Incorporating Causal Graphical Prior Knowledge into Predictive Modeling via Simple Data Augmentation

Takeshi Teshima[1,2] **Masashi Sugiyama**[2,1]

[1]Graduate School of Frontier Sciences, The University of Tokyo, JAPAN
[2]RIKEN, JAPAN

Abstract

Causal graphs (CGs) are compact representations of the knowledge of the data generating processes behind the data distributions. When a CG is available, e.g., from the domain knowledge, we can infer the conditional independence (CI) relations that should hold in the data distribution. However, it is not straightforward how to incorporate this knowledge into predictive modeling. In this work, we propose a model-agnostic data augmentation method that allows us to exploit the prior knowledge of the CI encoded in a CG for supervised machine learning. We theoretically justify the proposed method by providing an excess risk bound indicating that the proposed method suppresses overfitting by reducing the apparent complexity of the predictor hypothesis class. Using real-world data with CGs provided by domain experts, we experimentally show that the proposed method is effective in improving the prediction accuracy, especially in the small-data regime.

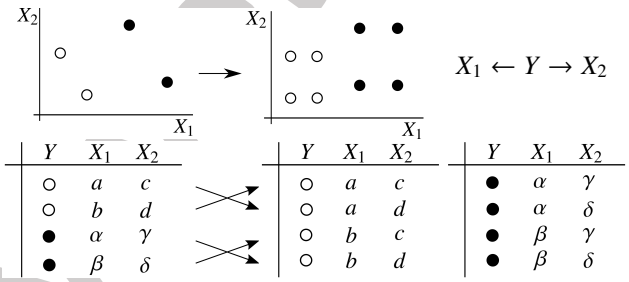


Figure 1: Visualization of the basic idea of the paper for the trivariate case $X_1 \leftarrow Y \rightarrow X_2$. In this case, the CI $X_1 \perp\!\!\!\perp X_2 \mid Y$ holds. One way to use this knowledge via data augmentation is to group the data according to Y and then to shuffle X_1 and X_2 within each group. Our method extends this idea to more general graphs.

1 INTRODUCTION

Causal graphs (CGs; Pearl, 2009) are compact representations of the knowledge of data generating processes. Such a CG is sometimes provided by domain experts in some problem instances, e.g., in biology (Sachs et al., 2005) or sociology (Shimizu et al., 2011). Otherwise, it may also be learned from data using the statistical causal discovery methods developed over the last decades (Spirtes et al., 2000; Pearl, 2009; Chickering, 2002; Shimizu et al., 2006; Peters et al., 2014; Peters et al., 2017). Once a CG is obtained, it can be used to infer the conditional independence (CI) relations that the data distribution should satisfy (Pearl, 2009).

The CI relations encoded in the CG could be strong prior knowledge for predictive tasks in machine learning, e.g., regression or classification, especially in the *small-data* regime where data alone may be insufficient to witness the CI relations (Spirtes et al., 2000, Section 5.2.2). However, it is not trivial how the CI relations should be directly incorporated into general supervised learning methods. In previous research, methods that leverage the causality for feature selection have been proposed (see, e.g., Yu et al. (2020) for a review). However, most of them are based on the notion of the *Markov blanket* or the *Markov boundary* (Tsamardinos et al., 2003). As a result, they only take into account partial information of all that is encoded in a CG, since a CG often entails more constraints on the data distribution than the specifications of Markov blankets or a Markov boundary (Richardson, 2003). Another approach to exploiting the prior knowledge of a CG is to build a *Bayesian network* (BN) model according to the CG structure (e.g., Lucas et al., 2004). However, constructing the predictors by employing BNs as the framework entails a specific modeling choice, e.g., it constructs a *generative* model as opposed to a *discriminative* model (Shalev-Shwartz et al., 2014, Chapter 24), precluding the choice of some flexible and effective models such as tree-based predictors (Friedman, 2001) and neural networks (Goodfellow et al., 2016) that may be preferred in

Accepted for the 37th Conference on Uncertainty in Artificial Intelligence (UAI 2021).

the application area of one's interest.

In this work, we propose a model-agnostic method to incorporate the CI relations implied by CGs directly into supervised learning via data augmentation. To illustrate our idea, let us consider the following trivariate case.

Illustrative example: trivariate case (Fig. 1). Suppose we want to predict a binary variable Y from (X_1, X_2). If the joint distribution follows the CG $X_1 \leftarrow Y \rightarrow X_2$, the CI $X_1 \perp\!\!\!\perp X_2 \mid Y$ holds (Pearl, 2009). If we know this relation, a natural idea is to stratify the sample by Y and then to take all combinations of X_1 and X_2 within each stratum.

In this trivariate example, it is straightforward to derive such a plausible data augmentation procedure to incorporate the CI relations since the relation $X_1 \perp\!\!\!\perp X_2 \mid Y$ involves all three variables. On the other hand, deriving such a procedure for general graphs is not straightforward as they may encode a multitude of CI relations each of which may involve only a subset of all variables.

Our contributions. (i) We propose a method to augment data based on the prior knowledge expressed as CGs, assuming that an estimated CG is available. (ii) We theoretically justify the proposed method via an excess risk bound based on the Rademacher complexity (Bartlett et al., 2002). The bound indicates that the proposed method suppresses overfitting at the cost of introducing additional complexity and bias into the problem. (iii) We empirically show that the proposed method yields consistent performance improvements especially in the small-data regime, through experiments using real-world data with CGs obtained from the domain knowledge.

2 PROBLEM SETUP

In this section, we describe the problem setup, the goal, and the main assumption exploited in our proposed method.

Basic notation. For the standard notation, namely $\mathbb{R}$, $\mathbb{R}_{\geq 0}$, $\mathbb{R}_{>0}$, $\mathbb{Z}$, $\mathbb{N}$, and $\mathbb{1}[\cdot]$, see Table 1 in Appendix that also provides a summary of notation. For $N, M \in \mathbb{N}$ with $N \leq M$, define $[N : M] := \{N, N+1, \ldots, M\}$ and $[N] := [1 : N]$. For an N-dimensional vector $\boldsymbol{x} = (x^1, \ldots, x^N)$ and $S \subset [N]$, we let $\boldsymbol{x}^S = (x^{s_1}, \ldots, x^{s_{|S|}})$ denote its sub-vector with indices in $S = \{s_1, \ldots, s_{|S|}\}$ with $s_1 < \cdots < s_{|S|}$. By abuse of notation, we write $\boldsymbol{x}^j := \boldsymbol{x}^{\{j\}}$ for $j \in [N]$. To simplify the notation, we let $[0] = \emptyset$, $\mathbb{R}^0 := \{0\}$, $\boldsymbol{x}^\emptyset = 0$, and $[N]^0 = \{0\}$.

Problem setup and goal. Throughout the paper, we fix $D \in \mathbb{N}$, and let $\mathcal{Z} = \times_{j=1}^D \mathcal{Z}^j$ where each $\mathcal{Z}^j$ is a subset of $\overline{\mathcal{Z}}^j$ that is $\mathbb{R}$, $\mathbb{Z}$, or a finite set. Let p be the joint probability density of $\mathbf{Z} := (Z^1, \ldots, Z^D)$ taking values in $\mathcal{Z}$. One of the variables, e.g., Z^{j^*} ($j^* \in [D]$), is the target variable that we want to predict. Let $\mathcal{X} = \times_{j \in [D] \setminus \{j^*\}} \overline{\mathcal{Z}}^j$ and $\mathcal{Y} = \overline{\mathcal{Z}}^{j^*}$. Let $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class and $\ell : \mathcal{F} \times \left(\times_{j=1}^D \overline{\mathcal{Z}}^j \right) \to \mathbb{R}$ be a loss function. We consider the supervised learning setting; that is, given the training data $\mathcal{D} = \{\mathbf{Z}_i\}_{i=1}^n$ that is an independently and identically distributed sample from p, our goal is to find a predictor $\hat{f} \in \mathcal{F}$ with a small risk $R(\hat{f}) = \mathbb{E}[\ell(\hat{f}, \mathbf{Z})]$, where $\mathbb{E}$ denotes the expectation with respect to p.

