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

A Preprint

Takeshi Teshima
The University of Tokyo, RIKEN
teshima@ms.k.u-tokyo.ac.jp

Masashi Sugiyama
RIKEN, The University of Tokyo
sugi@k.u-tokyo.ac.jp

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.

1 Introduction

Causal graphs (CGs; [1]) 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 [2] or sociology [3]. Otherwise, it may also be learned from data using the statistical causal discovery methods developed over the last decades [1, 4–8]. Once a CG is obtained, it can be used to infer the conditional independence (CI) relations that the data distribution should satisfy [1].

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 [4, 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. [9] for a review). However, most of them are based on the notion of the *Markov blanket* or the *Markov boundary* [10]. 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 [11]. Another approach to exploiting the prior knowledge of a CG is to build a *Bayesian network* model according to the CG structure (e.g., [12]). However, this approach inevitably restricts the modeling choice, and it can prohibit the use of flexible and effective models such as tree-based predictors [13] and neural networks [14] that may be preferred in 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. 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 [1]. 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 (Figure 1).

This procedure seems intuitively plausible, but extending it to general graphs is not straightforward since general CI statements involve only subsets of the variables; naively applying the above procedure to each CI statement would generate only a partial data vector.

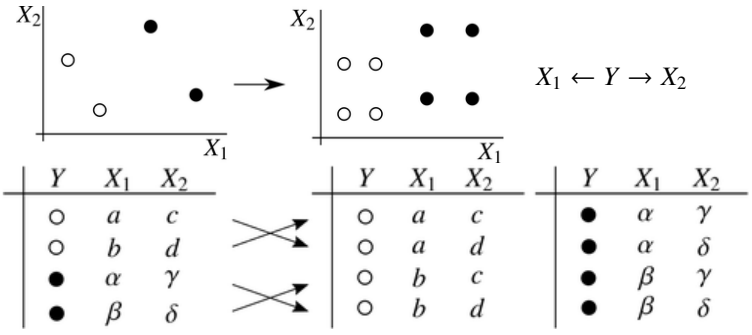


Figure 1: Visualization of the basic idea of the paper for the trivariate case $X_1 \leftarrow Y \rightarrow X_2$. In this case, the CI $X_1 \perp\!\!\!\perp X_2 \mid Y$ holds. One way to use this knowledge via data augmentation is to group the data according to Y and then to shuffle X_1 and X_2 within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

Our contributions. (i) We propose a method to augment data based on the prior knowledge expressed as CGs. (ii) We theoretically justify the proposed method via an excess risk bound based on the Rademacher complexity [15]. 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 2 in Supplementary Material 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; [11, 16]), 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 $\mathcal{G}$ [17], 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* [17] of $j \in [D]$ in $\mathcal{G}$ (see Appendix A for the definition), and $p_{j|\mathrm{mp}(j)}$ denotes the conditional density of Z^j given $\mathbf{Z}^{\mathrm{mp}(j)}$. The Markov pillow is a generalization of the notion of parents for graphs with bi-directed edges; if all edges are uni-directed, $\mathrm{mp}(j)$ matches the parent of j, and hence Equation (1) is a generalization of the usual Markov factorization with respect to directed acyclic graphs (DAGs; [1, 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, Equation (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.

[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.

Details on the assumption. ADMGs are used to represent *semi-Markovian causal graphical models* (CGMs; [18]), which are CGMs allowing for the existence of hidden confounders. The assumption of topological ADMG factorization is satisfied by such CGMs [18]. We refer the readers to Section 2 of Richardson et al. [16] for an overview of ADMGs and their use in CGMs involving latent variables. Note that the topological ADMG factorization, in general, captures only part of the equality constraints imposed by an ADMG on a semi-Markov model [17]. Indeed, Bhattacharya et al. [17] 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.

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 a whole set of new data vector, we successively resample each variable from the *conditional empirical distribution* [19, 20] 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 Figure 1.

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

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

Then, to formulate the nested conditional resampling procedure, we select a kernel function $K^j : \overline{\mathcal{Z}}^{\mathfrak{mp}(j)} \to \mathbb{R}_{\geq 0}$ for each $j \in [D]$.[3] Using this kernel function in the spirit of kernel-type function estimators [21–23], we approximate each conditional density $(*)$ as

$$\hat{p}_{j|\mathfrak{mp}(j)}(Z^j|\mathbf{Z}^{\mathfrak{mp}(j)}) := \frac{\sum_{i=1}^{n} \delta_{Z_i^j}(Z^j) K^j(\mathbf{Z}^{\mathfrak{mp}(j)} - \mathbf{Z}_i^{\mathfrak{mp}(j)})}{\sum_{k=1}^{n} K^j(\mathbf{Z}^{\mathfrak{mp}(j)} - \mathbf{Z}_k^{\mathfrak{mp}(j)})} I_{\neq 0},$$

where $\delta_{\boldsymbol{z}}$ denotes Dirac's delta function centered at $\boldsymbol{z}$ (e.g., [24, Section E.4.1]), and $I_{\neq 0} := \mathbb{1}\Big[\sum_{k=1}^{n} K^j(\mathbf{Z}^{\mathfrak{mp}(j)} - \mathbf{Z}_k^{\mathfrak{mp}(j)}) \neq 0\Big]$. The resulting approximation to the risk functional $R(f)$, denoted by $\hat{R}_{\mathrm{aug}}(f)$, is

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

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 $\hat{p}_{j|\mathfrak{mp}(j)}$ 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 $\hat{p}_{j|\mathfrak{mp}(j)}$, namely $\frac{K^j(\mathbf{Z}^{\mathfrak{mp}(j)} - \mathbf{Z}_i^{\mathfrak{mp}(j)}) I_{\neq 0}}{\sum_{k=1}^{n} K^j(\mathbf{Z}^{\mathfrak{mp}(j)} - \mathbf{Z}_k^{\mathfrak{mp}(j)})}$, depends on $\mathbf{Z}^{\mathfrak{mp}(j)}$, 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 $\mathfrak{mp}(j) \subset [j-1]$.

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

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

where

$$\hat{w}_{\boldsymbol{i}} = \prod_{j=1}^{D} \frac{K^j(\mathbf{Z}_{\boldsymbol{i}_{1:j-1}}^{\mathfrak{mp}(j)} - \mathbf{Z}_i^{\mathfrak{mp}(j)})}{\sum_{k=1}^{n} K^j(\mathbf{Z}_{\boldsymbol{i}_{1:j-1}}^{\mathfrak{mp}(j)} - \mathbf{Z}_k^{\mathfrak{mp}(j)})} I_{\neq 0}^{\boldsymbol{i}_{1:j-1}}, \tag{2}$$

$$\mathbf{Z}_{\boldsymbol{i}} = (Z_{i_1}^1, \ldots, Z_{i_D}^D), \quad \mathbf{Z}_{\boldsymbol{i}_{1:j-1}} = (Z_{i_1}^1, \ldots, Z_{i_{j-1}}^{j-1}),$$

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

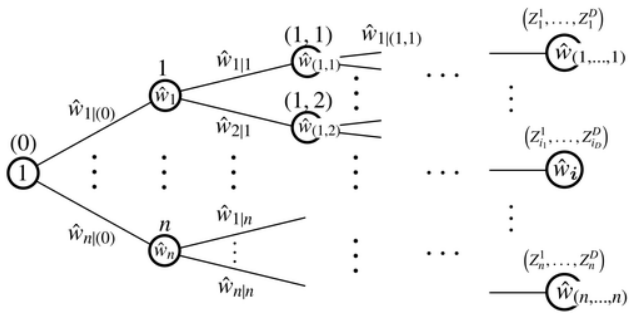


Figure 2: Probability tree to compute the weights of the augmented instances. At each depth j, the index i_j is selected and the weight is updated as $\hat{w}_{\boldsymbol{i}_{1:j}} = \hat{w}_{i_j|\boldsymbol{i}_{1:j-1}} \cdot \hat{w}_{\boldsymbol{i}_{1:j-1}}$.

and $I_{\neq 0}^{\boldsymbol{i}_{1:j-1}} := \mathbb{1}\left[\sum_{k=1}^n K^j(\mathbf{Z}_{\boldsymbol{i}_{1:j-1}}^{\mathrm{mp}(j)} - \mathbf{Z}_k^{\mathrm{mp}(j)}) \neq 0\right]$, for $\boldsymbol{i} = (i_1, \ldots, i_D) \in [n]^D$ and $\boldsymbol{i}_{1:j-1} = (i_1, \ldots, i_{j-1})$. Here, we use the convention $\mathbf{Z}_{\boldsymbol{i}_{1:0}}^{\mathrm{mp}(1)} := 0$ to be consistent with the notation.

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 $\mathcal{W}_{\mathrm{aug}} := \{\hat{w}_{\boldsymbol{i}}\}_{\boldsymbol{i}\in[n]^D}$, 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

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

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 Figure 1, $\mathcal{W}_{\mathrm{aug}}$ places equal weights on the augmented data, essentially yielding the same augmented data set as that in Figure 1.

Practical implementation. To reduce the computation cost of calculating the weights $\mathcal{W}_{\mathrm{aug}}$, we exploit the recursive structure in Equation (2) that can be represented by a probability tree [25], where we sequentially select the values $i_1, \ldots, i_D \in [n]$ (Figure 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

$$\hat{w}_{i_j|\boldsymbol{i}_{1:j-1}} := \frac{K^j(\mathbf{Z}_{\boldsymbol{i}_{1:j-1}}^{\mathrm{mp}(j)} - \mathbf{Z}_i^{\mathrm{mp}(j)})}{\sum_{i=1}^n K^j(\mathbf{Z}_{\boldsymbol{i}_{1:j-1}}^{\mathrm{mp}(j)} - \mathbf{Z}_i^{\mathrm{mp}(j)})} I_{\neq 0}^{\boldsymbol{i}_{1:j-1}}.$$

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 $\mathcal{W}_{\mathrm{aug}}$ (Figure 2). Algorithm 1 summarizes the procedure of the proposed method.

To reduce the computational complexity, 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 $\hat{w}_{\boldsymbol{i}} < \theta$. 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* [26] or stochastic optimization methods such as *stochastic gradient descent* [14, Section 5.9].

4 Theoretical Justification

In this section, we provide a theoretical justification of the proposed method in the form of an excess risk bound. The goal here is to elucidate how the proposed data augmentation procedure facilitates the statistical learning from a

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.

// Fill in the probability tree (Figure 2) to compute $\mathcal{W}_{\mathrm{aug}}$.

for $j \in [D]$ **do**

  for $(\boldsymbol{i}_{1:j-1}, i_j) \in [n]^{j-1} \times [n]$ **do**

    $\hat{w}_{\boldsymbol{i}_{1:j-1}} \leftarrow \hat{w}_{\boldsymbol{i}_{1:j-1}} \mathbb{1}\left[\hat{w}_{\boldsymbol{i}_{1:j-1}} \geq \theta\right]$ // Pruning

    $\hat{w}_{\boldsymbol{i}_{1:j}} \leftarrow \hat{w}_{i_j|\boldsymbol{i}_{1:j-1}} \cdot \hat{w}_{\boldsymbol{i}_{1:j-1}}$

  end for

end for

Let $\tilde{R}_\lambda(f) := (1-\lambda)\hat{R}_{\mathrm{emp}}(f) + \lambda \hat{R}_{\mathrm{aug}}(f) + \Omega(f)$.

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

theoretical perspective. We focus on the case that $\overline{\mathcal{Z}}^j = \mathbb{R}$ for all $j \in [D]$. Select $\tilde{K}^j$ and $\boldsymbol{h} = (\boldsymbol{h}^1, \ldots, \boldsymbol{h}^D) \in \mathbb{R}_{>0}^D$, 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*

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

Under additional assumptions on the boundedness and smoothness of the kernels and the underlying densities (see Theorem 2 in Supplementary Material C), 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 Supplementary Material C. 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 the 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. To see the complexity reduction effect, note that $\mathcal{L}^j_{\mathcal{F}}$ consists of univariate functions. In Section 5, we empirically confirm that the complexity reduction effect is worth the newly introduced bias and complexity of the kernel approximation in practice.

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 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 Supplementary Material B.

Data sets. We employ 6 data sets for the experiment, namely *Sachs* [2], *GSS* [3], *Boston Housing* [27], *Auto MPG* [28], *White Wine* [29], and *Red Wine* [29]. Table 1 summarizes these data sets. The *Sachs* data and the *GSS* data are accompanied by the ADMGs obtained from domain experts, and hence we use them in the experiments. For the other data sets, we first perform *DirectLiNGAM* [3] on the entire data set to obtain the estimated CGs, simulating a situation that we have background knowledge from domain experts.

Predictor model class. We employ the gradient boosted regression trees [13, 30] 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 $\boldsymbol{h} = (\boldsymbol{h}^1, \ldots, \boldsymbol{h}^D) \in \mathbb{R}^D_{>0}$ and use the product kernel $K^j(\boldsymbol{x} - \boldsymbol{y}) := \prod_{j' \in \mathrm{mp}(j)} \frac{1}{\boldsymbol{h}^{j'}} K^j_{j'} \left(\frac{\boldsymbol{x}^{j'} - \boldsymbol{y}^{j'}}{\boldsymbol{h}^{j'}} \right)$ for the proposed method. For each $j' \in \mathrm{mp}(j)$, 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* [31, 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:

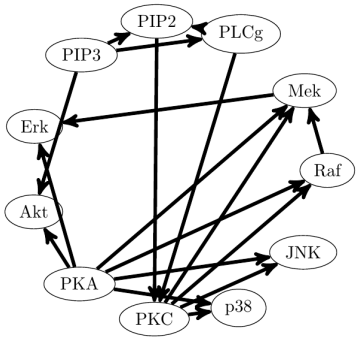
$$\hat{f} \in \arg\min_{f \in \mathcal{F}} \{\hat{R}_{\text{emp}}(f) + \Omega(f)\}.$$

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

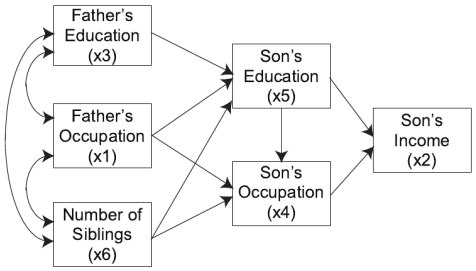
Evaluation procedure. The prediction accuracy is measured by the mean squared error (MSE). For each data set, we randomly subsample a fraction of the data as the training set and use the rest as the testing set. The fraction of the training set is varied in $\{.1, .15, \ldots, .85\}$. For each training 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, *Domain*: domain knowledge of the status attainment model, *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



(a) Reference graph for Sachs data. Figure excerpted from Mooij et al. [32].



