal{D}}(\mathcal{I}^{\epsilon}_{f_\circ}(\Theta_{train}) \notin \mathcal{I}^{\epsilon}_{f_\circ}(\theta)) \leq \delta'.$$

In other words, an ϵ- approximate invariant representation on the training environments generalizes with high probability to the test environment.

Connections to Causal Bayesian Networks: Invariance Principle or (less popularly known as) the modularity condition can be equivalently used to define causal Bayesian Networks (Bareinboim et al., 2012) - the central object in Pearlian Causal Models. With respect to all possible interventions wherein variable y has not been intervened on, only the representation that involves the true causal parents of y is invariant. In other words, the property of invariance can be taken to be the signature of causality. An open question in the setting has been the following: In how many interventional environments does this invariance property need to hold before it generalizes to most unseen environments. With respect to a random sampling on interventional environments for linear SEMs, our result answers this via high probability generalization guarantees without any faithfulness assumptions.

Structure Learning and Faithfulness: Invariance testing between a given set of interventional environments has been used to constrain the space of causal models (Yang et al., 2018) in the literature. However, the learning algorithm that synthesizes various invariances does require some notion of interventional faithfulness, i.e. observed invariances imply topological constraints on the true causal graph. In contrast, we do not make any such assumptions about faithfulness.

Comparison to Regularity Conditions in Invariant Prediction: Recently, invariant prediction (Arjovsky et al., 2019) has been used for out of distribution generalization. However, exact invariance holding deterministically in *all* unseen distributions from a family of linear SEMs was de-

sired. This required some general position conditions on the population covariances of the training interventional environments. The number of environments despite these additional conditions, required was $O(n^2)$. Our result on the interventional complexity is weaker but holds when we need no such assumptions on covariances and it holds under random sampling for approximate invariances. Similar technical general position conditions were also required in (Ahuja et al., 2020) for generalization.

In another recent work (Kamath et al., 2021), if the mapping between the intervention index set Θ and the observed training distributions on X is analytic, then only two training environments (almost surely) suffice for exact invariance in the population setting (i.e., infinite samples) for the entire index set. Here, we analyze approximate invariance without any restrictions on the mapping being analytic. For instance, their result does not apply in Example 3.3, where we need two specific interventions (one in which each node is assigned the value of the parent, and any other) to certify invariance. In fact, any set of interventions needs to have the specific intervention highlighted in the example to certify invariance, and this happens with very low probability.

3.3.2. Atomic interventions

Here we consider the instance in which our family of interventions consists of atomic interventions on a total of k nodes $\texttt{At}$. An atomic intervention at node i replaces the generative equation $X_i = \gamma_i(e)^\top \texttt{Pa}_i + \eta_i$ by the *assignment* $X_i = a_i$ for some scalar a_i. Note that we can interpret this as zeroing out the corresponding entries of B and changing the noise variable to be a constant a_i.

Theorem 3.8. *Given $|\Theta_{train}| = O(\frac{k^4+\log\frac{1}{\delta}}{\delta'})$ interventions drawn independently from a distribution $\mathcal{D}$ over all atomic interventions on some fixed set $\texttt{At}$ of k nodes that with probability at least $1-\delta$ over the randomness in Θ_{train}, with probability at least $1-\delta'$ over the randomness in D, we have for $\theta \sim \mathcal{D}$, $\mathcal{I}^\epsilon(\Theta_{train}) \subseteq \mathcal{I}^\epsilon(\theta)$*

3.3.3. Soft Interventions

Soft interventions are those in which the *weights* of the SEM are modified while the underlying causal structure remains the same. That is, a soft intervention at node v_i is equivalent to replacing the equation $X_i = (\beta_i^{\theta_\circ})^\top X + \eta_i$ with $X_i = (\beta_i^{\theta})^\top X + \eta_i$. Note that this is equivalent also to changing the ith row of B_θ.

Theorem 3.9. *Given $O(\frac{d^{4k}+\log\frac{1}{\delta}}{\delta'})$ interventions drawn independently from a distribution $\mathcal{D}$ over all soft interventions on some fixed set $\texttt{At}$ of k nodes such that each intervened node has in-degree at most d, we have that any $\epsilon-$approximately invariant representation Φ satisfies that with probability at least $1-\delta$ over the randomness in Θ_{train}, with probability at least $1-\delta'$ over the randomness in D, we have for $\theta \sim \mathcal{D}$,$\mathcal{I}^\epsilon(\Theta_{train}) = \mathcal{I}^\epsilon(\theta)$*

3.3.4. Indirect Observations

Finally, we consider the setting in which there is an underlying SEM which we observe after a linear transformation. In particular, we consider the setting in which we observe $(S_{-t}(X), S_t(X))$ for some linear transformations S_{-t} and S_t. Here, X is generated as before, according to 1. Note that the target variable is now S_t. We assume for realizability (in particular, for Lemma 3.5) the following:

Assumption 3.10. There is some linear representation Φ such that $\Phi S_{-t}(X) = \texttt{Pa}_t$.

Theorem 3.11. *Given $O(\frac{n^8+\log\frac{1}{\delta}}{\delta'})$ interventions drawn independently from a distribution $\mathcal{D}$ interventions, we have that any $\epsilon-$approximately invariant representation Φ satisfies that with probability at least $1-\delta$ over the randomness in Θ_{train}, with probability at least $1-\delta'$ over the randomness in D, we have for $\theta \sim \mathcal{D}$,$\mathcal{I}^\epsilon(\Theta_{train}) = \mathcal{I}^\epsilon(\theta)$*

Remark 3.12. Note that Theorem 3.11 encompasses models that have latent variables, so long as the causal parents are observed. Note that the existence of latent variables typically precludes any guarantees for algorithms that rely on CI-testing. One could consider this to be a factor in favor of using IRM-based solutions for OOD generalization.

Remark 3.13. Note that we only need Assumption 3.10 to be satisfied, and not that the functions S_{-t} and S_t themselves are linear. This is weaker than a linearity assumption, since we could have non-linearity in the spurious components.

4. Finite samples

Note that the above discussion was about population statistics. In practice, we only see finitely many samples from each interventional distribution. We use an estimator similar to that of (Ahuja et al., 2020) to get finite sample guarantees. We need the following assumption for scale.

Assumption 4.1. The following bounds hold, $\|\Phi\|_2 \leq 1$, $\|X\|_\infty < L$.

Lemma 4.2. *Given $\frac{4nL^2}{\epsilon^2}\left(\log\frac{2n}{\delta} + n^2\log(1+\frac{8n^{3/2}}{\epsilon})\right)$ samples from Δ_θ, we have that with probability $1-\delta$ over the samples drawn in each interventional distribution*[1]

$$\hat{\mathcal{I}}^\epsilon_{f_\circ}(\theta) \subseteq \mathcal{I}^{2\epsilon}_{f_\circ}(\theta).$$

Proof Sketch. We show this using the standard concentration arguments for a single fixed representation Φ. We take a union bound over an $\epsilon-$net of $\mathcal{R}^\theta_\Phi$ to get a uniform bound over all representations. See Lemma 10.3 for a proof. □

[1] In contrast to claims over the randomness $\mathcal{D}$ over Θ from previous theorems, this is a statement about the randomness Δ_θ for X.

We can now take a union bound over all $\theta \in \Theta_{\text{train}}$. In conclusion, we have shown that with high probability over the randomness in the samples we see in each of our interventional distributions, $\epsilon-$approximate invariance over the training data certifies $\epsilon-$approximate invariance over a $1-\delta'$ fraction of out full intervention set. The proof is defered to the Appendix Theorem 10.4

Theorem 4.3. *Given*

$$m = \begin{cases} O(\frac{k^4+\log\frac{1}{\delta}}{\delta'}) & \Theta \text{ is } k \text{ nodes, hard interventions} \\ O(\frac{d^{4k}+\log\frac{1}{\delta}}{\delta'}) & \Theta \text{ is } k \text{ nodes, soft interventions} \\ O(\frac{n^4+\log\frac{1}{\delta}}{\delta'}) & \Theta \text{ any interventions} \end{cases}$$

interventional datasets, and, in each dataset, at least $\frac{4nL^2}{\epsilon^2}\left(\log\frac{2nm}{\delta} + n^2\log(1+\frac{8n^{3/2}}{\epsilon})\right)$ *samples, we have that with probability* $1-\delta$*, with probability* $1-\delta'$ *for* $\theta \sim \mathcal{D}$*,*

$$\hat{\mathcal{I}}^{\epsilon}_{f_\circ}(\Theta_{\textit{train}}) \subseteq \mathcal{I}^{2\epsilon}_{f_\circ}(\theta).$$

5. Empirical Study

In this section we highlight the above results empirically. The experiment is described in detail in Appendix 11.

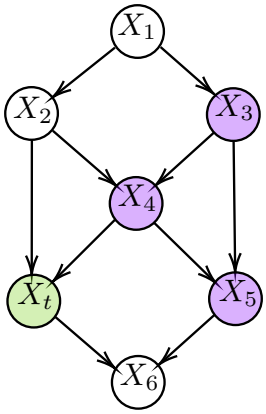

Figure 3. Linear SEM for Section 5. Nodes $\{v_3, v_4, v_5\}$ (purple) are used as intervention sites for $\mathcal{D}_{\text{hard}}$ and $\mathcal{D}_{\text{soft}}$. Each edge represents a weight of 1. Node X_t (green) is taken to be the target node.

