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

Takeshi Teshima$^2$  
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 $\mathbb{Z}^j$ is a finite set, we identify it with $\mathbb{Z}/m\mathbb{Z}$ where $m$ is the cardinality of $\mathbb{Z}^j$, to justify the subtractions inside the kernel functions.

A PRELIMINARIES ON ADMG

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

District. For $v \in V$, define $\mathcal{b}\mathcal{i}\mathcal{s}(v)$ 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 $\mathcal{p}a(A) := \bigcup_{v \in A} \mathcal{p}a(v) \setminus A$ where $\mathcal{p}a(v)$ denotes the parent of $v$ in the usual sense.

Markov pillow. For $v \in V$, define $G_{\preceq v}$ to be the subgraph of $G$ that is composed of only the vertices that precede $v$. Then, the Markov pillow of $v \in V$ is $\mathcal{m}\mathcal{p}(v) := \mathcal{b}\mathcal{i}\mathcal{s}(v) \cup \mathcal{p}a(\mathcal{b}\mathcal{i}\mathcal{s}(v)) \setminus \{v\}$ in $G_{\preceq v}$. Throughout the paper, we use the fact that $\mathcal{m}\mathcal{p}(v)$ 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.


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 $\text{Raf}$, $\text{Mek}$, $\text{Plcg}$, $\text{PIP2}$, $\text{PIP3}$, $\text{Erk}$, $\text{Akt}$, $\text{PKA}$, $\text{PKC}$, $\text{P38}$, and $\text{Jnk}$. Among these, for demonstration purposes, we considered $\text{PKA}$ as the target attribute. As preprocessing, we log-transformed $\text{Raf}$, $\text{Mek}$, and $\text{PKA}$.
### Table 1: Abbreviations and Symbols in the Paper.

<table>
<thead>
<tr>
<th>ABBREVIATION / SYMBOL</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>CG/CGM</td>
<td>Causal Graph / Causal Graphical Model</td>
</tr>
<tr>
<td>ADMG</td>
<td>Acyclic Directed Mixed Graph</td>
</tr>
<tr>
<td>DAG/PAG</td>
<td>Directed Acyclic Graph / Partial Ancestral Graph</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean Squared Error</td>
</tr>
<tr>
<td>( \mathbb{R}, \mathbb{R}<em>{\geq 0}, \mathbb{Z}, \mathbb{Z}</em>{\leq 0}, \mathbb{N} )</td>
<td>Set of all real numbers, nonnegative real numbers, positive real numbers, integers, nonnegative integers, and positive integers.</td>
</tr>
<tr>
<td>( I[A] )</td>
<td>Indicator function, i.e., 1 if ( A ) holds true and 0 otherwise.</td>
</tr>
<tr>
<td>( X \perp Y \mid Z )</td>
<td>( X ) and ( Y ) are conditionally independent given ( Z ).</td>
</tr>
<tr>
<td>( ] )</td>
<td>Disjoint union of sets.</td>
</tr>
<tr>
<td>( \text{diag}(x_1,\ldots,x_d) )</td>
<td>Diagonal matrix with diagonal elements ( (x_1,\ldots,x_d) ) ((d \in \mathbb{N})).</td>
</tr>
<tr>
<td>( | \cdot |, | \cdot |<em>{\text{op}}, | \cdot |</em>{\text{F}}, | \cdot |<em>{\text{S}}, | \cdot |</em>{\infty} )</td>
<td>( | \cdot | ) denotes the Euclidean norm of a vector, the operator norm of a matrix, the supremum norm of a function, and the determinant of a matrix.</td>
</tr>
<tr>
<td>( [\cdot]_z )</td>
<td>( [a] := \max{z \in \mathbb{R} : z \leq a} ) for ( a \in \mathbb{R} ).</td>
</tr>
<tr>
<td>( \Delta_k )</td>
<td>((K-1))-dimensional probability simplex (Boyd et al., 2004, Example 2.5).</td>
</tr>
<tr>
<td>([N : M], [N] )</td>
<td>([N : M] := {N, N+1, \ldots, M}) and ([N] := {1 : N}), where ( N, M \in \mathbb{N} ) and ( N \leq M ).</td>
</tr>
<tr>
<td>( x^S )</td>
<td>( x^S := (x^{s_1},\ldots,x^{s_S}) ) where ( x = (x^{1},\ldots,x^{D}) ) is an ( n )-dimensional vector and ( S = {s_1,\ldots,s_S} \subseteq \mathbb{N} ) with ( s_1 &lt; \ldots &lt; s_S ).</td>
</tr>
<tr>
<td>([0] = 0, \mathbb{R}^0 := {0}, x^0 = 0, [N]^0 := {0} )</td>
<td>Conventions used in the paper.</td>
</tr>
<tr>
<td>( D \in \mathbb{R}^n )</td>
<td>Overall data dimensionality (with ( X ) and ( Y ) combined).</td>
</tr>
<tr>
<td>( Z = x_{j=1}^D Z_j )</td>
<td>Overall data space (without distinguishing ( X ) and ( Y )).</td>
</tr>
<tr>
<td>( X = x_{j \in [D]\setminus{j}} Z_j ), ( Y = Z_j^\perp )</td>
<td>Input variable space and target variable space.</td>
</tr>
<tr>
<td>( p )</td>
<td>Joint probability density of ( Z := (Z^1,\ldots,Z^D) ) taking values in ( Z ).</td>
</tr>
<tr>
<td>( \text{Rad}_{md} )</td>
<td>Rademacher complexity of a function class.</td>
</tr>
<tr>
<td>( \mathcal{F} \subset \mathcal{Y} )</td>
<td>Hypothesis set.</td>
</tr>
<tr>
<td>( \ell : \mathcal{F} \times (\mathcal{X}_{p}^n,\mathcal{Z}) \rightarrow \mathbb{R} )</td>
<td>Loss function.</td>
</tr>
<tr>
<td>( R(f) = \mathbb{E}[\ell(f, Z)] )</td>
<td>Risk functional for ( f \in \mathcal{F} ).</td>
</tr>
<tr>
<td>( \mathcal{D} = [Z_i]_{i=1}^n )</td>
<td>Independently and identically distributed sample from ( p ).</td>
</tr>
<tr>
<td>( \mathcal{G} = ({D}, \mathcal{E}, \mathcal{B}) )</td>
<td>Underlying ADMG for which ( p ) satisfies the topological ADMG factorization and its estimator.</td>
</tr>
<tr>
<td>( \text{dis}(\cdot), \text{pa}(\cdot), \text{mp}(\cdot) )</td>
<td>District, parents, and Markov pillow of vertex ( j \in {D} ).</td>
</tr>
<tr>
<td>( P_{\text{pm}(j)}, P_{j, \text{mp}(j)}, P_{\text{mp}(j)} )</td>
<td>Conditional density of ( Z_j ) given ( Z^{\text{mp}(j)} ), the joint density of ( (Z^j, Z^{\text{mp}(j)}) ), and the marginal density of ( Z^{\text{mp}(j)} ).</td>
</tr>
<tr>
<td>( K_j : Z^{\text{mp}(j)} \rightarrow \mathbb{R} )</td>
<td>Kernel function (we define ( K_j := 1 ) if ( \text{mp}(j) = \emptyset )).</td>
</tr>
<tr>
<td>( Z_i = (Z_i^1,\ldots,Z_i^D) ) for ( i = (i_1,\ldots,i_D) \in [n]^D ).</td>
<td>Augmented data set and the instance weights.</td>
</tr>
<tr>
<td>( \mathcal{D}<em>{\text{emp}}, \mathcal{D}</em>{\text{aug}} )</td>
<td>Ordinary empirical risk estimator and the proposed risk estimator.</td>
</tr>
<tr>
<td>( \Omega(f) )</td>
<td>Regularization term for ( f \in \mathcal{F} ).</td>
</tr>
<tr>
<td>( \lambda )</td>
<td>Convex combination coefficient used in ( (1 - \lambda) \hat{R}<em>{\text{emp}}(f) + \lambda \hat{R}</em>{\text{aug}}(f) + \Omega(f) ).</td>
</tr>
<tr>
<td>( K_f^{c} )</td>
<td>Component of the product kernel ( K_f ) for ( f \in \text{mp}(j) ).</td>
</tr>
<tr>
<td>( \theta )</td>
<td>Pruning threshold of the small weights in Algorithm 1 [ \text{(1d)} ].</td>
</tr>
</tbody>
</table>

