Proceedings of Machine Learning Research 246:253–272, 2024 Probabilistic Graphical Models (PGM)

Kernel-Based Differentiable Learning of Non-Parametric Directed Acyclic Graphical Models

Yurou Liang YUROU.LIANG@TUM.DE
Oleksandr Zadorozhnyi OLEKSANDR.ZADOROZHNYI@TUM.DE
Mathias Drton MATHIAS.DRTON@TUM.DE
Technical University of Munich and Munich Center for Machine Learning, Germany

Editors: J.H.P. Kwisthout & S. Renooij

Abstract

Causal discovery amounts to learning a directed acyclic graph (DAG) that encodes a causal model. This model selection problem can be challenging due to its large combinatorial search space, particularly when dealing with non-parametric causal models. Recent research has sought to bypass the combinatorial search by reformulating causal discovery as a continuous optimization problem, employing constraints that ensure the acyclicity of the graph. In non-parametric settings, existing approaches typically rely on finite-dimensional approximations of the relationships between nodes, resulting in a score-based continuous optimization problem with a smooth acyclicity constraint. In this work, we develop an alternative approximation method by utilizing reproducing kernel Hilbert spaces (RKHS) and applying general sparsity-inducing regularization terms based on partial derivatives. Within this framework, we introduce an extended RKHS representer theorem. To enforce acyclicity, we advocate the log-determinant formulation of the acyclicity constraint and show its stability. Finally, we assess the performance of our proposed RKHS-DAGMA procedure through simulations and illustrative data analyses.

Keywords: Causal discovery; graphical model; kernel methods; RKHS; structural equation model.

1. Introduction

Structural equation models (SEMs) based on directed acyclic graphs (DAGs) have found widespread applications ranging from computational biology (Zhang et al., 2023) to manufacturing (Göbler et al., 2024) and finance (Ji et al., 2018). To represent the joint dependence structure, each variable is modeled as a function of a subset of the other variables and noise. In this setting, DAG-based models assume the absence of causal feedback loops. Although this assumption can be restrictive, it is crucial for defining non-linear models. Indeed, it is often unclear whether cyclic systems of structural equations have a unique solution or if they admit a solution at all.

In many applications, the underlying DAG is unknown, and methods for causal discovery, which learn the DAG from data, offer useful insights. Numerous algorithms have been proposed for causal discovery; see, e.g., Drton and Maathuis (2017) or Spirtes and Zhang (2019). A classical constraint-based approach relies on testing conditional independences (Spirtes and Glymour, 1991; Tsamardinos et al., 2003; Margaritis and Thrun, 1999). Another prominent approach is score-based algorithms (Heckerman et al., 1995; Chickering, 2002). In this work, we focus on a score-based approach. Specifically, we will take up a recent theme that aims to find a DAG minimizing a model selection score through a continuous optimization problem with a continuous acyclicity constraint applied to a weighted adjacency matrix W. This approach was initiated in the NOTEARS algorithm (Zheng et al., 2018), which assumes a linear SEM and uses an exponential acyclicity constraint transforming the combinatorial optimization problem into a continuous one that is solved using an augmented Lagrangian scheme. Several follow-up works have proposed alternative characterizations of acyclicity (Yu et al., 2019; Nazaret et al., 2023; Ng et al., 2020). One such work introduced

the DAGMA algorithm (Bello et al., 2022), which is generally faster than NOTEARS. In line with related literature, we refer to this broad approach as *differentiable causal discovery*.

Zheng et al. (2020) extended this methodology to non-parametric settings. Specifically, the authors modeled each variable as a non-parametric function of the other variables and noise, proposing to approximate the non-parametric functions by multi-layer perceptions or via a (truncated) basis expansion in the original functional space—i.e., the space of functions whose derivatives are square-integrable over the domain (an expansion is then possible via the trigonometric basis of functions). They then minimize the corresponding residual loss subject to the exponential acyclicity constraint. In our work, we build on this approach under additional assumptions that the non-parametric functions are continuously differentiable, taking up a perspective of (Gaussian) RKHS.

In the MLP framework of Zheng et al. (2020), the entry W_{kj} of the weighted adjacency matrix is defined as the L_2 norm of the kth column of the weight matrix in the first hidden layer of the jth MLP. When approximating by basis expansions, the general non-parametric model is assumed to be an additive model. The entries of the weighted adjacency matrix are defined as the L_2 norm of the coefficients corresponding to the basis approximation.

Both the MLP and the basis expansion methods in Zheng et al. (2020) lead to finite-dimensional optimization problems in terms of neural network weights or basis coefficients. The current MLP approximation is sensitive to the number of hidden units: while increasing the size of the hidden layers increases the flexibility of MLP functions, larger networks require more samples for accurate estimation (Zheng et al., 2020). Moreover, the MLP approximation relies on random initialization of weights, which introduces randomness in results (see Figure 2 in Waxman et al., 2024). Fine-tuning the architecture of a neural network is, thus, a non-trivial task. On the other hand, the current basis expansion approximations are restricted through a focus on additive models.

Contributions. In this work, we present a novel kernel-based methodology for differentiable causal discovery in non-parametric settings. Our contributions can be summarized as follows:

- We approximate each non-parametric function that represents the dependency structure between random variables using an RKHS with a differentiable kernel k. We establish a version of an RKHS Representer Theorem for an empirical acyclicity-constrained optimization problem, similar to that of Rosasco et al. 2013 in the statistical learning context. Given data $(x^i)_{i=1}^n$, this leads to optimizing functions that are combinations of evaluations of the kernel and its partial derivatives:

$$\sum_{i=1}^{n} \alpha_i k(x, x^i) + \sum_{i=1}^{n} \sum_{a=1}^{d} \beta_{ai} \frac{\partial k(x, s)}{\partial s^a}\bigg|_{s=x^i}. \tag{1}$$

- Let f_j represent the dependency structure between j-th random variable and the other random variables. If $x \mapsto f_j(x)$ is continuously differentiable for all x in a connected and compact sample space $\mathcal{X}$, then $\frac{\partial f_j}{\partial x_k} = 0$ implies that f_j does not depend on x_k. Thus, we define the weighted adjacency matrix directly via the partial derivatives of the functions f_j. This approach is model-agnostic in the sense that the weights do not refer to approximations to f_j.

- We base our optimization on the DAGMA method (Bello et al., 2022) and adopt the log-determinant acyclicity constraint h_{ldet}, for which we demonstrate stable optimization behavior on the boundary of the domain.

- We explore the behavior of kernel-based differentiable causal discovery in simulation experiments as well as the collection of cause-effect datasets from Mooij et al. (2016). The code for our experiments is available on the first author's GitHub site.[1]

1. https://github.com/yurou-liang/RKHS-DAGMA

Outline. Section 2 sets up notation, reviews the DAG learning problem, and presents basic facts about RKHS. Section 3 summarizes existing acyclicity constraints and discusses the stability of the log-determinant acyclicity constraint, on which our work is based. Section 4 introduces the sparsity regularizer and examines the extended RKHS representer theorem for the constrained optimization problem with acyclicity constraint. Section 5 outlines the resulting constrained optimization problem and introduces the RKHS-DAGMA algorithm for learning sparse nonparametric graphs. In Section 6, we compare the performance of RKHS-DAGMA with different versions of nonparametric NOTEARS (Zheng et al., 2020) in numerical experiments. Additional details on the experiments and proofs are given in the Appendix.

2. Background

2.1. Structural Equation Model and Differentiable Causal Discovery

Let $X = (X_1, \dots, X_d)$ be a random vector taking values in $\mathcal{X} \subset \mathbb{R}^d$ defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We assume $\mathcal{X}$ is a bounded connected non-empty open set. Let $[d] := \{1, 2, \dots, d\}$. Consider a directed acyclic graph $G = (V, E)$ with vertex set $V = [d]$ and edge set $E \subset V \times V$. As usual, $\mathrm{pa}(i) = \{j \in V : (j, i) \in E\}$ is the set of *parents* of a vertex $i \in V$. For a subset $A \subset [d]$, let $X_A = (X_i)_{i \in A}$. When $A = \emptyset$, we set $X_A \equiv 0$.

In a graphical model, each variable X_j exhibits a constrained stochastic relationship with the other coordinates of random vector $X := (X_j)_{j=1}^d$ (Maathuis et al., 2019). Presenting the model through structural equations and assuming additive noise, the model for DAG G postulates that

$$X_j = f_j(X) + \varepsilon_j, \quad j = 1, \dots, d, \tag{2}$$

where each measurable function $f_j : \mathbb{R}^d \mapsto \mathbb{R}$ depends only on the subvector $X_{\mathrm{pa}(i)}$, and $(\varepsilon_j)_{j \in [d]}$ are mutually independent stochastic error terms. In this model, the conditional expectations are $\mathbb{E}[X_j | X_{\mathrm{pa}(j)}] = f_j(X)$, where we note again that $f_j \colon \mathcal{X} \to \mathbb{R}$ does not depend on X_k if $k \notin \mathrm{pa}(j)$.

The DAG learning problem for the considered models may be formulated as follows. Let $\mathbb{X} \in \mathbb{R}^{n \times d}$ be a data matrix whose rows $(x^i)_{i=1}^n$ represent n i.i.d. observations. Let x_j^i be the j-th coordinate of the i-th observation. We denote the loss function by $\ell : \mathcal{X}^n \times \mathcal{Y}^n \mapsto \mathbb{R}_+$, where $\mathcal{Y} \subset \mathbb{R}$ is the image space for predictions. The typical loss function is the least squares loss $\ell(y, \widehat{y}) = \|y - \widehat{y}\|_2^2$. The goal is to estimate $f = (f_1, \dots, f_d)$ by minimizing the score function

$$L(f) := \frac{1}{2n} \sum_{j=1}^{d} \ell(X_j, f_j(\mathbb{X})) \text{ subject to the dependencies in } f \text{ corresponding to a DAG}, \tag{3}$$

where, in slight abuse of notation, $f_j(\mathbb{X}) = \big(f_j(x^1), \dots, f_j(x^n)\big)$. We assume that each f_j is continuously differentiable and that f_j and its derivative are both square-integrable. As in Waxman et al. (2024), we define the weighted adjacency matrix $W \in \mathbb{R}^{d \times d}$ with entries $W(f)_{kj} := \left\| \frac{\partial f_j(\cdot)}{\partial x_k} \right\|_{L^2}$. This definition is model-agnostic in that it does not refer to any particular family of approximation models for the function f_j (as was done, e.g., in the MLP setup of Zheng et al., 2020).