Assumption. Let $\mathcal{G} = ([D], \mathcal{E}, \mathcal{B})$ be an *acyclic directed mixed graph*[1] (ADMG; Richardson, 2003; Richardson et al., 2017), where $[D]$ is the set of the vertices, $\mathcal{E}$ is the uni-directed edges, and $\mathcal{B}$ is the bi-directed edges. For the simplicity of exposition, in this paragraph, we temporarily assume that $[D]$ is concordant with *topological order* of $\mathcal{G}$ without loss of generality.[2] Our main assumption is that p satisfies the *topological ADMG factorization* property with respect to $\mathcal{G}$ (Bhattacharya et al., 2020), i.e.,

$$p(\mathbf{Z}) = \prod_{j=1}^{D} p_{j|\mathrm{mp}(j)}(Z^j | \mathbf{Z}^{\mathrm{mp}(j)}), \tag{1}$$

where $\mathrm{mp}(j) \subset [j-1]$ denotes the *Markov pillow* of $j \in [D]$ in $\mathcal{G}$, and $p_{j|\mathrm{mp}(j)}$ denotes the conditional density of Z^j given $\mathbf{Z}^{\mathrm{mp}(j)}$. The Markov pillow $\mathrm{mp}(j)$ is the collection of the following vertices: (1) those connected to j via bi-directed paths (including j itself), and (2) all parents of such vertices (see Appendix A or Bhattacharya et al. (2020) for the definition). Markov pillow generalizes the notion of parents; if all edges are uni-directed, $\mathrm{mp}(j)$ matches the parents of j, and hence Eq. (1) is a generalization of the usual Markov factorization with respect to directed acyclic graphs (DAGs; Pearl, 2009, p.16) to ADMGs. In the special case that the ADMG is *uninformative*, i.e., when the graph is complete and all edges are bi-directed, Eq. (1) reduces to the ordinary *chain rule* of probability: $p(\mathbf{Z}) = \prod_{j=1}^D p(Z^j | \mathbf{Z}^{[j-1]})$, since $\mathrm{mp}(j) = [j-1]$ in this case. We assume that we are given an ADMG $\hat{\mathcal{G}} = ([D], \hat{\mathcal{E}}, \hat{\mathcal{B}})$ that is an estimator of $\mathcal{G}$, and hereafter we assume that $[D]$ is concordant with topological order of $\hat{\mathcal{G}}$ without loss of generality.

Details on the assumption. ADMGs with bi-directed edges appear in the case where unobserved confounders exist; they are used to represent *semi-Markovian causal graphical models* (CGMs; Tian et al., 2002), which are CGMs allowing for the existence of hidden confounders. The assumption of topological ADMG factorization is satisfied by such CGMs (Tian et al., 2002). We refer the readers to Section 2 of Richardson et al. (2017) for an overview of ADMGs and their use in CGMs involving latent variables. By accommodating not only DAGs (i.e., those without bi-directed edges) but also general ADMGs in the assumption,

[1] Here, *mixed* indicates that the graph may contain bi-directed edges in addition to uni-directed ones.

[2] That is, if $1 \leq i < j \leq D$, there is no directed path from j to i.

the applicability of the proposed method is extended to the case where there are unobserved confounders. Note that the topological ADMG factorization, in general, captures only part of the equality constraints imposed by an ADMG on a semi-Markov model (Bhattacharya et al., 2020). Indeed, Bhattacharya et al. (2020) proposed a simple sufficient condition called the *mb-shieldedness* (*mb* stands for "Markov blanket") under which the topological ADMG factorization captures all the equality constraints. Also note that a CG encodes more information/assumptions than the CI relations, namely, it encodes causal assumptions that describe how the data distribution should shift under an intervention (Pearl, 2009). In this work, we only exploit the statistical assumptions, namely the CI relations, implied by a given CG. Although our method does not directly exploit the causal interpretation of the DAGs/ADMGs, the causal modeling perspective can be useful in obtaining the DAGs/ADMGs from domain experts, i.e., one may be able to draw the DAGs/ADMGs by considering the (non-parametric) structural equations (Pearl, 2009).

3 PROPOSED METHOD

In this section, we explain the proposed data augmentation method to directly incorporate the prior knowledge of an ADMG into supervised learning. The method generalizes the intuitive data augmentation method described in the trivariate DAG example in Section 1, making it applicable to general ADMGs whose encoded CI relations do not necessarily involve all variables. The idea is to consider a *nested conditional resampling*; instead of trying to generate all elements of the new data vector at once, we successively resample each variable from the *conditional empirical distribution* (Stute, 1986; Horváth et al., 1988) conditioning on its Markov pillow. Then, our proposed method *ADMG data augmentation* is obtained by considering all possible resampling paths simultaneously. We later confirm that the proposed method indeed generalizes the previous procedure considered in the trivariate case of Fig. 1.

Derivation of the proposed method. Recall, given Eq. (1), we can express the risk functional as

$$R(f) = \int_{\mathcal{Z}} \ell(f, \mathbf{Z}) \prod_{j=1}^{D} \underbrace{p_{j|\mathrm{mp}(j)}(Z^j|\mathbf{Z}^{\mathrm{mp}(j)})}_{(*)} \mathrm{d}\mathbf{Z}.$$

Then, to formulate the nested conditional resampling procedure, we select a kernel function $K^j : \overline{\mathcal{Z}}^{\mathrm{mp}(j)} \to \mathbb{R}_{\geq 0}$ for each $j \in [D]$.[3] Using this kernel function in the spirit of kernel-type function estimators (Nadaraya, 1964; Watson, 1964; Einmahl et al., 2000), we approximate each conditional density $(*)$ as

$$\hat{p}_{j|\mathrm{mp}(j)}(Z^j|\mathbf{Z}^{\mathrm{mp}(j)}) := \frac{\sum_{i=1}^{n} \delta_{Z_i^j}(Z^j) K^j(\mathbf{Z}^{\mathrm{mp}(j)} - \mathbf{Z}_i^{\mathrm{mp}(j)})}{\sum_{k=1}^{n} K^j(\mathbf{Z}^{\mathrm{mp}(j)} - \mathbf{Z}_k^{\mathrm{mp}(j)})},$$

where $\delta_{\boldsymbol{z}}$ denotes Dirac's delta function centered at $\boldsymbol{z}$ (e.g., Zorich, 2015, Section E.4.1), and the right-hand side is defined to be zero when the denominator is zero. The resulting approximation to the risk functional $R(f)$, denoted by $\hat{R}_{\mathrm{aug}}(f)$, is

$$\hat{R}_{\mathrm{aug}}(f) := \int_{\mathcal{Z}} \ell(f, \mathbf{Z}) \prod_{j=1}^{D} \hat{p}_{j|\mathrm{mp}(j)}(Z^j|\mathbf{Z}^{\mathrm{mp}(j)}) \mathrm{d}\mathbf{Z}.$$

Here, the right-hand side can be interpreted as representing a nested conditional resampling procedure, in which we sequentially select $i_1, \ldots, i_D \in [n]$. Indeed, since each $\hat{p}_{j|\mathrm{mp}(j)}$ places its mass on $\{Z_i^j\}_{i=1}^n$, the integration for Z^j amounts to substituting $Z^j = Z_{i_j}^j$ and summing over the choices $i_j \in [n]$ with appropriate weights. The weight placed on Z_i^j by $\hat{p}_{j|\mathrm{mp}(j)}$, namely $\frac{K^j(\mathbf{Z}^{\mathrm{mp}(j)} - \mathbf{Z}_i^{\mathrm{mp}(j)}) I_{\neq 0}}{\sum_{k=1}^{n} K^j(\mathbf{Z}^{\mathrm{mp}(j)} - \mathbf{Z}_k^{\mathrm{mp}(j)})}$, depends on $\mathbf{Z}^{\mathrm{mp}(j)}$, and it can be computed from $(Z_{i_1}^1, \ldots, Z_{i_{j-1}}^{j-1})$ which are already selected at the time we select $Z_{i_j}^j$ since $\mathrm{mp}(j) \subset [j-1]$.