**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, \( x_4 \): mother’s education, and \( x_5 \): father’s age, 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, \( x_4 \): mother’s education, and \( x_5 \): father’s age.

[^1]: [https://gss.norc.org/](https://gss.norc.org/)
We perform causal discovery with the 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) therein) implemented in the statsmodels (2020) package of Python. For the pruning threshold, we use \( \theta = 10^{-3} \cdot n^{-1} \).
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., \( \frac{\text{MSE}_{\text{prop}} - \text{MSE}_{\text{base}}}{\text{MSE}_{\text{base}}} \times 100 \% \) where MSE_{base} and MSE_{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 \( \mathbb{Z}_{\leq 0} \) that of non-negative integers. For \( (x_1, \ldots, x_k) \in \mathbb{R}^k \), 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, \( | \cdot | \) 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 \( x = (x_1, \ldots, x_n) \) and \( S \subset [n] \), we let \( x^S = (x_{s_1}, \ldots, x_{s_j}) \) denote its sub-vector with indices in \( S = \{s_1, \ldots, s_j\} \) with \( s_1 < \cdots < s_j \). Similarly, for \( j \in [n] \), we let \( x^j := x^\{j\} \). For \( S \subset [n] \), we also define \( Z^S := x_{s \in S} Z^j \). To simplify the notation, we use the convention of \( \mathbb{R}^0 := \{0\} \), \( x^0 = 0 \), and \( [n]^{i-1} = \{0\} \).

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

\[
    p(z) = \prod_{j=1}^D p_{j, mp(j)}(z^j|z^{mp(j)}) = \prod_{j=1}^D p_{j, \text{model}}(z^j, z^{mp(j)}) p_{j, \text{prior}}(z^{mp(j)})
\]

Learning problem. Let \( \mathcal{F} \) denote a hypothesis class, and let \( \ell : \mathcal{F} \times \mathbb{R}^D \to \mathbb{R}_{\geq 0} \) be a loss function. To simplify the notation, we define \( \ell_f := \ell(f, \cdot) \) and \( \ell_F := \{\ell_f : f \in \mathcal{F}\} \). For each \( f \in \mathcal{F} \), we define the risk functional \( R(f) := \mathbb{E}[\ell_f(Z)] \). The learning problem is to find a hypothesis \( \hat{f} \in \mathcal{F} \) for which \( R \) is small, given the training data \( D \) and the graph \( \mathcal{G} \).
Proposed method. For each $j \in [D]$, we fix a kernel function $K_j : \mathbb{R}^{[m(h_j)]} \to \mathbb{R}$. For notation simplicity, we define $K_j := 1$ for $j$ such that $m(p(j)) = 0$. We also fix $h = (h^1, \ldots, h^D) \in \mathbb{R}^{D_0}$. Then, we define

$$H_j := \text{diag}(h_m(j)), \quad K_j(u) := \frac{1}{|\det H_j|} K_j(H_j^{-1} u).$$

For $i = (i_1, \ldots, i_D)$ and $z^{m(p)} \in \mathbb{R}^{[m(p)]}$, define

$$\hat{w}_i(z^{m(p)}) := \frac{K_i(z^{m(p)} - Z_i^{m(p)})}{\sum_{i=1}^n K_i(z^{m(p)} - Z_i^{m(p)})} \left[ \sum_{i=1}^n K_i(z^{m(p)} - Z_i^{m(p)}) \neq 0 \right]$$

where $i = (i_1, \ldots, i_D)$, $z^{m(p)} \in \mathbb{R}^{[m(p)]}$. Then, we recursively define

$$\hat{w}_{i_0} = 1, \quad \hat{w}_{i_1} = \hat{w}_{i_0}(i_{1,j-1}) : \hat{w}_{i_1,j-1} (j \in [D], i_{1,j-1} \in [n])^{j-1},$$

where

$$\hat{w}_{i_0}(i_{1,j-1}) := \hat{w}_i(Z_i^{m(p)}), \quad Z_{i,j-1} = (Z_{i_1}, \ldots, Z_{i_{j-1}}).$$

Here, we use the convention $Z_{i,0} := 0$ to be consistent with the notation. Using this notation, for $f \in \mathcal{F}$, define the augmented empirical risk estimator

$$\hat{R}_\text{aug}(f) := \sum_{i \in [n]} \hat{w}_i \ell_f(z_i).$$

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

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

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

$$\ell_{f,j} : \left( z^1, \ldots, z^{j-1}, z^j \right) \mapsto \int_{z^{[j+1:D]}} \ell_f(z) \left( \prod_{j'=j+1}^D p_{(m(p))}(z^j | z^{[j']} \right) dz^j \cdots dz^D.$$ 

Also define

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

$$\mathcal{K}_H := \left\{ K_i(z^{m(p)} - (\cdot)) : z^{m(p)} \in \mathbb{Z}^{[m(p)]} \right\}.$$ 

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 $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{Z}_d$ multi-index. For a multi-index $\alpha$, 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 $\partial^{\alpha}$ denote the partial differential operator defined by

$$\partial^{\alpha} = \frac{\partial |\alpha|}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}.$$
When \( \Omega = \mathbb{R}^d \), we define the following class of functions.