In the sequel, $\mathbb{P}_X$ denotes the joint distribution of the random vector X. We write $C^1(\mathcal{X})$ for the space of continuously differentiable functions over $\mathcal{X}$ and $L_2(\mathcal{X})$ for the space of (equivalence classes of) square-integrable functions on $\mathcal{X}$, i.e., functions $f : \mathcal{X} \to \mathbb{R}$ with $\int_{\mathcal{X}} f^2(x) \mathbb{P}_X(dx) < \infty$. We will mostly drop dependence on the domain and on the underlying distribution $\mathbb{P}_X$ and simply write $L_2 := L_2(\mathcal{X})$ or $L_2 := L_2(\mathbb{R}^d)$ when the domain is clear from the context. Furthermore, we use $\|\cdot\|_\infty$ to denote the essential supremum norm w.r.t. Lebesque measure λ on a (subset) $\mathcal{X} \subset \mathbb{R}$. For $g : \mathcal{X} \mapsto \mathbb{R}$ in $C^1(\mathcal{X})$, we write $\frac{\partial}{\partial x_k} g(x)$ for its partial derivative (as a map $x \mapsto \frac{\partial}{\partial x_k} g(x)$) and

$\frac{\partial}{\partial x_k} g(x)|_{x=s}$ for the derivative's value at point $x = s$. Finally, we denote the $i-$th smallest eigenvalue of a real matrix $A \in \mathbb{R}^{d\times d}$ by $\lambda_i(A)$; so, $|\lambda_1(A)| \leq |\lambda_2(A)| \leq \cdots \leq |\lambda_d(A)|$. The spectral radius $\rho(A) = |\lambda_d(A)|$ is the largest eigenvalue of A in magnitude.

2.2. Kernels and RKHS

To estimate the structural equation model in (2), we use the toolbox of kernel-based algorithms (Steinwart and Christmann, 2008, Chap. 4). For a given vector space $\mathcal{X}$, let $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ be a symmetric function such that for any $n \in \mathbb{N}$, $\{x^i\}_{i=1}^n \in \mathcal{X}^n$, and $c \in \mathbb{R}^n$, it holds that $c^\top Kc \geq 0$, where $K = \big(k(x^i, x^j)\big)_{i,j=1}^n \in \mathbb{R}^{n\times n}$ is the *kernel matrix*. Such a function $k(\cdot,\cdot)$ is termed a *kernel*. Its values can be represented as an inner product in a Hilbert space H_0, i.e., $k(x^i, x^j) = \big\langle \phi(x^i), \phi(x^j)\big\rangle_{H_0}$, where $\phi : \mathcal{X} \mapsto H_0$ is the *feature map*. To every kernel k, we can associate a reproducing kernel Hilbert space (RKHS) H, which is a space of functions $f : \mathcal{X} \mapsto \mathbb{R}$ for which the *reproducing property* holds: $f(x) = \langle f, k(\cdot, x)\rangle_H$ for all $x \in \mathcal{X}$, $f \in H$.

Let H be an RKHS, and let k being its correspondent reproducing kernel $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$. Let $\mathcal{R}_{E,\mathcal{D}}(f_j) := \frac{1}{n}\sum_{i=1}^n E(x_j^i, f_j(x^i))$ be the empirical risk for convex loss-function $E : \mathbb{R} \times \mathcal{Y} \mapsto \mathbb{R}_+$. Consider the empirical risk with function complexity regularizer based on an arbitrary non-decreasing function $J : \mathbb{R} \mapsto \mathbb{R}_+$:

$$\mathcal{R}_{E,\mathcal{D}}(f_j) + J\big(\|f_j\|_H\big). \tag{4}$$

Then (see, e.g. Schölkopf et al., 2001, Thm 1), the minimizer of (4) can be written as

$$\widehat{f}_j(\cdot) = \sum_{i=1}^n \alpha_i^j k(\cdot, x^i),$$

where $\alpha^j \in \mathbb{R}^n$. Furthermore, the reproducing property of the RKHS directly implies that $\widehat{f}_j(x) = \sum_{i=1}^n \alpha_i^j k(x, x^i)$ for all $x \in \mathcal{X}$.

3. Acyclicity Constraints

The earliest work that transfers the discrete acyclicity constraint in Problem (3) to a continuous differentiable constraint is NOTEARS (Zheng et al., 2018). It introduced the following acyclicity constraint based on the trace-exponential of the weighted adjacency matrix $W \in \mathbb{R}^{d\times d}$:

$$h_{\text{exp}}(W) = \operatorname{Tr}(\exp(W \circ W)) - d,$$

where $\circ$ is the Hadamard product. Since $W \circ W$ has nonnegative entries, we may consider h_{exp} also as a function of nonnegative matrices $A \in \mathbb{R}_{\geq 0}^{d\times d}$, $A = W \circ W$. Using the fact that $x \mapsto \exp(x)$ can be represented by uniformly over $\mathbb{R}$ convergent power series and by the linearity of the trace one can generalize the acyclicity constraint to the so-called *Power Series Trace Family* (PST) (see also Nazaret et al., 2023, Def. 1). The members of the family can be expressed as follows:

$$h(A) = \sum_{k=1}^\infty a_k \operatorname{Tr}[A^k],$$

where $(a_k)_{k\in\mathbb{N}^*} \in \mathbb{R}_{\geq 0}^{\mathbb{N}^*}$, and $A \in \mathbb{R}_{\geq 0}^{d\times d}$. Table 1 lists some examples.

Name	α_k	h_a
h_{expm}	$\frac{1}{k!}$	$\operatorname{Tr}\exp(A) - d$
h_{log}	$\frac{1}{k}$	$-\log\det(I_d - A)$
h_{inv}	1	$\operatorname{Tr}(I_d - A)^{-1}$
h_{binom}	$\binom{d}{k}$	$\operatorname{Tr}(I_d + A)^d - d$

Table 1: PST acyclicity constraints

Name	h_a
h_{spectral}	$\lvert\lambda_d(A)\rvert$
h_{ldet}	$-\log\det(sI_d - A) + d\log s$

Table 2: Spectral-based acyclicity constraints

Another class of constraints, the *spectral-based acyclicity constraints*, was developed based on the fact that the spectral radius of a non-negative weighted adjacency matrix A is zero if and only if A corresponds to a DAG (Nazaret et al., 2023; Bello et al., 2022). Table 2 gives examples.

Nazaret et al. (2023) introduced three criteria for constraints to ensure stable optimization behavior and showed that the PTS acyclicity constraints are not stable during the optimization process. Bello et al. (2022) studied the acyclicity constraint h_{ldet}, based on the log-determinant of the weight-matrix A. More precisely, for $s > 0$, consider space of matrices $\mathbb{W}^s := \{A \in \mathbb{R}^{d\times d} : s > \rho(A)\}$ and define $h^s_{\text{ldet}} : \mathbb{W}^s \to \mathbb{R}$ as

$$h^s_{\text{ldet}}(A) := -\log\det(sI_d - A) + d\log s. \tag{5}$$

Bello et al. (2022) show favorable performance of h_{ldet} in comparison to the constraints h_{exp} and h_{poly}. We will adopt the constraint h_{ldet} in the sequel. We explain our choice from the optimization perspective in the following theorem.

Theorem 1 *The constraint function $h^s_{ldet}(\cdot)$ from (5) satisfies the following stability properties: For all $A \in \mathbb{R}^{d\times d}_{\geq 0}$ it holds:*

- ***V-stable*** *If $h^s_{ldet}(A) \neq 0$, then for $\varepsilon \to 0^+$, $h^s_{ldet}(\varepsilon A) \geq c\varepsilon$ for some positive constant c.*
- ***D-stable*** *$h^s_{ldet}(A)$ and $\nabla h^s_{ldet}(A)$ are well-defined where $\nabla h^s_{ldet}(A) = (sI_d - A)^{-T}$, with $\nabla h^s_{ldet}(A) = 0$ if and only if A is a DAG.*

Finally, h^s_{ldet} is an acyclicity constraint in the sense that

$$h^s_{ldet}(A) \geq 0\,,\ \textit{with}\ h^s_{ldet}(A) = 0\ \textit{if and only if graph generated by matrix}A\ \textit{is a DAG.}$$

No PST constraint is V-stable for $d \geq 2$, i.e., there exists $A \in \mathbb{R}^{d\times d}$, $h(\varepsilon A) = \mathcal{O}(\varepsilon^d)$ as $\varepsilon \to 0$, $\varepsilon > 0$ (Nazaret et al., 2023, Theorem 2). In contrast, V-stability ensures h^s_{ldet} does not vanish rapidly to 0, and D-stability ensures that h^s_{ldet} and its gradient exists. Although h_{spectral} is both V-stable and D-stable (Nazaret et al., 2023), we observe that in practice, $h_{\text{spectral}}(W)$ and $h^s_{\text{ldet}}(W)$ operate on distinctly different scales, with $h_{\text{spectral}}(W)$ being substantially larger than $h^s_{\text{ldet}}(W)$. This considerable difference in magnitude raises challenges in assessing whether $h_{\text{spectral}}(W)$ can be regarded as sufficiently small to be considered zero. Thus, we choose $h^s_{\text{ldet}}(W)$.