We consider the 7-node linear SEM in Figure 3. The target variable is taken to be X_t. Each edge weight is set to 1 for the observational distribution. We consider an interventional distribution $\mathcal{D}_{\text{hard}}$ with support over the set of hard interventions on nodes $\{v_3, v_4, v_5\}$. Recall that a hard intervention consists of assigning a value to a node. We draw m interventional distributions from $\mathcal{D}_{\text{hard}}$ as our training interventions, and draw a sample consisting of $N = 200000$ datapoints from each distribution. The corresponding plots with $N \in \{10000, 15000, 30000, 45000\}$ are available in Appendix 11. We use a notion of inter-dataset variance of the least squares solutions to construct a set $s_{\text{hard}} = \{S_1, S_2, \cdots,\}$ where each S_i is an approximately invariant representation derived from the training datasets. Rather than using variance, we can use the norm of the gradient; please see Appendix 11 for details. We then generate an additional $m^{\text{test}} = 100$ test interventions from $\mathcal{D}_{\text{hard}}$, and generate datasets of size $N = 200000$ from each of them. We use the same inter-dataset variance to count the percentage of subsets in s_{hard} that continue to have low variation between the least squares solutions on the new *test distributions* and the *average least squares solution on the training distributions*. We repeat the above for soft interventions drawn from $\mathcal{D}_{\text{soft}}$. We consider soft interventions that modify the weights of the edges into v_3, v_4, and v_5. Please see Appendix 11 for exact details about how the hard and soft interventional ensembles are defined.

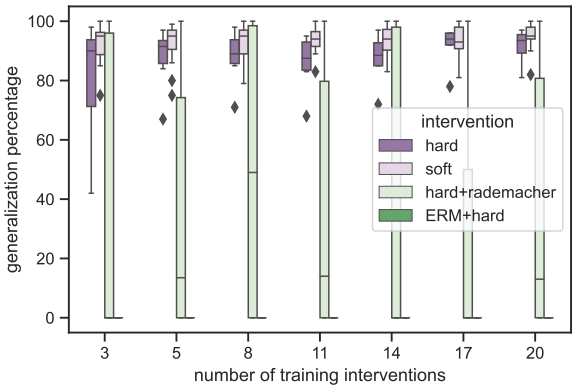

Figure 4. Generalization of approximately invariant representations under different combinations of train and test interventional distributions for the experiment described in Section 5. The ERM solution almost always fails to generalize, and the corresponding box plot is at '0'.

We also consider a different interventional distribution $\mathcal{D}_{\text{rad}}$ to test the generalization of representations that are approximately invariant given training distributions drawn from $\mathcal{D}_{\text{hard}}$. These interventions are generated by randomly flipping the signs of the edge weights in the original SEM.

Finally, to confirm that there is indeed variation across the least squares solutions in the different environments, we also plot the percentage of test distributions to which the ERM solution on the training distributions is similar to ERM solution of the test distributions (this percentage is almost always 0; meaning that the ERM solution always fails to generalize). We repeat this for m varying beteen 3 and 20. The results are plotted in Figure 4.

The code is available at `https://github.com/`

advaitparulekar/PAC_IRM

Interpretation: While not every subset in s_{hard} and s_{soft} generalizes to every test dataset, we see that *most* subsets generalize to *large percentage* of test distributions. The percentage that generalize when the train and test interventions are identically distributed exceeds the percentage that generalizes to datasets that come from the ("adversarial") rademacher interventional family. This phenomenon is captured in our bounds, which describe when *most* approximately invariant representations will continue to be approximately invariant on *most* test distributions.

Acknowledgements

This research is supported in part by NSF Grants 2019844, 2107037 and 2112471, ONR Grant N00014-19-1-2566, the Machine Learning Lab (MLL) at UT Austin, and the Wireless Networking and Communications Group (WNCG) Industrial Affiliates Program.

References

Ahuja, K., Wang, J., Dhurandhar, A., Shanmugam, K., and Varshney, K. R. Empirical or invariant risk minimization? a sample complexity perspective. In *International Conference on Learning Representations*, 2020.

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

Bareinboim, E., Brito, C., and Pearl, J. Local characterizations of causal bayesian networks. In *Graph Structures for Knowledge Representation and Reasoning*, pp. 1–17. Springer, 2012.

Beery, S., Horn, G. V., and Perona, P. Recognition in terra incognita. In *ECCV*, 2018.

Ben-David, S., Blitzer, J., Crammer, K., Pereira, F., et al. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19:137, 2007.

Brouillard, P., Lachapelle, S., Lacoste, A., Lacoste-Julien, S., and Drouin, A. Differentiable causal discovery from interventional data. *arXiv preprint arXiv:2007.01754*, 2020.

Chen, Y., Rosenfeld, E., Sellke, M., Ma, T., and Risteski, A. Iterative feature matching: Toward provable domain generalization with logarithmic environments. In *Advances in Neural Information Processing Systems*, 2022.

Chen, Y., Zhou, K., Bian, Y., Xie, B., Wu, B., Zhang, Y., KAILI, M., Yang, H., Zhao, P., Han, B., et al. Pareto invariant risk minimization: Towards mitigating the optimization dilemma in out-of-distribution generalization. In *The Eleventh International Conference on Learning Representations*, 2023.

Chickering, M. Statistically efficient greedy equivalence search. In *Conference on Uncertainty in Artificial Intelligence*, pp. 241–249. PMLR, 2020.

Duchi, J. C., Glynn, P. W., and Namkoong, H. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 2021.

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.

Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. In *International Conference on Learning Representations*, 2020.

Hauser, A. and Bühlmann, P. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *The Journal of Machine Learning Research*, 13(1):2409–2464, 2012.

Kamath, P., Tangella, A., Sutherland, D., and Srebro, N. Does invariant risk minimization capture invariance? In *International Conference on Artificial Intelligence and Statistics*, pp. 4069–4077. PMLR, 2021.

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. *arXiv preprint arXiv:1707.06422*, 2017.

Mansour, Y., Mohri, M., and Rostamizadeh, A. Domain adaptation with multiple sources. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L. (eds.), *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008.

Mansour, Y., Mohri, M., Ro, J., Suresh, A. T., and Wu, K. A theory of multiple-source adaptation with limited target labeled data. In *International Conference on Artificial Intelligence and Statistics*, pp. 2332–2340. PMLR, 2021.

Muandet, K., Balduzzi, D., and Schölkopf, B. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pp. 10–18. PMLR, 2013.

Pearl, J. *Causality*. Cambridge University Press, 2 edition, 2009. doi: 10.1017/CBO9780511803161.

Peters, J., Bühlmann, P., and Meinshausen, N. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pp. 947–1012, 2016.

Rosenfeld, E., Ravikumar, P. K., and Risteski, A. The risks of invariant risk minimization. In *International Conference on Learning Representations*, 2020.

Rothenhäusler, D., Meinshausen, N., Bühlmann, P., and Peters, J. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society Series B*, 83(2):215–246, April 2021. doi: 10.1111/rssb.12398.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Shalev-Shwartz, S. and Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014. doi: 10.1017/CBO9781107298019.

Solus, L., Wang, Y., and Uhler, C. Consistency guarantees for greedy permutation-based causal inference algorithms. *Biometrika*, 108(4):795–814, 2021.

Spirtes, P., Glymour, C., and Scheines, R. *Causation, Prediction, and Search*. MIT Press, 2000.

Squires, C., Wang, Y., and Uhler, C. Permutation-based causal structure learning with unknown intervention targets. In *Conference on Uncertainty in Artificial Intelligence*, pp. 1039–1048. PMLR, 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.

Sullivant, S., Talaska, K., and Draisma, J. Trek separation for gaussian graphical models. *The Annals of Statistics*, 38(3):1665–1685, 2010.

Uhler, C., Raskutti, G., Bühlmann, P., and Yu, B. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, pp. 436–463, 2013.

Verma, T. and Pearl, J. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, UAI '90, pp. 255–270, USA, 1990. Elsevier Science Inc. ISBN 0444892648.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices, 2011.

Wang, H., Si, H., Li, B., and Zhao, H. Provable domain generalization via invariant-feature subspace recovery. In *International Conference on Machine Learning*, pp. 23018–23033. PMLR, 2022.

Yang, K., Katcoff, A., and Uhler, C. Characterizing and learning equivalence classes of causal dags under interventions. In *International Conference on Machine Learning*, pp. 5541–5550. PMLR, 2018.

Zhao, H., Des Combes, R. T., Zhang, K., and Gordon, G. On learning invariant representations for domain adaptation. In *International Conference on Machine Learning*, pp. 7523–7532. PMLR, 2019.

X_i	random variable in SEM, see Equation 1
β_i	linear weight vector, see Equation 1
$\texttt{Pa}_i$	vector of random variables corresponding to parents of X_i, see 1
Θ	set of possible interventions
Θ_{train}	set of training interventions
$\mathcal{D}$	distribution from which training and test distributions are sampled
Δ_θ	joint distribution over variables induced by intervention θ
$\mathcal{I}(\Theta')$	set of invariant representations with respect to Θ'
$\mathcal{I}^\epsilon(\Theta')$	set of approximately invariant representations with respect to Θ'
$\hat{\mathcal{I}}^\epsilon(\Theta')$	set of approximately invariant representations with respect to Θ' (finite samples)
Φ	Diagonal matrix encoding a feature selector
d	in-degree of nodes in Section 9.3
k	number of nodes that have interventions in Section 9.2 and 9.3
δ	randomness in samples in the interventional datasets
δ'	randomness in interventions drawn from $\mathcal{D}$

Table 1. Notation table

6. Notation

For any quantity that depends on population statistics, A, we use $\hat{A}$ to denote the corresponding sample estimate. For a vector $a \in \mathbb{R}^n$, we use $a^{(\dot{k})}$ to denote the vector consisting of all monomials of degree *less than or equal to* k with variables in a. Thus, we have $a^{(\dot{2})} = (1, a_1, a_2, \cdots, a_n, a_1^2, a_1a_2, \cdots, a_n^2)^\top$. We use $a^{(k)}$ (that is, with no dot) to denote the vector consisting of all monomials of *exactly* degree k with variables in a, so $a^{(2)} = (a_1^2, a_1a_2, a_1a_3, \cdots a_n^2)^\top$. For a matrix A, we use $r(A)$ to denote the vector formed by flattening the matrix, so $r(A) = (A_{11}, A_{12}, A_{13}, \cdots, A_{n-1,n}, A_{nn})^\top$. Please see Table 1 for a summary of notation used in the paper.