(b) Reference graph for GSS data. Figure excerpted from Shimizu et al. [3].

Figure 3: Reference CGs for the data sets used in our experiments. a Consensus graph in Sachs et al. [2]. b Domain-knowledge graph based on the status attainment model [33].

5.2 Results

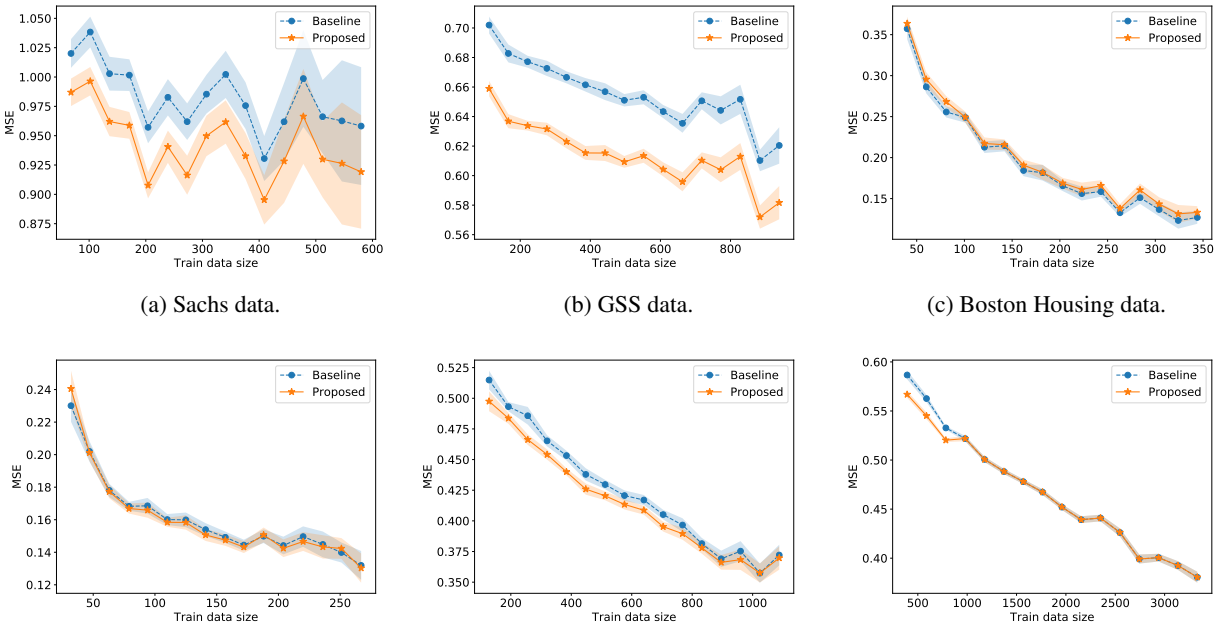
Figure 4 shows the experimental result. We observe a consistent performance improvement in most of the data sets. For the data sets for which the domain knowledge CG is provided (i.e., *Sachs* and *GSS*), we can see clear relative improvement ranging from 3% to 7% on average, especially in the small-data 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 Supplementary Material B.5.

6 Related Work and Discussion

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

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 [10]. For a recent comprehensive review on MB estimation, see Yu et al. [9]. The present paper is orthogonal to this line of work.



(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.

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 is 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. [34] 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. [35] 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 make good predictions in a single fixed distribution.

Regularization and model selection. Kyono et al. [36] proposed a model selection criterion that can reflect the structure of a CG. The goal of Kyono et al. [36] 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. [36] first modifies 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* [37]. 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. [36] 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 [8, 38, Section 8.2]. Schölkopf et al. [39] considered removing systematic errors using *half-sibling regression* inspired by the CG of the observation mechanism found in *exoplanet search*. Pitis et al. [40] 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. [41] proposed *causal bootstrapping*, a weighted bootstrap-type algorithm that is relevant to our method. While, methodologically, both the present paper and Little et al. [41] can be seen to be based on kernel-type function estimators [19, 20, 23] and CGs [1], the two works are complementary in that the problem setups differ. Causal bootstrapping of Little et al. [41] 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. [42] 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. [43] proposed a 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 (Equation (1)) is one natural way to exploit this prior knowledge [12]. This approach has the limitation that it inevitably restricts the modeling choice, 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) [44], e.g., the *fast causal inference* (FCI) algorithm [44, 45], enumerate the ADMGs in the equivalence class (e.g., by the *Pag2admg* algorithm; [46]), select a plausible candidate ADMG, and apply the proposed method. Such an assumption of a common causal mechanism has been exploited in recent work of causal discovery [47–49] and transfer learning [34, 50, 51], and it is based on a common belief that a causal mechanism remains invariant unless explicitly intervened in [52].

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.

Acknowledgments

The authors are grateful to Prof. Shohei Shimizu for providing them with the preprocessed GSS data set used in Shimizu et al. [3]. 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

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

[2] K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan, "Causal protein-signaling networks derived from multiparameter single-cell data," *Science*, vol. 308, no. 5721, pp. 523–529, 2005.

[3] S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvärinen, Y. Kawahara, T. Washio, P. O. Hoyer, and K. Bollen, "DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model," *Journal of Machine Learning Research*, vol. 12, no. 33, pp. 1225–1248, 2011.

[4] P. Spirtes, C. N. Glymour, and R. Scheines, *Causation, Prediction, and Search*, Second. Cambridge, Massachusetts: MIT Press, 2000.

[5] D. M. Chickering, "Optimal structure identification with greedy search," *The Journal of Machine Learning Research*, vol. 3, pp. 507–554, 2002.

[6] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. J. Kerminen, "A linear non-Gaussian acyclic model for causal discovery," *The Journal of Machine Learning Research*, vol. 7, no. 72, pp. 2003–2030, 2006.

[7] J. Peters and B. Sch, "Causal Discovery with Continuous Additive Noise Models," *Journal of Machine Learning Research*, vol. 15, no. June, pp. 2009–2053, 2014.

[8] J. Peters, D. Janzing, and B. Schölkopf, *Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge, Massachusetts: The MIT Press, 2017.

[9] K. Yu, X. Guo, L. Liu, J. Li, H. Wang, Z. Ling, and X. Wu, "Causality-based feature selection: Methods and evaluations," *ACM Computing Surveys*, vol. 53, no. 5, pp. 1–36, 2020.

[10] I. Tsamardinos and C. Aliferis, "Towards principled feature selection: Relevancy, filters and wrappers," in *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, 2003.

[11] T. Richardson, "Markov properties for acyclic directed mixed graphs," *Scandinavian Journal of Statistics*, vol. 30, no. 1, pp. 145–157, 2003.

[12] P. J. F. Lucas, L. C. van der Gaag, and A. Abu-Hanna, "Bayesian networks in biomedicine and health-care," *Artificial Intelligence in Medicine*, vol. 30, no. 3, pp. 201–214, 2004.

[13] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[15] P. L. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: Risk bounds and structural results," *Journal of Machine Learning Research*, vol. 3, no. Nov, pp. 463–482, 2002.

[16] T. S. Richardson, R. J. Evans, J. M. Robins, and I. Shpitser, "Nested Markov properties for acyclic directed mixed graphs," *arXiv:1701.06686 [stat.ME]*, 2017.

[17] R. Bhattacharya, R. Nabi, and I. Shpitser, "Semiparametric inference for causal effects in graphical models with hidden variables," *arXiv:2003.12659 [stat.ML]*, 2020.

[18] J. Tian and J. Pearl, "A general identification condition for causal effects," in *Proceedings of the Eighteenth National Conference on Artificial Intelligence*, 2002, pp. 567–573.

[19] W. Stute, "Conditional Empirical Processes," *Annals of Statistics*, vol. 14, no. 2, pp. 638–647, 1986.

[20] L. Horváth and B. S. Yandell, "Asymptotics of conditional empirical processes," *Journal of Multivariate Analysis*, vol. 26, no. 2, pp. 184–206, 1988.

[21] E. A. Nadaraya, "On estimating regression," *Theory of Probability & Its Applications*, vol. 9, no. 1, pp. 141–142, 1964.

[22] G. S. Watson, "Smooth regression analysis," *Sankhyā: The Indian Journal of Statistics, Series A*, vol. 26, no. 4, pp. 359–372, 1964.

[23] U. Einmahl and D. M. Mason, "An empirical process approach to the uniform consistency of kernel-type function estimators," *Journal of Theoretical Probability*, vol. 13, no. 1, pp. 1–37, 2000.

[24] V. A. Zorich, *Mathematical Analysis I*, Second. Berlin, Heidelberg: Springer Berlin Heidelberg, 2015.

[25] C. H. Brase and C. P. Brase, *Understanding Basic Statistics*. Cengage Learning, 2012.

[26] R. Bisiani, "Beam search," *Encyclopedia of Artificial Intelligence*, pp. 56–58, 1987.

[27] D. Harrison and D. L. Rubinfeld, "Hedonic housing prices and the demand for clean air," *Journal of Environmental Economics and Management*, vol. 5, no. 1, pp. 81–102, 1978.

[28] J. R. Quinlan, "Combining instance-based and model-based learning," in *Proceedings of the Tenth International Conference on Machine Learning*, 1993, pp. 236–243.

[29] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Modeling wine preferences by data mining from physicochemical properties," *Decision support systems*, vol. 47, no. 4, pp. 547–553, 2009.

[30] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

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

[32] J. M. Mooij and T. Heskes, "Cyclic causal discovery from continuous equilibrium data," in *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, 2013, pp. 431–439.

[33] O. D. Duncan, D. L. Featherman, and B. Duncan, *Socioeconomic Background and Achievement*. New York: Seminar Press, 1972.

[34] S. Magliacane, T. van Ommen, T. Claassen, S. Bongers, P. Versteeg, and J. M. Mooij, "Domain adaptation by using causal inference to predict invariant conditional distributions," in *Advances in Neural Information Processing Systems 31*, 2018, pp. 10 846–10 856.

[35] M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters, "Invariant models for causal transfer learning," *Journal of Machine Learning Research*, vol. 19, no. 36, pp. 1–34, 2018.

[36] T. Kyono and M. van der Schaar, "Improving model robustness using causal knowledge," *arXiv:1911.12441 [cs.LG]*, 2019.

[37] T. Kyono, Y. Zhang, and M. van der Schaar, "CASTLE: Regularization via auxiliary causal graph discovery," in *Advances in Neural Information Processing Systems 33*, 2020.

[38] L. Bottou, J. Peters, J. Quiñonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson, "Counterfactual reasoning and learning systems: The example of computational advertising," *Journal of Machine Learning Research*, vol. 14, no. 65, pp. 3207–3260, 2013.

[39] B. Schölkopf, D. Hogg, D. Wang, D. Foreman-Mackey, D. Janzing, C.-J. Simon-Gabriel, and J. Peters, "Removing systematic errors for exoplanet search via latent causes," in *Proceedings of the 32nd International Conference on Machine Learning*, 2015, pp. 2218–2226.

[40] S. Pitis, E. Creager, and A. Garg, "Counterfactual data augmentation using locally factored dynamics," in *Advances in Neural Information Processing Systems 33*, 2020.

[41] M. A. Little and R. Badawy, "Causal bootstrapping," *arXiv:1910.09648 [cs.LG]*, 2020.

[42] R. Evans and T. Richardson, "Markovian acyclic directed mixed graphs for discrete data," *The Annals of Statistics*, vol. 42, no. 4, pp. 1452–1482, 2014.

[43] R. Silva, C. Blundell, and Y. W. Teh, "Mixed cumulative distribution networks," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011, pp. 670–678.

[44] J. Zhang, "On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias," *Artificial Intelligence*, vol. 172, no. 16, pp. 1873–1896, 2008.

[45] P. Spirtes, C. Meek, and T. Richardson, "Causal inference in the presence of latent variables and selection bias," in *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, 1995, pp. 499–506.

[46] N. Subramani, "Pag2admg: An algorithm for the complete causal enumeration of a Markov equivalence class," in *Proceedings of the CausalML Workshop at ICML*, 2018.

[47] L. Xu, T. Fan, X. Wu, K. Chen, X. Guo, J. Zhang, and L. Yao, "A pooling-LiNGAM algorithm for effective connectivity analysis of fMRI data," *Frontiers in Computational Neuroscience*, vol. 8, p. 125, 2014.

[48] A. Ghassami, S. Salehkaleybar, N. Kiyavash, and K. Zhang, "Learning causal structures using regression invariance," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 3011–3021.

[49] R. P. Monti, K. Zhang, and A. Hyvärinen, "Causal discovery with general non-linear relationships using non-linear ICA," in *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence*, 2019, pp. 186–195.

[50] J. Pearl and E. Bareinboim, "Transportability of causal and statistical relations: A formal approach," in *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, 2011, pp. 247–254.

[51] T. Teshima, I. Sato, and M. Sugiyama, "Few-shot domain adaptation by causal mechanism transfer," in *Proceedings of the 37th International Conference on Machine Learning*, 2020, pp. 9458–9469.

[52] P. Hünermund and E. Bareinboim, "Causal inference and data-fusion in econometrics," *arXiv:1912.09104 [econ.EM]*, 2019.

