Department of Mathematics
TUM School of Computation, Information and Technology
Technical University of Munich

Differentiable Learning of Non-Linear Directed Graphical Models

Yurou Liang

Thesis for the attainment of the academic degree

Master of Science

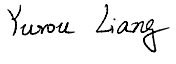
at the TUM School of Computation, Information and Technology of the Technical University of Munich

Supervisor:
Prof. Dr. Mathias Drton

Advisors:
Dr. Oleksandr Zadorozhnyi

Submitted:
Munich, July 13, 2024

I hereby declare that this thesis is my own work and that no other sources have been used except those clearly indicated and referenced.

Munich, July 13, 2024 Yurou Liang 

Zusammenfassung

Kausale Entdeckung bedeutet, einen gerichteten azyklischen Graphen (DAG) zu lernen, der ein kausales Modell kodiert. Aufgrund des großen kombinatorischen Suchraums kann dieses Modellauswahlproblem schwierig zu lösen sein, insbesondere wenn nicht-parametrische Kausalmodelle betrachtet werden. Eine neuere Forschungsrichtung versucht, die kombinatorische Suche zu umgehen, indem sie die kausale Entdeckung als kontinuierliches Optimierungsproblem betrachtet. Zu diesem Zweck werden Beschränkungen angenommen, die die Azyklizität des Graphen charakterisieren. Bestehende Arbeiten zu nicht-parametrischen Einstellungen basieren auf endlich-dimensionalen Approximationen der nicht-parametrischen Beziehung zwischen den Knoten, was zu einem Score-basierten kontinuierlichen Optimierungsproblem unter einer glatten Azyklizitätsbeschränkung führt. In dieser Arbeit entwickeln wir einen alternativen Approximationsansatz, indem wir mit reproduzierenden Kernel-Hilbert-Räumen (RKHS) arbeiten und allgemeine sparsamkeitsinduzierende Regularisierungsterme verwenden, die auf partiellen Ableitungen beruhen. In diesem Rahmen führen wir einen erweiterten RKHS-Repräsentantensatz ein. Um Azyklizität zu erhalten, befürworten wir die log-determinante Formulierung der Azyklizitätsbeschränkung und zeigen ihre Stabilität. Schließlich untersuchen wir die Leistung unseres resultierenden RKHS-DAGMA-Verfahrens in einer Reihe von Simulationen sowie einer illustrativen Datenanalyse.

Abstract

Causal discovery amounts to learning a directed acyclic graph (DAG) that encodes a causal model. Due to its large combinatorial search space, this model selection problem can be challenging to solve, especially when considering non-parametric causal models. A line of recent research seeks to sidestep the combinatorial search by framing causal discovery as a continuous optimization problem. To this end, one adopts constraints that characterize the acyclicity of the graph. Existing works on non-parametric settings are based on finite-dimensional approximations of the non-parametric relationship between the nodes, resulting in a score-based continuous optimization problem under a smooth acyclicity constraint. In this work, we develop an alternative approximation approach by working with reproducing kernel Hilbert spaces (RKHS) and using general sparsity-inducing regularization terms that are based on partial derivatives. Under this setting, we introduce an extended RKHS Representer Theorem. To hold acyclicity, we advocate the log-determinant formulation of the acyclicity constraint and show its stability. Finally, we examine the performance of our resulting RKHS-DAGMA procedure in a set of simulations as well as illustrative data analysis.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my supervisor, Prof. Dr. Mathias Drton, for providing me with the opportunity to work on this thesis. His insightful suggestions and direction were crucial in shaping the main focus and structure of my research.

I am also profoundly grateful to my advisor, Dr. Oleksandr Zadorozhnyi for his detailed advice, and encouragement, and for helping me refine the specifics of my work. His expertise and constructive feedback have been important in shaping the finer details of my work, and his mentorship has significantly contributed to the successful completion of this thesis.

I am particularly thankful to the Konrad Zuse School of Excellence in Reliable AI (relAI) for their generous financial support and resources, which were crucial for the completion of my research and this thesis. Their funding provided me with the opportunity to pursue my studies, conduct extensive research, and achieve my academic goals.

I would like to extend my appreciation to my friends and colleagues at the university, particularly Lei Cheng, for their invaluable assistance with the coding aspects of this thesis. Their expertise and willingness to help have been greatly appreciated.

Finally, I would like to extend my heartfelt thanks to my family and friends for their unwavering love, patience, and encouragement throughout my studies. Their support has been my greatest strength.

Contents

1 Introduction

Structural equation models (SEMs) based on directed acyclic graphs (DAGs, also known as Bayesian networks) have found wide-spread applications ranging from computational biology [Zha+23] to manufacturing [G̈24] and finance [Ji+18]. In the realm of computational biology, Zhang et al. [Zha+23] developed a causal active learning strategy that employs a Bayesian update for the causal model to identify interventions that are optimal. Within the financial sector, Ji et al. [Ji+18] used DAG to uncover the contemporaneous and lagged relations between Bitcoin and other asset classes. In the framework of SEMs, to represent a joint dependence structure, every variable is modeled as a function of a subset of the other variables as well as noise. In this setting, models based on DAGs assume that there are no causal feedback loops. Such an assumption can be restrictive but is key for allowing the definition of non-linear models. Indeed, it is generally unclear whether cyclic systems of structural equations admit a unique solution or a solution at all.

In many applications, the underlying DAG is not known, and methods for causal discovery, which learn the DAG from data, promise to offer useful insights. Numerous algorithms have been proposed for causal discovery, see, e.g., Drton and Maathuis [DM17] or Spirtes and Zhang [SZ19]. One classical approach is provided by constraint-based algorithms which are based on testing conditional independences [MT99; SG91; Tsa+03]. A second prominent approach is given by score-based algorithms [Chi02; HGC95]. In this work, our focus will be on a score-based approach. Specifically, we will take up a recent theme whose aim is to find a DAG that minimizes a model selection score based on a continuous optimization problem with a continuous acyclicity constraint imposed on a weighted adjacency matrix W. This theme was initiated in the NOTEARS algorithm [Zhe+18] which assumes a linear SEM and uses an exponential acyclicity constraint changing the combinatorial optimization problem into a continuous optimization problem that is solved in an augmented Lagrangian scheme. More precisely, consider $W \in \mathbb{R}^{d\times d}$ as the weighted adjacency matrix of a graph G with d nodes whose edges correspond to the direct effects in a linear SEM. Let $W \circ W$ be the Hadamard product. Then the constraint of Zheng et al. [Zhe+18] is posed through the exponential acyclicity function $h_{\exp}(W) = \operatorname{Tr}(e^{W\circ W}) - d$. With the resulting constraint, score-based causal discovery becomes amenable to the application of commonly used gradient-proxy algorithms for (non-) linear optimization problems. Several follow-up works proposed other acyclicity characterizations [BAR22; Naz+23; NGZ20; Yu+19]. Following related literature, we refer to this general approach as differentiable causal discovery.

Recent work also extended this methodology to non-parametric settings. Suppose we observe a d-dimensional stochastic system comprised of the real-valued random variables $(X_j)_{j=1}^d$. In a directed graphical model presented through structural equations, each variable X_j exhibits a relationship with the other coordinates of random vector $X := (X_j)_{j=1}^d$. Assuming additive noise ε_j, but allowing for non-linearity, it then holds that

$$X_j = f_j(X) + \varepsilon_j, \quad j = 1, \dots, d. \tag{1.1}$$

Here, each function $f_j : \mathbb{R}^d \mapsto \mathbb{R}$ is measurable and is modeled to belong to a function class of interest. In particular, f_j does not depend on coordinate X_j. Zheng et al. [Zhe+20] treat what we refer to as *differentiable causal discovery* in a nonparametric setting. Namely, the authors propose to approximate the functions f_j by multi-layer perceptrons or via a (truncated) basis expansion in the original functional space, i.e., in the space of functions where both the functions and their derivatives are square-integrable over domain $\mathcal{X}$ (an expansion is then possible via the trigonometric basis of functions) and minimize the corresponding residual loss subject to trace-based acyclicity constraint for the expansions of functions f_j. In the sequel, we work under additional assumptions that the functions are continuously differentiable (which as we will show is for example ensured when considering Gaussian RKHS). In the MLP framework of Zheng et al. [Zhe+20], 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. In the case of approximation by basis expansions, the general nonparametric model is assumed to be an additive model, so $f_j(X) = \sum_{k\neq j} f_{jk}(X_k)$, with each f_{jk} approximated by a finite number of orthonormal elements from the basis in the space of differentiable functions which

derivatives are square-integrable. Then W_{kj} is defined as the L_2 norm of the coefficients corresponding to the basis approximation of f_{jk}.

Both the MLP approximation and the basis expansions in Zheng et al. [Zhe+20] yield finite-dimensional optimization problems in terms of neural network weights or basis coefficients. The current MLP approximation is sensitive to the size of hidden units. Although increasing the size of the hidden layers increases the flexibility of MLP functions, larger networks require more samples to estimate the parameters [Zhe+20]. Moreover, the current MLP approximation relies on random initialization for the weights which causes obvious randomness in results [see the illustration from WBD24, Figure 2 therein]. Fine-tuning the architecture of a neural network is, thus, a non-trivial task. On the other hand, the basis expansion approximation adopted in the earlier work is restricted through its focus on additive models.

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

- We approximate each function in (1.1) with the help of an RKHS given by a differentiable kernel k and establish a version of an RKHS Representer Theorem for an empirical acyclicity-constrained optimization problem (similar to that of Rosasco et al. [Ros+13] in the statistical learning scenario). Given data $(x^i)_{i=1}^n$, this leads to optimizing functions that are combinations of evaluations of the kernel and its partial derivatives:

$$\sum_{i=1}^{n} \alpha_i k(x, x^i) + \sum_{i=1}^{n} \sum_{a=1}^{d} \beta_{ai} \frac{\partial k(x, s)}{\partial s^a}\bigg|_{s=x^i}. \tag{1.2}$$

- We use the implication that 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, which is model-agnostic as one does not define the weights in terms of approximations to f_j.

- We base our optimization on the DAGMA method [BAR22] and adopt the log-determinant acyclicity constraint h_{ldet}, for which we demonstrate stable optimization behaviour 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 [Moo+16]. The code for our experiments is available at the author's GitHub site.[1]

Outline. Chapter 2 sets up notation and reviews the basic settings for directed acyclic graph (DAG) as well as the DAG learning problem. In chapter 3, we investigate the DAG learning problem under the assumption of the linear structure equation model (SEM), demonstrating that it can be reformulated as an equality constrained optimization problem. Then we review two essential optimization methods, namely the Lagrange rule and the augmented Lagrange multiplier method [Nem99], to solve equality constrained optimization problems. Finally, we study the application of the augmented Lagrange scheme to the DAG learning problem within the framework of linear SEM, resulting in the linear NOTEARS algorithm [Zhe+18]. Chapter 4 modifies the linear NOTEARS algorithm by incorporating adaptive Lasso to facilitate learning a sparse DAG. This modification allows for adaptive penalty levels applied to different coefficients of the weighted adjacency matrix [Xu+22]. Chapter 5 extends the linear NOTEARS algorithm into general non-parametric DAGs [Zhe+20]. Moreover, we summarize the existing acyclicity constraints and discuss their stability [Naz+23]. We also introduce an alternative optimization scheme "DAGMA" using the central path approach [BAR22]. In chapter 6, we review fundamental facts for kernels and reproducing kernel Hilbert space (RKHS) as well as the RKHS Representer Theorem [SC08]. In chapter 7, we introduce the sparsity regularizer and develop a novel extended RKHS Representer Theorem for the constrained optimization problem with acyclicity constraint. Based on this, we formulate the overall learning objective and assemble the "DAGMA" optimization scheme. Finally, we compare the performance of RKHS-DAGMA with different versions of nonparametric NOTEARS [Zhe+20] in numerical experiments in chapter 8. Additional details on the optimization methods are given in the Appendix.

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

2 Graphical Modelling

A Bayesian network is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG) whose nodes represent the random variables. Each edge represents a direct conditional dependency. In this chapter, we review some basic settings for DAGs from Lauritzen [Lau96] and the DAG learning problem.

Definition 2.1 (DAG)**.** *Let G be a graph with vertices $V = \{v_1, \dots, v_d\}$ and edges $E \subseteq V \times V$, i.e. $G = (V, E)$. If E is a set of ordered pairs, then G is called a directed graph. A walk is a finite or infinite sequence of edges which joins a sequence of vertices. A finite walk of G is a sequence of edges $(e_1, e_2, \dots, e_{k-1})$ for which there is a sequence of vertices $(v_1, v_2, \dots, v_n)$ such that $e_i = (v_i, v_{i+1})$ for $i \in \{1, \dots, k-1\}$. The walk is called closed if $v_1 = v_k$. If a directed graph has no closed walk, it is called a directed acyclic graph (DAG).*

Definition 2.2 (Adjacency matrix)**.** *Suppose a graph $G = (V, E)$, a square $d \times d$ matrix A is called an adjacency matrix of G if $A_{ij} = 1$ when there is a directed edge from vertex i to j and $A_{ij} = 0$ otherwise.*

Definition 2.3 (Local Markov property)**.** *A DAG model G satisfies the local Markov property, if for all $\alpha \in V : \alpha \perp\!\!\!\perp \{nd(\alpha) \setminus pa(\alpha) \mid pa(\alpha)\}$ where $nd(\alpha)$ are the non-descendants of α, $pa(\alpha)$ are the parents of α.*

Example 2.4. *The following DAG implies that $B \perp\!\!\!\perp C|A$ and $B \perp\!\!\!\perp D|(A, C)$.*

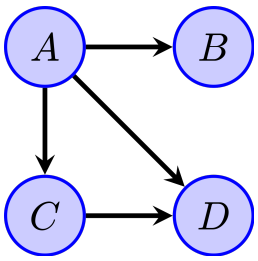


Figure 2.1

Denote $[d] := \{1, 2, \dots, d\}$. Consider a directed graph $G = (V, E)$ with vertex set $V = [d]$ and edge set $E \subset V \times V$. As usual, we define the set of *parents* of a vertex $i \in V$ as $\mathrm{pa}(i) = \{j \in V : (j, i) \in E\}$. We associate with the graph G a random vector $X = (X_1, \dots, X_d)$ taking values in $\mathcal{X} \subset \mathbb{R}^d$ with its distribution $\mathbb{P}$ over on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and we assume $\mathcal{X}$ is a bounded connected non-empty open set.

Proposition 2.5. For a subset $A \subset [d]$, let $X_A = (X_i)_{i \in A}$. Similarly, for a rectangular Borel set $B \in \mathcal{B}(\mathbb{R}^d)$, $B = \times_{i=1}^d B_i$, we define $B_A = \times_{i \in A} B_i$. When $A = \emptyset$, we set $X_A \equiv 0$ and $B_A = \{0\}$. The *probabilistic graphical model* [Maa+19] corresponding to graph G is the family of joint distributions $\mathbb{P}_X$ for X under which

$$\mathbb{P}_X(B) = \mathbb{P}(X \in B) = \prod_{i=1}^{d} \mathbb{P}\big(X_i \in B_i \,|\, X_{\mathrm{pa}(i)} \in B_{\mathrm{pa}(i)}\big) \quad \forall B = \times_{i=1}^d B_i \in \mathcal{B}\big(\mathbb{R}^d\big).$$

Proof. Assume the statement holds for all DAGs with k vertices. For DAG G with $k+1$ vertices, we reorder the index of vertices so that $(k+1)-$th vertex has no descendant. By Bayes' Rule,

$$\begin{aligned}\frac{\mathbb{P}((X_1, \dots, X_{k+1}) \in B_{[k+1]})}{\mathbb{P}(X_{\mathrm{pa}(k+1)} \in B_{\mathrm{pa}(k+1)})} &= \mathbb{P}((X_1, \dots, X_{k+1}) \in B_{[k+1]} \mid X_{\mathrm{pa}(k+1)} \in B_{\mathrm{pa}(k+1)}) \\ &= \mathbb{P}(X_{k+1} \in B_{k+1} \mid X_{\mathrm{pa}(k+1)} \in B_{\mathrm{pa}(k+1)}) \cdot \mathbb{P}((X_1, \dots, X_k) \in B_{[k]} \mid X_{\mathrm{pa}(k+1)} \in B_{\mathrm{pa}(k+1)}),\end{aligned}$$

where we use the local Markov property for the second equation. Then

$$\begin{aligned}\mathbb{P}((X_1,\dots,X_{k+1}) \in B_{[k+1]}) &= \mathbb{P}(X_{k+1} \in B_{k+1} \mid X_{\mathrm{pa}(k+1)} \in B_{\mathrm{pa}(k+1)}) \cdot \mathbb{P}((X_1,\dots,X_k) \in B_{[k]}) \\ &= \prod_{i=1}^{k+1} \mathbb{P}(X_i \in B_i \mid X_{\mathrm{pa}(i)}),\end{aligned}$$

where the second equality follows by assumption. The statement follows by induction. □

Suppose for all $j \in [d]$, the conditional expectations have the form $\mathbb{E}[X_j|X_{\mathrm{pa}(j)}] = f_j(X) + \varepsilon_j$, where f_j: $\mathcal{X} \to \mathbb{R}$ does not depend on X_k if $k \notin \mathrm{pa}(j)$, and $\left(\varepsilon_j\right)_{j\in[d]}$ are stochastic error terms that are independent over j. Let $\mathbb{X} \in \mathbb{R}^{n\times d}$ be a data matrix whose rows x^i, $i = 1,\dots,n$ 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 $\ell : \mathbb{R}\times\mathbb{R} \mapsto \mathbb{R}_+$. The typical loss function is the quadratic loss: $\ell(y,\widehat{y}) = (y-\widehat{y})^2$. The goal is to estimate the functional dependency structure of random vector X which can be represented as a DAG on the space of vertices V. More precisely, the goal is to estimate $f = (f_1,\dots,f_d)$ by minimizing the score function

$$L(f) := \frac{1}{2n}\sum_{j=1}^{d}\sum_{i=1}^{n} \ell(x_j^i, f_j(x^i)) \text{ subject to the dependencies in } f \text{ corresponding to a DAG.} \tag{2.1}$$

We describe structure learning in two cases: linear structural equation models [Zhe+18] and general non-parametric structural equation models [Zhe+20].

3 Structure Learning in Linear Structural Equation Models

In this chapter, we introduce the linear NOTEARS algorithm developed by Zheng et al. [Zhe+18] based on the linear structural equation model (SEM). Let $W = \{w_1|\dots|w_d\} \in \mathbb{R}^{d\times d}$. For random vector $X = (X_1, \dots, X_d)$, we assume a linear SEM: $X_j = w_j^\top X + \varepsilon_j$, and $\left(\varepsilon_j\right)_{j\in[d]}$ are stochastic error terms that are independent over j. Thus, W define a graph $G(W)$ on d nodes with the adjacency matrix $A(W)_{ij} = \mathbb{1}_{\{w_{ij}\neq 0\}}$. In causal inference, one usually expect that a random variable depends only on a few other random variables. Since W_{ij} represents the direct impact of X_i to X_j, we favor a weight W for which each entry W_{ij} is small. Thus, ℓ_1-regularization $\|W\|_1 = \sum_{i,j=1}^d |W_{ij}|$ is added to the score function. Thus, the aim is to solve

$$\min_{W\in\mathbb{R}^{d\times d}} \frac{1}{2n}\|\mathbb{X} - \mathbb{X}W\|_F^2 + \lambda\|W\|_1 \text{ subject to } G(W) \text{ is a DAG}, \tag{3.1}$$

where $\|A\|_F \coloneqq \sqrt{\sum_{i,j=1} |A_{ij}|^2}$.

3.1 Acyclicity Characterization

It is shown that the optimization problem (3.1) is NP-hard due to the combinatorial search space caused by the acyclicity constraint [CHM04]. To solve this problem, Zheng et al. [Zhe+18] introduced a continuous acyclicity constraint, transforming (3.1) into a continuous optimization problem. Ideally, an acyclicity constraint h should satisfy following desiderata:

D1. $h(W) = 0$ if and only if W is acyclic;

D2. The values of h quantify the "DAG-ness" of the graph;

D3. h is smooth. More precisely, h is at least twice continuous differentiable everywhere;

D4. h and its derivative are easy to compute.