7. Auxillary proofs

7.1. General Results

Lemma 7.1. *Consider a sequence of i.i.d. random vectors $V_1, V_2, \cdots, V_m \in \mathbb{R}^d$ with mean μ such that $\|V_i\|_\infty < L$ always. Then for $m > \frac{dL^2}{\epsilon^2} \log \frac{d}{\delta}$, we have*

$$\mathbb{P}[\|\frac{1}{m}\sum_{i=1}^m V_i - \mu\| < \epsilon] > 1 - \delta.$$

Proof. By a standard Hoeffding bound, with probability $1-\delta$, each entry of $\frac{1}{m}\sum_{i=1}^m V_i$ will be within $\frac{\epsilon}{\sqrt{d}}$ of its mean for $m \geq \frac{dL^2}{\epsilon^2} \log \frac{d}{\delta}$. Then the squared norm of the error $\frac{1}{m}\sum_{i=1}^m V_i - \mu$ will be less than ϵ. □

Lemma 7.2 ((Vershynin, 2011))**.** *The covering number of the Euclidean ball of radius R is $(1 + \frac{2R}{\epsilon})^d$.*

Lemma 7.3. *Let $A \in \mathbb{R}^{n\times n}$ be a matrix and $x \in \mathbb{R}^n$ be a column vector, then we have$\|Ax\|_2^2 = r(A^\top A)x^{(2)}$.*

Proof. This follows by writing the norm squared as a quadratic form. □

Lemma 7.4. *Consider a diagonal matrix $\Phi \in \mathbb{R}^{n\times n}$. Then there exists a fixed matrix $V \in \mathbb{R}^{n^2\times n}$ such that $r(\Phi) = V diag(\Phi)$.*

Proof. To construct V, we take the $n \times n$ identity matrix and insert with $n-1$ rows of all 0s between every row of the

identity matrix obtain $V \in \mathbb{R}^{n^2 \times n}$ which looks as follows:

$$V = \begin{pmatrix} 1 & 0 & 0 \cdots & 0 \\ 0 & 0 & 0 \cdots & 0 \\ \vdots & & & \\ 0 & 1 & 0 \cdots & 0 \\ 0 & 0 & 0 \cdots & 0 \\ \vdots & & & \\ 0 & 0 & 0 \cdots & 0 \\ 0 & 0 & 0 \cdots & 1 \end{pmatrix}$$

. □

Lemma 7.5 ((Sullivant et al., 2010), (Uhler et al., 2013))**.** *For the class of linear SEMs defined in Equation 1, we have the following*

- $\Sigma_\theta = J(1 - B_\theta)^{-1}\Xi_\theta \left((1 - B_\theta)^{-1}\right)^\top J^\top$ *where* $\Xi_\theta = \mathbb{E}_\theta[\eta\eta^\top]$ *and* J *is identity concatenated with one column of zeros at the index corresponding to* v_t*.*
- $(I - B_\theta)^{-1} = \sum_{k=0}^{n-1} B_\theta^k$
- $\Sigma_\theta = \sum_{k=0}^{2n-2} \sum_{\substack{r+s=k \\ r,s<n-1}} B^r \Xi_\theta (B^\top)^s$.

Proof. Note that the SEM equations can all be combined into a single vector equation of the form

$$X = B_\theta X + \eta \implies X = (I - B_\theta)^{-1}\eta$$

for some lower triangular matrix $B_\theta \in \mathbb{R}^{n \times n}$, and a noise vector η. We can then write the covariance of the features X_{-t}, Σ_θ, in terms of B_θ. Let $J \in \mathbb{R}^{n \times n+1}$ denote the matrix that is identity concatenated with an additional column consisting of all zeros in the index corresponding to v_t, such that it selects the submatrix of $\mathbb{E}_\theta[XX^\top]$ corresponding to $\Sigma_\theta = \mathbb{E}_\theta[X_{-t}X_{-t}^\top]$

$$\begin{aligned} \Sigma_\theta &:= J\,\mathbb{E}_\theta[XX^\top]J^\top \\ &= J(I - B_\theta)^{-1}\,\mathbb{E}_\theta[\eta\eta^\top]\left((I - B_\theta)^{-1}\right)^\top J^\top \end{aligned}$$

Since $\|B_\theta\| < 1$, we have $(I - B_\theta)^{-1} = \sum_{k=0}^{\infty} B_\theta^k$. Because B_θ is lower triangular, $B_\theta^k = 0$ for $k \geq n$. So $(I - B_\theta)^{-1} = \sum_{k=0}^{n-1} B_\theta^k$. Similarly $((I - B_\theta)^{-1})^\top = \sum_{k=0}^{n-1} (B_\theta^\top)^k$. The result follows by considering every cross term in the product of the sums.

$$\Sigma_\theta = \sum_{r=0}^{n-1} (B_\theta)^r \Xi_\theta \sum_{s=0}^{n-1} (B_\theta^\top)^s = \sum_{k=0}^{2n-2} \sum_{\substack{r+s=k \\ r,s<n-1}} B^r \Xi_\theta (B^\top)^s.$$

□

8. Invariant Representations

Consider the SEM model of Section 2.

$$\begin{aligned} X_i &= \left(\beta_i^\theta\right)^\top \mathtt{Pa}_i + \eta_i \\ X_t &= \beta^\top \mathtt{Pa}_t + \eta_t \end{aligned}$$

We see that a representation that projects onto the parents is indeed invariant, $\mathbb{E}_\theta[X_t|\mathtt{Pa}_t] = \beta^\top \mathtt{Pa}_t$ and the RHS has no dependence on θ. If X is independent of η_t, actually more can be said: the identity representation itself is invariant!

Lemma 8.1. *If $X \perp\!\!\!\perp \eta_t$, then $I \in \mathcal{I}(\Theta)$.*

Proof. We begin with Equation (8)

$$\nabla \mathcal{R}^{\theta}_{\Phi}(f_\circ) = \Phi^\top \Sigma_\theta \Phi f_\circ - \Phi^\top \mathbb{E}_\theta[X_\theta X_t^\top]$$

Plugging in $\Phi = I$, we have

$$\begin{aligned}
\nabla \mathcal{R}^{\theta}_{I}(f_\circ) &= \Sigma_\theta f_\circ - \mathbb{E}_\theta[X_\theta X_t^\top] \\
&= \Sigma_\theta f_\circ - \mathbb{E}_\theta[X\left(\mathtt{Pa}_t^\top \beta + \eta_t\right)] \\
&= \Sigma_\theta f_\circ - \mathbb{E}_\theta[X\left(X^\top \overline{\beta} + \eta_t\right)] \\
&= \Sigma_\theta f_\circ - \Sigma_\theta \beta + \mathbb{E}_\theta[X\eta_t] \\
&= \Sigma_\theta f_\circ - \Sigma_\theta \beta + \mathbb{E}_\theta[X]\,\mathbb{E}_\theta[\eta_t] \\
&= \Sigma_\theta (f_\circ - \beta)
\end{aligned}$$

So the loss is minimized at the common $f_\circ = \beta$ (that is, independent of intervention). □

Remark 8.2. In the absence of anti-causal variables (variables other than the target variable X_t that are "downstream" from X_t), the condition of Lemma 8.1 is satisfied, and we have the convenient result that the ERM solutions are actually already invariant.

8.1. Representations

In this section, we will look more closely at the set of representations.

We show that when the space of representations is all matrix transformations, then the invariant regressor can be taken to be $f_\circ = (1, 0 \dots 0)$.

Lemma 8.3. *Let the set of candidate representations be from $\mathbb{R}^{n\times n}$. Let the set of candidate invariant regressors be $\mathbb{R}^{n\times 1}$. Any approximately invariant solution $(\Phi, f) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times 1}$, corresponds to another invariant solution $(\Phi', f_\circ)$ where $\Phi' = \Phi A^{-1}$ for some invertible A, and $f_\circ = (1, 0, \cdots, 0)^T$.*

Proof. We will construct A as follows. For the first column, take $A_{:,0} = f$. Take the subspace orthogonal to f and consider a basis $b_2 \dots b_n$ that spans this. Then use Gram-Schmidt on b_i's to complete A, iteratively using columns orthogonal to previous columns. This ensures that $Af_\circ = f$, while also keeping A invertible. Furthermore, we have that $\|\nabla \mathcal{R}^{\theta}_{\Phi'}(f_\circ)\| \le \epsilon \iff \|\nabla \mathcal{R}^{\theta}_{\Phi' A^{-1}}(Af_\circ)\| \le \epsilon \iff \|\nabla \mathcal{R}^{\theta}_{\Phi}(f)\| \le \epsilon$ □

We show that when the space of representations is all diagonal matrices, then the invariant regressor can be taken to be $f_\circ = (1, 1 \dots 1)$.

Lemma 8.4. *Let the set of candidate representations be diagonal matrices of order n denoted by $\mathcal{D}_n$. Let the set of candidate invariant regressors be $\mathbb{R}^{n\times 1}$ where the regressor has non-zero entries.Consider an approximately invariant solution $(\Phi, f) \in \mathcal{D}_n \times \mathbb{R}^{n\times 1} : f_i \neq 0$ where f_i is the i-th coordinate of vector f. This corresponds to another invariant solution $(\Phi', f_\circ)$ where $f_\circ = (1, 1 \dots 1)^T$*

Proof. Consider the vector $f^T\Phi$ that is an invariant solution. Since Φ acts as feature selector, without loss of generality $f_i \neq 0,\ \forall i$. Now, put the entries into a diagonal matrix denoted by $\mathrm{diag}(f^T\Phi) = \mathrm{diag}(f)\Phi$. Now let $\Phi' = \Phi\mathrm{diag}(f)$. Clearly, $\mathrm{diag}(f)f_\circ = f$ where $f_\circ = (1, 1 \dots 1)^T$. Therefore, for $A = \mathrm{diag}(f)$, we have $\|\nabla \mathcal{R}^{\theta}_{\Phi'}(f_\circ)\| \le \epsilon \iff \|\nabla \mathcal{R}^{\theta}_{\Phi' A^{-1}}(Af_\circ)\| \le \epsilon \iff \|\nabla \mathcal{R}^{\theta}_{\Phi}(f)\| \le \epsilon$ □

In Sections 9.2,9.1 and 9.3 where diagonal representations (used as feature selectors) are used, $f_\circ = (1, 1 \dots 1)$. In Sections 9.4 where general representations are used, then $f_\circ = (1, 0 \dots 0)$.