[53] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[54] O. Yadan, *Hydra - A framework for elegantly configuring complex applications*, Github, 2019.

[55] *Statsmodels*, statsmodels, 2020.

[56] A. Hyvärinen and S. M. Smith, "Pairwise likelihood ratios for estimation of non-Gaussian structural equation models," *Journal of Machine Learning Research*, vol. 14, no. Jan, pp. 111–152, 2013.

[57] C. J. Stone, "Optimal global rates of convergence for nonparametric regression," *The Annals of Statistics*, vol. 10, no. 4, pp. 1040–1053, 1982.

[58] A. B. Tsybakov, *Introduction to Nonparametric Estimation*. New York ; London: Springer, 2009.

[59] U. Einmahl and D. M. Mason, "Uniform in bandwidth consistency of kernel-type function estimators," *Annals of Statistics*, vol. 33, no. 3, pp. 1380–1403, 2005.

[60] J. Dony, U. Einmahl, and D. M. Mason, "Uniform in bandwidth consistency of local polynomial regression function estimators," *Austrian Journal of Statistics*, vol. 35, no. 2, p. 16, 2006.

[61] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*, Second. Cambridge, Massachusetts: The MIT Press, 2018.

Appendices

Table 2 summarizes the abbreviations and the symbols used in the paper. For notation simplicity, when $\overline{\mathcal{Z}}^j$ is a finite set, we identify it with $\mathbb{Z}/m\mathbb{Z}$ where m is the cardinality of $\overline{\mathcal{Z}}^j$, to justify the subtractions inside the kernel functions.

Table 2: Abbreviations and Symbols in the Paper.

ABBREVIATION / SYMBOL	DESCRIPTION
CG/CGM	Causal Graph / Causal Graphical Model
ADMG	Acyclic Directed Mixed Graph
DAG/PAG	Directed Acyclic Graph / Partial Ancestral Graph
MSE	Mean Squared Error
$\mathbb{R}, \mathbb{R}_{\geq 0}, \mathbb{R}_{>0}, \mathbb{Z}, \mathbb{Z}_{\geq 0}, \mathbb{N}$	Set of all real numbers, nonnegative real numbers, positive real numbers, integers, nonnegative integers, and positive integers.
$\mathbb{1}[A]$	Indicator function, i.e., 1 if A holds true and 0 otherwise.
$X \perp\!\!\!\perp Y \mid Z$	X and Y are conditionally independent given Z.
$\coprod$	Disjoint union of sets.
$\mathrm{diag}((x_1, \ldots, x_d))$	Diagonal matrix with diagonal elements $(x_1, \ldots, x_d)$ ($d \in \mathbb{N}$).
$\lVert\cdot\rVert, \lVert\cdot\rVert_{\mathrm{op}}, \lVert\cdot\rVert_\infty, \det$	Euclidean norm of a vector, the operator norm of a matrix, the supremum norm of a function, and the determinant of a matrix.
$\lfloor\cdot\rfloor$	$\lfloor a \rfloor := \max\{z \in \mathbb{Z} : z \leq a\}$ for $a \in \mathbb{R}$.
$\delta_{\boldsymbol{z}}$	Dirac's delta function centered at $\boldsymbol{z}$ (e.g., [24, Section E.4.1]).
Δ_K	$(K-1)$-dimensional probability simplex [53, Example 2.5].
$[N:M], [N]$	$[N:M] := \{N, N+1, \ldots, M\}$ and $[N] := [1:N]$, where $N, M \in \mathbb{N}$ and $N \leq M$.
$\boldsymbol{x}^S$	$\boldsymbol{x}^S := (x^{s_1}, \ldots, x^{s_{\lvert S\rvert}})$ where $\boldsymbol{x} = (x^1, \ldots, x^n)$ is an n-dimensional vector and $S = \{s_1, \ldots, s_{\lvert S\rvert}\} \subset [n]$ with $s_1 < \cdots < s_{\lvert S\rvert}$.
$[0] = \emptyset, \mathbb{R}^0 := \{0\}, \boldsymbol{x}^\emptyset = 0, [N]^0 := \{0\}$	Conventions used in the paper.
$D \in \mathbb{N}$	Overall data dimensionality (with X and Y combined).
$\mathcal{Z} = \times_{j=1}^D \mathcal{Z}_j$	Overall data space (without distinguishing X and Y).
$\mathcal{X} = \times_{j \in [D]\setminus\{j^*\}} \overline{\mathcal{Z}}^j, \mathcal{Y} = \overline{\mathcal{Z}}^{j^*}$	Input variable space and target variable space.
p	Joint probability density of $\mathbf{Z} := (Z^1, \ldots, Z^D)$ taking values in $\mathcal{Z}$.
$\mathrm{Rad}_{m,q}$	Rademacher complexity of a function class.
$\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$	Hypothesis set.
$\ell : \mathcal{F} \times \left(\times_{j=1}^D \overline{\mathcal{Z}}^j\right) \to \mathbb{R}$	Loss function.
$R(f) = \mathbb{E}[\ell(f, \mathbf{Z})]$	Risk functional for $f \in \mathcal{F}$.
$\mathcal{D} = \{\mathbf{Z}_i\}_{i=1}^n$	Independently and identically distributed sample from p.
$\mathcal{G} = ([D], \mathcal{E}, \mathcal{B}), \hat{\mathcal{G}} = ([D], \hat{\mathcal{E}}, \hat{\mathcal{B}})$	Underlying ADMG for which p satisfies the topological ADMG factorization and its estimator.
$\mathfrak{dis}(\cdot), \mathfrak{pa}(\cdot), \mathfrak{mp}(j)$	District, parents, and Markov pillow of vertex $j \in [D]$.
$p_{j\mid\mathfrak{mp}(j)}, p_{j,\mathfrak{mp}(j)}, p_{\mathfrak{mp}(j)}$	Conditional density of Z^j given $\mathbf{Z}^{\mathfrak{mp}(j)}$, the joint density of $(Z^j, \mathbf{Z}^{\mathfrak{mp}(j)})$, and the marginal density of $\mathbf{Z}^{\mathfrak{mp}(j)}$.
$K^j : \overline{\mathcal{Z}}^{\mathfrak{mp}(j)} \to \mathbb{R}$	Kernel function (we define $K^j := 1$ if $\mathfrak{mp}(j) = \emptyset$).
$\mathbf{Z}_{\boldsymbol{i}}$	$\mathbf{Z}_{\boldsymbol{i}} = (Z^1_{i_1}, \ldots, Z^D_{i_D})$ for $\boldsymbol{i} = (i_1, \ldots, i_D) \in [n]^D$.
$\mathcal{D}_{\mathrm{aug}} := \{\mathbf{Z}_{\boldsymbol{i}}\}_{\boldsymbol{i} \in [n]^D}, \mathcal{W}_{\mathrm{aug}} := \{\hat{w}_{\boldsymbol{i}}\}_{\boldsymbol{i} \in [n]^D}$	Augmented data set and the instance weights.
$\hat{R}_{\mathrm{emp}}, \hat{R}_{\mathrm{aug}}$	Ordinary empirical risk estimator and the proposed risk estimator.
$\Omega(f)$	Regularization term for $f \in \mathcal{F}$.
$\lambda \in [0, 1]$	Convex combination coefficient used in $(1-\lambda)\hat{R}_{\mathrm{emp}}(f) + \lambda\hat{R}_{\mathrm{aug}}(f) + \Omega(f)$.
$K^j_{j'}$	Component of the product kernel K^j for $j' \in \mathfrak{mp}(j)$.
θ	Pruning threshold of the small weights in Algorithm 1.

A Preliminaries on ADMG

Given an ADMG $\mathcal{G}$ with the vertex set V and a topological order $\preceq$, we use the following terminologies [17].

District. For $v \in V$, define $\mathfrak{dis}(v)$ as the collection of $v' \in V$ that is connected to v via a bi-directed path.

Parents. For a subset $A \subset V$, we define its parents as $\mathfrak{pa}(A) := \bigcup_{v \in A} \mathfrak{pa}(v) \setminus A$ where $\mathfrak{pa}(v)$ denotes the parent of v in the usual sense.

Markov pillow. For $v \in V$, define $\mathcal{G}_{\preceq v}$ to be the subgraph of $\mathcal{G}$ that is composed of only the vertices that precede v. Then, the Markov pillow of $v \in V$ is $\mathfrak{mp}(v) := \mathfrak{dis}(v) \cup \mathfrak{pa}(\mathfrak{dis}(v)) \setminus \{v\}$ in $\mathcal{G}_{\preceq v}$. Throughout the paper, we use the fact that $\mathfrak{mp}(v)$ consists only of variables that are precedent to v.

B Experiment Details

Here, we describe the implementation details of the experiment. The experiment was implemented using the *hydra* package of Python [54]. All experiments were carried out on a 2.60 GHz Intel® Xeon® CPUs with 132 GB memory.

B.1 Data Set details

Following are the data acquisition procedures, the sample sizes, the variable definitions, and the preprocessing procedures used in our experiment. In all the data sets, after preprocessing as described below, we independently normalized each variable as a final preprocessing step.

Sachs data [2]. This data set consists of continuous measurements from the flow cytometry of proteins and phospholipids in human immune system cells. The *consensus graph* is provided in Sachs et al. [2] based on the conventionally accepted cellular signaling networks (Figure 3a). Among the eight data sets corresponding to different intervention conditions [2], we use the one that is *observational*, i.e., without any interventions. The data set contains 853 observations of 11 variables, namely *Raf*, *Mek*, *Plcg*, *PIP2*, *PIP3*, *Erk*, *Akt*, *PKA*, *PKC*, *P38*, and *Jnk*. Among these, for demonstration purposes, we considered *PKA* as the target attribute. As preprocessing, we log-transformed *Raf*, *Mek*, and *PKA*.

GSS data [3]. This data set is concerning the status attainment theory in sociology. This data set is originally part of the General Social Survey (GSS)[4], and we used a subset of the data that was previously used in the causal discovery literature [3]. The reference graph is based on domain knowledge of the status attainment model ([33]; Figure 3b). The acquired data set consists of 1380 observations of 6 variables, namely x_1: father's occupation level, x_2: son's income, x_3: father's education, x_4: son's occupation, x_5: son's education, and x_6: the number of siblings. We consider x_4 as the target variable.

Boston Housing data [27]. This data set is concerning the house prices in Boston, and the objective is to predict the prices of the house from its attributes. We acquired the data from `https://github.com/adityatiwari13/Boston_Dataset`. The acquired data set consists of 506 observations of 13 variables, namely *CRIM*, *ZN*, *INDUS*, *CHAS*, *NOX*, *RM*, *AGE*, *DIS*, *RAD*, *TAX*, *PTRATIO*, *B*, *LSTAT*, and *MEDV*. The objective is to predict the value of prices of the house, i.e., *MEDV*, using the given features.

Auto MPG data [28]. This data set concerns the city-cycle fuel consumption in miles per gallon (MPG). We acquired the data from `https://archive.ics.uci.edu/ml/datasets/Auto+MPG`. The acquired data set consists of 398 observations of 9 variables, namely *mpg*, *cylinders*, *displacement*, *horsepower*, *weight*, *acceleration*, *model year*, *origin*, and *car name*. Among these, we discard *origin* and *car name*, and we consider *mpg* as the predicted variable.

White Wine data [29]. This data set is concerning the prediction of wine quality from its physicochemical attributes. We acquired the data from `https://archive.ics.uci.edu/ml/datasets/wine+quality`. The acquired data set consists of 4898 observations of 12 variables, namely *fixed acidity*, *volatile acidity*, *citric acid*, *residual sugar*, *chlorides*, *free sulfur dioxide*, *total sulfur dioxide*, *density*, *pH*, *sulphates*, *alcohol*, and *quality*. Among the variables, we consider the *quality* variable as the target.

[4] `https://gss.norc.org/`

Red Wine data [29]. This data set is concerning the prediction of wine quality from its physicochemical attributes. We acquired the data from `https://archive.ics.uci.edu/ml/datasets/wine+quality`. The acquired data set consists of 1599 observations of 12 variables, namely *fixed acidity*, *volatile acidity*, *citric acid*, *residual sugar*, *chlorides*, *free sulfur dioxide*, *total sulfur dioxide*, *density*, *pH*, *sulphates*, *alcohol*, and *quality*. Among these, we consider the *quality* variable as the target.

B.2 Predictor model details

For the implementation of the predictor model, we employed the *xgboost* library of Python [30]. See Chen et al. [30] for the optimization method and the other details.

B.3 Proposed method implementation details

For continuous variables, we compute the kernel bandwidths as follows. We first specify the *bandwidth temperature* $\gamma > 0$ as a hyper-parameter. Then we calculate the rule-of-thumb bandwidth h_j^{thumb} for each $j \in [D]$ using the training data $\{\mathbf{Z}_i^j\}_{i=1}^n$. Finally, we set $h_j = \gamma \cdot h_j^{\text{thumb}}$. In the experiment, we fix $\gamma = 10^{-3}$ throughout all runs.

For the rule-of-thumb kernel bandwidth, we employed *Silverman*'s rule-of-thumb [31, pp.45–47, Equations (3.28) and (3.30) therein] implemented in the *statsmodels* package of Python [55], namely, $h^{\text{thumb}} = \left(\frac{4}{3}\right)^{1/5} An^{-1/5}$ where $A = \min\{\hat{\sigma}, \text{IQR}/1.349\}$, $\hat{\sigma}$ is the square root of the unbiased estimator of the variance, and IQR is the interquantile range.

For the pruning threshold, we use $\theta = 10^{-3} \cdot n^{-1}$.

B.4 Causal Discovery Method Configuration

We perform *DirectLiNGAM* [3] on the data sets to simulate a situation where we have access to domain knowledge. As the independence measure used in the DirectLiNGAM framework, we employ the pairwise likelihood ratio score [56] that is based on a nonparametric approximation to the mutual information.