Proposed method. By simultaneously considering all the possible resampling candidates, we reach at the *instance-weighted data augmentation* procedure:

$$\hat{R}_{\mathrm{aug}}(f) = \sum_{\boldsymbol{i} \in [n]^D} \hat{w}_{\boldsymbol{i}} \cdot \ell(f, \mathbf{Z}_{\boldsymbol{i}}), \tag{2}$$

where

$$\hat{w}_{\boldsymbol{i}} = \prod_{j=1}^{D} \frac{K^j(\mathbf{Z}_{\boldsymbol{i}_{1:j-1}}^{\mathrm{mp}(j)} - \mathbf{Z}_i^{\mathrm{mp}(j)})}{\sum_{k=1}^{n} K^j(\mathbf{Z}_{\boldsymbol{i}_{1:j-1}}^{\mathrm{mp}(j)} - \mathbf{Z}_k^{\mathrm{mp}(j)})}, \tag{3}$$
$$\mathbf{Z}_{\boldsymbol{i}} = (Z_{i_1}^1, \ldots, Z_{i_D}^D), \quad \mathbf{Z}_{\boldsymbol{i}_{1:j-1}} = (Z_{i_1}^1, \ldots, Z_{i_{j-1}}^{j-1}),$$

for $\boldsymbol{i} = (i_1, \ldots, i_D) \in [n]^D$ and $\boldsymbol{i}_{1:j-1} = (i_1, \ldots, i_{j-1})$, and the right-hand side of Eq. (3) is defined to be zero when the denominator is zero. Here, we use the convention $\mathbf{Z}_{\boldsymbol{i}_{1:0}}^{\mathrm{mp}(1)} := 0$ to be consistent with the notation.

Here, Eq. (2) represents a data-augmentation procedure in which new data points are created (see Fig. 1). Each new data point $\mathbf{Z}_{\boldsymbol{i}}$ is generated by the following procedure. First, D training data points are selected with replacement (specified by $\boldsymbol{i} = (i_1, \ldots, i_D) \in [n]^D$). Then, $\mathbf{Z}_{\boldsymbol{i}}$ is constructed by copying the j-th element $Z_{i_j}^j$ from $\mathbf{Z}_{i_j}$ ($j \in [D]$). Eq. (2) performs this procedure for all combinations of the indices $\boldsymbol{i} \in [n]^D$.

In the proposed data augmentation method, which we call *ADMG data augmentation*, we consider $\mathcal{D}_{\mathrm{aug}} := \{\mathbf{Z}_{\boldsymbol{i}}\}_{\boldsymbol{i} \in [n]^D}$ to be a weighted training data whose weights are $\mathcal{W}_{\mathrm{aug}} :=$

[3] For notational simplicity, we define $K^j := 1$ where j is such that $\mathrm{mp}(j) = \emptyset$.

$\{\hat{w}_{\boldsymbol{i}}\}_{\boldsymbol{i}\in[n]^D}$, and we perform supervised learning using $\mathcal{D}_{\mathrm{aug}}$ and $\mathcal{W}_{\mathrm{aug}}$, where any standard method that incorporates instance weights can be employed. As a practical device, to account for the possibility that $\hat{\mathcal{G}}$ is only an inaccurate approximation of $\mathcal{G}$, we propose to use a convex combination of the *empirical risk estimator* $\hat{R}_{\mathrm{emp}}(f) := \frac{1}{n}\sum_{i=1}^{n} \ell(f, \mathbf{Z}_i)$ and the *augmented empirical risk estimator* $\hat{R}_{\mathrm{aug}}(f)$, that is to use

$$\hat{f} \in \operatorname*{arg\,min}_{f\in\mathcal{F}} \{(1-\lambda)\hat{R}_{\mathrm{emp}}(f) + \lambda \hat{R}_{\mathrm{aug}}(f) + \Omega(f)\}$$

as the predictor, where $\lambda \in [0, 1]$ is a hyper-parameter and Ω is a regularization term for $f \in \mathcal{F}$. In the experiments in Section 5, we used a fixed parameter $\lambda = .5$ and observed that it performs reasonably for all data sets.

The ADMG data augmentation generalizes the idea described in the trivariate example $X_1 \leftarrow Y \rightarrow X_2$ in Section 1. In fact, in the trivariate example of Fig. 1, $\mathcal{W}_{\mathrm{aug}}$ places equal weights on the augmented data, essentially yielding the same augmented data set as that in Fig. 1.

Practical implementation. To reduce the computation cost of calculating the weights $\mathcal{W}_{\mathrm{aug}}$, we exploit the recursive structure in Eq. (3) that can be represented by a probability tree (Brase et al., 2012), where we sequentially select the values $i_1, \ldots, i_D \in [n]$ (Fig. 2). To see this, recursively define

$$\hat{w}_{\boldsymbol{i}_{1:0}} = 1, \quad \hat{w}_{\boldsymbol{i}_{1:j}} = \hat{w}_{i_j|\boldsymbol{i}_{1:j-1}} \cdot \hat{w}_{\boldsymbol{i}_{1:j-1}} \ (j \in [D], \boldsymbol{i}_{1:j-1} \in [n]^{j-1}),$$

where

$$\hat{w}_{i_j|\boldsymbol{i}_{1:j-1}} := \frac{K^j(\mathbf{Z}^{\mathrm{mp}(j)}_{\boldsymbol{i}_{1:j-1}} - \mathbf{Z}^{\mathrm{mp}(j)}_{i})}{\sum_{i=1}^{n} K^j(\mathbf{Z}^{\mathrm{mp}(j)}_{\boldsymbol{i}_{1:j-1}} - \mathbf{Z}^{\mathrm{mp}(j)}_{i})},$$

and the right-hand side is defined to be zero when the denominator is zero. Then, we have $\hat{w}_{\boldsymbol{i}} = \hat{w}_{\boldsymbol{i}_{1:D}}$.

With this recursive structure in mind, we construct the probability tree as follows: we index the root node by 0 and the nodes at depth $j \in [D]$ by $\boldsymbol{i}_{1:j}$ in a standard manner, assign the weight $\hat{w}_{i_j|\boldsymbol{i}_{1:j-1}}$ to each edge $(\boldsymbol{i}_{1:j-1}, \boldsymbol{i}_{1:j})$, and assign to each node $\boldsymbol{i}_{1:j}$ the product of the weights of the edges on the path from the root to $\boldsymbol{i}_{1:j}$. Then, by recursively computing the weights of the nodes on this weighted tree, we can obtain $\mathcal{W}_{\mathrm{aug}}$ (Fig. 2). Algorithm 1 summarizes the procedure of the proposed method.

To reduce the computation cost, we specify a threshold $\theta \in (0, 1)$, and we prune the branches once the node weight becomes lower than θ along the course of the recursive computation. Since the edge weights satisfy $\sum_{i_j=1}^{n} \hat{w}_{i_j|\boldsymbol{i}_{1:j-1}} \in \{0, 1\}$ and $\hat{w}_{i_j|\boldsymbol{i}_{1:j-1}} \geq 0$ for each $\boldsymbol{i}_{1:j-1}$, the node weight $\hat{w}_{\boldsymbol{i}_{1:j}}$ is monotonically decreasing in j. Therefore, the above pruning procedure only discards the nodes for which $\hat{w}_{\boldsymbol{i}} < \theta$. The worst-case computational complexity of Algorithm 1 is $O\left(n^D\right)$ (see Appendix D), and it is important in future work to explore how to effectively reduce the computation complexity. Apart from the pruning procedure, to reduce the computation time by taking advantage of the probability-tree structure, one may well consider employing heuristic top candidate search methods such as *beam search* (Bisiani, 1987) or stochastic optimization methods such as *stochastic gradient descent* (Goodfellow et al., 2016, Section 5.9).