Since the least squares objective is strongly convex, using gradients instead of loss values does not change the set of invariant representations

Lemma 8.5.

$$f_\circ = \arg\min \mathcal{R}^\theta_\Phi(f) \quad \forall \theta \in \Theta \iff \nabla \mathcal{R}^\theta_\Phi(f_\circ) = 0 \quad \forall \theta \in \Theta$$

Proof. This follows from strong convexity of the loss function.

□

Lemma 8.6.

$$\|\nabla \mathcal{R}^\theta_\Phi(f_\circ)\| = 2\|\Phi^\top \Sigma_\theta \Phi f_\circ - \Phi^\top(\Sigma_\theta \beta - J(1-B_\theta)^{-1} e_t)\|$$

Proof.

$$\begin{aligned}
&\|\nabla \mathbb{E}[\left(f^\top \Phi^\top X_{-t} - X_t\right)^2 \big|_{f=f_\circ}\|_2 \\
&= 2\|\mathbb{E}[\Phi^\top X_{-t}\left(X_{-t}^\top \Phi f_\circ - X_t\right)]\| \qquad (6)\\
&= 2\|\Phi^\top \Sigma_\theta \Phi f_\circ - \Phi^\top \mathbb{E}_\theta[X_{-t} X_t^\top]\| \\
&= 2\|\Phi^\top \Sigma_\theta (\Phi f_\circ - \overline{\beta}_t) - \Phi^\top \mathbb{E}_\theta[X_{-t}^\top \eta_t]\| \\
&= 2\|\Phi^\top \Sigma_\theta (\Phi f_\circ - \beta) - \Phi^\top \mathbb{E}_\theta[J(1-B_\theta)^{-1}\eta\eta_t]\| \\
&= 2\|\Phi^\top \Sigma_\theta (\Phi f_\circ - \beta) - \Phi^\top J(1-B_\theta)^{-1}\mathbb{E}_\theta[\eta\eta_t]\| \\
&= 2\|\Phi^\top \Sigma_\theta \Phi f_\circ - \Phi^\top(\Sigma_\theta \beta - J(1-B_\theta)^{-1} e_t)\| \qquad (7)
\end{aligned}$$

□

Lemma 8.7. *There is a function class $\mathcal{G}$ mapping interventions $\theta \in \Theta$ to $\{0,1\}$, such that*

$$\Phi \in \mathcal{I}^\epsilon_{f_\circ}(\Theta_{train}) \iff f^\epsilon_\Phi(\theta) = 1 \; \forall \theta \in \Theta_{train}$$

Proof. By Lemmas 8.6 and 7.5, the invariance condition can be simplified as follows (for approriate J, e_t defined in the Lemma):

$$\|\nabla \mathbb{E}[\left(f^\top \Phi^\top X_{-t} - X_t\right)^2 \big|_{f_\circ}\|_2 = 2\|\Phi^\top \Sigma_\theta \Phi f_\circ - \Phi^\top(\Sigma_\theta \beta - J(1-B_\theta)^{-1} e_t)\| \le \epsilon. \qquad (8)$$

Our desired function class is $\mathcal{G} = \{f^\epsilon_\Phi(\theta)\}$ parameterized by Φ where

$$f^\epsilon_\Phi(\theta) := \mathbb{1}\{2\|\Phi^\top \Sigma_\theta \Phi f_\circ - \Phi^\top(\Sigma_\theta \beta - J(1-B_\theta)^{-1} e_t)\| \le \epsilon\},$$

is a function mapping an intervention to $\{0,1\}$ such that $\Phi \in \mathcal{I}^\epsilon_{f_\circ}(\theta) \iff f^\epsilon_\Phi(\theta) = 1$. □

Corollary 8.8. *If the VC dimension of $\mathcal{G}$ is bounded by d, then given at least $O(\frac{d+\log\frac{1}{\delta}}{\delta'})$ training interventions, if $\mathcal{I}^\epsilon(\Theta)$ is not empty, we have that with probability $1-\delta$ over the set of training interventions,*

$$\mathbb{P}_{\theta\sim\mathcal{D}}(\mathcal{I}^\epsilon_{f_\circ}(\Theta_{train}) \not\subset \mathcal{I}^\epsilon_{f_\circ}(\theta)) \le \delta'.$$

Proof. Let $\tilde{D}$ be a distribution over $\Theta \times \{\pm 1\}$ such that the marginal over Θ is $\mathcal{D}$, and the marginal over $\{\pm 1\}$ is the dirac-delta on the element '1' (i.e. it always takes value 1). We can now apply Lemma 3.1 to our function class $\mathcal{G} = \{g^\epsilon_\Phi(\theta)\}$ and distribution $\tilde{\mathcal{D}}$. By assumption, there exists a representation $\Phi \in \mathcal{I}^\epsilon_{f_\circ}(\Theta) = \bigcap_{\theta\in\Theta_{\text{train}}} \mathcal{I}^\epsilon(\theta)$, which means $f^\epsilon_\Phi(\theta) = 1$ for all $\theta \in \Theta$. Since $\Theta_{\text{train}} \subseteq \Theta$, we have $f^\epsilon_\Phi(\theta) = 1$ for all $\theta \in \Theta_{\text{train}}$. PAC learnability now states that given $O(\frac{d+\log\frac{1}{\delta}}{\delta'})$ interventions, if $\tilde{\Phi}$ is such that $\tilde{\Phi} \in g^\epsilon_\Phi(\Theta_{\text{train}})$ (so that it is a hypothesis that fits the training data exactly), we have that with probability at least $1-\delta$, $\mathbb{P}_{\theta\sim\mathcal{D}}[g^\epsilon_{\tilde{\Phi}}(\theta) = 1] > 1-\delta'$. □

9. Intervention Models

9.1. General Interventions

Lemma 9.1. *For some matrix U_θ, we have*

$$\Phi^\top \Sigma_\theta \Phi f_\circ - \Phi^\top (\Sigma_\theta \beta - J(1-B_\theta)^{-1} e_t) = U_\theta r(\Phi)^{(2)}.$$

Proof. We will see that both of these terms can be written as $U_\theta r(\Phi)^{(\dot{2})}$ for some (different) choices of U_θ. We will enumerate the coordinates of $r(\Phi)^{(\dot{2})}_{(ij),(kl)}$ as $\Phi_{ij}\Phi_{kl}$. Note that $r(\Phi)^{(\dot{2})}$ also contains the individual Φ_{ij}, and we will use $\cdot$ to indicate this extra alphabet. To clarify, some of the coordinates are then of the form $r(\Phi)^{(\dot{2})}_{(\cdot,(kl))} = \Phi_{kl}$. Now observe that

$$\Phi^\top \Sigma_\theta \Phi f_\circ = U^1_\theta r(\Phi)^{(\dot{2})}$$

where

$$(U^1_\theta)_{a,((ij),(kl))} = \begin{cases} \Sigma_{jk}(f_\circ)_l, & a = i \text{ and } (ij) \neq \cdot \text{ and } (kl) \neq \cdot \\ 0, & a \neq i \text{ or } (ij) = \cdot \text{ or } (kl) = \cdot \end{cases}$$

The second term is almost in the form we would like already

$$\Phi^\top (\Sigma_\theta \beta - J(1-B_\theta)^{-1} e_t) = U^2_\theta r(\Phi)^{(\dot{2})}$$

where

$$(U^2_\theta)_{a,((ij),(kl))} = \begin{cases} (\Sigma_\theta \beta - J(1-B)^{-1} e_t)_{ij} & (ij) \neq \cdot \text{ and } (kl) = \cdot \\ 0, & (ij) = \cdot \text{ or } (kl) \neq \cdot \end{cases}$$

Our result follows by setting $U_\theta = U^1_\theta + U^2_\theta$. □

Theorem 9.2. *Suppose that we are given $O(\frac{n^4 + \log\frac{1}{\delta}}{\delta'})$ interventions drawn independently from distribution $\mathcal{D}$ over the intervention index set Θ. Then with probability at least $1-\delta$ over the randomness in Θ_{train}, the following statement holds:*

$$\mathbb{P}_{\theta\sim\mathcal{D}}(\mathcal{I}^\epsilon_{f_\circ}(\Theta_{train}) \notin \mathcal{I}^\epsilon_{f_\circ}(\theta)) \leq \delta'.$$

In other words, an ϵ- approximate invariant representation on the training environments generalizes with high probability to the test environment.

Proof. We again begin with our expression for the gradient of the loss given by Lemma 8.6.

$$\nabla \mathcal{R}^\theta_\Phi(f_\circ) = \Phi^\top \Sigma_\theta \Phi f_\circ - \Phi^\top (\Sigma_\theta \beta - J(1-B_\theta)^{-1} e_t).$$

We show in Lemma 9.1 that we can write this as $\nabla \mathcal{R}^\theta_\Phi(f_\circ) = U_\theta r(\Phi)^{(\dot{2})}$ for some matrix U_θ. From Lemma 7.3 we can write the squared norm as $\|\nabla \mathcal{R}^\theta_\Phi(f_\circ)\|_2^2 = r(U_\theta^\top U_\theta)(r(\Phi)^{(\dot{2})})^{(2)}$. Finally, we can now write our function class as a class of halfspace classifiers.

$$g^\epsilon_\Phi(\theta) = 1 \iff r(U_\theta^\top U_\theta)(r(\Phi)^{(\dot{2})})^{(2)} \leq \epsilon.$$

Now consider the map $\Psi : \text{diag}(\Phi^{(\dot{4})}) \to r(\Phi)^{(\dot{2})(2)}$. Because Φ is diagonal, this is well defined. Each term in $r(\Phi)^{(\dot{2})(2)}$ contains at most 4 terms of Φ, and each of these is an element of $\Phi^{(\dot{4})}$. Also, since the relation between entries of $r(\Phi)^{(\dot{2})(2)}$ and entries of $\text{diag}(\Phi^{(\dot{4})})$ is fixed, Ψ is actually a linear transformation (that is, it is some sparse $0/1-$matrix). We can then write

$$g^\epsilon_\Phi(\theta) = 1 \iff r(U_\theta^\top U_\theta)\Psi \text{diag}(\Phi^{(\dot{4})}) \leq \epsilon.$$

Finally, note that $\text{diag}(\Phi^{(\dot{4})}) \subset \mathbb{R}^{(n+1)^4}$, that is, we have written the function class $\mathcal{G}$ as corresponding to some *subset* of all halfspace classifiers in $\mathbb{R}^{(n+1)^4}$. By Lemma 3.2, we know that the VC dimension of $\mathcal{G}$ is bounded by $(n+1)^4$. The result follows from Corollary 3.4.