B.5 Supplementary experiment results

Figure 5 shows the average improvement achieved by the proposed method relative to the baseline without a device. The improvement in the small-data regime is consistently observed except in a few cases in the *Auto MPG* and the *Boston Housing* data. In the *Boston Housing* data set, the performance loss may be due to the failure of the CG estimation since the performance loss is magnified as the training set size is increased. In the *Auto MPG* data, the performance degradation for the smallest training set fraction may be due to the additional complexity and bias introduced by the kernel approximation.

C Details and Proof of the Main Theorem

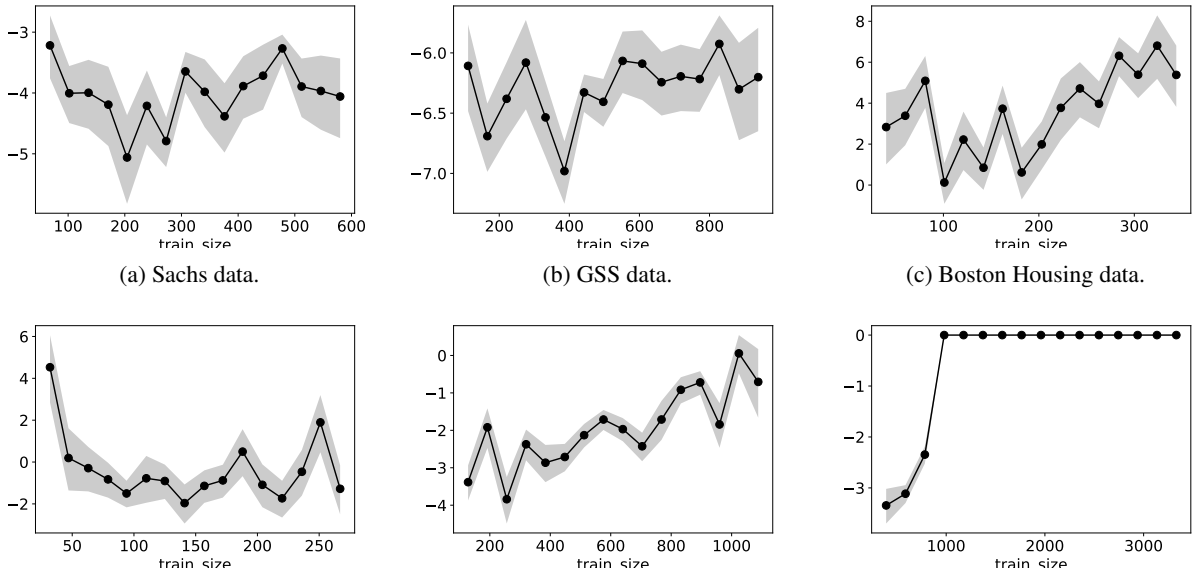
Here, we detail the assumptions, the statement, and a proof of Theorem 1.

Basic notation. Let $\mathbb{R}$ denote the set of real numbers, $\mathbb{N}$ that of positive integers, $\mathbb{R}_{>0}$ that of positive real numbers, $\mathbb{Z}$ that of integers, and $\mathbb{Z}_{\geq 0}$ that of non-negative integers. For $(x_1, \ldots, x_k) \in \mathbb{R}^k$, $\text{diag}((x_1, \ldots, x_k))$ denotes the diagonal matrix whose diagonal elements are $(x_1, \ldots, x_k)$. For a vector, $\|\cdot\|$ denotes its Euclidean norm. For a matrix, det denotes its determinant, and $\|\cdot\|_{\text{op}}$ its operator norm. For a function, $\|\cdot\|_\infty$ denotes its supremum norm over a suitable set of inputs when the domain is clear from the context. For a finite set, $|\cdot|$ denotes its cardinality.

Utility notation. For $n \in \mathbb{N}$, define $[n] := \{1, 2, \ldots, n\}$. For $n, m \in \mathbb{N}$ with $n \leq m$, define $[n : m] := \{n, n+1, \ldots, m\}$. For an n-dimensional vector $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $S \subset [n]$, we let $\boldsymbol{x}^S = (x_{s_1, \ldots, s_{|S|}})$ denote its sub-vector with indices in $S = \{s_1, \ldots, s_{|S|}\}$ with $s_1 < \cdots < s_{|S|}$. Similarly, for $j \in [n]$, we let $\boldsymbol{x}^j := \boldsymbol{x}^{\{j\}}$. For $S \subset [n]$, we also define $\mathcal{Z}^S := \times_{k \in S} \mathcal{Z}^k$. To simplify the notation, we use the convention of $\mathbb{R}^0 := \{0\}$, $\boldsymbol{x}^\emptyset = 0$, and $[n]^{j-1} = \{0\}$.

C.1 Recap of the Problem Setup

Distribution and sample. Let $D \in \mathbb{N}$. In this theoretical analysis, we assume that $\mathcal{Z}^j$ is a measurable subset of $\mathbb{R}$ ($j \in [D]$). We consider a probability distribution over $\mathcal{Z} := \times_{j=1}^D \mathcal{Z}^j$, and let p denote its density function (assuming



(a) Sachs data. (b) GSS data. (c) Boston Housing data.

(d) Auto MPG data. (e) Red Wine data. (f) White Wine data.

Figure 5: Average relative improvement in percentage. In all figures, the horizontal axis is the varied sizes of the original training data before augmentation. The vertical axis is the relative MSE improvement in percentage, i.e., $\frac{\mathrm{MSE}_{\mathrm{prop}}-\mathrm{MSE}_{\mathrm{base}}}{\mathrm{MSE}_{\mathrm{base}}} \times 100$ % where $\mathrm{MSE}_{\mathrm{base}}$ and $\mathrm{MSE}_{\mathrm{prop}}$ are the MSE of the baseline and that of the proposed method, respectively (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. In most of the cases, the proposed method shows a consistently improved performance compared to the baseline based on the empirical risk minimization with the same hypothesis class, particularly in the small-data regime.

it exists). We are given $\mathcal{D} = \{\mathbf{Z}_i\}_{i=1}^n$, an independently and identically distributed sample from p. Let $\mathbb{E}$ denote the expectation with respect to p. Additionally, we are given an ADMG $\mathcal{G} = ([D], \mathcal{E}, \mathcal{B})$. Let $\mathrm{mp}(j) \subset [D]$ denote the Markov pillow of $j \in [D]$. Throughout this section, we assume p satisfies the topological ADMG factorization relation according to $\mathcal{G}$ [17]:

$$p(\boldsymbol{z}) = \prod_{j=1}^{D} p_{j|\mathrm{mp}(j)}(\boldsymbol{z}^j|\boldsymbol{z}^{\mathrm{mp}(j)}) \quad \left(= \prod_{j=1}^{D} \frac{p_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j, \boldsymbol{z}^{\mathrm{mp}(j)})}{p_{\mathrm{mp}(j)}(\boldsymbol{z}^{\mathrm{mp}(j)})}\right).$$

Learning problem. Let $\mathcal{F}$ denote a hypothesis class, and let $\ell : \mathcal{F} \times \mathbb{R}^D \to \mathbb{R}_{>0}$ be a loss function. For each $f \in \mathcal{F}$, we define the risk functional $R(f) := \mathbb{E}[\ell(f, \mathbf{Z})]$. The learning problem is to find a hypothesis $\hat{f} \in \mathcal{F}$ for which R is small, given the training data $\mathcal{D}$ and the graph $\mathcal{G}$.

Proposed method. For each $j \in [D]$, we fix a kernel function $K^j : \mathbb{R}^{|\mathrm{mp}(j)|} \to \mathbb{R}$. For notation simplicity, we define $K^j := 1$ for j such that $\mathrm{mp}(j) = \emptyset$. We also fix $\boldsymbol{h} = (\boldsymbol{h}^1, \ldots, \boldsymbol{h}^D) \in \mathbb{R}_{>0}^D$. Then, we define

$$\mathbf{H}_j := \mathrm{diag}(\boldsymbol{h}^{\mathrm{mp}(j)}), \qquad K_{\mathbf{H}}^j(u) := \frac{1}{|\det \mathbf{H}_j|} K^j(\mathbf{H}_j^{-1} u).$$

For $\boldsymbol{i} = (i_1, \ldots, i_D)$ and $\boldsymbol{z}^{\mathrm{mp}(j)} \in \mathbb{R}^{|\mathrm{mp}(j)|}$, define

$$\hat{w}_i^j(\boldsymbol{z}^{\mathrm{mp}(j)}) := \frac{K_{\mathbf{H}}^j(\boldsymbol{z}^{\mathrm{mp}(j)} - \mathbf{Z}_i^{\mathrm{mp}(j)})}{\sum_{i=1}^n K_{\mathbf{H}}^j(\boldsymbol{z}^{\mathrm{mp}(j)} - \mathbf{Z}_i^{\mathrm{mp}(j)})} \mathbb{1}\left[\sum_{i=1}^n K_{\mathbf{H}}^j(\boldsymbol{z}^{\mathrm{mp}(j)} - \mathbf{Z}_i^{\mathrm{mp}(j)}) \neq 0\right]$$

where $\boldsymbol{i} = (i_1, \ldots, i_D)$, $\boldsymbol{z}^{\mathrm{mp}(j)} \in \mathbb{R}^{|\mathrm{mp}(j)|}$. Then, we 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

$$\hat{w}_{i_j|\boldsymbol{i}_{1:j-1}} := \hat{w}^{j}_{i_j}\left(\mathbf{Z}^{\mathrm{mp}(j)}_{\boldsymbol{i}_{1:j-1}}\right), \quad Z_{\boldsymbol{i}_{1:j-1}} = \left(Z^{1}_{i_1}, \ldots, Z^{j-1}_{i_{j-1}}\right).$$

Here, we use the convention $Z^{\mathrm{mp}(1)}_{\boldsymbol{i}_{1:0}} := 0$ to be consistent with the notation. Using this notation, for $f \in \mathcal{F}$, define the augmented empirical risk estimator

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

Target of the theoretical analysis. We aim to provide a stochastic upper bound on $R(\hat{f}) - R(f^*)$, where

$$\hat{f} \in \arg\min_{f\in\mathcal{F}}\{\hat{R}_{\mathrm{aug}}(f)\}, \text{ and } f^* \in \arg\min_{f\in\mathcal{F}}\{R(f)\},$$

assuming both exist.

C.2 Preliminaries

We use the following convenient *multi-index* notation (see, e.g., [57]).

Multi-index notation. For $d \in \mathbb{N}$, we call a d-tuple $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{Z}^d_{\geq 0}$ *multi-index*. For a multi-index α, let $|\alpha| := \sum_{j=1}^d \alpha_j$ and $\alpha! := \prod_{j=1}^d \alpha_j!$, and $x^\alpha = x_1^{\alpha_1}\cdots x_d^{\alpha_d}$ for $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$. Also, let ∂^α denote the partial differential operator defined by

$$\partial^\alpha = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1}\cdots\partial x_d^{\alpha_d}}.$$

Convolution. Let $d \in \mathbb{N}$ and $\Omega \subset \mathbb{R}^d$ be a measurable subset. For continuous bounded functions $f, g : \Omega \to \mathbb{R}$, we define a function $(f \underset{[\Omega]}{*} g) : \Omega \to \mathbb{R}$ by

$$f \underset{[\Omega]}{*} g(\boldsymbol{x}) := \int_\Omega f(\boldsymbol{x} - \boldsymbol{y})g(\boldsymbol{y})\mathrm{d}\boldsymbol{y}.$$

When $\Omega = \mathbb{R}^d$, we drop Ω from the notation and denote $f * g$.

We define the following class of functions.

Definition 2 (Hölder class; [57, 58])**.** *Let $d \in \mathbb{N}$, $\beta > 1$, $L > 0$, and let $\Omega \subset \mathbb{R}^d$ be an open subset. The (β, L)-Hölder class $\Sigma(\beta, L, \Omega)$ is defined as the set of $k = \lfloor\beta\rfloor$-times continuously differentiable functions $f : \Omega \to \mathbb{R}$ satisfying*

$$|\partial^\alpha f(x) - \partial^\alpha f(x')| \leq L\left\|x - x'\right\|^{\beta-k} \quad \text{for} \quad x, x' \in \Omega \text{ and } |\alpha| = k,$$

where $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{Z}^d_{\geq 0}$ is a multi-index, and $\lfloor a\rfloor = \max\{z \in \mathbb{Z} : z \leq a\}$ for $a \in \mathbb{R}$. When $\Omega = \mathbb{R}^d$, we also drop $\mathbb{R}^d$ from the notation and denote $\Sigma(\beta, L)$ when the dimension is clear from the context.

Remark 1. *In the 1-dimensional case, a related analysis based on the notion of the Hölder class is presented in Section 1.2.3 of Tsybakov [58].*

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

Definition 3 (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 $m \in \mathbb{N}$, $\{\sigma_i\}_{i=1}^m$ are independent uniform $\{\pm 1\}$-valued random variables, and $\{X_i\}_{i=1}^m \overset{i.i.d.}{\sim} q$.

C.3 Assumptions

For simplicity, throughout this theoretical analysis, we assume that all quantities appearing in the proof satisfy sufficient measurability conditions.