**Assumption 2** (Smoothness assumptions)

\( \beta > 1 \) and \( L > 0 \), and let \( \Omega \subset \mathbb{R}^d \) be an open subset. The \((\beta, L)\)-Hölder class \( \Sigma(\beta, L, \Omega) \) is defined as the set of \( k = [\beta] \)-times continuously differentiable functions \( f : \Omega \to \mathbb{R} \) satisfying

\[
|\partial^a f(x) - \partial^a f(x')| \leq L||x - x'||^{|a| - k} \quad \text{for} \quad x, x' \in \Omega \quad \text{and} \quad |a| = k,
\]

where \( \alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{Z}^d \) is a multi-index, and \( |\alpha| = \max\{z \in \mathbb{Z} : z \leq a\} \) for \( a \in \mathbb{R} \). When \( \Omega = \mathbb{R}^d \), we also drop \( \Omega \) 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 \( \mathcal{F} \subset \mathbb{R}^X \), define

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

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

### 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., \( B_f := \sup_{f \in \mathcal{F}} \sup_{z \in \mathcal{Z}} |f(f, z)| < \infty \).
- \( K := \{K_j\}_{j=1}^D \) are uniformly bounded from above, i.e., \( B_K := \max\{||K_j||_{\infty} : j \in [D]\} < \infty \).
- For each \( j \in [D] \), \( \mathcal{Z}_j \subset \mathbb{R} \) is a compact subset. Let \( B_j := \int_{\mathcal{Z}_j} dz < \infty \).
- For all \( j \in [D] \), \( p_{\text{w/o}(j)} \) is bounded away from zero over \( \mathcal{Z}_{\text{w/o}(j)} \). Define \( \epsilon_{\text{w/o}(j)} := \inf_{z \in \mathcal{Z}_{\text{w/o}(j)}} p_{\text{w/o}(j)}(z) \).
- For each \( j \in [D] \), \( K_j \) is continuous and strictly positive. We define

\[
\phi_{K_j}(H_j) := \sup_{z \in \mathcal{Z}_{\text{w/o}(j)} \cap H_j} \left| K_j'(z) - z \right| = \sup_{z \in \mathcal{Z}_{\text{w/o}(j)} \cap H_j} \left| K_j'(z) - z \right| \|H_j^{-1}\|^{-1}
\]

and assume \( \phi_{K_j}(H_j) < \infty \).

**Remark 2.** Since \( \mathcal{Z}_{\text{w/o}(j)} \) is compact and \( K_j \) is continuous, if we define

\[
\epsilon_{K_j}(H_j) := |\det H_j| \left( \inf_{x, x' \in \mathcal{Z}_{\text{w/o}(j)}} K_j'(x) - x' \right) = \inf_{x, x' \in \mathcal{Z}_{\text{w/o}(j)}} K_j'(x) - x',
\]

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 \( j \in [D] \):
• $p_{n_0(j)}$ has an extension $\tilde{p}_{n_0(j)} \in \Sigma(\beta, L)$ such that $\tilde{I}_{n_0(j)} := \int_{\mathbb{R}^{\dim(j)} \times \mathbb{R}^{\dim(j)}} |\tilde{p}_{n_0(j)}(z^{\dim(j)}, \tilde{z}^{\dim(j)})|d\tilde{z}^{\dim(j)} < \infty$.

• For all $\mathcal{Z}$, $p_{n_0,(\mathcal{Z}^j)}(z^j, \cdot)$ has an extension $\tilde{p}_{n_0,(\mathcal{Z}^j)}(z^j, \cdot) \in \Sigma(\beta, L)$ such that $\tilde{I}_{n_0,(\mathcal{Z}^j)} := \int_{\mathbb{R}^{\dim(j)}} |\tilde{p}_{n_0,(\mathcal{Z}^j)}(z^j, \tilde{z}^{\dim(j)})|d\tilde{z}^{\dim(j)} < \infty$.

• $K'$ is of order $k = \lfloor k \rfloor$, i.e.,

$$\int_{\mathbb{R}^{\dim(j)}} K'(u)du = 1, \quad \int_{\mathbb{R}^{\dim(j)}} K'(u)u^a \, du = 0 \quad (1 \leq |a| \leq k),$$

where $a \in \mathbb{Z}^{\dim(j)}$ is a multi-index, and $K'$ satisfies $\int_{\mathbb{R}^{\dim(j)}} |K'(u)| \cdot ||u||^a \, du < \infty$.

Remark 3 (Existence of the smooth extensions). The smooth extensions in Assumption 1 exist, for example, if we consider a smooth density function $\tilde{p}_{n_0(j)}$ on $\mathbb{R}^{\dim(j)}$ and regard its restriction to $\mathbb{Z}^{\dim(j)}$ with appropriate scaling as $p_{n_0(j)}$.

C.2.3 Statement and Proof

We prove the following theorem. Theorem 1 is obtained by changing $\delta$ to $\frac{\delta}{2 \delta}$ in the following theorem, substituting $\|\mathbf{H}\|_1 = \max_{f \in \mathcal{F}(j)} h^f$, 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

$$C_H := B_1 \sum_{j=1}^{D} \frac{1}{\epsilon_{n,j}(H)} \left( B_j + \frac{B_k}{\epsilon_{K,H}(H)} \right) \Phi(\beta, L, K') \|H\|_1, \quad C_p := B_1 \sum_{j=1}^{D} \frac{\Phi_{K,H}(I)}{\epsilon_{n,j}(H)} \mathbf{I}_{n,j} + \frac{B_k}{\epsilon_{K,H}(H)} \mathbf{I}_{n,j}.'$$

$$C_K := \max_{j \in [D]} \left\{ \frac{1}{\epsilon_{K,H}(H)} \cdot \frac{B_k}{\epsilon_{K,H}(H)} \right\}, \quad R_F := \sum_{j=1}^{D} \text{det} \mathbf{H} \cdot \text{Rad}_{a,h} \left( \mathcal{L}^f \otimes \mathcal{K}' \mathbf{H} \right), \quad R_K := \sum_{j=1}^{D} \text{det} \mathbf{H} \cdot \text{Rad}_{a,h} \left( \mathcal{K}' \mathbf{H} \right).$$

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

$$R(\hat{f}) - R(f^*) \leq 2(C_H + C_p) + 4C_K R_F + B_1 R_K + 2DB_1 B_2 C_K \sqrt{\frac{\log(2/\delta)}{2n}}.$$

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,

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

(1)

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

$$\frac{|A|}{|B|} - \frac{|C|}{|D|} \leq \frac{|A - C|}{|B|} \cdot \frac{C}{|B - D|} \leq \frac{|A - C|}{|B - D|} \cdot |B - D|. \quad (2)$$