□

9.2. Atomic interventions

Here we consider the instance in which our family of interventions consists of atomic interventions on a total of k nodes $\texttt{At}$. An atomic intervention at node i replaces the generative equation $X_i = \gamma_i(e)^\top \texttt{Pa}_i + \eta_i$ by the *assignment* $X_i = a_i$ for some scalar a_i. Note that we can interpret this as zeroing out the corresponding entries of B and changing the noise variable to be a constant a_i.

Lemma 9.3. *For some matrix U, we have*

$$\|\Phi^\top \Sigma_\theta \Phi f_\circ - \Phi^\top (\Sigma_\theta \beta - J(1-B_\theta)^{-1} e_t)\|_2^2 = r(\Phi)^{(\dot{4})} U a^{(\dot{2})}.$$

Proof. We will examine the dependence of $\mathbb{E}_\theta[X_{-t}X_{-t}^\top]$ and $\mathbb{E}_\theta[X_{-t}X_t^\top]$ on a_θ^i starting from Lemma 7.5.

For the former, note that $(\Sigma_\theta)_{ij}$ can be interpreted as having one term corresponding to every path from i *backwards* to some node k and then forwards to j. Because atomic interventions essentially cut off the connections between a node and its parents, any path that consists initially of backwards edges and then forward edges can only have a quadratic dependence on the parameters of the intervention, regardless of how many nodes are intervened on. It is only such paths that contribute to the covariance matrix, and thus to the thresholded polynomials that determine approximate invariance. We will make this explicit below.

The nature of atomic interventions is that $B_\theta = B_\Theta = BI_\Theta$ for each of the interventions $\theta \in \Theta$, where I_Θ is an identity matrix with zeros on the diagonal entries corresponding to $\texttt{At}$ (we are disconnecting each of the sites of intervention from their parents). In addition, the noise covariance changes. Previously, with independent unit variance noise, the covariance was $\Xi_\theta = \mathbb{E}_\theta[\eta\eta^\top] = I$. Since we are incorporating the assignments for the atomic interventions into the noise, it is now given by

$$(\Xi_\theta)_{ij} = \begin{cases} 1, & i = j \notin \texttt{At} \\ 0, & i \in \texttt{At}, j \notin \texttt{At}, i \neq j \\ 0, & i \notin \texttt{At}, j \in \texttt{At}, i \neq j \\ a_i^\theta a_j^\theta & i, j \in \texttt{At} \end{cases}$$

That is, we have replaced the submatrix corresponding to the sites of the interventions with $a^\theta (a^\theta)^\top$ where a^θ denotes the vector of assigments $a^\theta = (a_1^\theta, a_2^\theta, \cdots, a_{|\texttt{At}|}^\theta)$.

$$\begin{aligned}
(\Phi^\top \Sigma_\theta \Phi f_\circ)_j &= \sum_{k=0}^{2n-2} \sum_{\substack{r+s=k \\ r,s<n}} (\Phi^\top J B_\Theta^r \Xi_\theta (B_\Theta^\top)^s J^\top \Phi f_\circ)_j \\
&= \sum_{k=0}^{2n-2} \sum_{\substack{r+s=k \\ r,s<n}} \sum_{j_1,j_2,j_3,j_4,j_5,j_6,j_7} \Phi_{j_1,j} J_{j_1,j_2} (B_\Theta^r)_{j_2,j_3} (\Xi_\theta)_{j_3,j_4} (B_\Theta^s)_{j_5,j_4} J_{j_5,j_6}^\top \Phi_{j_6,j_7} (f_\circ)_{j_7} \\
&= \sum_{k=0}^{2n-2} \sum_{\substack{r+s=k \\ r,s<n}} \sum_{\substack{j_1,j_2,j_3,j_5,j_6,j_7 \\ j_2 \notin \texttt{At}}} \Phi_{j_1,j} J_{j_1,j_2} (B_\Theta^r)_{j_2,j_3} (\Xi_\theta)_{j_3,j_3} (B_\Theta^s)_{j_5,j_4} J_{j_5,j_6}^\top \Phi_{j_6,j_7} (f_\circ)_{j_7} + \\
&\qquad \sum_{\substack{j_1,j_2,j_3,j_4,j_5,j_6,j_7 \\ j_2,j_3 \in \texttt{At}}} \Phi_{j_1,j} J_{j_1,j_2} (B_\Theta^r)_{j_2,j_3} (\Xi_\theta)_{j_3,j_4} (B_\Theta^s)_{j_5,j_4} J_{j_5,j_6}^\top \Phi_{j_6,j_7} (f_\circ)_{j_7} \\
&= \sum_{k=0}^{2n-2} \sum_{\substack{r+s=k \\ r,s<n}} \sum_{\substack{j_1,j_2,j_3,j_5,j_6,j_7 \\ j_2 \notin \texttt{At}}} \Phi_{j_1,j} J_{j_1,j_2} (B_\theta^r)_{j_2,j_3} (B_\Theta^s)_{j_5,j_4} J_{j_5,j_6}^\top \Phi_{j_6,j_7} (f_\circ)_{j_7} + \\
&\qquad \sum_{\substack{j_1,j_2,j_3,j_4,j_5,j_6,j_7 \\ j_2,j_3 \in \texttt{At}}} \Phi_{j_1,j} J_{j_1,j_2} (B_\Theta^r)_{j_2,j_3} a_{j_3}^\theta a_{j_4}^\theta (B_\Theta^s)_{j_5,j_4} J_{j_5,j_6}^\top \Phi_{j_6,j_7} (f_\circ)_{j_7} \\
&= \sum_{k=0}^{2n-2} \sum_{\substack{r+s=k \\ r,s<n}} r(\Phi)^{(2)} U_j^{1,\Theta,r,s} a^{(2)}
\end{aligned}$$

$$= r(\Phi)^{(2)} U_j^{1,\Theta} a^{(2)}$$

Stacking the $U_j^{1,\Theta}$ together gives us $\Phi^\top \Sigma_\theta \Phi f_\circ = r(\Phi)^{(2)} U^{1,\Theta} a^{(2)}$.

Similarly to Lemma 9.1, we have

$$\begin{aligned}
\Phi^\top \Sigma_\theta \beta &= \sum_{k=0}^{2n-2} \sum_{\substack{r+s=k \\ r,s<n}} \sum_{j_1,j_2,j_3,j_4} \Phi_{j_1,j} (B_\theta^r)_{j_1,j_2} (\Xi_\theta)_{j_2,j_3} (B_\theta^s)_{j_4,j_3} \overline{\beta}_{j_4} \\
&= \sum_{k=0}^{2n-2} \sum_{\substack{r+s=k \\ r,s<n}} \sum_{\substack{j_1,j_2,j_4 \\ j_2 \notin \mathtt{At}}} \Phi_{j_1,j} (B_\theta^r)_{j_1,j_2} (\Xi_\theta)_{j_2,j_3} (B_\theta^s)_{j_4,j_3} \overline{\beta}_{j_4} + \sum_{\substack{j_1,j_2,j_3,j_4 \\ j_2,j_3 \in \mathtt{At}}} \Phi_{j_1,j} (B_\theta^r)_{j_1,j_2} (\Xi_\theta)_{j_2,j_3} (B_\theta^s)_{j_4,j_3} \overline{\beta}_{j_4} \\
&= \sum_{k=0}^{2n-2} \sum_{\substack{r+s=k \\ r,s<n}} \sum_{\substack{j_1,j_2,j_4 \\ j_2 \notin \mathtt{At}}} \Phi_{j_1,j} (B_\theta^r)_{j_1,j_2} (B_\theta^s)_{j_4,j_3} \overline{\beta}_{j_4} + \sum_{\substack{j_1,j_2,j_3,j_4 \\ j_2,j_3 \in \mathtt{At}}} \Phi_{j_1,j} (B_\theta^r)_{j_1,j_2} a_{j_2} a_{j_3} (B_\theta^s)_{j_4,j_3} \overline{\beta}_{j_4} \\
&= \sum_{k=0}^{2n-2} \sum_{\substack{r+s=k \\ r,s<n}} r(\Phi)^{(2)} U_j^{2,r,s} a^{(2)} \\
&= r(\Phi)^{(2)} U_j^2 a^{(2)}
\end{aligned}$$

Stacking the U_j^2 together, we get $\Phi^\top \Sigma_\theta \overline{\gamma}^\top = U^2 a^{(2)}$. Now,

$$\Phi^\top J(1-B_\theta)^{-1} e_t = r(\Phi)^{(2)} U^{3,\Theta} a^{(\dot{2})}$$

simply by noting that there is no dependence on the parameters a_i. Taking the difference, we have for some $U^{4,\Theta}$,

$$\Phi^\top \Sigma_\theta \Phi f_\circ - \Phi^\top (\Sigma_\theta \beta - J(1-B_\theta)^{-1} e_t) = r(\Phi)^{(2)} U^{4,\Theta} a^{(2)}.$$

Finally, the squared norm consists of terms that contain up to two terms of $r(\Phi)^{(\dot{2})}$ and two terms of $a^{(\dot{2})}$, this can be written as

$$\|\Phi^\top \Sigma_\theta \Phi f_\circ - \Phi^\top (\Sigma_\theta \beta - J(1-B_\theta)^{-1} e_t)\|_2^2 = r(\Phi)^{(\dot{4})} U^\Theta a^{(\dot{4})}.$$