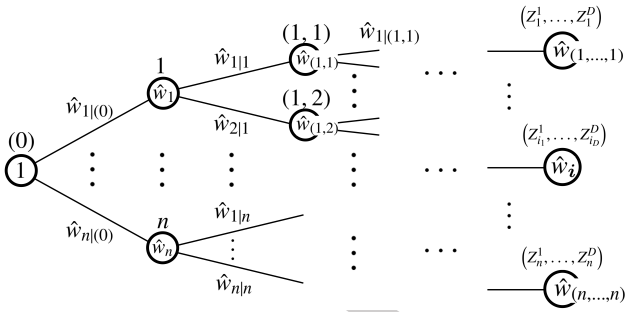


Figure 2: Probability tree to compute the weights of the augmented instances. At each depth j, the index i_j is selected and the weight is updated as $\hat{w}_{\boldsymbol{i}_{1:j}} = \hat{w}_{i_j|\boldsymbol{i}_{1:j-1}} \cdot \hat{w}_{\boldsymbol{i}_{1:j-1}}$.

Algorithm 1 Proposed method: ADMG data augmentation

Input: Training data $\mathcal{D}$, ADMG $\hat{\mathcal{G}}$, coefficient $\lambda \in [0, 1]$, regularization functional Ω, pruning threshold $\theta \in [0, 1)$, hypothesis class $\mathcal{F}$, kernel functions $\{K^j\}_{j=1}^D$, loss function ℓ.

function FillProbTree($\mathcal{D}, \hat{\mathcal{G}}, \theta, \{K^j\}_{j=1}^D$) ▷ see Fig. 2
  for $j \in [D]$ ▷ for each variable j
    for $\boldsymbol{i}_{1:j-1} \in [n]^{j-1}$ ▷ current node (depth j)
      for $i_j \in [n]$ ▷ next node (depth $j+1$)
        $\hat{w}_{\boldsymbol{i}_{1:j-1}} \leftarrow \hat{w}_{\boldsymbol{i}_{1:j-1}} \mathbb{1}\left[\hat{w}_{\boldsymbol{i}_{1:j-1}} \geq \theta\right]$ ▷ pruning
        $\hat{w}_{\boldsymbol{i}_{1:j}} \leftarrow \hat{w}_{i_j|\boldsymbol{i}_{1:j-1}} \cdot \hat{w}_{\boldsymbol{i}_{1:j-1}}$
  return $\mathcal{W}_{\mathrm{aug}} := \{\hat{w}_{\boldsymbol{i}}\}_{\boldsymbol{i}\in[n]^D}$
Let $\mathcal{W}_{\mathrm{aug}}$ = FillProbTree($\mathcal{D}, \hat{\mathcal{G}}, \theta, \{K^j\}_{j=1}^D$).
Let $\hat{R}_{\mathrm{aug}}(f) := \sum_{\boldsymbol{i}\in[n]^D} \hat{w}_{\boldsymbol{i}} \cdot \ell(f, \mathbf{Z}_{\boldsymbol{i}})$.
Let $\tilde{R}_\lambda(f) := (1-\lambda)\hat{R}_{\mathrm{emp}}(f) + \lambda \hat{R}_{\mathrm{aug}}(f) + \Omega(f)$.

Output: $\hat{f} \in \operatorname*{arg\,min}_{f\in\mathcal{F}} \tilde{R}_\lambda(f)$: the predictor.

4 THEORETICAL JUSTIFICATION

In this section, we provide a theoretical justification of the proposed method in the form of an excess risk bound, under the assumption that the CG is perfectly estimated. The goal here is to elucidate how the proposed data augmentation procedure facilitates statistical learning from a theoretical perspective. We focus on the case that $\overline{\mathcal{Z}}^j = \mathbb{R}$ for all $j \in [D]$. Select $\tilde{K}^j$ and $\boldsymbol{h} = (\boldsymbol{h}^1, \ldots, \boldsymbol{h}^D) \in \mathbb{R}^D_{>0}$, and define $K^j(u) := \frac{1}{|\det \mathbf{H}_j|} \tilde{K}^j(\mathbf{H}_j^{-1} u)$, where $\mathbf{H}_j := \mathrm{diag}(\boldsymbol{h}^{\mathrm{mp}(j)})$ is a diagonal matrix with elements $\boldsymbol{h}^{\mathrm{mp}(j)}$.

For function classes, we quantify their complexities using the Rademacher complexity.

Definition 1 (Rademacher complexity)**.** *Let q denote a probability distribution on some measurable space $\mathcal{X}$. For a function class $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$, define*

$$\mathrm{Rad}_{m,q}(\mathcal{F}) := \mathbb{E}_q \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(X_i) \right| \right],$$

where $\{\sigma_i\}_{i=1}^m$ are independent uniform $\{\pm 1\}$-valued random variables, and $\{X_i\}_{i=1}^m \overset{i.i.d.}{\sim} q$.

To state our result, let us define the set of marginalized functions and that of the shifted kernel functions as

$$\mathcal{L}_{\mathcal{F}}^{j} := \left\{ \ell_{f,j}(\boldsymbol{z}^1, \ldots, \boldsymbol{z}^{j-1}, \cdot) : f \in \mathcal{F}, (\boldsymbol{z}^1, \ldots, \boldsymbol{z}^{j-1}) \in \mathcal{Z}^{[1:j-1]} \right\},$$

$$\left(\ell_{f,j} : \begin{pmatrix} \boldsymbol{z}^1 \\ \vdots \\ \boldsymbol{z}^j \end{pmatrix} \mapsto \int \ell(f, \boldsymbol{z}) \left(\prod_{k=j+1}^{D} p_{k|\mathrm{mp}(k)}(\boldsymbol{z}^k | \boldsymbol{z}^{\mathrm{mp}(k)}) \right) \mathrm{d}\boldsymbol{z}^{[j+1:D]} \right),$$

$$\mathcal{K}_{\mathbf{H}}^{j} := \left\{ K^j(\boldsymbol{z}^{\mathrm{mp}(j)} - (\cdot)) : \boldsymbol{z}^{\mathrm{mp}(j)} \in \mathcal{Z}^{\mathrm{mp}(j)} \right\},$$

where the integration is over $\mathcal{Z}^{[j+1:D]}$.

Theorem 1 (Excess risk bound)**.** *Let $\hat{f} \in \arg\min_{f \in \mathcal{F}} \{\hat{R}_{\mathrm{aug}}(f)\}$ and $f^* \in \arg\min_{f \in \mathcal{F}} \{R(f)\}$, assuming both exist. Assume $\hat{\mathcal{G}} = \mathcal{G}$ and also assume that $\mathcal{Z}^j \subset \mathbb{R}$ is compact. Let $p_{\mathrm{mp}(j)}$ and $p_{j,\mathrm{mp}(j)}$ denote the marginal density of $\mathbf{Z}^{\mathrm{mp}(j)}$ and the joint density of $(Z^j, \mathbf{Z}^{\mathrm{mp}(j)})$, respectively, and assume $p_{\mathrm{mp}(j)}$ and $p_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j, \cdot)$ $(\boldsymbol{z}^j \in \mathcal{Z}^j)$ have extensions to the entire $\mathbb{R}^{|\mathrm{mp}(j)|}$ belonging to $\Sigma(\beta, L)$, where $\Sigma(\beta, L)$ denotes the Hölder class of functions, $\beta > 1$, and $L > 0$. Define*

$$R_{\mathbf{H}} := \sum_{j=1}^{D} \left(\max_{j' \in \mathrm{mp}(j)} \boldsymbol{h}^{j'} \right)^{\beta}, \quad R_K := \sum_{j=1}^{D} \left| \det \mathbf{H}_j \right| \mathrm{Rad}_{n,p}(\mathcal{K}_{\mathbf{H}}^{j}),$$

$$R_{\mathcal{F},K} := \sum_{j=1}^{D} \left| \det \mathbf{H}_j \right| \mathrm{Rad}_{n,p}\left(\mathcal{L}_{\mathcal{F}}^{j} \otimes \mathcal{K}_{\mathbf{H}}^{j} \right).$$

Under additional assumptions on the boundedness and smoothness of the kernels and the underlying densities (see Theorem 1 in Appendix C.2), there exist $C_1, C_p, C_2, C_3, C_4 > 0$ depending on the boundedness and the smoothness of $p, \ell, \{\tilde{K}^j\}_{j=1}^D$, and $\mathbf{H}$, such that for any $\delta \in (0,1)$, we have with probability at least $1 - \delta$,

$$R(\hat{f}) - R(f^*) \leq \underbrace{C_1 R_{\mathbf{H}} + C_p}_{\textit{Kernel Bias}} + \underbrace{C_2 R_K}_{\textit{Kernel Complexity}} + \underbrace{C_3 R_{\mathcal{F},K}}_{\textit{Hypothesis Complexity}} + \underbrace{C_4 \sqrt{\frac{\log(4D/\delta)}{2n}}}_{\textit{Uncertainty}}.$$

A proof is provided in Appendix C.2. Note that the existence of a smooth extension is satisfied by, e.g., a truncated version of a smooth density on $\mathbb{R}^{|\mathrm{mp}(j)|}$.

