Learning Joint Interventional Effects from Single-Variable Interventions in Additive Models

Armin Kekić [1] **Sergio Hernan Garrido Mejia** [1,2] **Bernhard Schölkopf** [1]

[1] Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2] Amazon Research
{armin.kekic,shgm,bs}@tue.mpg.de

Abstract

Estimating joint causal effects is crucial in many domains, but obtaining data from multiple simultaneous interventions can be challenging. Our study explores how to learn joint interventional effects using only observational data and single-variable interventions. We present an identifiability result for this problem, showing that for a class of nonlinear additive outcome mechanisms, joint effects can be inferred without access to joint interventional data. We propose a practical estimator that decomposes the causal effect into confounded and unconfounded contributions for each intervention variable. Experiments on synthetic data demonstrate that our method achieves performance comparable to models trained directly on joint interventional data, outperforming a purely observational estimator.

1 Introduction

Understanding the effects of interventions is fundamental across many domains, from designing public health policies to optimizing business operations and administering medical treatments. Particularly challenging and important are joint interventional effects, where simultaneous interventions on multiple action variables influence a target outcome. Such scenarios are common in epidemiology [1], e-commerce [2, 3], and medicine [4].

Such joint effects can be estimated from observational data, assuming a known causal structure and that all variables necessary for identification are observed [5]. However, real-world scenarios frequently involve unobserved confounding factors, introducing biases that render observational models unreliable. The gold standard approach is to run fully randomized experiments. However, this can be challenging for ethical or practical reasons, particularly as the number of intervention settings grows combinatorially with the number of intervenable variables.

Our work addresses a middle ground between these two approaches: learning joint causal effects using observational and single-intervention data, where only one variable is intervened upon at a time. This is an instance of the *Intervention Generalization Problem* [6]: predicting treatment effects in previously unseen interventional settings. Causal models encode additional structural relationships between variables that allow us to generalize to settings in which non-causal machine learning approaches assuming independent, identically distributed (i.i.d.) data fail.

However, this problem setting is not solvable in its most general form and in order to achieve Intervention Generalization, we need to restrict the causal model class [7]. In this study, we focus on causal models with real-valued variables, where each action contributes to the outcome variable in a nonlinear way and is subject to confounding. We assume that these complex individual effects combine additively to produce the outcome. Within this model class, we show that the joint interventional effect is identifiable from observational and single-intervention data using an estimator that decomposes the causal effect into confounded and unconfounded contributions for each intervention variable.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

2 Problem Statement

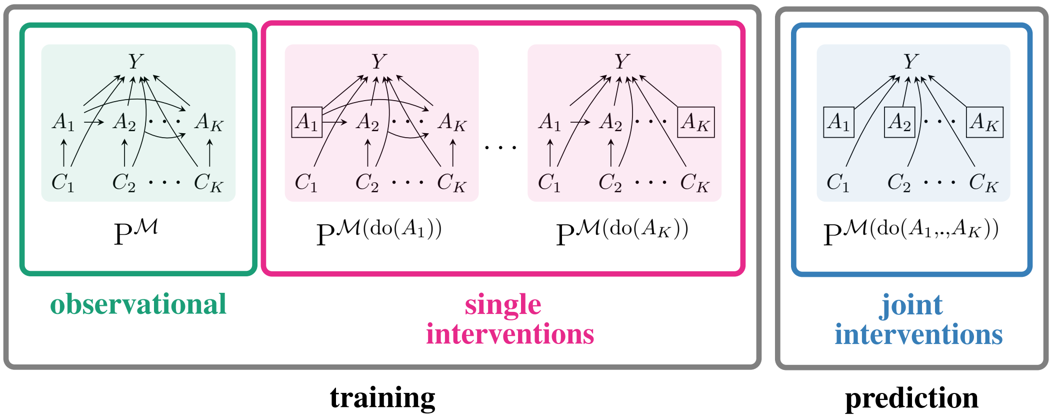

Y
A_1
A_2
A_K
C_1
C_2
C_K
$\mathrm{P}^{\mathcal{M}}$
$\mathrm{P}^{\mathcal{M}(\mathrm{do}(A_1))}$
$\mathrm{P}^{\mathcal{M}(\mathrm{do}(A_K))}$
$\mathrm{P}^{\mathcal{M}(\mathrm{do}(A_1,..,A_K))}$
observational
single interventions
joint interventions
training
prediction

Figure 1: **The Intervention Generalization Problem.** The figure shows the different interventional regimes. Our goal is to estimate the joint interventional effect of the action variables $\{A_1, \dots, A_K\}$ on Y, that is, $\mathbb{E}[Y \mid \mathrm{do}(A_1, \dots, A_K)]$ (right). However, during training we only have access to observational (left) and single-interventional data (middle). There are unobserved counfounders $\{C_1, \dots, C_K\}$ between the actions and the outcome variable Y. A box around a variable indicates that it was intervened on.

2.1 Background and Notation

We use boldface for vector-valued or sets of random variables. We denote random variables with capital letters and realizations thereof in lowercase.

Definition 1 (SCM [5, 8])**.** An N-dimensional structural causal model is a triplet $\mathcal{M} = (\mathcal{G}, \mathbb{S}, P_{\mathbf{U}})$ consisting of:

- a joint distribution $\mathrm{p}_{\mathbf{U}}$ over the jointly independent "exogenous" variables $\mathbf{U} = \{U_1, \dots, U_N\}$,
- a directed acyclic graph $\mathcal{G}$ with N vertices,
- a set $\mathbb{S} = \{X_j := f_j(\mathbf{Pa}_j, U_j), j = 1, \dots, N\}$ of structural assignments, where each f_j is a scalar-valued function and $\mathbf{Pa}_j$ are the variables indexed by the set of parents of node j in $\mathcal{G}$,

such that for every $\mathbf{u}$, the system $\{x_j := f_j(\mathbf{pa}_j, u_j)\}$ has a unique solution. The SCM thus entails a joint distribution over the "endogenous" random variables $\mathbf{X} = \{X_1, \dots, X_N\}$.

Interventions in SCMs. Interventions on one or more endogenous variables in SCMs are encoded by replacing the corresponding structural assignments. The $\mathrm{do}(\cdot)$-operator represents perfect interventions. For example, applying $\mathrm{do}(X_i{=}x_i)$ describes an intervention where the corresponding structural assignment is replaced by a constant x_i, leading to an interventional model $\mathcal{M}(\mathrm{do}(X_i{=}x_i))$, or $\mathcal{M}(\mathrm{do}(x_i))$ for short. In general, $\mathcal{M}(\mathrm{do}(x_i))$ entails a different distribution over the endogenous variables, which we represent through a superscript, that is $\mathrm{p}^{\mathcal{M}(\mathrm{do}(x_i))}$. When there is no superscript, the distribution is assumed to come from the unintervened or observational model $\mathcal{M}$. In conditioning sets, $\{x_1, \dots, \mathrm{do}(x_i), \dots, x_n\}$ denotes a set where X_i was intervened on and all other variables have observational realizations.

2.2 Setting

Let $\mathbf{A} = \{A_1, \dots, A_K\}$ be a set of treatment or action variables, $\mathbf{C} = \{C_1, \dots, C_K\}$ be a set of unobserved confounders and Y be an outcome variable. The actions are direct causes of the outcome, and we allow for an arbitrary acyclic causal structure among the actions. For notational simplicity, we assume the actions are in topological order and write the causal structure as a fully connected DAG.[1]

[1] This means that each variable A_k can potentially depend on all preceding variables $\{A_1, \dots, A_{k-1}\}$.

