Incorporating Causal Graphical Prior Knowledge into Predictive Modeling via Simple Data Augmentation (Supplementary Material)

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

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

Table 1 summarizes the abbreviations and the symbols used in the paper. For notation simplicity, when $\bar{\mathcal{Z}}^j$ is a finite set, we identify it with $\mathbb{Z}/m\mathbb{Z}$ where m is the cardinality of $\bar{\mathcal{Z}}^j$, to justify the subtractions inside the kernel functions.

A PRELIMINARIES ON ADMG

Given an ADMG $\mathcal{G}$ with the vertex set V and topological order $\leq$, we use the following terminologies (Bhattacharya et al., 2020).

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 $\mathcal{G}_{\leq v}$ 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 $\mathcal{G}_{\leq v}$. 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 (Yadan, 2019). All experiments were carried out on a 2.60 GHz Intel® Xeon® CPUs with 132 GB memory.

Our experiment code can be found at `https://github.com/takeshi-teshima/incorporating-causal-graphical-prior-knowledge-into-predictive-modeling-via-simple-data-augmentation`.

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 (Sachs et al., 2005). 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. (2005) based on the conventionally accepted cellular signaling networks (Figure 3(a)). Among the eight data sets corresponding to different intervention conditions (Sachs et al., 2005), 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*.

Table 1: 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., Zorich, 2015, Section E.4.1).
Δ_K	$(K-1)$-dimensional probability simplex (Boyd et al., 2004, 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 $\mathfrak{mp}(j) = \emptyset$).
$\mathbf{Z}_{\boldsymbol{i}}$	$\mathbf{Z}_{\boldsymbol{i}} = (Z_{i_1}^1, \ldots, Z_{i_D}^D)$ 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.

GSS data (Shimizu et al., 2011). This data set is concerning the status attainment theory in sociology. This data set is originally part of the General Social Survey (GSS)[1], and we used a subset of the data that was previously used in the causal discovery literature (Shimizu et al., 2011). The reference graph is based on domain knowledge of the status attainment model (Duncan et al., 1972; Figure 3(b)). 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,

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

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 (Harrison et al., 1978). 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 (Quinlan, 1993). 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 (Cortez et al., 2009). 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.

Red Wine data (Cortez et al., 2009). 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 (Chen et al., 2016). See Chen et al. (2016) 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 (Silverman, 1986, pp.45–47, Equations (3.28) and (3.30) therein) implemented in the *statsmodels* package of Python (*Statsmodels* 2020), namely, $h^{\text{thumb}} = \left(\frac{4}{3}\right)^{1/5} A n^{-1/5}$ where $A = \min\{\hat{\sigma}, \mathrm{IQR}/1.349\}$, $\hat{\sigma}$ is the square root of the unbiased estimator of the variance, and IQR is the interquantile range.

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

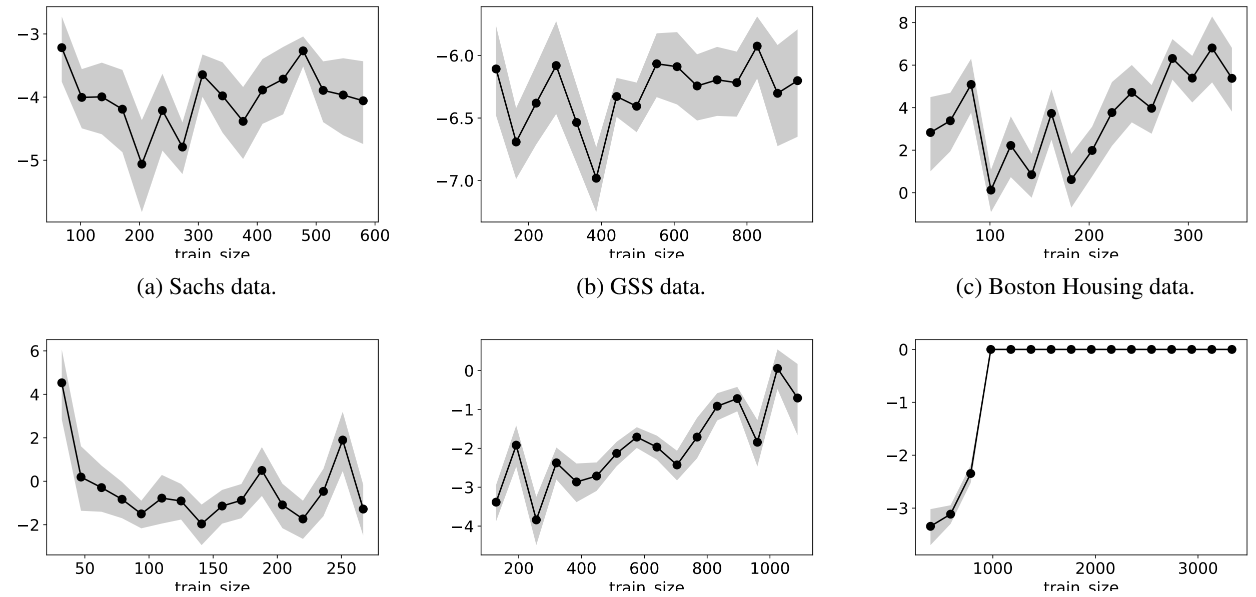
B.4 CAUSAL DISCOVERY METHOD CONFIGURATION