Implications. Theorem 1 implies that the proposed method contributes to statistical learning by reducing the apparent complexity of the hypothesis class at the cost of introducing the additional complexity and bias arising from the kernel approximations. In the interest of space, we provide a formal assessment of this complexity reduction effect in Proposition 2 in Appendix C.3 under some additional Lipschitz-continuity assumptions. In the derivation of Proposition 2 indicating the complexity reduction effect, the fact that $\mathcal{L}_{\mathcal{F}}^{j}$ consists of univariate functions is critical. In Section 5, we empirically confirm that the complexity reduction effect is worth the newly introduced bias and complexity due to the kernel approximation in practice.

Scope of the analysis. It should be noted that the present theoretical guarantee only covers the case that the conditional independence assumptions implied by the CG are correct. The robustness of the proposed method to the conditional independence assumptions is an important area of research in future work.

5 REAL-WORLD DATA EXPERIMENT

In this section, we report the results of the real-world data experiments to demonstrate the effectiveness of the proposed method in improving the prediction accuracy.

5.1 EXPERIMENT SETUP

The goal of this experiment is to confirm that the proposed method contributes to the performance of the trained predictor, especially in the small-data regime. To investigate the performance improvement, we make a comparison between the two cases: training with and without the proposed device, using the same hypothesis class and the same training algorithm. To analyze the performance improvement in relation to the sample size, we vary the fraction of the data used for training the predictor and compare the performances of the proposed method and that of the baseline without a device. For further details omitted here for the space limitation, please refer to Appendix B.

Data sets. We employ 6 data sets for the experiment, namely *Sachs* (Sachs et al., 2005), *GSS* (Shimizu et al., 2011), *Boston Housing* (Harrison et al., 1978), *Auto MPG* (Quinlan, 1993), *White Wine* (Cortez et al., 2009), and *Red Wine* (Cortez et al., 2009). Table 1 summarizes these data sets. The *Sachs* data and the *GSS* data are accompanied by the ADMGs obtained from domain experts (Fig. 3(b) and Fig. 3(a), respectively), and hence we use them in the experiment. For the other data sets, we first perform *DirectLiNGAM* (Shimizu et al., 2011) on the entire data set to obtain the estimated CGs, simulating a situation that we have background knowledge from domain experts. Since

DirectLiNGAM produces DAGs, the CGs used in this experiment are DAGs except for the case of *GSS* data set which is accompanied by an ADMG produced by domain experts (Fig. 3(b)).

Predictor model class. We employ the gradient boosted regression trees (Friedman, 2001; Chen et al., 2016) as the predictor model class. The hypothesis class consists of the convex combinations of binary regression trees with at most M leaves:

$$\mathcal{F}_{M,K} := \left\{ \sum_{k=1}^{K} \alpha^k w_k^{h_k(\cdot)} : \alpha \in \Delta_K, T_k \in [M], w_k \in \mathbb{R}^{T_k}, h_k \in \mathcal{T}_{T_k} \right\},$$

where $M, K \in \mathbb{N}$, $\mathcal{T}_T$ represents the set of binary tree structures mapping $\mathcal{X}$ to $[T]$, and Δ_K is the $(K-1)$-dimensional probability simplex. The loss function is the squared error $\ell(f, \mathbf{Z}) = (Y - f(\mathbf{X}))^2$ where $Y = Z^{j^*}$ and $\mathbf{X} = \mathbf{Z}^{[D]\setminus\{j^*\}}$, and the regularization function is $\Omega(f) = \sum_{k=1}^{K} \frac{\rho}{2} \|w_k\|^2$ $(\rho > 0)$. We fix $M = 64$ and search the number of boosting rounds K in $\{10, 50, 250, 1250\}$ and the ℓ_2-regularization coefficient ρ in $\{1, 10, 100, 1000\}$. The hyper-parameters are selected by the grid-search based on 3-fold weighted cross-validation. Note that, for the proposed method, we perform cross-validation on the union of the original training data and the augmented data with the weights adjusted by λ, namely $\mathcal{D} \coprod \mathcal{D}_{\text{aug}}$ with weights $(1-\lambda)\mathcal{W}_{\text{orig}} \coprod \lambda \mathcal{W}_{\text{aug}}$ where $\mathcal{W}_{\text{orig}} = (\frac{1}{n}, \ldots, \frac{1}{n})$.

Configurations of the proposed method. We select $\boldsymbol{h} = (\boldsymbol{h}^1, \ldots, \boldsymbol{h}^D) \in \mathbb{R}_{>0}^D$ and use the product kernel $K^j(\boldsymbol{x} - \boldsymbol{y}) := \prod_{j' \in \mathfrak{mp}(j)} \frac{1}{\boldsymbol{h}^{j'}} K_{j'}^j \left(\frac{\boldsymbol{x}^{j'} - \boldsymbol{y}^{j'}}{\boldsymbol{h}^{j'}} \right)$ for the proposed method. For each $j' \in \mathfrak{mp}(j)$, if the variable is continuous (i.e., $\overline{\mathcal{Z}}^{j'} = \mathbb{R}$), we use the Gaussian kernel $K_{j'}^j(x - y) := (2\pi)^{-1/2} \exp\left(-\frac{(x-y)^2}{2}\right)$. Otherwise, i.e., if the variable is discrete, we use the identity kernel $K_{j'}^j(x - y) := \mathbb{1}[x = y]$ and $\boldsymbol{h}^{j'} = 1$. For the Gaussian kernels, we select the *kernel bandwidth* $\boldsymbol{h}^{j'}$ based on *Silverman's rule-of-thumb* (Silverman, 1986, pp.45–47). In the experiment, we fix $\lambda = .5$ throughout all runs and find that it yields reasonable performances in all data sets.

Compared methods. We compare the performances of the proposed method and the naive baseline method without a device:

$$\hat{f} \in \arg\min_{f \in \mathcal{F}} \{\hat{R}_{\text{emp}}(f) + \Omega(f)\}.$$

In Section 5.2 where we report the results, the two methods are referred to as *Proposed* and *Baseline*, respectively.

Evaluation procedure. The prediction accuracy is measured by the mean squared error (MSE). For each data set, we randomly subsample a fraction of the data as the training set and use the rest as the testing set. The fraction of the training set is varied in $\{.1, .15, \ldots, .85\}$. For each training set fraction, random train-test splits are performed 20 times. Subsequently, for each split, *Proposed* and *Baseline* are trained on the training set, and then evaluated on the testing set. We report the average performances as well as the standard errors over the 20 runs for each training set fraction.

Table 1: Summary of Data Sets (*NAME*: name of the data set, *#VAR*: number of variables in the data set, *#OBS*: number of observations, *GRAPH*: CG used for the proposed method, *Consensus*: consensus network (Fig. 3(b)), *Domain*: domain knowledge of the status attainment model (Fig. 3(a)), *LiNGAM*: CG is estimated by performing DirectLiNGAM on the entire data set).