Note that the desiderata D1 characterizes the acyclicity. D2 means that the values of h quantify how severe violations from acyclicity become as $G(W)$ moves further from DAG. Furthermore, D3 and D4 ensure the continuous optimization problem resulted by h can be solved by numerous common optimization methods, for example, the augmented Lagrangian scheme which will be investigated in detail in section 3.2.

To study the acyclicity constraint, let us consider first the simpler case of binary adjacency matrices $B \in \{0,1\}^{d\times d}$.

Special case: Binary adjacency matrices

Proposition 3.1.1. *[Zhe+18, Proposition 1] Suppose $B \in \{0,1\}^{d\times d}$. Let $\lambda_i(B) \in \mathbb{C}$ be the $i-$th large eigenvalue of B w.r.t. complex magnitude, i.e., $|\lambda_1(B)| \le |\lambda_2(B)| \le \cdots \le |\lambda_d(B)|$, and let $r(B)$ be the spectral radius of B, i.e., $r(B) = |\lambda_d(B)|$. Let $G(B)$ denote the graph with adjacency matrix B. Suppose $r(B) < 1$, then $G(B)$ is a DAG if and only if $tr(I-B)^{-1} = d$.*

Proof. Since $tr(B^k)$ counts the number of length-k closed walks in a directed graph, then $G(B)$ is acyclic if and only if $tr(B^k) = 0$ for all $k = 1, \dots, \infty$, which is equivalent to $\sum_{k=1}^\infty \sum_{i=1}^d (B^k)_{ii} = \sum_{k=1}^\infty tr(B^k) = 0$. Note that under the assumption $r(B) < 1$, we have $(I-B)^{-1} = \sum_{k=0}^\infty B^k$. Then

$$tr(I-B)^{-1} = tr(I) + \sum_{k=1}^\infty tr(B^k) = d + \sum_{k=1}^\infty \sum_{i=1}^d (B^k)_{ii} = d + 0 = d.$$

We conclude the statement. □

However, the condition $r(B) < 1$ is generally not true. The following acyclicity characterization holds for all possible B.

Proposition 3.1.2. *[Zhe+18, Proposition 2] For binary matrix $B \in \{0,1\}^{d\times d}$, $G(B)$ is a DAG if and only if $tr(e^B) = d$.*

Proof. As in the proof of Proposition 3.1.1, B is acyclic if and only if $(B^k)_{ii} = 0$ for all $k = 1, \dots, \infty$ and $i = 1, \dots, d$, which is equivalent to

$$0 = \sum_{k=1}^{\infty} \sum_{i=1}^{d} (B^k)_{ii}/k! = \sum_{k=0}^{\infty} \frac{tr(B^k)}{k!} - d = tr(e^B) - d.$$

□

For the map $h_{\text{exp}} : \{0,1\}^{d\times d} \to \mathbb{R}, B \mapsto tr(e^B) - d$, the domain B is defined on discrete space, hence the derivative of h is not well-defined, which means the desiderata D4 is not satisfied. The following steps extend this acyclicity characterization to continuous domain $\mathbb{R}^{d\times d}$.

The general case: Weighted adjacency matrices

Theorem 3.1.3 (Exponential acyclicity constraint)**.** *[Zhe+18, Theorem 1] A matrix $W \in \mathbb{R}^{d\times d}$ is the weight matrix of a DAG $G(W)$ if and only if the exponential acyclicity constraint $h_{exp}(W) := tr(e^{W\circ W}) - d = 0$, where $W \circ W$ denotes the Hadamard product. Moreover, $h_{exp}(W)$ has a simple gradient*

$$\nabla h_{exp}(W) = (e^{W\circ W})^\top \circ 2W$$

and satisfies all desiderata (D1)-(D4)*.*

Proof. Let $A(W)$ be the adjacency matrix of $G(W)$. $G(W)$ is acyclic if and only if $(A(W)^k)_{ii} = 0$ for all $k \geq 1, i = 1, \dots, d$. Since

$$A(W)_{ij} = \mathbb{1}_{\{w_{ij}\neq 0\}} = \mathbb{1}_{\{w_{ij}^2 > 0\}},$$

then $(A(W)^k)_{ii} = 0$ for all $k \geq 1, i = 1, \dots, d$ if and only if $[(W \circ W)^k]_{ii} = 0$ for all $k \geq 1, i = 1, \dots, d$. Similar to the proof of Proposition 3.1.2, we conclude that $h(W) := tr(e^{W\circ W}) - d = 0$ if and only if $G(W)$ is a DAG. The gradient of h is followed by numerical calculations. Consequently, desiderata (D1), (D3), and (D4) are satisfied. Note that $tr(A(W) + A(W)^2 + \dots)$ counts the number of closed walks in $G(W)$. Thus $\exp(tr(A(W)) = \sum_{k=1}^{\infty} \frac{tr(A(W)^k)}{k!}$ re-weights these counts. Replacing $A(W)$ by $W \circ W$ amounts to counting weighted closed walks, where the weight of each edge is w_{ij}^2. Thus, the desiderata (D2) holds. □

By Theorem 3.1.3, the optimization problem (3.1) is equivalent to the following equality constrained problem:

$$\min_{W\in\mathbb{R}^{d\times d}} \frac{1}{2n}\|\mathbb{X} - \mathbb{X}W\|_F^2 + \lambda\|W\|_1 \text{ s.t. } h_{\text{exp}}(W) = 0. \tag{3.2}$$

In the next section, we review some optimization methods to solve (3.2).

3.2 Optimization

First, we consider the following general equality constrained problem (ECP):

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) \text{ subject to } h(\boldsymbol{x}) = \begin{pmatrix} h_1(\boldsymbol{x}) \\ \vdots \\ h_m(\boldsymbol{x}) \end{pmatrix} = 0, \tag{3.3}$$

where $f(\boldsymbol{x}), h_i(\boldsymbol{x}), i = 1, \dots, m$ are smooth (at least twice continuously differentiable) real-valued functions. First, we review some optimization problems to solve (3.3) and then apply them to solve the constrained optimization problem (3.2).

3.2.1 Lagrange Rule

Definition 3.2.1. 1. The feasible surface is defined as $S := \{\boldsymbol{x} \in \mathbb{R}^d : h(\boldsymbol{x}) = \begin{pmatrix} h_1(\boldsymbol{x}) \\ \vdots \\ h_m(\boldsymbol{x}) \end{pmatrix} = 0\}$.

2. A feasible solution $\boldsymbol{x}$ is called regular for the system of constraints of the problem (3.3) if $\nabla h_i(\boldsymbol{x}), i = 1, \ldots, m$ are linearly independent vectors in $\mathbb{R}^m$ or equivalently, the $m \times d$ matrix $\nabla h(\boldsymbol{x})$ has full row rank m.

3. A continuous map $f : \mathbb{R}^{d\times d} \to \mathbb{R}$ is said to be differentiable at a point $\boldsymbol{u} \in \mathbb{R}^{d\times d}$ if there exists a linear functional T that maps from $\mathbb{R}^{d\times d}$ to $\mathbb{R}$, such that

$$\lim_{\boldsymbol{v}\to 0} \frac{|f(\boldsymbol{u}+\boldsymbol{v}) - f(\boldsymbol{u}) - T\boldsymbol{v}|}{\|\boldsymbol{v}\|_2} = 0.$$

If f is differentiable at $\boldsymbol{u}$, the operator T is called the derivative of f at $\boldsymbol{u}$. The operator is usually denoted by $Df(\boldsymbol{u})$ or $\nabla f(\boldsymbol{u})$. If f is differentiable at every point of an open subset $U \subseteq \mathbb{R}^{d\times d}$, then f is called differentiable on U. If f is differentiable at $\boldsymbol{u}$, then for every $\boldsymbol{v} \in \mathbb{R}^{d\times d}$, $Df(\boldsymbol{u})(\boldsymbol{v}) = \frac{d}{dt}\Big|_{t=0} f(\boldsymbol{u} + t\boldsymbol{v})$.

4. For a continuous map $f : \mathbb{R}^d \to \mathbb{R}$, if f is differentiable at $\boldsymbol{u} \in \mathbb{R}^d$, the directional derivative of f along a vector $\boldsymbol{v} \in \mathbb{R}^d$ valued at $\boldsymbol{u}$ is defined by the limit

$$\nabla_{\boldsymbol{v}} f(\boldsymbol{u}) := \lim_{t\to 0} \frac{f(\boldsymbol{u}+t\boldsymbol{v}) - f(\boldsymbol{u})}{t} = Df(\boldsymbol{u})(\boldsymbol{v}) = \nabla f(\boldsymbol{u}) \cdot \boldsymbol{v}.$$

5. A point $\boldsymbol{x}$ is called critical point of a function f if $\nabla f(\boldsymbol{x}) = 0$.

Theorem 3.2.2 (Taylor's Theorem)**.** *[Sar]*

1. *Univariate version: Let $k \geq 1$ be an integer and let the function $f : \mathbb{R} \to \mathbb{R}$ be $(k+1)$ times continuously differentiable on some open interval around the point $a \in \mathbb{R}$. Then for any x on this interval:*

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(k)}(a)}{k!}(x-a)^k + \frac{f^{(k+1)}(\xi)}{(k+1)!}(x-a)^{k+1}$$

for some real number ξ between a and x.

2. *Multivariate version: Let $k \geq 1$ be an integer and let $f : \mathbb{R}^d \to \mathbb{R}$ be a $(k+1)$ times continuously differentiable function on some open interval around the point $\boldsymbol{a} \in \mathbb{R}^d$. Then for any x on this interval:*

$$f(\boldsymbol{x}) = f(a) + \sum_{|\alpha|=1}^{k} \frac{D^\alpha f(\boldsymbol{a})}{\alpha!}(\boldsymbol{x}-\boldsymbol{a})^\alpha + \sum_{|\alpha|=k+1} \frac{D^\alpha f(\xi)}{\alpha!}[\boldsymbol{x}-\boldsymbol{a}]^\alpha,$$

where ξ is some point on the line segment connecting $\boldsymbol{a}$ and $\boldsymbol{x}$. The notations mean:

- $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ *is a multi-index of non-negative integers*;
- $|\alpha| = \alpha_1 + \cdots + \alpha_n$;
- $\alpha! = \alpha_1!\alpha_2! \ldots \alpha_n!$;
- $D^\alpha f = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \ldots \partial x_n^{\alpha_n}}$;
- $(\boldsymbol{x}-\boldsymbol{a})^\alpha = (x_1 - a_1)^{\alpha_1}(x_2 - a_2)^{\alpha_2} \ldots (x_n - a_n)^{\alpha_n}$.

3. *Matrix version: Let $k \geq 1$ be an integer and let $f : \mathbb{R}^{d\times d} \to \mathbb{R}$ be a $(k+1)$ times continuously differentiable function on some open interval around the point $W_0 \in \mathbb{R}^{d\times d}$. Then for any W on this interval:*

$$f(W) = f(W_0) + \sum_{|\alpha|=1}^{k} \frac{D^\alpha f(W_0)}{\alpha!}(W - W_0)^\alpha + \sum_{|\alpha|=k+1} \frac{D^\alpha f(W_\xi)}{\alpha!}(W - W_0)^\alpha,$$

where W_ξ is some point on the line segment between W and W_0. The notations mean:

- $\alpha = (\alpha_{11}, \alpha_{12}, \ldots, \alpha_{dd})$ *is a multi-index of non-negative integers;*
- $|\alpha| = \sum_{i,j} \alpha_{ij}$;
- $\alpha! = \prod_{i,j} \alpha_{ij}!$;
- $[W - W_0]^\alpha = \prod_{i,j} (w_{ij} - (w_0)_{ij})^{\alpha_{ij}}$.

To get "geometrically tractable" feasible sets, it is assumed that at every point of the feasible sets, it should admit a "tangent plane", which means the plane is the best approximate of the surface near that point. So a natural candidate of the tangent plane to the surface S at a point $\boldsymbol{x} \in S$ is the set of those points y obtained by the linearization at $\boldsymbol{x}$ of the smooth equations defining S:

$$\bar{h}_i(\boldsymbol{y}) = h_i(\boldsymbol{x}) + (\boldsymbol{y} - \boldsymbol{x})^\top \bigtriangledown h_i(\boldsymbol{x}) \overset{\boldsymbol{x}\in S}{=} (\boldsymbol{y} - \boldsymbol{x})^\top \bigtriangledown h_i(\boldsymbol{x}).$$

Hence, we denote

$$T_{\boldsymbol{x}} := \{\boldsymbol{y} \mid (\boldsymbol{y} - \boldsymbol{x})^\top \bigtriangledown h_i(\boldsymbol{x}) = 0, i = 1, \ldots, m\} = \{\boldsymbol{y} \mid \bigtriangledown h(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}) = 0\},$$
$$\text{where } \bigtriangledown h(\boldsymbol{x}) = \begin{pmatrix} [\bigtriangledown h_1(\boldsymbol{x})]^\top \\ \vdots \\ [\bigtriangledown h_m(\boldsymbol{x})]^\top \end{pmatrix} = 0 \in \mathbb{R}^{m\times d}.$$

The next theorem says if $\boldsymbol{x}$ is a regular solution, then $T_{\boldsymbol{x}}$ indeed defines the tangent plane to S at point $\boldsymbol{x}$.

Theorem 3.2.3 (Theorem on Tangent plane)**.** *[Nem99, Theorem 8.2.1] Let $\boldsymbol{x} \in S$ be a regular solution of $h(\boldsymbol{x}) = 0$ where h_i is twice differentiable for all $i = 1, \ldots, m$. Then $T_{\boldsymbol{x}}$ is the tangent plane to S at $\boldsymbol{x}$, namely, there exists constant C s.t.*

1. *the distance from an arbitrary point $\boldsymbol{x}' \in S$ of the surface to $T_{\boldsymbol{x}}$ is of the second order of magnitude as compared to the distance from $\boldsymbol{x}'$ to $\boldsymbol{x}$:*

$$\forall \boldsymbol{x}' \in S \; \exists (\boldsymbol{y}' \in T_{\boldsymbol{x}}) : \|\boldsymbol{x}' - \boldsymbol{y}'\|_2 \leq C\|\boldsymbol{x}' - \boldsymbol{x}\|_2^2;$$

2. *the distance from an arbitrary point $\boldsymbol{y}' \in T_{\boldsymbol{x}}$ of the tangent plane to the surface is of the second order of magnitude as compared to the distance from y' to x:*

$$\forall \boldsymbol{y}' \in T_{\boldsymbol{x}} \; \exists (\boldsymbol{x}' \in S) : \|\boldsymbol{x}' - \boldsymbol{y}'\|_2 \leq C\|\boldsymbol{y}' - \boldsymbol{x}\|_2^2.$$

Proof. 1. For any $\boldsymbol{x}' \in S$, by Taylor's Theorem:

$$h_i(\boldsymbol{x}') = h_i(\boldsymbol{x}) + (\boldsymbol{x}' - \boldsymbol{x})^\top \bigtriangledown h_i(\boldsymbol{x}) + \frac{1}{2}(\boldsymbol{x}' - \boldsymbol{x})^\top D^2 h_i(\xi)(\boldsymbol{x}' - \boldsymbol{x}),$$

where ξ is on the line segment between $\boldsymbol{x}$ and $\boldsymbol{x}'$. Since $\boldsymbol{x}, \boldsymbol{x}' \in S$, i.e., $h_i(\boldsymbol{x}) = h_i(\boldsymbol{x}') = 0$, then

$$|(\boldsymbol{x}' - \boldsymbol{x})^\top \bigtriangledown h_i(\boldsymbol{x})| = \frac{1}{2}(\boldsymbol{x}' - \boldsymbol{x})^\top D^2 h_i(\xi)(\boldsymbol{x}' - \boldsymbol{x}).$$

In (3.3), h_i is assumed to be twice continuously differentiable. Thus $D^2 h(\xi)$ is continuous on the compact set defined by the line segment between $\boldsymbol{x}$ and $\boldsymbol{x}'$, and $\left\|D^2 h(\xi)\right\|_2 \leq C_1$ consequently. Then

$$|(\boldsymbol{x}' - \boldsymbol{x})^\top \bigtriangledown h_i(\boldsymbol{x})| = \frac{1}{2}(\boldsymbol{x}' - \boldsymbol{x})^\top D^2 h_i(\xi)(\boldsymbol{x}' - \boldsymbol{x}) \leq C_1\|\boldsymbol{x}' - \boldsymbol{x}\|_2^2$$

for some positive constant C_1. For a fixed $\boldsymbol{x}'$, we can take a $\boldsymbol{y}' \in T_{\boldsymbol{x}}$ s.t.:

$$\|\boldsymbol{x}' - \boldsymbol{y}'\|_2 \leq C_2 |(\boldsymbol{x}' - \boldsymbol{x})^\top \nabla h_i(\boldsymbol{x}) - \underbrace{(\boldsymbol{y}' - \boldsymbol{x})^\top \nabla h_i(\boldsymbol{x})}_{=0}| = C_1 C_2 \|\boldsymbol{x}' - \boldsymbol{x}\|_2^2,$$

where C_2 denotes a positive constant.
2. By Taylor's Theorem, for any $\boldsymbol{y}' \in T_{\boldsymbol{x}}$:

$$h_i(\boldsymbol{y}') = h_i(\boldsymbol{x}) + (\boldsymbol{y}' - \boldsymbol{x})^\top \nabla h_i(\boldsymbol{x}) + \frac{1}{2}(\boldsymbol{y}' - \boldsymbol{x})^\top D^2 h_i(\xi)(\boldsymbol{y}' - \boldsymbol{x}),$$

where ξ is on the line segment between $\boldsymbol{x}$ and $\boldsymbol{y}'$. Since $\boldsymbol{x}$ is feasible and $\boldsymbol{y}' \in T_{\boldsymbol{x}}$, similarly to the proof of statement 1, we can obtain that $|h_i(\boldsymbol{y}')| \leq C_3 \|\boldsymbol{y}' - \boldsymbol{x}\|_2^2$ for some positive constant C_3. We take $\boldsymbol{x}' \in S$ s.t. $\|\boldsymbol{x}' - \boldsymbol{y}'\|_2 \leq C_4 |h(\boldsymbol{x}') - h(\boldsymbol{y}')|$ for some positive constant C_4. Since $\boldsymbol{x}' \in S$, then $\|\boldsymbol{x}' - \boldsymbol{y}'\|_2 \leq C_4 |h(\boldsymbol{y}')| \leq C_3 C_4 \|\boldsymbol{y}' - \boldsymbol{x}\|_2^2$.

□

Remark 3.2.4. Note that we assume for every $\boldsymbol{x} \in S$, there exists $T_{\boldsymbol{x}}$, Theorem 3.2.3 shows that the plane approximates the surface S locally within accuracy which is "infinitesimal of higher order" as compared to the distance to $\boldsymbol{x}$.

Theorem 3.2.5 (First Order Necessary Optimality conditions for ECP - the Lagrange rule)**.** *[Nem99, Theorem 8.2.2] Let $\boldsymbol{x}^*$ be a locally optimal solution of ECP* (3.3) *and let it be a regular point for the system of equality constraints of the problem* (3.3)*. Note that $T_{\boldsymbol{x}^*}$ is well-defined by regularity, then following properties hold:*

1. *The directional derivative of f taken at $\boldsymbol{x}^*$ in every direction along the plane $T_{\boldsymbol{x}^*}$ is zero:*

   For all $\boldsymbol{e}$ with $\boldsymbol{x}^* + \boldsymbol{e} \in T_{\boldsymbol{x}^*}$: $\boldsymbol{e}^\top \nabla f(\boldsymbol{x}^*) = 0$,i.e., if $\nabla h(\boldsymbol{x}^*) \cdot \boldsymbol{e} = 0$, then $\boldsymbol{e}^\top \nabla f(\boldsymbol{x}^*) = 0$.