Assumption 1 (Boundedness assumptions)**.** *We assume that the following hold:*

- *The loss function is bounded, i.e.,* $B_\ell := \sup_{f\in\mathcal{F}} \sup_{\mathbf{Z}\in\mathbb{R}^D} |\ell(f,\mathbf{Z})| < \infty$.
- $\mathbf{K} := \{K^j\}_{j=1}^D$ *are uniformly bounded from above, i.e.,* $B_{\mathbf{K}} := \max\left\{\left\|K^j\right\|_\infty : j \in [D]\right\} < \infty$.
- *For each* $j \in [D]$, $\mathcal{Z}^j \subset \mathbb{R}$ *is a compact subset. Let* $B_j := \int_{\mathcal{Z}^j} \mathrm{d}z^j < \infty$.
- *For all* $j \in [D]$, $p_{\mathrm{mp}(j)}$ *is bounded away from zero over* $\mathcal{Z}^{\mathrm{mp}(j)}$. *Define* $\epsilon_{\mathrm{mp}(j)} := \inf_{\boldsymbol{z}^{\mathrm{mp}(j)}\in\mathcal{Z}^{\mathrm{mp}(j)}} p_{\mathrm{mp}(j)}(\boldsymbol{z}^{\mathrm{mp}(j)})$.
- *For each* $j \in [D]$, K^j *is continuous and strictly positive. We define*

$$\phi_{K^j,\mathbf{H}_j} := \sup_{\substack{\boldsymbol{z}^{\mathrm{mp}(j)}\in\mathcal{Z}^{\mathrm{mp}(j)},\\ \boldsymbol{z}^{\mathrm{mp}(j)\prime}\in\mathbb{R}^{|\mathrm{mp}(j)|}\setminus\mathcal{Z}^{\mathrm{mp}(j)}}} \left|K^j_{\mathbf{H}}(\boldsymbol{z}^{\mathrm{mp}(j)} - \boldsymbol{z}^{\mathrm{mp}(j)\prime})\right| = \sup_{\substack{\boldsymbol{z}^{\mathrm{mp}(j)}\in\mathbf{H}_j^{-1}\mathcal{Z}^{\mathrm{mp}(j)},\\ \boldsymbol{z}^{\mathrm{mp}(j)\prime}\in\mathbf{H}_j^{-1}(\mathbb{R}^{|\mathrm{mp}(j)|}\setminus\mathcal{Z}^{\mathrm{mp}(j)})}} \left|K^j(\boldsymbol{z}^{\mathrm{mp}(j)} - \boldsymbol{z}^{\mathrm{mp}(j)\prime})\right| |\det\mathbf{H}_j|^{-1}$$

  and assume $\phi_{K^j,\mathbf{H}_j} < \infty$.

Remark 2. *Since* $\mathcal{Z}^{\mathrm{mp}(j)}$ *is compact and* K^j *is continuous, if we define*

$$\epsilon_{K^j}(\mathbf{H}_j) := \left|\det\mathbf{H}_j\right| \left(\inf_{\boldsymbol{x},\boldsymbol{x}'\in\mathcal{Z}^{\mathrm{mp}(j)}} K^j_{\mathbf{H}}(\boldsymbol{x}-\boldsymbol{x}')\right) = \inf_{\boldsymbol{x},\boldsymbol{x}'\in\mathbf{H}_j^{-1}\mathcal{Z}^{\mathrm{mp}(j)}} K^j(\boldsymbol{x}-\boldsymbol{x}'),$$

this quantity is strictly positive under Assumption 1.

From here, we fix $\beta > 1$ and $L > 0$.

Assumption 2 (Smoothness assumptions)**.** *We assume that the following hold for all* $j \in [D]$*:*

- $p_{\mathrm{mp}(j)}$ *has an extension* $\check{p}_{\mathrm{mp}(j)} \in \Sigma(\beta, L)$ *such that* $\check{I}_{\mathrm{mp}(j)} := \int_{\mathbb{R}^{|\mathrm{mp}(j)|}\setminus\mathcal{Z}^{\mathrm{mp}(j)}} |\check{p}_{\mathrm{mp}(j)}(\boldsymbol{z}^{\mathrm{mp}(j)})| \mathrm{d}\boldsymbol{z}^{\mathrm{mp}(j)} < \infty$.
- *For all* $\boldsymbol{z}^j \in \mathcal{Z}^j$, $p_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j,\cdot)$ *has an extension* $\check{p}_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j,\cdot) \in \Sigma(\beta,L)$ *such that* $\check{I}_{j,\mathrm{mp}(j)} := \int_{\mathcal{Z}^j}\left(\int_{\mathbb{R}^{|\mathrm{mp}(j)|}\setminus\mathcal{Z}^{\mathrm{mp}(j)}} |\check{p}_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j,\boldsymbol{z}^{\mathrm{mp}(j)})| \mathrm{d}\boldsymbol{z}^{\mathrm{mp}(j)}\right)\mathrm{d}\boldsymbol{z}^j < \infty$.
- K^j *is of order* $k = \lfloor\beta\rfloor$*, i.e.,*

$$\int_{\mathbb{R}^{|\mathrm{mp}(j)|}} K^j(u)\mathrm{d}u = 1, \qquad \int_{\mathbb{R}^{|\mathrm{mp}(j)|}} K^j(u)u^\alpha \mathrm{d}u = 0 \quad (1 \le |\alpha| \le k),$$

  where $\alpha \in \mathbb{Z}_{\ge 0}^{|\mathrm{mp}(j)|}$ *is a multi-index, and* K^j *satisfies* $\int_{\mathbb{R}^{|\mathrm{mp}(j)|}} |K^j(u)| \cdot \|u\|^\beta \mathrm{d}u < \infty$.

Remark 3 (Existence of the smooth extensions)**.** *The smooth extensions in Assumption 2 exist, for example, if we consider a smooth density function* $\check{p}_{\mathrm{mp}(j)}$ *on* $\mathbb{R}^{|\mathrm{mp}(j)|}$ *and regard its restriction to* $\mathcal{Z}^{\mathrm{mp}(j)}$ *with appropriate scaling as* $p_{\mathrm{mp}(j)}$*.*

C.4 Statement and Proof

To state the main theorem, we use the following notation. For each $j \in [D]$ and $f \in \mathcal{F}$, define

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

Also define

$$\mathcal{L}^j_{\mathcal{F}} := \left\{\ell_{f,j}(\boldsymbol{z}^1,\ldots,\boldsymbol{z}^{j-1},\cdot) : f\in\mathcal{F}, (\boldsymbol{z}^1,\ldots,\boldsymbol{z}^{j-1}) \in \mathcal{Z}^{[1:j-1]}\right\},$$
$$\mathcal{K}^j_{\mathbf{H}} := \left\{K^j_{\mathbf{H}}(\boldsymbol{z}^{\mathrm{mp}(j)} - (\cdot)) : \boldsymbol{z}^{\mathrm{mp}(j)} \in \mathcal{Z}^{\mathrm{mp}(j)}\right\}.$$

We prove the following theorem. Theorem 1 is obtained by changing δ to $\frac{\delta}{2D}$ in the following theorem, substituting $\left\|\mathbf{H}_j\right\|_{\mathrm{op}} = \max_{j'\in\mathrm{mp}(j)} \boldsymbol{h}^{j'}$, and defining the appropriate constants.

Theorem 2 (Excess risk bound)**.** *Assume that Assumptions 1 and 2 hold. Let $n \in \mathbb{N}$. For $j \in [D]$, define*

$$C_{\mathbf{H}} := B_\ell \sum_{j=1}^{D} \frac{1}{\epsilon_{\mathrm{mp}(j)}} \left(B_j + \frac{B_{\mathbf{K}}}{\epsilon_{K^j}(\mathbf{H}_j)} \right) \Phi(\beta, L, K^j) \left\| \mathbf{H}_j \right\|_{\mathrm{op}}^{\beta}, \quad C_p := B_\ell \sum_{j=1}^{D} \frac{\phi_{K^j,\mathbf{H}_j}}{\epsilon_{\mathrm{mp}(j)}} \left(\check{I}_{j,\mathrm{mp}(j)} + \frac{B_{\mathbf{K}}}{\epsilon_{K^j}(\mathbf{H}_j)} \check{I}_{\mathrm{mp}(j)} \right),$$

$$C_{\mathbf{K}} := \max_{j \in [D]} \left\{ \frac{1}{\epsilon_{K^j}(\mathbf{H}_j)}, \frac{B_{\mathbf{K}}}{(\epsilon_{K^j}(\mathbf{H}_j))^2} \right\}, \quad R_{\mathcal{F},\mathbf{K}} := \sum_{j=1}^{D} \left| \det \mathbf{H}_j \right| \mathrm{Rad}_{n,p} \left(\mathcal{L}_{\mathcal{F}}^{j} \otimes \mathcal{K}_{\mathbf{H}}^{j} \right), \quad R_{\mathbf{K}} := \sum_{j=1}^{D} \left| \det \mathbf{H}_j \right| \mathrm{Rad}_{n,p}(\mathcal{K}_{\mathbf{H}}^{j}).$$

Then, for any $\delta \in (0,1)$, we have with probability at least $1 - 2D\delta$,

$$R(\hat{f}) - R(f^*) \leq 2(C_{\mathbf{H}} + C_p) + 4C_{\mathbf{K}}(R_{\mathcal{F},\mathbf{K}} + B_\ell R_{\mathbf{K}}) + 2DB_\ell B_{\mathbf{K}} C_{\mathbf{K}} \sqrt{\frac{\log(2/\delta)}{2n}}.$$

Proof overview. Our proof borrows ideas from the literature on *local empirical processes* and *kernel-type estimators*, namely Einmahl et al. [23, 59] and Dony et al. [60]. Two elementary calculations are essential in the proof. The first one handles a difference between two products: let $N \in \mathbb{N}$, $(a_1, \ldots, a_N) \in \mathbb{R}^N$, and $(b_1, \ldots, b_N) \in \mathbb{R}^N$, then,

$$\left(\prod_{j=1}^{N} a_i \right) - \left(\prod_{j=1}^{N} b_i \right) = \sum_{j=1}^{N} a_1 \cdots a_{j-1}(a_j - b_j) b_{j+1} \cdots b_N. \tag{3}$$

The second one bounds a difference between two ratios from above: for $A, B, C, D \in \mathbb{R}$ with $B, D \neq 0$,

$$\left| \frac{A}{B} - \frac{C}{D} \right| = \left| \frac{A}{B} - \frac{C}{B} + \frac{C}{B} - \frac{C}{D} \right| \leq \left| \frac{1}{B} \right| \cdot |A - C| + \left| \frac{C}{BD} \right| \cdot |B - D|. \tag{4}$$

Proof of Theorem 2. First, note

$$R(\hat{f}) - R(f^*) = R(\hat{f}) - \hat{R}_{\mathrm{aug}}(\hat{f}) + \hat{R}_{\mathrm{aug}}(\hat{f}) - R(f^*) \leq R(\hat{f}) - \hat{R}_{\mathrm{aug}}(\hat{f}) + \hat{R}_{\mathrm{aug}}(f^*) - R(f^*) \leq \underbrace{2 \sup_{f \in \mathcal{F}} |R(f) - \hat{R}_{\mathrm{aug}}(f)|}_{(*)}.$$

For ease of notation, define $\hat{p}_j(z^j | z^{\mathrm{mp}(j)}) = \sum_{i=1}^{n} \delta_{Z_i^j}(z^j) \hat{w}_i^j(z^{\mathrm{mp}(j)})$ and temporarily denote $p_k := p_{k|\mathrm{mp}(k)}$. With this notation, $\hat{R}_{\mathrm{aug}}(f) = \int_{\mathcal{Z}} \ell(f, z) \prod_{j=1}^{D} \hat{p}_j(z^j | z^{\mathrm{mp}(j)}) \mathrm{d}z$. Then, applying the argument of Equation (3), we have

$$\begin{aligned}
(*) &= \sup_{f \in \mathcal{F}} \left| \int_{\mathcal{Z}} \ell(f, z) \prod_{j=1}^{D} p_j(z^j | z^{\mathrm{mp}(j)}) \mathrm{d}z - \int_{\mathcal{Z}} \ell(f, z) \prod_{j=1}^{D} \hat{p}_j(z^j | z^{\mathrm{mp}(j)}) \mathrm{d}z \right| \\
&= \sup_{f \in \mathcal{F}} \left| \int_{\mathcal{Z}} \ell(f, z) \sum_{j=1}^{D} \left(\prod_{k=j+1}^{D} p_k(z^k | z^{\mathrm{mp}(k)}) \right) (p_j(z^j | z^{\mathrm{mp}(j)}) - \hat{p}_j(z^j | z^{\mathrm{mp}(j)})) \left(\prod_{k=1}^{j-1} \hat{p}_k(z^k | z^{\mathrm{mp}(k)}) \right) \mathrm{d}z \right| \\
&\leq \sum_{j=1}^{D} \underbrace{\sup_{f \in \mathcal{F}} \left| \int_{\mathcal{Z}} \ell(f, z) \left(\prod_{k=j+1}^{D} p_k(z^k | z^{\mathrm{mp}(k)}) \right) (p_j(z^j | z^{\mathrm{mp}(j)}) - \hat{p}_j(z^j | z^{\mathrm{mp}(j)})) \left(\prod_{k=1}^{j-1} \hat{p}_k(z^k | z^{\mathrm{mp}(k)}) \right) \mathrm{d}z \right|}_{(*j)}.
\end{aligned}$$