We can write the structural assignments as follows:

$$\begin{aligned} Y &:= f(A_1, \dots, A_K, C_1, \dots, C_K, U) && && (1) \\ A_k &:= g_k(A_1, \dots, A_{k-1}, C_1, \dots, C_{k-1}, V_k) && \text{for } k \in \{1, \dots, K\} && (2) \\ C_k &:= W_k && \text{for } k \in \{1, \dots, K\} && (3) \end{aligned}$$

where $\{U, V_1, \dots, V_K, W_1, \dots, W_K\}$ are mutually independent exogenous noise variables.

We are given a dataset of observational i.i.d. samples

$$\mathrm{D}_{\text{obs}} \sim \mathrm{P}^{\mathcal{M}}_{(Y,\mathbf{A})} \tag{4}$$

and K datasets of i.i.d. samples for single-variable interventions on each action variable

$$\{\mathrm{D}^k_{\text{int}} \sim \mathrm{P}^{\mathcal{M}_{\text{do}(A_k)}}_{(Y,\mathbf{A})}\}_{k=1}^K \, . \tag{5}$$

Our objective is to estimate the *joint interventional effect*

$$\mathbb{E}\left[Y \mid \text{do}(A_1, \dots, A_K)\right] \, . \tag{6}$$

2.3 General Non-Identifiability

In general, the setting described in Section 2.2 is not identifiable. In this section, we show that two distinct SCMs can induce identical observational distributions and single-variable interventional distributions, but exhibit different behaviors under joint interventions.

Example 1. Consider the following two SCMs over binary variables:

$$\begin{aligned} \mathcal{M}: \quad Y &:= A_1 \wedge A_2 \wedge C \wedge U & \qquad \widetilde{\mathcal{M}}: \quad Y &:= A_2 \wedge C \wedge U \\ A_2 &:= A_1 \wedge C \wedge V_2 & A_2 &:= A_1 \wedge C \wedge V_2 \\ A_1 &:= C & A_1 &:= C \\ C &:= W & C &:= W \end{aligned}$$

where $U, V_2, W{\sim}\text{Bernoulli}(p)$, with $0{<}p{<}1$. These two models induce the same observational distribution over the observed variables; that is, $\mathrm{P}^{\mathcal{M}}(Y, A_1, A_2) = \mathrm{P}^{\widetilde{\mathcal{M}}}(Y, A_1, A_2)$. They also lead to the same single-variable interventional distributions; $\mathrm{P}^{\mathcal{M}(\text{do}(A_1))}(Y, A_1, A_2) = \mathrm{P}^{\widetilde{\mathcal{M}}(\text{do}(A_1))}(Y, A_1, A_2)$ and $\mathrm{P}^{\mathcal{M}(\text{do}(A_2))}(Y, A_1, A_2) = \mathrm{P}^{\widetilde{\mathcal{M}}(\text{do}(A_2))}(Y, A_1, A_2)$.[2] However, they induce different distributions when A_1 and A_2 are jointly intervened:

$$\mathrm{P}^{\mathcal{M}(\text{do}(A_1=0, A_2=1))}(Y{=}1) = 0 \neq p = \mathrm{P}^{\widetilde{\mathcal{M}}(\text{do}(A_1=0, A_2=1))}(Y{=}1) \, . \tag{7}$$

This example demonstrates the need for additional assumptions on the ground-truth SCM to identify joint interventional effects from single variable interventions.

2.4 Assumptions

Assumption 1. The variables are continuous.

Assumption 2 (Intervention Support)**.** The distributions of the action variables have identical support across all interventional regimes. That is,

$$\text{supp}_{\mathrm{P}^{\mathcal{M}}_{(\mathbf{A})}}(\mathbf{A}) = \text{supp}_{\mathrm{P}^{\mathcal{M}_{\text{do}(A_1,.,A_K)}}_{(\mathbf{A})}}(\mathbf{A}) = \text{supp}_{\mathrm{P}^{\mathcal{M}_{\text{do}(A_k)}}_{(\mathbf{A})}}(\mathbf{A}) \quad \text{for any } k \in \{1, \dots, K\} \, . \tag{8}$$

Assumption 3 (Additive Outcome Mechanism)**.** There is pair-wise confounding between the actions and the outcome. The outcome is generated by an additive combination of separate nonlinear functions for each action and its associated confounder. The structural assignments can be written as:

$$\begin{aligned} Y &:= \textstyle\sum_{k=1}^K f_k(A_k, C_k) + U && && (9) \\ A_k &:= g_k(A_1, \dots, A_{k-1}, C_k, V_k) && \text{for } k \in \{1, \dots, K\} && (10) \\ C_k &:= W_k && \text{for } k \in \{1, \dots, K\} && (11) \end{aligned}$$

where $\{U, V_1, \dots, V_K, W_1, \dots, W_K\}$ are mutually independent exogenous noise variables.

[2] These probability distributions are shown in Tables 1 to 3 in Appendix D.

For the outcome mechanism (9), we assume that both the contributions of each action-confounder pair (A_k, C_k) and the exogenous noise U are additive. The latter is an assumption which has gained popularity in causal structure learning [9]. Note that for the other structural assignments we do not introduce such constraints. The resulting causal structure is illustrated in Figure 1 (left).

3 Identifiability

In this section, we show that in the causal model class with an additive outcome mechanism, outlined in Section 2.4, we can achieve Intervention Generalization.

Theorem 1 (Identifiability)**.** Under the assumptions in Section 2.4, the joint interventional effect (6) is identifiable from single-variable interventions and observational data in the infinite data regime.

Proof Sketch We first note that we can decompose the joint interventional effect (6) which we want to estimate, as well as the conditional expectations, for which we have data, as

$$\mathbb{E}[Y \mid \mathrm{do}(a_1, \dots, a_K)] = \textstyle\sum_k \mathbb{E}_{C_k \sim \mathrm{p}(C_k)} \left[f_k(a_k, C_k)\right] \tag{12}$$

$$\begin{aligned}\mathbb{E}[Y \mid a_1, \dots, \mathrm{do}(a_j), \dots, a_K] = {} & \mathbb{E}_{C_j \sim \mathrm{p}(C_j)}[f_j(a_j, C_j)] \\ & + \textstyle\sum_{k \neq j} \mathbb{E}_{C_k \sim \mathrm{p}(C_k \mid a_1,.,a_K)}[f_k(a_k, C_k)] \quad \text{for } j \in \{1, \dots, K\}\end{aligned} \tag{13}$$

$$\mathbb{E}[Y \mid a_1, \dots, a_K] = \textstyle\sum_k \mathbb{E}_{C_k \sim \mathrm{p}(C_k \mid a_1,.,a_K)}[f_k(a_k, C_k)]\,. \tag{14}$$

These decompositions correspond to the terms f_k in the additive outcome mechanism (9). In each term, the expectation over the confounder C_k is taken with respect to a measure that depends on whether the corresponding action variable A_k was intervened on. When A_k is not intervened on, the confounding can introduce an additional dependence on the other action variables, entangling their influences.