4. Overall Learning Objective

4.1. Sparsity Regularizer

In applications, we often expect the random variables X_j to have only a few parent variables and, thus, invoke a sparsity regularizer. Recall that we assume that the functions f_j: $\mathcal{X} \to \mathbb{R}$ are in the

class $C^1(\mathcal{X})$, and the functions f_j and their derivatives are both square-integrable. We consider the regularizer (compare also Rosasco et al. 2013)

$$\Omega_1(f_j) = \sum_{k=1}^{d} \left\| \frac{\partial f_j(\cdot)}{\partial x_k} \right\|_{L_2} = \sum_{k=1}^{d} \sqrt{\int_{\mathcal{X}} \left(\frac{\partial f_j(x)}{\partial x_k} \right)^2 \mathbb{P}_X(dx)}. \tag{6}$$

To develop a data-based decision rule we need to consider an empirical counterpart for the $\|\cdot\|_{L^2}$ norm of the derivative. Thus, we form an empirical estimate based on the data $\mathbb{X} = \{x^i\}_{i=1}^n$, by setting

$$\left\| \frac{\partial f_j(\cdot)}{\partial x_k} \right\|_n := \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\frac{\partial f_j(x^i)}{\partial x_k} \right)^2}.$$

Then the empirical estimate of (6) is

$$\Omega_1^{\mathcal{D}}(f_j) = \sum_{k=1}^{d} \left\| \frac{\partial f_j(\cdot)}{\partial x_k} \right\|_n.$$

Similarly, the empirical estimate of the coefficient W_{jk} of the weighted adjacency matrix is

$$W_{kj}^{\mathcal{D}} = \left\| \frac{\partial f_j(\cdot)}{\partial x_k} \right\|_n.$$

4.2. Constrained Empirical Optimization Problem Solved by Kernel Methods

For each $j \in [d]$, we further assume $f_j \colon \mathcal{X} \to \mathbb{R}, j \in [d]$ to be in a reproducing kernel Hilbert space (RKHS) H generated by a bounded continuously-differentiable kernel k on $\mathcal{X}$ and use the term $\lambda \|f_j\|_H^2$ to penalize function complexity. Then we aim to minimize the following loss function:

$$\sum_{j=1}^{d} \frac{1}{2n} \ell(\mathbb{X}^j, f_j(\mathbb{X})) + \tau(2\Omega_1^{\mathcal{D}}(f_j) + \lambda \|f_j\|_H^2) \text{ s.t. } h_{\text{ldet}}^s(W^{\mathcal{D}}) = 0, \tag{7}$$

where τ, λ are positive numbers and $s > 0$ is some fixed number (typically set to 1). Following Rosasco et al. (2013), we show a version of the RKHS Representer Theorem for problem (7) with the log-determinant acyclicity constraint. The main point of the result below is to show that the solution of the log-determinant constrained empirical minimization problem (7) admits the form of a finite linear combination, for which we may then optimize the weights.

Theorem 2 *Let $\mathcal{X}$ be a bounded connected non-empty open set in $\mathbb{R}^d$, $k(\cdot,\cdot)$ be a bounded countiniously differentiable kernel. Then the constrained minimizer of (7) can be written as*

$$\widehat{f}_j^{\,\tau}(x) = \sum_{i=1}^{n} \alpha_i^j k(x, x^i) + \sum_{i=1}^{n} \sum_{a=1}^{d} \beta_{ai}^j \frac{\partial k(x,s)}{\partial s_a}\bigg|_{s=x^i}, \quad x \in \mathcal{X}, \tag{8}$$

where $\alpha^j, (\beta_{ai}^j)_{i=1}^n \in \mathbb{R}^n$ and $a, j \in [d]$. Then,

$$\left\| \widehat{f}_j^{\,\tau} \right\|_H^2 = \sum_{i,l=1}^{n} \alpha_i^j \alpha_l^j k(x^i, x^l) + 2 \sum_{i,l=1}^{n} \sum_{a=1}^{d} \alpha_i^j \beta_{al}^j \frac{\partial k(x^i, x^l)}{\partial x_a^l} + \sum_{i,l=1}^{n} \sum_{a,b=1}^{d} \beta_{ai}^j \beta_{bl}^j \frac{\partial k(x^i, x^l)}{\partial x_a^i \partial x_b^l}. \tag{9}$$

Via Theorem 2, we may estimate every function f_j by a kernel estimator. As no random variable should cause itself, f_j should not depend on x_j as its input. Thus, to estimate f_j we replace k in (8) with a restricted kernel k^{-j} that depends only on the subvector $(x_k)_{k\neq j}$ and use the representation

$$\widehat{f}_j(x) = \sum_{i=1}^{n} \alpha_i^j k^{-j}(x, x^i) + \sum_{i=1}^{n}\sum_{a=1}^{d} \beta_{ai}^j \frac{\partial k^{-j}(x,s)}{\partial s_a}\bigg|_{s=x^i}.$$

Let $\theta_j = \{\alpha^j, \beta_{ai}^j : a \in [d], i \in [n]\}$ denote the parameters for each f_j, and let $\theta = (\theta_1, \ldots, \theta_d)$. Then the loss function is constructed as follows:

$$\sum_{j=1}^{d} \frac{1}{2n} \ell(\mathbb{X}^j, \widehat{f}_j^{\,\theta}(\mathbb{X})) + \tau[2\Omega_1^{\mathcal{D}}(\widehat{f}_j^{\,\theta}) + \lambda \|\widehat{f}_j^{\,\theta}\|_H^2]. \tag{10}$$

To evaluate the acyclicity constraint on the dataset, using the Representer Theorem for the function $\widehat{f}_j$, we consequently obtain for every $k, j \in [d]$:

$$\begin{aligned} W_{kj}^{\mathcal{D}}(\widehat{f}_j^{\,\theta}) = W^{\mathcal{D}}(\theta)_{kj} &= \left\| \frac{\partial \widehat{f}_j^{\,\theta}(x)}{\partial x_k} \right\|_n \\ &= \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[\sum_{l=1}^{n} \alpha_l^j \frac{\partial k^{-j}(x^i, x^l)}{\partial x_k^i} + \sum_{l=1}^{n}\sum_{a=1}^{d} \beta_{al}^j \frac{\partial k^{-j}(x^i, x^l)}{\partial x_k^i \partial x_a^l} \right]^2 \right\}^{\frac{1}{2}}, \end{aligned} \tag{11}$$

and consequently

$$\Omega_1^{\mathcal{D}}(\widehat{f}_j^{\,\theta}) = \sum_{k=1}^{d} \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left(\frac{\partial \widehat{f}_j^{\,\theta}(x^i)}{\partial x_k} \right)^2} = \sum_{k=1}^{d} W^{\mathcal{D}}(\theta)_{kj}. \tag{12}$$

As a default, we base our algorithm on the Gaussian kernel $k_\gamma(x, x') = \exp(-\frac{1}{\gamma^2}\|x - x'\|^2)$, based on the Euclidean distance between x and x'. For $\gamma > 0$, $j \in [d]$, we define the restricted *Gaussian kernel* $k_\gamma^{-j} : \mathbb{R}^{d-1} \mapsto \mathbb{R}$ as

$$k_\gamma^{-j}(x, x') := \exp\left(-\frac{1}{\gamma^2} \sum_{i\neq j}^{d} (x_i - x_i')^2 \right). \tag{13}$$

It corresponds to the Gaussian kernel $k_\gamma(x, x')$ when fixing the j-th coordinate to a constant.

Remark 3 *We want the decision rule returned by our estimation algorithm to belong to the space of continuously differentiable functions, as this guarantees that a vanishing partial derivative implies that the function does not depend on the considered coordinate. To this end, we claim that any function that belongs to Gaussian RKHS $H_\gamma(\mathcal{X})$ also belongs to $C^1(\mathcal{X})$. To see this, let H_γ be the RKHS of the real-valued Gaussian RBF kernel k_γ for $\gamma > 0$. Since $k_\gamma(\cdot,\cdot)$ is (infinitely many times) differentiable, Theorem 10.45 in Wendland (2004) implies that the associated RKHS $H_\gamma(\mathcal{X})$ is a subset of the space of continuously differentiable functions. Therefore, $H_\gamma(\mathcal{X}) \subset C^1(\mathcal{X})$, which implies that $\|f\|_{C^1(\mathcal{X})} < \infty$ for every $f \in H_\gamma$.*

5. Optimization

Combining Equations (10), (11), (12) together with least squares loss and log-determinant acyclicity constraint, we obtain the following constrained empirical optimization problem:

$$\min_{\theta} \sum_{j=1}^{d} \left\{ \frac{1}{2n} \left\| \mathbb{X}^j - \widehat{f}_j^{\,\theta}(\mathbb{X}) \right\|_2^2 + \tau[2\Omega_1^{\mathcal{D}}(\widehat{f}_j^{\,\theta}) + \lambda \|\widehat{f}_j^{\,\theta}\|_H^2] \right\}$$
$$\text{s.t. } -\log\det(sI_d - W^{\mathcal{D}}(\theta) \circ W^{\mathcal{D}}(\theta)) + d\log(s) = 0.$$

As in DAGMA, we use a central path method to solve the constrained optimization problem. We give our method in Algorithm 1 and refer to it as RKHS-DAGMA. Note that line 3 of Algorithm 1 means that starting at $\theta = \theta^{(t)}$, $\theta^{(t+1)}$ is obtained by the ADAM optimizer (Kingma and Ba, 2015).

Algorithm 1: RKHS-DAGMA

Input: Data matrix $\mathbb{X}$, initial coefficient (learning step) $\mu^{(0)}$ (e.g., 1), decay factor $\alpha \in (0, 1)$(e.g., 0.1), sparsity parameter τ(e.g., 1×10^{-4}), function complexity parameter λ (e.g., 1×10^{-3}), log-det parameter $s > 0$ (e.g., 1), number of iterations T (e.g., 6), threshold ω (e.g., 0.1).

