

Out-of-Variable Generalization

Siyuan Guo ^{*†‡}Jonas Wildberger [†]Bernhard Schölkopf [†]

Abstract

The ability of an agent to perform well in new and unseen environments is a crucial aspect of intelligence. In machine learning, this ability is referred to as *strong* or *out-of-distribution* generalization. However, simply considering differences in data distributions is not sufficient to fully capture differences in environments. In the present paper, we assay *out-of-variable* generalization, which refers to an agent's ability to handle new situations that involve variables never jointly observed before. We expect that such ability is important also for AI-driven scientific discovery: humans, too, explore "Nature" by probing, observing and measuring subsets of variables at one time. Mathematically, it requires efficient re-use of past marginal knowledge, i.e., knowledge over subsets of variables. We study this problem, focusing on prediction tasks that involve observing overlapping, yet distinct, sets of causal parents. We show that the residual distribution of one environment encodes the partial derivative of the true generating function with respect to the unobserved causal parent. Hence, learning from the residual allows zero-shot prediction even when we never observe the outcome variable in the other environment.

1 Introduction

An animal's natural intelligence differs from machine intelligence in their ability to generalize from one problem to the next. Much of modern machine learning can be seen as large scale pattern recognition on suitably collected *independent and identically distributed* (i. i. d.) data. Its success boils down to its ability to generalize from a set of data points to the next one sampled from the same distribution. Though both using the term *generalization*, generalization in the i. i. d. setting is not the same as generalization from one problem to the next. The machine learning community studies the latter under the term *out-of-distribution* generalization [52, 24, 48, 2, 31, 66, 64], where training and test data deviate in their distributions. However, differences in distributions may not fully describe differences in environments [6]. In the present work, we investigate generalization across environments containing different sets of variables, referring to the problem as *out-of-variable* (OOV) generalization.

Out-of-variable generalization aims to transfer knowledge learned from a set of source environments to a target environment that contains variables never jointly present in any of the source environments. This type of learning requires synthesizing information from multiple sources that contain different sets of variables. For example, human doctors can often accurately diagnose a patient with a disease even if their symptoms, medical history and set of available clinical tests differ from past patients, whereas machine learning models struggle in prediction when facing overlapping, yet distinct, sets of variables. We can think of 'Nature' as a complex system of interrelated processes and variables. Humans, due to limited resources and knowledge, only observe and measure a subset of variables at a time. Scientific discovery [50, 23] can thus be seen as a problem in which scientist need to generalize *out-of-variable*. It is crucial to understand how and to what extent knowledge can be re-utilized.

*Correspondence to: sguo26v@gmail.com

[†]Max Planck Institute for Intelligent Systems, Tübingen, Germany

[‡]Department of Computer Science, University of Cambridge, United Kingdom

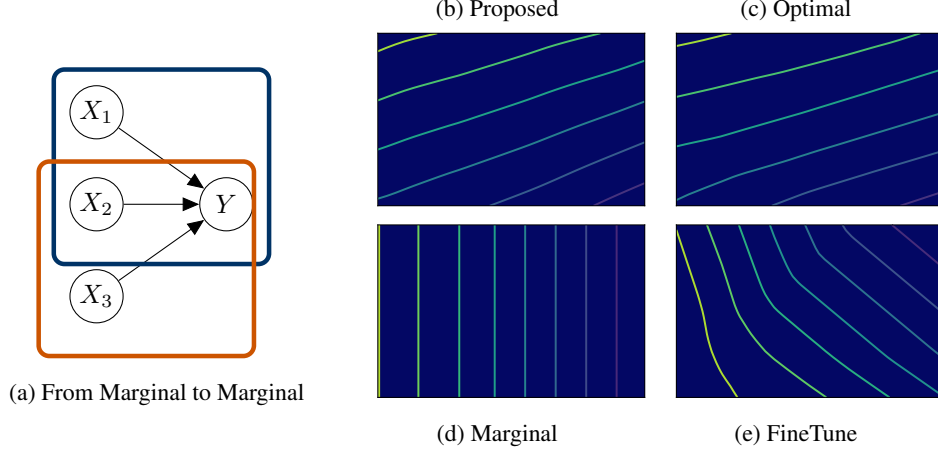


Figure 1: **Toy example:** (a) the blue box includes variables observed in the source environment, and the orange box those in the target environment. A directed edge represents a causal relationship between two variables. The goal is to improve the zero-shot (i.e., without additional data) prediction of Y in the target environment using the source environment. (b)-(e) Learning from the residual distribution in the source environment, our proposed method exhibits zero-shot performance close to the optimal solution compared to two baselines, as shown via the predictors’ contour plots (for details, cf. § 4.1).

Recent work in transfer learning [59, 63, 34, 18, 65] has focused on the development of discriminative models. Such models are preferred [57] over generative models in prediction tasks as they more directly solve the task at hand. They learn a direct map from inputs X to the target variable Y , aiming to find $\mathbb{E}[Y | X]$. Common strategies are to either extract re-usable features [33, 38, 56] or fine-tune model parameters [15, 51, 25, 12, 60] across tasks. Empirical performance shows such reuse of knowledge generally performs better than starting from a random state and is also more data-efficient. Marginal problems [58, 53, 44, 27, 36, 17, 20], on the other hand, often apply generative modelling approaches, i.e., to search joint distributions that are compatible with marginal observations. *Out-of-variable* generalization combines marginal problems and transfer learning. The goal is to re-use marginal observations for prediction tasks. In the present work, we propose a method that generalizes to variables never jointly present in the source environments more efficiently than by modelling the joint distribution. **Our contributions** are:

- We introduce and formalize *out-of-variable* generalization (§ 2) and contextualize this concept within the existing literature (§ 5).
- We prove impossibility and possibility theorems to show limitations of discriminative approaches in *out-of-variable* scenarios (§ 3.2).
- Modelling moments from the residual distribution of the source environment has predictive power for variables never jointly observed before. Via a motivating example (§ 3.3), we propose a method (§ 3.4) for general nonlinear smooth functions. We find in experiments (§ 4) that our approach achieves strong zero-shot transfer performance.

Fig. 1 provides a toy examples giving an overview of the main idea.

2 Out-of-Variable Generalization

2.1 Motivation

Definition 1 (Structural Causal Model (SCM)) A structural causal model \mathcal{M} consists of a collection of random variables X_1, \dots, X_n and its corresponding structural assignments

$$X_i := f_i(\mathbf{PA}_i, U_i), \forall i = 1, \dots, n, \quad (1)$$

where \mathbf{PA}_i is the parents or direct causes of X_i and U_i are jointly independent noise variables. Given a SCM \mathcal{M} , one can find its corresponding directed acyclic graph (DAG) where each node has all its incoming edges from its parents set. Any joint distribution generated from some SCM \mathcal{M} satisfying the DAG structure \mathcal{G} allows the **Markov factorization**

$$p(x_1, \dots, x_n) = \prod_{i=1}^n p(x_i \mid \mathbf{pa}_i^{\mathcal{G}}) \quad (2)$$

where $\mathbf{pa}_i^{\mathcal{G}}$ are parents of node x_i in graph \mathcal{G} .

Principle 1 (Independent Causal Mechanisms (ICM)) A change in one mechanism $p(x_i \mid \mathbf{pa}_i^{\mathcal{G}})$ does not inform [21, 27] or influence [47] any of the other mechanisms $p(x_j \mid \mathbf{pa}_j^{\mathcal{G}}) (i \neq j)$.

Structural causal model (SCM) [39] establishes a mathematical framework modelling causal relationships exist in Nature and ICM principle [49] postulates that true causal mechanisms should be independent of each other. Prior work [1, 22, 4, 45] about causality and generalization exploit the invariance property of causal mechanisms across domains. Beyond causal frameworks, learning distributed representations [8, 55], e.g. taking the form of Eq. 2, has also long been argued to offer advantages in efficient parameter estimation [30, 11] and interpretability [28, 32]. In this paper, we highlight another advantage:

The Markov factorization frees us from the need of observing all variables of interest at the same time.

Observation For each observed variable i , suppose there exists some environment that contains X_i and its causal parents $\mathbf{PA}_i^{\mathcal{G}}$. Then by Eq. 2, one can recover the joint distribution by piecing together learnt conditional distributions $p(x_i \mid \mathbf{pa}_i^{\mathcal{G}})$ in each environment, even though there may never exist an environment that contains all variables of interest at once. This phenomenon also occurs in undirected probabilistic graphical models [30], where the joint distribution⁴ is recoverable given potentials learnt from environments that only contain variables appearing in each clique.