However, these decompositions still allow us to learn a representation that enables us to generalize from the observational and single-interventional setting to the joint-interventional effect. We define K estimator functions

$$\hat{f}_k(a_1, \dots, a_K, R_k), \qquad k \in \{1, \dots, K\}\,, \tag{15}$$

where $R_k \in \{0, 1\}$ indicates an intervention on A_k. Each R_k can be thought of as selecting one of two functions

$$\hat{f}_k(a_1, \dots, a_K, R_k) = \begin{cases} \hat{f}_k^{\mathrm{obs}}(a_1, \dots, a_K) & \text{if } R_k = 0 \\ \hat{f}_k^{\mathrm{int}}(a_1, \dots, a_K) & \text{if } R_k = 1 \end{cases} \tag{16}$$

where $\hat{f}_k^{\mathrm{int}}$ represent the terms in the decompositions (12)–(14) where the corresponding action A_k is intervened on, and $\hat{f}_k^{\mathrm{obs}}$ are the factors with A_k observational. We then define an overall estimator

$$\hat{f}(a_1, \dots, a_K, R_1, \dots, R_K) = \textstyle\sum_{k=1}^{K} \hat{f}_k(a_1, \dots, a_K, R_k) \tag{17}$$

to represent all regimes, depending on the setting of the indicator variables $R_1, \dots, R_K$:

- When $R_1{=}1, \dots, R_K{=}1$, the function $\hat{f}$ is an estimator for the joint interventional regime.
- $R_1{=}0, \dots, R_j{=}0, \dots, R_K{=}1$ corresponds to the single-interventional setting of $\mathcal{M}(\mathrm{do}(a_j))$.
- $R_1{=}0, \dots, R_K{=}0$ is the observational setting.

In the outcome mechanism (9) each term f_k depends only on one action A_k.[3] In contrast, each model factor $\hat{f}_k$ has to take all actions into account due to the entanglement introduced through confounding.

Now if we fit the estimator $\hat{f}$ in the observational and the single-interventional regimes, that is,

$$\hat{f}(a_1, \dots, a_K, R_1{=}0, \dots, R_K{=}0) = \mathbb{E}[Y \mid a_1, \dots, a_K] \tag{18}$$

$$\hat{f}(a_1, \dots, a_K, R_1{=}0, \dots, R_j{=}1, \dots, R_K{=}0) = \mathbb{E}[Y \mid a_1, \dots, \mathrm{do}(a_j), \dots, a_K], \quad \text{for } j \in \{1, \dots, K\}\,, \tag{19}$$

we can show that the estimator also identifies the joint interventional effect:

$$\hat{f}(a_1, \dots, a_K, R_1{=}1, \dots, R_K{=}1) = \mathbb{E}[Y \mid \mathrm{do}(a_1, \dots, a_K)]\,. \tag{20}$$

The full proof is shown in Appendix A. ■

[3] f_k also depends on the confounder C_k.

Note that, while our approach assumes that the action variables are direct causes of the outcome and that there is no confounding between the actions, we do not assume a particular causal structure between the actions. The approach we present here is agnostic to causal relationships among the actions and does not use this information to infer the joint interventional effect (6).

Moreover, we can extend our results to any combination of intervened and observational actions:

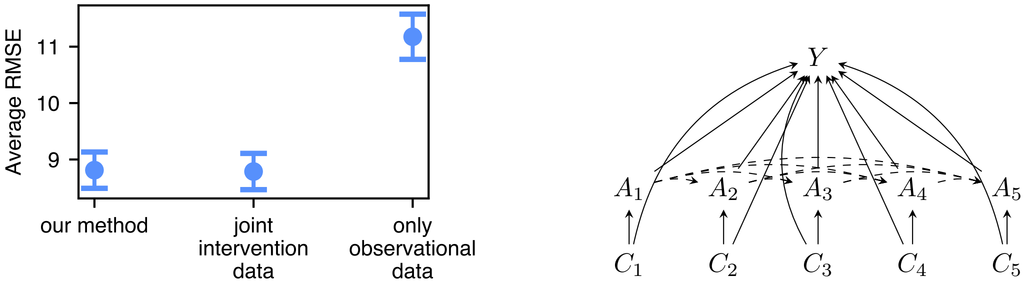
Proposition 1 (Identifiability of Mixed Interventional Effects)**.** Let $\mathbf{A}_{\text{int}} \cup \mathbf{A}_{\text{obs}} = \{A_1, \dots, A_K\}$ be a partition of the action variables into intervened and observational actions. Under the assumptions in Section 2.4, the effect

$$\mathbb{E}[Y \mid \text{do}(\mathbf{a}_{\text{int}}), \mathbf{a}_{\text{obs}}] \tag{21}$$

is identifiable from single-variable interventions and observational data in the infinite data regime. The proof is given in Appendix B.

Theorem 1 implies that for an additive outcome mechanism (1), the number of interventional datasets required for identification of the joint effect (6) grows only linearly with the number of actions. In Appendix C, we show that even when (1) is only additive with respect to the effect of subsets of actions, identification is possible as long as we have joint interventional data on each subset.

4 Experiments



(a)

(b)

Figure 2: **Experiments on Synthetic Data**. (a) Average root mean squared error (RMSE) for predicting the joint interventional effect $\mathbb{E}[Y \mid \text{do}(A_1, \dots, A_5)]$, averaged over 100 experiment runs. Each run uses a randomly generated ground truth SCM. We compare three approaches: (i) Our Intervention Generalization method, training the estimator (17) on observational and single-intervention data (Section 3). (ii) An estimator trained directly on joint interventional data (topline). (iii) An estimator trained solely on observational data. The error bars show the standard error of the mean. (b) Causal graph structure of the ground truth latent SCMs. Dashed edges between actions represent probabilistic dependencies that may or may not exist in each sampled SCM.

Synthetic data-generating process. We sample a structural causal model with five actions and confounders and causal relationships as shown in Figure 2b. The structural assignments are linear, and the exogenous noises are lognormal. The corresponding parameters are drawn at random before each experiment run. The dependencies between actions are probabilistic, with each potential edge having a probability p_{edge} of being active. We sample 100 SCMs, where for each run we sample $100,000$ data points for the observational, the single-interventional- and joint-interventional datasets. We split each dataset into 80% training- and 20% test data. Further details about the experimental setup are given in Appendix E.

Models and benchmarks. We train linear estimator functions (17) as outlined in Section 3. We compare that model to two baselines. (i) A linear model that is directly trained on joint interventional data. That is, we directly fit $\mathbb{E}[Y \mid \text{do}(A_1, \dots, A_5)]$. This comparison represents the minimal error that our method could achieve. (ii) A linear model that only considers the observational data. That is, we use a model fit to $\mathbb{E}[Y \mid A_1, \dots, A_5]$ to predict $\mathbb{E}[Y \mid \text{do}(A_1, \dots, A_5)]$. This is a typical approximation made in the absence of interventional data in real-world applications [2].[4]

[4] The additional error incurred through making this simplifying assumption quantifies the *Causal Risk* [10].

Results. The mean root mean squared error (RMSE) over all sampled SCMs in each of the three settings is shown in Figure 2a. We observe that our method achieves a similarly low error as the minimally achievable error of the topline model that was trained on joint interventional data. Both our approach and the topline benchmark significantly outperform the naive observational-only model. The results empirically validate the effectiveness of our Intervention Generalization technique in leveraging single-intervention data to predict joint interventional effects.

5 Related Work

The most closely related prior work is that of Saengkyongam and Silva [7]. They show that generalization from single-intervention and observational data to the joint interventional effect with continuous variables is possible. Their set of assumptions is more restrictive in some aspects and more general in others. It is more general in that they allow for full confounding between all variables and has fewer restrictions on the functional form of the outcome mechanism (1). The biggest restriction in [7] is the assumption of Gaussian additive noise for all causal mechanisms. In contrast, our method only requires additive noise in the outcome mechanism (1), without parametric assumptions on the distribution of exogenous noise variables.

Bravo-Hermsdorff et al. [6] present a factor graph approach for the Intervention Generalization Problem, investigating identifiability of joint interventional effects given specific factorizations of observational and interventional probability distributions. Similarly, Jung et al. [11] present graphical conditions for non-parametric identification of joint interventional effects (which they term Multiple Treatment Interactions, MTI) and employ double machine learning techniques for estimation from marginal interventional data.

The complementary problem of generalizing from joint interventional data to single intervention effects was studied by [12] and [13], and generalization to unseen interventions without identifiability was explored in the context of stationary diffusion models [14].