2. *There exists uniquely defined (by $\boldsymbol{x}^*$) Lagrange multipliers $\lambda_i^*, i = 1, \ldots, m$, s.t.*

   $$\nabla f(\boldsymbol{x}^*) + \sum_{i=1}^{m} \lambda_i^* \nabla h_i(\boldsymbol{x}^*) = 0,$$

   or equivalently, for the Lagrange function

   $$\mathcal{L}(\boldsymbol{x}; \lambda) := f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i h_i(\boldsymbol{x}) = f(\boldsymbol{x}) + \lambda^\top h(\boldsymbol{x}) : \nabla_{\boldsymbol{x}} L(\boldsymbol{x}^*; \lambda^*) = 0.$$

Proof. 1. The statement can be proved by contradiction. Assume there exists a direction $\boldsymbol{e}$ with $\nabla h(\boldsymbol{x}^*)\boldsymbol{e} = 0$ and $\boldsymbol{e}^\top \nabla f(\boldsymbol{x}^*) \neq 0$. W.l.o.g. we can assume $\boldsymbol{e}^\top \nabla f(\boldsymbol{x}^*) < 0$ (otherwise we replace $\boldsymbol{e}$ by $-\boldsymbol{e}$). Consider $\boldsymbol{y}_t := \boldsymbol{x}^* + t\boldsymbol{e}$, then $\nabla h(\boldsymbol{x}^*)(\boldsymbol{y}_t - \boldsymbol{x}^*) = \nabla h(\boldsymbol{x}^*) \cdot t\boldsymbol{e} = 0$ since $\boldsymbol{x}^*$ is feasible. Thus, $\boldsymbol{y}_t \in T_{\boldsymbol{x}^*}$. Note that $\frac{d}{dt}|_{t=0} f(\boldsymbol{y}_t) = \boldsymbol{e}^\top \nabla f(\boldsymbol{x}^*) \stackrel{\text{def}}{=:} -\alpha < 0$, then for all small enough t :

$$f(\boldsymbol{y}_t) \leq f(\boldsymbol{x}^*) - \frac{\alpha}{2} t. \tag{3.4}$$

By the regularity of $\boldsymbol{x}^*$ and Theorem 3.2.3, for every t, there exists a $\boldsymbol{x}_t \in S$ s.t.

$$\|\boldsymbol{y}_t - \boldsymbol{x}_t\|_2 \leq C \|\boldsymbol{y}_t - \boldsymbol{x}^*\|_2^2 = \underbrace{C\|\boldsymbol{e}\|_2^2}_{=:C_1} t^2. \tag{3.5}$$

Since f is continuously differentiable, then by Mean Value Theorem, f is Lipschitz continuous in the neighbourhood of $\boldsymbol{x}^*$. Thus for small enough $t > 0$,

$$|f(\boldsymbol{x}_t) - f(\boldsymbol{y}_t)| \leq C_2 \|\boldsymbol{x}_t - \boldsymbol{y}_t\|_2 \stackrel{(3.5)}{\leq} C_1 C_2 t^2.$$

By (3.4): $f(\boldsymbol{y}_t) - f(\boldsymbol{x}_t) \leq f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t) - \frac{\alpha}{2}t$ which implies

$$f(\boldsymbol{x}_t) \leq f(\boldsymbol{x}^*) + f(\boldsymbol{x}_t) - f(\boldsymbol{y}_t) - \frac{\alpha}{2}t \leq f(\boldsymbol{x}^*) + |f(\boldsymbol{x}_t) - f(\boldsymbol{y}_t)| - \frac{\alpha}{2}t \leq f(\boldsymbol{x}^*) - \frac{\alpha}{2}t + C_1C_2t^2. \tag{3.6}$$

Since (3.6) holds for arbitrary small and positive t, it follows $f(\boldsymbol{x}_t) < f(\boldsymbol{x}^*)$. Note that $\boldsymbol{x}_t$ is a feasible solution of ECP (3.3), thus $f(\boldsymbol{x}_t) < f(\boldsymbol{x}^*)$ which contradicts to that $\boldsymbol{x}^*$ is optimal solution of ECP (3.3).

2. Since $\boldsymbol{x}^*$ is locally optimal of (ECP), by statement 1: For all $\boldsymbol{e}$ with $\boldsymbol{x}^* + \boldsymbol{e} \in T$, it holds $\boldsymbol{e}_{\boldsymbol{x}^*}^\top \triangledown f(\boldsymbol{x}^*) = 0$, i.e., $\triangledown f(\boldsymbol{x}^*)$ is orthogonal to $\{\boldsymbol{e} | \triangledown h(\boldsymbol{x}^*)\boldsymbol{e} = 0\}$. Recall from linear algebra: a vector $\boldsymbol{p}$ is orthogonal to $\{\boldsymbol{e} : A\boldsymbol{e} = 0\}$ if and only if $\boldsymbol{p} + A^\top\lambda = 0$ for a certain vector λ. Denote $\boldsymbol{p} := \triangledown f(\boldsymbol{x}^*) = 0, A := \triangledown h(\boldsymbol{x}^*)$, then there exists Lagrange multipliers $\lambda^* \in \mathbb{R}^m$ s.t.

$$0 = \triangledown f(\boldsymbol{x}^*) + [\triangledown h(\boldsymbol{x}^*)]^\top\lambda^* = \triangledown f(\boldsymbol{x}^*) + \sum_{i=1}^{m} \lambda_i^* \triangledown h_i(\boldsymbol{x}^*).$$

Uniqueness of λ^*: From the above proof, $-\triangledown f(\boldsymbol{x}^*) = [\triangledown h(\boldsymbol{x}^*)]^\top\lambda^*$. Since $\boldsymbol{x}^*$ is regular, $\triangledown h(\boldsymbol{x}^*)$ has full rank, thus λ^* is unique. □

Remark 3.2.6. To see the necessity of the regularity of a feasible solution, we introduce the following example. Let's consider the ECP:

$$\min_{(x_1,x_2)\in\mathbb{R}^2} x_2 \text{ s.t. } \begin{cases} h_1(\boldsymbol{x}) := (x_1-1)^2 + x_2^2 = 1 \\ h_2(\boldsymbol{x}) := (x_1+1)^2 + x_2^2 = 1 \end{cases}.$$

Then,

$$\triangledown h_1(\boldsymbol{x}) = \begin{pmatrix} 2(x_1-1) \\ 2x_2 \end{pmatrix} = 2\binom{x_1}{x_2} - 2\binom{1}{0},$$
$$\triangledown h_2(\boldsymbol{x}) = \begin{pmatrix} 2(x_1+1) \\ 2x_2 \end{pmatrix} = 2\binom{x_1}{x_2} + 2\binom{1}{0}.$$

So $\triangledown h_1(x)$ and $\triangledown h_2(x)$ are linearly dependent at feasible point $(x_1, x_2) = (0, 0)$, which means the feasible point $(x_1, x_2) = (0, 0)$ is not regular. Denote $\lambda = (\lambda_1, \lambda_2)^\top$, the Lagrange function is

$$\mathcal{L}(\boldsymbol{x}, \lambda) = \mathcal{L}(\boldsymbol{x}, \lambda_1, \lambda_2) = x_2 + \lambda_1[(x_1-1)^2 + x_2^2 - 1] + \lambda_2[(x_1+1)^2 + x_2^2 - 1] = 0.$$

Then,

$$\frac{\partial\mathcal{L}(\boldsymbol{x}, \lambda_1, \lambda_2)}{\partial x} = \binom{0}{1} + 2\lambda_1[\binom{x_1}{x_2} - \binom{1}{0}] + 2\lambda_2[\binom{x_1}{x_2} + \binom{1}{0}] = 2(\lambda_1+\lambda_2)\binom{x_1}{x_2} + \begin{pmatrix} 2\lambda_2 - 2\lambda_1 \\ 1 \end{pmatrix}.$$

Note that the only feasible point is $(x_1, x_2) = (0, 0)$, so there doesn't exists (λ_1, λ_2) s.t. $\frac{\partial\mathcal{L}(\boldsymbol{x},\lambda_1,\lambda_2)}{\partial\boldsymbol{x}} = 0$.

Theorem 3.2.7 (Second Order Optimality conditions for ECP)**.** *[Nem99, Theorem 8.2.3] Let $\boldsymbol{x}^*$ be a feasible solution for the ECP* (3.3)*, let $\boldsymbol{x}^*$ be regular for the system of constraints of the problem* (3.3)*, and let*

$$\mathcal{L}(\boldsymbol{x}; \lambda) := f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i h_i(\boldsymbol{x}) = f(\boldsymbol{x}) + \lambda^\top h(\boldsymbol{x})$$

be the Lagrange function of the problem. Then.

1. *[Second Order Necessary Optimality condition]*
   Let x^* be a locally optimal solution to the problem. Then $\boldsymbol{x}^*$ satisfies the First Order Optimality condition:

$$\exists\lambda^* : \ \triangledown_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}^*; \lambda^*) = 0; \ \triangledown_\lambda\mathcal{L}(\boldsymbol{x}^*; \lambda^*) = 0$$

   and, besides this, the Hessian of the Lagrangian with respect to $\boldsymbol{x}$, reduced to the tangent plane is positive semidefinite:

$$\triangledown h(\boldsymbol{x}^*)e = 0 \implies \boldsymbol{e}^\top[\triangledown_{\boldsymbol{x}}^2\mathcal{L}(\boldsymbol{x}^*; \lambda^*)]\boldsymbol{e} \geq 0.$$

2. *[Second Order Sufficient Optimality condition]*
   Let $\boldsymbol{x}^*$ satisfy the First Order Optimality condition

$$\exists \lambda^* : \ \nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^*; \lambda^*) = 0; \ \nabla_{\lambda} \mathcal{L}(\boldsymbol{x}^*; \lambda^*) = 0$$

   and let, besides this, the Hessian of the Lagrangian with respect to $\boldsymbol{x}$, reduced to the tangent plane, be positive definite:

$$\nabla h(\boldsymbol{x}^*) e = 0, \boldsymbol{e} \neq 0 \implies \boldsymbol{e}^\top [\nabla^2_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^*; \lambda^*)] \boldsymbol{e} > 0.$$

   Then $\boldsymbol{x}^*$ is locally optimal solution to the problem.

3.2.2 Augmented Lagrange Multiplier Method

Although the Lagrange rule is powerful, in practice, it is computationally costly or even impossible to get the closed-form solution from the Lagrange rule. In this section, we introduce an iterative way to solve ECP (3.3) which is called the augmented Lagrange multiplier method. We consider the same ECP (3.3).

Definition 3.2.8 (Non-degenerate solution)**.** A feasible solution $\boldsymbol{x}^*$ to ECP (3.3) is called a non-degenerate solution to the ECP if $\boldsymbol{x}^*$

1. is regular for the constraints of the problem;
2. satisfies the second-order sufficient optimality conditions, i.e.,

$$\exists \lambda^* : \ \nabla_{\boldsymbol{x}} L(\boldsymbol{x}^*; \lambda^*) = 0; \ \nabla_{\lambda} L(\boldsymbol{x}^*; \lambda^*) = 0, \tag{3.7}$$

   and the Hessian of the Lagrange function w.r.t. $\boldsymbol{x}$ reduced to the tangent plane is positive definite:

$$\nabla h(\boldsymbol{x}^*) e = 0, \boldsymbol{e} \neq 0 \implies \boldsymbol{e}^\top [\nabla^2_{\boldsymbol{x}} L(\boldsymbol{x}^*; \lambda^*)] \boldsymbol{e} > 0.$$

Remark 3.2.9. Above definition involves both $\boldsymbol{x}^*$ and the Lagrange multiplier λ^*, which is a property of the pair $(\boldsymbol{x}^*, \lambda^*)$. Since $\boldsymbol{x}^*$ is regular for the constraints, by Theorem 3.2.5, there exists a Lagrange multiplier λ^* satisfying (3.7), and is uniquely defined by $\boldsymbol{x}^*$, so the definition is indeed a property of $\boldsymbol{x}^*$ alone.

Back to the ECP (3.3), let us add a quadratic penalty term so that the penalized objective: $f_\rho(\boldsymbol{x}) = f(\boldsymbol{x}) + \frac{1}{2}\rho \sum_{i=1}^m h_i^2(\boldsymbol{x})$. And we define ECP$_\rho$:

$$\min_{\boldsymbol{x} \in \mathbb{R}^d} f_\rho(\boldsymbol{x}) \text{ subject to } h(\boldsymbol{x}) = 0. \tag{3.8}$$

Note that the penalty term is zero on the feasible surface, so ECP$_\rho$ (3.8) is equivalent to ECP (3.3). The following lemma shows the advantage of ECP$_\rho$ (3.8).

Lemma 3.2.10. *Let $\boldsymbol{x}^*_\rho$ be a non-degenerate local solution to ECP$_\rho$* (3.8)*, and let*

$$\mathcal{L}_\rho(\boldsymbol{x}, \lambda) := f_\rho(\boldsymbol{x}) + \lambda^\top h(\boldsymbol{x}) = f_\rho(\boldsymbol{x}) + \sum_{i=1}^m \lambda_i h_i(\boldsymbol{x})$$

be the Lagrange function of ECP$_\rho$ (3.8)*. Let λ^*_ρ denote the vector of Lagrange multipliers corresponding to the solution $\boldsymbol{x}^*_\rho$ so that $\nabla_{\boldsymbol{x}} \mathcal{L}_\rho(\boldsymbol{x}^*, \lambda^*) = 0$,*

1. *let $\boldsymbol{x}^*$ denote the non-degenerate solution to ECP* (3.3)*, and let λ^* be the Lagrange multiplier corresponding to $\boldsymbol{x}^*$ s.t. $\nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^*, \lambda^*) = 0$, then $\boldsymbol{x}^* = \boldsymbol{x}^*_\rho$, $\lambda^* = \lambda^*_\rho$.*
2. *for ρ large enough, the matrix $\nabla^2_{\boldsymbol{x}} \mathcal{L}_\rho(\boldsymbol{x}^*, \lambda^*)$ is positive definite.*

Proof. 1. Since ECP (3.3) and ECP$_\rho$ (3.8) are equivalent on the feasible surface, then $\boldsymbol{x}^* = \boldsymbol{x}^*_\rho$. We write always $\boldsymbol{x}^*$ for simplicity. Note that

$$\nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^*, \lambda^*) = \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^*) + \sum_{i=1}^m \lambda_i^* \nabla h_i(\boldsymbol{x}^*) = 0,$$

$$\nabla_{\boldsymbol{x}} f_\rho(\boldsymbol{x}) = \nabla_{\boldsymbol{x}} f(\boldsymbol{x}) + \rho \sum_{i=1}^m h_i(\boldsymbol{x}) \nabla h_i(\boldsymbol{x}) \overset{h(\boldsymbol{x}^*)=0}{\Longrightarrow} \nabla_{\boldsymbol{x}} f_\rho(\boldsymbol{x}^*) = \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^*).$$

   Since λ^*_ρ is uniquely defined by $\nabla_{\boldsymbol{x}} \mathcal{L}_\rho(\boldsymbol{x}^*, \lambda^*_\rho) = \nabla_{\boldsymbol{x}} f_\rho(\boldsymbol{x}^*) + \sum_{i=1}^m \lambda^*_{\rho_i} \nabla_i (\boldsymbol{x}^*) = 0$, then $\lambda^*_\rho = \lambda^*$, thus we write always λ^* for simplicity.

2. Note that

$$\mathcal{L}_\rho(\boldsymbol{x},\lambda) = f_\rho(\boldsymbol{x}) + \sum_{i=1}^m \lambda_i h_i(\boldsymbol{x}) = f(\boldsymbol{x}) + \sum_{i=1}^m \lambda_i h_i(\boldsymbol{x}) + \frac{1}{2}\rho \sum_{i=1}^m h_i(\boldsymbol{x})^2 = \mathcal{L}(\boldsymbol{x},\lambda) + \frac{1}{2}\rho \sum_{i=1}^m h_i(\boldsymbol{x})^2,$$

then

$$\nabla_{\boldsymbol{x}} \mathcal{L}_\rho(\boldsymbol{x},\lambda) = \nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x},\lambda) + \rho \sum_{i=1}^m h_i(\boldsymbol{x})\, \nabla_{\boldsymbol{x}}\, h_i(\boldsymbol{x}).$$

Furthermore,

$$\nabla^2_{\boldsymbol{x}} \mathcal{L}_\rho(\boldsymbol{x},\lambda) = \underbrace{\nabla^2_{\boldsymbol{x}} \mathcal{L}_(\boldsymbol{x},\lambda)}_{=:H(\boldsymbol{x},\lambda)} + \rho \sum_{i=1}^m \nabla_{\boldsymbol{x}} h_i(\boldsymbol{x}) [\nabla_{\boldsymbol{x}} h_i(\boldsymbol{x})]^\top + \underbrace{\rho \sum_{i=1}^m h_i(\boldsymbol{x})\, \nabla^2_{\boldsymbol{x}}\, h_i(\boldsymbol{x})}_{=:B(\boldsymbol{x})}.$$

Denote $A(\boldsymbol{x}) := \nabla_{\boldsymbol{x}} h(\boldsymbol{x}) = \begin{pmatrix} [\nabla h_1(\boldsymbol{x})]^\top \\ \vdots \\ [\nabla h_m(\boldsymbol{x})]^\top \end{pmatrix}$, then

$$\nabla^2_{\boldsymbol{x}} \mathcal{L}_\rho(\boldsymbol{x},\lambda) = H(\boldsymbol{x},\lambda) + \rho A^\top(\boldsymbol{x}) A(x) + \rho B(\boldsymbol{x}).$$

By assumption, $\boldsymbol{x}^*$ is a non-degenerate solution of (ECP), i.e., $H(\boldsymbol{x}^*,\lambda^*)$ is positive definite at the subspace T of directions tangent to feasible surface at point $\boldsymbol{x}^*$. Note that for any $\boldsymbol{s} \in \mathbb{R}^d$, $\boldsymbol{s}^\top A^\top(\boldsymbol{x}^*) A(\boldsymbol{x}^*)\boldsymbol{s} = |A(\boldsymbol{x}^*)\boldsymbol{s}|^2 \geq 0$, i.e., $A^\top(\boldsymbol{x}^*)A(\boldsymbol{x}^*)$ is positive semidefinite. Note that $\boldsymbol{x}^*$ is feasible, then for any $\boldsymbol{s} \in \mathbb{R}^d$:

$$\boldsymbol{s}^\top\, \nabla^2_{\boldsymbol{x}}\, \mathcal{L}_\rho(\boldsymbol{x}^*,\lambda^*)\boldsymbol{s} = \boldsymbol{s}^\top H(\boldsymbol{x}^*,\lambda^*)\boldsymbol{s} + \rho \boldsymbol{s}^\top A^\top(\boldsymbol{x}^*) A(\boldsymbol{x}^*)\boldsymbol{s}. \tag{3.9}$$

Let T be the kernel of matrix $A(\boldsymbol{x}^*)$: $T := \{\boldsymbol{s} | A(\boldsymbol{x}^*)\boldsymbol{s} = 0\} = \{\boldsymbol{s} |\, \nabla\, h(\boldsymbol{x}^*)\boldsymbol{s} = 0\}$.
If $\boldsymbol{s} \in T$, (3.9) $= \boldsymbol{s}^\top H(\boldsymbol{x}^*,\lambda^*)\boldsymbol{s} > 0$ by the non-degeneration of $\boldsymbol{x}^*$.
If $\boldsymbol{s} \notin T : \boldsymbol{s}^\top A^\top(\boldsymbol{x}^*)A(\boldsymbol{x}^*)\boldsymbol{s} > 0$, then (3.9) > 0 for ρ large enough.