2.2 OOV Generalization

Formally, an environment \mathcal{E} consists of a domain \mathcal{D} and a task \mathcal{T} . The domain is defined by two components: a feature variable space, denoted as $\mathcal{X} := (X_1, X_2, \dots)$, and its corresponding marginal probability distribution, represented by $P(\mathcal{X})$. Given a domain, a task is defined by: a target variable space, represented by $\mathcal{Y} := Y$, and a predictive function, represented as $f(\cdot)$. The variable space in an environment contains both the feature and target variable spaces. Note, in our setting, there could be multiple source environments, e.g. $\mathcal{E}_S := \{\mathcal{E}_s\}_{s \in S}$, whereas the target environment only comprises a single environment \mathcal{E}_T . To differentiate between components belonging to the source and target environments, the subscripts s and T are used, respectively.

What does it mean to have same variables? In probability and statistics, one considers a random variable as a function that maps the outcome of an experiment to a real number. It is equipped with a probability space, $(\Omega, \mathcal{F}, \mathbb{P})$. One can interpret the space as a model for experiment where Ω is the set of possible outcomes, \mathcal{F} is the set of observable sets of outcomes, and $\mathbb{P}(A)$ is the probability of an event A . In practice, random variables may share the same physical meaning but have different distributions, e.g. blood pressure measured in different age groups. Under certain pathological case, one can even construct random variables that have different physical meanings but share the same mathematical object. For example, the risk of a disease is governed by law in Nature relating to clinical variables, like blood pressure, heart rate and lifestyle. Such law remains the same across different patients in different hospitals. The variables recorded in each hospital share the same physical meanings, though may differ in distributions. Such inconsistency between statistical and physical interpretation of variables motivates us to formalize what we mean by *out-of-variables*.

Definition 2 Two random variables share physical meaning if they represent the same physical property [9] of the same physical system [7]. Further, they also share the same measurement process. We say \mathcal{A} is not a subset of \mathcal{B} , if there exists a random variable X in \mathcal{A} such that it does not share physical meaning with any of the random variables in \mathcal{B} .

⁴The joint distribution can be represented as $p(x_1, \dots, x_n) = \frac{1}{Z} \prod_{c \in \text{cliques}} \Psi_c(x_c)$.

Definition 3 (OOV Generalization) *Out-of-variable uncertainty arises when the variable space of source environment(s) differs from the variable space of the target environment, i.e. $\forall s \in S, \{\mathcal{X}_T, \mathcal{Y}_T\} \not\subseteq \{\mathcal{X}_s, \mathcal{Y}_s\}$. The need for generalization happens when we want to improve a quantity Q_T in the target environment using the related information from source environment(s).*

For this paper, $Q_T = f_T(\cdot)$, the target predictive function. Appendix A provides a detailed discussion of the need to introduce Def. 2 and its connection with statistics and physics. Though the applicability of the work set out in this paper is not limited to a causal framework, for ease of illustration, from here onwards we consider variables are causally related with each other.

3 Transfer Learning From Residual

3.1 Problem Formulation

Consider a structural causal model that is generated by an additive noise model (ANM):

$$\begin{aligned} Y &:= \phi(X_1, X_2, X_3) + \epsilon \\ X_1 &:= U_1, X_2 := U_2, X_3 := U_3 \end{aligned} \quad (3)$$

where ϕ is some unknown function and U_1, U_2, U_3 are jointly independent random variables and $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Assume we do not have access to the joint environment J that contains all variables of interest, namely (X_1, X_2, X_3, Y) . Instead, we consider:

- Source environment S with variables (X_1, X_2, Y)
- Target environment T with variables (X_2, X_3, Y)

The optimal predictive functions achievable via minimizing mean squared error loss are $f_S(x_1, x_2) = \mathbb{E}[Y \mid x_1, x_2]$ and $f_T(x_2, x_3) = \mathbb{E}[Y \mid x_2, x_3]$, where each corresponds to a mapping learnt under discriminative modelling from environment S and T respectively. We consider the residual distribution from the source environment as the distribution of $Y - f_S(X_1, X_2) \mid X_1, X_2$. The optimal predictive function satisfies **marginal consistency** condition:

$$\mathbb{E}_{X_1}[f_S(x_1, x_2)] = \mathbb{E}_{X_3}[f_T(x_2, x_3)] \quad (4)$$

Suppose Y from the target environment is never observed. The challenge is that directly training a classifier in the target environment is not possible. Our goal is to learn the target predictive function f_T in a data-efficient way by transferring knowledge learned from the source environment.

3.2 Theorem: Possibility and Impossibility

In this section, we show limitations of discriminative modelling, which considers $\mathbb{E}[Y \mid X]$ only without taking into account of its residual distribution. Under such modelling, Theorem 3.1 shows it is impossible to identify the optimal predictive function in the target environment for scenarios shown in Fig. 2. In this context, non-identifiability means non-uniqueness of a target function satisfying marginal consistency conditions (as described in Eq. 4). See Appendix C.1 and C.2 for detailed proofs.

Theorem 3.1 (Impossibility Theorem) *Consider out-of-variable scenarios illustrated in Fig. 2 (a)-(c). They are governed by the structural causal model described in § 3.1. Suppose all variables considered are continuous variables, except X_1, X_3 in Fig. 2c are binary. Assume 1) for all i , $p(x_i)$ is known and let the support set be $S_i := \{x \in \mathbb{R} \mid p_i(B(x, \epsilon)) > 0\}$ where p_i denotes the marginal distribution of x_i and $B(x, \epsilon)$ is the ball centred around x with radius ϵ . 2) Suppose for all i there exists two distinct points $x, x' \in S_i$. Given fixed f_S , for any function f_T that satisfies $\mathbb{E}_{X_1}[f_S(x_1, x_2)] = \mathbb{E}_{X_3}[f_T(x_2, x_3)]$, there always exists a different function f'_T such that it satisfies the above condition and for any chosen $R > 0$, $\|f_T - f'_T\|_2 \geq R$.*

Theorem 3.1 says enforcing marginal consistency between source and target predictive function is (in general) not sufficient to identify a unique solution, and the learnt solution may deviate arbitrarily large from the optimal one. While such impossibility theorem is discouraging, Theorem 3.2 illustrates special scenarios where exploiting information contained in $\mathbb{E}[Y \mid X]$ for each environment is sufficient to identify the optimal predictive function in the target environment.

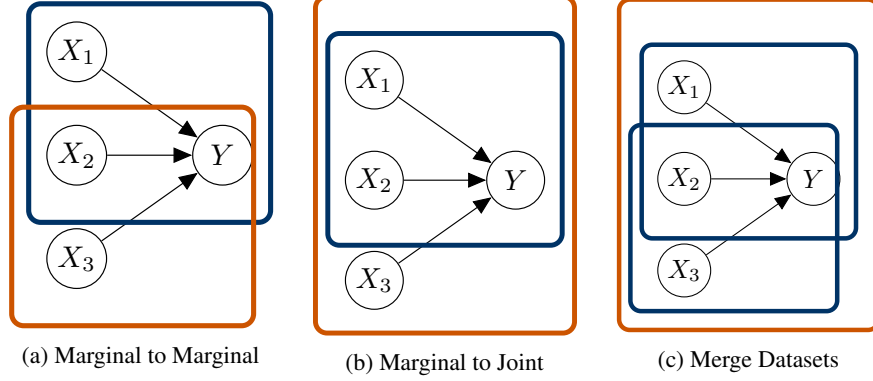


Figure 2: Examples of OOV Scenarios

Theorem 3.2 Consider a target variable Y and its corresponding causal parents \mathbf{PA}_Y . Let Z be a common child of all variables in \mathbf{PA}_Y and it satisfies $Z \perp\!\!\!\perp Y \mid \mathbf{PA}_Y$. Given:

- Environment S_1 contains variables (Z, Y) and $f_{S_1}(z) = \mathbb{E}[Y \mid Z]$,
- Environment S_2 contains variables (\mathbf{PA}_Y, Z) and $f_{S_2}(\mathbf{pa}_Y) = \mathbb{E}[Y \mid \mathbf{PA}_Y]$

Assume the SCM is generated by an additive noise model and f_{S_2} is an invertible function. With small noise in the functional relationship between Z and \mathbf{PA}_Y , composing $f_{S_1} \cdot f_{S_2}$ identifies the optimal predictive function in the target environment that contains variables never jointly observed previously, i.e., (\mathbf{PA}_Y, Y) . Fig. 3 in Appendix C.2 shows an example of such scenario where $Z = X_1$ and $\mathbf{PA}_Y = X_2$.