Proof of Theorem 1. First, note

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

For ease of notation, define $\hat{p}(z|z^{\dim(j)}) = \sum_{j=1}^{D} \delta^j(z^j) \hat{w}^j(z^{\dim(j)})$ and temporarily denote $p_k := p_{\text{km}(k)}$. With this notation, $\hat{R}_{\text{aug}}(f) = \int_{\mathcal{Z}} \ell_j(z) \prod_{j=1}^{D} \hat{p}(z^j|z^{\dim(j)}) \, dz$. Then, applying the argument of Eq. (1), we have

$$\begin{align*}
\left. \sum_{f \in \mathcal{F}} \int_{\mathcal{Z}} \ell_j(z) \prod_{j=1}^{D} p_j(z^j|z^{\dim(j)}) \, dz - \int_{\mathcal{Z}} \ell_j(z) \prod_{j=1}^{D} \hat{p}(z^j|z^{\dim(j)}) \, dz \right. \\
= \left. \sup_{f \in \mathcal{F}} \int_{\mathcal{Z}} \ell_j(z) \left( \prod_{j=1}^{D} p_j(z^j|z^{\dim(j)}) \right) \left( p_j(z^j|z^{\dim(j)}) - \hat{p}_j(z^j|z^{\dim(j)}) \right) \left( \prod_{k=1}^{D} \hat{p}_k(z^j|z^{\dim(j)}) \right) \, dz \right. \\
\leq \left. \sum_{j=1}^{D} \sup_{f \in \mathcal{F}} \int_{\mathcal{Z}} \ell_j(z) \left( \prod_{j=1}^{D} p_j(z^j|z^{\dim(j)}) \right) \left( p_j(z^j|z^{\dim(j)}) - \hat{p}_j(z^j|z^{\dim(j)}) \right) \left( \prod_{k=1}^{D} \hat{p}_k(z^j|z^{\dim(j)}) \right) \, dz \right..
\end{align*}$$
Now, for \( f \in \mathcal{F} \) and \( j \in [D] \), we define \( \ell_{f,j}^{(1)} : z' \mapsto \ell_{f,j}(Z_{k_1-1}, z') \). Then, for each \( j \in [D] \), applying Lemma 5, we obtain

\[
(*) = \sup_{f \in \mathcal{F}} \left| \sum_{i_1=1}^n \cdots \sum_{i_{k_1-1}=1}^n \left( \int Z_i \ell_{f,i}^{(1)}(z') p_j(z' | Z_{k_1-1}^{\text{mix}}) dz' - \sum_{i_{k_1-1}=1}^n \ell_{f,i}^{(1)}(Z_i') \phi_i k_{1,j}(Z_i') \right) \phi_j k_{1,j-2} \cdots \phi_1 \right| \leq 1 \cdot \sup_{f \in \mathcal{F}} \max \left( \int Z_i \ell_{f,i}^{(1)}(z') p_j(z' | Z_{k_1-1}^{\text{mix}}) dz' - \sum_{i_{k_1-1}=1}^n \ell_{f,i}^{(1)}(Z_i') \phi_i k_{1,j}(Z_i') \right) \leq \max_{k_1,j \in [n]} \sup_{f \in \mathcal{F}} \max Z_{k_1-1}^{\text{mix}} Z_{k_1}^{\text{mix}} \left( \int Z_i \ell_{f,i}^{(1)}(z') p_j(z' | Z_{k_1-1}^{\text{mix}}) dz' - \sum_{i_{k_1-1}=1}^n \ell_{f,i}^{(1)}(Z_i') \phi_i k_{1,j}(Z_i') \right) \leq \sup_{f \in \mathcal{F}} \sup_{Z_{k_1-1}^{\text{mix}} Z_{k_1}^{\text{mix}}} \left( \int Z_i \ell_{f,i}^{(1)}(z') p_j(z' | Z_{k_1-1}^{\text{mix}}) dz' - \sum_{i_{k_1-1}=1}^n \ell_{f,i}^{(1)}(Z_i') \phi_i k_{1,j}(Z_i') \right),
\]

where we used that \( \left\{ Z_{k_1-1}^{\text{mix}} \right\}_{k_1,j \in [n]} \subset Z_{k_1}^{\text{mix}} \) that follows from \( \left\{ Z_i^{\text{mix}} \right\}_{i=1}^n \subset Z_{k_1}^{\text{mix}} \). Define

\[
r^j(f, Z_{k_1}^{\text{mix}}) := \int Z_i f(z') p_j(z | Z_{k_1}^{\text{mix}}) dz', \quad \hat{r}^j(f, Z_{k_1}^{\text{mix}}) := \frac{1}{n} \sum_{i=1}^n f(Z_i) K_{ij}(Z_{k_1}^{\text{mix}} - Z_i^{\text{mix}}),
\]

\[
g^j(Z_{k_1}^{\text{mix}}) := p_{\text{mix}}(Z_{k_1}^{\text{mix}}), \quad \hat{g}^j(Z_{k_1}^{\text{mix}}) := \frac{1}{n} \sum_{i=1}^n K_{ij}(Z_{k_1}^{\text{mix}} - Z_i^{\text{mix}}).
\]

Then, for each \( \ell_{f,j}^i \in L_{f,j}^i \) and \( Z_{k_1}^{\text{mix}} \in Z_{k_1}^{\text{mix}} \),

\[
(**) = \frac{r^j(\ell_{f,j}^i, Z_{k_1}^{\text{mix}})}{g^j(Z_{k_1}^{\text{mix}})} = \frac{\hat{r}^j(\ell_{f,j}^i, Z_{k_1}^{\text{mix}})}{\hat{g}^j(Z_{k_1}^{\text{mix}})} \leq \frac{r^j(\ell_{f,j}^i, Z_{k_1}^{\text{mix}})}{g^j(Z_{k_1}^{\text{mix}})} + \frac{\mathbb{E} \hat{r}^j(\ell_{f,j}^i, Z_{k_1}^{\text{mix}})}{\mathbb{E} \hat{g}^j(Z_{k_1}^{\text{mix}})} + \frac{\mathbb{E} \hat{r}^j(\ell_{f,j}^i, Z_{k_1}^{\text{mix}})}{\mathbb{E} \hat{g}^j(Z_{k_1}^{\text{mix}})} - \frac{\hat{r}^j(\ell_{f,j}^i, Z_{k_1}^{\text{mix}})}{\hat{g}^j(Z_{k_1}^{\text{mix}})} - \frac{\hat{g}^j(Z_{k_1}^{\text{mix}})}{\hat{g}^j(Z_{k_1}^{\text{mix}})}.
\]

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

\[
\rho_1 \leq \frac{1}{\hat{g}^j(Z_{k_1}^{\text{mix}})} \cdot r^j(\ell_{f,j}^i, Z_{k_1}^{\text{mix}}) - \hat{r}^j(\ell_{f,j}^i, Z_{k_1}^{\text{mix}}) + \frac{\mathbb{E} \hat{r}^j(\ell_{f,j}^i, Z_{k_1}^{\text{mix}})}{\mathbb{E} \hat{g}^j(Z_{k_1}^{\text{mix}})} \cdot |g^j(Z_{k_1}^{\text{mix}}) - \hat{g}^j(Z_{k_1}^{\text{mix}})|,
\]

\[
\rho_2 \leq \frac{1}{\mathbb{E} \hat{g}^j(Z_{k_1}^{\text{mix}})} \cdot |\mathbb{E} \hat{r}^j(\ell_{f,j}^i, Z_{k_1}^{\text{mix}}) - \hat{r}^j(\ell_{f,j}^i, Z_{k_1}^{\text{mix}})| + \frac{\mathbb{E} \hat{r}^j(\ell_{f,j}^i, Z_{k_1}^{\text{mix}})}{\mathbb{E} \hat{g}^j(Z_{k_1}^{\text{mix}})} \cdot |\hat{g}^j(Z_{k_1}^{\text{mix}}) - \hat{g}^j(Z_{k_1}^{\text{mix}})|.
\]

Applying Lemma 1 to the coefficients, Lemma 2 to the deterministic difference terms bounding \( \rho_1 \), Lemma 3 to the stochastic difference terms bounding \( \rho_2 \) along with the union bound, for any \( \delta \in (0, 1) \), we have with probability at least \( 1 - 2D\delta \)

\[
R(f) - R(f^*) \leq 2 \sum_{j=1}^D \left( \frac{1}{\epsilon_{\text{mix}}^j} \left( B_j B_k \Phi(\beta, L, K') \| H \|_0^\beta + B_k \phi_k \| I_{\text{mix}} \|_{\text{mix}} \right) + \frac{B_j B_k}{\epsilon_{K}(H)} \left( \Phi(\beta, L, K') \| H \|_0^\beta + \phi_k \| I_{\text{mix}} \|_{\text{mix}} \right) \right)
\]

By reorganizing the terms, we obtain the assertion. □
C.2.4 Lemmas

Here, we prove the lemmas used in the proof of Theorem \[\text{1}\].