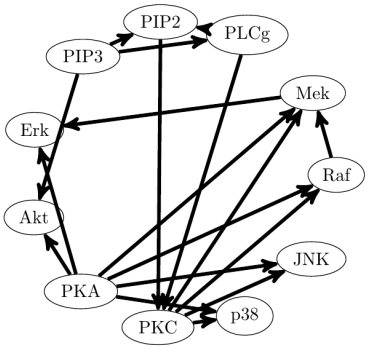
NAME	#VAR	#OBS	GRAPH
Sachs	11	853	Consensus
GSS	6	1380	Domain
Boston Housing	14	506	LiNGAM
Auto MPG	7	392	LiNGAM
White Wine	12	4898	LiNGAM
Red Wine	12	1599	LiNGAM

5.2 RESULTS

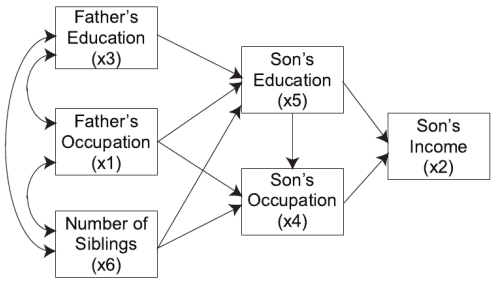
Fig. 4 shows the experimental result. We observe a consistent performance improvement in most of the data sets. For the data sets for which the domain knowledge CG is provided (i.e., *Sachs* and *GSS*), we can see clear relative improvement ranging from 3% to 7% on average, especially in the small-data regime where approximately 10–40% is the training set fraction. In the other data sets without the background knowledge, relatively little improvement is observed except in the small-data regions of *Red Wine* and *White Wine*, where up to 4% relative improvement on average is observed. The lack of relative improvement in the majority of these cases emphasizes the importance of having accurate domain knowledge in the proposed approach, and it motivates the development of effective causal discovery methods. In the *White Wine* data, the proposed method coincides with the baseline in the larger-data region as the augmentation did not effectively take place due to the adaptive bandwidth that is narrowed according to the sample size. For supplementary figures visualizing the average relative improvements, see Appendix B.5.

6 RELATED WORK AND DISCUSSION

In this section, we explain the context of the paper in relation to existing work.

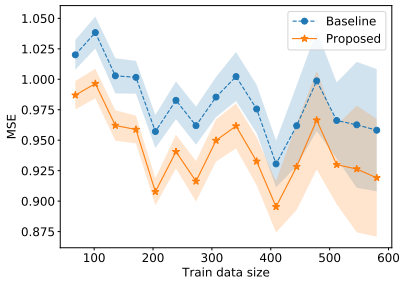


(a) Reference graph for Sachs data. Figure excerpted from Mooij et al. (2013).

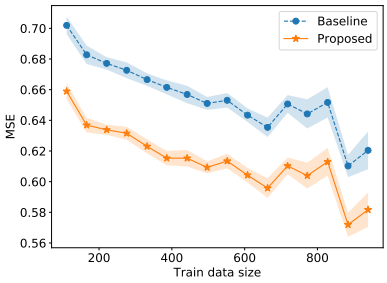


(b) Reference graph for GSS data. Figure excerpted from Shimizu et al. (2011).

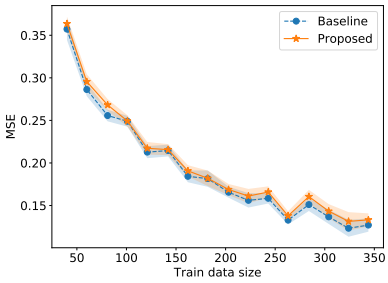
Figure 3: Reference CGs for the data sets used in our experiments. (a) Consensus graph in Sachs et al. (2005). (b) Domain-knowledge graph based on the status attainment model (Duncan et al., 1972).



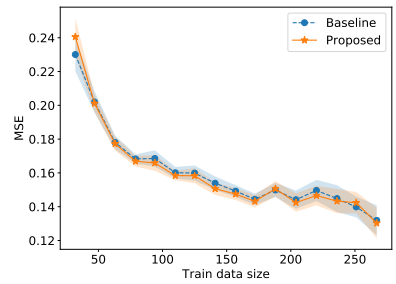
(a) Sachs data.



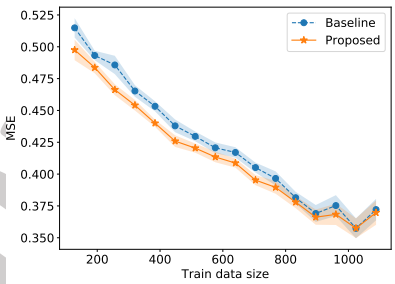
(b) GSS data.



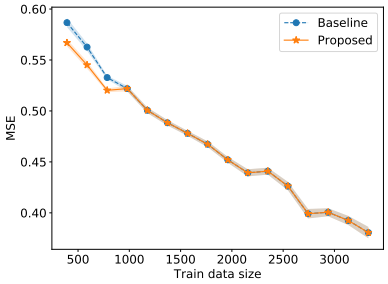
(c) Boston Housing data.



(d) Auto MPG data.



(e) Red Wine data.



(f) White Wine data.

Figure 4: Illustration of the experimental results. In all figures, the horizontal axis is the varied size of the training data before augmentation, and the vertical axis is the performance metric (MSE; the lower the better). The markers and the lines indicate the average over the 20 independent runs, and the shades are drawn for the width of the standard errors both above and below the lines. The proposed method shows a consistent improvement over the naive baseline based on the empirical risk minimization with the same hypothesis class, particularly in the small-data regime.

6.1 CGMS AND PREDICTIVE MODELING

Variable selection in a single-distribution setting. The background knowledge encoded in a CG can be used for variable selection by identifying a *Markov boundary* of the target variable. Here, $\mathfrak{mb}(j) \subset [D]$ is called a *Markov blanket* of j if Z^j is conditionally independent of all the other variables given $\mathbf{Z}^{\mathfrak{mb}(j)}$. If, moreover, $\mathfrak{mb}(j)$ is minimal, i.e., if none of its proper subsets are Markov blankets, it is called a *Markov boundary* (MB). Under certain assumptions, the MB of a target variable is known to be the minimal set of variables with optimal predictive performance (Tsamardinos et al., 2003). For a recent comprehensive review on MB estimation, see Yu et al. (2020). The present paper is

orthogonal to this line of work. In fact, the CGs can encode more information than a specification of the Markov boundary of the predicted variable; for example, consider the CG $X_1 \leftarrow Y \rightarrow X_2$ where Y is the target variable and (X_1, X_2) are the predictors. In this case, the Markov boundary of Y is (X_1, X_2), and hence the variable selection does not reduce the number of the predictors. On the other hand, the proposed method still leverages the factorization structure of the data distribution entailing the CG. In practice, the two approaches can be combined straightforwardly. In our experiments, we do not perform variable selection using the data regarding the possibility that the obtained CGs are inaccurate.

Variable selection in distribution-shift setting. Another line of research is concerned with making predictions under distribution shift and leverage feature selection based on causal background knowledge or causal discovery. Magliacane et al. (2018) considered the case that a distribution shift is due to intervention in some variables, and they proposed a method to perform domain adaptation by identifying a set of variables that is likely to perform well regardless of the intervention. Rojas-Carulla et al. (2018) assume that if the conditional distribution of the predicted variable given some subset of features is invariant across different distributions, then this conditional distribution is the same in the *target distribution* for which one wants to make good predictions, and leveraged it to find the set of variables for which the relation to the target variable does not change. The present paper is complementary to this line of work since our goal is to make good predictions in a single fixed distribution.

Regularization and model selection. Kyono et al. (2019) proposed a model selection criterion that can reflect the structure of a CG. The goal of Kyono et al. (2019) is *domain generalization* and *out-of-distribution prediction*, i.e., making good predictions under a distribution shift without access to any samples from the target distribution or making good predictions for the data that is outside the support of the training data distribution. To achieve it, given a DAG as prior knowledge, Kyono et al. (2019) first modify it so that the edges coming out of the target variable are removed. Then, to score the predictor model candidates, it generates a data set whose predicted variables are replaced by the predictions of the model and computes the *Bayes Information Criterion* (BIC) that evaluates the fitness of the modified DAG structure to the generated data set. Another approach for using the background knowledge of a CG is the *CASTLE regularization* (Kyono et al., 2020). CASTLE regularization regularizes a neural network while performing the CG discovery as an auxiliary task. The method imposes a reconstruction loss using the internal layers of the predictor implemented by neural networks under a DAG constraint. The present paper is orthogonal to these researches and can be straightforwardly combined in practice. Also note that our method has a theoretical justification while Kyono et al. (2019) provided no theoretical justifications.