We perform *DirectLiNGAM* (Shimizu et al., 2011) 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 (Hyvärinen et al., 2013) that is based on a nonparametric approximation to the mutual information.

B.5 SUPPLEMENTARY EXPERIMENT RESULTS

Figure 1 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 THEORETICAL ANALYSIS



(a) Sachs data.

(b) GSS data.

(c) Boston Housing data.

(d) Auto MPG data.

(e) Red Wine data.

(f) White Wine data.

Figure 1: 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.

C.1 NOTATION AND PROBLEM SETUP

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 $(x_1, \ldots, x_k) \in \mathbb{R}^k$, $\mathrm{diag}((x_1, \ldots, x_k))$ 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 $\|\cdot\|_{\mathrm{op}}$ 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 $[n]^{j-1} = \{0\}$.

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 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 $\mathfrak{mp}(j) \subset [D]$ denote the Markov pillow of $j \in [D]$. Throughout this section, we assume p satisfies the topological ADMG factorization relation according to $\mathcal{G}$ (Bhattacharya et al., 2020):

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

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. To simplify the notation, we define $\ell_f := \ell(f, \cdot)$ and $\mathcal{L}_{\mathcal{F}} := \{\ell_f : f \in \mathcal{F}\}$. 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 $\boldsymbol{h} = (\boldsymbol{h}^1, \dots, \boldsymbol{h}^D) \in \mathbb{R}^D_{>0}$. Then, we define

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

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

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

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

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

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

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

assuming both exist.

Notation for stating the results. 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

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

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

C.2 MAIN THEOREM

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

C.2.1 Preliminaries

We use the following convenient *multi-index* notation (see, e.g., Stone, 1982).

Definition 1 (Multi-index notation). *For $d \in \mathbb{N}$, we call a d-tuple $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{Z}^d_{\geq 0}$* 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 $x = (x_1, \dots, x_d) \in \mathbb{R}^d$. Also, let ∂^α denote the partial differential operator defined by*

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

Definition 2 (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 3 (Hölder class; Stone, 1982; Tsybakov, 2009)**.** *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*

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

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 (2009).*

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

Definition 4 (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.2.2 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 $B_j := \int_{\mathcal{Z}^j} \mathrm{d}\boldsymbol{z}^j < \infty$.*
- *For all $j \in [D]$, $p_{\mathfrak{mp}(j)}$ is bounded away from zero over $\mathcal{Z}^{\mathfrak{mp}(j)}$. Define $\epsilon_{\mathfrak{mp}(j)} := \inf_{\boldsymbol{z}^{\mathfrak{mp}(j)} \in \mathcal{Z}^{\mathfrak{mp}(j)}} p_{\mathfrak{mp}(j)}(\boldsymbol{z}^{\mathfrak{mp}(j)})$.*
- *For each $j \in [D]$, K^j is continuous and strictly positive. We define*

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

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

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

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

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 $j \in [D]$:*

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

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

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

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

C.2.3 Statement and Proof

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

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

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

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

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 derives ideas from the literature on *local empirical processes* and *kernel-type estimators*, namely Einmahl et al. (2000), Einmahl et al. (2005), and Dony et al. (2006). 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,

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

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

$$\left| \frac{A}{B} - \frac{C}{D} \right| = \left| \frac{A}{B} - \frac{C}{B} + \frac{C}{B} - \frac{C}{D} \right| \leq \left| \frac{1}{B} \right| \cdot |A - C| + \left| \frac{C}{BD} \right| \cdot |B - D|. \tag{2}$$