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

\[
\sup_{z \in \mathbb{Z}^{m(j)}} \left| \frac{1}{\hat{g}'(z^{\text{mp}(j)})} \right| \leq \frac{1}{\epsilon_{K'}(H)},
\]

\[
\sup_{z \in \mathbb{Z}^{m(j)}} \left| \frac{1}{\hat{g}'(z^{\text{mp}(j)})} \right| \leq \frac{1}{\epsilon_{K'}(H)}.
\]

Proof. By Assumption \[\text{1}\] we have

\[
\sup_{z \in \mathbb{Z}^{m(j)}} \left| \frac{1}{\hat{g}'(z^{\text{mp}(j)})} \right| = \frac{1}{\inf_{z \in \mathbb{Z}^{m(j)}} |\hat{g}'(z^{\text{mp}(j)})|} \leq \frac{1}{\epsilon_{K'}(H)}.
\]

Also,

\[
\sup_{z \in \mathbb{Z}^{m(j)}} \left| \frac{1}{\hat{g}'(z^{\text{mp}(j)})} \right| \leq \frac{1}{\inf_{z \in \mathbb{Z}^{m(j)}} |\hat{g}'(z^{\text{mp}(j)})|} \leq \frac{1}{\epsilon_{K'}(H)}.
\]

where we used the positivity of the integrand. Now,

\[
\sup_{z \in \mathbb{Z}^{m(j)}} \left| \frac{1}{\hat{g}'(z^{\text{mp}(j)})} \right| = \frac{1}{\inf_{z \in \mathbb{Z}^{m(j)}} |\hat{g}'(z^{\text{mp}(j)})|} \leq \frac{1}{\epsilon_{K'}(H)}.
\]

Similarly, we have \( \inf_{z \in \mathbb{Z}^{m(j)}} |\det H| \cdot |\hat{g}'(z^{\text{mp}(j)})| \geq \epsilon_{K'}(H) \). Therefore,

\[
\sup_{z \in \mathbb{Z}^{m(j)}} \left| \frac{1}{\hat{g}'(z^{\text{mp}(j)})} \right| \leq \frac{1}{\epsilon_{K'}(H)}.
\]

\[
\sup_{z \in \mathbb{Z}^{m(j)}} \left| \frac{1}{\hat{g}'(z^{\text{mp}(j)})} \right| \leq \frac{1}{\epsilon_{K'}(H)}.
\]

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

\[
\sup_{z \in \mathbb{Z}^{m(j)}} \left| \frac{r'(\ell_{j}, z^{\text{mp}(j)}) - \hat{g}'(z^{\text{mp}(j)})}{\hat{g}'(z^{\text{mp}(j)})} \right| \leq B_{r}B_{\Phi}(\beta, L, K') \sup_{z \in \mathbb{Z}^{m(j)}} \left| \frac{1}{\hat{g}'(z^{\text{mp}(j)})} \right| \leq B_{r}B_{\Phi}(\beta, L, K') \epsilon_{K'}(H).
\]

\[
\sup_{z \in \mathbb{Z}^{m(j)}} \left| \frac{r'(\ell_{j}, z^{\text{mp}(j)}) - \hat{g}'(z^{\text{mp}(j)})}{\hat{g}'(z^{\text{mp}(j)})} \right| \leq B_{r}B_{\Phi}(\beta, L, K') \epsilon_{K'}(H) \leq B_{r}B_{\Phi}(\beta, L, K') \epsilon_{K'}(H).
\]
Proof. By applying Lemma 4 under Assumption 2
\[
\sup_{z^m \in \mathcal{Z}^m} |r^f(z^m) - \mathbb{E} \tilde{r}^f(z^m)| = \sup_{z^m \in \mathcal{Z}^m} \left| p_{m,\tilde{z}}(z^{\text{mp}}) - \left( K^f_{\text{mp}}(z^{\text{mp}}) \right) + \sup_{\gamma \in \mathcal{Z}^m} \left| \int_{\mathcal{Z}^m} K^f_{\gamma}(z^{\text{mp}}) - z^{\text{mp}} \right| \right| \leq \Phi(\beta, L, K') \left| \mathbf{H} \right| \left| \delta \right|
\]
Similarly, for each \( \ell_{j,f} \in \mathcal{L}_f \) and \( z^{\text{mp}} \in \mathcal{Z}^{\text{mp}} \),
\[
|r^f(\ell_{j,f}, z^{\text{mp}}) - \mathbb{E} \tilde{r}^f(\ell_{j,f}, z^{\text{mp}})| = \left| \int_{\mathcal{Z}} \ell_{j,f}(z^{\text{mp}}) p_{m,\tilde{z}}(z^{\text{mp}}) dz - \int_{\mathcal{Z}} \ell_{j,f}(z^{\text{mp}}) \left( \int_{\mathcal{Z}^m} K^f_{\gamma}(z^{\text{mp}}) - \left( K^f_{\gamma}(z^{\text{mp}}) \right) \right) dz \right| \leq B_f \sup_{z^m \in \mathcal{Z}^m} \left| \int_{\mathcal{Z}^m} K^f_{\gamma}(z^{\text{mp}}) - \left( K^f_{\gamma}(z^{\text{mp}}) \right) \right| dz \]
Applying Lemma 4 under Assumption 2 for each \( z^i \in \mathcal{Z}^i \), we obtain
\[
\sup_{z^m \in \mathcal{Z}^m} \left| \tilde{p}_{\text{mp}}(z^i, z^{\text{mp}}) - \left( K^f_{\gamma}(z^{\text{mp}}) \right) \right| \leq \Phi(\beta, L, K') \left| \mathbf{H} \right| \left| \delta \right|
\]
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
\[
\sup_{\ell_{j,f} \in \mathcal{L}_f} \mathbb{E} \left( r^f(\ell_{j,f}, z^{\text{mp}}) - \tilde{r}^f(\ell_{j,f}, z^{\text{mp}}) \right) \leq 2 \text{Rad}_n \left( \mathcal{L}_f \otimes \mathbf{K}_j \right) + \frac{B_f K}{\left| \mathbf{H} \right|} \sqrt{\frac{\log(2/\delta)}{2n}}
\]
Similarly, for any \( \delta \in (0, 1) \), with probability at least \( 1 - \delta \), we have
\[
\sup_{z^{\text{mp}} \in \mathcal{Z}^{\text{mp}}} \mathbb{E} \left( r^f(z^{\text{mp}}, \ell_{j,f}) - \tilde{r}^f(z^{\text{mp}}, \ell_{j,f}) \right) \leq 2 \text{Rad}_n \left( \mathbf{K}_j \right) + \frac{B_f}{\left| \mathbf{H} \right|} \sqrt{\frac{\log(2/\delta)}{2n}}
\]
Proof. Note
\[
\sup_{\ell_{j,f} \in \mathcal{L}_f} \sup_{z^{\text{mp}} \in \mathcal{Z}^{\text{mp}}} \mathbb{E} \left( r^f(\ell_{j,f}, z^{\text{mp}}) - \tilde{r}^f(\ell_{j,f}, z^{\text{mp}}) \right) = \sup_{\ell_{j,f} \in \mathcal{L}_f} \sup_{k \in \mathcal{K}_j} \left| \frac{1}{n} \sum_{i=1}^n \ell_{j,f}(z^i) k(z^{\text{mp}}, z^i) \right| - \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \ell_{j,f}(z^i) k(z^{\text{mp}}, z^i) \right]
\]
and
\[
\sup_{z^{\text{mp}} \in \mathcal{Z}^{\text{mp}}} \mathbb{E} \left( r^f(z^{\text{mp}}, \ell_{j,f}) - \tilde{r}^f(z^{\text{mp}}, \ell_{j,f}) \right) = \sup_{k \in \mathcal{K}_j} \left| \frac{1}{n} \sum_{i=1}^n k(z^{\text{mp}}, z^i) \right| - \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n k(z^{\text{mp}}, z^i) \right]
\]
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\textsuperscript{[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:

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

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 $n \in \mathbb{N}$, $x_1, \ldots, x_n \geq 0$, and $w_1, \ldots, w_n \geq 0$. Define $w := w_1 + \cdots + w_n$ and assume $w > 0$. Then,

$$\frac{w_1x_1 + \cdots + w_nx_n}{w} \geq \left(\frac{x_1^{w_1} \cdots x_n^{w_n}}{w}\right)^{\frac{1}{w}}.$$ 

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 $G$ 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 $\{z_i\}_{i=1}^m \sim z$ the following holds:

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

C.2.6 Basic Lemmas

Here, we prove the basic lemmas used in the proof of Theorem\textsuperscript{[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 $k = [\beta]$ and satisfies

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

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

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

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

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

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

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

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

(*)
We apply Fact 1 to obtain

\[
(*) = \left| f(x) - \int_{\mathbb{R}^d} K(u) \left( f(x) + \sum_{|\alpha| \leq k} \frac{\partial^{|\alpha|} f(x)}{\alpha!} (-Hu)^{\alpha} + \sum_{|\alpha| = k} \frac{|\alpha|}{\alpha!} (-Hu)^{\alpha} \int_0^1 (1 - t)|^{\beta - 1} \partial^{\alpha}f(x + t(-Hu)) dt \right) du \right|
\]

\[
= \left| \int_{\mathbb{R}^d} K(u) \left( \sum_{|\alpha| = k} \frac{|\alpha|}{\alpha!} (-Hu)^{\alpha} \int_0^1 (1 - t)|^{\beta - 1} \partial^{\alpha}f(x + t(-Hu)) dt \right) du \right|
\]

\[
\leq \int_{\mathbb{R}^d} |K(u)| \left( \sum_{|\alpha| = k} \frac{|\alpha|}{\alpha!} |Hu|^{\alpha} \int_0^1 (1 - t)|^{\beta - 1} L \|Hu\|^{\beta - k} dt \right) du,
\]

where \( \alpha = (\alpha_1, \ldots, \alpha_d) \) is a multi-index and \( |Hu|^\alpha := |h_{1\alpha_1}| \cdots |h_{d\alpha_d}|. \) Now, by the Hölder-condition of \( \partial^{\alpha}f \), we have \( |\partial^{\alpha}f(x + t(-Hu)) - \partial^{\alpha}f(x)| \leq L \|Hu\|^{\beta - k} \). Also, by applying Fact 2 we have

\[
|Hu|^\alpha = |h_{1\alpha_1}| \cdots |h_{d\alpha_d}| \leq \left( \frac{1}{|\alpha|} \sum_{j=1}^d \alpha_j |h_{j\alpha_j}| \right)^{||\alpha||} \leq \left( \frac{1}{|\alpha|} \|\alpha\| \cdot \|Hu\| \right)^{||\alpha||} = \frac{\|\alpha\|^{|\alpha||}}{K^{|\alpha||}} \|Hu\|^{|\alpha||}.
\]

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

\[
(**) \leq \int_{\mathbb{R}^d} |K(u)| \left( \sum_{|\alpha| = k} \frac{|\alpha|}{\alpha!} |Hu|^{\alpha} \int_0^1 (1 - t)^{|\beta - 1|} L \|Hu\|^{\beta - k} dt \right) du
\]

\[
L \int_0^1 (1 - t)^{|\beta - 1|} \|Hu\|^{\beta - k} dt \sum_{|\alpha| = k} \frac{|\alpha|}{\alpha!} \int_0^1 |K(u)| \cdot \|Hu\|^{\alpha} du.
\]

Finally, applying \( \|Hu\| \leq \|H\|_{op} \|u\| \), we have the assertion. \( \square \)

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

\[
\sum_{i=1}^n \cdots \sum_{i=1}^n \tilde{w}_{j_1 \cdots j_{i-1}} \tilde{w}_{1} \in [0, 1].
\]

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

\[
\sum_{i=1}^n \tilde{w}_i(z^{mp(j)}) = \begin{cases} 
\sum_{i=1}^n \frac{1}{n} & \text{if } mp(j) = \emptyset, \\
\sum_{i=1}^n 0 & \text{if } K_i(z^{mp(j)} - Z_i^{mp(j)}) = 0, \forall i, \\
\sum_{i=1}^n \frac{K_i(z^{mp(j)} - Z_i^{mp(j)})}{\sum_{i=1}^n K_i(z^{mp(j)} - Z_i^{mp(j)})} & \text{otherwise,}
\end{cases}
\]

\( \in [0, 1] \).

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

\[
\sum_{i=1}^n \tilde{w}_i = \sum_{i=1}^n \frac{1}{n} = 1.
\]

For \( j \geq 2 \),

\[
\sum_{i=1}^n \cdots \sum_{i=1}^n \tilde{w}_{j_1 \cdots j_{i-1}} \tilde{w}_i = \sum_{i=1}^n \cdots \sum_{i=1}^n \tilde{w}_{j_1 \cdots j_{i-1}} \tilde{w}_i \left( \sum_{i=1}^n \tilde{w}_{j_1 \cdots j_{i-1}} \right)
\]

\( \in \left\{ 0, \left( \sum_{i=1}^n \cdots \sum_{i=1}^n \tilde{w}_{j_1 \cdots j_{i-1}} \right) \right\} \).