□

Lemma 3.2.11. *[Nem99, Theorem 11.1.1] If $\nabla^2_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x},\lambda)$ is positive definite on entire space, then there exists a convex neighbourhood V of $\boldsymbol{x}^*$ in $\mathbb{R}^d$ and a convex neighbourhood Λ of λ^* in $\mathbb{R}^m$ s.t.*

1. *for every $\lambda \in \Lambda$, $\mathcal{L}_\lambda(\boldsymbol{x}) : \boldsymbol{x} \mapsto \mathcal{L}(\boldsymbol{x},\lambda)$ is strongly convex in V, so the critical point $\boldsymbol{x}^*(\lambda)$ is uniquely defined in V, and $\boldsymbol{x}^*(\lambda)$ is a non-degenerate minimizer of $\mathcal{L}_\lambda$ in V.*
2. *Define optimal value of $\mathcal{L}_\lambda(\boldsymbol{x})$ in V as: $\underline{\mathcal{L}}(\lambda) = \mathcal{L}_\lambda(\boldsymbol{x}^*(\lambda))$, $\underline{\mathcal{L}}(\lambda) : \lambda \mapsto \mathcal{L}_\lambda(\boldsymbol{x}^*(\lambda))$ is concave and twice continuous differentiable in Λ with gradient $\nabla_\lambda \underline{\mathcal{L}}(\lambda) = \overline{h}(\boldsymbol{x}^*(\lambda))$.*

Remark 3.2.12. Since $\boldsymbol{x}^*(\lambda^*)$ is feasible, by Lemma 3.2.11 statement 2 it follows $\nabla_\lambda \underline{\mathcal{L}}(\lambda^*) = 0$. Due to the concavity of $\underline{\mathcal{L}}(\lambda)$, $\underline{\mathcal{L}}$ attains its maximum over Λ at point λ^*. By Lemma 3.2.11: $(\boldsymbol{x}^*,\lambda^*)$ is local saddle point of Lagrange function, more precisely, there exists a neighbourhood of $(\boldsymbol{x}^*,\lambda^*)$ (namely, $V \times \Lambda$) s.t. $\mathcal{L}|_{(V\times\Lambda)}$ attains at $(\boldsymbol{x}^*,\lambda^*)$ its minimum in $\boldsymbol{x}$ and its maximum in λ. Therefore, to find $\boldsymbol{x}^*$, it's same as to solve the dual problem: (D^ρ): $\max\limits_{\lambda} \underline{\mathcal{L}}(\lambda)$ where $\underline{\mathcal{L}}(\lambda) = \min\limits_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x},\lambda)$.

Theorem 3.2.13 (Augmented Lagrange Multiplier Method)**.** *[Nem99, section 11.2] To solve ECP* (3.3)*, augmented Lagrange function is defined as: $\mathcal{L}_\rho(\boldsymbol{x},\lambda) := f(\boldsymbol{x}) + \lambda^\top h(\boldsymbol{x}) + \frac{\rho}{2}\|h(\boldsymbol{x})\|^2$. At $k-$th iteration, we update*

$$\begin{aligned} \boldsymbol{x}_k &= \arg\min_{\boldsymbol{x}} \mathcal{L}_\rho(\boldsymbol{x},\lambda_{k-1}), \\ \lambda_k &= \lambda_{k-1} + \rho h(\boldsymbol{x}_k). \end{aligned}$$

Proof. Consider ECP$_\rho$ (3.8): $\min\limits_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) + \frac{\rho}{2}\|h(\boldsymbol{x})\|^2$ s.t. $h(\boldsymbol{x}) = 0$ which is equivalent to ECP (3.3) on the feasible surface. ECP$_\rho$ has the Lagrange function:

$$\mathcal{L}_\rho(\boldsymbol{x},\lambda) = f(\boldsymbol{x}) + \frac{\rho}{2} \sum_{i=1}^m h_i^2(\boldsymbol{x}) + \sum_{i=1}^m \lambda_i h_i(\boldsymbol{x})$$

By Lemma 3.2.11: Take ρ large enough s.t. the Hessian $\nabla^2_{\boldsymbol{x}}\mathcal{L}_\rho(\boldsymbol{x},\lambda)$ taken at point $(\boldsymbol{x}^*,\lambda^*)$ is positive definite on the entire space, where $\boldsymbol{x}^*$ denotes the non-degenerate solution of (ECP) to be approximated, and let λ^* be the corresponding vector of Lagrange multipliers (existence by Theorem 3.2.5). By Remark 3.2.12: To find $\boldsymbol{x}^*$, it's enough to solve the dual problem (D^ρ) :

$$\max_{\lambda\in\mathbb{R}^m} \underline{\mathcal{L}_\rho}(\lambda) \text{ where } \underline{\mathcal{L}_\rho}(\lambda) = \min_{\boldsymbol{x}\in\mathbb{R}^d} \mathcal{L}_\rho(\boldsymbol{x},\lambda).$$

So at $k-$th step: we update

$$\begin{aligned}\boldsymbol{x}_k &= \arg\min_{\boldsymbol{x}} \mathcal{L}_\rho(\boldsymbol{x},\lambda_{k-1}),\\ \lambda_k &= \arg\min_{\lambda} \mathcal{L}_\rho(\boldsymbol{x}_k,\lambda).\end{aligned}$$

By Lemma 3.2.11: $\nabla_\lambda\mathcal{L}_\rho(\boldsymbol{x}_k,\lambda) = h(\boldsymbol{x}_k)$, so λ_k can be updated by gradient descent:

$$\lambda_k = \lambda_{k-1} + \rho h(\boldsymbol{x}_k).$$

□

3.2.3 Linear NOTEARS Algorithm

Recall that to estimate the dependency structure under linear SEM, we aim at solving the ECP (3.2):

$$\min_{W\in\mathbb{R}^{d\times d}} \frac{1}{2n}\|\mathbb{X}-\mathbb{X}W\|_F^2 + \lambda\|W\|_1 \text{ s.t. } h_{\exp}(W) = 0,$$

where $h_{\exp}(W) := tr(e^{W\circ W}) - d$. Note that (3.2) is a non-convex constrained problem since the feasible set $\{W : h_{\exp}(W) = 0\}$ is a non-convex set (see following counter-example).

Example 3.2.14. Consider $W_1 = \begin{pmatrix}0 & 1\\ 0 & 0\end{pmatrix}$ and $W_2 = \begin{pmatrix}0 & 0\\ 1 & 0\end{pmatrix}$. Then for W_1, we have following linear SEM:

$$\begin{aligned}X_1 &= \varepsilon_1,\\ X_2 &= X_1 + \varepsilon_2.\end{aligned}$$

For W_2, we obtain

$$\begin{aligned}X_1 &= X_2 + \varepsilon_1,\\ X_2 &= \varepsilon_2.\end{aligned}$$

Note that

$$W_1^k = 0\ \forall k \geq 2, \text{ so } h_{\exp}(W_1) = \sum_{k=0}^{\infty}\frac{tr(W_1^k)}{k!} - 2 = \frac{tr(W_1^0)}{0!} - 2 = h_{\exp}(W_2) = 0.$$

However,

$$h_{\exp}(\frac{1}{2}W_1 + \frac{1}{2}W_2) = \sum_{k=0}^{\infty}\frac{tr(\begin{pmatrix}0 & \frac{1}{2}\\ \frac{1}{2} & 0\end{pmatrix}^k)}{k!} - 2 \geq \frac{tr(\begin{pmatrix}0 & \frac{1}{2}\\ \frac{1}{2} & 0\end{pmatrix}^0)}{0!} + \frac{tr(\begin{pmatrix}\frac{1}{4} & 0\\ 0 & \frac{1}{4}\end{pmatrix})}{1!} - 2 \geq 2 + \frac{1}{2} \neq 0.$$

Zheng et al. [Zhe+18] developed an algorithm called linear NOTEARS based on the augmented Lagrange multiplier methods to solve the ECP (3.2) whose Lagrange function is

$$\mathcal{L}^\rho(W,\alpha) = \frac{1}{2n}\|\mathbb{X}-\mathbb{X}W\|_F^2 + \lambda\|W\|_1 + \frac{\rho}{2}|h_{\exp}(W)|^2 + \alpha h_{\exp}(W),$$

where $\rho > 0$ and α denotes the Lagrange multiplier.

Algorithm 1 Linear NOTEARS

1: **Input:** Data matrix $\mathbb{X}$, initial guess (W_0, α_0) with $W_0 \in \mathbb{R}^{d\times d}$, $\alpha_0 \in \mathbb{R}$, progress rate $\xi \in (0, 1)$, tolerance $\varepsilon > 0$, threshold $\omega > 0$.
2: **Output:** $\widehat{W}$, the estimated weighted adjacency matrix.
3: **for** $t \leftarrow 0, 1, 2 \ldots$ **do**
4: Solve primal $W_{t+1} \leftarrow \arg\min_W \mathcal{L}^\rho(W, \alpha_t)$ with ρ such that $h_{\exp}(W_{t+1}) < \xi h_{\exp}(W_t)$.
5: Dual ascent $\alpha_{t+1} \leftarrow \alpha_t + \rho h(W_{t+1})$.
6: **if** $h_{\exp}(W_{t+1}) < \varepsilon$, **then**
7: set $\widetilde{W} = W_{t+1}$ and break.
8: Threshold matrix $\widehat{W} = \widetilde{W} \cdot \mathbb{1}(|\widetilde{W}| > \omega)$.

Remark 3.2.15.

1. The subproblem in line 4 of algorithm 1 can be solved by the proximal quasi-Newton (PQN) method [Zho+14] which is given in detail in the appendix A.1.

2. In the realization of the linear NOTEARS algorithm from Zheng et al. [Zhe+18], the initial guess of (W_0, α_0) is simple zero matrix and constant 0 respectively. Meanwhile, the progress rate ξ is set to 0.25, if the condition $h(W_{t+1}) < \xi h(W_t)$ is not satisfied, then ρ is updated by 10 times of itself. The default value of the tolerance ε and threshold ω are 1×10^{-8} and 0.3 respectively.

4 Structure Learning in Linear Structural Equation Models Based on Adaptive Lasso

The linear NOTEARS algorithm uses strict thresholding which is not flexible for practical data. On the other hand, the penalty term for sparsity in loss function from (3.2) treats all coefficients with indistinctive penalty levels. This construction puts much more penalty on large coefficients rather than false positive ones (the weight of the edges be learned as existence but actually not). This may result in false positive edges that cannot be zeroed out and miss-learned results for edges with large weights. To avoid these problems, Xu et al. [Xu+22] developed the linear NOTEARS-AL algorithm which extends the linear NOTEARS algorithm with adaptive Lasso to achieve learning the sparse DAG structure from purely observational data and showed its asymptotic oracle properties.

One intuition is to apply adaptive penalty levels to different coefficients. More precisely, one can modify (3.2) to

$$\min_{W\in\mathbb{R}^{d\times d}} \frac{1}{2n}\|\mathbb{X}-\mathbb{X}W\|_F^2 + \lambda_n \sum_{i,j=1}^{d} |c_{ij}W_{ij}| \text{ subject to } h_{\exp}(W)=0, \tag{4.1}$$

where c_{ij} represents the specified penalty for W_{ij}. More details about how to choose c_{ij} are given in the section 4.2.

Expecting minor false positive edges can be shrunk to zeros while reserving the true edges at the same time, we would like the corresponding c_{ij} to be larger for the former and small for the latter so that minor false edges are heavily penalized and true positive edges are lightly penalized. Xu et al. [Xu+22] shows, with a proper choice of c_{ij} and under some mild conditions, the adaptive Lasso method satisfies asymptotic oracle properties that are described in detail in section 4.1.

ECP (4.1) can be solved in a similar way to solving the ECP (3.2) corresponding to the linear NOTEARS algorithm, namely, by the augmented Lagrange multiplier method. The first step of each iteration is to find the local minimum of the augmented Lagrange function with a fixed Lagrange multiplier α from the last iteration:

$$\min_{W\in\mathbb{R}^{d\times d}} L_n(W) = \frac{1}{2n}\|\mathbb{X}-\mathbb{X}W\|_F^2 + \lambda_n \sum_{i,j=1}^{d} |c_{ij}W_{ij}| + \frac{\rho}{2}|h_{\exp}(W)|^2 + \alpha h_{\exp}(W).$$

4.1 Asymptotic Oracle Properties

Definition 4.1.1 (Convergence in probability)**.** A sequence $\{D_n\}$ of random matrices converges to a random matrix D in probability (denoted by $D_n \xrightarrow{P} D$) if for all $\varepsilon > 0 : \lim_{n\to\infty} \mathbb{P}(\|D_n - D\|_F > \varepsilon) = 0$.

Xu et al. [Xu+22] considered the following two conditions:

C1. Under the assumed linear SEM: $X_j = w_j^\top X + \varepsilon_j, j = 1,\dots,d$, the random noise ε_j are identically independent distributed with mean 0 and variance $\sigma^2 < \infty$.

C2. For data matrix $\mathbb{X} \in \mathbb{R}^{n\times d}$, $\frac{1}{n}\mathbb{X}^\top\mathbb{X}$ converges to a positive definite matrix $D \in \mathbb{R}^{d\times d}$ in probability.

Notation 4.1.2. $W \in \mathbb{R}^{d\times d}$ denotes underlying true weighted adjacency matrix. Define

$\mathcal{A} = \{(i,j) : W_{ij} \neq 0\}\widehat{=}\{\text{indices for terms whose true parameters are non-zero}\}$,
$\mathcal{A}_c = \{(i,j) : W_{ij} = 0\}\widehat{=}\{\text{indices for terms that do not exist in the underlying true model}\}$,
$\mathcal{A}_i = \{(i,j) : (i,j) \in \mathcal{A}\}$, and assume w.l.o.g $\mathcal{A}_i = \{(i,1),(i,2),\dots,(i,d_i)\}$.

We assume that condition C2 holds, i.e., $\frac{1}{n}\mathbb{X}^\top\mathbb{X} \overset{n\to\infty}{\longrightarrow} D$. Let $D = \begin{pmatrix} D_{i0} & D_{i1} \\ D_{i2} & D_{i3} \end{pmatrix}$ where D_{i0} is a $d_i \times d_i$ symmetric matrix corresponding to the coefficients with index in $\mathcal{A}_i$. Then let $\mathbb{X}_{\mathcal{A}_i} := \begin{pmatrix} x_1^1 & \dots & x_{d_i}^1 \\ \vdots & \ddots & \vdots \\ x_1^n & \dots & x_{d_i}^n \end{pmatrix}$ denote the corresponding submatrix of the data matrix $\mathbb{X}$. Similarly, denote $W_{\mathcal{A}_i} = \begin{pmatrix} W_{1i} \\ \vdots \\ W_{d_i i} \end{pmatrix}$. Let $\boldsymbol{\varepsilon}_i := \begin{pmatrix} \varepsilon_{1i} \\ \vdots \\ \varepsilon_{ni} \end{pmatrix} \in \mathbb{R}^{n\times 1}$ be the random noise of n observations corresponding to the random variable X_i.
A sequence of random variables Y_n is said to be $O_p(n^m)$ from some $m \in \mathbb{R}$ if for any $\varepsilon > 0$, there exists a constant $M > 0$ and a positive integer N such that:

$$\mathbb{P}(|Y_n| \leq M \cdot n^m) \geq 1 - \varepsilon \text{ for all } n \geq N.$$

$\{Y_n\}$ is said to be $o_p(n^m)$ if

$$\lim_{n\to\infty} \mathbb{P}(|\frac{Y_n}{n^m}| \geq \varepsilon) = 0 \text{ for every positive } \varepsilon.$$

Lemma 4.1.3 (Local minimizer)**.** *[Xu+22, Theorem 1] Under the conditions C1 and C2, let $a_n := \lambda_n c_{\mathcal{A}}$ where $c_{\mathcal{A}}$ represents the random c_{ij}'s associated with the non-zero coefficients in W. If $a_n = o_p(n^q)$ for some $q \leq -\frac{1}{2}$, then there exists a local minimizer of $L_n(W)$, denoted by $\widehat{W}_n$, which means, there exists $\varepsilon > 0$ such that $L_n(W) \geq L_n(\widehat{W}_n)$ for all W with $\left\|W - \widehat{W}_n\right\|_F < \varepsilon$. Moreover, it satisfies $\left\|\widehat{W}_n - W\right\|_F = O_p(n^{-\frac{1}{2}})$.*

Proof. Let $\eta_n := n^p$ with $p < 0$, $\boldsymbol{\delta} = \begin{pmatrix} \delta_{1,1} & \delta_{1,2} & \dots & \delta_{1,d} \\ \delta_{2,1} & \delta_{2,2} & \dots & \delta_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{d,1} & \delta_{d,2} & \dots & \delta_{d,d} \end{pmatrix}$ with $\|\boldsymbol{\delta}\|_F = c$, where c is a constant.

$$\begin{aligned} D_n(\boldsymbol{\delta}) &:= L_n(W + \eta_n\boldsymbol{\delta}) - L_n(W) \\ &= l_n(W + \eta_n\boldsymbol{\delta}) - l_n(W) + \lambda_n \sum_{i,j} c_{ij}\{|W_{ij} + \eta_n\delta_{i,j}| - |W_{ij}|\} \\ &\quad + \frac{\rho}{2}\{|h_{\exp}(W + \eta_n\boldsymbol{\delta})|^2 - |h_{\exp}(W)|^2\} + \alpha\{h_{\exp}(W + \eta_n\boldsymbol{\delta}) - h_{\exp}(W)\}, \end{aligned}$$

where $l_n(W) = \frac{1}{2n}\|\mathbb{X} - \mathbb{X}W\|_F^2$. Let $a_{n,ij} := \lambda_n c_{ij}$, $(i,j) \in \mathcal{A}$, we obtain that for $n \to \infty$, i.e., $W + \eta_n\boldsymbol{\delta} \to W$,

$$\begin{aligned} D_n(\boldsymbol{\delta}) &\geq l_n(W + \eta_n\boldsymbol{\delta}) - l_n(W) + \sum_{(i,j)\in\mathcal{A}} a_{n,ij}\{|W_{ij} + \eta_n\delta_{i,j}| - |W_{ij}|\} + \\ &\quad \frac{\rho}{2}\{|h_{\exp}(W + \eta_n\boldsymbol{\delta})|^2 - |h_{\exp}(W)|^2\} + \alpha\{h_{\exp}(W + \eta_n\boldsymbol{\delta}) - h_{\exp}(W)\} \\ &\overset{\triangle\text{-eq.}}{\geq} l_n(W + \eta_n\boldsymbol{\delta}) - l_n(W) - \eta_n \sum_{(i,j)\in\mathcal{A}} a_{n,ij}|\delta_{i,j}| + \\ &\quad \frac{\rho}{2}\{|h_{\exp}(W + \eta_n\boldsymbol{\delta})|^2 - |h(W)|^2\} + \alpha\{h_{\exp}(W + \eta_n\boldsymbol{\delta}) - h_{\exp}(W)\} \\ &\overset{\text{Taylor}}{=} \nabla l_n(W)^\top(\eta_n\boldsymbol{\delta}) + \frac{1}{2}(\eta_n\boldsymbol{\delta})^\top[\nabla^2 l_n(W)](\eta_n\boldsymbol{\delta})(1 + o_p(1)) - \eta_n \sum_{(i,j)\in\mathcal{A}} a_{n,ij}|\delta_{i,j}| + \\ &\quad \frac{\rho}{2}\{|h_{\exp}(W + \eta_n\boldsymbol{\delta})|^2 - |h_{\exp}(W)|^2\} + \alpha\{h_{\exp}(W + \eta_n\boldsymbol{\delta}) - h_{\exp}(W)\} \\ &=: I + II + III + IV, \end{aligned}$$

where we define

$$\begin{aligned} I &:= \nabla l_n(W)^\top(\eta_n\boldsymbol{\delta}); \\ II &:= \frac{1}{2}(\eta_n\boldsymbol{\delta})^\top[\nabla^2 l_n(W)](\eta_n\boldsymbol{\delta})(1 + o_p(1)); \\ III &:= -\eta_n \sum_{(i,j)\in\mathcal{A}} a_{n,ij}|\delta_{i,j}|; \\ IV &:= \frac{\rho}{2}\{|h_{\exp}(W + \eta_n\boldsymbol{\delta})|^2 - |h_{\exp}(W)|^2\} + \alpha\{h_{\exp}(W + \eta_n\boldsymbol{\delta}) - h_{\exp}(W)\}. \end{aligned}$$

Denote $U := \mathbb{X} - \mathbb{X}W$, then by the Chain Rule

$$\begin{aligned}\frac{\partial l_n(W)}{\partial W_{ij}} &= \frac{1}{2n}\sum_{k,l=1}^{d}\frac{\partial\|U\|_F^2}{\partial U_{kl}}\cdot\frac{\partial(\mathbb{X}-\mathbb{X}W)_{kl}}{\partial W_{ij}} = \frac{1}{2n}\sum_{k,l=1}^{d}2U_{kl}\cdot\frac{\partial(-\mathbb{X}W)_{kl}}{\partial W_{ij}}\\ &= -\frac{1}{n}\sum_{k,l=1}^{d}(\mathbb{X}-\mathbb{X}W)_{kl}\cdot\frac{\partial\mathbb{X}_{k\cdot}W_{\cdot l}}{\partial W_{ij}} = -\frac{1}{n}\sum_{k=1}^{d}(\mathbb{X}-\mathbb{X}W)_{kj}\mathbb{X}_{ki}\\ &= -\frac{1}{n}\sum_{k=1}^{d}(\mathbb{X}^\top)_{ik}(\mathbb{X}-\mathbb{X}W)_{kj} = -\frac{1}{n}[\mathbb{X}^\top(\mathbb{X}-\mathbb{X}W)]_{ij}.\end{aligned}$$

In other words, $\frac{\partial l_n(W)}{\partial W} = -\frac{1}{n}[\mathbb{X}^\top(\mathbb{X}-\mathbb{X}W)]$.