Output: $\widehat{W}$, the estimated weighted adjacency matrix.

1 Initialize $\theta^{(0)}$ so that $W^{\mathcal{D}}(\theta^{(0)}) \in \mathbb{W}^s$.
2 **for** $t \leftarrow 0$ **to** $T - 1$ **do**
3 Starting at $\theta^{(t)}$, solve $\theta^{(t+1)} =$
$\arg\min_\theta \mu^{(t)}(\sum_{j=1}^{d}\{\frac{1}{2n}\|\mathbb{X}^j - \widehat{f}_j^{\,\theta}(\mathbb{X})\|_2^2 + \tau[2\Omega_1^{\mathcal{D}}(\widehat{f}_j^{\,\theta}) + \lambda\|\widehat{f}_j^{\,\theta}\|_H^2]\}) + h^s_{\text{ldet}}(W^{\mathcal{D}}(\theta))$.
4 Set $\mu^{(t+1)} = \alpha\mu^{(t)}$.
5 **end**
6 Threshold matrix $\widehat{W} = W^{\mathcal{D}}(\theta^{(T)}) \cdot \mathbb{1}(W^{\mathcal{D}}(\theta^{(T)}) > \omega)$.

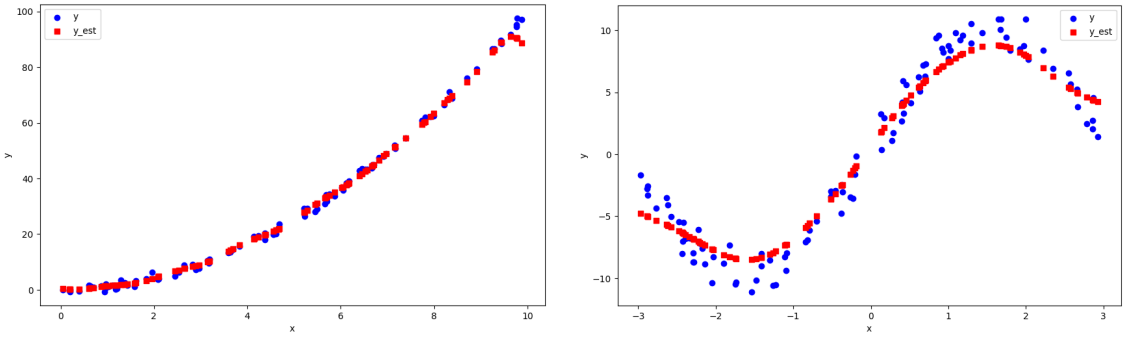
6. Experiments

This section is divided into three parts. First, we analyze the performance of RKHS-DAGMA in a simple bivariate prediction setting, distinguishing cause-effect within artificially constructed toy models. Second, we evaluate and compare the properties of RKHS-DAGMA with non-parametric NOTEARS algorithms, including NOTEARS-MLP and NOTEARS-SOB, as well as the score-based FGES algorithm (Ramsey, 2015) on sampled directed Erdős-Rényi graphs of growing dimension. Finally, we assess the performance of RKHS-DAGMA against NOTEARS-MLP, NOTEARS-SOB, and FGES on real-world bivariate datasets (Mooij et al., 2016).

6.1. Toy Example

We illustrate the performance of RKHS-DAGMA by two simple simulations with $d = 2$ nodes and $n = 100$ data points. Figure 1 plots the ground truth data points (blue) and estimated function values obtained by RKHS-DAGMA (red) in the bivariate causal models (a) $Y = X^2 + \varepsilon$, $X \sim \mathcal{U}[0, 10]$ and (b) $Y = 10\sin(X) + \varepsilon$, $X \sim \mathcal{U}[-3, 3]$. Let W_{est} be the estimated weighted adjacency matrix without any thresholding. The results of Figure 1(*a*) and Figure 1(*b*) correspond to the estimated matrices

$$W_{\text{est}} = \begin{pmatrix} 0 & 10.35 \\ 6.22 \times 10^{-4} & 0 \end{pmatrix}, \qquad W_{\text{est}} = \begin{pmatrix} 0 & 4.91 \\ 8.49 \times 10^{-4} & 0 \end{pmatrix},$$



(*a*) quadratic relationship (*b*) sine relationship

Figure 1: Illustrations of RKHS-DAGMA by toy examples with two nodes.

respectively. In both cases, W_{12} is large, whereas W_{21} is small enough to be ignored after thresholding, indicating that RKHS-DAGMA finds correct causal relationships.

6.2. Structure Learning

Next, we compare RKHS-DAGMA to non-parametric NOTEARS methods by comparing the estimated DAG with the ground truth, generated as Erdős-Rényi (ER) directed graphs with given topological ordering. Zheng et al. (2020) consider several graph models, including ER and scale-free graphs. As one of the hardest settings (Zheng et al., 2020, Fig. 4), we take up ER4 graphs for which NOTEARS algorithms are less competitive compared to algorithms like fast greedy equivalence search (Ramsey et al., 2017), DAG-GNN (Yu et al., 2019), or greedy equivalence search with generalized scores (Huang et al., 2018). ERm ($m = 4$) denotes an ER graph with $m \times d$ edges.

Given the ground truth, we simulate from an SEM with $X_j = g_j(X_{\mathrm{pa}(j)}) + \varepsilon_j$ with $\varepsilon_j \sim \mathcal{N}(0, 1)$. We consider the following functional relationships for g_j: Additive models with Gaussian processes, Gaussian processes, and MLPs according to the procedure of Zheng et al. (2020). We also add a simulation type called the combinatorial model, which is an additive model with non-linear relationships random picked from the following common non-linear functions: $g(x) = \exp(-|x|)$, $g(x) = 0.05 \cdot x^2$, $g(x) = \sin(x)$. We refer to Appendix B for comprehensive details on the simulations.

As noted in Bello et al. (2022), the initial point $W(\theta^{(0)})$ is required to be inside $\mathbb{W}^s$. Since the zero matrix is always inside $\mathbb{W}^s$ for any $s > 0$, we set the parameters $\theta^{(0)}$ be 0. Given that our approximation method and sparsity regularizer fundamentally differ from those in NOTEARS algorithms, the hyperparameters λ and τ are tuned via grid search. In RKHS-DAGMA, we take sparsity parameter $\tau = 1 \times 10^{-4}$, function complexity parameter $\lambda = 1 \times 10^{-3}$, and threshold $\omega = 0.1$. Additionally, we take $\mu^{(0)} = 1$ and the default value $T = 6$; if the resulting weighted adjacency matrix is not a DAG, we enhance T to 7. We set $\gamma = 0.4d$ for the Gaussian kernel (see Appendix B.2 for an explanation of this choice). Due to the explicit computation of derivatives and the Hessian of the kernel function, we limit the maximum number of iterations of the ADAM optimizer to 10% of corresponding values in DAGMA to compensate for the additional cost. For the NOTEARS algorithms, we choose the default hyperparameters as described in Waxman et al. (2024) and Zheng et al. (2020).

To evaluate model performance, we use the structural Hamming distance (SHD), which measures the total number of edge additions, deletions, and reversals needed to convert the estimated graph into the true graph. A lower SHD indicates better model performance.

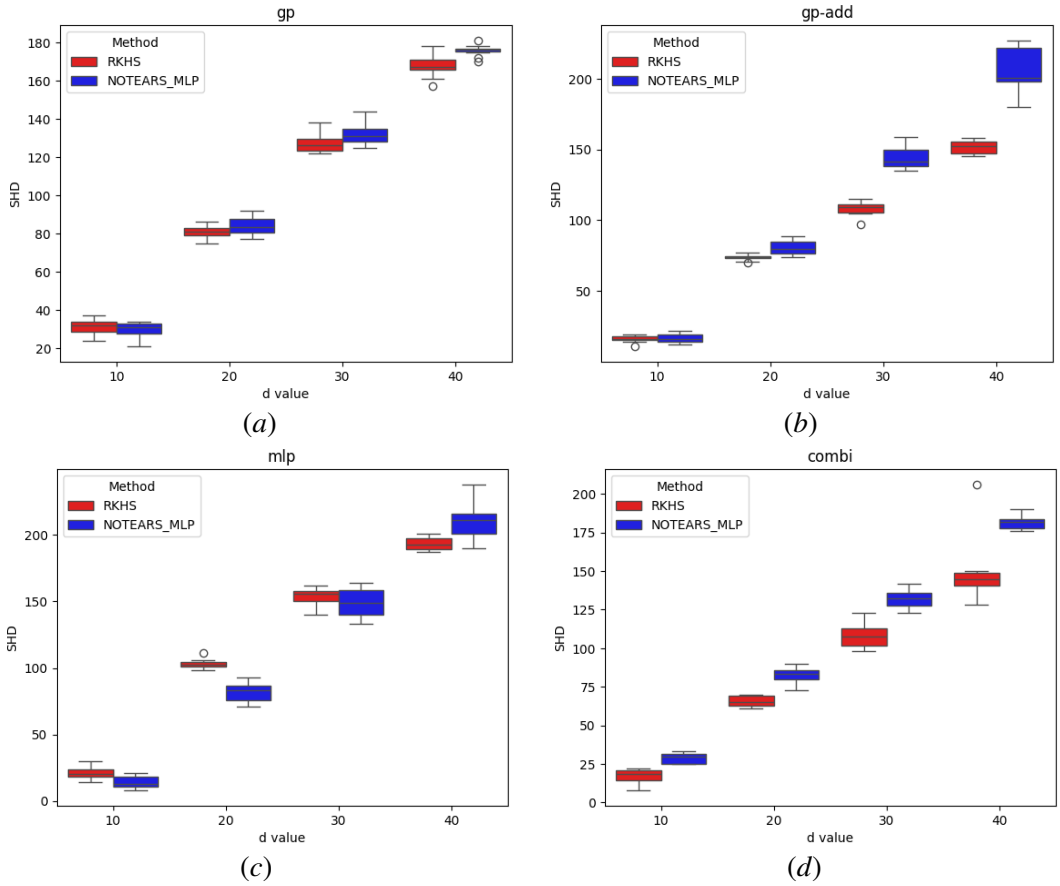


Figure 2: Comparison between RKHS-DAGMA and NOTEARS-MLP by SHD (lower is better) for random data generated from 2(*a*) the ER-4 GP model, 2(*b*) the ER-4 GP-additive model, 2(*c*) the ER-4 MLP model, 2(*d*) the model with combination of functions. Boxplots show the median and quartiles across 10 different simulations for each simulation model.