By recursively applying the above argument for a finite number of times, we obtain the assertion for all \( j \in [D] \). \( \square \)
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 \(R_{\mathcal{F},K}\) and \(R_K\)).

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 \(1 - 2D\delta\),

\[
R(\hat{f}) - R(f^*) \leq 2(C_H + C_p) + 4C_K(R_{\mathcal{F},K} + B_KR_K) + 2DB_KC_K \sqrt{\frac{\log(2/\delta)}{2n}}.
\]

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

**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

\[
\hat{f}_{\text{emp}} \in \arg \min_{f \in \mathcal{F}} [\hat{R}_{\text{emp}}(f)]
\]

satisfies

\[
R(\hat{f}_{\text{emp}}) - R(f^*) \leq 4\text{Rad}_{n,p}(\mathcal{L}_F) + 2B_L \sqrt{\frac{\log(2/\delta)}{2n}}.
\]

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

\[
R(\hat{f}_{\text{emp}}) - R(f^*) = R(\hat{f}_{\text{emp}}) - \hat{R}_{\text{emp}}(\hat{f}_{\text{emp}}) + \hat{R}_{\text{emp}}(\hat{f}_{\text{emp}}) - R(f^*)
\leq R(\hat{f}_{\text{emp}}) - \hat{R}_{\text{emp}}(\hat{f}_{\text{emp}}) + \hat{R}_{\text{emp}}(f^*) - R(f^*) \leq 2 \sup_{f \in \mathcal{F}} |R(f) - \hat{R}_{\text{emp}}(f)|.
\]

\(\blacksquare\)

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

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

- The functions in \(\mathcal{L}_F\) are \(L_1\)-Lipschitz continuous.
- The functions \(K^t\) are \(L_{K^t}\)-Lipschitz continuous.
- The functions \(p_{t|\text{init}}(z^t|\cdot)\) are \(L_{p,z}\)-Lipschitz continuous for all \(z^t\).

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

Under this assumption, we have the following:

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

\[
\text{Rad}_{n,p}(\mathcal{L}_F) \leq O\left(n^{-\frac{1}{2D}}\right), \quad R_{\mathcal{F},K} + B_KR_K \leq O\left(n^{-1/3}\right).
\]
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 \( \mathcal{L}_F \) and the boundedness of \( Z \), we can apply Fact 6 to obtain

\[
\log N(\|\cdot\|_\infty)(\mathcal{L}_F) \leq C \left( \frac{L}{t} \right)^D
\]

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,

\[
\log N(\|\cdot\|_\infty)(\mathcal{L}_F) \leq \log N(\|\cdot\|_\infty)(\mathcal{L}_F^j) \leq \log N(\|\cdot\|_\infty)(\mathcal{L}_F^j) + \log N(\|\cdot\|_\infty)(K_{ij})
\]

where \( t_1, t_2 \) are such that \( B_k t_1 + B_j t_2 = t \). Now, applying Lemma 8,

\[
\log N(\|\cdot\|_\infty)(\mathcal{L}_F) \leq \log \sup_{z \in \mathbb{Z}^{r+1}} N(t_1, \|\cdot\|_\infty)(\mathcal{F}_j) + \log N(t_1, \|\cdot\|_\infty)(\mathbb{B}^{j-1}(R_Z))
\]

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

\[
\log \sup_{z \in \mathbb{Z}^{r+1}} N(t_1, \|\cdot\|_\infty)(\mathcal{F}_j) \leq C \frac{L_2}{t_1}, \quad \log N(t_1, \|\cdot\|_\infty)(\mathbb{B}^{j-1}(R_Z)) \leq (j - 1) \log \left( 1 + \frac{2R_Z}{t_1} \right),
\]

where \( t_1, t_2 \) are such that \( t_1 = t_1 + L_2 t_2, \) and \( t_2 = L_1 + B_k \sum_k L_{jk} \).

On the other hand, by Lemma 10 we have

\[
\log N(\|\cdot\|_\infty)(K_{ij}) \leq |mp(j)| \log \left( 1 + \frac{2L_{K,j,1} R_Z}{t_2} \right),
\]

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

Therefore, we have

\[
\log N(\|\cdot\|_\infty)(\mathcal{L}_F^j) \leq C \frac{L_2}{t_1} + (j - 1) \log \left( 1 + \frac{2R_Z}{t_1} \right) + |mp(j)| \log \left( 1 + \frac{2L_{K,j,1} R_Z}{t_2} \right).
\]

By applying Fact 4, letting

\[
t_1 = \frac{t}{3B_k}, \quad t_2 = \frac{3B_k L_2}{t}, \quad t_2 = \frac{t}{3B_k},
\]

and minimizing the upper bound for \( t \), we have

\[
|\text{det} H_{ij}^p| \text{Rad}_n(\mathcal{L}_F^j \otimes K_{ij}) \leq O \left( n^{-1/3} \right).
\]

Therefore, we have

\[
R_{F,K} = \sum_{j=1}^D |\text{det} H_{ij}^p| \text{Rad}_n(\mathcal{L}_F^j \otimes K_{ij}) \leq O \left( n^{-1/3} \right)
\]

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 $\|f\|_\infty \leq M_\mathcal{F}(f \in \mathcal{F})$ and $\|g\|_\infty \leq M_\mathcal{G}(g \in \mathcal{G})$. Then, we have for any $t_1, t_2 > 0$,

$$\log N(t_1\|\|_{\infty})(\mathcal{F} \otimes \mathcal{G}) \leq \log N(t_1, 1\|\|_{\infty})(\mathcal{F}) + \log N(t_2\|\|_{\infty})(\mathcal{G})$$

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

**Proof.** Let $(f_i), (g_i)$ 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

$$\|f \otimes g - f_i \otimes g_j\|_\infty \leq \|f \otimes g - f_i \otimes g\|_\infty + \|f_i \otimes g - f_i \otimes g_j\|_\infty \leq M_\mathcal{F}t_1 + M_\mathcal{G}t_2.$$

This implies the assertion. \qed

**Lemma 7** (Lipschitz continuity of marginalized function class). Assume that $p_{\|\|_{(\cdot,t)\|\|_{\infty}}}(z^k)$ is $L_{p,k}$-Lipschitz continuous for all $z^k$. Then, the elements of $\mathcal{L}_t^j$ are Lipschitz continuous with the constant $L_1 + B_t \sum_k L_{p,k}$.