For I,

$$\begin{aligned}I &= \nabla l_n(W)^\top(\eta_n\boldsymbol{\delta}) = \eta_n(\nabla l_n(W)^\top)\boldsymbol{\delta}\\ &= \eta_n(-\frac{1}{n}(\mathbb{X}-\mathbb{X}W)^\top\mathbb{X})\boldsymbol{\delta} = -\frac{1}{\sqrt{n}}\eta_n\underbrace{(\sqrt{n}\frac{1}{n}(\mathbb{X}-\mathbb{X}W)^\top\mathbb{X})}_{:=A}\boldsymbol{\delta}.\end{aligned}$$

Since

$$\begin{aligned}\|A\|_F &= \left\|\sqrt{n}(I-W^\top)\frac{1}{n}(\mathbb{X}^\top\mathbb{X}-D+D)\right\|_F \overset{\triangle}{\leq} \sqrt{n}(\left\|\frac{1}{n}\mathbb{X}^\top\mathbb{X}-D\right\|_F+\frac{1}{n}\|D\|_F)\underbrace{\left\|(I-W^\top)\right\|_F}_{:=B}\\ &= \sqrt{n}\left\|\frac{1}{n}\mathbb{X}^\top\mathbb{X}-D\right\|_F\cdot B+\frac{\|D\|_F}{\sqrt{n}}B,\end{aligned}$$

the second term is $O_p(n^{-\frac{1}{2}})$. Consider the first term, for any $\varepsilon>0$:

$$\mathbb{P}(\frac{\sqrt{n}\left\|\frac{1}{n}\mathbb{X}^\top\mathbb{X}-D\right\|_F\cdot B}{n^{-\frac{1}{2}}}\geq\varepsilon) = \mathbb{P}(n\left\|\frac{1}{n}\mathbb{X}^\top\mathbb{X}-D\right\|_F\cdot B\geq\varepsilon)\geq\mathbb{P}(\left\|\frac{1}{n}\mathbb{X}^\top\mathbb{X}-D\right\|_F\geq\frac{\varepsilon}{B}) = 0,$$

for $n\to\infty$ by condition C2, then $\sqrt{n}(\frac{1}{n}\mathbb{X}^\top\mathbb{X}-D)(I-W^\top) = O_p(n^{-\frac{1}{2}})$. Therefore, $I = -O_p(\frac{1}{\sqrt{n}}\eta_n)\boldsymbol{\delta}$.

For II,

$$II = \frac{1}{2}\eta_n^2\{\boldsymbol{\delta}^\top[\frac{1}{n}\mathbb{X}^\top\mathbb{X}]\boldsymbol{\delta}\}(1+o_p(1))>0.$$

For III,

$$III = -\eta_n\sum_{(i,j)\in\mathcal{A}}a_{n,ij}|\delta_{ij}|\geq-\eta_nAn^qd,$$

where $A=\max\{a_{n,ij}/q^n:(i,j)\in\mathcal{A}\}$, and $A\in O(1)$ by the assumption $a_n=o_p(n^q)$.

For IV, since $h(W)\geq0$ for all $W\in\mathbb{R}^{d\times d}$, we can drive that

$$IV = [h_{\exp}(W+\eta_n\boldsymbol{\delta})-h_{\exp}(W)]\{\frac{\rho}{2}[h_{\exp}(W+\eta_n\boldsymbol{\delta})+h_{\exp}(W)]+\alpha\}.$$

So for $\rho>0$ large enough, $IV\geq0$. Thus,

$$D_n(\boldsymbol{\delta})\geq-O_p(\frac{1}{\sqrt{n}}\eta_n)\boldsymbol{\delta}+\frac{1}{2}\eta_n^2\{\boldsymbol{\delta}^\top[\frac{1}{n}\mathbb{X}^\top\mathbb{X}]\boldsymbol{\delta}\}(1+o_p(1))-\eta_nAn^qd+IV.$$

For n large enough, $\frac{1}{n}\mathbb{X}^\top\mathbb{X}\to D$ by condition C2, then $II=O_p(n^{2p})$. When $p>-\frac{1}{2}$, $\frac{II}{-I}=O_p(n^{p+\frac{1}{2}})>1$ for $n\to\infty$, which means that II dominates I. Similarly, by $q\leq-\frac{1}{2}$, we obtain that II dominates III, then for any given $\varepsilon>0$, there exists a large enough constant c such that

$$\mathbb{P}\{\inf_{\|\delta\|=c}L_n(W+\eta_n\delta)>L_n(W)\}\geq1-\varepsilon. \tag{4.2}$$

By Extreme Value Theorem and the continuity of $L_n(W)$, there exists a minimizer of $L_n(W)$ over the compact set $\{W + \eta_n\delta : \|\delta\|_F \leq c\}$. Then by (4.2), with probability at least $1 - \varepsilon$, the minimizer is inside the ball $\{W + \eta_n\delta : \|\delta\|_F < c\}$, which implies the minimizer is a local minimizer of $L_n(W)$ over $\mathbb{R}^{d\times d}$. Hence, we can conclude that there exists a local minimizer of $L_n(W)$ (denoted by $\widehat{W}_n$) such that for $0 > p > -\frac{1}{2}$,

$$\left\|\widehat{W}_n - W\right\|_F = O_p(\eta_n) = O_p(n^{-\frac{1}{2}}).$$

□

Definition 4.1.4 (Triangular array)**.** A triangular array of random variables is of the form

$$\begin{array}{lll} Y_{11} & & \\ Y_{21} & Y_{22} & \\ Y_{31} & Y_{32} & Y_{33} \\ \vdots & & \end{array}$$

where the random variables in each row

1. are independent of each other;
2. have zero mean;
3. have finite variance.

Lemma 4.1.5 (Multivariate Lindeberg-Feller CLT)**.** *([Vaa00]) Suppose $\{\boldsymbol{y}_{ni}\}_{i\leq n}$ is a triangular array of $d \times 1$ random vectors such that $\boldsymbol{s}_n = \frac{1}{n}\sum_{i=1}^n(\boldsymbol{y}_{ni})$, $V_n := \frac{1}{n}\sum_{i=1}^n Var(\boldsymbol{y}_{ni}) \to V$ where V is positive definite. If for every $\varepsilon > 0$, $\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|\boldsymbol{y}_{ni}\|_2^2 \mathbb{1}_{\{\|\boldsymbol{y}_{ni}\|_2 \geq \varepsilon\sqrt{n}\}}] \to 0$, then $\sqrt{n}\boldsymbol{s}_n \xrightarrow{d} \mathcal{N}(0, V)$.*

Lemma 4.1.6. *Assume the mild condition: For all fixed $i = 1, \ldots, n$, it holds*

$$\max_{1\leq k\leq n} \left\|(\mathbb{X}_{\mathcal{A}_i})_k\right\|_2 = \max_{1\leq k\leq n} [(x_1^k)^2 + \cdots + (x_{d_i}^k)^2]^{\frac{1}{2}} = o(n^{\frac{1}{2}}),$$

and additionally let conditions C1 and C2 hold, then $\frac{1}{\sqrt{n}}\mathbb{X}_{\mathcal{A}_i}^\top \boldsymbol{\varepsilon}_i \xrightarrow{d} \mathcal{N}(0, \sigma^2 D_{i0})$.

Proof. Let $\boldsymbol{x}_{\mathcal{A}_i}^k := \begin{pmatrix} x_1^k \\ \vdots \\ x_{d_i}^k \end{pmatrix} \in \mathbb{R}^{d_i}$. Consider the vector $\boldsymbol{x}^k\varepsilon_{ki}$ where ε_{ki} is the random noise scalar, then

$$\begin{aligned} \boldsymbol{s}_n &:= \frac{1}{n}\sum_{k=1}^n \boldsymbol{x}_{\mathcal{A}_i}^k \varepsilon_{ki} = \frac{1}{n}\mathbb{X}_{\mathcal{A}_i}^\top \boldsymbol{\varepsilon}_i, \\ V_n &:= \frac{1}{n}\sum_{k=1}^n \mathrm{Var}(\boldsymbol{x}_{\mathcal{A}_i}^k \varepsilon_{ki}) = \frac{1}{n}\sum_{k=1}^n \sigma^2 \boldsymbol{x}_{\mathcal{A}_i}^k (\boldsymbol{x}_{\mathcal{A}_i}^k)^T = \frac{\sigma^2}{n}\mathbb{X}_{\mathcal{A}_i}^\top \mathbb{X}_{\mathcal{A}_i} \overset{\text{condition 2}}{\longrightarrow} \sigma^2 D_{i0}. \end{aligned}$$

1) $\boldsymbol{x}_{\mathcal{A}_i}^k \varepsilon_{ki}$ is a triangular array: By condition 1, $\varepsilon_{ki} \perp\!\!\!\perp \varepsilon_{kl}$. By condition 2, it follows

$$\mathbb{E}[\boldsymbol{x}_{\mathcal{A}_i}^k \varepsilon_{ki}] = 0 \ ; \ \mathrm{Var}(\boldsymbol{x}_{\mathcal{A}_i}^k \varepsilon_{ki}) = \sigma^2 \boldsymbol{x}_{\mathcal{A}_i}^k (\boldsymbol{x}_{\mathcal{A}_i}^k)^T < \infty.$$

2) Lindeberg-Feller condition:

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\left\|\boldsymbol{x}_{\mathcal{A}_i}^k \varepsilon_{ki}\right\|_2^2 \mathbb{1}_{\{\left\|\boldsymbol{x}_{\mathcal{A}_i}^k \varepsilon_{ki}\right\|_2 \geq \varepsilon\sqrt{n}\}}] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[|\varepsilon_{ki}|^2 \left\|\boldsymbol{x}_{\mathcal{A}_i}^k\right\|_2^2 \mathbb{1}_{\{|\varepsilon_{ki}|\left\|\boldsymbol{x}_{\mathcal{A}_i}^k\right\|_2 \geq \varepsilon\sqrt{n}\}}] \\ &\overset{\text{ass.}}{=} \frac{1}{n}\sum_{i=1}^n \mathbb{E}[|\varepsilon_{ki}|^2 o(n) \mathbb{1}_{\{|\varepsilon_{ki}| o(1) \geq \varepsilon\}}] = \mathbb{E}[\underbrace{|\varepsilon_i|^2 o(n) \mathbb{1}_{\{|\varepsilon_i| o(1) \geq \varepsilon\}}}_{:=T_n}]. \end{aligned}$$

Since $T_n \xrightarrow{P} 0$ and consequently, $\mathbb{E}[T_n] \to 0$, then by multivariate Lindeberg-Feller CLT, the statement holds. □

Remark 4.1.7. The mild condition in Lemma 4.1.6 is satisfied by the design since $\frac{(x_1^k)^2+\cdots+(x_{d_i}^k)^2}{n} \to 0$ as $n \to \infty$ for all $k = 1, \ldots, n$ and fixed i.

Theorem 4.1.8 (Oracle properties)**.** *[Xu+22, Theorem 1] Under conditions 1 and 2, let $b_n := \lambda_n c_{\mathcal{A}_c}$ where $c_{\mathcal{A}_c}$ represents the c_{ij}'s associated with the zero coefficients in W, let $\widehat{W_n}$ be the local minimizer from Lemma 4.1.3, if $\sqrt{n}b_n \to \infty$ as $n \to \infty$, then $\widehat{W_n}$ satisfies following properties:*

(Sparsity) $\lim_{n\to\infty} \mathbb{P}(\widehat{W}_{\mathcal{A}_c} = 0) = 1$,

(Asymptotic normality) For all $i \in \{1, 2, \ldots, d\}$: $\sqrt{n}(\widehat{W}_{\mathcal{A}_i} - W_{\mathcal{A}_i}) \xrightarrow{d} \mathcal{N}(0, \sigma^2 D_{i0}^{-1})$ as $n \to \infty$.

Proof. Sparsity: It's sufficient to show that for all $(i, j) \in \mathcal{A}_c$:

$$\begin{cases} \left.\frac{\partial L_n(W)}{\partial W_{ij}}\right|_{W=\widehat{W}} < 0 & \text{if } -\varepsilon_n < \widehat{W}_{ij} < 0, \\ \left.\frac{\partial L_n(W)}{\partial W_{ij}}\right|_{W=\widehat{W}} > 0 & \text{if } 0 < \widehat{W}_{ij} < \varepsilon_n, \end{cases}$$

with probability going to 1 where $\varepsilon_n = O(n^{-\frac{1}{2}})$. From the proof of Lemma 4.1.3: For $l_n(W) := \frac{1}{2n}\|\mathbb{X} - \mathbb{X}W\|_F^2$,

$$\begin{aligned} \frac{\partial l_n(\widehat{W})}{\partial W_{ij}} &= -\frac{1}{n}[\mathbb{X}^\top(\mathbb{X} - \mathbb{X}\widehat{W})]_{ij} = -\frac{1}{n}[\mathbb{X}^\top(\mathbb{X} - \mathbb{X}W + \mathbb{X}W - \mathbb{X}\widehat{W})]_{ij} \\ &= -\frac{1}{n}[(\mathbb{X}^\top)_{i\cdot}(\mathbb{X}_{\cdot j} - \mathbb{X}W_{\cdot j})] + \frac{1}{n}[\mathbb{X}^\top\mathbb{X}(\widehat{W} - W)]_{ij} \\ &= -\frac{1}{n}[(\mathbb{X}^\top)_{i\cdot}(\mathbb{X}_{\cdot j} - \mathbb{X}W_{\cdot j})] + \frac{1}{n}(\mathbb{X}^\top)_{i\cdot}[\mathbb{X}(\widehat{W} - W)]_{\cdot j} \\ &= -\frac{1}{n}[(\mathbb{X}^\top)_{i\cdot}(\mathbb{X}_{\cdot j} - \mathbb{X}W_{\cdot j})] + \frac{1}{n}(\mathbb{X}^\top)_{i\cdot}\sum_k \mathbb{X}_{\cdot k}(\widehat{W}_{kj} - W_{kj}). \end{aligned}$$

Together with Theorem 3.1.3,

$$\begin{aligned} \frac{\partial L_n(\widehat{W})}{\partial W_{ij}} =& -\frac{1}{n}[(\mathbb{X}^\top)_{i\cdot}(\mathbb{X}_{\cdot j} - \mathbb{X}W_{\cdot j})] + \frac{1}{n}(\mathbb{X}^\top)_{i\cdot}\sum_k \mathbb{X}_{\cdot k}(\widehat{W}_{kj} - W_{kj}) \\ &+ \lambda_n c_{ij}\,\mathrm{sign}\{\widehat{W}_{ij}\} + 2\widehat{W}_{ij}(\rho h_{\exp}(\widehat{W}) + \alpha)(e^{\widehat{W}\circ\widehat{W}})_{ij}^\top. \end{aligned}$$

By Lemma 4.1.3 and condition C2,

$$\begin{aligned} &\frac{1}{n}[(\mathbb{X}^\top)_{i\cdot}(\mathbb{X}_{\cdot j} - \mathbb{X}W_{\cdot j})] = O_p(n^{-\frac{1}{2}}), \\ &\frac{1}{n}(\mathbb{X}^\top)_{i\cdot}\sum_k \mathbb{X}_{\cdot k}(\widehat{W}_{kj} - W_{kj}) = O_p(n^{-\frac{1}{2}}). \end{aligned}$$

By Theorem 3.1.3, $h(W) \geq 0$, then for ρ large enough:

$$2\widehat{W}_{ij}(\rho h_{\exp}(\widehat{W}) + \alpha)(e^{\widehat{W}\circ\widehat{W}})_{ij}^\top \begin{cases} < 0 & \text{if } -\varepsilon_n < \widehat{W}_{ij} < 0, \\ > 0 & \text{if } 0 < \widehat{W}_{ij} < \varepsilon_n. \end{cases}$$

Since $\varepsilon_n = O(n^{-\frac{1}{2}})$, then $|2\widehat{W}_{ij}(\rho h_{\exp}(\widehat{W}) + \alpha)(e^{\widehat{W}\circ\widehat{W}})_{ij}^\top| = O(n^{-\frac{1}{2}})$.
Consider the term $\lambda_n c_{ij}\,\mathrm{sign}\{\widehat{W}_{ij}\}$, let $b_{n,ij} = \lambda_n c_{ij}$. Therefore, for all $(i, j) \in \mathcal{A}_c$, if $\sqrt{n}b_{n,ij} \to \infty$, then the sign of $\frac{\partial L_n(\widehat{W})}{\partial W_{ij}}$ is dominated by $\mathrm{sign}\{\widehat{W}_{ij}\}$.