Näive Model Given the problem described in § 3.1, one approach is to train separate models for each variable that is observed, e.g. (X_1, Y) and (X_2, Y) . This hopes to additively re-use the model trained on the overlapping variable X_2 . When facing the prediction task with variables (X_2, X_3, Y) , the remaining step is to train a model on (X_3, Y) . One, then, hopes to recover the optimal predictive function by flexibly composing trained models on different sets of variables. This approach also circumvents the need to jointly observe variables of interest (X_2, X_3, Y) . Here we highlight, this method relies on the fact that there is no interaction between different variables. We show the model in detail in Appendix B. Above argument still applies when variable considered are replaced to any subsets of variables.

In this paper, we ask: how and to what extent can we generalize given we do not observe all variables of interest. We illustrate our method via a motivating example § 3.3 with details in § 3.4. The idea is: 1) general to work for scenarios in Fig.2, and 2) is not restricted to strict causal assumptions, e.g. covariates need not be strict causal parents.

3.3 Motivating example

Consider the problem described in § 3.1 where ϕ is a polynomial function:

$$Y := \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_1 X_2 + \alpha_5 X_1 X_3 + \alpha_6 X_2 X_3 + \alpha_7 X_1 X_2 X_3 + \epsilon \quad (5)$$

Let X_i has mean μ_i , variance σ_i and skew k_i for all i . The optimal predictive functions for environment S and T are:

$$\begin{aligned} f_S(x_1, x_2) &= (\alpha_3 \mu_3) + (\alpha_1 + \alpha_5 \mu_3)x_1 + (\alpha_2 + \alpha_6 \mu_3)x_2 + (\alpha_4 + \alpha_7 \mu_3)x_1 x_2 \\ f_T(x_2, x_3) &= (\alpha_1 \mu_1) + (\alpha_3 + \alpha_5 \mu_1)x_3 + (\alpha_2 + \alpha_4 \mu_1)x_2 + (\alpha_6 + \alpha_7 \mu_1)x_2 x_3 \end{aligned} \quad (6)$$

A common strategy in transfer learning is to **fine-tune** model parameters. We see from the above example that such strategy is reasonable, as the coefficients in each term does overlap in Eq. 6. We can thus hope that the model parameters can be adapted quickly when training in the target environment. Though it is clear that we cannot uniquely determine the coefficients of f_T without

observing any data from environment T . Since even with known representation, the system of equations is under-determined with eight unknown coefficients and four estimated values.

To illustrate our idea, consider the skew from the residual distribution in environment S :

$$\mathbb{E}[(Y - f_S(x_1, x_2))^3 \mid x_1, x_2] = (\alpha_3 + \alpha_5 x_1 + \alpha_6 x_2 + \alpha_7 x_1 x_2)^3 \mathbb{E}[(X_3 - \mu_3)^3] \quad (7)$$

We observe the term in bracket coincides exactly with the first-order partial derivative of the generating function ϕ with respect to variable X_3 evaluated at x_1, x_2, μ_3 :

$$\left. \frac{\partial \phi}{\partial X_3} \right|_{x_1, x_2, \mu_3} = \alpha_3 + \alpha_5 x_1 + \alpha_6 x_2 + \alpha_7 x_1 x_2 \quad (8)$$

Under the polynomial above, let $Z = \frac{1}{k_3}(Y - f_S(x_1, x_2))^3$ and $g(x_1, x_2) = [x_1, x_2, x_1 x_2]$. Learning θ from fitting Z with $(\theta^T g(x_1, x_2))^3$ enables estimation of coefficients $\alpha_3, \alpha_5, \alpha_6, \alpha_7$. Combining with the coefficients estimated from learning f_S , we are able to recover the exact coefficient about f_T without the need to train data from environment T .

Discussion The intuition behind this seemingly surprising ability is rather straightforward. The variable X_3 though unobserved by the source environment is a generating factor of the target variable Y . Its information is not only contained in the marginalized mean but also in the residual distribution. Looking at moments of the residual distribution provides an additional set of equation that helps to determine exact coefficients. Even without the polynomial assumption, the moment also provides additional information about the partial derivative of the generating function with respect to X_3 . This is learnable as the skew of our current residual distribution is a propagated effect of the skew of X_3 with some interference by the noise variable. We can estimate the skew of X_3 from samples in the target environment without the need to jointly observe its corresponding Y . Such phenomenon, in fact, extends to general nonlinear smooth functions.

3.4 Method

Theorem 3.1 *Consider the problem setup in § 3.1. Assume the generating function ϕ is 2-times differentiable with respect to the third covariate everywhere and source environment S contains variables (X_1, X_2, Y) with $f_S(x_1, x_2) = \mathbb{E}[Y \mid x_1, x_2]$. For fixed x_1, x_2 , approximate the function $\phi : x_1 \times x_2 \times X_3 \rightarrow \mathbb{R}$ with first-order Taylor approximation, we have:*

$$\mathbb{E}[(Y - f_S(x_1, x_2))^n \mid x_1, x_2] = \sum_{k=0}^n \binom{n}{k} \mathbb{E}[\epsilon^k] \left(\left. \frac{\partial \phi}{\partial X_3} \right|_{x_1, x_2, \mu_3} \right)^{n-k} \mathbb{E}[(X_3 - \mu_3)^{n-k}] \quad (9)$$

When $n = 3$:

$$\mathbb{E}[(Y - f_S(x_1, x_2))^3 \mid x_1, x_2] = \left(\left. \frac{\partial \phi}{\partial X_3} \right|_{x_1, x_2, \mu_3} \right)^3 \mathbb{E}[(X_3 - \mu_3)^3] + \mathbb{E}[\epsilon^3] \quad (10)$$

Intuition Under first-order Taylor approximation about the underlying generating function, the moments of the residual distribution is a mixed effect between the moments of noise variable and the propagated effects caused by the unobserved variable. In particular, when $n = 3$, most terms involving undesired noise variable disappear which allows us to estimate desired information. See Appendix C.3 for detailed proof.

Discussion One can consider this first-order approximation to the true generating function as a product of the function learned in the source environment with a linear interaction with the unobserved variable. This means if the true generating function is indeed linear with respect to the unobserved variable, for example when ϕ is as in § 3.3, then our solution will be exact. This results in a function space that respects any general function with respect to the observed variables X_1, X_2 while varying linearly in X_3 , i.e., $H(x_1, x_2) \times L(x_3)$ where L is a linear function of X_3 . Under model misspecification [61], our method finds a function in this function space that is closest to the true generating function. For ease of illustration, we demonstrate under first-order Taylor approximation with a single unobserved variable. This approach provides us the flexibility to choose the trade-off between approximation accuracy and computational efficiency. Model users can determine the degree of approximation accuracy by choosing the number of moments to model, as well as the degree of the Taylor approximations. See Appendix D.1 and D.2 for extensions of our method to higher-order approximations and to situations that involve multiple unobserved variables.

3.4.1 Zero-shot Learning

Applying Theorem 3.1, we propose zero-shot estimate \tilde{f}_T of optimal predictive function in the target environment:

$$\tilde{f}_T(x_2, x_3) = \frac{1}{n} \sum_{i=1}^n f_S(x_{1,i}, x_2) + h_\theta(x_{1,i}, x_2)(x_3 - \mu_3), \quad \text{where } x_{1,i} \sim p(x_1) \quad (11)$$

where h_θ is a neural network parameterized by θ and is learnt to estimate the partial derivative via modelling the third moment of the residual distribution from the source environment. One can consider the proposed method as a first-order approximation to the true generating function around its third variable. This will thus be strictly better than naïvely marginalizing f_S from the source environment. Further the proposed estimate automatically satisfies the marginal consistency condition as the second term in Eq. 11 vanishes by taking expectation, i.e., $\mathbb{E}_{X_3}[\tilde{f}_T(x_2, x_3)] = \mathbb{E}_{X_1}[f_S(x_1, x_2)]$. Interestingly, our result shows even though identification may be impossible, there is prediction power from marginal observations. See Algorithm 1 for detailed procedure and Appendix E for derivation.

Algorithm 1: Zero-shot learning

Input : Source environment S with variables X_1, X_2 and Y ; Target environment T with variables X_2 and X_3 .