Inference under specific CGs. Under some specific problem settings with known specific underlying CGs, methods to take advantage of the prior knowledge have been developed. For example, in the instance weight estimation for episodic reinforcement learning, methods to perform *state simplification* based on the CGs have been proposed (Bottou et al., 2013; Peters et al., 2017, Section 8.2). Schölkopf et al. (2015) considered removing systematic errors using *half-sibling regression* inspired by the CG of the observation mechanism found in the *exoplanet search*. Pitis et al. (2020) proposed a method to enhance the sample efficiency in reinforcement learning (RL) by a procedure to exchange the realizations of the variables within the (conditionally) disconnected components in the CG of the *Markov decision process* of specific RL instances. This line of work and the present work are complementary in that our approach is widely applicable to general ADMGs whereas these analyses have the potential to exploit the characteristics of the specific problem setups.

Causal bootstrapping. Recently, Little et al. (2020) proposed *causal bootstrapping*, a weighted bootstrap-type algorithm that is relevant to our method. While, methodologically, both the present paper and Little et al. (2020) can be seen to be based on kernel-type function estimators (Stute, 1986; Horváth et al., 1988; Einmahl et al., 2000) and CGs (Pearl, 2009), the two works are complementary in that the problem setups differ. Causal bootstrapping of Little et al. (2020) aims at mitigating the performance degradation due to a distribution shift arising from an intervention, and it uses kernel-type function estimators to simulate sampling from an interventional distribution. On the other hand, we investigate the performance improvement yielded from using the background knowledge of a CG in a scenario without a distribution shift.

Constructing probabilistic graphical models. Evans et al. (2014) provided a smooth parametrization of the set of distributions that are *Markov with respect to* an ADMG $\mathcal{G}$ in the binary case: $\overline{\mathcal{Z}}^j = \{0, 1\}$ $(j \in [D])$. Complementarily, for the case of $\overline{\mathcal{Z}}^j = \mathbb{R}$ $(j \in [D])$, Silva et al. (2011) proposed the construction of flexible probability models that are Markov with respect to a given ADMG. Similarly, in the case that the ADMG has no bi-directed edges, constructing a Bayesian network by specifying the conditional distributions appearing in the Markov factorization (Eq. (1)) is one natural way to exploit this prior knowledge (Lucas et al., 2004). This approach has the limitation that it inevitably restricts the modeling choice, e.g., the constructed predictor is a generative model as opposed to a discriminative model (Shalev-Shwartz et al., 2014, Chapter 24), whereas our approach has the virtue of being model-agnostic.

6.2 CAUSAL DISCOVERY AND TRANSFER LEARNING

Our method provides a channel through which an estimated CG can be used for enhancing the predictive modeling. In this sense, the proposed method can serve as a transfer learning method under a *transfer assumption* of *common CG*, i.e., an assumption that one is given many samples from another distribution sharing the same CG with the distribution for which we want to make the predictions. Under such an assumption, one may first estimate the ADMG using causal discovery methods to estimate the *Markov equivalence class* of ADMGs expressed as a *partial ancestral graph* (PAG) (Zhang, 2008), e.g., the *fast causal inference* (FCI) algorithm (Spirtes et al., 1995; Zhang, 2008), enumerate the ADMGs in the equivalence class (e.g., by the *Pag2admg* algorithm; Subramani, 2018), select a plausible candidate ADMG that is concordant with the domain knowledge, and apply the proposed method. Such an assumption of a common causal mechanism has been exploited in recent work of causal discovery (Xu et al., 2014; Ghassami et al., 2017; Monti et al., 2019) and transfer learning (Pearl et al., 2011; Magliacane et al., 2018; Teshima et al., 2020), and it is based on a common belief that a causal mechanism remains invariant unless explicitly intervened in (Hünermund et al., 2019).

6.3 CGMS AND EFFICIENT ESTIMATION

Our method could be also seen as a method to perform sample-efficient inference given a CG. In the existing work, the knowledge of a CG has been used for deriving efficient estimators for *identifiable causal estimands* (Pearl, 2009) such as the *interventional distributions* (Jung et al., 2021b; Jung et al., 2021a) or the *average causal effect* (Bhattacharya et al., 2020). For instance, Jung et al. (2021b) and Jung et al. (2021a) derived expressions of efficient estimators of the identifiable interventional distributions given an ADMG and a PAG, respectively, by leveraging the knowledge of the CG in the *double/debiased machine learning* (Chernozhukov et al., 2018) framework. Another line of research provided graphical criteria for selecting the *efficient adjustment sets*, the set of covariates to be adjusted for producing a valid estimator of a causal effect with the minimal asymptotic variance (Henckel et al., 2020; Rotnitzky et al., 2020; Witte et al., 2020; Smucler et al., 2021). Our goal differs from the goals of these lines of research; we are interested in improving the sample efficiency of training the predictor whereas they aimed to improve the sample efficiency of causal inference. Nevertheless, it is an interesting direction of future research to elucidate whether the proposed method is optimally efficient in estimating the risk functional given the CG.

7 CONCLUSION

In this paper, we proposed a general method for exploiting the causal prior knowledge in predictive modeling. We theoretically provided an excess risk bound indicating that the proposed method has a complexity reduction effect that mitigates overfitting while it introduces additional complexity and bias arising from the kernel approximations. Through the experiments using real-world data, we demonstrated that the proposed method consistently improves the predictive performance especially in the small-data regime, which implies that the complexity reduction effect is worth the newly introduced bias and complexity in practice. Important areas in future work include incorporating the equality constraints imposed by an ADMG but not captured by the topological ADMG factorization and handling more relaxed assumptions such as those expressed as PAGs.

Author Contributions

TT contributed to the conception of the research, conducted the theoretical analysis and experiments, and drafted the paper. MS contributed to the conception of the research and revised the draft.

Acknowledgements

The authors are grateful to Prof. Shohei Shimizu for providing them with the preprocessed GSS data set used in Shimizu et al. (2011). We also thank Han Bao and Kenshin Abe for proofreading the manuscript. We would also like to thank Kento Nozawa and Yoshihiro Nagano for maintaining the computational resources used for our experiments. This work was supported by RIKEN Junior Research Associate Program. TT was supported by Masason Foundation. MS was supported by JST CREST Grant Number JPMJCR18A2.

REFERENCES

Bartlett, P. L. et al. (2002). “Rademacher and Gaussian Complexities: Risk Bounds and Structural Results.” In: *Journal of Machine Learning Research* 3.Nov, pp. 463–482.

Bhattacharya, R. et al. (2020). “Semiparametric Inference for Causal Effects in Graphical Models with Hidden Variables.” In: *arXiv:2003.12659 [stat.ML]*.

Bisiani, R. (1987). “Beam Search.” In: *Encyclopedia of Artificial Intelligence*, pp. 56–58.

Bottou, L. et al. (2013). “Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising.” In: *Journal of Machine Learning Research* 14.65, pp. 3207–3260.

Brase, C. H. et al. (2012). *Understanding Basic Statistics*. Cengage Learning.

Chen, T. et al. (2016). "XGBoost: A Scalable Tree Boosting System." In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794.

Chernozhukov, V. et al. (2018). "Double/Debiased Machine Learning for Treatment and Structural Parameters." In: *The Econometrics Journal* 21.1, pp. C1–C68.

Chickering, D. M. (2002). "Optimal Structure Identification with Greedy Search." In: *The Journal of Machine Learning Research* 3, pp. 507–554.

Cortez, P. et al. (2009). "Modeling Wine Preferences by Data Mining from Physicochemical Properties." In: *Decision support systems* 47.4, pp. 547–553.