Proof of Theorem 1. First, note

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

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

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

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

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

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

$$r^j(f, \boldsymbol{z}^{\mathfrak{mp}(j)}) := \int_{\mathcal{Z}^j} f(\boldsymbol{z}^j) p_{j,\mathfrak{mp}(j)}(\boldsymbol{z}^j, \boldsymbol{z}^{\mathfrak{mp}(j)}) \mathrm{d}\boldsymbol{z}^j, \qquad \hat{r}^j(f, \boldsymbol{z}^{\mathfrak{mp}(j)}) := \frac{1}{n} \sum_{i=1}^{n} f(Z_i^j) K_{\mathbf{H}}^j (\boldsymbol{z}^{\mathfrak{mp}(j)} - \mathbf{Z}_i^{\mathfrak{mp}(j)}),$$

$$g^j(\boldsymbol{z}^{\mathfrak{mp}(j)}) := p_{\mathfrak{mp}(j)}(\boldsymbol{z}^{\mathfrak{mp}(j)}), \qquad \hat{g}^j(\boldsymbol{z}^{\mathfrak{mp}(j)}) := \frac{1}{n} \sum_{i=1}^{n} K_{\mathbf{H}}^j (\boldsymbol{z}^{\mathfrak{mp}(j)} - \mathbf{Z}_i^{\mathfrak{mp}(j)}).$$

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

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

By applying the argument of Eq. (2), we can bound each ratio difference term as

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

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 $1 - 2D\delta$,

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

By reorganizing the terms, we obtain the assertion. □

C.2.4 Lemmas

Here, we prove the lemmas used in the proof of Theorem 1.

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

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

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

Proof. By Assumption 1, we have

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

Also,

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

where we used the positivity of the integrand. Now,

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

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

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

□

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

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

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

Proof. By applying Lemma 4 under Assumption 2,

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

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

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

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

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

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*

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

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

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

Proof. Note

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

and

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

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

C.2.5 Facts

Here, we state some facts used in the proof of Theorem 1. The following is Taylor's formula with the integral form of the remainder, stated using the multi-index notation.

Fact 1 (Taylor's theorem; Zorich, 2015, 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 (Mohri et al., 2018, Theorem 3.3).

Fact 3 (Rademacher complexity bound; Theorem 3.3 in Mohri et al., 2018)**.** *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 $\{z_i\}_{i=1}^m \overset{\text{i.i.d.}}{\sim} z$, 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.2.6 Basic Lemmas

Here, we prove the basic lemmas used in the proof of Theorem 1.

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

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

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 $\boldsymbol{z}^{\mathfrak{mp}(j)} \in \mathcal{Z}^{\mathfrak{mp}(j)}$,

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

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 $j \in [D]$. □

C.3 SUPPLEMENTARY THEORY: COMPARISON OF COMPLEXITY MEASURES

Here, we formally demonstrate the complexity reduction effect explained in Section 4. More concretely, as an example in which the effect can be demonstrated, we take the example represented by Assumption 3 where the Lipschitz continuity of the functions are assumed and compare the upper bounds on the complexity terms appearing in the generalization error bound of the usual empirical risk minimization (ERM) and those in Theorem 1 (namely $R_{\mathcal{F},\mathbf{K}}$ and $R_{\mathbf{K}}$).

The complexity reduction effect in this example is demonstrated by the different dependencies of the upper bounds on the sample size, both derived based on the metric-entropy method; the one corresponding to ERM yields a bound of order $O(n^{-1/(2+D)})$ whereas the one for the proposed method yields $O(n^{-1/3})$. Although the comparison between the two upper bounds only provides circumstantial evidence, we believe that the reduced exponent demonstrates the complexity reduction effect as they are derived based on the same proof technique.

First, recall that the proposed method enjoys Theorem 1 which states, for any $\delta \in (0, 1)$, we have with probability at least $1 - 2D\delta$,

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

On the other hand, the usual empirical risk minimization algorithm enjoys the following theoretical guarantee. Recall $\hat{R}_{\mathrm{emp}}(f) := \frac{1}{n}\sum_{i=1}^n \ell(f, \mathbf{Z}_i)$.