Output : Zero-shot predictive function $\tilde{f}_T(x_2, x_3)$

- 1 **Step 1:** Learn $\mathbb{E}[Y \mid X_1, X_2]$
 - 2 Train a neural network f_S via minimizing its mean squared error $\|Y - f_S(x_1, x_2)\|_2^2$
 - 3 **Step 2:** Learn partial derivative h_θ from modelling conditional skew
 - 4 Compute $Z = (Y - f_S(X_1, X_2))^3$.
 - 5 Estimate the skew of X_3 : $k_3 = \mathbb{E}[(X_3 - \mu_3)^3]$, where $\mu_3 = \mathbb{E}[X_3]$.
 - 6 Train a neural network h_θ via minimizing $\|Z - k_3 h_\theta(x_1, x_2)^3\|_2^2$
 - 7 **Step 3:** Monte Carlo Estimation
 - 8 Uniformly sample n observations of X_1 from Env S : $\{x_{1,i}\}_{i=1}^n$.
 - 9 For fixed x_2, x_3 , calculate the proposed zero-shot estimate in Eq. 11.
-

4 Experiments

We perform experiments to evaluate our algorithm’s zero-shot learning performance. We find that our method demonstrates both qualitative and quantitative empirical success. We compare the proposed method with several baselines:

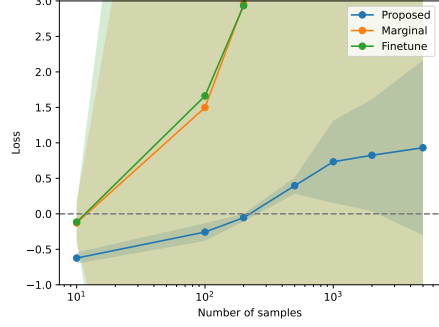
- **Optimal:** Model learnt if there is sufficient data from the target environment
- **Marginal:** Marginalize model learnt from the source environment to overlapping variable(s)
- **FineTune:** Model learnt from the source environment (zero-shot prediction at the start of the fine-tuning process)

4.1 Synthetic datasets

We generate synthetic data described in § 3.1 where for all i , $X_i \sim \Gamma(1, 1)$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.1$. We standardize the covariates into range $[0, 5]$ and generate $100k$ observations for the source environment and 50 covariates samples from the target environment, i.e., there is no observation of Y in the target environment. To learn f_S and the partial derivative h_θ , we use a 2-layer MLP with ReLU activation function. All MLPs are trained to minimize its mean squared error loss between the target and predicted values. We sample $n = 1000$ observations of X_1 from the source environment to perform monte-carlo approximation according to Eq. 11.

Qualitative Fig. 1 (b)-(e) shows qualitative comparisons for contour plots of the zero-shot prediction of our proposed method compared to optimal and the other two baselines. We see our method is almost identical to the optimal baseline, whereas the other two baselines deviate far from optimal. This means our method achieves similar prediction performance to that trained on the target environment with sufficient data. This shows learning the third moment of the residual distribution in the source environment is sufficient under certain function classes to achieve accurate zero-shot estimation.

Quantitative To evaluate our method’s performance on general nonlinear functions, we generate functions in each function class over 5 random seeds. We then aggregate results uniformly over polynomial, nonlinear and trigonometric functions (described in Appendix F). We evaluate the quantitative sample efficiency advantage brought by zero-shot estimation of our method. In the figure from the right-hand side, we compare the number of samples that contain joint variables required to observe from the target environment (x-axis) with the percentage advantage offered by zero-shot estimation instead of training from scratch. Loss (y-axis) is calculated as the percentage difference between zero-shot loss estimated by a specific method (loss_m) and loss achieved over training on the number of samples observed (loss_o), i.e., $(\text{loss}_m - \text{loss}_o)/\text{loss}_o$. This means the lower the loss, the larger the advantage of our proposed method over a new model trained from scratch in the target environment on the given number of samples. We can see our method far outperforms the baselines while maintaining tighter error bars. This shows the generality of our proposed method on different function classes and the stability of our estimations. In particular, the method offers significant benefit over training from scratch until a sufficient number of data points have been collected in the target environment.



5 Related Work

Missing data [46] studies the problem when covariates are missing for individual data points. Most approaches [16, 14, 29] either omit data points that contain missing values or aim to impute with estimated values. Our problem differs in that we consider variables that are missing entirely from certain environments, but we still aim to do zero-shot prediction – which neither of these two approaches covers.

Marginal problems in the classical statistical [58] and causal literature [26, 36, 20, 17, 44] study the problem of learning over marginal models to infer properties about the joint model. In contrast with common approaches that enforce compatibility over joint distributions [36, 26], our work shows learning from the residual distribution is sufficient to achieve strong transfer performance and specifies the connection between residual distributions and functions.

Generalization has predominantly focused on *out-of-distribution* generalization [3, 31, 66, 2]. Causality-related approaches either study the transfer of causal effects across different experimental conditions [40, 6, 41, 5, 13] or aim to learn invariant causal information that are robust in different environments [4, 35, 22, 42, 43, 45]. Our work differs from the former as we study generalization problem to variables never jointly observed before and the latter implicitly assumes the causal information shared fully characterize the task of prediction. In our problem, however, different environments share fundamentally partial but related information. Thus, learning only shared information as shown by our marginal baseline is inefficient to achieve good transfer performance.

6 Discussion

Limitations and Extensions The performance criteria in Def. 3 is defined as the target predictive function. Here we highlight it can be extended to other criteria such as causal structure learning as investigated by [36, 62]. We also think it would be interesting to extend the notion of variables beyond observable and measurable quantities (e.g. latent variables). Our method is now demonstrated under first-order approximations with one unobserved variable. For extensions to achieve higher approximation accuracy and more unobserved variables, see Appendix D.

Conclusion In this paper, we introduced *out-of-variable* generalization. We show that learning moments from the residual distribution in the source environment allows us to achieve strong zero-shot performance on variables never jointly observed before and is efficient to estimate given complex datasets. We hope our work inspires future work to explore and develop methodologies that could apply to different *out-of-variable* scenarios.

References

- [1] Causal inference using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society, Series B*:947–1012, 2016.
- [2] Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34: 3438–3450, 2021.
- [3] Martin Arjovsky. *Out of distribution generalization in machine learning*. PhD thesis, New York University, 2020.
- [4] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.
- [5] Elias Bareinboim and Judea Pearl. A general algorithm for deciding transportability of experimental results. *Journal of causal Inference*, 1(1):107–134, 2013.
- [6] Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.
- [7] Ori Belkind. Physical systems: conceptual pathways between spacetime and matter. 01 2004.
- [8] Y. Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35:1798–1828, 08 2013. doi: 10.1109/TPAMI.2013.50.
- [9] Mark Burgin. *Theory of knowledge: Structures and processes*. 12 2016. doi: 10.1142/8893.
- [10] Paul Busch and Pekka Lahti. Observable (compendium entry), July 2008. URL <http://philsci-archive.pitt.edu/4109/>.
- [11] Adam Coates and Andrew Y Ng. The importance of encoding versus training with sparse coding and vector quantization. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 921–928, 2011.
- [12] Corinna Cortes, Mehryar Mohri, and Andrés Muñoz Medina. Adaptation based on generalized discrepancy. *Journal of Machine Learning Research*, 20(1):1–30, 2019. URL <http://jmlr.org/papers/v20/15-192.html>.
- [13] Irina Degtiar and Sherri Rose. A review of generalizability and transportability. *Annual Review of Statistics and its Application*, 10:501–524, 2023.
- [14] AP Dempster, NM Laird, and DB Rubin. Maximal likelihood from incomplete data via the em algorithm. *J. Roy. Statist. Soc*, pages 1–38.
- [15] Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*, 2020.
- [16] Allan Donner. The relative effectiveness of procedures commonly used in multiple regression analysis for dealing with missing values. *The American Statistician*, 36(4):378–381, 1982.
- [17] Robin J Evans and Vanessa Didelez. Parameterizing and simulating from causal models. *arXiv preprint arXiv:2109.03694*, 2021.
- [18] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.
- [19] Nicolas Gisin, Grégoire Ribordy, Wolfgang Tittel, and Hugo Zbinden. Quantum cryptography. *Reviews of modern physics*, 74(1):145, 2002.

- [20] Luigi Gresele, Julius von Kügelgen, Jonas M. Kübler, Elke Kirschbaum, Bernhard Schölkopf, and Dominik Janzing. Causal inference through the structural causal marginal problem. 2 2022. URL <http://arxiv.org/abs/2202.01300>.
- [21] Siyuan Guo, Viktor Tóth, Bernhard Schölkopf, and Ferenc Huszár. Causal de finetti: On the identification of invariant causal structure in exchangeable data. *arXiv preprint arXiv:2203.15756*, 2022.
- [22] Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 2018.
- [23] Tony Hey, Stewart Tansley, and Kristin Tolle, editors. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, Redmond, Washington, 2009. ISBN 978-0-9825442-0-4. URL <https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/>.
- [24] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.
- [25] Judy Hoffman, Sergio Guadarrama, Eric S Tzeng, Ronghang Hu, Jeff Donahue, Ross Girshick, Trevor Darrell, and Kate Saenko. Lsda: Large scale detection through adaptation. *Advances in neural information processing systems*, 27, 2014.
- [26] Dominik Janzing. Merging joint distributions via causal model classes with low vc dimension, 2018. URL <https://arxiv.org/abs/1804.03206>.
- [27] Dominik Janzing and Bernhard Schölkopf. Causal inference using the algorithmic Markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010.
- [28] Ian Jolliffe. Principal component analysis. *Encyclopedia of statistics in behavioral science*, 2005.
- [29] Jae-On Kim and James Curry. The treatment of missing data in multivariate analysis. *Sociological Methods & Research*, 6(2):215–240, 1977.
- [30] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. 2010.
- [31] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021.
- [32] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019.
- [33] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015.
- [34] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. 3 conditional adversarial domain adaptation. 2018.
- [35] Chaochao Lu, Yuhuai Wu, José Miguel Hernández-Lobato, and Bernhard Schölkopf. Invariant causal representation learning for out-of-distribution generalization. In *International Conference on Learning Representations*, 2021.
- [36] Sergio Hernan Garrido Mejia, Elke Kirschbaum, and Dominik Janzing. Obtaining causal information by merging datasets with maxent. 7 2021. URL <http://arxiv.org/abs/2107.07640>.