Our results indicate that RKHS-DAGMA consistently outperforms NOTEARS-SOB in terms of structured Hamming distance (SHD). Additionally, compared to NOTEARS-MLP, RKHS-DAGMA shows superior performance in simulations based on Gaussian processes (GP), additive GP, and combinatorial models, while maintaining competitive results in MLP experiments (see Figure 2-3).

We also conduct a comparative analysis between RKHS-DAGMA and the FGES algorithm, implemented through the `py-causal` package [2]. Figure 4 indicates that RKHS-DAGMA significantly outperforms the FGES algorithm in simulations based on GP, additive GP, and combinatorial models for $d = 10$. It achieves comparable performance $d \geq 20$. However, in the MLP simulations, the performance of RKHS-DAGMA clearly falls behind for $d \geq 20$.

6.3. Real Data

Finally, we compare the performance of RKHS-DAGMA with NOTEARS-MLP, NOTEARS-SOB, and the FGES on a benchmark collection of datasets featuring cause-effect pairs (Mooij et al., 2016). These datasets are bivariate, each consisting of one pair of statistically dependent variables. We excluded six datasets containing multi-dimensional random variables and standardized the remaining

2. https://github.com/bd2kccd/py-causal

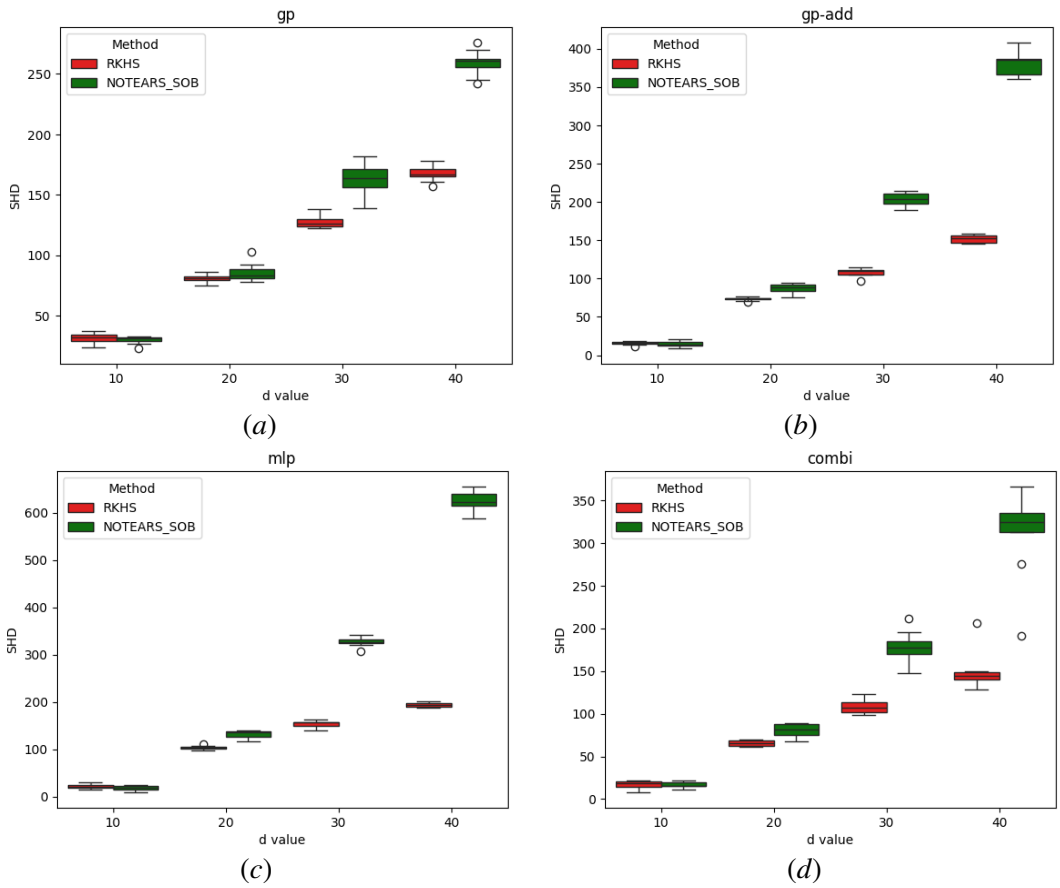


Figure 3: Comparison between RKHS-DAGMA and NOTEARS-SOB by SHD (lower is better) for random data generated from 3(*a*) the ER-4 GP model, 3(*b*) the ER-4 GP-additive model, 3(*c*) the ER-4 MLP model, 3(*d*) the model with combination of functions. Boxplots show the median and quartiles across 10 different simulations for each simulation model.

datasets. To reduce computational costs, datasets with more than 400 samples were organized in ascending order according to the first variable. These datasets were then partitioned into 300 grids according to the first variable, with the median value of the first variable in each grid and the corresponding value of the second variable used for model evaluation.

RKHS-DAGMA achieves the best accuracy of 55.88% among the remaining 102 datasets, while NOTEARS-SOB and NOTEARS-MLP achieve an accuracy of 45.10% and 0.98% correspondingly. We attribute the bad performance of NOTEARS-MLP to the small sample size with a relatively large number of hidden units compared to the number of nodes, and to the specific definition of the weighted adjacency matrix, which depends on the weights of the first hidden layer and may differ significantly from those defined by derivatives (Waxman et al., 2024). Furthermore, the FGES algorithm yields only undirected edges across all bivariate datasets, as it employs a conditional independence test to eliminate unnecessary edges. In the context of bivariate data, where each variable pair lacks additional conditioning variables, the algorithm's capacity to extract information is inherently limited.

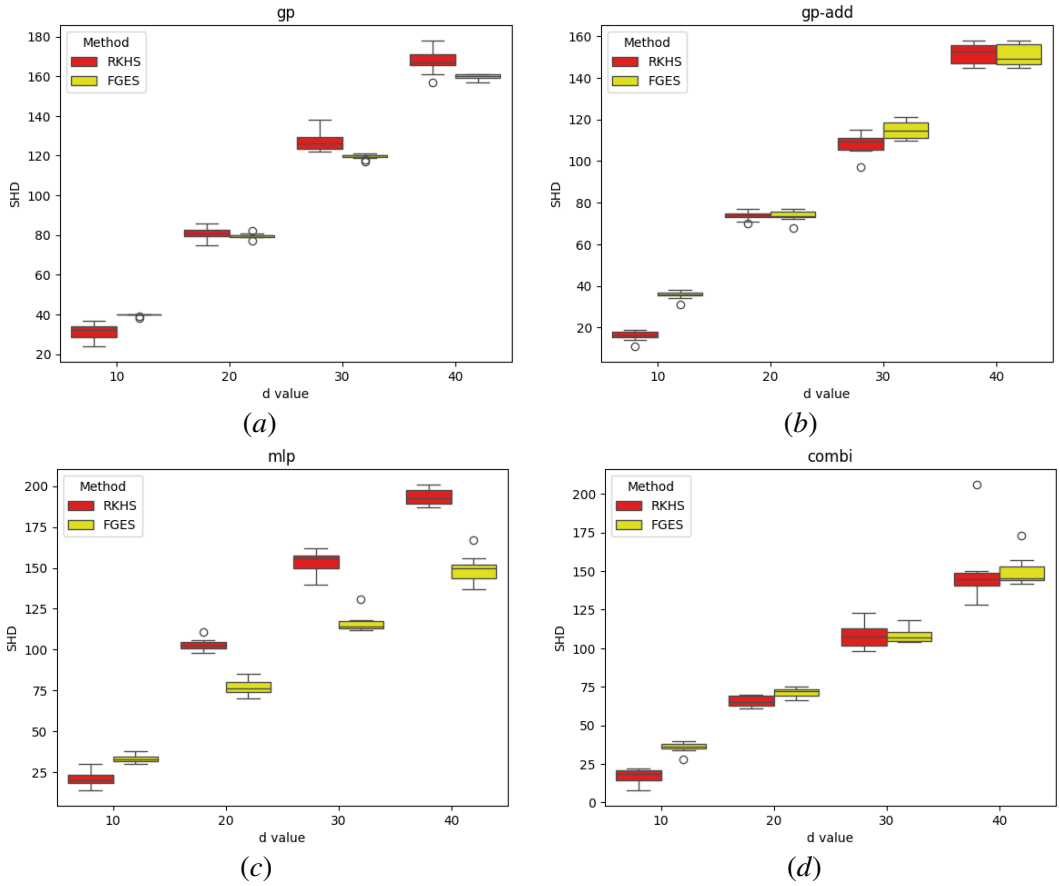


Figure 4: Comparison between RKHS-DAGMA and FGES by SHD (lower is better) for random data generated from 4(*a*) the ER-4 GP model, 4(*b*) the ER-4 GP-additive model, 4(*c*) the ER-4 MLP model, 4(*d*) the model with combination of functions. Boxplots show the median and quartiles across 10 different simulations for each simulation model.