Duncan, O. D. et al. (1972). *Socioeconomic Background and Achievement*. New York: Seminar Press.

Einmahl, U. et al. (2000). "An Empirical Process Approach to the Uniform Consistency of Kernel-Type Function Estimators." In: *Journal of Theoretical Probability* 13.1, pp. 1–37.

Evans, R. et al. (2014). "Markovian Acyclic Directed Mixed Graphs for Discrete Data." In: *The Annals of Statistics* 42.4, pp. 1452–1482.

Friedman, J. H. (2001). "Greedy Function Approximation: A Gradient Boosting Machine." In: *Annals of Statistics* 29.5, pp. 1189–1232.

Ghassami, A. et al. (2017). "Learning Causal Structures Using Regression Invariance." In: *Advances in Neural Information Processing Systems 30*, pp. 3011–3021.

Goodfellow, I. et al. (2016). *Deep Learning*. MIT Press.

Harrison, D. et al. (1978). "Hedonic Housing Prices and the Demand for Clean Air." In: *Journal of Environmental Economics and Management* 5.1, pp. 81–102.

Henckel, L. et al. (2020). "Graphical Criteria for Efficient Total Effect Estimation via Adjustment in Causal Linear Models." In: *arXiv:1907.02435 [math, stat]*.

Horváth, L. et al. (1988). "Asymptotics of Conditional Empirical Processes." In: *Journal of Multivariate Analysis* 26.2, pp. 184–206.

Hünermund, P. et al. (2019). "Causal Inference and Data-Fusion in Econometrics." In: *arXiv:1912.09104 [econ.EM]*.

Jung, Y. et al. (2021a). "Estimating Identifiable Causal Effects on Markov Equivalence Class through Double Machine Learning." In: *Proceedings of the 38th International Conference on Machine Learning*, p. 10.

– (2021b). "Estimating Identifiable Causal Effects through Double Machine Learning." In: *Proceedings of the 35th AAAI Conference on Artificial Intelligence*. Vol. 35, pp. 12113–12122.

Kyono, T. et al. (2019). "Improving Model Robustness Using Causal Knowledge." In: *arXiv:1911.12441 [cs.LG]*.

Kyono, T. et al. (2020). "CASTLE: Regularization via Auxiliary Causal Graph Discovery." In: *Advances in Neural Information Processing Systems 33*.

Little, M. A. et al. (2020). "Causal Bootstrapping." In: *arXiv:1910.09648 [cs.LG]*.

Lucas, P. J. F. et al. (2004). "Bayesian Networks in Biomedicine and Health-Care." In: *Artificial Intelligence in Medicine* 30.3, pp. 201–214.

Magliacane, S. et al. (2018). "Domain Adaptation by Using Causal Inference to Predict Invariant Conditional Distributions." In: *Advances in Neural Information Processing Systems 31*, pp. 10846–10856.

Monti, R. P. et al. (2019). "Causal Discovery with General Non-Linear Relationships Using Non-Linear ICA." In: *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 186–195.

Mooij, J. M. et al. (2013). "Cyclic Causal Discovery from Continuous Equilibrium Data." In: *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pp. 431–439.

Nadaraya, E. A. (1964). "On Estimating Regression." In: *Theory of Probability & Its Applications* 9.1, pp. 141–142.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Second. Cambridge, U.K. ; New York: Cambridge University Press.

Pearl, J. et al. (2011). "Transportability of Causal and Statistical Relations: A Formal Approach." In: *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, pp. 247–254.

Peters, J. et al. (2014). "Causal Discovery with Continuous Additive Noise Models." In: *Journal of Machine Learning Research* 15.June, pp. 2009–2053.

Peters, J. et al. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge, Massachusetts: The MIT Press.

Pitis, S. et al. (2020). "Counterfactual Data Augmentation Using Locally Factored Dynamics." In: *Advances in Neural Information Processing Systems 33*.

Quinlan, J. R. (1993). "Combining Instance-Based and Model-Based Learning." In: *Proceedings of the Tenth International Conference on Machine Learning*, pp. 236–243.

Richardson, T. (2003). "Markov Properties for Acyclic Directed Mixed Graphs." In: *Scandinavian Journal of Statistics* 30.1, pp. 145–157.

Richardson, T. S. et al. (2017). "Nested Markov Properties for Acyclic Directed Mixed Graphs." In: *arXiv:1701.06686 [stat.ME]*.

Rojas-Carulla, M. et al. (2018). "Invariant Models for Causal Transfer Learning." In: *Journal of Machine Learning Research* 19.36, pp. 1–34.

Rotnitzky, A. et al. (2020). "Efficient Adjustment Sets for Population Average Causal Treatment Effect Estimation

in Graphical Models." In: *Journal of Machine Learning Research* 21.188, pp. 1–86.

Sachs, K. et al. (2005). "Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data." In: *Science* 308.5721, pp. 523–529.

Schölkopf, B. et al. (2015). "Removing Systematic Errors for Exoplanet Search via Latent Causes." In: *Proceedings of the 32nd International Conference on Machine Learning*, pp. 2218–2226.

Shalev-Shwartz, S. et al. (2014). *Understanding Machine Learning: From Theory to Algorithms*. New York, NY, USA: Cambridge University Press.

Shimizu, S. et al. (2006). "A Linear Non-Gaussian Acyclic Model for Causal Discovery." In: *The Journal of Machine Learning Research* 7.72, pp. 2003–2030.

Shimizu, S. et al. (2011). "DirectLiNGAM: A Direct Method for Learning a Linear Non-Gaussian Structural Equation Model." In: *Journal of Machine Learning Research* 12.33, pp. 1225–1248.

Silva, R. et al. (2011). "Mixed Cumulative Distribution Networks." In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 670–678.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. 1st. Chapman and Hall/CRC.

Smucler, E. et al. (2021). "Efficient Adjustment Sets in Causal Graphical Models with Hidden Variables." In: *Biometrika*.

Spirtes, P. et al. (1995). "Causal Inference in the Presence of Latent Variables and Selection Bias." In: *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pp. 499–506.

Spirtes, P. et al. (2000). *Causation, Prediction, and Search*. Second. Cambridge, Massachusetts: MIT Press.

Stute, W. (1986). "Conditional Empirical Processes." In: *Annals of Statistics* 14.2, pp. 638–647.

Subramani, N. (2018). "Pag2admg: An Algorithm for the Complete Causal Enumeration of a Markov Equivalence Class." In: *Proceedings of the CausalML Workshop at ICML*.

Teshima, T. et al. (2020). "Few-Shot Domain Adaptation by Causal Mechanism Transfer." In: *Proceedings of the 37th International Conference on Machine Learning*, pp. 9458–9469.

Tian, J. et al. (2002). "A General Identification Condition for Causal Effects." In: *Proceedings of the Eighteenth National Conference on Artificial Intelligence*, pp. 567–573.

Tsamardinos, I. et al. (2003). "Towards Principled Feature Selection: Relevancy, Filters and Wrappers." In: *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*.

Watson, G. S. (1964). "Smooth Regression Analysis." In: *Sankhyā: The Indian Journal of Statistics, Series A* 26.4, pp. 359–372.

Witte, J. et al. (2020). "On Efficient Adjustment in Causal Graphs." In: *Journal of Machine Learning Research* 21.246, pp. 1–45.

Xu, L. et al. (2014). "A Pooling-LiNGAM Algorithm for Effective Connectivity Analysis of fMRI Data." In: *Frontiers in Computational Neuroscience* 8, p. 125.

Yu, K. et al. (2020). "Causality-Based Feature Selection: Methods and Evaluations." In: *ACM Computing Surveys* 53.5, pp. 1–36.

Zhang, J. (2008). "On the Completeness of Orientation Rules for Causal Discovery in the Presence of Latent Confounders and Selection Bias." In: *Artificial Intelligence* 172.16, pp. 1873–1896.

Zorich, V. A. (2015). *Mathematical Analysis I*. Second. Berlin, Heidelberg: Springer Berlin Heidelberg.