Proposition 1. *For any $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have that the solution to the usual empirical risk minimization*

$$\hat{f}_{\mathrm{emp}} \in \mathop{\arg\min}_{f \in \mathcal{F}} \{\hat{R}_{\mathrm{emp}}(f)\}$$

satisfies

$$R(\hat{f}_{\mathrm{emp}}) - R(f^*) \leq \underbrace{4\mathrm{Rad}_{n,p}(\mathcal{L}_{\mathcal{F}})}_{\textit{Complexity term}} + 2B_\ell \sqrt{\frac{\log(2/\delta)}{2n}}.$$

Proof. The assertion is immediate from Fact 3 and the following inequality:

$$\begin{aligned} R(\hat{f}_{\mathrm{emp}}) - R(f^*) &= R(\hat{f}_{\mathrm{emp}}) - \hat{R}_{\mathrm{emp}}(\hat{f}_{\mathrm{emp}}) + \hat{R}_{\mathrm{emp}}(\hat{f}_{\mathrm{emp}}) - R(f^*) \\ &\leq R(\hat{f}_{\mathrm{emp}}) - \hat{R}_{\mathrm{emp}}(\hat{f}_{\mathrm{emp}}) + \hat{R}_{\mathrm{emp}}(f^*) - R(f^*) \leq 2 \sup_{f \in \mathcal{F}} |R(f) - \hat{R}_{\mathrm{emp}}(f)|. \end{aligned}$$

□

From here, we compare the dependency of the complexity terms $\mathrm{Rad}_{n,p}(\mathcal{L}_{\mathcal{F}})$ and $R_{\mathcal{F},\mathbf{K}} + B_\ell R_{\mathbf{K}}$ on n. In addition to Assumptions 1 and 2, assume the following:

Assumption 3 (Complexity assumptions)**.** *We assume the following:*

- *The functions in $\mathcal{L}_{\mathcal{F}}$ are L_1-Lipschitz continuous.*
- *The functions K^j are $L_{K,j}$-Lipschitz continuous.*
- *The functions $p_{k|\mathfrak{mp}(k)}(\boldsymbol{z}^k|\cdot)$ are $L_{p,k}$-Lipschitz continuous for all $\boldsymbol{z}^k$.*

For simplicity, we also assume $\mathbf{H} = \mathrm{diag}((h, \ldots, h))$.

Under this assumption, we have the following:

Proposition 2 (Comparison of the complexity measures)**.** *Given Assumptions 1, 2, and 3, we have the following:*

$$\mathrm{Rad}_{n,p}(\mathcal{L}_{\mathcal{F}}) \leq O\left(n^{-\frac{1}{D+2}}\right), \qquad R_{\mathcal{F},\mathbf{K}} + B_\ell R_{\mathbf{K}} \leq O\left(n^{-1/3}\right).$$

Implications. Proposition 2 shows that the complexity terms appearing in Theorem 1 has a better dependency on the sample size compared to those in Proposition 1, demonstrating the complexity reduction effect in this example. Note here that we do not claim that the proposed method yields a rate-optimal predictor, but instead, we provide Theorem 1 and this supplementary analysis to obtain insights regarding how the proposed method may facilitate the learning.

Proof of Proposition 2. By the Lipschitz continuity of the functions in $\mathcal{L}_{\mathcal{F}}$ and the boundedness of $\mathcal{Z}$, we can apply Fact 6 to obtain

$$\log \mathcal{N}_{(t, \|\cdot\|_\infty)}(\mathcal{L}_{\mathcal{F}}) \leq C \left(\frac{L_1}{t} \right)^D$$

for a constant $C > 0$. By applying Fact 4, and minimizing the right-hand side for t, we have the first assertion.

On the other hand, by Lemma 6,

$$\log \mathcal{N}_{(t, \|\cdot\|_\infty)}(\mathcal{L}_{\mathcal{F}}^j \otimes \mathcal{K}_{\mathbf{H}}^j) \leq \log \mathcal{N}_{(t_1, \|\cdot\|_\infty)}(\mathcal{L}_{\mathcal{F}}^j) + \log \mathcal{N}_{(t_2, \|\cdot\|_\infty)}(\mathcal{K}_{\mathbf{H}}^j),$$

where t_1, t_2 are such that $B_{\mathbf{K}} t_1 + B_\ell t_2 = t$. Now, applying Lemma 8,

$$\log \mathcal{N}_{(t_1, \|\cdot\|_\infty)}(\mathcal{L}_{\mathcal{F}}^j) \leq \log \sup_{z \in \mathcal{Z}^{j-1}} \mathcal{N}_{(t_{1,1}, \|\cdot\|_\infty)}(\mathcal{F}_z) + \log \mathcal{N}_{(t_{1,2}, \|\cdot\|)}(\mathcal{B}^{j-1}(R_{\mathcal{Z}}))$$

By combining Lemma 7 and Lemma 9, and applying Fact 5, we have

$$\log \sup_{z \in \mathcal{Z}^{j-1}} \mathcal{N}_{(t_{1,1}, \|\cdot\|_\infty)}(\mathcal{F}_z) \leq C \frac{L_2}{t_{1,1}}, \quad \log \mathcal{N}_{(t_{1,2}, \|\cdot\|)}(\mathcal{B}^{j-1}(R_{\mathcal{Z}})) \leq (j-1) \log \left(1 + \frac{2R_{\mathcal{Z}}}{t_{1,2}} \right),$$

where $t_{1,1}, t_{1,2}$ are such that $t_1 = t_{1,1} + L_2 t_{1,2}$, and $L_2 = L_1 + B_\ell \sum_k L_{p,k}$.