7. Conclusion

In this work, we addressed the non-parametric DAG learning problem using a procedure that exploits the machinery of infinite-dimensional (Gaussian) RKHS. We showed in Theorem 2 that the RKHS-DAGMA Algorithm, which solves a (combined) constrained empirical optimization problem with log-determinant acyclicity constraint, admits an explicit solution as a finite-dimensional representation of the kernel elements of the data and their derivatives. Furthermore, this solution can be computed using central path methods similar to those in the DAGMA algorithm. We compared the accuracy of RKHS-DAGMA with the known baselines under non-parametric structural equation modeling, such as NOTEARS-MLP and NOTEARS-SOB (see ex. Zheng et al., 2020), as well as the score-based FGES algorithm (Ramsey, 2015) under settings comparable to those in Zheng et al. (2020). The RKHS-DAGMA Algorithm demonstrates utility across both simulated and empirical datasets.

Acknowledgments

Yurou Liang is supported by the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research.

References

K. Bello, B. Aragam, and P. Ravikumar. DAGMA: Learning DAGs via m-matrices and a log-determinant acyclicity characterization. *Advances in Neural Information Processing Systems*, 35: 8226–8239, 2022.

D. M. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

M. Drton and M. H. Maathuis. Structure learning in graphical modeling. *Annu. Rev. Stat. Appl.*, 4: 365–393, 2017.

C. Giraud. *Introduction to high-dimensional statistics*. Chapman and Hall/CRC, 2021.

K. Göbler, T. Windisch, M. Drton, T. Pychynski, M. Roth, and S. Sonntag. `causalAssembly`: Generating realistic production data for benchmarking causal discovery. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pages 609–642. PMLR, 01–03 Apr 2024.

L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer-Verlag, New York, 2002.

D. Heckerman, D. Geiger, and D. M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.

B. Huang, K. Zhang, Y. Lin, B. Schölkopf, and C. Glymour. Generalized score functions for causal discovery. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1551–1560, 2018.

Q. Ji, E. Bouri, R. Gupta, and D. Roubaud. Network causality structures among bitcoin and other financial assets: A directed acyclic graph approach. *The Quarterly Review of Economics and Finance*, 70:203–213, 2018.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

M. Maathuis, M. Drton, S. Lauritzen, and M. Wainwright, editors. *Handbook of graphical models*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, FL, 2019.

D. Margaritis and S. Thrun. Bayesian network induction via local neighborhoods. *Advances in Neural Information Processing Systems*, 12:505–511, 1999.

J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. Distinguishing cause from effect using observational data: Methods and benchmarks. *Journal of Machine Learning Research*, 17(32):1–102, 2016.

A. Nazaret, J. Hong, E. Azizi, and D. Blei. Stable differentiable causal discovery. *arXiv preprint arXiv:2311.10263*, 2023.

I. Ng, A. Ghassami, and K. Zhang. On the role of sparsity and DAG constraints for learning linear DAGs. *Advances in Neural Information Processing Systems*, 33:17943–17954, 2020.

J. Ramsey, M. Glymour, R. Sanchez-Romero, and C. Glymour. A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International Journal of Data Science and Analytics*, 3:121–129, 2017.

J. D. Ramsey. Scaling up greedy equivalence search for continuous variables. *CoRR*, abs/1507.07749, 2015. URL http://arxiv.org/abs/1507.07749.

L. A. Rosasco, S. Villa, S. Mosci, M. Santoro, and A. Verri. Nonparametric sparsity and regularization. *Journal of Machine Learning Research*, 14(1):1665–1714, 2013.

W. Rudin. *Functional analysis*, volume 2. McGraw-Hil, 1991.

B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. In D. P. Helmbold and R. C. Williamson, editors, *Computational Learning Theory, 14th Annual Conference on Computational Learning Theory, COLT 2001 and 5th European Conference on Computational Learning Theory, EuroCOLT 2001, Amsterdam, The Netherlands, July 16-19, 2001, Proceedings*, volume 2111 of *Lecture Notes in Computer Science*, pages 416–426. Springer, 2001.

P. Spirtes and C. Glymour. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9(1):62–72, 1991.

P. Spirtes and K. Zhang. Search for causal models. In *Handbook of graphical models*, Chapman & Hall/CRC Handb. Mod. Stat. Methods, pages 439–469. CRC Press, Boca Raton, FL, 2019.

I. Steinwart and A. Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

I. Tsamardinos, C. F. Aliferis, and A. R. Statnikov. Algorithms for large scale Markov blanket discovery. In I. Russell and S. M. Haller, editors, *Proceedings of the Sixteenth International Florida Artificial Intelligence Research Society Conference, May 12-14, 2003, St. Augustine, Florida, USA*, pages 376–381. AAAI Press, 2003.

D. Waxman, K. Butler, and P. M. Djurić. Dagma-DCE: Interpretable, non-parametric differentiable causal discovery. *IEEE Open Journal of Signal Processing*, 2024.

H. Wendland. *Scattered data approximation*, volume 17. Cambridge university press, 2004.

Y. Yu, J. Chen, T. Gao, and M. Yu. DAG-GNN: DAG structure learning with graph neural networks. In *International Conference on Machine Learning*, pages 7154–7163. PMLR, 2019.

J. Zhang, L. Cammarata, C. Squires, T. P. Sapsis, and C. Uhler. Active learning for optimal intervention design in causal models. *Nature Machine Intelligence*, 5(10):1066–1075, 2023.

X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing. DAGs with NO TEARS: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31: 9492–9503, 2018.

X. Zheng, C. Dan, B. Aragam, P. Ravikumar, and E. Xing. Learning sparse nonparametric DAGs. In *International Conference on Artificial Intelligence and Statistics*, pages 3414–3425. PMLR, 2020.

D.-X. Zhou. Derivative reproducing properties for kernel methods in learning theory. *Journal of Computational and Applied Mathematics*, 220(1-2):456–463, 2008.

Appendix A. Detailed Proofs

A.1. Proof of Theorem 1

- **D-stable:** Follows from Theorem 1 in Bello et al. (2022).
- **V-stable:** Note that

$$\log\det(sI - A) - d\log s = \log(s^d \det(I - s^{-1}A)) - d\log s = \log\det(I - s^{-1}A).$$

Since $s > \rho(A)$ is equivalent to $1 > \rho(s^{-1}A)$, we consider the case $s = 1$ w.l.o.g. For any $\varepsilon > 0$ with $\varepsilon|\rho(A)| \leq 1$, it holds that

$$h^s_{\text{ldet}}(\varepsilon A) = -\log\det(I - \varepsilon A) = -\log(\prod_{i=1}^{d} \lambda_i(I - \varepsilon A)) = \sum_{i=1}^{d} -\log(1 - \varepsilon\lambda_i(A)).$$

We prove the statement of the Theorem under the assumption that the eigenvalues $\{\lambda_i(A)\}_{i\in[d]}$ of A are complex numbers and under the additional assumption that $\sum_{i=1}^{d} \lambda_i(A) \neq 0$.

Applying the Cauchy-integral formula to the principal branch of the complex logarithm function $z \mapsto \log(1 - z)$, which is analytic within domain $|z| < 1$, we get:

$$-\log(1 - \varepsilon\lambda_i(A)) = \varepsilon\lambda_i(A) + \frac{\varepsilon^2\lambda_i^2(A)}{2\pi i}\int_{\eta_0} \frac{\log(1-w)}{(\varepsilon\lambda_i(A) - w)w^2}\, dw,$$

where η_0 is any closed circle of radius r_0 with $\varepsilon|\lambda_i(A)| < r_0 < 1$. Therefore, summing over all complex values $\lambda_i(A)$ and using triangle inequality $|a - b| \geq |a| - |b|$ for $a, b \in \mathbb{C}$, we deduce that

$$\left|\sum_{i=1}^{d} -\log(1 - \varepsilon\lambda_i(A))\right| \geq \varepsilon\left|\sum_{i=1}^{d}\lambda_i(A)\right| - \frac{1}{2\pi}\left|\sum_{i=1}^{d}\varepsilon^2\lambda_i^2(A)\int_{\eta_0}\frac{\log(1-w)}{(w - \varepsilon\lambda_i(A))w^2}\, dw\right|.$$

Since $w \mapsto \frac{\log(1-w)}{w-\varepsilon\lambda_i(A)}$ is continuous, Weierstrass Theorem implies that the function is bounded on the bounded domain η_0, yielding that $\left|\frac{\log(1-w)}{w-\varepsilon\lambda_i(A)}\right| \leq K$ for some $K > 0$. Since $\frac{\varepsilon|\lambda_i(A)|}{r_0} < 1$ and η_0 is a circle of radius r_0, the ML inequality for the complex integral gives

$$\left|\varepsilon^2\lambda_i^2(A)\int_{\eta_0}\frac{\log(1-w)}{w - \varepsilon\lambda_i(A)w^2}\, dw\right| \leq 2\pi\varepsilon^2\lambda_i^2(A)\frac{K}{r_0},$$

which in turn implies that

$$\begin{aligned}
\left|\sum_{i=1}^{d} -\log(1 - \varepsilon\lambda_i(A))\right| &\geq \varepsilon\left|\sum_{i=1}^{d}\lambda_i(A)\right| - \frac{K}{r_0}\varepsilon^2\left|\sum_{i=1}^{d}\lambda_i^2(A)\right| \\
&= \varepsilon\left|\sum_{i=1}^{d}\lambda_i(A)\right| - \frac{K}{r_0}\varepsilon^2\sum_{i=1}^{d}\lambda_i^2(A) \\
&= \varepsilon\left|\sum_{i=1}^{d}\lambda_i(A)\right| - \frac{K\varepsilon^2\|A\|_2^2}{r_0} \\
&\geq \varepsilon\frac{\left|\sum_{i=1}^{d}\lambda_i(A)\right|}{2},
\end{aligned}$$

where the last inequality holds provided ε is chosen small enough, with

$$\varepsilon \leq \min\left\{\frac{1}{|\rho(A)|}, \frac{r_0\left|\sum_{i=1}^{d}\lambda_i(A)\right|}{2K\|A\|_2^2}\right\}.$$

Thus, we proved the claim.

Finally, notice that Bello et al. (2022) showed that h^s_{ldet} is an acyclicity constraint and the instability of PST acyclicity constraint is shown in Nazaret et al. (2023).