- [37] James R. Norris. Probability and measure. 2013.
- [38] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1717–1724, 2014.
- [39] Judea Pearl. *Causality*. Cambridge university press, 2009.
- [40] Judea Pearl and Elias Bareinboim. Transportability across studies: A formal approach. Technical report, CALIFORNIA UNIV LOS ANGELES DEPT OF COMPUTER SCIENCE, 2011.
- [41] Judea Pearl and Elias Bareinboim. External validity: From do-calculus to transportability across populations. In *Probabilistic and causal inference: The works of Judea Pearl*, pages 451–482. 2022.
- [42] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 947–1012, 2016.
- [43] Niklas Pfister, Peter Bühlmann, and Jonas Peters. Invariant causal prediction for sequential data. *Journal of the American Statistical Association*, 114(527):1264–1276, 2019.
- [44] James M Robins. Association, causation, and marginal structural models. *Synthese*, 121(1/2): 151–179, 1999.
- [45] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning, 2018. URL <http://jmlr.org/papers/v19/16-432.html>.
- [46] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.
- [47] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. M. Mooij. On causal and anticausal learning. pages 1255–1262, 2012.
- [48] Bernhard Schölkopf. Causality for machine learning. *arXiv preprint 1911.10500, published: R. Dechter, J. Halpern, and H. Geffner. Probabilistic and Causal Inference: The Works of Judea Pearl*. ACM Books, 2019.
- [49] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Towards causal representation learning, 2021. URL <https://arxiv.org/abs/2102.11107>.
- [50] Martin G. Seneviratne, Michael G. Kahn, and Tina Hernandez-Boussard. Merging heterogeneous clinical data to enable knowledge discovery. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 24:439 – 443, 2018.
- [51] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.
- [52] Zheyang Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021.
- [53] Abe Sklar. Random variables, distribution functions, and copulas: a personal look backward and forward. *Lecture notes-monograph series*, pages 1–14, 1996.
- [54] Michael Spivak. *Calculus*. Publish or Perish, fourth edition, 2008.
- [55] R. Suter, D. Miladinovic, B. Schölkopf, and S. Bauer. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 6056–6065. PMLR, June 2019. URL <http://proceedings.mlr.press/v97/suter19a.html>.

- [56] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE international conference on computer vision*, pages 4068–4076, 2015.
- [57] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.
- [58] N. N. Vorob’ev. Consistent families of measures and their extensions. *Theory of Probability & its Applications*, 7(2):147–163, 1962. doi: 10.1137/1107014. URL <https://doi.org/10.1137/1107014>.
- [59] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.
- [60] Florian Wenzel, Andrea Dittadi, Peter Gehler, Carl-Johann Simon-Gabriel, Max Horn, Dominik Zietlow, David Kernert, Chris Russell, Thomas Brox, Bernt Schiele, Bernhard Schölkopf, and Francesco Locatello. Assaying out-of-distribution generalization in transfer learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 7181–7198. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/2f5acc925919209370a3af4eac5cad4a-Paper-Conference.pdf.
- [61] Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1): 1–25, 1982. ISSN 00129682, 14680262.
- [62] Jonas Wildberger, Siyuan Guo, Arnab Bhattacharyya, and Bernhard Schölkopf. On the interventional kullback-leibler divergence. *arXiv preprint arXiv:2302.05380*, 2023.
- [63] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.
- [64] Dinghuai Zhang, Kartik Ahuja, Yilun Xu, Yisen Wang, and Aaron Courville. Can subnetwork structure be the key to out-of-distribution generalization? In *International Conference on Machine Learning*, pages 12356–12367. PMLR, 2021.
- [65] K. Zhang, M. Gong, and B. Schölkopf. Multi-source domain adaptation: A causal view. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 3150–3157. AAAI Press, 2015. URL <http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/10052/9994>.
- [66] Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, and Zheyang Shen. Deep stable learning for out-of-distribution generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5372–5382, 2021.

A Random Variables

In probability measure theory, a random variable is formally defined as below. Here we inherit the notation and interpretation from [37].

Definition 4 (Random Variable) Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and (E, \mathcal{E}) be a measurable space. A measurable function $X : \Omega \rightarrow E$ is called a random variable in E .

Intuitively, one can consider the random variable as a function mapping the outcome of an experiment to some real number, where $E = \mathbb{R}$. The probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a way to model an experiment whose outcome is subject to chance, where one can interpret:

Ω is the set of possible outcomes, or *sample space*
 \mathcal{F} is the set of observable sets of outcomes, or *events*
 $\mathbb{P}(A)$ is the probability of an event A

The law or distribution of X is defined as $\mu_X = \mathbb{P} \circ X^{-1}$. For example, for some real-valued X , the distribution function of X is officially defined as: $F_X(x) = \mu_X((-\infty, x]) = \mathbb{P}(X \leq x)$.

Though well-defined as a mathematical object, it is unclear what does it mean for two random variables to be the same. It is known that two random variables can have identical distributions or further same probability space, but differs significantly. For example, suppose the experiment is to roll a dice twice. Let X_1 denotes the outcome of dice in the first roll and X_2 denotes the outcome of dice in the second roll. Though identically distributed, X_1 and X_2 are independent from each other.

The difference between statistical and physical perspective of random variables can be seen below. Consider X_1 and X_2 are blood pressures measured in different age groups. It is expected that they will have different distributions, as elders tend to have higher blood pressures. Mathematically, X_1 and X_2 represent different random variables though they both represent the same physical property - blood pressures. Under certain pathological cases, one can also construct two random variables that represent different physical meanings but have the same mathematical object. For example, a horse's body temperature is between 37.5 – 38.5 degree Celsius. Assume a population of horse that has body temperature uniformly distributed in this range. We know this temperature coincides with humans that have low-grade fever. Suppose further there is also a population of human patients that has body temperature uniformly distributed in this range. Random sampling from the two populations form random variables that share the same probability space though they represent temperature on different species.

In this work, we borrow well-established terminology from physics to specify what it means for two random variables are the same. Particularly, we restrict our attention to random variables that are observable [10], i.e., a physical property that can be measured. In quantum mechanics, the process that extracts information about the physical properties of a system is called measurement process. It is well-known that measurement process could affect the underlying physical properties of a system [19]. Thus, in this work, we consider two random variables are the same if they represent the same physical property of the same physical system and have the same measurement process.

Some examples:

1. Suppose in source environment, we observe blood pressured measured from elderly population. In the target environment, we observe blood pressured measured from the young. If the blood pressures measured share the same measurement process, we will consider the differences between environments as distribution shifts rather than *out-of-variable* scenarios.
2. Suppose a patient's first day's blood pressure (X_1) and oxygen saturation (X_2) constitutes as important predictors for the length of his/her stay in ICU (Y). If we artificially construct a third variable X_3 that is a linear combination of X_1 and X_2 , e.g. $X_3 = aX_1 + bX_2$, for some constant a, b . We note, either observing (X_1, X_2) or observing (X_2, X_3) allows the same prediction power to the target variable Y . Here we will consider them as *out-of-variable* scenarios since different variables can contain the same information.

B N  ive Model

Consider the structural causal model described in § 3.1 for ease of illustration. The method described below applies if we replace variable X_i to any subsets of variables. Consider the true generating function ϕ is additively composed with univariate functions about its covariates, i.e., there is no interaction term among different covariates. Then the structural causal model becomes:

$$Y := f_1(X_1) + f_2(X_2) + f_3(X_3) + \epsilon \quad (12)$$

where X_1, X_2, X_3 are jointly independent of each other and $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Further, f_1, f_2, f_3 are some unknown functions.