Intervention Generalization is akin to other types of generalization, which is aided by the structure encoded in causal models such as the *Causal Marginal Problem* [15–17]. There, instead of generalizing to a new interventional regime, the goal is to learn about the joint behavior and causal structure of variables that have only been observed in subsets, but never jointly. Another example is *Out-of-Variable Generalization* [18], where some variables have never been observed in training.

A key aspect of our method is the ability of causal models to combine information from different data sets. This aligns with *Causal Representation Learning*, where many approaches use datasets or observation pairs that differ by interventions in the latent variables [19–22], samples from scientific simulations [1, 23], and multiple views and modalities [24].

6 Discussion and Outlook

We have shown that by constraining the outcome mechanism (1) to an additive model class, we can successfully identify joint interventional effects from single-interventional and observational data. Our constructive identifiability proof provides a practical estimator for the joint interventional effect (6). The estimator function decomposes into terms for the confounded and unconfounded contribution of each action to the outcome.

This work opens up several avenues for future research. A primary direction is exploring potential generalizations of the function class for the outcome mechanism (1). Generalized Additive [25] or Postnonlinear Models [26] could be suitable candidates. Such an extension could broaden the applicability of our approach to a wider range of real-world scenarios where strict additivity may not hold.

The precise confounding structure can be difficult to assess and often there are additional covariate to account for. Therefore, another area for investigation is the adaptation of our estimation technique to more complex scenarios. This includes settings with additional non-intervened covariates or under more general confounding structures between action variables.

Acknowledgments

We thank Frederick Träuble and Amartya Sanyal for fruitful discussions, and Simon Buchholz for valuable feedback on the manuscript.

References

[1] Armin Kekić, Jonas Dehning, Luigi Gresele, Julius von Kügelgen, Viola Priesemann, and Bernhard Schölkopf. Evaluating vaccine allocation strategies using simulation-assisted causal modeling. *Patterns*, 2023.

[2] Manuel Kunz, Stefan Birr, Mones Raslan, Lei Ma, Zhen Li, Adele Gouttes, Mateusz Koren, Tofigh Naghibi, Johannes Stephan, Mariia Bulycheva, Matthias Grzeschik, Armin Kekić, Michael Narodovitch, Kashif Rasul, Julian Sieber, and Tim Januschowski. Deep Learning based Forecasting: a case study from the online fashion industry. 2023.

[3] Douglas Schultz, Johannes Stephan, Julian Sieber, Trudie Yeh, Manuel Kunz, Patrick Doupe, and Tim Januschowski. Causal forecasting for pricing. *arXiv preprint arXiv:2312.15282*, 2023.

[4] Mattia Prosperi, Yi Guo, Matt Sperrin, James S Koopman, Jae S Min, Xing He, Shannan Rich, Mo Wang, Iain E Buchan, and Jiang Bian. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence*, 2020.

[5] Judea Pearl. *Causality*. Cambridge university press, 2009.

[6] Gecia Bravo-Hermsdorff, David Watson, Jialin Yu, Jakob Zeitler, and Ricardo Silva. Intervention generalization: A view from factor graph models. *NeurIPS*, 2023.

[7] Sorawit Saengkyongam and Ricardo Silva. Learning joint nonlinear effects from single-variable interventions in the presence of hidden confounders. *UAI*, 2020.

[8] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[9] Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *NeurIPS*, 2008.

[10] Leena Chennuru Vankadara, Philipp Michael Faller, Michaela Hardt, Lenon Minorics, Debarghya Ghoshdastidar, and Dominik Janzing. Causal forecasting: generalization bounds for autoregressive models. *UAI*, 2022.

[11] Yonghan Jung, Jin Tian, and Elias Bareinboim. Estimating joint treatment effects by combining multiple experiments. *ICML*, 2023.

[12] Olivier Jeunen, Ciarán Gilligan-Lee, Rishabh Mehrotra, and Mounia Lalmas. Disentangling causal effects from sets of interventions in the presence of unobserved confounders. *NeurIPS*, 2022.

[13] Muhammad Qasim Elahi, Mahsa Ghasemi, and Murat Kocaoglu. Identification of Average Causal Effects in Confounded Additive Noise Models. *arXiv preprint arXiv:2407.10014*, 2024.

[14] Lars Lorch, Andreas Krause, and Bernhard Schölkopf. Causal Modeling with Stationary Diffusions. *AISTATS*, 2024.

[15] Luigi Gresele, Julius Von Kügelgen, Jonas Kübler, Elke Kirschbaum, Bernhard Schölkopf, and Dominik Janzing. Causal inference through the structural causal marginal problem. *ICML*, 2022.

[16] Sergio Hernan Garrido Mejia, Elke Kirschbaum, and Dominik Janzing. Obtaining Causal Information by Merging Datasets with MAXENT. *AISTATS*, 2022.

[17] Sergio Hernan Garrido Mejia, Elke Kirschbaum, Armin Kekić, and Atalanti Mastakouri. Estimating Joint interventional distributions from marginal interventional Data. 2024.

[18] Siyuan Guo, Jonas Bernhard Wildberger, and Bernhard Schölkopf. Out-of-Variable Generalisation for Discriminative Models. *ICLR*, 2024.

[19] Wendong Liang, Armin Kekić, Julius von Kügelgen, Simon Buchholz, Michel Besserve, Luigi Gresele, and Bernhard Schölkopf. Causal Component Analysis. *NeurIPS*, 2023.

[20] Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco S Cohen. Weakly supervised causal representation learning. *NeurIPS*, 2022.

[21] Julius von Kügelgen, Michel Besserve, Wendong Liang, Luigi Gresele, Armin Kekić, Elias Bareinboim, David Blei, and Bernhard Schölkopf. Nonparametric Identifiability of Causal Representations from Unknown Interventions. *NeurIPS*, 2023.

[22] Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. *NeurIPS*, 2023.

[23] Armin Kekić, Bernhard Schölkopf, and Michel Besserve. Targeted Reduction of Causal Models. *UAI*, 2023.

[24] Dingling Yao, Danru Xu, Sebastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-View Causal Representation Learning with Partial Observability. *ICLR*, 2024.

[25] T.J. Hastie and R.J. Tibshirani. *Generalized Additive Models*. Taylor & Francis, 1990.

[26] Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. *UAI*, 2009.

A Proof of Theorem 1

A.1 Lemmas

Before we prove the main proposition, we introduce two useful lemmas.

Lemma 1. Let $\mathcal{M}$ be an SCM as defined above. Then,

$$\mathrm{p}^{\mathcal{M}(\mathrm{do}(a_j))}(c_k \mid a_1, \ldots, \mathrm{do}(a_j), \ldots, a_k) = \mathrm{p}(c_k \mid a_1, \ldots, a_j, \ldots, a_k), \tag{22}$$

for all $k \in \{2, \ldots, K\}$, and $j \in \{1, \ldots, k-1\}$.