Now, for $f \in \mathcal{F}$ and $j \in [D]$, we define $\ell_{f,j}^{\boldsymbol{i}_{1:j-1}} : z^j \mapsto \ell_{f,j}(Z_{\boldsymbol{i}_{1:j-1}}, z^j)$. Then, for each $j \in [D]$, applying Lemma 5, we obtain

$$
\begin{aligned}
(*j) &= \sup_{f\in\mathcal{F}} \left| \sum_{i_1=1}^{n} \cdots \sum_{i_{j-1}=1}^{n} \left(\int_{\mathcal{Z}^j} \ell_{f,j}^{\boldsymbol{i}_{1:j-1}}(z^j) p_j(z^j|\mathbf{Z}^{\mathrm{mp}(j)}_{\boldsymbol{i}_{1:j-1}}) \mathrm{d}z^j - \sum_{i_j=1}^{n} \ell_{f,j}^{\boldsymbol{i}_{1:j-1}}(Z_{i_j}^j) \hat{w}_{i_j|\boldsymbol{i}_{1:j-1}} \right) \hat{w}_{i_{j-1}|\boldsymbol{i}_{1:j-2}} \cdots \hat{w}_{i_1}^1 \right| \\
&\leq 1 \cdot \left(\sup_{f\in\mathcal{F}} \max_{\boldsymbol{i}_{1:j-1}\in[n]^{j-1}} \left| \int_{\mathcal{Z}^j} \ell_{f,j}^{\boldsymbol{i}_{1:j-1}}(z^j) p_j(z^j|\mathbf{Z}^{\mathrm{mp}(j)}_{\boldsymbol{i}_{1:j-1}}) \mathrm{d}z^j - \sum_{i_j=1}^{n} \ell_{f,j}^{\boldsymbol{i}_{1:j-1}}(Z_{i_j}^j) \hat{w}_{i_j}^j(Z^{\mathrm{mp}(j)}_{\boldsymbol{i}_{1:j-1}}) \right| \right) \\
&\leq \max_{\boldsymbol{i}_{1:j-1}\in[n]^{j-1}} \sup_{f\in\mathcal{F}} \sup_{\boldsymbol{z}^{\mathrm{mp}(j)}\in\mathcal{Z}^{\mathrm{mp}(j)}} \left| \int_{\mathcal{Z}^j} \ell_{f,j}^{\boldsymbol{i}_{1:j-1}}(z^j) p_j(z^j|z^{\mathrm{mp}(j)}) \mathrm{d}z^j - \sum_{i_j=1}^{n} \ell_{f,j}^{\boldsymbol{i}_{1:j-1}}(Z_{i_j}^j) \hat{w}_{i_j}^j(z^{\mathrm{mp}(j)}) \right| \\
&\leq \sup_{\ell'_{f,j}\in\mathcal{L}_{\mathcal{F}}^j} \sup_{\boldsymbol{z}^{\mathrm{mp}(j)}\in\mathcal{Z}^{\mathrm{mp}(j)}} \underbrace{\left| \int_{\mathcal{Z}^j} \ell'_{f,j}(z^j) p_j(z^j|z^{\mathrm{mp}(j)}) \mathrm{d}z^j - \sum_{i_j=1}^{n} \ell'_{f,j}(Z_{i_j}^j) \hat{w}_{i_j}^j(z^{\mathrm{mp}(j)}) \right|}_{(**)},
\end{aligned}
$$

where we used that $\left\{ \mathbf{Z}^{\mathrm{mp}(j)}_{\boldsymbol{i}_{1:j-1}} \right\}_{\boldsymbol{i}_{1:j-1}\in[n]^{j-1}} \subset \mathcal{Z}^{\mathrm{mp}(j)}$ that follows from $\left\{ \mathbf{Z}_i^{\mathrm{mp}(j)} \right\}_{i=1}^n \subset \mathcal{Z}^{\mathrm{mp}(j)}$. Define

$$
\begin{aligned}
r^j(f, z^{\mathrm{mp}(j)}) &:= \int_{\mathcal{Z}^j} f(z^j) p_{j,\mathrm{mp}(j)}(z^j, z^{\mathrm{mp}(j)}) \mathrm{d}z^j, & \hat{r}^j(f, z^{\mathrm{mp}(j)}) &:= \frac{1}{n} \sum_{i=1}^{n} f(Z_i^j) K_{\mathbf{H}}^j(z^{\mathrm{mp}(j)} - \mathbf{Z}_i^{\mathrm{mp}(j)}), \\
g^j(z^{\mathrm{mp}(j)}) &:= p_{\mathrm{mp}(j)}(z^{\mathrm{mp}(j)}), & \hat{g}^j(z^{\mathrm{mp}(j)}) &:= \frac{1}{n} \sum_{i=1}^{n} K_{\mathbf{H}}^j(z^{\mathrm{mp}(j)} - \mathbf{Z}_i^{\mathrm{mp}(j)}).
\end{aligned}
$$

Then, for each $\ell'_{f,j} \in \mathcal{L}_{\mathcal{F}}^j$ and $z^{\mathrm{mp}(j)} \in \mathcal{Z}^{\mathrm{mp}(j)}$,

$$
\begin{aligned}
(**) &= \left| \frac{r^j(\ell'_{f,j}, z^{\mathrm{mp}(j)})}{g^j(z^{\mathrm{mp}(j)})} - \frac{\hat{r}^j(\ell'_{f,j}, z^{\mathrm{mp}(j)})}{\hat{g}^j(z^{\mathrm{mp}(j)})} \right| \\
&\leq \underbrace{\left| \frac{r^j(\ell'_{f,j}, z^{\mathrm{mp}(j)})}{g^j(z^{\mathrm{mp}(j)})} - \frac{\mathbb{E}\hat{r}^j(\ell'_{f,j}, z^{\mathrm{mp}(j)})}{\mathbb{E}\hat{g}^j(z^{\mathrm{mp}(j)})} \right|}_{\rho_1} + \underbrace{\left| \frac{\mathbb{E}\hat{r}^j(\ell'_{f,j}, z^{\mathrm{mp}(j)})}{\mathbb{E}\hat{g}^j(z^{\mathrm{mp}(j)})} - \frac{\hat{r}^j(\ell'_{f,j}, z^{\mathrm{mp}(j)})}{\hat{g}^j(z^{\mathrm{mp}(j)})} \right|}_{\rho_2}.
\end{aligned}
$$

By applying the argument of Equation (4), we can bound each ratio difference term as

$$
\begin{aligned}
\rho_1 &\leq \left| \frac{1}{g^j(z^{\mathrm{mp}(j)})} \right| \cdot |r^j(\ell'_{f,j}, z^{\mathrm{mp}(j)}) - \mathbb{E}\hat{r}^j(\ell'_{f,j}, z^{\mathrm{mp}(j)})| + \left| \frac{\mathbb{E}\hat{r}^j(z^{\mathrm{mp}(j)})}{g^j(z^{\mathrm{mp}(j)})\mathbb{E}\hat{g}^j(z^{\mathrm{mp}(j)})} \right| \cdot |g^j(z^{\mathrm{mp}(j)}) - \mathbb{E}\hat{g}^j(z^{\mathrm{mp}(j)})| \\
\rho_2 &\leq \left| \frac{1}{\mathbb{E}\hat{g}^j(z^{\mathrm{mp}(j)})} \right| \cdot |\mathbb{E}\hat{r}^j(\ell'_{f,j}, z^{\mathrm{mp}(j)}) - \hat{r}^j(\ell'_{f,j}, z^{\mathrm{mp}(j)})| + \left| \frac{\hat{r}^j(z^{\mathrm{mp}(j)})}{\mathbb{E}\hat{g}^j(z^{\mathrm{mp}(j)})\hat{g}^j(z^{\mathrm{mp}(j)})} \right| \cdot |\mathbb{E}\hat{g}^j(z^{\mathrm{mp}(j)}) - \hat{g}^j(z^{\mathrm{mp}(j)})|.
\end{aligned}
$$

Applying Lemma 1 to the coefficients, Lemma 2 to the deterministic difference terms bounding ρ_1, Lemma 3 to the stochastic difference terms bounding ρ_2 along with the union bound, for any $\delta \in (0, 1)$, we have with probability at least $1 - 2D\delta$,

$$
\begin{aligned}
R(\hat{f}) - R(f^*) \leq 2 \sum_{j=1}^{D} \Bigg(& \frac{1}{\epsilon_{\mathrm{mp}(j)}} \left(B_\ell B_j \Phi(\beta, L, K^j) \left\| \mathbf{H}_j \right\|_{\mathrm{op}}^{\beta} + B_\ell \phi_{K^j, \mathbf{H}_j} \check{I}_{j,\mathrm{mp}(j)} \right) \\
& + \frac{1}{\epsilon_{\mathrm{mp}(j)}} \cdot \frac{B_\ell B_{\mathbf{K}}}{\epsilon_{K^j}(\mathbf{H}_j)} \left(\Phi(\beta, L, K^j) \left\| \mathbf{H}_j \right\|_{\mathrm{op}}^{\beta} + \phi_{K^j, \mathbf{H}_j} \check{I}_{\mathrm{mp}(j)} \right) \\
& + \frac{\left| \det \mathbf{H}_j \right|}{\epsilon_{K^j}(\mathbf{H}_j)} \left(2\mathrm{Rad}_{n,p}\left(\mathcal{L}_{\mathcal{F}}^j \otimes \mathcal{K}_{\mathbf{H}}^j \right) + \frac{B_\ell B_{\mathbf{K}}}{\left| \det \mathbf{H}_j \right|} \sqrt{\frac{\log(2/\delta)}{2n}} \right) \\
& + \frac{\left| \det \mathbf{H}_j \right|}{\epsilon_{K^j}(\mathbf{H}_j)} \cdot \frac{B_\ell B_{\mathbf{K}}}{\epsilon_{K^j}(\mathbf{H}_j)} \left(2\mathrm{Rad}_{n,p}(\mathcal{K}_{\mathbf{H}}^j) + \frac{B_{\mathbf{K}}}{\left| \det \mathbf{H}_j \right|} \sqrt{\frac{\log(2/\delta)}{2n}} \right) \Bigg).
\end{aligned}
$$

By reorganizing the terms, we obtain the assertion. □

C.5 Lemmas

Here, we prove the lemmas used in the proof of Theorem 2.

Lemma 1 (Bounded coefficients)**.** *Assume Assumption 1 holds. Let $j \in [D]$. Then,*

$$\sup_{\boldsymbol{z}^{\mathrm{mp}(j)} \in \mathcal{Z}^{\mathrm{mp}(j)}} \left| \frac{1}{g^j(\boldsymbol{z}^{\mathrm{mp}(j)})} \right| \leq \frac{1}{\epsilon_{\mathrm{mp}(j)}}, \qquad \sup_{\ell'_{f,j} \in \mathcal{L}^j_{\mathcal{F}}} \sup_{\boldsymbol{z}^{\mathrm{mp}(j)} \in \mathcal{Z}^{\mathrm{mp}(j)}} \left| \frac{\mathbb{E}\hat{r}^j(\ell'_{f,j}, \boldsymbol{z}^{\mathrm{mp}(j)})}{\mathbb{E}\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})} \right| \leq \frac{B_\ell B_{\mathbf{K}}}{\epsilon_{K^j}(\mathbf{H}_j)},$$

$$\sup_{\boldsymbol{z}^{\mathrm{mp}(j)} \in \mathcal{Z}^{\mathrm{mp}(j)}} \left| \frac{1}{\mathbb{E}\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})} \right| \leq \frac{\left|\det \mathbf{H}_j\right|}{\epsilon_{K^j}(\mathbf{H}_j)}, \qquad \sup_{\ell'_{f,j} \in \mathcal{L}^j_{\mathcal{F}}} \sup_{\boldsymbol{z}^{\mathrm{mp}(j)} \in \mathcal{Z}^{\mathrm{mp}(j)}} \left| \frac{\hat{r}^j(\ell'_{f,j}, \boldsymbol{z}^{\mathrm{mp}(j)})}{\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})} \right| \leq \frac{B_\ell B_{\mathbf{K}}}{\epsilon_{K^j}(\mathbf{H}_j)}.$$

Proof. By Assumption 1, we have

$$\sup_{\boldsymbol{z}^{\mathrm{mp}(j)} \in \mathcal{Z}^{\mathrm{mp}(j)}} \left| \frac{1}{g^j(\boldsymbol{z}^{\mathrm{mp}(j)})} \right| = \frac{1}{\inf_{\boldsymbol{z}^{\mathrm{mp}(j)} \in \mathcal{Z}^{\mathrm{mp}(j)}} p_{\mathrm{mp}(j)}(\boldsymbol{z}^{\mathrm{mp}(j)})} \leq \frac{1}{\epsilon_{\mathrm{mp}(j)}}.$$

Also,

$$\begin{aligned}
\sup_{\boldsymbol{z}^{\mathrm{mp}(j)} \in \mathcal{Z}^{\mathrm{mp}(j)}} \left| \frac{1}{\mathbb{E}\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})} \right| &\leq \frac{1}{\inf_{\boldsymbol{z}^{\mathrm{mp}(j)} \in \mathcal{Z}^{\mathrm{mp}(j)}} \left|\mathbb{E}\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})\right|} \\
&= \frac{1}{\inf_{\boldsymbol{z}^{\mathrm{mp}(j)} \in \mathcal{Z}^{\mathrm{mp}(j)}} \left| \int_{\mathcal{Z}^{\mathrm{mp}(j)}} K^j_{\mathbf{H}}(\boldsymbol{z}^{\mathrm{mp}(j)} - \boldsymbol{z}^{\mathrm{mp}(j)\prime}) g^j(\boldsymbol{z}^{\mathrm{mp}(j)\prime}) \mathrm{d}\boldsymbol{z}^{\mathrm{mp}(j)\prime} \right|} \\
&= \frac{1}{\inf_{\boldsymbol{z}^{\mathrm{mp}(j)} \in \mathcal{Z}^{\mathrm{mp}(j)}} \int_{\mathcal{Z}^{\mathrm{mp}(j)}} K^j_{\mathbf{H}}(\boldsymbol{z}^{\mathrm{mp}(j)} - \boldsymbol{z}^{\mathrm{mp}(j)\prime}) g^j(\boldsymbol{z}^{\mathrm{mp}(j)\prime}) \mathrm{d}\boldsymbol{z}^{\mathrm{mp}(j)\prime}} \\
&\leq \frac{1}{|\det \mathbf{H}_j|^{-1} \epsilon_{K^j}(\mathbf{H}_j) \int_{\mathcal{Z}^{\mathrm{mp}(j)}} g^j(\boldsymbol{z}^{\mathrm{mp}(j)\prime}) \mathrm{d}\boldsymbol{z}^{\mathrm{mp}(j)\prime}} = \frac{\left|\det \mathbf{H}_j\right|}{\epsilon_{K^j}(\mathbf{H}_j)},
\end{aligned}$$

where we used the positivity of the integrand. Now,

$$\begin{aligned}
\sup_{\ell'_{f,j} \in \mathcal{L}^j_{\mathcal{F}}} \sup_{\boldsymbol{z}^{\mathrm{mp}(j)} \in \mathcal{Z}^{\mathrm{mp}(j)}} \left| \frac{\mathbb{E}\hat{r}^j(\ell'_{f,j}, \boldsymbol{z}^{\mathrm{mp}(j)})}{\mathbb{E}\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})} \right| &= \sup_{\ell'_{f,j} \in \mathcal{L}^j_{\mathcal{F}}} \sup_{\boldsymbol{z}^{\mathrm{mp}(j)} \in \mathcal{Z}^{\mathrm{mp}(j)}} \left| \frac{\left|\det \mathbf{H}_j\right| \mathbb{E}\hat{r}^j(\ell'_{f,j}, \boldsymbol{z}^{\mathrm{mp}(j)})}{\left|\det \mathbf{H}_j\right| \mathbb{E}\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})} \right| \\
&\leq \frac{\sup_{\ell'_{f,j} \in \mathcal{L}^j_{\mathcal{F}}} \sup_{\boldsymbol{z}^{\mathrm{mp}(j)} \in \mathcal{Z}^{\mathrm{mp}(j)}} \left\|\ell'_{f,j}\right\|_\infty \cdot \left\| \left(\left|\det \mathbf{H}_j\right| K^j_{\mathbf{H}} \right) \right\|_\infty}{\inf_{\boldsymbol{z}^{\mathrm{mp}(j)} \in \mathcal{Z}^{\mathrm{mp}(j)}} \left|\det \mathbf{H}_j\right| |\mathbb{E}\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})|} \leq \frac{B_\ell B_{\mathbf{K}}}{\epsilon_{K^j}(\mathbf{H}_j)}.
\end{aligned}$$