There are two environments:

- Source env S contains variables (X_1, X_2, Y)
- Target env T contains variables (X_2, X_3, Y)

Goal is to learn the most efficient way to transfer knowledge from env S to env T such that we can learn better predictor $\mathbb{E}[Y \mid X_2, X_3]$.

By the nature of no-interaction assumption, the solution is easy. One approach is to build separate neural networks for datasets (X_1, Y) and (X_2, Y) . With sufficient data, the learnt functions will be f_1, f_2 respectively due to independence between X_i . One solution to transfer is to migrate the learnt function f_2 to the target environment and then use data from the target environment to learn f_3 , i.e., functions that involve previously unobserved variable. Such solution will not hold if there is interaction between covariates, as function learnt is not an independent module with respect to the unobserved variable. Here, we highlight the ability to additive compose different modules relies on the assumption of no-interaction between unobserved variables and chosen variables.

C Proofs

C.1 Impossibility Theorem

Theorem 3.1 (Impossibility Theorem) *Consider out-of-variable scenarios illustrated in Fig. 2 (a)-(c). They are governed by the structural causal model described in § 3.1. Suppose all variables considered are continuous variables, except X_1, X_3 in Fig. 2c are binary. Assume 1) for all i , $p(x_i)$ is known and let the support set be $S_i := \{x \in \mathbb{R} \mid p_i(B(x, \epsilon)) > 0\}$ where p_i denotes the marginal distribution of x_i and $B(x, \epsilon)$ is the ball centred around x with radius ϵ . 2) Suppose for all i there exists two distinct points $x, x' \in S_i$. Given fixed f_S , for any function f_T that satisfies $\mathbb{E}_{X_1}[f_S(x_1, x_2)] = \mathbb{E}_{X_3}[f_T(x_2, x_3)]$, there always exists a different function f'_T such that it satisfies the above condition and for any chosen $R > 0$, $\|f_T - f'_T\|_2 \geq R$.*

Proof 1 *Proof by construction. Consider the scenario illustrated in Fig. 2b. Find two distinct values in the support set of $p(x_3)$, x'_3 and x''_3 . For some appropriate $\epsilon > 0$, consider their neighbourhoods as $N_1 = [x'_3 - \epsilon, x'_3 + \epsilon]$ and $N_2 = [x''_3 - \epsilon, x''_3 + \epsilon]$. Suppose we learnt the optimal predictive function in the source environment f_A , it can be written as:*

$$f_A(x_1, x_2) = \int_{\Omega} \phi(x_1, x_2, x_3) p(x_3) d(x_3) \quad (13)$$

$$= \underbrace{\int_{(\Omega \setminus N_1) \setminus N_2} \phi(x_1, x_2, x_3) p(x_3) d(x_3)}_{\text{Remainder}(x_1, x_2)} \quad (14)$$

$$+ \underbrace{\int_{N_1} \phi(x_1, x_2, x_3) p(x_3) d(x_3)}_{g(x_1, x_2)} + \underbrace{\int_{N_2} \phi(x_1, x_2, x_3) p(x_3) d(x_3)}_{h(x_1, x_2)} \quad (15)$$

We call the integration in the region excluding the specified neighbourhoods as $\text{Remainder}(x_1, x_2)$ and the integration over N_1 as $g(x_1, x_2)$ and that over N_2 as $h(x_1, x_2)$. Given any function $c(x_1, x_2)$, it is easy to find a function $d(x_1, x_2)$ such that $f_A(x_1, x_2) - \text{Remainder}(x_1, x_2) = c(x_1, x_2)g(x_1, x_2) +$

$d(x_1, x_2)h(x_1, x_2)$. This means whenever we find a function ϕ that satisfies Equation (13), it is always possible to slightly perturb ϕ such that ϕ' can also satisfy marginal consistency. For example, construct a ϕ' which is a result of proposed ϕ scaled by c elementwise over the neighbourhood N_1 and scaled by d elementwise over the neighbourhood N_2 . Moreover, the deviation of ϕ' with ϕ can be arbitrarily different:

$$\|\phi - \phi'\|_2 \geq \text{const}_1 \|c - 1\|_2 + \text{const}_2 \|d - 1\|_2 \quad (16)$$

which the lower bound can be arbitrarily large by the choice of $c(x_1, x_2)$.

Consider the scenario illustrated in Fig. 2a. Given the information obtained from the source environment $p(x_1), p(x_2), f_A(x_1, x_2)$, by argument above, we know it is always possible to perturb learnt ϕ appropriately to get ϕ' that satisfies the desired marginal consistency conditions. We will show it will also be impossible to identify the optimal predictive function f_B in the target environment. By the argument above, we can choose the function $c(x_1, x_2)$ freely. Then there always exists a function c such that $\text{sgn}(c(x_1, x_2)) = \text{sgn}(\phi(x_1, x_2, x'_3))$ and $|c(x_1, x_2)| \geq L, \forall x_1, x_2$, where $L > 1$. Consider a point x'_3 in the neighbourhood N_1 . Then under a learnt ϕ and perturbed ϕ' , its corresponding optimal predictive function f_B and f'_B can be written as:

$$f_B(x_2, x'_3) = \int_{\Omega} \phi(x_1, x_2, x'_3) p(x_1) d(x_1) \quad (17)$$

$$f'_B(x_2, x'_3) = \int_{\Omega} \underbrace{c(x_1, x_2) \phi(x_1, x_2, x'_3)}_{\geq 0 \text{ and } \neq \phi} p(x_1) d(x_1) \quad (18)$$

This implies $f_B(x_2, x'_3) \neq f'_B(x_2, x'_3)$ for all values in the neighbourhood N_1 . This means though f_B and f'_B are both marginally consistent with f_A (since ϕ and ϕ' are both consistent with f_A), but they are different functions. Moreover their difference can be arbitrarily large:

$$\|f_B - f'_B\|_2 \geq \int \int_{N_1} (f_B(x_2, x_3) - f'_B(x_2, x_3))^2 p(x_3) dx_3 p(x_2) dx_2 \quad (19)$$

If analyse the term that squared, we have:

$$(f_B(x_2, x_3) - f'_B(x_2, x_3))^2 = \left(\int_{\Omega} (c - 1) \phi(x_1, x_2, x_3) p(x_1) dx_1 \right)^2 \geq (|L| - 1)^2 \left(\int_{\Omega} \phi p(x_1) dx_1 \right)^2$$

The last inequality holds by construction of c . Thus substituting it into Eq. 19, we have $\|f_B - f'_B\|_2 \geq \text{const} * (|L| - 1)^2$, where the lower bound of the constructed function c can be arbitrarily large.

Consider the scenario illustrated in Fig. 2c. Here we restrict to cases when X_1 and X_3 are binary variables. Set $\gamma_i := P(X_i = 0)$. Then given source environments where one observes variables (X_1, X_2, Y) and the other observes variables (X_2, X_3, Y) . The potential generating function must satisfy below system of equations:

$$f_A(0, x_2) = \gamma_3 \phi(0, x_2, 0) + (1 - \gamma_3) \phi(0, x_2, 1) \quad (20)$$

$$f_A(1, x_2) = \gamma_3 \phi(1, x_2, 0) + (1 - \gamma_3) \phi(1, x_2, 1) \quad (21)$$

$$f_B(x_2, 0) = \gamma_1 \phi(0, x_2, 0) + (1 - \gamma_1) \phi(1, x_2, 0) \quad (22)$$

$$f_B(x_2, 1) = \gamma_1 \phi(0, x_2, 1) + (1 - \gamma_1) \phi(1, x_2, 1) \quad (23)$$

Perturb $\phi(0, x_2, 0)$ by $c(0, x_2, 0)$, then in order to still satisfy the above system of equations, the coefficients need to be correspondingly adjusted as:

$$c(0, x_2, 1) = \frac{f_A(0, x_2) - \gamma_3 c(0, x_2, 0) \phi(0, x_2, 0)}{(1 - \gamma_3) \phi(0, x_2, 1)} \quad (24)$$

$$c(1, x_2, 1) = \frac{f_B(x_2, 1) - \frac{\gamma_1 f_A(0, x_2) - \gamma_1 \gamma_3 c(0, x_2, 0) \phi(0, x_2, 0)}{(1 - \gamma_3)}}{(1 - \gamma_1) \phi(1, x_2, 1)} \quad (25)$$

$$c(1, x_2, 0) = \frac{f_A(1, x_2) - \frac{(1 - \gamma_3) f_B(x_2, 1) - \gamma_1 f_A(0, x_2) + \gamma_1 \gamma_3 c(0, x_2, 0) \phi(0, x_2, 0)}{1 - \gamma_1}}{\gamma_3 \phi(1, x_2, 0)} \quad (26)$$