A.2. Proof of Theorem 2

Proof (8): Consider arbitrary elements $f, g \in H$. Since H is complete, there exists a sequence $(g)_{n=1}^{\infty}$, $g_n \in H$, that converges to g in the norm of Hilbert space H. Furthermore, for any $n \in \mathbb{N}$,

$$\langle f, g_n\rangle_H - \langle f, g\rangle_H \leq |\langle f, g_n\rangle_H - \langle f, g\rangle_H| = |\langle f, g_n - g\rangle_H| \overset{\text{Cauchy-Schwarz}}{\leq} \|f\|_H \cdot \|g_n - g\|_H.$$

Since the kernel k is bounded, elements $f \in H$ are bounded (Steinwart and Christmann, 2008, Lemma 4.23). Furthermore, since $g_n \to g$ in the norm of the space H we have that

$$\lim_{n\to\infty}\langle f, g_n\rangle_H - \langle f, g\rangle = 0,$$

which implies that map $x \mapsto \langle \cdot, x\rangle_H$ is continuous. Similarly, one can prove that the scalar product is continuous in the first coordinate.

For coherence of the remainder of the proof, we first refine the proof that for the open set $\mathcal{X} \subset \mathbb{R}^d$, we have that $\frac{\partial}{\partial x_a}k(\cdot, x) \in H$ for every $a \in [d]$ and $x \in \mathcal{X}$, and that a "differential reproducing property" holds, which states that

$$\frac{\partial}{\partial x_a}f(x) = \left\langle f, \frac{\partial}{\partial s_a}k(\cdot, s)\Big|_{s=x}\right\rangle_H, \quad \forall x \in \mathcal{X}, \quad f \in H.$$

Specifically, we refine the proof of Theorem 1 a)-b) in Zhou (2008) in case $\alpha = 0$ to show that the "derivative element" exists in H. We deviate from that proof in the part that establishes the differential reproducing property, the difference being the use of completeness of H and the fact that weak convergence and pointwise convergence are equivalent in RKHS H.

Consider arbitrary $x \in \mathcal{X}$. As $\mathcal{X}$ is open, there exists $r > 0$ such that the ball $\{x + y, y \in \mathbb{R}^d, \|y\|_2 \leq r\} \subset \mathcal{X}$. For $a \in [d]$, let e_a be the a-th orthonormal vector in the standard Euclidean basis in $\mathbb{R}^d$. Let $\tilde{h}_{x,a} : \mathcal{X} \mapsto \mathbb{R}$ be the function given by $\tilde{h}_{x,a}(y) = \frac{\partial}{\partial s_a}k(s, y)|_{s=x}$ for all $y \in \mathcal{X}$. By definition of the RKHS, $k(\cdot, x) \in H$, and the set of functions $H \mapsto \mathbb{R}$ given by

$$\left\{\frac{1}{t}(k(\cdot, x + te_a) - k(\cdot, x)) : |t| \leq r\right\} \tag{14}$$

satisfies that for every $t, |t| \leq r$:

$$\begin{aligned}\left\|\frac{k(\cdot, x + te_a) - k(\cdot, x)}{t}\right\|_H^2 &= \frac{1}{t^2}\Big(k(x + te_a, x + te_a) - k(x, x + te_a) - k(x + te_a, x) + k(x, x)\Big)\\ &\leq \left\|\frac{\partial^2}{\partial x_a \partial x_a}k(\cdot, \cdot)\right\|_{\infty}^2,\end{aligned}$$

where the last inequality follows from the fact that $k(\cdot,\cdot)$ is a continuously differentiable function in every coordinate and an application of the Mean Value Theorem (twice, once in every coordinate). The latter inequality implies that set (14) lies within a ball of radius $\left\|\frac{\partial^2}{\partial x_a \partial x_a}k(\cdot,\cdot)\right\|_\infty$ in Hilbert space H. It is known (since $*$-weak convergence is equivalent to weak convergence in Hilbert spaces) that the ball in the Hilbert space is weakly sequentially compact; see, e.g., (Rudin, 1991, Chap. 3). Thus, there exists a sequence (t_n) such that $\lim_{n\to\infty} t_n = 0$ and the sequence $\frac{1}{t_n}(k(\cdot, x + t_n e_a) - k(\cdot, x))$ converges weakly to an element $h_x \in H$. The latter means that

$$\lim_{n\to\infty}\left\langle \frac{1}{t_n}(k(\cdot, x + t_n e_a) - k(\cdot, x)), f\right\rangle_H = \langle h_x, f\rangle_H \qquad \forall f \in H. \tag{15}$$

Consider $f(\cdot) = k(\cdot, y)$, where $y \in \mathcal{X}$ is arbitrary. Since $k(\cdot, y)$ is differentiable (as a function of the first coordinate), the left-hand side of (15) satisfies

$$\begin{aligned}\lim_{n\to\infty}\left\langle \frac{1}{t_n}(k(\cdot, x + t_n e_a) - k(\cdot, x)), k(\cdot, y)\right\rangle_H &= \lim_{n\to\infty}\frac{1}{t_n}(k(x + t_n e_a, y) - k(x, y)) \\ &= \frac{\partial}{\partial s_a}k(s, y)\Big|_{s=x} = \tilde{h}_{x,a}(y).\end{aligned}$$

But for the right-hand side of (15), it holds that

$$\lim_{n\to\infty}\left\langle \frac{1}{t_n}(k(\cdot, x + t_n e_a) - k(\cdot, x)), k(\cdot, y)\right\rangle_H = \langle h_x, k(y,\cdot)\rangle = h_x(y) \qquad \forall y \in \mathcal{X}.$$

We conclude that $\tilde{h}_{x,a} = h_x$ as a map $\mathcal{X} \mapsto \mathbb{R}$, and since h_x is in H, so is $\tilde{h}_x$. By identifying $\frac{\partial}{\partial x_a}k(\cdot, x) := h_x$, the existence of an element y with $\frac{\partial}{\partial x_a}k(x, y) = \left\langle \frac{\partial}{\partial x_a}k(x,\cdot), k(\cdot, y)\right\rangle$ follows.

Now we show the differentiable reproducing property, that is, the convergence to the limit $\frac{\partial}{\partial x_a}(k(\cdot, x))$ is pointwise (and not only as a weak limit) and we can exchange the differential and inner product sign. The latter is equivalent to the folklore fact that weak convergence is equivalent to pointwise convergence when the underlying space is RKHS. Indeed, consider any sequence g_n that converges weakly to an element $g \in H$, which means that $\lim_{n\to\infty}\langle g_n, f\rangle = \langle g, f\rangle$ for all $f \in H$. In particular, the latter holds for all $k(x,\cdot)$, $x \in \mathcal{X}$ yielding the necessity. To show the sufficiency, if $\lim_{n\to\infty} g_n(x) = g(x)$ for all $x \in X$ then $\lim_{n\to\infty}\langle g_n, f\rangle = \langle g, f\rangle$ for all $f \in \operatorname{span}\{k(x,\cdot)\}$, by linearity of the inner product and its continuity. The claim then follows since H is complete. From this statement, we deduce that the limit in (15) is actually a pointwise limit. Thus, for every $x \in \mathcal{X}$, we have $\lim_{t\to 0}\frac{1}{t}(k(\cdot, x + te_a) - k(\cdot, x)) = \frac{\partial}{\partial x_a}k(\cdot, x)$ and, moreover, it holds for every $f \in H$ that

$$\left\langle \frac{\partial}{\partial x_a}k(\cdot, x), f\right\rangle_H = \left\langle \lim_{t\to 0}\frac{1}{t}(k(\cdot, x + te_a) - k(\cdot, x)), f\right\rangle = \lim_{t\to 0}\frac{f(x + te_a) - f(x)}{t} = \frac{\partial f(x)}{\partial x_a},$$

where we used continuity of inner product and the reproducing property in the second equality. Hence, the derivative exists, and the differential reproducing property holds.

Let $X' := \{x^i : i = 1, \ldots, n\}$, and let

$$H|_{X'} := \operatorname{span}\left\{k(\cdot, x^i), \frac{\partial k(\cdot, s)}{\partial s_a}|_{s=x^i} : i = 1, \ldots, n,\ a = 1, \ldots, d\right\}.$$

Furthermore, let $H_{|X'}^{\perp}$ be the orthogonal complement of $H|_{X'}$ in H; it exists and is well-defined as every element in the span exists and is well-defined. By the Hilbert Projection Theorem, every

$f_j \in H$, $j \in [d]$ can be uniquely decomposed as $f_j = f_j^{\|} + f_j^{\perp}$, where $f_j^{\|} \in H|_{X'}$ and $f_j^{\perp} \in H^{\perp}_{|X'}$. Let $e_a \in \mathbb{R}^d$ be the a-th vector of the standard Euclidean basis in $\mathbb{R}^d$. By the reproducing property and the definition of $H^{\perp}_{|X'}$ in H, it holds that

$$f_j^{\perp}(x^i) = \Big\langle f_j^{\perp}, k(\cdot, x^i) \Big\rangle_H = 0.$$

Using the fact (proved above) that the differentiation reproducing property holds for $f_j \in H$, together with the orthogonal property we get:

$$\frac{\partial f_j^{\perp}}{\partial x_a}(x^i) = \Big\langle f_j^{\perp}, \frac{\partial k(\cdot, s)}{\partial s_a}|_{s=x^i} \Big\rangle_H \overset{f_j^{\perp} \in H^{\perp}_{|X'}}{=} 0.$$

By the reproducing property in H, we deduce that for every $j \in [d]$, it holds

$$\frac{1}{n}\ell\big(\mathbb{X}^j, f_j(\mathbb{X})\big)^2 = \frac{1}{n}\ell\Big(\mathbb{X}^j, {f_j}^{\|}(\mathbb{X})\Big)^2,$$

whereas the differential reproducing property implies that it holds that

$$W_{aj}^{\mathcal{D}}(f) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial f_j^2(x^i)}{\partial x_a^i} = \frac{1}{n}\sum_{i=1}^{n} \frac{\Big(\partial f_j^{\|}(x^i)\Big)^2}{\partial x_a^i} = W_{aj}^{\mathcal{D}}(f^{\|})$$