Similarly, we have $\inf_{\boldsymbol{z}^{\mathrm{mp}(j)} \in \mathcal{Z}^{\mathrm{mp}(j)}} \left|\det \mathbf{H}_j\right| \cdot \left|\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})\right| \geq \epsilon_{K^j}(\mathbf{H}_j)$. Therefore,

$$\begin{aligned}
\sup_{\ell'_{f,j} \in \mathcal{L}^j_{\mathcal{F}}} \sup_{\boldsymbol{z}^{\mathrm{mp}(j)} \in \mathcal{Z}^{\mathrm{mp}(j)}} \left| \frac{\hat{r}^j(\ell'_{f,j}, \boldsymbol{z}^{\mathrm{mp}(j)})}{\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})} \right| &= \sup_{\ell'_{f,j} \in \mathcal{L}^j_{\mathcal{F}}} \sup_{\boldsymbol{z}^{\mathrm{mp}(j)} \in \mathcal{Z}^{\mathrm{mp}(j)}} \left| \frac{\left|\det \mathbf{H}_j\right| \hat{r}^j(\ell'_{f,j}, \boldsymbol{z}^{\mathrm{mp}(j)})}{\left|\det \mathbf{H}_j\right| \hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})} \right| \\
&\leq \frac{\sup_{\ell'_{f,j} \in \mathcal{L}^j_{\mathcal{F}}} \sup_{\boldsymbol{z}^{\mathrm{mp}(j)} \in \mathcal{Z}^{\mathrm{mp}(j)}} \left|\det \mathbf{H}_j\right| \cdot \left|\hat{r}^j(\ell'_{f,j}, \boldsymbol{z}^{\mathrm{mp}(j)})\right|}{\inf_{\boldsymbol{z}^{\mathrm{mp}(j)} \in \mathcal{Z}^{\mathrm{mp}(j)}} \left|\det \mathbf{H}_j\right| \cdot \left|\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})\right|} \leq \frac{B_\ell B_{\mathbf{K}}}{\epsilon_{K^j}(\mathbf{H}_j)}.
\end{aligned}$$

□

Lemma 2 (Deterministic terms)**.** *Assume that Assumptions 1 and 2 hold. Let $j \in [D]$. Then,*

$$\sup_{\ell'_{f,j} \in \mathcal{L}^j_{\mathcal{F}}} \sup_{\boldsymbol{z}^{\mathrm{mp}(j)} \in \mathcal{Z}^{\mathrm{mp}(j)}} |r^j(\ell'_{f,j}, \boldsymbol{z}^{\mathrm{mp}(j)}) - \mathbb{E}\hat{r}^j(\ell'_{f,j}, \boldsymbol{z}^{\mathrm{mp}(j)})| \leq B_\ell B_j \Phi(\beta, L, K^j) \left\|\mathbf{H}_j\right\|^\beta_{\mathrm{op}} + B_\ell \phi_{K^j, \mathbf{H}_j} \check{I}_{j, \mathrm{mp}(j)},$$

$$\sup_{\boldsymbol{z}^{\mathrm{mp}(j)} \in \mathcal{Z}^{\mathrm{mp}(j)}} |g^j(\boldsymbol{z}^{\mathrm{mp}(j)}) - \mathbb{E}\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})| \leq \Phi(\beta, L, K^j) \left\|\mathbf{H}_j\right\|^\beta_{\mathrm{op}} + \phi_{K^j, \mathbf{H}_j} \check{I}_{\mathrm{mp}(j)}.$$

Proof. By applying Lemma 4 under Assumption 2,

$$
\begin{aligned}
&\sup_{\boldsymbol{z}^{\mathrm{mp}(j)}\in\mathcal{Z}^{\mathrm{mp}(j)}} |g^j(\boldsymbol{z}^{\mathrm{mp}(j)}) - \mathbb{E}\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})| \\
&= \sup_{\boldsymbol{z}^{\mathrm{mp}(j)}\in\mathcal{Z}^{\mathrm{mp}(j)}} \left| p_{\mathrm{mp}(j)}(\boldsymbol{z}^{\mathrm{mp}(j)}) - \int_{\mathcal{Z}^{\mathrm{mp}(j)}} K^j_{\mathbf{H}}(\boldsymbol{z}^{\mathrm{mp}(j)} - \boldsymbol{z}^{\mathrm{mp}(j)\prime}) p_{\mathrm{mp}(j)}(\boldsymbol{z}^{\mathrm{mp}(j)\prime}) \mathrm{d}\boldsymbol{z}^{\mathrm{mp}(j)\prime} \right| \\
&= \sup_{\boldsymbol{z}^{\mathrm{mp}(j)}\in\mathcal{Z}^{\mathrm{mp}(j)}} \left| \check{p}_{\mathrm{mp}(j)}(\boldsymbol{z}^{\mathrm{mp}(j)}) - \int_{\mathcal{Z}^{\mathrm{mp}(j)}} K^j_{\mathbf{H}}(\boldsymbol{z}^{\mathrm{mp}(j)} - \boldsymbol{z}^{\mathrm{mp}(j)\prime}) \check{p}_{\mathrm{mp}(j)}(\boldsymbol{z}^{\mathrm{mp}(j)\prime}) \mathrm{d}\boldsymbol{z}^{\mathrm{mp}(j)\prime} \right| \\
&\leq \sup_{\boldsymbol{z}^{\mathrm{mp}(j)}\in\mathcal{Z}^{\mathrm{mp}(j)}} \left| \check{p}_{\mathrm{mp}(j)}(\boldsymbol{z}^{\mathrm{mp}(j)}) - \left(K^j_{\mathbf{H}} * \check{p}_{\mathrm{mp}(j)}\right)(\boldsymbol{z}^{\mathrm{mp}(j)}) \right| + \sup_{\boldsymbol{z}^{\mathrm{mp}(j)}\in\mathcal{Z}^{\mathrm{mp}(j)}} \left| \int_{\mathbb{R}^{|\mathrm{mp}(j)|}\setminus\mathcal{Z}^{\mathrm{mp}(j)}} K^j_{\mathbf{H}}(\boldsymbol{z}^{\mathrm{mp}(j)} - \boldsymbol{z}^{\mathrm{mp}(j)\prime}) \check{p}_{\mathrm{mp}(j)}(\boldsymbol{z}^{\mathrm{mp}(j)\prime}) \mathrm{d}\boldsymbol{z}^{\mathrm{mp}(j)\prime} \right| \\
&\leq \Phi(\beta, L, K^j) \left\|\mathbf{H}_j\right\|_{\mathrm{op}}^{\beta} + \phi_{K^j,\mathbf{H}_j} \check{I}_{\mathrm{mp}(j)}.
\end{aligned}
$$

Similarly, for each $\ell'_{f,j} \in \mathcal{L}^j_{\mathcal{F}}$ and $\boldsymbol{z}^{\mathrm{mp}(j)} \in \mathcal{Z}^{\mathrm{mp}(j)}$,

$$
\begin{aligned}
&|r^j(\ell'_{f,j}, \boldsymbol{z}^{\mathrm{mp}(j)}) - \mathbb{E}\hat{r}^j(\ell'_{f,j}, \boldsymbol{z}^{\mathrm{mp}(j)})| \\
&= \left| \int_{\mathcal{Z}^j} \ell'_{f,j}(\boldsymbol{z}^j) p_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j, \boldsymbol{z}^{\mathrm{mp}(j)}) \mathrm{d}\boldsymbol{z}^j - \int_{\mathcal{Z}^j} \ell'_{f,j}(\boldsymbol{z}^j) \left(\int_{\mathcal{Z}^{\mathrm{mp}(j)}} K^j_{\mathbf{H}}(\boldsymbol{z}^{\mathrm{mp}(j)} - \boldsymbol{z}^{\mathrm{mp}(j)\prime}) p_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j, \boldsymbol{z}^{\mathrm{mp}(j)\prime}) \mathrm{d}\boldsymbol{z}^{\mathrm{mp}(j)\prime} \right) \mathrm{d}\boldsymbol{z}^j \right| \\
&= \left| \int_{\mathcal{Z}^j} \ell'_{f,j}(\boldsymbol{z}^j) \check{p}_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j, \boldsymbol{z}^{\mathrm{mp}(j)}) \mathrm{d}\boldsymbol{z}^j - \int_{\mathcal{Z}^j} \ell'_{f,j}(\boldsymbol{z}^j) \left(\int_{\mathcal{Z}^{\mathrm{mp}(j)}} K^j_{\mathbf{H}}(\boldsymbol{z}^{\mathrm{mp}(j)} - \boldsymbol{z}^{\mathrm{mp}(j)\prime}) \check{p}_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j, \boldsymbol{z}^{\mathrm{mp}(j)\prime}) \mathrm{d}\boldsymbol{z}^{\mathrm{mp}(j)\prime} \right) \mathrm{d}\boldsymbol{z}^j \right| \\
&\leq \left| \int_{\mathcal{Z}^j} \ell'_{f,j}(\boldsymbol{z}^j) \left(\check{p}_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j, \boldsymbol{z}^{\mathrm{mp}(j)}) - (K^j_{\mathbf{H}} * \check{p}_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j, \cdot))(\boldsymbol{z}^{\mathrm{mp}(j)}) \right) \mathrm{d}\boldsymbol{z}^j \right| \\
&\quad + \left| \int_{\mathcal{Z}^j} \ell'_{f,j}(\boldsymbol{z}^j) \left(\int_{\mathbb{R}^{|\mathrm{mp}(j)|}\setminus\mathcal{Z}^{\mathrm{mp}(j)}} K^j_{\mathbf{H}}(\boldsymbol{z}^{\mathrm{mp}(j)} - \boldsymbol{z}^{\mathrm{mp}(j)\prime}) \check{p}_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j, \boldsymbol{z}^{\mathrm{mp}(j)\prime}) \mathrm{d}\boldsymbol{z}^{\mathrm{mp}(j)\prime} \right) \mathrm{d}\boldsymbol{z}^j \right| \\
&\leq B_\ell \int_{\mathcal{Z}^j} \left| \check{p}_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j, \boldsymbol{z}^{\mathrm{mp}(j)}) - (K^j_{\mathbf{H}} * \check{p}_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j, \cdot))(\boldsymbol{z}^{\mathrm{mp}(j)}) \right| \mathrm{d}\boldsymbol{z}^j + B_\ell \phi_{K^j,\mathbf{H}_j} \check{I}_{j,\mathrm{mp}(j)} \\
&\leq B_\ell B_j \sup_{\boldsymbol{z}^j\in\mathcal{Z}^j} \left| \check{p}_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j, \boldsymbol{z}^{\mathrm{mp}(j)}) - (K^j_{\mathbf{H}} * \check{p}_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j, \cdot))(\boldsymbol{z}^{\mathrm{mp}(j)}) \right| + B_\ell \phi_{K^j,\mathbf{H}_j} \check{I}_{j,\mathrm{mp}(j)} \\
&\leq B_\ell B_j \sup_{\boldsymbol{z}^j\in\mathcal{Z}^j} \sup_{\boldsymbol{z}^{\mathrm{mp}(j)}\in\mathcal{Z}^{\mathrm{mp}(j)}} \left| \check{p}_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j, \boldsymbol{z}^{\mathrm{mp}(j)}) - (K^j_{\mathbf{H}} * \check{p}_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j, \cdot))(\boldsymbol{z}^{\mathrm{mp}(j)}) \right| + B_\ell \phi_{K^j,\mathbf{H}_j} \check{I}_{j,\mathrm{mp}(j)}.
\end{aligned}
$$

Applying Lemma 4 under Assumption 2, for each $\boldsymbol{z}^j \in \mathcal{Z}^j$, we obtain

$$
\sup_{\boldsymbol{z}^{\mathrm{mp}(j)}\in\mathcal{Z}^{\mathrm{mp}(j)}} \left| \check{p}_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j, \boldsymbol{z}^{\mathrm{mp}(j)}) - (K^j_{\mathbf{H}} * \check{p}_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j, \cdot))(\boldsymbol{z}^{\mathrm{mp}(j)}) \right| \leq \Phi(\beta, L, K^j) \left\|\mathbf{H}_j\right\|_{\mathrm{op}}^{\beta}.
$$