Asymptotic normality: Denote

$$I(W) := \frac{1}{2n}\|\mathbb{X} - \mathbb{X}W\|_F^2,\ II(W) := \lambda_n \sum_{i,j} |c_{ij}W_{ij}|,\ III(W) = \frac{\rho}{2}|h_{\exp}(W)|^2 + \alpha h_{\exp}(W),$$

so that

$$L_n(W) = I(W) + II(W) + III(W),$$

then

$$\nabla_{\mathcal{A}_i} L_n(\widehat{W}_{\mathcal{A}_i}) = \nabla_{\mathcal{A}_i} I(\widehat{W}_{\mathcal{A}_i}) + \nabla_{\mathcal{A}_i} II(\widehat{W}_{\mathcal{A}_i}) + \nabla_{\mathcal{A}_i} III(\widehat{W}_{\mathcal{A}_i}).$$

By the Taylor expansion at $\widehat{W}_{\mathcal{A}_i} = W_{\mathcal{A}_i}$, as $n \to \infty$:

$$\begin{aligned}\nabla_{\mathcal{A}_i} I(\widehat{W}_{\mathcal{A}_i}) &= \nabla_{\mathcal{A}_i} I(W_{\mathcal{A}_i}) + [\nabla^2_{\mathcal{A}_i} I(W_{\mathcal{A}_i})](\widehat{W}_{\mathcal{A}_i} - W_{\mathcal{A}_i}) + (\widehat{W}_{\mathcal{A}_i} - W_{\mathcal{A}_i}) o_p(1) \\ &= -\frac{1}{n}\mathbb{X}^\top_{\mathcal{A}_i}(\mathbb{X}_{\mathcal{A}_i} - \mathbb{X}_{\mathcal{A}_i} W_{\mathcal{A}_i}) + \frac{1}{n}\mathbb{X}^\top_{\mathcal{A}_i}\mathbb{X}_{\mathcal{A}_i}(\widehat{W}_{\mathcal{A}_i} - W_{\mathcal{A}_i}) + (\widehat{W}_{\mathcal{A}_i} - W_{\mathcal{A}_i}) o_p(1).\end{aligned}$$

Let $C_{\mathcal{A}_i}$ denote corresponding adaptive penalty weights c_{ij}s, then

$$\begin{aligned}\nabla_{\mathcal{A}_i} II(\widehat{W}_{\mathcal{A}_i}) &= \lambda_n C_{\mathcal{A}_i} \operatorname{sign}(\widehat{W}_{\mathcal{A}_i}) \\ &= \lambda_n C_{\mathcal{A}_i} \operatorname{sign}(W_{\mathcal{A}_i}) + (\widehat{W}_{\mathcal{A}_i} - W_{\mathcal{A}_i}) o_p(1).\end{aligned}$$

Moreover,

$$\begin{aligned}\nabla_{\mathcal{A}_i} III(\widehat{W}_{\mathcal{A}_i}) &= (\rho h_{\exp}(\widehat{W}_{\mathcal{A}_i}) + \alpha) \nabla h(\widehat{W}_{\mathcal{A}_i}) \\ &= (\rho(tr(e^{\widehat{W}_{\mathcal{A}_i} \circ \widehat{W}_{\mathcal{A}_i}}) - d) + \alpha)(e^{\widehat{W}_{\mathcal{A}_i} \circ \widehat{W}_{\mathcal{A}_i}})^\top \circ 2\widehat{W}_{\mathcal{A}_i} \\ &= (\widehat{W}_{\mathcal{A}_i} - W^*_{\mathcal{A}_i}) o_p(1).\end{aligned}$$

Since $\widehat{W}_{\mathcal{A}_i}$ is the sub weighted adjacency matrix of the local minimizer $\widehat{W}_n$ of $L_n(W)$, then

$$\begin{aligned}0 = \nabla_{\mathcal{A}_i} L_n(\widehat{W}_{\mathcal{A}_i}) = &-\frac{1}{n}\mathbb{X}^\top_{\mathcal{A}_i}(\mathbb{X}_{\mathcal{A}_i} - \mathbb{X}_{\mathcal{A}_i} W_{\mathcal{A}_i}) + \frac{1}{n}\mathbb{X}^\top_{\mathcal{A}_i}\mathbb{X}_{\mathcal{A}_i}(\widehat{W}_{\mathcal{A}_i} - W_{\mathcal{A}_i}) + \lambda_n C_{\mathcal{A}_i} \operatorname{sign}(W_{\mathcal{A}_i}) \\ &+ (\widehat{W}_{\mathcal{A}_i} - W_{\mathcal{A}_i}) o_p(1). \qquad (4.3)\end{aligned}$$

By $a_n = \lambda_n C_{\mathcal{A}}$, $a_{n,ij} = o_p(n^{-\frac{1}{2}})$, and $\left\|\widehat{W}_{\mathcal{A}} - W_{\mathcal{A}}\right\|_F = O_p(n^{-\frac{1}{2}})$,

$$0 = (4.3) = -\frac{1}{n}\mathbb{X}^\top_{\mathcal{A}_i}(\mathbb{X}_{\mathcal{A}_i} - \mathbb{X}_{\mathcal{A}_i} W_{\mathcal{A}_i}) + \frac{1}{n}\mathbb{X}^\top_{\mathcal{A}_i}\mathbb{X}_{\mathcal{A}_i}(\widehat{W}_{\mathcal{A}_i} - W_{\mathcal{A}_i}) + o_p(n^{-\frac{1}{2}}),$$

then

$$\sqrt{n}\frac{1}{n}\mathbb{X}^\top_{\mathcal{A}_i}\mathbb{X}_{\mathcal{A}_i}(\widehat{W}_{\mathcal{A}_i} - W_{\mathcal{A}_i}) = \frac{1}{\sqrt{n}}\mathbb{X}^\top_{\mathcal{A}_i}(\mathbb{X}_{\mathcal{A}_i} - \mathbb{X}_{\mathcal{A}_i} W_{\mathcal{A}_i}) - o_p(1),$$

and equivalently,

$$\sqrt{n}(\widehat{W}_{\mathcal{A}_i} - W_{\mathcal{A}_i}) = \underbrace{(\frac{1}{n}\mathbb{X}^\top_{\mathcal{A}_i}\mathbb{X}_{\mathcal{A}_i})^{-1}}_{\overset{\text{condition 2}}{\longrightarrow} D^{-1}_{i0}} \cdot \underbrace{[\frac{1}{\sqrt{n}}\mathbb{X}^\top_{\mathcal{A}_i}(\mathbb{X}_{\mathcal{A}_i} - \mathbb{X}_{\mathcal{A}_i} W_{\mathcal{A}_i}) - o_p(1)]}_{:=I}.$$

By Lemma 4.1.6 and it's remark: $I \xrightarrow{d} \mathcal{N}(0, \sigma^2 D_{i0})$ which implies that

$$\sqrt{n}(\widehat{W}_{\mathcal{A}_i} - W_{\mathcal{A}_i}) \to \mathcal{N}(0, \sigma^2 D^{-1}_{i0}).$$

□

Remark 4.1.9. The sparsity property shows that given $a_n = o_p(n^q)$ for some $q \le -\frac{1}{2}$, $\sqrt{n}b_n \to \infty$, then linear NOTEARS-AL (algorithm 2) can consistently remove all irrelevant variables with probability tending to 1. The asymptotic normality property shows that by magnifying the difference by $\sqrt{n}$ for non-zero estimators, the pattern turns out to be a normal distribution.

4.2 Proper Choice of Specified Penalty

It remains to determine proper c_{ij}'s so that the non-zero related part $a_n = o_p(n^q)$ for some $q \leq -\frac{1}{2}$ and zero related part $\sqrt{n}b_n \to \infty$ are satisfied. Define the ordinary least square (OLS) estimates

$$\widehat{W}_{ols} = \arg\min_{W \in \mathbb{R}^{d\times d}} \frac{1}{2n}\|\mathbb{X} - \mathbb{X}W\|_F^2 \text{ subject to } h_{\exp}(W) = 0 \tag{4.4}$$

Similar to Lemma 4.1.3, we can obtain that $\left\|\widehat{W}_{ols} - W\right\|_F = O_p(n^{-\frac{1}{2}})$. Suppose $\lambda_n\sqrt{n} \to 0, \lambda_n n^{\frac{r+1}{2}} \to \infty$ for some specific $r > 0$, we can define $c_{ij} = \frac{1}{|\widehat{W}_{ols_{ij}}|^r}$, then for $(i, j) \in \mathcal{A}$,

$$\frac{a_n}{n^{-\frac{1}{2}}} = \frac{1}{|\widehat{W}_{ols_{ij}}|^r}\frac{\lambda_n}{n^{-\frac{1}{2}}} \to 0.$$

For $(i, j) \in \mathcal{A}_c$, $\sqrt{n}b_n = \frac{\lambda_n}{|\widehat{W}_{ols_{ij}}|^r}\sqrt{n}$. Note that $\widehat{W}_{ols}$ is $\sqrt{n}-$consistent, therefore,

$$\begin{aligned}\sqrt{n}b_n &\geq \frac{\lambda_n}{(|W_{ij}| + |Cn^{-\frac{1}{2}}|)^r}\sqrt{n} \\ &\geq \frac{\lambda_n}{(Cn^{-\frac{1}{2}})^r}\sqrt{n} = \frac{\lambda_n}{C}n^{\frac{r+1}{2}} \overset{n\to\infty}{\longrightarrow} \infty.\end{aligned}$$

4.3 NOTEARS with Adaptive Lasso

Hence, the algorithm is divided into two parts:

1. Solve $\widehat{W}_{ols}$ from Equation (4.4);
2. Plug adaptive penalty parameter $C = (c_{ij})_{(i,j)\in[d]^2}$ in Equation (4.1) and solve it.

Part 1: Similar to the linear NOTEARS algorithm, (4.4) can be solved by the augmented Lagrange multiplier method. More precisely, at $(k+1)-$th iteration, we update

$$\begin{aligned}&W_{k+1} = \arg\min_{W\in\mathbb{R}^{d\times d}} \frac{1}{2n}\|\mathbb{X} - \mathbb{X}W\|_F^2 + \frac{\rho}{2}|h_{\exp}(W)|^2 + \alpha_k h_{\exp}(W), \\ &\alpha_{k+1} = \alpha_k + \rho h_{\exp}(W_{k+1}).\end{aligned}$$

Then we obtain the adaptive penalty parameters matrix $C \in \mathbb{R}^{d\times d}$ with entries $c_{ij} = \frac{1}{|\widehat{W}_{ols_{ij}}|^\gamma}$.

Part 2: Let $W_C := C \circ W$ where $\circ$ denotes the Hadamard product. Consequently, $W = W_C \oslash C$ where $\oslash$ denotes the Hadamard division. Then (4.1) can be transformed as

$$\arg\min_{W_C\in\mathbb{R}^{d\times d}} \frac{1}{2n}\|\mathbb{X} - \mathbb{X}W_C \oslash C\|_F^2 + \lambda_n\|W_C\|_1 \text{ subject to } h_{\exp}(W_C \oslash C) = 0. \tag{4.5}$$

Similarly, ECP (4.5) can be solved by the augmented Lagrange multiplier method, i.e., at $(t+1)-$th iteration:

$$\begin{aligned}&W_{C_{t+1}} = \arg\min_{W_C\in\mathbb{R}^{d\times d}} \frac{1}{2n}\|\mathbb{X} - \mathbb{X}W_C \oslash C\|_F^2 + \lambda_n\|W_C\|_F + \frac{\rho}{2}|h_{\exp}(W_C \oslash C)|^2 + \alpha_t h_{\exp}(W_C \oslash C), \\ &\alpha_{t+1} = \alpha_t + \rho h_{\exp}(W_{C_{t+1}} \oslash C).\end{aligned} \tag{4.6}$$

Note that (4.6) can be solved by PQN method (appendix A.1), and we obtain the following gradient terms similarly to the proof of Lemma 4.1.3 and by Theorem 3.1.3 respectively:

$$\frac{\partial\frac{1}{2n}\|\mathbb{X} - \mathbb{X}W_C \oslash C\|_F^2}{\partial W_C} = -\frac{1}{n}\mathbb{X}^\top(\mathbb{X} - \mathbb{X}W_C \oslash C) \oslash C.$$

Let $\widehat{W}_C$ denote the solution of ECP (4.5), then $\widehat{W}_n = \widehat{W}_C \oslash C$ is the final estimate for W. Summing up above steps, Xu et al. [Xu+22] concluded the following algorithm:

Algorithm 2 Linear NOTEARS with adaptive lasso

Input: Data matrix $\mathbb{X}$, initial guess (W_0, α_0) with $W_0 \in \mathbb{R}^{d\times d}, \alpha_0 \in \mathbb{R}$, progress rate $\xi \in (0,1)$, and tolerance $\varepsilon > 0$.
Output: $\widehat{W}_n$, the estimated weighted adjacency matrix.
OLS Loop:
for $k \leftarrow 0, 1, 2, \ldots$ **do**
    Solve $W_{k+1} = \arg\min_{W\in\mathbb{R}^{d\times d}} \frac{1}{2n}\|\mathbb{X} - \mathbb{X}W\|_F^2 + \frac{\rho}{2}|h_{\exp}(W)|^2 + \alpha_k h_{\exp}(W)$ with ρ such that

$$h_{\exp}(W_{k+1}) < \varepsilon h_{\exp}(W_k).$$

    Dual ascent $\alpha_{k+1} \leftarrow \alpha_k + \rho h_{\exp}(W_{k+1})$.
    if $h_{\exp}(W_{k+1}) < \varepsilon$, **then**
        set $\widehat{W}_{ols} = W_{k+1}, C := 1 \oslash |\widehat{W}_{ols}|^\gamma, W_C := C \circ W$ and break.
Adaptive lasso loop:
for $t \leftarrow 0, 1, 2, \ldots$ **do**
    Solve $W_{C,t+1} = \arg\min_{W_C\in\mathbb{R}^{d\times d}} \frac{1}{2n}\|\mathbb{X} - \mathbb{X}W_C \oslash C\|_F^2 + \lambda_n\|W_C\|_F + \frac{\rho}{2}|h_{\exp}(W_C \oslash C)|^2 + \alpha_t h_{\exp}(W_C \oslash C)$
    with ρ such that $h_{\exp}(W_{C,t+1} \oslash C) < \xi h_{\exp}(W_{C,t} \oslash C)$.
    Dual ascent $\alpha_{t+1} = \alpha_t + \rho h_{\exp}(W_{C,t+1} \oslash C)$.
    if $h_{\exp}(W_{t+1}) < \varepsilon$, **then**
        set $\widehat{W}_{C,n} = W_{C,t+1}$ and break.
Return the adaptive estimate $\widehat{W}_n := \widehat{W}_{C,n} \oslash C$.

The results of NOTEARS-based methods are sensitive to the value of hyper-parameter λ_n. Xu et al. [Xu+22] utilizes cross-validation to find the optimal λ_n. With a set of candidates of λ_n, by NOTEARS-AL, we can obtain a set of candidate models $S = \{s_n, n = 1, 2 \ldots, N\}$, where $s_n = \{(i,j) \in \{1,\ldots,d\}^2 : \widehat{W}_{n.ij} \neq 0\}$.

Algorithm 3 The Cross-validation method

Step 0. Divide the data $\mathbb{X}$ into validation set $\mathbb{X}_v$ and training set $\mathbb{X}_t$, and $|\mathbb{X}_v| = n_v, |\mathbb{X}_t| = n_t, n_v + n_t = n$.
For model s_n $(n = 1, 2, \ldots, N)$. Using the training set $\mathbb{X}_t$, compute the solution $\widetilde{W}_n^t$, where

$$\widetilde{W}_n^t = \arg\min_{W\in\mathbb{R}^{d\times d}, W_{s_n^c}=0} \frac{1}{2n}\left\|\mathbb{X}_t - W^\top \mathbb{X}_t\right\|_F^2.$$

Evaluate the prediction performance of $\widetilde{W}_n^t$ on the validation set $\mathbb{X}_v$ by loss function $\frac{1}{2n}\left\|\mathbb{X}_v - (\widetilde{W}_n^t)^\top \mathbb{X}_v\right\|_F^2$.
Take the model $s \in \{S_n : n = 1, \ldots, N\}$ with smallest loss on validation set $\mathbb{X}_v$.

A more sufficient validation set is needed to find the best model to achieve model-selection consistency, especially when the candidate model set is large. Instead of using the traditional K-fold cross-validation method, whose validation set is only $1/K$ of data, Xu et al. [Xu+22] suggested swapping the proportion of the validation set and training set.

5 Structure Learning in General Non-Parametric Structural Equation Models

The linear dependence structure is a stringent restriction on the class of models. Zheng et al. [Zhe+20] extended the linear NOTEARS to general non-parametric SEMs by approximating the non-parametric relationships between the random variables by either multi-layer perceptrons or via a (truncated) basis expansion, namely, NOTEARS-MLP and NOTEARS-SOB. As in the case of linear NOTEARS, Zheng et al. [Zhe+18] remained the exponential acyclicity constraint $h_{\exp}(W) := tr(e^{W \circ W}) - d$ and also utilized the augmented Lagrange scheme. In section 5.1, we investigate both NOTEARS-MLP and NOTEARS-SOB in detail. Several follow-up works proposed other acyclicity characterizations [BAR22; Naz+23; NGZ20; Yu+19]. In section 5.2, we summarize those existing acyclicity constraints and discuss their stability. Besides the augmented Lagrange scheme, Bello, Aragam, and Ravikumar [BAR22] developed a preferable optimization approach called DAGMA which is studied in section 5.3.

Recall that $X = (X_1, \ldots, X_d)$ is a random vector whose dependency structure is encoded by a DAG $G = (V, E)$ on d nodes. Under the setting of general non-parametric SEM, we assume that for all $j \in [d]$, the conditional expectations have the form $\mathbb{E}[X_j | X_{\mathrm{pa}(j)}] = f_j(X) + \varepsilon_j$, where $f_j \colon \mathcal{X} \to \mathbb{R}$ does not depend on X_k if $k \notin \mathrm{pa}(j)$, and $(\varepsilon_j)_{j \in [d]}$ are stochastic error terms that are independent over j. Let $G(f)$ be the graph defined by f. We denote by $C^1(\mathcal{X})$ the space of continuously differentiable functions over $\mathcal{X}$. Furthermore, let $\mathbb{P}_X$ be the distribution of the random vector X, we use the usual notation for the $\mathbb{P}_X-$ equivalence classes of square-integrable functions w.r.t. measure $\mathbb{P}_X$, i.e., we denote $L_2(\mathcal{X}, \mathbb{P}_X)$ the (equivalence) class of real-valued functions $[f]$ such that $\int_{\mathcal{X}} f^2(x) \mathbb{P}_X(dx)$ is finite. We will drop dependence on the domain and on the underlying distribution $\mathbb{P}_X$ and simply write $L_2 := L_2(\mathcal{X})$ or $L_2 := L_2(\mathbb{R}^d)$ in cases when the domain is clear from the context. Furthermore, we denote $\|\cdot\|_\infty$ to be the essential supremum norm w.r.t Lebesque measure on a (subset) $\mathcal{X} \subset \mathbb{R}$. For an element $g \in C^1(\mathcal{X})$, $g : \mathcal{X} \mapsto \mathbb{R}$ we denote $\frac{\partial}{\partial x_k} g(x)$ for its partial derivative (as a map $x \mapsto \frac{\partial}{\partial x_k} g(x)$) and denote $\frac{\partial}{\partial x_k} g(x)|_{x=s}$ for its value at point $x = s$.

Given data matrix $\mathbb{X} \in \mathbb{R}^{n \times d}$ whose rows $x^i, i = 1, \ldots, n$, represent n i.i.d. observations. We consider the quadratic loss: $\ell(y, \widehat{y}) = \|y - \widehat{y}\|_2^2$, the aim is to estimate $f = (f_1, \ldots, f_d)$ by minimizing the score function:

$$\min_f L(f) = \frac{1}{2n} \sum_{j=1}^{d} \sum_{i=1}^{n} \ell(x_j^i, f_j(x^i)) \text{ subject to } G(f) \in \mathrm{DAG}. \tag{5.1}$$

Let $H^1(\mathbb{R}^d)$ denote a subset of $C^1(\mathcal{X})$ where the function and its derivative are both square-integrable. We assume that each $f_j \in H^1(\mathbb{R}^d)$.

5.1 Non-Parametric NOTEARS Algorithms

5.1.1 Acyclicity Characterization

Definition 5.1.1 (Weighted adjacency matrix)**.** Let $\partial_k f_j$ denote the partial derivative of $f_j(x)$ w.r.t. x_k. The weighted adjacency matrix $W = W(f) = W(f_1, \ldots, f_d) \in \mathbb{R}^{d \times d}$ is defined with entries $[W(f)]_{k,j} := \left\|\partial_k f_j\right\|_{L^2}$.

Remark 5.1.2. $[W(f)]_{k,j} := \left\|\partial_k f_j\right\|_{L^2} = (\int |\partial_k f_j|^2 d\mu)^{\frac{1}{2}} = 0$ if and only if $\partial_k f_j = 0$. Since $f_j \in C^1(\mathcal{X})$, $\partial_k f_j = 0$ is equivalent to that f_j is independent of X_k. Therefore, $W(f)_{kj}$ encodes the dependency structure among X_j.

Similar to the linear NOTEARS algorithm in chapter 3, we have the exponential acyclicity constraint $h_{\exp}(W(f)) := tr(e^{W \circ W}) - d$. Then the ECP (5.1) can be transformed to

$$\min_f L(f) \text{ subject to } h_{\exp}(W(f)) = 0. \tag{5.2}$$

5.1.2 Approximation Families

Since ECP (5.2) is infinite-dimensional, the key is to reduce it to a finite dimension. The general recipe is as follows:

1. Choose a model family for the conditional expectation $\mathbb{E}[X_j|X_{pa(j)}]$ (e.g. general nonparametric, additive, index, etc.);
2. Choose a suitable family of approximations that can be parametrized by some parameters θ (e.g. neutral networks, orthogonal series, etc.);
3. Translate the loss function $L(f)$ and constraint $W(f)$ into parametric forms $L(\theta)$ and $W(\theta)$ using the approximating family;
4. Solve the resulting finite-dimensional problem.