Proof (Lemma 1) Using Bayes' rule on the left hand side of Equation (22) we have,

$$\mathrm{p}^{\mathcal{M}(\mathrm{do}(a_j))}(c_k \mid a_1, \ldots, \mathrm{do}(a_j), \ldots, a_k) \tag{23}$$

$$= \underbrace{\mathrm{p}^{\mathcal{M}(\mathrm{do}(a_j))}(a_k \mid a_1, \ldots, \mathrm{do}(a_j), \ldots, a_{k-1}, c_k)}_{\text{causal mechanism}} \frac{\overbrace{\mathrm{p}^{\mathcal{M}(\mathrm{do}(a_j))}(c_k \mid a_1, \ldots, \mathrm{do}(a_j), \ldots, a_{k-1})}^{\text{no open path between } c_k \text{ and } a_1,..,\mathrm{do}(a_j),..,a_{k-1}}}{\mathrm{p}^{\mathcal{M}(\mathrm{do}(a_j))}(a_k \mid a_1, \ldots, \mathrm{do}(a_j), \ldots, a_{k-1})} \tag{24}$$

$$= \mathrm{p}(a_k \mid a_1, \ldots, a_j, \ldots, a_{k-1}, c_k) \frac{\overbrace{\mathrm{p}^{\mathcal{M}(\mathrm{do}(a_j))}(c_k)}^{\text{root node}}}{\mathrm{p}^{\mathcal{M}(\mathrm{do}(a_j))}(a_k \mid a_1, \ldots, \mathrm{do}(a_j), \ldots, a_{k-1})} \tag{25}$$

$$= \mathrm{p}(a_k \mid a_1, \ldots, a_j, \ldots, a_{k-1}, c_k) \frac{\mathrm{p}(c_k)}{\mathrm{p}^{\mathcal{M}(\mathrm{do}(a_j))}(a_k \mid a_1, \ldots, \mathrm{do}(a_j), \ldots, a_{k-1})}. \tag{26}$$

We now focus on the denominator,

$$\mathrm{p}^{\mathcal{M}(\mathrm{do}(a_j))}(a_k \mid a_1, \ldots, \mathrm{do}(a_j), \ldots, a_{k-1}) \tag{27}$$

$$= \int \mathrm{p}^{\mathcal{M}(\mathrm{do}(a_j))}(a_k, c_k \mid a_1, \ldots, \mathrm{do}(a_j), \ldots, a_{k-1}) \, \mathrm{d}c_k \tag{28}$$

$$= \int \mathrm{p}^{\mathcal{M}(\mathrm{do}(a_j))}(a_k \mid a_1, \ldots, \mathrm{do}(a_j), \ldots, a_{k-1}, c_k) \, \mathrm{p}^{\mathcal{M}(\mathrm{do}(a_j))}(c_k \mid a_1, \ldots, \mathrm{do}(a_j), \ldots, a_{k-1}) \, \mathrm{d}c_k \tag{29}$$

$$= \int \mathrm{p}(a_k \mid a_1, \ldots, a_j, \ldots, a_{k-1}) \, \mathrm{p}^{\mathcal{M}(\mathrm{do}(a_j))}(c_k) \, \mathrm{d}c_k \tag{30}$$

$$= \int \mathrm{p}(a_k \mid a_1, \ldots, a_j, \ldots, a_{k-1}) \, \mathrm{p}(c_k) \, \mathrm{d}c_k \tag{31}$$

$$= \int \mathrm{p}(a_k \mid a_1, \ldots, a_j, \ldots, a_{k-1}) \, \mathrm{p}(c_k \mid a_1, \ldots, a_j, \ldots, a_{k-1}) \, \mathrm{d}c_k \tag{32}$$

$$= \mathrm{p}(a_k \mid a_1, \ldots, a_j, \ldots, a_{k-1}). \tag{33}$$

Using Equation (33) on Equation (26), and using Bayes' rule we obtain,

$$\mathrm{p}(a_k \mid a_1, \ldots, a_j, \ldots, a_{k-1}, c_k) \frac{\mathrm{p}(c_k \mid a_1, \ldots, a_j, \ldots, a_{k-1})}{\mathrm{p}(a_k \mid a_1, \ldots, a_j, \ldots, a_{k-1})} = \mathrm{p}(c_k \mid a_1, \ldots, a_j, \ldots, a_k), \tag{34}$$

as required. ■

Lemma 2. The following identities hold:

a) $\mathrm{p}^{\mathcal{M}(\mathrm{do}(a_1,..,a_K))}(c_1, \ldots, c_K \mid \mathrm{do}(a_1, \ldots, a_K)) = \prod_k \mathrm{p}(c_k)$.

b) $\mathrm{p}^{\mathcal{M}(\mathrm{do}(a_j))}(c_1, \ldots, c_k \mid a_1, \ldots, \mathrm{do}(a_j), \ldots, a_K) = \mathrm{p}(c_j) \prod_{k \neq j} \mathrm{p}(c_k \mid a_1, \ldots, a_k)$

c) $\mathrm{p}(c_1, \ldots, c_k \mid a_1, \ldots, a_k) = \prod_k \mathrm{p}(c_k \mid a_1, \ldots, a_k)$

Proof (Lemma 2)

a) Given the causal graph, and since all action variables A_k are intervened on, the only way in which we could introduce dependencies between the confounders is through conditioning on the collider Y. Hence, $C_i \perp\!\!\!\perp C_j \mid \mathrm{do}(A_1, \dots, A_k)$, for all $i, j \in \{1, \dots, K\}$. Thus, $\mathrm{p}(c_1, \dots, c_K \mid \mathrm{do}(a_1, \dots, a_k)) = \prod_k \mathrm{p}(c_k \mid \mathrm{do}(a_1, \dots, a_K))$. Furthermore, since intervention cuts the dependence to the action variables, and Y is not conditioned on, we have $C_K \perp\!\!\!\perp \mathrm{do}(A_1, \dots, A_K)$ for all K, giving us the first identity.

b) We have $C_i \perp\!\!\!\perp C_k \mid A_1, \dots, \mathrm{do}(A_j), \dots, A_k$ for all $i, k \in \{1, \dots, K\}$ since the conditioning set blocks all paths between the confounders. Either A_k block the outgoing path from C_k for unintervened actions, or there is no outgoing edge (other than to Y) for C_j.

Additionally, we have $C_j \perp\!\!\!\perp A_1, \dots, \mathrm{do}(A_j), \dots, A_K$. Hence,

$$\mathrm{p}^{\mathcal{M}(\mathrm{do}(a_j))}(c_1, \dots, c_K \mid a_1, \dots, \mathrm{do}(a_j), \dots, a_K) \tag{35}$$

$$= \overbrace{\mathrm{p}^{\mathcal{M}(\mathrm{do}(a_j))}(c_j)}^{\text{root node}} \prod_{k \neq j} \mathrm{p}^{\mathcal{M}(\mathrm{do}(a_j))}(c_k \mid a_1, \dots, \mathrm{do}(a_j), \dots, a_K) \tag{36}$$

$$= \mathrm{p}(c_j) \prod_{k \neq j} \mathrm{p}^{\mathcal{M}(\mathrm{do}(a_j))}(c_k \mid a_1, \dots, \mathrm{do}(a_j), \dots, a_K) \tag{37}$$

$$= \mathrm{p}(c_j) \prod_{k \neq j} \mathrm{p}(c_k \mid a_1, \dots, a_K), \tag{38}$$

where in the last line we use Lemma 1.

c) Follows from an argument analogous to b).