On the other hand, by Lemma 10, we have

$$\log \mathcal{N}_{(t_2, \|\cdot\|_\infty)}(\mathcal{K}_{\mathbf{H}}^j) \leq |\mathfrak{mp}(j)| \log \left(1 + \frac{2L_{K,H,j} R_{\mathcal{Z}}}{t_2} \right).$$

where $L_{K,H,j} = h^{-|\mathfrak{mp}(j)|-1} L_{K,j}$.

Therefore, we have

$$\log \mathcal{N}_{(t, \|\cdot\|_\infty)}(\mathcal{L}_{\mathcal{F}}^j \otimes \mathcal{K}_{\mathbf{H}}^j) \leq C \frac{L_2}{t_{1,1}} + (j-1) \log \left(1 + \frac{2R_{\mathcal{Z}}}{t_{1,2}} \right) + |\mathfrak{mp}(j)| \log \left(1 + \frac{2L_{K,H,j} R_{\mathcal{Z}}}{t_2} \right).$$

By applying Fact 4, letting

$$t_{1,1} = \frac{t}{3B_{\mathbf{K}}}, \; t_{1,2} = \frac{t}{3B_{\mathbf{K}} L_2}, \; t_2 = \frac{t}{3B_\ell},$$

and minimizing the upper bound for t, we have

$$\left| \det \mathbf{H}_j \right| \mathrm{Rad}_{n,p} \left(\mathcal{L}_{\mathcal{F}}^j \otimes \mathcal{K}_{\mathbf{H}}^j \right) \leq O\left(n^{-1/3} \right).$$

Therefore, we have

$$R_{\mathcal{F}, \mathbf{K}} = \sum_{j=1}^{D} \left| \det \mathbf{H}_j \right| \mathrm{Rad}_{n,p} \left(\mathcal{L}_{\mathcal{F}}^j \otimes \mathcal{K}_{\mathbf{H}}^j \right) \leq O\left(n^{-1/3} \right),$$

$$R_{\mathbf{K}} = \sum_{j=1}^{D} \left| \det \mathbf{H}_j \right| \mathrm{Rad}_{n,p}(\mathcal{K}_{\mathbf{H}}^j) \leq O\left(n^{-1/2} \right),$$

and obtain the second assertion. □

C.3.1 Lemmas and Facts

Lemma 6 (Metric entropy of products)**.** *Let $\mathcal{F}, \mathcal{G}$ be two classes of bounded measurable functions satisfying $\|f\|_\infty \le M_\mathcal{F} (f \in \mathcal{F})$ and $\|g\|_\infty \le M_\mathcal{G} (g \in \mathcal{G})$. Then, we have for any $t_1, t_2 > 0$,*

$$\log \mathcal{N}_{(t,\|\cdot\|_\infty)}(\mathcal{F} \otimes \mathcal{G}) \le \log \mathcal{N}_{(t_1,\|\cdot\|_\infty)}(\mathcal{F}) + \log \mathcal{N}_{(t_2,\|\cdot\|_\infty)}(\mathcal{G})$$

where $t = M_\mathcal{G} t_1 + M_\mathcal{F} t_2$.

Proof. Let $\{f_i\}_i$ ($\{g_j\}_j$) be the t_1- (resp. t_2-)covering of $\mathcal{F}$ (resp. $\mathcal{G}$). Then, for any $f \in \mathcal{F}$ and $g \in \mathcal{G}$, we have for some i, j that

$$\begin{aligned}
\|f \otimes g - f_i \otimes g_j\|_\infty &\le \|f \otimes g - f_i \otimes g\|_\infty + \|f_i \otimes g - f_i \otimes g_j\|_\infty \\
&\le \|f - f_i\|_\infty M_\mathcal{G} + M_\mathcal{F} \|g - g_j\|_\infty \\
&\le M_\mathcal{G} t_1 + M_\mathcal{F} t_2.
\end{aligned}$$