Zheng et al. [Zhe+20] developed two approximation families, the multilayer perceptrons and basis expansions.

Multilayer Perceptrons

Consider a multilayer perceptron (MLP) with h hidden layers and a single activation function $\sigma : \mathbb{R} \to \mathbb{R}$:

$$\text{MLP}(u; A^{(1)}, \ldots, A^{(h)}) = \sigma(A^{(h)}\sigma(\ldots A^{(2)}\sigma(A^{(1)}u)))$$

where $A^{(l)} \in \mathbb{R}^{m_l \times m_{l-1}}$, $m_0 = d$.

Theorem 5.1.3 (Universal Approximation Theorem [HSW89])**.** *For any $f_j \in H^1(\mathbb{R}^d)$, a MLP can approximate f_j arbitrarily well by increasing the number of the hidden layers h or the number of hidden neurons m_l in each layer.*

Proposition 5.1.4 (Independency identification)**.** *[Zhe+20, Proposition 1] Consider*

$$\mathcal{F} := \{f | f(u) = \text{MLP}(u; A^{(1)}, \ldots, A^{(h)}), f \text{ independent of } u_k\},$$
$$\mathcal{F}_0 := \{f | f(u) = \text{MLP}(u; A^{(1)}, \ldots, A^{(h)}), k\text{th col of } A^{(1)} = 0\},$$

then $\mathcal{F} = \mathcal{F}_0$.

Proof. 1. $\mathcal{F}_0 \subseteq \mathcal{F}$: for any $f_0 \in \mathcal{F}$, $f_0(u) = \text{MLP}(u; A^{(1)}, \ldots, A^{(h)})$, where $A^{(1)}_{bk} = 0$ for all $b = 1, \ldots, m_1$. Thus the linear function $A^{(1)}u$ is independent of u_k. Therefore,

$$f_0(u) = \text{MLP}(u; A^{(1)}, \ldots, A^{(h)}) = \sigma(A^{(h)}\sigma(\ldots A^{(2)}\sigma(A^{(1)}u)))$$

is independent of u_k, which means $f_0 \in \mathcal{F}$.

2. $\mathcal{F} \subseteq \mathcal{F}_0$: for any $f \in \mathcal{F}$, $f(u) = \text{MLP}\big(u; A^{(1)}, \ldots, A^{(h)}\big)$ and f is independent of u_k. We will show that $f \in \mathcal{F}_0$ by constructing a matrix $\tilde{A}^{(1)}$, such that

$$f(u) = \text{MLP}\Big(u; \tilde{A}^{(1)}, A^{(2)}, \ldots, A^{(h)}\Big)$$

and $\tilde{A}^{(1)}_{bk} = 0$ for all $b = 1, \ldots, m_1$. Let $\tilde{u}$ be the vector such that $\tilde{u}_k = 0$ and $\tilde{u}_{k'} = u_k$ for all $k' \neq k$.

Since $\tilde{u}$ and u differ only on the kth dimension, and f is independent of u_k, we have

$$f(u) = f(\tilde{u}) = \mathrm{MLP}\Big(\tilde{u}; A^{(1)}, \ldots, A^{(h)}\Big). \tag{5.3}$$

Now define $\tilde{A}^{(1)}$ be the matrix such that $\tilde{A}^{(1)}_{bk} = 0$ and $\tilde{A}^{(1)}_{bk'} = A^{(1)}_{bk}$ for all $k' \neq k$. Then we have the following observation: for each entry $s \in \{1, \ldots, m_1\}$,

$$\begin{aligned}\left(\tilde{A}^{(1)} u\right)_s &= \sum_{k'=1}^{d} \tilde{A}_{sk'} u_{k'} = \sum_{k' \neq k} A_{sk'} u_{k'} \\ &= \sum_{k'=1}^{d} A_{sk'} \tilde{u}_{k'} = \left(A^{(1)} \tilde{u}\right)_s.\end{aligned}$$

Hence,

$$\tilde{A}^{(1)} u = A^{(1)} \tilde{u}.$$

Therefore, by (5.3)

$$\begin{aligned} f(u) &= f(\tilde{u}) \\ &= \mathrm{MLP}\Big(\tilde{u}; A^{(1)}, \ldots, A^{(h)}\Big) \\ &= \sigma\Big(A^{(h)} \sigma\Big(\cdots A^{(2)} \sigma\Big(A^{(1)} \tilde{u}\Big)\Big)\Big) \\ &= \sigma\Big(A^{(h)} \sigma\Big(\cdots A^{(2)} \sigma\Big(\tilde{A}^{(1)} u\Big)\Big)\Big) \\ &= \mathrm{MLP}\Big(u; \tilde{A}^{(1)}, A^{(2)}, \ldots, A^{(h)}\Big). \end{aligned}$$

By definition of $\mathcal{F}_0$, we know that $\mathrm{MLP}\Big(u; \tilde{A}^{(1)}, A^{(2)}, \ldots, A^{(h)}\Big) \in \mathcal{F}_0$. Thus, $f \in \mathcal{F}_0$ which completes the proof.

□

Let $\theta_j = (A^{(1)}, \ldots, A^{(h)})$ denote the parameters for the jth MLP and $\theta = (\theta_1, \ldots, \theta_d)$. By Proposition 5.1.4, it follows $\left\|k\text{th-column}(A_j^{(1)})\right\|_2 = 0$ if and only if $[W(f)]_{kj} := \left\|\partial_k f_j\right\|_{L^2} = 0$. Therefore, Zheng et al. [Zhe+20] define $[W(\theta)]_{kj} := \left\|k\text{th-column}(A_j^{(1)})\right\|_2$. Moreover, to enforce sparsity, the regularization term $\left\|A_j^{(1)}\right\|_1 := \sum_{i,k} |(A_j^{(1)})_{ik}|$ is added to the score function. The problem (5.2) can be reduced to

$$\min_{\theta} \frac{1}{2n} \sum_{j=1}^{d} \sum_{i=1}^{n} (x_j^i - \mathrm{MLP}(x^i; \theta_j))^2 + \lambda \left\|A_j^{(1)}\right\|_1 \text{ s.t. } h_{\exp}(W(\theta)) = 0. \tag{5.4}$$

Basis Expansions

Theorem 5.1.5 (Approximation theorem [Sch67])**.** *Let $\{\varphi_r\}_{r=1}^{\infty}$ be the orthonormal basis of $H^1(\mathbb{R})$ s.t. $\mathbb{E}[\varphi_r(X)] = 0$ for all r, then for any $f \in H^1(\mathbb{R})$ can be written uniquely:*

$$f(u) = \sum_{r=1}^{\infty} \alpha_r \varphi_r(u), \ \alpha_r = \int_{\mathbb{R}^d} \varphi_r(u) f(u) du.$$

Zheng et al. [Zhe+20] assumed an additive model with one-dimensional expansions. More precisely,

$$f_j(u_1, \ldots, u_d) = \sum_{k \neq j} f_{jk}(u_k) = \sum_{r=1}^{\infty} \alpha_{jkr} \varphi_r(u_k).$$

Proposition 5.1.6 (Approximation error [Efr08]). *For any $f(u) = \sum_{r=1}^{\infty} \alpha_r \varphi_r(u)$, define finite series $\widehat{f}^R := \sum_{r=1}^{R} \alpha_r \varphi_r$. Given integers R_k, assume f_{jk} is sufficiently smooth, which means, f_{jk} is m times continuously differentiable for m large enough. Then $\left\| f_{jk} - \widehat{f}_{jk}^{R_k} \right\|_{L^2} = O(1/R_k)$, so the overall approximation error is on the order $O(d/\min_k R_k)$.*

Note that $[W(f)]_{kj} = \frac{\partial f_j}{\partial X_k} = 0$ if and only if $\alpha_{jkr} = 0$ for all r and in practice, we approximate f_{jk} by finite basis expansion, i.e., $f_{jk}(u_k) = \sum_{r=1}^{R_k} \alpha_{jkr} \varphi_r(u_k)$. It suffices to check that only $\alpha_{jkr} = 0$ for all $r = 1, \ldots, R_k$ which is equivalent to $\sum_{r=1}^{R_k} \alpha_{jkr}^2 = 0$, so let $\theta = \{\alpha_{jkr}\ \forall j, k, r\}$, Zheng et al. [Zhe+20] define $[W(\theta)]_{kj} = [\sum_{r=1}^{R_k} \alpha_{jkr}^2]^{\frac{1}{2}}$. Let Φ_k be the matrix $[\Phi_k]_{ir} = \varphi_r(X_k^{(i)}) \in \mathbb{R}^{n \times R_k}$. After adding orthogonal series smoothing [Rav+09] and ℓ_1-regularization term to enforce sparsity, the ECP (5.2) can be reduced to:

$$\min_{\theta} \frac{1}{2n} \sum_{j=1}^{d} \sum_{i=1}^{n} (x_j^i - \sum_{k \neq j} [\Phi_k]_{i\cdot} \cdot \alpha_{jk})^2 + \lambda_1 \sum_{k \neq j} \frac{1}{2n} \alpha_{jk}^\top \Phi^\top \Phi_k \alpha_{jk} + \lambda_2 \sum_{k \neq j} \left\| \alpha_{jk} \right\|_1 \text{ s.t. } h_{\exp}(W(\theta)) = 0. \quad (5.5)$$

5.1.3 Optimization

Both (5.4) and (5.5) are optimization problems with ℓ_1-regularization, which can be written in the following generic form:

$$\min_{\theta} L(\theta) + \lambda \|\theta\|_1 \text{ s.t. } h(W(\theta)) = 0. \quad (5.6)$$

As in the linear case, (5.6) can be solved by an augmented Lagrange scheme with the Lagrange function:

$$L^\rho(\theta_t, \alpha_t) = L(\theta_t) + \frac{\rho}{2} |h(W(\theta_t))|^2 + \alpha_t h(W(\theta_t)) + \lambda \|\theta\|_1.$$

Zheng et al. [Zhe+20] concluded the following non-parametric NOTEARS algorithm:

Algorithm 4 Non-parametric NOTEARS algorithm

1: **Input:** Data matrix $\mathbb{X}$, initial guess (W_0, α_0) with $W_0 \in \mathbb{R}^{d \times d}$, $\alpha_0 \in \mathbb{R}$, progress rate $\xi \in (0, 1)$, tolerance $\varepsilon > 0$ and threshold $\omega > 0$.
2: **Output:** $\widehat{W}$, the estimated weighted adjacency matrix.
3: **for** $t \leftarrow 0, 1, 2 \ldots$ **do**
4: Solve primal $\theta_{t+1} \leftarrow \arg\min_W L^\rho(\theta_t, \alpha_t)$ with ρ such that $h(\theta_{t+1}) < \xi h(\theta_t)$.
5: Dual ascent $\alpha_{t+1} \leftarrow \alpha_t + \rho h(W(\theta_t))$.
6: **if** $h(W(\theta_{t+1})) < \varepsilon$, **then**
7: set $\widetilde{W} = W(\theta_{t+1})$ and break.
8: Threshold matrix $\widehat{W} = \widetilde{W} \cdot \mathbb{1}(|\widetilde{W}| > \omega)$.

Remark 5.1.7.

1. Similarly, the subproblem in line 4 of the algorithm 4 can be solved by the proximal quasi-Newton (PQN) method [Zho+14] in the appendix A.1. Moreover, since the ℓ_1-regularizer is not differentiable at 0. Zheng et al. [Zhe+20] cast (5.6) into a box-constrained form:

$$\min_{\theta} F(\theta) + \lambda \|\theta\|_1 \iff \min_{\theta^+ \geq 0, \theta^- \geq 0} F(\theta^+ - \theta^-) + \lambda \mathbb{1}^\top (\theta^+ - \theta^-).$$

2. In the realization of the NOTEARS-MLP and NOTEARS-SOB algorithms from Zheng et al. [Zhe+20], the initial guess of α_0 is simple constant 0, and W_0 is randomly initialized. In case of NOTEARS-MLP, the sigmoid function is used as the activation function. Meanwhile, the progress rate ξ is set to 0.25, if the condition $h(W_{t+1}) < \xi h(W_t)$ is not satisfied, then ρ is updated by 10 times of itself. The default value of the tolerance ε and threshold ω are 1×10^{-8} and 0.3 respectively.

3. Let $H^1([-\pi, \pi])$ denote the space of functions over $[-\pi, \pi]$ that the functions and their derivatives are both square-integrable. Considering the norm $\|f\|_{H^1(S)} = (\int_S |f(x)|^2 dx + \int_S |f'(x)|^2 dx)^{\frac{1}{2}}$, $\{\sin(kx), \cos(kx), k = 1, \ldots, \infty\}$ is a basis of $H^1([-\pi, \pi])$.

Proof. For any $f \in H^1([-\pi, \pi])$, define

$$f_n(x) = \frac{a_0}{2} + \sum_{k=1}^{n}(a_k \cos(kx) + b_k \sin(kx)),$$

where

$$\begin{aligned} a_0 &= \frac{1}{\pi}\int_{\pi}^{-\pi} f(x)dx; \\ a_k &= \frac{1}{\pi}\int_{\pi}^{-\pi} f(x)\cos(nx)dx; \\ b_n &= \frac{1}{\pi}\int_{\pi}^{-\pi} f(x)\sin(nx)dx. \end{aligned}$$

By Dirichlet's Theorem, $f(x) = \lim_{n\to\infty} f_n(x)$. Hence, we obtain that

$$\begin{aligned} f'(x) &= \sum_{k=1}^{\infty}(-ka_k \sin(kx) + kb_k \cos(kx)); \\ f_n'(x) &= \sum_{k=1}^{n}(-ka_k \sin(kx) + kb_k \cos(kx)). \end{aligned}$$

By Parseval's Theorem:

$$\left\|f' - f_n'\right\|_{L^2}^2 = \pi \sum_{k=n+1}^{\infty} k^2(a_k^2 + b_k^2). \tag{5.7}$$

Since $f' \in L^2$, by Parseval's Theorem, $\sum_{k=n+1}^{\infty} k^2(a_k^2 + b_k^2) < \infty$. Therefore, $\left\|f' - f_n'\right\|_{L^2}^2 = (5.7) \to 0$ for $n \to \infty$. Similarly, $\|f - f_n\|_{L^2}^2 \to 0$ for $n \to \infty$. Thus,

$$\|f - f_n\|_{H^1} \to 0 \text{ as } n \to \infty,$$

which means $\{\sin(kx), \cos(kx), k = 1, \ldots, \infty\}$ is a basis of $H^1([-\pi, \pi])$. The orthogonality follows from the orthogonality of $\sin(kx)$ and $\cos(kx)$ for $k > 0$. □

When constructing NOTEARS-SOB, Zheng et al. [Zhe+20] used the orthogonal series $\varphi_r(u) = s_r \sin(u/s_r)$, $s_r = 2/((2r-1)\pi), r = 1, 2, \ldots, 10$ [HS12] to approximate functions in $H_1(\mathbb{R})$.

5.2 Stability of Acyclicity Constraints

Besides the exponential acyclicity constraint in the NOTEARS algorithm, there are several other acyclicity constraints. In this section, we summarize the class of existing acyclicity constraints and discuss their stability during optimization. Since $W \circ W$ has non-negative entries, we write $A \in \mathbb{R}_{\geq 0}^{d\times d}$ instead of $W \circ W$ in this chapter for simplicity.

Nazaret et al. [Naz+23] developed the definition of power series trace constraints and gave three criteria for stability. This paper showed that the power series trace constraints are unstable. Moreover, it also introduced the spectral acyclicity constraint and showed its stability.

Definition 5.2.1 (Power Series Trace Family)**.** [Naz+23, Definition 1] For any non-negative coefficients $(a_k)_{k\in\mathbb{N}^*} \in \mathbb{R}_{\geq 0}^{\mathbb{N}^*}$, consider the power series $f_a(x) = \sum_{k=1}^{\infty} a_k x^k$. Then for any matrix $A \in \mathbb{R}_{\geq 0}^{d\times d}$ with non-negative entries, the Power Series Trace (PST) function is defined as

$$h_a(A) = Tr[f_a(A)] = \sum_{k=1}^{\infty} a_k Tr[A^k].$$

Remark 5.2.2. Since $Tr(A^k)$ represents the total weight of all length-k cycles in $G(A)$ where the weight of a cycle is the product of its edge weights. $h_a(A)$ can be seen as a linear combination of weights of cycles in all possible lengths in the graph represented by A.

Theorem 5.2.3 (PST constraint)**.** *[Naz+23, Theorem 1] For any sequence $(a_k)_{k\in\mathbb{N}^*} \in \mathbb{R}_{\geq 0}^{\mathbb{N}^*}$, if we have $a_k > 0$ for all $k \in [d]$, then for any matrix $A \in \mathbb{R}_{\geq 0}^{d\times d}$, we have*

$$\begin{cases} h_a(A) = 0 \iff A \text{ is acyclic}, \\ h_a(A) \geq 0, \\ \nabla h_a(A) = \sum_{k=1}^{\infty} k a_k (A^\top)^{k-1}. \end{cases}$$

We say that h_a is a PST constraint.

Proof. For any matrix $A \in \mathbb{R}_{\geq 0}^{d\times d}$, we have

$$Tr[A^k] = \sum_{\substack{(i_0,\dots,i_k)\in[d]^{k+1} \\ i_0=i_k}} \prod_{\ell=1}^{k} A_{i_{\ell-1},i_\ell}. \tag{5.8}$$

$G(A)$ has a cycle of length k if and only if there exists $(i_0,\dots,i_k) \in [d]^{k+1}$ such that $i_0 = i_k$ and for all $\ell \in [k] : A_{i_{\ell-1},i_\ell} > 0$. Then by (5.8), $G(A)$ has a cycle of length-k if and only if $Tr[A^k] > 0$.
Since $a_k > 0$ for all $k \in [d]$, then

$$\begin{aligned} h_a(A) = 0 &\iff Tr[A^k] = 0.\ \forall k \in [d]. \\ &\iff A \text{ doesn't contain any cycle of any length } k \in [d]. \\ &\iff A \text{ is acyclic.} \end{aligned}$$

Since $A \geq 0$, then $Tr[A^k] \geq 0$ and consequently, $h_a(A) \geq 0$. The gradient of h_a follows directly by computation. □

Example 5.2.4. By Theorem 5.2.3, several standard power series are PST constraints. For example,

Name	α_k	f_a	h_a	$\nabla h_a^\top$
$h_{\exp}$	$\frac{1}{k!}$	$\exp(x) - 1$	$\operatorname{Tr}\exp(A) - d$	$\exp(A)$
$h_{\log}$	$\frac{1}{k}$	$\log\left(\frac{1}{1-x}\right)$	$-\log\det(I - A)$	$(I - A)^{-1}$
h_{inv}	1	$\frac{1}{1-x}$	$\operatorname{Tr}(I - A)^{-1}$	$(I - A)^{-2}$
h_{binom}	$\binom{d}{k}$	$(1 + x)^d - 1$	$\operatorname{Tr}(I + A)^d - d$	$d(I + A)^{d-1}$

Nazaret et al. [Naz+23] introduced following three criteria for acyclicity constraints to exhibit stable optimization.

Definition 5.2.5. [Naz+23, Definition 2] An acyclicity constraint h is stable if these criteria hold for almost every $A \in \mathbb{R}_{\geq 0}^{d\times d}$:

- **E-stable** $h(tA) = O_{t\to\infty}(t)$.
- **V-stable** If $h(A) \neq 0$, then $h(\varepsilon A) = \Omega_{\varepsilon\to 0^+}(\varepsilon)$, which means, for $\varepsilon \to 0^+$, $h(\varepsilon A) \geq c\varepsilon$ for some positive constant c.
- **D-stable** $h(A)$ and $\nabla h(A)$ are well-defined.

Remark 5.2.6. E-stability ensures that h does not explode to infinity; D-stability ensures that h and its gradient exist. V-stability ensures h does not vanish rapidly to 0. Without V-stability, $h(A_\theta)$ can shrink quickly very close to 0 during the optimization process, while $G(A_\theta)$ remains far from a DAG.