■

A.2 Proof of main result

Proof (Theorem 1) First note that

a)

$$\mathbb{E}[Y \mid \mathrm{do}(a_1, \dots, a_K)] = \int y \, \mathrm{p}^{\mathcal{M}(\mathrm{do}(a_1,\text{-},a_K))}(y \mid \mathrm{do}(a_1, \dots, a_K)) \, \mathrm{d}y \tag{39}$$

$$= \int \dots \int y \, \mathrm{p}^{\mathcal{M}(\mathrm{do}(a_1,\text{-},a_K))}(y \mid \mathrm{do}(a_1, \dots, a_K), c_1, \dots, c_K) \, \mathrm{d}y \\ \times \mathrm{p}^{\mathcal{M}(\mathrm{do}(a_1,\text{-},a_K))}(c_1, \dots c_K \mid \mathrm{do}(a_1, \dots, a_K)) \, \mathrm{d}c_1 \dots \mathrm{d}c_K \tag{40}$$

$$= \int \dots \int \mathbb{E}[Y \mid \mathrm{do}(a_1, \dots, a_K), c_1, \dots, c_K] \\ \times \mathrm{p}^{\mathcal{M}(\mathrm{do}(a_1,\text{-},a_K))}(c_1, \dots c_K \mid \mathrm{do}(a_1, \dots, a_K)) \, \mathrm{d}c_1 \dots \mathrm{d}c_K \tag{41}$$

$$= \int \dots \int \left(\sum_k f_k(a_k, c_k) \right) \mathrm{p}^{\mathcal{M}(\mathrm{do}(a_1,\text{-},a_K))}(c_1, \dots c_K \mid \mathrm{do}(a_1, \dots, a_K)) \, \mathrm{d}c_1 \dots \mathrm{d}c_K \tag{42}$$

$$\overset{\text{Lemma 2a)}}{=} \int \dots \int \left(\sum_k f_k(a_k, c_k) \right) \prod_k \mathrm{p}(c_k) \, \mathrm{d}c_1 \dots \mathrm{d}c_K \tag{43}$$

$$= \sum_k \int f_k(a_k, c_k) \, \mathrm{p}(c_k) \, \mathrm{d}c_k \tag{44}$$

$$= \sum_k \mathbb{E}_{C_k \sim \mathrm{p}(C_k)} \left[f_k(a_k, C_k) \right]. \tag{45}$$

Second,

b) For every $j \in \{1, \ldots, K\}$ we have

$$\mathbb{E}[Y \mid a_1, \ldots, \mathrm{do}(a_j), \ldots, a_K] = \int y \, \mathrm{p}^{\mathcal{M}(\mathrm{do}(a_j))}(y \mid a_1, \ldots, \mathrm{do}(a_j), \ldots, a_K) \, \mathrm{d}y \tag{46}$$

$$\begin{aligned} &= \int \ldots \int y \, \mathrm{p}^{\mathcal{M}(\mathrm{do}(a_j))}(y \mid a_1, \ldots, \mathrm{do}(a_j), \ldots, a_K, c_1, \ldots, c_K) \, \mathrm{d}y \\ &\qquad \times \mathrm{p}^{\mathcal{M}(\mathrm{do}(a_j))}(c_1, \ldots c_K \mid a_1, \ldots, \mathrm{do}(a_j), \ldots, a_K) \, \mathrm{d}c_1 \ldots \mathrm{d}c_K \end{aligned} \tag{47}$$

$$\begin{aligned} &= \int \ldots \int y \, \mathrm{p}(y \mid a_1, \ldots, a_j, \ldots, a_K, c_1, \ldots, c_K) \, \mathrm{d}y \\ &\qquad \times \mathrm{p}^{\mathcal{M}(\mathrm{do}(a_j))}(c_1, \ldots c_K \mid a_1, \ldots, \mathrm{do}(a_j), \ldots, a_K) \, \mathrm{d}c_1 \ldots \mathrm{d}c_K \end{aligned} \tag{48}$$

$$\begin{aligned} &= \int \ldots \int \mathbb{E}[Y \mid a_1, \ldots, a_j, \ldots, a_K, c_1, \ldots, c_K] \\ &\qquad \times \mathrm{p}^{\mathcal{M}(\mathrm{do}(a_j))}(c_1, \ldots c_K \mid a_1, \ldots, \mathrm{do}(a_j), \ldots, a_K) \, \mathrm{d}c_1 \ldots \mathrm{d}c_K \end{aligned} \tag{49}$$

$$= \int \ldots \int \left(\sum_k f_k(a_k, c_k) \right) \, \mathrm{p}^{\mathcal{M}(\mathrm{do}(a_j))}(c_1, \ldots c_K \mid a_1, \ldots, \mathrm{do}(a_j), \ldots, a_K) \, \mathrm{d}c_1 \ldots \mathrm{d}c_K \tag{50}$$

$$\overset{\text{Lemma 2b)}}{=} \int \ldots \int \left(\sum_k f_k(a_k, c_k) \right) \mathrm{p}(c_j) \prod_{k \neq j} \mathrm{p}(c_k \mid a_1, \ldots, a_K) \, \mathrm{d}c_1 \ldots \mathrm{d}c_K \tag{51}$$

$$= \int f_j(a_j, c_j) \mathrm{p}(c_j) \, \mathrm{d}c_j + \sum_{k \neq j} \int f_k(a_k, c_k) \mathrm{p}(c_k \mid a_1, \ldots, a_K) \, \mathrm{d}c_k \tag{52}$$

$$= \mathbb{E}_{C_j \sim \mathrm{p}(C_j)}[f_j(a_j, C_j)] + \sum_{k \neq j} \mathbb{E}_{C_k \sim \mathrm{p}(C_k \mid a_1, \ldots, a_K)}[f_k(a_k, C_k)]. \tag{53}$$

And finally,

c)

$$\mathbb{E}[Y \mid a_1, \ldots, a_K] = \int y \, \mathrm{p}(y \mid a_1, \ldots, a_K) \, \mathrm{d}y \tag{54}$$

$$= \int \ldots \int y \, \mathrm{p}(y \mid a_1, \ldots, a_K, c_1, \ldots, c_K) \, \mathrm{d}y \, \mathrm{p}(c_1, \ldots c_K \mid a_1, \ldots, a_K) \, \mathrm{d}c_1 \ldots \mathrm{d}c_K \tag{55}$$

$$= \int \ldots \int \mathbb{E}[Y \mid a_1, \ldots, a_K, c_1, \ldots, c_K] \, \mathrm{p}(c_1, \ldots c_K \mid a_1, \ldots, a_K) \, \mathrm{d}c_1 \ldots \mathrm{d}c_K \tag{56}$$

$$= \int \ldots \int \left(\sum_k f_k(a_k, c_k) \right) \, \mathrm{p}(c_1, \ldots c_K \mid a_1, \ldots, a_K) \, \mathrm{d}c_1 \ldots \mathrm{d}c_K \tag{57}$$

$$\overset{\text{Lemma 2c)}}{=} \int \ldots \int \left(\sum_k f_k(a_k, c_k) \right) \prod_k \mathrm{p}(c_k \mid a_1, \ldots, a_K) \, \mathrm{d}c_1 \ldots \mathrm{d}c_K \tag{58}$$

$$= \sum_k \int f_k(a_k, c_k) \, \mathrm{p}(c_k \mid a_1, \ldots, a_K) \, \mathrm{d}c_k \tag{59}$$

$$= \sum_k \mathbb{E}_{C_k \sim \mathrm{p}(C_k \mid a_1, \ldots, a_K)}[f_k(a_k, C_k)]. \tag{60}$$

We learn K functions

$$\hat{f}_k(a_1, \ldots, a_K, R_k), \qquad k \in \{1, \ldots, K\} \tag{61}$$

where $R_k \in \{0, 1\}$ is an indicator for whether A_k was intervened on. R_k can be thought of as selecting one of two functions

$$\hat{f}_k(a_1, \ldots, a_K, R_k) = \begin{cases} \hat{f}_k^{\mathrm{obs}}(a_1, \ldots, a_K) & \text{if } R_k = 0 \\ \hat{f}_k^{\mathrm{int}}(a_1, \ldots, a_K) & \text{if } R_k = 1 \end{cases} \tag{62}$$

where $\hat{f}_k^{\text{obs}}$ and $\hat{f}_k^{\text{int}}$ are universal function approximators.