This implies the assertion. □

Lemma 7 (Lipschitz continuity of marginalized function class)**.** *Assume that $p_{k|\mathrm{mp}(k)}(\boldsymbol{z}^k|\cdot)$ is $L_{p,k}$-Lipschitz continuous for all $\boldsymbol{z}^k$. Then, the elements of $\bar{\mathcal{L}}^j_\mathcal{F}$ are Lipschitz continuous with the constant $L_1 + B_\ell \sum_k L_{p,k}$.*

Proof. Since the functions in $\mathcal{L}_\mathcal{F}$ are L_1-Lipschitz continuous, the elements of $\bar{\mathcal{L}}^j_\mathcal{F}$ are also Lipschitz continuous:

$$\begin{aligned}
|\ell_{f,j}(x) - \ell_{f,j}(y)| &= \left| \int \ell_f((x,z)) \prod_k p_{k|\mathrm{mp}(k)}(\boldsymbol{z}^k|(x,z)^{\mathrm{mp}(k)}) \mathrm{d}\boldsymbol{z} - \int \ell_f((y,z)) \prod_k p_{k|\mathrm{mp}(k)}(\boldsymbol{z}^k|(y,z)^{\mathrm{mp}(k)}) \mathrm{d}\boldsymbol{z} \right| \\
&\le \int |\ell_f((x,z)) - \ell_f((y,z))| \prod_k p_{k|\mathrm{mp}(k)}(\boldsymbol{z}^k|(y,z)) \mathrm{d}\boldsymbol{z} \\
&\quad + \sum_{k \ge j+1} \int |\ell_f((x,z))| p_{j+1|\mathrm{mp}(j+1)}(\boldsymbol{z}^{j+1}|(x,z)) \cdots (p_{k|\mathrm{mp}(k)}(\boldsymbol{z}^k|(x,z)) - p_{k|\mathrm{mp}(k)}(\boldsymbol{z}^{\mathrm{l}}(y,z))) \cdots p_{D|\mathrm{mp}(D)}(\boldsymbol{z}^D|(y,z)) \mathrm{d}\boldsymbol{z} \\
&\le L_1 \|x - y\| \cdot 1 + B_\ell \sum_k 1 \cdot L_{p,k} \|x - y\| \cdot 1 \\
&\le (L_1 + B_\ell \sum_k L_{p,k}) \|x - y\|.
\end{aligned}$$

□

Lemma 8 (Lipschitz continuity of curried function class)**.** *Let $j \in [2 : D]$ and $R_\mathcal{Z} = \sup_{z \in \mathcal{Z}} \|z\|$. Also let $\mathcal{B}^{j-1}(R)$ denote the radius-R ball in the $(j-1)$-dimensional Euclidean space, and define $\mathcal{F}_z := \{\ell_{f,j}(z,\cdot) : \ell_{f,j} \in \bar{\mathcal{L}}^j_\mathcal{F}\}$ for $z \in \mathcal{Z}^{j-1}$. Assume $\bar{\mathcal{L}}^j_\mathcal{F}$ consist of L_2-Lipschitz continuous functions. Then, we have*

$$\log \mathcal{N}_{(t,\|\cdot\|_\infty)}(\mathcal{L}^j_\mathcal{F}) \le \log \sup_{z \in \mathcal{Z}^{j-1}} \mathcal{N}_{(u,\|\cdot\|_\infty)}(\mathcal{F}_z) + \log \mathcal{N}_{(v,\|\cdot\|)}(\mathcal{B}^{j-1}(R_\mathcal{Z}))$$

where $t, u, v > 0$ are such that $t = u + L_2 v$.

Proof. Let $\{z_\mu\}_\mu \subset \mathcal{Z}^{j-1}$ be a v-covering of $\mathcal{Z}^{j-1}$. For each z_μ, consider the set $\mathcal{F}_\mu = \{\ell_{f,j}(z_\mu,\cdot) : \ell_{f,j} \in \bar{\mathcal{L}}^j_\mathcal{F}\}$. Let $\{\ell^{\mu,k}_{f,j}\}_k \subset \mathcal{F}_\mu$ be a u-covering of $\mathcal{F}_\mu$. Then, for any $\ell_{f,j} \in \bar{\mathcal{L}}^j_\mathcal{F}$ and $z \in \mathcal{Z}^{j-1}$, there exists z_μ such that $\|z_\mu - z\| \le v$. Moreover, since we have $\ell_{f,j}(z_\mu,\cdot) \in \mathcal{F}_\mu$, there exists $\ell^{\mu,k}_{f,j}$ such that $\|\ell_{f,j}(z_\mu,\cdot) - \ell^{\mu,k}_{f,j}(z_\mu,\cdot)\|_\infty \le u$. For such a pair $(z_\mu, \ell^{\mu,k}_{f,j})$, we have

$$\|\ell_{f,j}(z,\cdot) - \ell^{\mu,k}_{f,j}(z_\mu,\cdot)\|_\infty \le \|\ell_{f,j}(z,\cdot) - \ell_{f,j}(z_\mu,\cdot)\|_\infty + \|\ell_{f,j}(z_\mu,\cdot) - \ell^{\mu,k}_{f,j}(z_\mu,\cdot)\|_\infty \le L_2 v + u$$