**Proof.** Since the functions in $\mathcal{L}_t^j$ are $L_1$-Lipschitz continuous, the elements of $\mathcal{L}_t^j$ are also Lipschitz continuous:

$$|\ell_{f_j}(x) - \ell_{f_j}(y)| = \left|\int \ell_{f_j}(x, z) \prod_p p_{\|\|_{(z^k,t)\|\|_{\infty}}}(z^k | (x, z)^{\text{mp}}) dz - \int \ell_{f_j}(y, z) \prod_p p_{\|\|_{(z^k,t)\|\|_{\infty}}}(z^k | (y, z)^{\text{mp}}) dz\right|$$

$$\leq \int |\ell_{f_j}(x, z) - \ell_{f_j}(y, z)| \prod_p p_{\|\|_{(z^k,t)\|\|_{\infty}}}(z^k | (y, z)) dz$$

$$+ \sum_{k \geq j+1} \int |\ell_{f_j}(x, z)| \prod_p p_{\|\|_{(z^k,t)\|\|_{\infty}}}(z^k | (x, z)) \cdots (p_{\|\|_{(z^k,t)\|\|_{\infty}}}(z^k | (y, z))) \cdots p_{\|\|_{(z^k,t)\|\|_{\infty}}}(z^k | (y, z)) dz$$

$$\leq L_z \|x - y\|_t \cdot 1 + B_t \sum_k 1 \cdot L_{p,k} \|x - y\|_t \cdot 1$$

$$\leq (L_1 + B_t \sum_k L_{p,k}) \|x - y\|_t.$$

\qed

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

$$\log N(t, 1\|\|_{\infty})(\mathcal{L}_t^j) \leq \log \sup_{z \in \mathcal{Z}^{j-1}} N_{u, 1\|\|_{\infty}}(\mathcal{F}_z) + \log N(t, \|\|_{\infty})(\mathcal{B}^{j-1}(R_z))$$

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

$$\|\ell_{f_j}(z_\mu, \cdot) - \ell_{f_j}^\mu(z_\mu, \cdot)\|_\infty \leq \|\ell_{f_j}(z_\mu, \cdot) - \ell_{f_j}(z_\mu, \cdot)\|_\infty + \|\ell_{f_j}(z_\mu, \cdot) - \ell_{f_j}^\mu(z_\mu, \cdot)\|_\infty \leq L_2 v + u$$

Therefore, the set $\bigcup_\mu \{z_\mu\}_\mu \times \{\ell_{f_j}^\mu\}_k$ induces a $(L_2 v + u)$-covering of $\mathcal{L}_t^j$. Noting that the cardinality of $\bigcup_\mu \{z_\mu\}_\mu$ is bounded by $N(t, \|\|_{\infty})(\mathcal{B}^{j-1}(R_z))$ and that of $\{\ell_{f_j}^\mu\}_k$ by $\sup_{z \in \mathcal{Z}^{j-1}} N_{u, 1\|\|_{\infty}}(\mathcal{F}_z)$, we have the assertion. \qed

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

$$\sup_{z \in \mathcal{Z}^{j-1}} N_{u, 1\|\|_{\infty}}(\mathcal{F}_z) \leq C \frac{L_2}{u}.$$
Proof. Since the elements of $\mathcal{L}_j^l$ are $L_2$-Lipschitz continuous, so are the elements of $\mathcal{F}_\epsilon$ with Lipschitz constant $L_2$. Indeed, for any $x, y \in \mathcal{Z}_j$ and $z \in \mathcal{Z}_{j+1}$, we have
\[
|\ell_{j,l}(z,x) - \ell_{j,l}(z,y)| \leq L_2 \|z - (z)| = L_2 \|x - y\|.
\]
Therefore, by applying Lemma 5, we have the assertion. □

**Lemma 10** (Shifted kernel complexity). Assume that $K^l : \mathbb{R}^{m \times p(j)} \to \mathbb{R}$ is $L_{K,j}$-Lipschitz continuous. Let $L_{K,j} = \frac{1}{|\operatorname{det}(H_j)|} \|H_j^{-1}\|_{\operatorname{op}}$. Then, we have the following:
\[
\log N_{d(2,\|\cdot\|_{\infty})}(\mathcal{K}_n^j) \leq |m\nu(j)| \log \left(1 + \frac{2L_{K,j} R \|x\|}{t_2}\right).
\]

Proof. Recalling $K_n^j(u) = \frac{1}{|\operatorname{det}(H_j)|} K^l(H_j^{-1} u)$, for any $K_n^j(z_1 - \cdot), K_n^j(z_2 - \cdot) \in \mathcal{K}_n^j$, we have
\[
\|K_n^j(z_1 - \cdot) - K_n^j(z_2 - \cdot)\|_{\infty} \leq \frac{1}{|\operatorname{det}(H_j)|} L_{K,j} \|H_j^{-1}\|_{\operatorname{op}} \|z_1 - z_2\|
\]
\[
\leq \frac{1}{|\operatorname{det}(H_j)|} L_{K,j} \|H_j^{-1}\|_{\operatorname{op}} \|z_1 - z_2\|
\]
Therefore, we have
\[
\log N_{d(2,\|\cdot\|_{\infty})}(\mathcal{K}_n^j) \leq \log N_{d(2;L_{K,j},1)}(\mathcal{Z}^{m\nu(j)}).
\]
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:
\[
\operatorname{Rad}_m(\mathcal{F}) \leq t + c \sqrt{\frac{\log N_{d(\|\cdot\|_{\infty})}(\mathcal{F})}{m}}
\]

**Fact 5** (Euclidean ball metric entropy bound; Wainwright, 2019 Example 5.8, p.126). Let $R > 0$ and $d \in \mathbb{N}$. Let $B(R)$ denote the radius-$R$ ball in the $d$-dimensional Euclidean space. Then, we have the following metric entropy bound:
\[
\log N_{d(\|\cdot\|_{\infty})}(B(R)) \leq d \log \left(1 + \frac{2R}{\delta}\right).
\]

**Fact 6** (Lipschitz functions metric entropy bound; Wainwright, 2019 Example 5.10, p.129). Let $L, R > 0$ and $d \in \mathbb{N}$. Let $\operatorname{Lip}(R, L)$ denote the set of $L$-Lipschitz functions on $[0, R]^d$. Then, we have the following metric entropy bound for sufficiently small $\delta > 0$:
\[
\log N_{d(\|\cdot\|_{\infty})}(\operatorname{Lip}(R, L)) \leq C \left(\frac{LR}{\delta}\right)^d,
\]
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 $O\left(n^p\right)$. The main computation cost of Algorithm 1 comes from the computation of the weights $\hat{w}_{j,|\delta|+1}$. 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, $\{\hat{w}_{j,|\delta|+1} \mid i_{|\delta|} \in \mathcal{Z}_j\}$, 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 $|\operatorname{m\nu}(j)| \leq j - 1$. 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 $n' \times n$ divisions. The same computation takes place for each of the $i_{|\delta|+1} \in \mathcal{Z}_{j+1}$ nodes at depth $j$, therefore, the edge weights between depth $j$ and depth $j + 1$ can be computed by $O(n')$ operations. The edge weights are multiplied to obtain the node weights, which requires $O(n')$ multiplications since the number of multiplications that take place is equal to the number of edges in Fig. 2 Overall, Algorithm 1 requires $O\left(n^p\right)$ operations for the edge weight computation and $O\left(n^p\right)$ for the node weight computation, amounting to $O\left(n^p\right)$ operations in total, in the worst case that no edge is pruned by the threshold $\theta$. 
REFERENCES