Therefore, we have the assertion. □

Lemma 3 (Probabilistic terms)**.** *Assume that Assumption 1 holds. Let $j \in [D]$. For any $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$
\sup_{\ell'_{f,j}\in\mathcal{L}^j_{\mathcal{F}}} \sup_{\boldsymbol{z}^{\mathrm{mp}(j)}\in\mathcal{Z}^{\mathrm{mp}(j)}} |\mathbb{E}\hat{r}^j(\ell'_{f,j}, \boldsymbol{z}^{\mathrm{mp}(j)}) - \hat{r}^j(\ell'_{f,j}, \boldsymbol{z}^{\mathrm{mp}(j)})| \leq 2\mathrm{Rad}_{n,p}\left(\mathcal{L}^j_{\mathcal{F}} \otimes \mathcal{K}^j_{\mathbf{H}}\right) + \frac{B_\ell B_{\mathbf{K}}}{\left|\det \mathbf{H}_j\right|} \sqrt{\frac{\log(2/\delta)}{2n}}.
$$

Similarly, for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$
\sup_{\boldsymbol{z}^{\mathrm{mp}(j)}\in\mathcal{Z}^{\mathrm{mp}(j)}} |\mathbb{E}\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)}) - \hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})| \leq 2\mathrm{Rad}_{n,p}(\mathcal{K}^j_{\mathbf{H}}) + \frac{B_{\mathbf{K}}}{\left|\det \mathbf{H}_j\right|} \sqrt{\frac{\log(2/\delta)}{2n}}.
$$

Proof. Note

$$
\sup_{\ell'_{f,j}\in\mathcal{L}^j_{\mathcal{F}}} \sup_{\boldsymbol{z}^{\mathrm{mp}(j)}\in\mathcal{Z}^{\mathrm{mp}(j)}} |\mathbb{E}\hat{r}^j(\ell'_{f,j}, \boldsymbol{z}^{\mathrm{mp}(j)}) - \hat{r}^j(\ell'_{f,j}, \boldsymbol{z}^{\mathrm{mp}(j)})| = \sup_{\ell'_{f,j}\in\mathcal{L}^j_{\mathcal{F}}} \sup_{k\in\mathcal{K}^j_{\mathbf{H}}} \left| \frac{1}{n}\sum_{i=1}^n \ell'_{f,j}(Z^j_i) k(\mathbf{Z}^{\mathrm{mp}(j)}_i) - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \ell'_{f,j}(Z^j_i) k(\mathbf{Z}^{\mathrm{mp}(j)}_i) \right] \right|
$$

and

$$
\sup_{\boldsymbol{z}^{\mathrm{mp}(j)}\in\mathcal{Z}^{\mathrm{mp}(j)}} |\mathbb{E}\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)}) - \hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})| = \sup_{k\in\mathcal{K}^j_{\mathbf{H}}} \left| \frac{1}{n}\sum_{i=1}^n k(\mathbf{Z}^{\mathrm{mp}(j)}_i) - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n k(\mathbf{Z}^{\mathrm{mp}(j)}_i) \right] \right|.
$$

Now, applying Fact 3 to these expressions, we obtain the assertions of the lemma. □

C.6 Facts

Here, we state some facts used in the proof of Theorem 2. The following is the Taylor's formula with integral form of the remainder, stated using the multi-index notation.

Fact 1 (Taylor's theorem; [24], Section 8.4.4)**.** *Let $\Omega \subset \mathbb{R}^n$ be an open subset. Let $n \in \mathbb{N}$, and let $f : \Omega \to \mathbb{R}$ be k-times continuously differentiable. Then, for any $x, u \in \Omega$ such that $x + tu \in \Omega$ for all $t \in [0, 1]$, the following equality holds:*

$$f(x+u) - f(x) = \sum_{1 \le |\alpha| < k} \frac{\partial^\alpha f(x)}{\alpha!} u^\alpha + \sum_{|\alpha|=k} \frac{|\alpha|}{\alpha!} u^\alpha \int_0^1 (1-t)^{|\alpha|-1} \partial^\alpha f(x+tu) \mathrm{d}t.$$

The following elementary inequality is easily proved by using the strict convexity and the strict monotonicity of the logarithm function.

Fact 2 (Weighted AM-GM inequality)**.** *Let $n \in \mathbb{N}$, $x_1, \ldots, x_n \ge 0$, and $w_1, \ldots, w_n \ge 0$. Define $w := w_1 + \cdots + w_n$ and assume $w > 0$. Then,*

$$\frac{w_1 x_1 + \cdots + w_n x_n}{w} \ge \left(x_1^{w_1} \cdots x_n^{w_n} \right)^{\frac{1}{w}}.$$

The following standard Rademacher complexity bound is essentially due to McDiarmid's inequality, which is applied twice with the union bound [61, Theorem 3.3].

Fact 3 (Rademacher complexity bound; Theorem 3.3 in [61])**.** *Let $B > 0$ and $m \in \mathbb{N}$. Let $\mathcal{G}$ be a family of functions mapping from $\mathcal{Z}$ to $[0, B]$, and let z be a $\mathcal{Z}$-valued random variable. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over the draw of an independent and identically distributed sample $\{z_i\}_{i=1}^m \overset{i.i.d.}{\sim} z$, the following holds:*

$$\sup_{g \in \mathcal{G}} \left| \frac{1}{m} \sum_{i=1}^m g(z_i) - \mathbb{E}[g(z)] \right| \le 2\mathrm{Rad}_{m,p}(\mathcal{G}) + B\sqrt{\frac{\log(2/\delta)}{2m}}.$$

C.7 Basic Lemmas

Here, we prove the basic lemmas used in the proof of Theorem 2.

Lemma 4 (Convolution error bound for Hölder class)**.** *Let $d \in \mathbb{N}$, $\beta > 1$, and $L > 0$. Assume that the kernel function $K : \mathbb{R}^d \to \mathbb{R}$ is of order $k = \lfloor \beta \rfloor$ and satisfies*

$$\int_{\mathbb{R}^d} |K(u)| \cdot \|u\|^\beta \mathrm{d}u < \infty.$$

Let $\mathbf{H} = \mathrm{diag}(h_1, \ldots, h_d)$ with $h_1, \ldots, h_d > 0$, and define $K_{\mathbf{H}}(u) := \frac{1}{|\det \mathbf{H}|} K(\mathbf{H}^{-1}u)$. Then, for any $f \in \Sigma(\beta, L)$, the following holds:

$$\sup_{\boldsymbol{x} \in \mathbb{R}^d} |f(\boldsymbol{x}) - (K_{\mathbf{H}} * f)(\boldsymbol{x})| \le \Phi(\beta, L, K) \|\mathbf{H}\|_{\mathrm{op}}^\beta,$$

where $\Phi(\beta, L, K)$ is defined as

$$\Phi(\beta, L, K) := L\left(\int_0^1 (1-t)^{k-1} t^{\beta-k} \mathrm{d}t \right) \sum_{|\alpha|=k} \frac{\|\alpha\|^k}{\alpha! k^{k-1}} \int_{\mathbb{R}^d} |K(u)| \cdot \|u\|^\beta \mathrm{d}u$$

and $\alpha \in \mathbb{Z}_{\ge 0}^d$ runs over multi-indices.

Proof. First, we fix $x \in \mathbb{R}^d$. We apply the change of variables formula and obtain

$$|f(x) - (K_{\mathbf{H}} * f)(x)| = \left| f(x) - \int_{\mathbb{R}^d} K(u) f(x - \mathbf{H}u) \mathrm{d}u \right|. \tag{*}$$

We apply Fact 1 to obtain

$$
\begin{aligned}
(*) &= \left| f(x) - \int_{\mathbb{R}^d} K(u) \left(f(x) + \sum_{1 \le |\alpha| < k} \frac{\partial^\alpha f(x)}{\alpha!} (-\mathbf{H}u)^\alpha + \sum_{|\alpha|=k} \frac{|\alpha|}{\alpha!} (-\mathbf{H}u)^\alpha \int_0^1 (1-t)^{|\alpha|-1} \partial^\alpha f(x + t(-\mathbf{H}u)) \mathrm{d}t \right) \mathrm{d}u \right| \\
&= \left| \int_{\mathbb{R}^d} K(u) \left(\sum_{|\alpha|=k} \frac{|\alpha|}{\alpha!} (-\mathbf{H}u)^\alpha \int_0^1 (1-t)^{|\alpha|-1} \partial^\alpha f(x - t\mathbf{H}u) \mathrm{d}t \right) \mathrm{d}u \right| \\
&= \left| \int_{\mathbb{R}^d} K(u) \left(\sum_{|\alpha|=k} \frac{|\alpha|}{\alpha!} (-\mathbf{H}u)^\alpha \int_0^1 (1-t)^{|\alpha|-1} (\partial^\alpha f(x - t\mathbf{H}u) - \partial^\alpha f(x)) \mathrm{d}t \right) \mathrm{d}u \right| \\
&\le \int_{\mathbb{R}^d} |K(u)| \left(\sum_{|\alpha|=k} \frac{|\alpha|}{\alpha!} |\mathbf{H}u|^\alpha \int_0^1 (1-t)^{|\alpha|-1} |\partial^\alpha f(x - t\mathbf{H}u) - \partial^\alpha f(x)| \mathrm{d}t \right) \mathrm{d}u, \qquad (**)
\end{aligned}
$$

where $\alpha = (\alpha_1, \ldots, \alpha_d)$ is a multi-index and $|\mathbf{H}u|^\alpha := |h_1 u_1|^{\alpha_1} \cdots |h_d u_d|^{\alpha_d}$. Now, by the Hölder-condition of $\partial^\alpha f$, we have $|\partial^\alpha f(x - t\mathbf{H}u) - \partial^\alpha f(x)| \le L \|t\mathbf{H}u\|^{\beta-k}$. Also, by applying Fact 2, we have

$$
|\mathbf{H}u|^\alpha = |h_1 u_1|^{\alpha_1} \cdots |h_d u_d|^{\alpha_d} \le \left(\frac{1}{|\alpha|} \sum_{j=1}^d \alpha_j |h_j u_j| \right)^{|\alpha|} \le \left(\frac{1}{|\alpha|} \|\alpha\| \cdot \|hu\| \right)^{|\alpha|} = \frac{\|\alpha\|^k}{k^k} \|hu\|^k .
$$

By applying these inequalities and imputing $|\alpha| = k$, we obtain

$$
\begin{aligned}
(**) &\le \int_{\mathbb{R}^d} |K(u)| \left(\sum_{|\alpha|=k} \frac{\|\alpha\|^k}{\alpha! k^{k-1}} \|\mathbf{H}u\|^k \int_0^1 (1-t)^{k-1} L \|t\mathbf{H}u\|^{\beta-k} \mathrm{d}t \right) \mathrm{d}u \\
&= L \left(\int_0^1 (1-t)^{k-1} t^{\beta-k} \mathrm{d}t \right) \sum_{|\alpha|=k} \frac{\|\alpha\|^k}{\alpha! k^{k-1}} \int_{\mathbb{R}^d} |K(u)| \cdot \|\mathbf{H}u\|^\beta \mathrm{d}u.
\end{aligned}
$$

Finally, applying $\|\mathbf{H}u\| \le \|\mathbf{H}\|_{\mathrm{op}} \|u\|$, we have the assertion. □

Lemma 5 (Bounded weights)**.** *For all $j \in [D]$,*

$$
\sum_{i_1=1}^n \cdots \sum_{i_j=1}^n \hat{w}_{i_j | \boldsymbol{i}_{1:j-1}} \cdots \hat{w}^1_{i_1} \in \{0, 1\}.
$$

Proof. By direct computation, we have for any $\boldsymbol{z}^{\mathfrak{mp}(j)} \in \mathcal{Z}^{\mathfrak{mp}(j)}$,

$$
\begin{aligned}
\sum_{i=1}^n \hat{w}^j_i(\boldsymbol{z}^{\mathfrak{mp}(j)}) &= \begin{cases} \sum_{i=1}^n \frac{1}{n} & \text{if } \mathfrak{mp}(j) = \emptyset, \\ \sum_{i=1}^n 0 & \text{if } K^j_{\mathbf{H}}(\boldsymbol{z}^{\mathfrak{mp}(j)} - \mathbf{Z}_i^{\mathfrak{mp}(j)}) = 0, \forall i, \\ \sum_{i=1}^n \frac{K^j_{\mathbf{H}}(\boldsymbol{z}^{\mathfrak{mp}(j)} - \mathbf{Z}_i^{\mathfrak{mp}(j)})}{\sum_{i=1}^n K^j_{\mathbf{H}}(\boldsymbol{z}^{\mathfrak{mp}(j)} - \mathbf{Z}_i^{\mathfrak{mp}(j)})} & \text{otherwise,} \end{cases} \\
&\in \{0, 1\}.
\end{aligned}
$$

For $j = 1$, since $\mathfrak{mp}(1) = \emptyset$, we can directly show the assertion as

$$
\sum_{i_1=1}^n \hat{w}^1_{i_1} = \sum_{i_1=1}^n \frac{1}{n} = 1.
$$

For $j \ge 2$,

$$
\begin{aligned}
\sum_{i_1=1}^n \cdots \sum_{i_j=1}^n \hat{w}_{i_j | \boldsymbol{i}_{1:j-1}} \cdots \hat{w}^1_{i_1} &= \sum_{i_1=1}^n \cdots \sum_{i_{j-1}=1}^n \hat{w}_{i_{j-1} | \boldsymbol{i}_{1:j-2}} \cdots \hat{w}^1_{i_1} \left(\sum_{i_j=1}^n \hat{w}_{i_j | \boldsymbol{i}_{1:j-1}} \right) \\
&\in \left\{ 0, \left(\sum_{i_1=1}^n \cdots \sum_{i_{j-1}=1}^n \hat{w}_{i_{j-1} | \boldsymbol{i}_{1:j-2}} \cdots \hat{w}^1_{i_1} \right) \right\}.
\end{aligned}
$$

By recursively applying the above argument for a finite number of times, we obtain the assertion for all $j \in [D]$. □