Therefore, the set $\bigcup_\mu \{z_\mu\}_\mu \times \{\ell^{\mu,k}_{f,j}\}_k$ induces a $(L_2 v + u)$-covering of $\mathcal{L}^j_\mathcal{F}$. Noting that the cardinality of $\bigcup_\mu \{z_\mu\}_\mu$ is bounded by $\mathcal{N}_{(v,\|\cdot\|)}(\mathcal{B}^{j-1}(R_\mathcal{Z}))$ and that of $\{\ell^{\mu,k}_{f,j}\}_k$ by $\sup_{z \in \mathcal{Z}^{j-1}} \mathcal{N}_{(u,\|\cdot\|_\infty)}(\mathcal{F}_z)$, we have the assertion. □

Lemma 9 (Metric entropy of functions curried by a specific input)**.** *Assume that the elements of $\bar{\mathcal{L}}^j_\mathcal{F}$ are L_2-Lipschitz continuous. Then, there exists a constant $C > 0$ such that for sufficiently small $u > 0$,*

$$\sup_{z \in \mathcal{Z}^{j-1}} \mathcal{N}_{(u,\|\cdot\|_\infty)}(\mathcal{F}_z) \le C \frac{L_2}{u}.$$

Proof. Since the elements of $\bar{\mathcal{L}}^{j}_{\mathcal{F}}$ are L_2-Lipschitz continuous, so are the elements of $\mathcal{F}_z$ with Lipschitz constant L_2. Indeed, for any $x, y \in \mathcal{Z}^j$ and $z \in \mathcal{Z}^{j-1}$, we have

$$|\ell_{f,j}(z,x) - \ell_{f,j}(z,y)| \le L_2 \left\| \begin{pmatrix} z \\ x \end{pmatrix} - \begin{pmatrix} z \\ y \end{pmatrix} \right\| = L_2 \|x - y\|.$$

Therefore, by applying Lemma 6, we have the assertion. □

Lemma 10 (Shifted kernel complexity)**.** *Assume that $K^j : \mathbb{R}^{|\mathfrak{mp}(j)|} \to \mathbb{R}$ is $L_{K,j}$-Lipschitz continuous. Let $L_{K,H,j} = \frac{1}{|\det \mathbf{H}_j|} L_{K,j} \left\|\mathbf{H}_j^{-1}\right\|_{\mathrm{op}}$. Then, we have the following:*

$$\log \mathcal{N}_{(t_2, \|\cdot\|_\infty)}(\mathcal{K}^j_{\mathbf{H}}) \le |\mathfrak{mp}(j)| \log\left(1 + \frac{2 L_{K,H,j} R_{\mathcal{Z}}}{t_2}\right).$$

Proof. Recalling $K^j_{\mathbf{H}}(u) = \frac{1}{|\det \mathbf{H}_j|} K^j(\mathbf{H}_j^{-1} u)$, for any $K^j_{\mathbf{H}}(z_1 - \cdot), K^j_{\mathbf{H}}(z_2 - \cdot) \in \mathcal{K}^j_{\mathbf{H}}$, we have

$$\begin{aligned} \|K^j_{\mathbf{H}}(z_1 - \cdot) - K^j_{\mathbf{H}}(z_2 - \cdot)\|_\infty &\le \frac{1}{|\det \mathbf{H}_j|} L_{K,j} \|\mathbf{H}_j^{-1}(z_1 - z_2)\| \\ &\le \frac{1}{|\det \mathbf{H}_j|} L_{K,j} \left\|\mathbf{H}_j^{-1}\right\|_{\mathrm{op}} \|z_1 - z_2\| \end{aligned}$$

Therefore, we have

$$\log \mathcal{N}_{(t_2, \|\cdot\|_\infty)}(\mathcal{K}^j_{\mathbf{H}}) \le \log \mathcal{N}_{(t_2/L_{K,H,j}, \|\cdot\|)}(\mathbf{Z}^{\mathfrak{mp}(j)}).$$