□

Theorem 9.4. *Given $|\Theta_{train}| = O(\frac{k^4 + \log \frac{1}{\delta}}{\delta'})$ interventions drawn independently from a distribution $\mathcal{D}$ over all atomic interventions on some fixed set $\mathtt{At}$ of k nodes that with probability at least $1-\delta$ over the randomness in Θ_{train}, with probability at least $1-\delta'$ over the randomness in D, we have for $\theta \sim \mathcal{D}$, $\mathcal{I}^\epsilon(\Theta_{train}) \subseteq \mathcal{I}^\epsilon(\theta)$*

Proof. Recall Equation (8.6):

$$\|\nabla \mathcal{R}_\Phi^\theta(f_\circ)\|_2 = 2\|\Phi^\top \Sigma_\theta \Phi f_\circ - \Phi^\top (\Sigma_\theta \beta - J(1-B_\Theta)^{-1} e_t)\|$$

We show in Lemma 9.3 that we can simplify this expression down to one of the form

$$\|\nabla \mathcal{R}_\Phi^\theta(f_\circ)\|_2^2 = r(\Phi)^{(\dot{4})} U^\Theta a^{(4)}$$

for some different matrix U^θ. The function class $\mathcal{G}$ now consists of half-spaces,

$$f_\Phi^\epsilon(\theta) = 1 \iff r(\Phi)^{(\dot{4})} U^\Theta a^{(4)} \leq \epsilon$$

We can thus re-parameterize $\mathcal{G}$ using $r(\Phi)^{(4)}$ rather than Φ. Since these half-spaces live in $\mathbb{R}^{k^4}$ rather than $\mathbb{R}^{n^2}$, by Lemma 3.2 the VC-dimension is bounded by k^4. From Corollary 3.4, we have that $O((k^4 + \log \frac{1}{\delta})/\delta')$ interventions suffices to ensure that any $\tilde{\Phi}$ satisfying $f_{\tilde{\Phi}}^\epsilon(\theta) = 1$ for all $\theta \in \Theta_{\text{train}}$ satisfies $f_\Phi^\epsilon(\theta) = 1$ with probability $1-\delta'$ for $\theta \sim \mathcal{D}$. □

9.3. Soft Interventions

Soft interventions are those in which the *weights* of the SEM are modified while the underlying causal structure remains the same. That is, a soft intervention at node v_i is equivalent to replacing the equation $X_i = (\beta_i^{\theta_\circ})^\top X + \eta_i$ with $X_i = (\beta_i^\theta)^\top X + \eta_i$. Note that this is equivalent also to changing the ith row of B_θ. $\beta \in \mathbb{R}^{\sum_{i \in \mathtt{At}} |\mathtt{Pa}_i|}$ denote the vector of all weights involved in the set of soft interventions we are considering. Let $\underline{\beta}^{(2k)}$ denote the vector consisting of every combination of up to $2k$ entries of β modulo terms of the form $\beta^\theta_{l_2,l_1}\beta^\theta_{l_3,l_1}\beta^\theta_{l_4,l_1}$. In other words, $\underline{\beta}^{(2k)}$ does not contain any three weights from the same site.

Lemma 9.5. *For some matrix U, we have*

$$\|\Phi^\top \Sigma_\theta \Phi f_\circ - \Phi^\top(\Sigma_\theta \beta - J(1-B_\theta)^{-1}e_t)\|^2 = r(\Phi)^{(\dot{4})} U \underline{\beta}^{(2k)(\dot{2})}.$$

Proof. We will look at how Σ_θ and $(1-B_\theta)^{-1}$ depend on the parameters of the intervention.

Denote by Γ_{ki} the set of directed paths from node v_k to v_i in G. From Lemma 7.5 we have

$$(\Sigma_\theta)_{ij} = \sum_{\substack{\gamma_{ki}\in\Gamma_{ki}\\ \gamma_{kj}\in\Gamma_{kj}}} \prod_{\substack{(l_1,l_2)\in\gamma_{ki}\\ l_2\in\mathtt{At}}} \beta^\theta_{l_1,l_2} \prod_{\substack{(l_1,l_2)\in\gamma_{ki}\\ l_2\notin\mathtt{At}}} \beta_{l_1,l_2} \prod_{\substack{(l_1',l_2')\in\gamma_{kj}\\ l_2'\in\mathtt{At}}} \beta^\theta_{l_1',l_2'} \prod_{\substack{(l_1',l_2')\in\gamma_{ki}\\ l_2'\notin\mathtt{At}}} \beta_{l_1,l_2}$$

We can bound the number of terms in this expansion that depend on the intervention using a simple counting argument. There are no terms that are multiples of $\beta^\theta_{l_2,l_1}\beta^\theta_{l_3,l_1}\beta^\theta_{l_4,l_1}$ since there can be no directed path that passes through a single node twice (else we could find a cycle in the acyclic graph G). Thus each term in $(\Sigma_\theta)_{ij}$ consists of at most two terms from each site of intervention. There are k total sites, and d^2 choices of two terms at each, for a total of $(d^2)^k$ distinct possible terms in the covariance.

So each entry in the covariance matrix is an inner product of some vector with $\underline{\beta}^{(2k)}$.

Similarly, consider $(1-B_\theta)^{-1}$. We can write this as $\sum_{k=0}^{n-1} B_\theta^r$ (since B_θ is lower triangular, higher order terms are 0). We have

$$((1-B_\theta)^{-1})_{ij} = \sum_{\gamma_{ki}\in\Gamma_{ki}} \prod_{\substack{(l_1,l_2)\in\gamma_{ki}\\ l_2\in\mathtt{At}}} \beta^\theta_{l_1,l_2} \prod_{\substack{(l_1,l_2)\in\gamma_{ki}\\ l_2\notin\mathtt{At}}} \beta_{l_1,l_2} \tag{9}$$

Similarly, each entry here is an inner product of some vector with $\underline{\beta}^{(k)}$. In this case, no terms of the form $\beta^\theta_{l_2,l_1}\beta^\theta_{l_3,l_1}$ occur, since each backwards path can only pass through each site of intervention once.

We can then argue analogously to Lemma 9.3 that the complete each term in the gradient of the loss can be written using terms consisting of elements of $\underline{\beta}^{(2k)}$ and elements of $r(\Phi)^{(\dot{2})}$. The squared norm of the gradient can then be written using terms consisting of pairs of $\underline{\beta}^{(2k)}$ and pairs of elements of $r(\Phi)^{(\dot{2})}$. In other words,

$$\|\Phi^\top \Sigma_\theta \Phi f_\circ - \Phi^\top(\Sigma_\theta \beta - J(1-B_\theta)^{-1}e_t)\|^2 = r(\Phi)^{(\dot{4})} U \underline{\beta}^{(2k)(\dot{2})}$$

□

Theorem 9.6. *Given $O(\frac{d^{4k}+\log\frac{1}{\delta}}{\delta'})$ interventions drawn independently from a distribution $\mathcal{D}$ over all soft interventions on some fixed set $\mathtt{At}$ of k nodes such that each intervened node has in-degree at most d, we have that any $\epsilon-$approximately invariant representation Φ satisfies that with probability at least $1-\delta$ over the randomness in Θ_{train}, with probability at least $1-\delta'$ over the randomness in D, we have for $\theta \sim \mathcal{D}$,$\mathcal{I}^\epsilon(\Theta_{train}) = \mathcal{I}^\epsilon(\theta)$*

Proof. We show in the Lemma 9.5 that we can now write the norm of the gradient of the loss in terms of a product in a lower dimensional space again, similar to what we did previously.

$$\|\nabla \mathcal{R}^\theta_\Phi(f_\circ)\|_2^2 = r(\Phi)^{(4)} V \underline{\beta}^{(4k)}.$$

to get an upper bound on the VC dimension of d^{4k}. The theorem follows from Corollary 3.4. □

9.4. Indirect Observations

For this section, we change the observation model. There is assumed to be an underlying linear SEM, defined as in 1, but we observe $Z_{-t}, Z_t = (S_{-t}(X), S_t(X))$ for some S_{-t}, S_t. Here Z_t denotes the target variable, and Z_{-t} denotes the "features". For instance, If $S_t(X)$ was just X_t, and S_{-t} was diagonal with elements in $\{0, 1\}$, then this would correspond to a setting with latent (that is, hidden) variables. We need the following assumption on S_{-t} to ensure that the function class is still realizable, as in Lemma 3.5.

Assumption 9.7. There exists a matrix Φ such that $\Phi S_{-t}(X) = \texttt{Pa}_t$.

In the latent variables setting, this corresponds to the assumption that each of the true parents are actually visible. Finally, because we are in a more general setting, we no longer restrict to representations Φ that are diagonal (that is, feature selectors).

We use the same definition for the invariant representations, namely,

$$\Phi \in \mathcal{I}^{\epsilon}_{f_\circ}(\theta) \iff \|\nabla \mathbb{E}_\theta[f^\top \Phi Z_{-t} - Z_t] \mid_{f_\circ} \|_2 \leq \epsilon.$$

Here lemma 8.6 should be modified to account for the indirect observations.

Lemma 9.8. *We have,*

$$\|\nabla \mathbb{E}_\theta[\left(f^\top \Phi Z_{-t} - Z_t\right)^2] \mid_{f_\circ} \|_2 = 2\|\Phi^\top \mathbb{E}_\theta[S_{-t}(X)S_{-t}(X)^\top]\Phi f_\circ - \Phi^\top \mathbb{E}_\theta[S_{-t}(X)S_t(X)]\|_2$$

Proof. The proof is immediate. The expansions in Lemma 8.6 are due to linearity, which we do not leverage here. □

The following Lemma is similar to Lemma 9.1; we have included it for completeness.