We define

$$\hat{f}(a_1, \dots, a_K, R_1, \dots, R_K) = \sum_{k=1}^{K} \hat{f}_k(a_1, \dots, a_K, R_k). \tag{63}$$

Since we are in the infinite data regime and have universal function approximators we can fit $\hat{f}$ such that

$$\hat{f}(a_1, \dots, a_K, R_1{=}0, \dots, R_K{=}0) = \mathbb{E}[Y \mid a_1, \dots, a_K] \tag{64}$$

$$\hat{f}(a_1, \dots, a_K, R_1{=}0, \dots, R_j{=}1, \dots, R_K{=}0) = \mathbb{E}[Y \mid a_1, \dots, \mathrm{do}(a_j), \dots, a_K], \quad \text{for } j \in \{1, \dots, K\}. \tag{65}$$

From the definition of Equation (63), we have for each $j \in \{1, \dots, K\}$

$$\hat{f}(a_1, \dots, a_K, R_1{=}0, \dots, R_j{=}1, \dots, R_K{=}0) - \hat{f}(a_1, \dots, a_K, R_1{=}0, \dots, R_j{=}0, \dots, R_K{=}0) \tag{66}$$

$$= \hat{f}_j(a_1, \dots, a_K, R_j{=}1) - \hat{f}_j(a_1, \dots, a_K, R_j{=}0). \tag{67}$$

After training the estimator we have

$$\hat{f}(a_1, \dots, a_K, R_1{=}0, \dots, R_j{=}1, \dots, R_K{=}0) - \hat{f}(a_1, \dots, a_K, R_1{=}0, \dots, R_j{=}0, \dots, R_K{=}0) \tag{68}$$

$$= \mathbb{E}[Y \mid a_1, \dots, \mathrm{do}(a_j), \dots, a_K] - \mathbb{E}[Y \mid a_1, \dots, a_K] \tag{69}$$

$$\overset{(53),(60)}{=} \mathbb{E}_{C_j \sim \mathrm{p}(C_j)}[f_j(a_j, C_j)] - \mathbb{E}_{C_j \sim \mathrm{p}(C_j \mid a_1,.,a_K)}[f_j(a_j, C_j)] \tag{70}$$

where in the last step we have plugged in the decomposition of the expectation in the single-intervention (53) and observational (60) setting.

Combining the definition of the estimator (63), and Equations (67) and (70), we get an expression for the joint interventional effect:

$$\hat{f}(a_1, \dots, a_K, R_1{=}1, \dots, R_K{=}1) = \sum_{j=1}^{K} \hat{f}_j(a_1, \dots, a_K, R_j{=}1) \tag{71}$$

$$= \sum_{j=1}^{K} \Big(\mathbb{E}_{C_j \sim \mathrm{p}(C_j)}[f_j(a_j, C_j)] - \mathbb{E}_{C_j \sim \mathrm{p}(C_j \mid a_1,.,a_K)}[f_j(a_j, C_j)] + \hat{f}_j(a_1, \dots, a_K, R_j{=}0) \Big) \tag{72}$$

$$= \sum_{j=1}^{K} \mathbb{E}_{C_j \sim \mathrm{p}(C_j)}[f_j(a_j, C_j)] - \sum_{j=1}^{K} \mathbb{E}_{C_j \sim \mathrm{p}(C_j \mid a_1,.,a_K)}[f_j(a_j, C_j)] + \sum_{j=1}^{K} \hat{f}_j(a_1, \dots, a_K, R_j{=}0) \tag{73}$$

$$\overset{(45),(60),(63)}{=} \mathbb{E}[Y \mid \mathrm{do}(a_1, \dots, a_K)] \underbrace{- \mathbb{E}[Y \mid a_1, \dots, a_K] + \hat{f}(a_1, \dots, a_K, R_1{=}0, \dots, R_K{=}0)}_{\overset{(64)}{=} 0} \tag{74}$$

$$= \mathbb{E}[Y \mid \mathrm{do}(a_1, \dots, a_K)]\,. \tag{75}$$

■

B Proof of Proposition 1

Proof (Proposition 1) As in the proof of Theorem 1, we can decompose the interventional effect as

$$\mathbb{E}[Y \mid \mathrm{do}(\mathbf{a}_{\text{int}}), \mathbf{a}_{\text{obs}}] = \sum_{\substack{j \\ A_j \in \mathbf{A}_{\text{int}}}} \mathbb{E}_{C_j \sim \mathrm{p}(C_j)}[f_j(a_j, C_j)] + \sum_{\substack{k \\ A_k \in \mathbf{A}_{\text{obs}}}} \mathbb{E}_{C_k \sim \mathrm{p}(C_k \mid a_1,.,a_K)}[f_k(a_k, C_k)]. \tag{76}$$

We learn K functions and fit them to the observational and single-interventional expectation (Equations (61) to (65)).

Let

$$R_k = \begin{cases} 1 \text{ if } A_k \in \mathbf{A}_{\text{int}} \\ 0 \text{ if } A_k \in \mathbf{A}_{\text{obs}} \end{cases} \quad \text{for all } k \in \{1, \ldots, K\}. \tag{77}$$

Then our estimator identifies the interventional effect (21), since

$$\begin{aligned}
&\hat{f}(a_1, \ldots, a_K, R_1, \ldots, R_K) &(78)\\
&\overset{(62),(63)}{=} \sum_{\substack{k\\ A_k \in \mathbf{A}_{\text{obs}}}} \hat{f}_k(a_1, \ldots, a_K, R_k{=}0) + \sum_{\substack{j\\ A_j \in \mathbf{A}_{\text{int}}}} \hat{f}_j(a_1, \ldots, a_K, R_j{=}1) &(79)\\
&= \sum_{\substack{k\\ A_k \in \mathbf{A}_{\text{obs}}}} \hat{f}_k(a_1, \ldots, a_K, R_k{=}0) + \sum_{\substack{l\\ A_l \in \mathbf{A}_{\text{int}}}} \hat{f}_l(a_1, \ldots, a_K, R_l{=}0)\\
&\qquad - \sum_{\substack{l\\ A_l \in \mathbf{A}_{\text{int}}}} \hat{f}_l(a_1, \ldots, a_K, R_l{=}0) + \sum_{\substack{j\\ A_j \in \mathbf{A}_{\text{int}}}} \hat{f}_j(a_1, \ldots, a_K, R_j{=}1) &(80)\\
&= \underbrace{\sum_{k=1}^{K} \hat{f}_k(a_1, \ldots, a_K, R_k{=}0)}_{\overset{(63),(64)}{=} \mathbb{E}[Y | a_1, \ldots, a_K]} + \sum_{j,\ A_j \in \mathbf{A}_{\text{int}}} \underbrace{\Big(\hat{f}_j(a_1, \ldots, a_K, R_j{=}1) - \hat{f}_j(a_1, \ldots, a_K, R_j{=}0)\Big)}_{\overset{(67),(70)}{=} \mathbb{E}_{C_j \sim \mathrm{p}(C_j)}[f_j(a_j, C_j)] - \mathbb{E}_{C_j \sim \mathrm{p}(C_j | a_1, \ldots, a_K)}[f_j(a_j, C_j)]} &(81)\\
&\overset{(60)}{=} \sum_{k=1}^{K} \mathbb{E}_{C_k \sim \mathrm{p}(C_k | a_1, \ldots, a_K)}[f_k(a_k, C_k)]\\
&\qquad + \sum_{j,\ A_j \in \mathbf{A}_{\text{int}}} \big(\mathbb{E}_{C_j \sim \mathrm{p}(C_j)}[f_j(a_j, C_j)] - \mathbb{E}_{C_j \sim \mathrm{p}(C_j | a_1, \ldots, a_K)}[f_j(a_j, C_j)]\big) &(82)\\
&= \sum_{\substack{j\\ A_j \in \mathbf{A}_{\text{int}}}} \mathbb{E}_{C_j \sim \mathrm{p}(C_j)}[f_j(a_j, C_j)] + \sum_{\substack{k\\ A_k \in \mathbf{A}_{\text{obs}}}} \mathbb{E}_{C_k \sim \mathrm{p}(C_k | a_1, \ldots, a_K)}[f_k(a_k, C_k)] &(83)\\
&\overset{(76)}{=} \mathbb{E}[Y \mid \mathrm{do}(\mathbf{a}_{\text{int}}), \mathbf{a}_{\text{obs}}] &(84)
\end{aligned}$$