Notice further that

$$\left\| \frac{\partial f_j(x)}{\partial x_a} \right\|_n = \left\| \frac{\partial f_j^{\|}(x)}{\partial x_a} \right\|_n,$$

which in turn implies that $\Omega_1^{\mathcal{D}}(f_j) = \Omega_1^{\mathcal{D}}(f_j^{\|})$. Thus, by denoting

$$\mathcal{R}_{L,\mathcal{D}}(f) + \tau(2\Omega_1^{\mathcal{D}}(f) + \lambda\|f\|_{H^d}^2) = \sum_{j=1}^{d}\left\{\frac{1}{2n}\big\|\mathbb{X}^j - f_j(\mathbb{X})\big\|^2 + \tau[2\Omega_1^{\mathcal{D}}(f_j) + \lambda\big\|f_j\big\|_H^2]\right\},$$

for $f = (f_1, \ldots, f_d) \in H^{\otimes d}$, where $H^{\otimes d}$ denotes the direct product of d copies of H, we get that over the acyclicity constraint the following chain of the equalities holds:

$$\begin{aligned}
&\inf_{f \in H^d, h^s_{\text{ldet}}(W^{\mathcal{D}}(f))=0} \mathcal{R}_{L,\mathcal{D}}(f) + \tau(2\Omega_1^{\mathcal{D}}(f) + \lambda\|f\|_{H^d}^2) \\
&= \inf_{\substack{f \in H^d, f = f^{\|} + f^{\perp}, \\ f^{\|} \in H|_{X'}^d, h^s_{\text{ldet}}(W^{\mathcal{D}}(f^{\|}))=0}} \mathcal{R}_{L,\mathcal{D}}(f^{\|}) + \tau(2\Omega_1^{\mathcal{D}}(f^{\|}) + \lambda\big\|f^{\|}\big\|_{H^d}^2 + \lambda\big\|f^{\perp}\big\|_{H^d}^2) \\
&= \inf_{\substack{f \in H^d, f = f^{\|} + f^{\perp}, f^{\|} \in H|_{X'}^d, \\ \|f^{\perp}\|_{H^d}=0, h^s_{\text{ldet}}(W^{\mathcal{D}}(f^{\|}))=0}} \mathcal{R}_{L,\mathcal{D}}(f^{\|}) + \tau(2\Omega_1^{\mathcal{D}}(f^{\|}) + \lambda\big\|f^{\|}\big\|_{H^d}^2) \\
&= \inf_{f \in H|_{X'}^d, h^s_{\text{ldet}}(W^{\mathcal{D}}(f))=0} \mathcal{R}_{L,\mathcal{D}}(f) + \tau(2\Omega_1^{\mathcal{D}}(f) + \lambda\|f\|_{H^d}^2).
\end{aligned}$$

Thus, we showed that

$$\inf_{\substack{f \in H^d, \\ h^s_{\text{ldet}}(W^{\mathcal{D}}(f))=0}} \mathcal{R}_{L,\mathcal{D}}(f) + \tau(2\Omega_1^{\mathcal{D}}(f) + \lambda\|f\|_{H^d}^2) = \inf_{\substack{f \in H|_{X'}^d, \\ h^s_{\text{ldet}}(W^{\mathcal{D}}(f))=0}} \mathcal{R}_{L,\mathcal{D}}(f) + \tau(2\Omega_1^{\mathcal{D}}(f) + \lambda\|f\|_{H^d}^2),$$

establishing the first claim of the Theorem holds.

To prove (9), we first note that, by the differential reproducing property,

$$\langle \frac{\partial k(\cdot, x)}{\partial x_a}, \frac{\partial k(\cdot, y)}{\partial y_b} \rangle_H = \frac{\partial^2 k(x, y)}{\partial x_a \partial y_b}. \tag{16}$$

Plugging in the formula for the solution of the constrained minimization problem and using reproducing and differential reproducing properties we obtain:

$$\begin{aligned}
\left\| \widehat{f}_j^{\,\tau} \right\|_H^2 &= \left\langle \sum_{i=1}^n \alpha_i^j k\big(\cdot, x^i\big) + \sum_{i=1}^n \sum_{a=1}^d \beta_{ai}^j \frac{\partial k(\cdot, x^i)}{\partial x_a^i}, \sum_{i=1}^n \alpha_i^j k\big(\cdot, x^i\big) + \sum_{i=1}^n \sum_{a=1}^d \beta_{ai}^j \frac{\partial k(\cdot, x^i)}{\partial x_a^i} \right\rangle_H \\
&= \sum_{i,l=1}^n \alpha_i^j \alpha_l^j k\Big(x^i, x^l\Big) + 2 \sum_{i,l=1}^n \sum_{a=1}^d \alpha_i^j \beta_{a_l}^j \left\langle \frac{\partial k\big(\cdot, x^l\big)}{\partial x_a^l}, k\big(\cdot, x^i\big) \right\rangle_H \\
&+ \sum_{i,l=1}^n \sum_{a,b=1}^d \beta_{ai}^j \beta_{bl}^j \left\langle \frac{\partial k\big(\cdot, x^i\big)}{\partial x_a^i}, \frac{\partial k\big(\cdot, x^l\big)}{\partial x_b^l} \right\rangle_H \\
&= \sum_{i,l=1}^n \alpha_i^j \alpha_l^j k\Big(x^i, x^l\Big) + 2 \sum_{i,l=1}^n \sum_{a=1}^d \alpha_i^j \beta_{al}^j \frac{\partial k\big(x^i, x^l\big)}{\partial x_a^l} + \sum_{i,l=1}^n \sum_{a,b=1}^d \beta_{ai}^j \beta_{bl}^j \frac{\partial k\big(x^i, x^l\big)}{\partial x_a^i \partial x_b^l},
\end{aligned} \tag{17}$$

which finalizes the proof. ■

Appendix B. Details of Numerical Experiments and Intuition for the Choice of the Bandwidth of the Gaussian Kernel γ.

B.1. Simulations.

We simulate ER4 DAGs following the procedure outlined by Zheng et al. (2020) with the functional relationship established in three distinct ways and add an additional simulation type.

1. The first way employs the Gaussian process:

   $$f_j(X) = g_j(X_{\mathrm{pa}(j)}) + \varepsilon_j \quad \forall\, j \in [d],$$

   where g_j is sampled from RBF GP with lengthscale 1. In detail, for the sub-data matrix $\mathbb{X}_{\mathrm{pa}(j)} \in \mathbb{R}^{n \times p}$ where $p \leq d$, we have $g_j(X_{\mathrm{pa}(j)}) \sim \mathcal{N}(0, K(\mathbb{X}_{\mathrm{pa}(j)}, \mathbb{X}_{\mathrm{pa}(j)}))$ with

   $$K(\mathbb{X}_{\mathrm{pa}(j)}, \mathbb{X}_{\mathrm{pa}(j)})_{a,b} = \exp\left(-\frac{\left\| \boldsymbol{x}^a - \boldsymbol{x}^b \right\|_2^2}{2l^2} \right).$$

   Here, $\boldsymbol{x}^a, \boldsymbol{x}^b$ denote the a-th and b-th row of $\mathbb{X}_{\mathrm{pa}(j)}$, respectively, and the lengthscale is $l = 1$. Moreover, $\varepsilon_j \sim \mathcal{N}(0, I_n)$ is a standard Gaussian noise.

2. The second way is referred to as the additive Gaussian process (additive GP). It sets

   $$f_j(X) = \sum_{k \in \mathrm{pa}(j)} g_{kj}(X_k) + \varepsilon_j,$$

   where each g_{kj} is sampled from RBF GP with lengthscale 1.

3. The third way involves using an MLP network with a hidden layer size of 100 and a sigmoid activation function, where all weights are sampled from $\mathcal{U}((-2.0,-0.5)\cup(0.5,2.0))$.

4. Moreover, we introduce an additional simulation type called the combinatorial model, where the non-linear relationship is a linear combination of various common non-linear functions:

$$f_j(X) = \sum_{k\in \mathrm{pa}(j)} g_{kj}(X_k) + \varepsilon_j,$$

where g_{kj} is randomly picked from following non-linear functions:

$$g(x) = \exp(-|x|),\ g(x) = 0.05x^2,\ g(x) = \sin(x).$$

B.2. Hyperparameter choice

While considering the experimental setting with different dimensions of the underlying Erdős-Rényi graphs, we notice that the the complexity of the decision rule increases with the grows of the dimension. Thus, one observes a "classical" phenomenon of the curse of dimensionality (see for example Giraud, 2021; Györfi et al., 2002). In a nutshell, high-dimensional i.i.d. observations are "essentially" equidistant from each other, while the distance between the points grows with the growing dimension, which poses a problem for high-dimensional metric-based methods. To handle this problem, we seek to reduce the estimation error for the signal by decision rules with higher regularity.

In our case, we employ the machinery of RKHS rules based on the Gaussian kernel (13) with parameter γ. Let H_γ be the Gaussian RKHS with reproducing kernel $k(x,\cdot) := k_\gamma(x,\cdot) = \exp\left(-\frac{\|x-\cdot\|^2}{\gamma^2}\right)$. It then holds that $H_{\gamma_2} \subset H_{\gamma_1}$ for $\gamma_2 > \gamma_1 > 0$ (see Steinwart and Christmann, 2008, Proposition 4.46). Hence, higher regularity results from a larger choice of γ. Notice that for bounded domains $\mathcal{X}$, the same effect (i.e., restricting to the spaces of larger smoothness) can be ensured by considering the re-scaled domain with parameter $\frac{1}{\gamma}$.