Lemma 9.9. *For some matrix U_θ, we have*

$$\Phi^\top \mathbb{E}_\theta[S_{-t}(X)S_{-t}(X)^\top]\Phi f_\circ - 2\Phi^\top \mathbb{E}_\theta[S_{-t}(X)S_t(X)] = U_\theta r(\Phi)^{(2)}.$$

Proof. Observe that

$$\Phi^\top \mathbb{E}_\theta[S_{-t}(X)S_{-t}(X)]\Phi f_\circ = U^1_\theta r(\Phi)^{(\dot{2})}$$

where

$$(U^1_\theta)_{a,((ij),(kl))} = \begin{cases} \left(\mathbb{E}_\theta[S_{-t}(X)S_{-t}(X)^\top]\right)_{jk}(f_\circ)_l, & a = i \text{ and } (ij) \neq \cdot \text{ and } (kl) \neq \cdot \\ 0, & a \neq i \text{ or } (ij) = \cdot \text{ or } (kl) = \cdot \end{cases}$$

The second term is almost in the form we would like already

$$\Phi^\top \mathbb{E}_\theta[S_{-t}(X)S_t(X)] = U^2_\theta r(\Phi)^{(\dot{2})}$$

where

$$(U^2_\theta)_{a,((ij),(kl))} = \begin{cases} \left(\mathbb{E}_\theta[S_{-t}(X)S_t(X)]\right)_{ij} & (ij) \neq \cdot \text{ and } (kl) = \cdot \\ 0, & (ij) = \cdot \text{ or } (kl) \neq \cdot \end{cases}$$

Our result follows by setting $U_\theta = U^1_\theta + U^2_\theta$. □

Theorem 9.10. *Suppose that we are given $O(\frac{n^8 + \log\frac{1}{\delta}}{\delta'})$ interventions drawn independently from distribution $\mathcal{D}$ over the intervention index set Θ. Then with probability at least $1 - \delta$ over the randomness in Θ_{train}, the following statement holds:*

$$\mathbb{P}_{\theta\sim\mathcal{D}}(\mathcal{I}^{\epsilon}_{f_\circ}(\Theta_{train}) \notin \mathcal{I}^{\epsilon}_{f_\circ}(\theta)) \leq \delta'.$$

Proof. We again begin with our expression for the gradient of the loss given by Lemma 9.8. We show in Lemma 9.9 that we can write this as $\nabla \mathcal{R}^\theta_\Phi(f_\circ) = U_\theta r(\Phi)^{(\dot{2})}$ for some matrix U_θ. From Lemma 7.3 we can write the squared norm as $\|\nabla \mathcal{R}^\theta_\Phi(f_\circ)\|_2^2 = r(U^\top_\theta U_\theta)(r(\Phi)^{(\dot{2})})^{(2)}$. The result follows identically to Theorem 3.7. Note that the representation now has n^2 parameters already, so the half-space embedding is in dimension n^8. □

10. Finite sample bounds

Lemma 10.1. *Given $\frac{dL^2}{\epsilon^2}\log\frac{2d}{\delta}$ samples from Δ_θ, we can compute $\|\hat{\nabla}\mathcal{R}^\theta_\Phi(f_\circ)\|$ such that with probability $1-\delta$ over the samples drawn in each interventional distribution, for any fixed Φ,*

$$\|\nabla\mathcal{R}^\theta_\Phi(f_\circ)\| - \|\hat{\nabla}\mathcal{R}^\theta_\Phi(f_\circ)\|_2 \le \epsilon.$$

Proof. Suppose we had $2m$ samples from Δ_θ. Split these samples into m pairs $(X^{2i}, X^{2i+1})_{i=1}^m$ randomly. Denote by $R^\theta_\Phi(f, X)$ the loss at a point X, given by

$$R^\theta_\Phi(f, X) = \Phi^\top XX^\top\Phi f_\circ - \Phi^\top XX_t$$

Then we can estimate $\|\nabla\mathcal{R}^\theta_\Phi(f_\circ)\|$ as

$$\|\hat{\nabla}\mathcal{R}^\theta_\Phi(f_\circ)\| = \frac{1}{m}\sqrt{\left(\sum_{i=1}^m R^\theta_\Phi(f_\circ, X^{2i})\right)^\top\left(\sum_{i=1}^m R^\theta_\Phi(f_\circ, X^{2i+1})\right)}$$

Each of the inner terms concentrate about their mean as given by Lemma 7.1.

$$\begin{aligned}&\frac{1}{m}\sqrt{\left(\sum_{i=1}^m R^\theta_\Phi(f_\circ, X^{2i})\right)^\top\left(\sum_{i=1}^m R^\theta_\Phi(f_\circ, X^{2i+1})\right)} - \|\mathcal{R}^\theta_\Phi(f_\circ)\| \\ &\le \frac{1}{m}\sqrt{\left(\sum_{i=1}^m R^\theta_\Phi(f_\circ, X^{2i}) - \mathcal{R}^\theta_\Phi(f_\circ)\right)^\top\left(\sum_{i=1}^m R^\theta_\Phi(f_\circ, X^{2i+1}) - \mathcal{R}^\theta_\Phi(f_\circ)\right)} \le \epsilon\end{aligned}$$

where the final inequality is true for $m \ge \frac{dL^2}{\epsilon^2}\log\frac{2d}{\delta}$ with probability $1-\delta$. □

Lemma 10.2. *Consider the function class consisting of functions parameterized by Φ for a fixed θ*

$$\{\mathcal{R}^\theta_\Phi(f) := \Phi^\top\Sigma_e\Phi f - \Phi^\top\,\mathbb{E}_\theta[XX_t] \quad \Big| \quad \|\Phi\|_2 = 1, \|f\| = \sqrt{n}\}$$

Let $\mathcal{N}_\epsilon$ denote the covering number of this set with respect to the metric dist$(\Phi_1, \Phi_2) = \|\mathcal{R}^\theta_{\Phi_1}(f) - \mathcal{R}^\theta_{\Phi_2}(f)\|$. Then $\log\mathcal{N}_\epsilon = n^2\log\left(1 + \frac{4n^{3/2}}{\epsilon}\right)$.

Proof. We have

$$\begin{aligned}\text{dist}(\Phi_1, \Phi_2) &= \|\mathcal{R}^\theta_{\Phi_1}(f) - \mathcal{R}^\theta_{\Phi_2}(f)\| \\ &= \|\Phi_1^\top\Sigma_\theta\Phi_1 f - \Phi_2^\top\Sigma_\theta\Phi_2 f + \Phi_2^\top\,\mathbb{E}_\theta[X_{-t}X_t] - \Phi_1^\top\,\mathbb{E}_\theta[X_{-t}X_t]\| \\ &\le \|\Phi_1^\top\Sigma_\theta\Phi_1 f - \Phi_2^\top\Sigma_\theta\Phi_2 f\| + \|\Phi_2^\top\,\mathbb{E}_\theta[_{-t}XX_t] - \Phi_1^\top\,\mathbb{E}_\theta[X_{-t}X_t]\| \\ &\le \|\Phi_1^\top\Sigma_\theta\Phi_1 - \Phi_2^\top\Sigma_\theta\Phi_2\|\|f\| + \|\Phi_1 - \Phi_2\|\|\,\mathbb{E}_\theta[X_{-t}X_t]\| \\ &\le \|(\Phi_1 - \Phi_2)^\top\Sigma_\theta(\Phi_1 - \Phi_2)\|\|f\| + \|\Phi_1 - \Phi_2\|\|\,\mathbb{E}_\theta[X_{-t}X_t]\| \\ &\le \|\Phi_1 - \Phi_2\|^2\|\Sigma_\theta\|\|f\| + \|\Phi_1 - \Phi_2\|\|\,\mathbb{E}_\theta[X_{-t}X_t]\|\end{aligned}$$

We have $\|f_\circ\| = \sqrt{n}$. Because $\|X\|_\infty \le 1$, we know that $\|\Sigma_\theta\| \le n$ and $\|\,\mathbb{E}_\theta[X_{-t}X_t]\| \le \sqrt{n}$. Now we take ϕ to be an $\frac{\epsilon}{2n^{3/2}}$-covering of the space of representations Φ. We know from Lemma 7.2 that we can find one containing at most $\left(1 + \frac{4n^{3/2}}{\epsilon}\right)^{n^2}$ representations. That is, given any Φ, $\|\Phi\| \le 1$, we can find $\Phi' \in \phi$ such that $\|\Phi - \Phi'\| \le \frac{\epsilon}{2n^{3/2}}$. Then we have (assuming $\epsilon < 1, n > 1$)

$$\text{dist}(\Phi, \Phi') \le \left(\frac{\epsilon}{2n^{3/2}}\right)^2 n^{\frac{3}{2}} + \frac{\epsilon}{2n^{3/2}}\sqrt{n} \le \epsilon.$$

□

Putting these together,

Lemma 10.3. *Given* $\frac{4dL^2}{\epsilon^2}\left(\log\frac{2d}{\delta}+\log\mathcal{N}_{\epsilon/2}\right)$ *samples in a dataset, with probability* $1-\delta$*, for every representation* Φ *simultaneously we have*

$$\hat{\mathcal{I}}^{\frac{\epsilon}{2}}_{f_\circ}(\theta)\subseteq\mathcal{I}^{\epsilon}_{f_\circ}(\theta)$$

Proof. Consider $\Phi\in\hat{\mathcal{I}}^{\frac{\epsilon}{2}}_{f_\circ}(\theta)$. From the definition,

$$\|\hat{\nabla}\mathcal{R}^{\theta}_{\Phi}(f_\circ)\|<\frac{\epsilon}{2}.$$

By Lemma 7.2 and Lemma 10.1, we know that

$$\|\nabla R^{\theta}_{\Phi}(f_\circ)\|-\|\hat{\nabla}\mathcal{R}^{\theta}_{\Phi}(f_\circ)\|<\frac{\epsilon}{2}$$

Putting these together, we see that

$$\|\nabla R^{\theta}_{\Phi}(f_\circ)\|\leq\epsilon\implies\Phi\in\mathcal{I}^{\epsilon}_{f_\circ}(\theta).$$