■

C Additional Restults

Definition 2. Let $\mathfrak{B}$ be a partition of the index set $\{1, \ldots, K\}$. For an SCM of the form discussed in Section 2.2, we call the outcome mechanism (1) *additive with respect to* $\mathfrak{B}$, if we can write the structural assignments as:

$$\begin{aligned}
Y &:= \sum_{B \in \mathfrak{B}} f_B(\mathbf{A}_B, \mathbf{C}_B) + U & &(85)\\
A_k &:= g_k(A_1, \ldots, A_{k-1}, C_k, V_k) & \text{for } k \in \{1, \ldots, K\} &(86)\\
C_k &:= W_k & \text{for } k \in \{1, \ldots, K\}\,. &(87)
\end{aligned}$$

Corollary 1. Let $\mathfrak{B}$ be a partition of the index set $\{1, \ldots, K\}$ such that the ground-truth SCM $\mathcal{M}$ has an outcome mechanism which is additive with respect to $\mathfrak{B}$. Then the joint interventional effect (6) can be identified from observational data

$$\mathrm{D}_{\text{obs}} \sim \mathrm{P}^{\mathcal{M}}_{(Y,\mathbf{A})} \tag{88}$$

and $|\mathfrak{B}|$ interventional datasets

$$\{\mathrm{D}^{b}_{\text{int}} \sim \mathrm{P}^{\mathcal{M}_{\mathrm{do}(\mathbf{A}_B)}}_{(Y,\mathbf{A})}\}_{B \in \mathfrak{B}}\,. \tag{89}$$

Proof (Corollary 1) Similar to the proof of Theorem 1, we can decompose the outcome expectations in the joint interventional regime and the settings for which we have data as

$$\mathbb{E}[Y \mid \mathrm{do}(a_1, \dots, a_K)] = \sum_{B \in \mathfrak{B}} \mathbb{E}_{\mathbf{C}_B \sim \prod_{k \in B} \mathrm{p}(C_k)} \left[f_B(\mathbf{a}_B, \mathbf{C}_B) \right] \tag{90}$$

$$\begin{aligned} \mathbb{E}[Y \mid \mathrm{do}(\mathbf{a}_B), \mathbf{a}_{\neg B}] &= \mathbb{E}_{\mathbf{C}_B \sim \prod_{k \in B} \mathrm{p}(C_k)} \left[f_B(\mathbf{a}_B, \mathbf{C}_B) \right] \\ &\quad + \sum_{\tilde{B} \neq B} \mathbb{E}_{\mathbf{C}_{\tilde{B}} \sim \prod_{k \in \tilde{B}} \mathrm{p}(C_k \mid a_1,..,a_K)} [f_{\tilde{B}}(\mathbf{a}_{\tilde{B}}, \mathbf{C}_{\tilde{B}})] \quad \forall B \in \mathfrak{B} \end{aligned} \tag{91}$$

$$\mathbb{E}[Y \mid a_1, \dots, a_K] = \sum_{B \in \mathfrak{B}} \mathbb{E}_{\mathbf{C}_B \sim \prod_{k \in \tilde{B}} \mathrm{p}(C_k \mid a_1,..,a_K)} [f_B(\mathbf{a}_B, \mathbf{C}_B)] \,, \tag{92}$$

where $\mathbf{a}_{\neg B}$ denotes all actions that are not in subset B.

We define $|\mathfrak{B}|$ estimator functions as

$$\sum_{B \in \mathfrak{B}} \hat{f}_B(a_1, \dots, a_K, R_B) \,, \tag{93}$$

where $R_B \in \{0, 1\}$ indicates whether the actions in the subset B were intervened on. Then, following the analogous steps as in Theorem 1, the joint interventional effect (6) is identified through

$$\sum_{B \in \mathfrak{B}} \hat{f}_B(a_1, \dots, a_K, R_B{=}1) \,. \tag{94}$$

■

D Probability Distributions for Example 1

Table 1: Observational distribution: $\mathrm{P}^{\mathcal{M}}(Y, A_1, A_2) = \mathrm{P}^{\widetilde{\mathcal{M}}}(Y, A_1, A_2)$

$\mathrm{P}^{\mathcal{M}}(Y, A_1, A_2)$	$Y = 0$	$Y = 1$
$A_1 = 0, A_2 = 0$	$(1 - p)$	0
$A_1 = 1, A_2 = 0$	$p(1 - p)$	0
$A_1 = 0, A_2 = 1$	0	0
$A_1 = 1, A_2 = 1$	$p^2(1 - p)$	p^3

Table 2: Single-intervention distribution: $\mathrm{P}^{\mathcal{M}(\mathrm{do}(A_1))}(Y, A_2) = \mathrm{P}^{\widetilde{\mathcal{M}}(\mathrm{do}(A_1))}(Y, A_2)$

	$\mathrm{P}^{\mathcal{M}(\mathrm{do}(A_1))}(Y, A_2)$	$Y = 0$	$Y = 1$
$\mathrm{do}(A_1 = 0)$	$A_2 = 0$	1	0
	$A_2 = 1$	0	0
$\mathrm{do}(A_1 = 1)$	$A_2 = 0$	$(1 - p)$	0
	$A_2 = 1$	0	p

Table 3: Single-intervention distribution: $\mathrm{P}^{\mathcal{M}(\mathrm{do}(A_2))}(Y, A_1) = \mathrm{P}^{\widetilde{\mathcal{M}}(\mathrm{do}(A_2))}(Y, A_1)$

	$\mathrm{P}^{\mathcal{M}(\mathrm{do}(A_2))}(Y, A_1)$	$Y = 0$	$Y = 1$
$\mathrm{do}(A_2 = 0)$	$A_1 = 0$	$(1 - p)$	0
	$A_1 = 1$	$p(1 - p)$	p^2
$\mathrm{do}(A_2 = 1)$	$A_1 = 0$	$(1 - p)$	0
	$A_1 = 1$	$p(1 - p)$	p^2

E Experiments

Sampling SCMs We sample SCMs of the form

$$Y \coloneqq \sum_{k=1}^{5} \underbrace{\alpha_k A_k + \beta_k C_k}_{f_k(A_k, C_k)} + U \tag{95}$$

$$A_k \coloneqq \gamma_k C_k + \sum_{j<k} \delta_{kj} M_{kj} A_j + V_k \qquad \text{for } k \in 1, \ldots, 5 \tag{96}$$

$$C_k \coloneqq W_k \qquad \text{for } k \in 1, \ldots, 5, , \tag{97}$$

where $M_{kj} \sim \text{Bernoulli}(p_{\text{edge}})$ are binary masks determining which edges between actions are active, and the exogenous noises are lognormal distributions with zero mean and standard deviation drawn uniformly from $\sigma_U, \sigma_{V_1}, \ldots, \sigma_{V_5}, \sigma_{W_1}, \ldots, \sigma_{W_5} \sim \mathrm{Uniform}(0.5, 1.5)$. The parameters of the causal mechanisms are also drawn uniformly from $\alpha_1, \ldots, \alpha_5, \gamma_1, \ldots, \gamma_5, \delta_{kj} \sim \mathrm{Uniform}(0.5, 1.5)$. The edge probability p_{edge} is set to 0.5 in our experiments.