Note we have the adjusted coefficients are consistent with each other:

$$(1 - \gamma_1)c(1, x_2, 0)\phi(1, x_2, 0) = \frac{1}{\gamma_3}[(1 - \gamma_1)f_A(1, x_2) - (1 - \gamma_3)f_B(x_2, 1)] \quad (27)$$

$$+ \gamma_1 f_A(0, x_2) - \gamma_1 \gamma_3 c(0, x_2, 0)\phi(0, x_2, 0)] \quad (28)$$

$$= f_B(x_2, 0) - \gamma_1 c(0, x_2, 0)\phi(0, x_2, 0) \quad (29)$$

Thus it is possible to find a new ϕ' such that it still satisfies the system of equations. More over ϕ' deviates from ϕ arbitrarily large by the choice of $c(0, x_2, 0)$:

$$\|\phi - \phi'\|_2 \geq \|\phi(0, x_2, 0) - c(0, x_2, 0)\phi(0, x_2, 0)\|_2 \geq \|c - 1\|_2 * \text{const} \quad (30)$$

C.2 Proof of possibility

Theorem 3.2 Consider a target variable Y and its corresponding causal parents \mathbf{PA}_Y . Let Z be a common child of all variables in \mathbf{PA}_Y and it satisfies $Z \perp\!\!\!\perp Y \mid \mathbf{PA}_Y$. Given:

- Environment S_1 contains variables (Z, Y) and $f_{S_1}(z) = \mathbb{E}[Y \mid Z]$,
- Environment S_2 contains variables (\mathbf{PA}_Y, Z) and $f_{S_2}(\mathbf{pa}_Y) = \mathbb{E}[Y \mid \mathbf{PA}_Y]$

Assume the SCM is generated by an additive noise model and f_{S_2} is an invertible function. With small noise in the functional relationship between Z and \mathbf{PA}_Y , composing $f_{S_1} \cdot f_{S_2}$ identifies the optimal predictive function in the target environment that contains variables never jointly observed previously, i.e., (\mathbf{PA}_Y, Y) . Fig. 3 in Appendix C.2 shows an example of such scenario where $Z = X_1$ and $\mathbf{PA}_Y = X_2$.

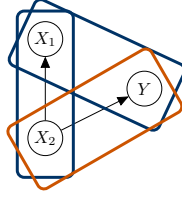


Figure 3: An example of scenarios considered in Theorem 3.2

Proof 2 The scenario illustrated satisfy

$$Y := \phi(\mathbf{PA}_Y) + \epsilon_Y \quad (31)$$

$$Z := g(\mathbf{PA}_Y) + \epsilon_1 \quad (32)$$

$$X_2 := U_2 \quad (33)$$

where ϕ and g are some reasonably continuous functions and g is invertible, $\epsilon_Y \sim \mathcal{N}(0, \sigma_Y^2)$ and $\epsilon_1 \sim \mathcal{N}(0, \sigma_1^2)$. The SCM is illustrated by the DAG in Fig. 3.

The optimal predictive function in the target environment coincides with the true generating function, written as: $f_B(\mathbf{PA}_Y) = \mathbb{E}[Y \mid \mathbf{PA}_Y] = \phi(\mathbf{PA}_Y)$. Given g is continuous and invertible, we have g^{-1} exists and is continuous. Since $\mathbf{pa}_Y = g^{-1}(z - \epsilon_1)$, given ϵ_1 is small enough, by continuity in g^{-1} , approximate \mathbf{pa}_Y as $g^{-1}(z)$. By $Z \perp\!\!\!\perp Y \mid \mathbf{PA}_Y$, we have

$$p(y \mid z) = \int p(y \mid \mathbf{pa}_Y, z) p(\mathbf{pa}_Y \mid z) d\mathbf{pa}_Y \quad (34)$$

$$= \int p(y \mid \mathbf{pa}_Y) p(\mathbf{pa}_Y \mid z) d\mathbf{pa}_Y \quad (35)$$

Performing point estimation over the integral where we assume $p(\mathbf{pa}_Y \mid z) = \delta(\mathbf{pa}_Y = g^{-1}(z))$, we have $p(y \mid z) = \mathcal{N}(y \mid \phi \cdot g^{-1}(z), \sigma_Y^2)$ and $f_{S_2}(z) = \mathbb{E}[Y \mid z] = \phi \cdot g^{-1}(z)$. This implies, we can recover the optimal predictive function in the target environment by composing learnt functions from the two source environments, where $f_B = \phi = f_{S_1} \cdot f_{S_2}$.

C.3 Proof of Theorem 3.1

Theorem 3.1 Consider the problem setup in § 3.1. Assume the generating function ϕ is 2-times differentiable with respect to the third covariate everywhere and source environment S contains variables (X_1, X_2, Y) with $f_S(x_1, x_2) = \mathbb{E}[Y \mid x_1, x_2]$. For fixed x_1, x_2 , approximate the function $\phi : x_1 \times x_2 \times \mathcal{X}_3 \rightarrow \mathbb{R}$ with first-order Taylor approximation, we have:

$$\mathbb{E}[(Y - f_S(x_1, x_2))^n \mid x_1, x_2] = \sum_{k=0}^n \binom{n}{k} \mathbb{E}[\epsilon^k] \left(\frac{\partial \phi}{\partial X_3} \Big|_{x_1, x_2, \mu_3} \right)^{n-k} \mathbb{E}[(X_3 - \mu_3)^{n-k}] \quad (9)$$

When $n = 3$:

$$\mathbb{E}[(Y - f_S(x_1, x_2))^3 \mid x_1, x_2] = \left(\frac{\partial \phi}{\partial X_3} \Big|_{x_1, x_2, \mu_3} \right)^3 \mathbb{E}[(X_3 - \mu_3)^3] + \mathbb{E}[\epsilon^3] \quad (10)$$

Theorem C.1 (Taylor's Theorem [54]) Let $k \geq 1$ be an integer and let the function $f : \mathbb{R} \rightarrow \mathbb{R}$ be $k + 1$ -times differentiable everywhere and $f^{(k)}$ is continuous on the closed interval $[a, x]$. Then there exists a point $\xi \in [a, x]$ such that

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(k)}(a)}{k!}(x-a)^k + \underbrace{\frac{f^{(k+1)}(\xi)}{(k+1)!}(x-a)^{k+1}}_{:=R_{k+1}} \quad (36)$$

Proof 3 Let $\mu_x = \mathbb{E}[X]$, $\sigma_x^2 = \text{Var}[X]$, $k_x = \text{Skew}[X]$ and $a = \mu_x$, then Theorem C.1 states that

$$f(x) = f(\mu_x) + f'(\mu_x)(x - \mu_x) + C(x - \mu_x)^2 \quad (37)$$

Take first-order Taylor approximation over the generating function, we suppose $f(x) \approx f(\mu_x) + f'(\mu_x)(x - \mu_x)$. Taking expectations, then

$$\mathbb{E}[f(x)] \approx f(\mu_x) \quad (38)$$

Consider the residuals of $f(x)$ without its expectations and raise it to the power of n , we have:

$$\left(f(x) - \mathbb{E}[f(x)] \right)^n = \left(f'(\mu_x)(x - \mu_x) \right)^n \quad (39)$$

Taking expectations, on Eq. 39, we have:

$$\mathbb{E} \left[\left(f(x) - \mathbb{E}[f(x)] \right)^n \right] = (f'(\mu_x))^n \mathbb{E}[(x - \mu_x)^n] \quad (40)$$

Take $f_{x_1, x_2} : \mathcal{X}_3 \rightarrow \mathbb{R}$ to be the function $\phi : x_1 \times x_2 \times \mathcal{X}_3 \rightarrow \mathbb{R}$ where values x_1, x_2 are fixed. We can substitute f as f_{x_1, x_2} to Eq. 40. Since $Y = \phi(x_1, x_2, x_3) + \epsilon$, we have

$$\mathbb{E}[(Y - f_S(x_1, x_2))^n \mid x_1, x_2] = \mathbb{E} \left[\left(f(x) + \epsilon - \mathbb{E}[f(x)] \right)^n \right] \quad (41)$$

$$= \sum_{k=0}^n \binom{n}{k} \mathbb{E}[\epsilon^k] \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^{n-k}] \quad (42)$$

The second equation is due to independence of ϵ and $f(x) - \mathbb{E}[f(x)]$. This is well known that $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ if X and Y are independent from each other.