□

Theorem 10.4. *Given*

$$m=\begin{cases}O(\frac{k^4+\log\frac{1}{\delta}}{\delta'}) & \Theta\text{ is }k\text{ nodes, hard interventions}\\ O(\frac{d^{4k}+\log\frac{1}{\delta}}{\delta'}) & \Theta\text{ is }k\text{ nodes, soft interventions}\\ O(\frac{n^4+\log\frac{1}{\delta}}{\delta'}) & \Theta\text{ any interventions}\end{cases}$$

interventional datasets, and $\frac{4nL^2}{\epsilon^2}\left(\log\frac{2nm}{\delta}+n^2\log(1+\frac{8n^{3/2}}{\epsilon})\right)$ *samples in each dataset, we have that with probability* $1-\delta$*, with probability* $1-\delta'$ *for* $\theta\sim\mathcal{D}$*,*

$$\hat{\mathcal{I}}^{\epsilon}_{f_\circ}(\Theta_{train})\subseteq\mathcal{I}^{2\epsilon}_{f_\circ}(\theta).$$

Proof. From Lemma 4.2 we know that

$$\hat{\mathcal{I}}^{\epsilon}_{f_\circ}(\theta)\subseteq\mathcal{I}^{2\epsilon}_{f_\circ}(\theta). \tag{10}$$

With a union bound over Θ_{train}, we can establish Equation (10) at once over all interventions in the training set. Since $\mathcal{I}^{\epsilon}_{f_\circ}(\Theta_{\text{train}})=\bigcap_{\theta\in\Theta_{\text{train}}}\mathcal{I}^{\epsilon}_{f_\circ}(\theta)$ and $\hat{\mathcal{I}}^{\epsilon}_{f_\circ}(\Theta_{\text{train}})=\bigcap_{\theta\in\Theta_{\text{train}}}\hat{\mathcal{I}}^{\epsilon}_{f_\circ}(\theta)$, we get that $\hat{\mathcal{I}}^{\epsilon}_{f_\circ}(\Theta_{\text{train}})\in\mathcal{I}^{\epsilon}_{f_\circ}(\Theta_{\text{train}})$. Finally, from Theorems 3.7, 3.8, 3.9, we have $\mathcal{I}^{\epsilon}_{f_\circ}(\Theta_{\text{train}})\subseteq\mathcal{I}^{\epsilon}_{f_\circ}(\Theta)$ for each of the cases listed. □

11. Empirical study

Construction of training and test interventional distributions: For the class of hard interventions, we consider assignments drawn from Gaussians on nodes $\{v_1,v_2,v_3\}$. Specifically, we take $X_i=a_i$ for $i\in\{3,4,5\}$ with $a_i\sim\mathcal{N}(0,\sigma^2)$ independently for each node. For the class of soft interventions, we assign each nonzero entry of $\beta^{\theta}_i\sim\mathcal{N}(0,\sigma^2)$ for $i\in\{3,4,5\}$ independently. In our experiments, $\sigma^2=1$. In this way, we generate m atomic-interventional datasets $\{N_{j,\text{hard}}\}_{j\in[m]}$ and m soft-interventional datasets $\{N_{j,\text{soft}}\}_{j\in[m]}$ for m varying from 3 to 20.

Subsets as Representations: We iterate over the power-set of the nodes to make a list of approximately invariant representations. For every subset S of the non-target nodes, we consider Φ_S to be the corresponding representation, diagonal, with a one on each index in S (a projection onto S). We count the fraction of these representations that are approximately invariant, as defined below.

Approximately Invariant Representations: Here, these are defined as being such that there is not much variation over the least squares solutions on top of the representations. For every such representation Φ_S, we denote by $f_{S,j}$ the least squares solution on top of the representation Φ_S for dataset N_j. We take as a measure of invariance the quantity

$$\rho_S=\frac{\sum_{j_1,j_2\in[m]}\|f_{S,j_1}-f_{S,j_2}\|_2}{(m-1)\|\sum_{j\in[m]}f_{S,j}\|_2}.$$

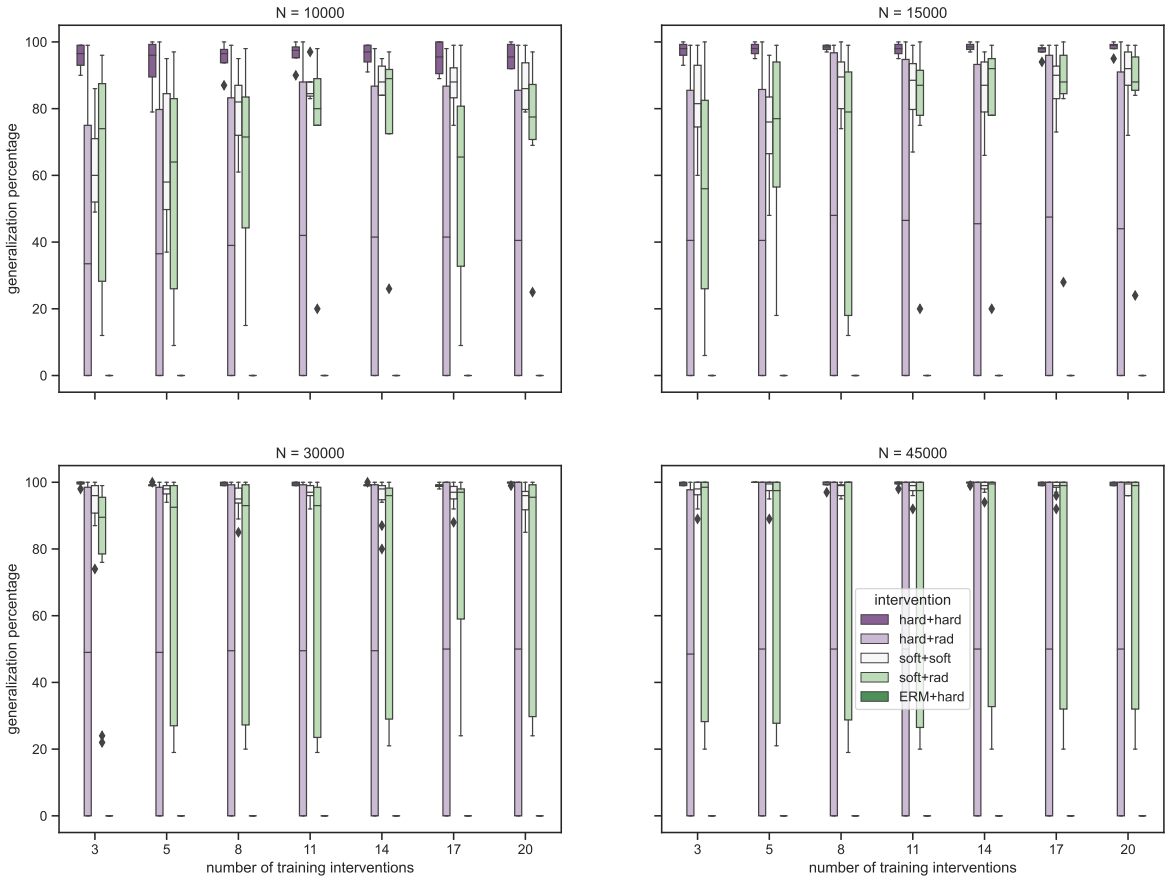

Figure 5. Comparison of generalization for various combinations of training and test interventional distributions. Here "hard" refers to the i.i.d. interventions setting with hard interventions from the same distribution in both training and test interventions, similarly "soft" refers to i.i.d. soft interventions, while "hard+rad" (respectively "soft+rad") refer to hard interventions (respectively soft) in training with rademacher interventions in test. We repeat the above for $N = 10,000, 15,000, 30,000$, and $45,000$.

Note that this is the ratio of the average distance between the various least squares solutions and the norm of the average least squares solution. We expect this to be low when the various least squares solutions are close to one another. An approximately invariant subset S is then taken to be one such that $\rho_S < 0.02$. Let s_{hard}, s_{soft} denote the approximately invariant subsets given the training sets $\{N_{j,\text{hard}}\}_{j\in[m]}$, $\{N_{j,\text{soft}}\}_{j\in[m]}$

Test distributions: We construct m_{test} test datasets induced by interventions drawn from each of the same families, that is, $\{N^{\text{test}}_{j,\text{hard}}\}_{j\in m_{\text{test}}}$ for atomic interventions, and $\{N^{\text{test}}_{j,\text{hard}}\}_{j\in m_{\text{test}}}$ for soft interventions, as well as a set of datasets $\{N^{\text{test}}_{\alpha,j,\text{rad}}\}_{j\in m_{\text{test}}}$, drawn from the linear SEM constructed by flipping each edge weight in the original SEM with probability α (except the ones connecting the parents of the target to the target). Note that this last dataset is not drawn from interventions that are drawn from the same distribution as the training datasets. For each set in s_{hard}, we return the percentage of datasets in $\{N^{\text{test}}_{j,\text{hard}}\}_{j\in m_{\text{test}}}$, $\{N^{\text{test}}_{j,\text{soft}}\}_{j\in m_{\text{test}}}$, and $\{N^{\text{test}}_{\alpha,j,\text{rad}}\}_{j\in m_{\text{test}}}$ such that the least squares solution on top of the subsets continues to be approximately the same. The labels of the plots reflect in order the trianing and test interventional ditributions, so "soft+rad(0.5)" means we are evaluating s_{soft} on 100 datasets $\{N_{0.5,j,\text{rad}}\}$.

ERM generalization: To demonstrate that generalization in this sense is indeed non-trivial, we also plot the fraction of least squares solutions to the observational datasets (where we pool together the training distributions) that exhibit low variance in test interventional distributions. This box plot is essentially trivial at 0.

Norm of gradient: Rather than using an RMS variance, we can use the norm of the gradient. The empirical results are similar (in fact, these are related to each other, since $\|\nabla R_1(f_1)\|, \|\nabla R_2(f_2)\| = 0, \|f_1 - f_2\|_2 \leq \epsilon \implies \|\nabla R_1(\frac{f_1+f_2}{2})\|, \|\nabla R_2(\frac{f_1+f_2}{2})\| \leq \beta\epsilon$ where β is the smoothness of the loss functions).