Applying Fact 5, we obtain the assertion. □

Fact 4 (One-step discretization bound)**.** *Let $\mathcal{F}$ be a class of measurable functions. There exist constants c and B such that for any $t \in (0, B]$, the following relation between the Rademacher complexity and the metric entropy holds:*

$$\mathrm{Rad}_{m,q}(\mathcal{F}) \le t + c \sqrt{\frac{\log \mathcal{N}_{(t, \|\cdot\|_\infty)}(\mathcal{F})}{m}}$$

Fact 5 (Euclidean ball metric entropy bound; Wainwright, 2019, Example 5.8, p.126)**.** *Let $R > 0$ and $d \in \mathbb{N}$. Let $\mathcal{B}(R)$ denote the radius-R ball in the d-dimensional Euclidean space. Then, we have the following metric entropy bound:*

$$\log \mathcal{N}_{(\delta, \|\cdot\|)}(\mathcal{B}(R)) \le d \log\left(1 + \frac{2R}{\delta}\right).$$

Fact 6 (Lipschitz functions metric entropy bound; Wainwright, 2019, Example 5.10, p.129)**.** *Let $L, R > 0$ and $d \in \mathbb{N}$. Let $\mathrm{Lip}(R, L)$ denote the set of L-Lipschitz functions on $[0, R]^d$. Then, we have the following metric entropy bound for sufficiently small $\delta > 0$:*

$$\log \mathcal{N}_{(\delta, \|\cdot\|_\infty)}(\mathrm{Lip}(R, L)) \le C \left(\frac{LR}{\delta}\right)^d,$$

where $C > 0$ is a constant.

D COMPUTATIONAL COMPLEXITY OF ALGORITHM 1

Here, we remark why the worst-case computational complexity of Algorithm 1 is $O\left(n^D\right)$. The main computation cost of Algorithm 1 comes from the computation of the weights $\hat{w}_{i_j|\boldsymbol{i}_{1:j-1}}$. There are n^{j-1} nodes at depth j (Fig. 2), each with n weighted edges connected to depth $j + 1$. The set of weights corresponding to each node, $\{\hat{w}_{i_j|\boldsymbol{i}_{1:j-1}}\}_{i_j \in [n]}$, is computed by constructing a matrix of shape $n \times n^{j-1}$ each of whose element is the kernel value for two vectors of dimensionality $|\mathfrak{mp}(j)| (\le j - 1)$. In the case of Gaussian kernels, each kernel value requires $O(j - 1)$ operations to compute. Subsequently, the kernel matrix is normalized by the column sum, which requires $O(n)$ summations and n^j divisions. The same computation takes place for each of the $\boldsymbol{i}_{1:j-1} \in [n]^{j-1}$ nodes at depth j, therefore, the edge weights between depth j and depth $j + 1$ can be computed by $O\left(n^j\right)$ operations. The edge weights are multiplied to obtain the node weights, which requires $O\left(n^D\right)$ multiplications since the number of multiplications that take place is equal to the number of edges in Fig. 2. Overall, Algorithm 1 requires $O\left(n^D\right)$ operations for the edge weight computation and $O\left(n^D\right)$ for the node weight computation, amounting to $O\left(n^D\right)$ operations in total, in the worst case that no edge is pruned by the threshold θ.

REFERENCES

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

Boyd, S. et al. (2004). *Convex Optimization*. Cambridge University Press.

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

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

Dony, J. et al. (2006). “Uniform in Bandwidth Consistency of Local Polynomial Regression Function Estimators.” In: *Austrian Journal of Statistics* 35.2, p. 16.

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

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

– (2005). “Uniform in Bandwidth Consistency of Kernel-Type Function Estimators.” In: *Annals of Statistics* 33.3, pp. 1380–1403.

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

Hyvärinen, A. et al. (2013). “Pairwise Likelihood Ratios for Estimation of Non-Gaussian Structural Equation Models.” In: *Journal of Machine Learning Research* 14.Jan, pp. 111–152.

Mohri, M. et al. (2018). *Foundations of Machine Learning*. Second. Cambridge, Massachusetts: The MIT Press.

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

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

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

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

Statsmodels (2020). statsmodels.

Stone, C. J. (1982). “Optimal Global Rates of Convergence for Nonparametric Regression.” In: *The Annals of Statistics* 10.4, pp. 1040–1053.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. New York ; London: Springer.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. 1st. Cambridge University Press.

Yadan, O. (2019). *Hydra - A Framework for Elegantly Configuring Complex Applications*. Github.

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