D Extensions

D.1 Higher-order Taylor Approximations

Appendix C.3 shows the proof of Theorem 3.1 where we approximated the function $\phi : x_1 \times x_2 \times \mathcal{X}_3 \rightarrow \mathbb{R}$ with first-order Taylor approximations. Here we demonstrate its flexibility to extend to higher-order Taylor approximations, e.g. second-order and for ease of illustration and computation, we focus our work on first-order Taylor approximations. Let $\mu_x = \mathbb{E}[X]$, $\sigma_x^2 = \text{Var}[X]$, $k_x = \text{Skew}[X]$

and $a = \mu_x$. Theorem C.1 states that with second order approximations, the function can be written as:

$$f(x) = f(\mu_x) + f'(\mu_x)(x - \mu_x) + f''(\mu_x)(x - \mu_x)^2 + C(x - \mu_x)^3 \quad (43)$$

The moment of the residuals of $f(x)$ is expressed as:

$$\left(f(x) - \mathbb{E}[f(x)]\right)^n = \left(f'(\mu_x)(x - \mu_x) + f''(\mu_x)((x - \mu_x)^2 - \sigma_x^2)\right)^n \quad (44)$$

When $n = 2$ and $n = 3$:

$$\begin{aligned} \left(f(x) - \mathbb{E}[f(x)]\right)^2 &= f'(\mu_x)^2(x - \mu_x)^2 + 2f'(\mu_x)f''(\mu_x)(x - \mu_x)((x - \mu_x)^2 - \sigma_x^2) \\ &\quad + f''(\mu_x)^2((x - \mu_x)^2 - \sigma_x^2)^2 \\ \left(f(x) - \mathbb{E}[f(x)]\right)^3 &= f'(\mu_x)^3(x - \mu_x)^3 + 3f'(\mu_x)^2f''(\mu_x)(x - \mu_x)^2((x - \mu_x)^2 - \sigma_x^2) \\ &\quad + 3f'(\mu_x)f''(\mu_x)^2(x - \mu_x)((x - \mu_x)^2 - \sigma_x^2)^2 + f''(\mu_x)^3((x - \mu_x)^2 - \sigma_x^2)^3 \end{aligned}$$

Taking expectations on both sides:

$$\mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2] = f'(\mu_x)^2\sigma_x^2 + 2f'(\mu_x)f''(\mu_x)k_x + f''(\mu_x)^2(\mathbb{E}[(x - \mu_x)^4] - \sigma_x^4) \quad (45)$$

$$\mathbb{E}[(f(x) - \mathbb{E}[f(x)])^3] = f'(\mu_x)^3k_x + 3f'(\mu_x)^2f''(\mu_x)(\mathbb{E}[(x - \mu_x)^4] - \sigma_x^4) \quad (46)$$

$$+ 3f'(\mu_x)f''(\mu_x)^2(\mathbb{E}[(x - \mu_x)^5] - 2\sigma_x^2k_x) \quad (47)$$

$$+ f''(\mu_x)^3(\mathbb{E}[(x - \mu_x)^6] - 3\sigma_x^2\mathbb{E}[(x - \mu_x)^4] + 2\sigma_x^6) \quad (48)$$

Take $f_{x_1, x_2} : \mathcal{X}_3 \rightarrow \mathbb{R}$ to be the function $\phi : x_1 \times x_2 \times \mathcal{X}_3 \rightarrow \mathbb{R}$ where values x_1, x_2 are fixed. We can substitute f as f_{x_1, x_2} . Recall when $n = 2$ and $n = 3$ in Theorem 3.1,

$$\mathbb{E}[(Y - f_S(x_1, x_2))^2 | x_1, x_2] = \mathbb{E}[(\phi(x_1, x_2, X_3) - f_S(x_1, x_2))^2 | x_1, x_2] + \mathbb{E}[\epsilon^2] \quad (49)$$

$$\mathbb{E}[(Y - f_S(x_1, x_2))^3 | x_1, x_2] = \mathbb{E}[(\phi(x_1, x_2, X_3) - f_S(x_1, x_2))^3 | x_1, x_2] + \mathbb{E}[\epsilon^3] \quad (50)$$

Assume ϵ has small variance and negligible skew and given we can estimate the higher moments of variable X_3 , with two equations and two unknowns, we are able to estimate the unknowns, namely

$$\frac{\partial \phi}{\partial X_3} \Big|_{x_1, x_2, \mu_3} \text{ and } \frac{\partial^2 \phi}{\partial^2 X_3} \Big|_{x_1, x_2, \mu_3}.$$

D.2 More than one unobserved variables

Here, we consider a function with two variables $f(x, y)$ where variables can be considered as two unobserved variables from the source environment. Note, the same argument can easily extend to multivariate functions. Let $\mathbb{E}[X] = \mu_x, \mathbb{E}[Y] = \mu_y$. Expand multivariate Taylor approximations around the point $\mathbf{a} = (\mu_x, \mu_y)$, we have:

$$f(x, y) = f(\mu_x, \mu_y) + \frac{\partial f}{\partial x} \Big|_{\mathbf{a}}(x - \mu_x) + \frac{\partial f}{\partial y} \Big|_{\mathbf{a}}(y - \mu_y) \quad (51)$$

$$+ C_1(x - \mu_x)^2 + C_2(x - \mu_x)(y - \mu_y) + C_3(y - \mu_y)^3 \quad (52)$$

With first-order Taylor approximations, we ignore the higher order terms. Taking expectations on both sides, we have $\mathbb{E}[f(x, y)] = f(\mu_x, \mu_y)$. Similarly,

$$(f(x, y) - \mathbb{E}[f(x, y)])^n = \left(\frac{\partial f}{\partial x} \Big|_{\mathbf{a}}(x - \mu_x) + \frac{\partial f}{\partial y} \Big|_{\mathbf{a}}(y - \mu_y)\right)^n \quad (53)$$

$$= \sum_{k=0}^n \binom{n}{k} \left(\frac{\partial f}{\partial x} \Big|_{\mathbf{a}}\right)^k (x - \mu_x)^k \left(\frac{\partial f}{\partial y} \Big|_{\mathbf{a}}\right)^{n-k} (y - \mu_y)^{n-k} \quad (54)$$

Taking expectations on both sides, assuming we can estimate the cross-moments between two unobserved variables from data, with two unknowns and two equations, we can estimate the unknowns. Note here we also relax the independence assumption among the unobserved variables.

E Algorithm

To estimate the target predictive function, one can consider it as the true generating function marginalized over the unobserved variable X_1 in environment T . The optimal predictive function in the target environment can be evaluated as:

$$f_T(x_2, x_3) = \int \phi(x_1, x_2, x_3)p(x_1)dx_1 \approx \frac{1}{n} \sum_{i=1}^n \phi(x_{1,i}, x_2, x_3), \quad \text{where } x_{1,i} \sim p(x_1)$$

If the true-generating function is known, the above term can be Monte Carlo estimated as environment S contains the marginal distribution of X_1 . Though identifying the generating function is hard, Theorem 3.1 shows that we can approximate it with first-degree Taylor approximation around its third variable:

$$\phi(x_1, x_2, x_3) = \phi(x_1, x_2, \mu_3) + \left. \frac{\partial \phi}{\partial X_3} \right|_{x_1, x_2, \mu_3} (x_3 - \mu_3) + C(x_3 - \mu_3)^2 \quad (55)$$

for some constant C . Taking expectations over X_3 on both sides, we have $f_S(x_1, x_2) = \phi(x_1, x_2, \mu_3) + C\sigma_3^2$. Approximating the first term in Eq. 55 with $f_S(x_1, x_2)$ and estimate the partial derivative term by Theorem 3.1 gives us Algorithm 1.

F Data Generation Details

We generate synthetic data according to the following equations:

$$\forall i \in [3] : X_i \stackrel{i.i.d}{\sim} \Gamma(1, 1), \tilde{X}_i \leftarrow \text{standardize}(X_i) \quad (56)$$

$$\forall j \in [7] : \alpha_j \stackrel{i.i.d}{\sim} \mathcal{N}(0, 1) \quad (57)$$

$$Y_{poly} = \alpha^T g_{poly}(\tilde{X}_1, \tilde{X}_2, \tilde{X}_3) + \epsilon_1 \quad (58)$$

$$Y_{nonlinear} = \sqrt{(\alpha_1 \tilde{X}_1)^2 + (\alpha_2 \tilde{X}_2)^2 + (\alpha_3 \tilde{X}_3)^2} + \epsilon_2 \quad (59)$$

$$Y_{trig} = \alpha_1 \cos(\tilde{X}_1) + \alpha_2 \sin(\tilde{X}_2) + \alpha_3 \cos(\tilde{X}_3) + \alpha_4 \cos(\tilde{X}_1 \tilde{X}_3) + \epsilon_3 \quad (60)$$

where $g_{poly}(X_1, X_2, X_3) = [X_1, X_2, X_3, X_1 X_2, X_1 X_3, X_2 X_3, X_1 X_2 X_3]$ and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ for all i and $\sigma = 0.1$. The covariates are standardized into range between 0 and 5.