Theorem 5.2.7 (PST unstability)**.** *[Naz+23, Theorem 2] For $d \geq 2$, any PST constraint h_a is both E-unstable and V-unstable. More precisely,*

- ***E-unstable*** $\exists A \in \mathbb{R}_{\geq 0}^{d\times d}, h_a(tA) = \Omega_{s\to\infty}(t^d)$.
- ***V-unstable*** $\exists A \in \mathbb{R}_{\geq 0}^{d\times d}, h_a(\varepsilon A) = O_{\varepsilon\to 0^+}(\varepsilon^d)$.

Also, any PST constraint for which f_a has a finite radius of convergence is ***D-unstable*** *(e.g., $h_{\log}, h_{inv}$).*

Proof. Take a PST constraint h_a for some $(a_k)_k \in \mathbb{R}_{\geq 0}^{d\times d}$ with $a_k > 0$ for $k \in [d]$. The E-unstability and V-unstability are shown by a particular adjacency matrix C. Define C as the adjacency matrix of the cycle $1 \to 2 \to \ldots \to d \to 1$ with edges weights of 1. That is:

$$C = \begin{bmatrix} 0 & 1 & 0 & \ddots & \ddots & 0 \\ & 0 & 1 & 0 & \ddots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ & \vdots & & \ddots & \ddots & 0 \\ & & & & \ddots & 1 \\ 0 & & & & & \\ 1 & 0 & & \ddots & & 0 \end{bmatrix}.$$

We have $C^d = I_d$ and $\mathrm{Tr}\left[C^k\right] = \begin{cases} d & \text{if } k = 0 \bmod d \\ 0 & \text{if } k \neq 0 \bmod d \end{cases}$. We obtain for any $t \in \mathbb{R}_{\geq 0}$,

$$h_a(tC) = d\sum_{\ell=1}^{\infty} a_{\ell d} t^{\ell d}.$$

In particular, since the coefficient $(a_k)_{k\in\mathbb{N}^*} \geq 0$ and $a_d > 0$, we have for any

$$t \geq 0, h_a(tC) \geq d a_d t^d = \Omega_{t\to+\infty}(t^d).$$

This proves the E-instability.
Define $u = \min(1, r_a/2)$ where r_a is the radius of convergence of f_a. Then, for any $\varepsilon \in [0, u^2)$,

$$\begin{aligned} h_a(\varepsilon C) &= d\sum_{\ell=1}^{+\infty} a_{\ell d}\varepsilon^{\ell d} \\ &= \varepsilon^d d\left(\sum_{\ell=1}^{+\infty} a_{\ell d}\varepsilon^{(\ell-1)d}\right) \\ &\leq \varepsilon^d d\left(\sum_{\ell=1}^{+\infty} a_{\ell d}u^{2(\ell-1)d}\right) \\ &\leq \varepsilon^d d\left(\sum_{\ell=1}^{+\infty} a_{\ell d}u^{\ell d} + a_d\right) \qquad (5.9) \\ &\leq \varepsilon^d d(f_a(u) + a_d). \\ &= O_{\varepsilon\to 0^+}\left(\varepsilon^d\right), \end{aligned}$$

where we obtain (5.9) by noting that $2(\ell - 1) \geq \ell$ and $u \leq 1$. Finally, since $u < r_a$, $f_a(u)$ is finite. Hence the result.
The D-instability result follows from the definition of the radius of convergence. □

Another class of acyclicity constraints is based on the spectrum of the weighted adjacency matrix.

Proposition 5.2.8. *For any matrix $A \in \mathbb{R}_{\geq 0}^{d\times d}$, the followings are equivalent:*

1. *$G(A) \in$ DAGs.*
2. *$A^k = 0$ for some $k \geq 1$, i.e. A is nilpontent.*
3. *The spectral radius of A: $r(A) = 0$.*

Proof. 1 $\iff$ 2:
"$\Leftarrow$": Assume $A^k = 0$ for some $k \geq 1$, then there does not exist a walk of arbitrary high length.
If $G(A)$ is a cycle, there exists a walk of arbitrarily high length, which is a contradiction to the above.
"$\Rightarrow$": If $G(A)$ is acyclic, then similarly to the case based on linear SEMs, all diagonal elements of A are zeros, which implies that A is nilpotent.
2 $\iff$ 3: A nilpontent $\iff$ all eigenvalues $\lambda_i(A) = 0\ \forall i \in [d]$ $\iff$ spectral radius $r(A) = 0$. □

Based on Proposition 5.2.8, two other acyclicity constraints are developed. We write $W \circ W$ instead of A explicitly for the following theorem.

Theorem 5.2.9 (Log-determinant acyclicity characterization)**.** *[BAR22, Theorem 1] Let $s > 0$, define $\mathbb{W}^s = \{W \in \mathbb{R}^{d\times d} | s > r(W \circ W)\}$ where $r(W \circ W)$ denotes the spectral radius of $W \circ W$, and let $h^s_{ldet} : \mathbb{W}^s \to \mathbb{R}$ be defined as $h^s_{ldet}(W) := -\log\det(sI - W \circ W) + d \log s$. Then the following holds:*

(i) $h^s_{ldet}(W) \geq 0$, with $h^s_{ldet}(W) = 0$ if and only if $G(W)$ is a DAG.

(ii) $\triangledown h^s_{ldet}(W) = 2(sI - W \circ W)^{-T} \circ W$, with $\triangledown h^s_{ldet}(W) = 0$ if and only if $G(W)$ is a DAG.

Proof. Note that for any $s > 0$ and matrix $B \in \mathbb{R}^{d\times d}$, we have $\det(sB) = s^d B$. Then, $\log\det(sI - W \circ W) - d \log s = \log(s^d \det(I - s^{-1}W \circ W)) - d \log s = \log\det(I - s^{-1}W \circ W)$. Moreover, since $W \circ W \in \mathbb{W}^s$, we have that $s > r(W \circ W)$ or equivalently $1 > r(s^{-1}W \circ W)$. Therefore, we set $s = 1$ w.l.o.g.
(ii): By matrix calculus,

$$\triangledown h^s_{\text{ldet}}(W) = 2(sI - W \circ W)^{-T} \circ W.$$

Note that the gradient is well-defined since $(sI - W \circ W)$ is an M-matrix and by Berman and Plemmons [BP94], the inverse $(sI - W \circ W)^{-1}$ exists. By Taylor's Theorem,

$$(sI - W \circ W)^{-1} = \frac{1}{s}I + \frac{1}{s^2}(W \circ W) + \frac{1}{s^3}(W \circ W)^2 + \ldots,$$

Therefore, after taking the transpose, $[(sI - W \circ W)^{-T}]_{ij}$ is non-zero if and only if there exists a walk from j to i. After taking the Hadamard product, $[(sI - W \circ W)^{-T} \circ W]_{ij}$ is non-zero if and only if $W_{ij} \neq 0$ and $[(sI - W \circ W)^{-T}]_{ij} \neq 0$, which means there exists a direct edge from i to j and also a walk from j to i, which implies that there exists a closed walk from i to i passing through j. Thus, $\triangledown h^s_{\text{ldet}}(W) = 0$ if and only if $G(W)$ is a DAG.
(i): It holds

$$h^{s=1}_{\text{ldet}}(W) = -\log\det(I - W \circ W) = -\log(\prod_{i=1}^{d} \lambda_i(I - W \circ W)) = \sum_{i=1}^{d} -\log(1 - \lambda_i(I - W \circ W)),$$

thus at the boundary of $\mathbb{W}^s$, $h^{s=1}_{\text{ldet}}(W) \to \infty$. Therefore, the global minima of $h^{s=1}_{\text{ldet}}$ must be in the interior of $\mathbb{W}^s$ and corresponds to the set of the stationary points, i.e., $\triangledown h^{s=1}_{\text{ldet}}(W) = 0$. Together with (ii), DAGs are local and global minima of $h^{s=1}_{\text{ldet}}$.
By Proposition 5.2.8 if W is a DAG, then $\lambda_i(W \circ W) = 0$ for all i, which implies $\det(I - W \circ W) = 1$, and consequently $h^{s=1}_{\text{ldet}}(W) = 0$. Since DAGs are global minima, this implies that for all $W \in \mathbb{W}^s$, $h^{s=1}_{\text{ldet}}(W) \geq 0$. □

Theorem 5.2.10. *$h^s_{ldet}(\cdot)$ 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 $\triangledown h^s_{ldet}(A) = (sI_d - A)^{-T}$, with $\triangledown h^s_{ldet}(A) = 0$ if and only if A is a DAG.*

Proof. • **D-stable:** Proved by Theorem 5.2.9.

- **V-stable:** Note that $\log\det(sI - A) - d\log s = \log(s^d \det(I - s^{-1}A)) - d\log s = \log\det(I - s^{-1}A)$. Moverover, $s > \rho(A)$ is equivalent to $1 > \rho(s^{-1}A)$. So let's consider $s = 1$ w.l.o.g. For any $\varepsilon > 0$ such that $\varepsilon|\rho(A)| \leq 1$ it holds

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

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 $\sum_{i=1}^{d} \lambda_i(A) \neq 0$. Using the Cauchy-integral formula applied for the principal branch of the complex logarithm function $z \mapsto \log(1 - z)$ (which is analytic within domain $|z| < 1$) we get:

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

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

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

Now, notice that $w \mapsto \frac{\log(1-w)}{w-\varepsilon\lambda_i(A)}$ is continuous, therefore, by Weierstrass Theorem, it is bounded on η_0 (since η_0 is bounded domain), yielding that $\left|\frac{\log(1-w)}{w-\varepsilon\lambda_i(A)}\right| \leq K$, for some $K > 0$. Finally, since $\frac{\varepsilon|\lambda_i(A)|}{r_0} < 1$ and η_0 is a circle of radius r_0 we obtain by using ML inequality for the complex integral:

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

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 (more precisely if we choose $\varepsilon \leq \min\left\{|\rho(A)|^{-1}, \frac{r_0\left|\sum_{i=1}^{d} \lambda_i(A)\right|}{2K\|A\|_2^2}\right\}$). Thus, we proved the claim.

□

Theorem 5.2.11 (Spectral acyclicity constraint)**.** *[Naz+23, Theorem 3] The spectral radius is an acyclicity constraint*

$$h_{spectral}(A) := |\lambda_d(A)| = 0 \iff G(A) \text{ is a DAG.}$$

It's differentiable almost everywhere, with gradient

$$\nabla h_{spectral}(A) = v_d u_d^\top / v_d^\top u_d,$$

where u_d, v_d are respectively the right and left eigenvectors associated with $\lambda_d(A)$.

Proof. By Proposition 5.2.8, $h_{\text{spectral}}(A) = 0$ if and only if $G(A)$ is a DAG. Magnus [Mag85] shows that h_{spectral} is differentiable at every A that has mutually distinct eigenvalues, with the formula provided in the Theorem. The set of matrices with all distinct eigenvalues is dense in the set of matrices [HJ12, Theorem 2.4.7.1], which finalizes the proof. □

Remark 5.2.12. Since $\mathbb{C}$ is algebraically closed, then every matrix has eigenvalues in $\mathbb{C}$. So h_{spectral} is well defined everywhere.

Theorem 5.2.13 (Stability of spectrum based constraints)**.** *[Naz+23, Theorem 4] $h_{spectral}$ is stable.*

Proof.
- **E-stable:** For any $s \geq 0$ and matrix A, $h_{\text{spectral}}(sA) = |s|h_{\text{spectral}}(A) = O_{s\to\infty}(s)$.
- **V-stable:** For any $\varepsilon > 0$ and matrix A such that $h_{\text{spectral}}(A) > 0$,

$$h_{\text{spectral}}(\varepsilon A) = |\varepsilon|h_{\text{spectral}}(A) = \Omega_{\varepsilon\to 0^+}(\varepsilon).$$

- **D-stable:** Proved by Theorem 5.2.11.

□

Remark 5.2.14. Note that PST constraints are V-unstable for $d \geq 2$. Furthermore, although h_{spectral} is E-stable, V-stable and D-stable, we observe that in practice, for the same weighted adjacency matrix W, $h_{\text{spectral}}(W)$ 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 $h^s_{\text{ldet}}(W)$ as our acyclicity constraint.

5.3 Alternative Optimization Scheme: DAGMA

Besides the augmented Lagrange scheme, Bello, Aragam, and Ravikumar [BAR22] developed a simpler optimization method named DAGMA which resembles the central path approach of barrier methods [BV04]. We consider the same ECP (5.6) from Section 5.1.3 with the log-determinant acyclicity constraint.

$$\min_{\theta} L(\theta) + \lambda\|\theta\|_1 \text{ s.t. } h^s_{\text{ldet}}(W(\theta)) = 0. \tag{5.10}$$

Algorithm 5 DAGMA

1: **Input:** Data matrix $\mathbb{X}$, initial central path coefficient $\mu^{(0)}$, decay factor $\alpha \in (0, 1)$, ℓ_1 parameter $\lambda > 0$, log-det parameter $s > 0$, number of iterations T and threshold $\omega > 0$.
2: **Output:** $\widehat{W}$, the estimated weighted adjacency matrix.
3: Initialize $\mu^{(0)}$ so that $W(\mu^{(0)}) \in \mathbb{W}^s$.
4: **for** $t \leftarrow 0, 1, 2 \ldots, T-1$ **do**
5: Starting at $\mu^{(t)}$, solve $\mu^{(t+1)} = \arg\min_\theta \mu^{(t)}(L(\theta) + \lambda\|\theta\|_1) + h^s_{\text{ldet}}(W(\theta))$.
6: Set $\mu^{(t+1)} = \alpha\mu^{(t)}$.
7: Threshold matrix $\widehat{W} = W(\theta^{(T)}) \cdot \mathbb{1}(|W(\theta^{(T)})| > \omega)$.

Lemma 5.3.1. *[BAR22, Lemma 6] Algorithm 5 returns a DAG whenever $\mu^{(t)} \to 0$.*

Proof. In the proof of Theorem 5.2.9, it is shown that DAGs are global minima of h^s_{ldet}. Note that at the limit of the central path ($\mu^{(t)} \to 0$), we solve the following problem:

$$\widehat{\theta} = \arg\min_\theta h^s_{\text{ldet}}(W(\theta)).$$

Thus, the solution $W(\widehat{\theta})$ must be a DAG. □

Remark 5.3.2.
1. We require that the initial point $W(\theta^{(0)})$ be inside $\mathbb{W}^s$. Since the zero matrix is in the interior of $\mathbb{W}^s$ for any $s > 0$, one can simply set $\theta^{(0)} = 0$.
2. The default value of the hyperparameters are $\mu^{(0)} = 1, \alpha = 0.1, \beta_1 = 0.01, s = 1, T = 4$.
3. Line 5 can be solved by ADAM optimizer [KB14].

6 Kernel Methods and RKHS Representer Theorem

In this chapter, we review the kernel method, reproducing kernel Hilbert space (RKHS) and their relevant properties, and most importantly, the RKHS Representer Theorem as presented in Steinwart and Christmann [SC08].

6.1 Kernel Methods and their Properties

Definition 6.1.1 (Kernel and its feature map)**.** [SC08, Definition 4.1] Let $\mathbb{K}$ denote the set $\mathbb{R}$ or $\mathbb{C}$, and let $\mathcal{X}$ be a non-empty set. Then a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$ is called a **kernel** on $\mathcal{X}$ if there exists a $\mathbb{K}$-Hilbert space H_0 and a map $\Phi : \mathcal{X} \to H_0$ such that for all $x, x' \in \mathcal{X}$ we have

$$k(x, x') = \langle \Phi(x'), \Phi(x) \rangle_{H_0}.$$

We call Φ a **feature map** and H_0 a **feature space** of k.

Remark 6.1.2. Note that kernel k does not determine feature map Φ nor feature space H_0 uniquely.

Definition 6.1.3 (Kernel matrix)**.** [SC08, Definition 6.1.2] For fixed $x^1, \ldots, x^n \in \mathcal{X}$, the $n \times n$ matrix $K := (k(x^i, x^j))_{i,j}$ is called the **kernel matrix**.
A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called **positive semi-definite** if, for all $n \in \mathbb{N}$, $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ and all $x^1, \ldots, x^n \in \mathcal{X}$, the kernel matrix is positive semi-definite, i.e.,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x^i, x^j) \geq 0. \tag{6.1}$$

Furthermore, k is said to be **positive definite** if, for mutually distinct $x^1, \ldots, x^n \in \mathcal{X}$, equality in (6.1) holds only for $\alpha_1 = \ldots = \alpha_n = 0$. Finally, k is called **symmetric** if $k(x, x') = k(x', x)$ for all $x, x' \in \mathcal{X}$.

Remark 6.1.4. A kernel k with feature map $\Phi : \mathcal{X} \to H$ is always positive definite, since for $n \in \mathbb{N}, \alpha_1, \ldots, \alpha_n \in \mathbb{R}$, and $x^1, \ldots, x^n \in \mathcal{X}$ we have

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x^i, x^j) = \left\langle \sum_{i=1}^{n} \alpha_i \Phi(x^i), \sum_{j=1}^{n} \alpha_j \Phi(x^j) \right\rangle_{H_0} \geq 0.$$

If k is an $\mathbb{R}$-valued kernel, then k is always symmetric.

Proposition 6.1.5 (Sufficient and necessary condition for kernels)**.** *[SC08, Theorem 4.16] A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel if and only if it is symmetric and positive definite.*

6.2 RKHS and its Properties

Definition 6.2.1. [SC08, Definition 4.18] Let $\mathcal{X} \neq \emptyset$ and H be a $\mathbb{K}$-Hilbert function space over $\mathcal{X}$, i.e., a $\mathbb{K}$-Hilbert space that consists of functions mapping from $\mathcal{X}$ into $\mathbb{K}$.

1. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$ is called a **reproducing kernel** of H if we have $k(\cdot, x) \in H$ for all $x \in \mathcal{X}$ and the **reproducing property**
$$f(x) = \langle f, k(\cdot, x) \rangle_H$$
holds for all $f \in H$ and all $x \in \mathcal{X}$.

2. The space H is called a **reproducing kernel Hilbert space (RKHS)** over $\mathcal{X}$ if for all $x \in \mathcal{X}$ the Dirac functional $\delta_x : H \to \mathbb{K}$ defined by

$$\delta_x(f) := f(x), \quad f \in H,$$

is continuous.

Remark 6.2.2 (Reproducing kernels are kernels)**.** [SC08, Lemma 4.19] Let H be a Hilbert function space over $\mathcal{X}$ that has a reproducing kernel k. Then H is an RKHS and H is also a feature space of k, where the feature map $\Phi : \mathcal{X} \to H$ is given by

$$\Phi(x) = k(\cdot, x), \quad x \in \mathcal{X}.$$

We call Φ the **canonical feature map**.

Theorem 6.2.3 (Every RKHS has a unique reproducing kernel)**.** *[SC08, Theorem 4.20] Let H be an RKHS over $\mathcal{X}$. Let H' be the space of linear functionals that map from H to $\mathbb{K}$. Then $k : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$ defined by*

$$k(x, x') := \langle \delta_x, \delta_{x'} \rangle_{H'}, \quad x, x' \in \mathcal{X},$$

is the only reproducing kernel of H.

Theorem 6.2.4 (Every kernel has a unique RKHS)**.** *[SC08, Theorem 4.21] Let $\mathcal{X} \neq \emptyset$ and k be a kernel over $\mathcal{X}$ with feature space H_0 and feature map $\phi_0 : \mathcal{X} \to H_0$. Then*

$$H := \{f : \mathcal{X} \to \mathbb{K} \mid \exists w \in H_0 \text{ with } f(x) = \langle w, \phi_0(x) \rangle_{H_0} \text{ for all } x \in \mathcal{X}\}$$

equipped with the norm

$$\|f\|_H := \inf\{\|w\|_{H_0} : w \in H_0 \text{ with } f = \langle w, \phi_0(\cdot) \rangle_{H_0}\}$$

is the only RKHS for which k is a reproducing kernel. Consequently, both definitions are independent of the choice of H_0 and Φ_0. Moreover, RKHS H of k is "smallest" feature space of k in the sense that, the operator $V : H_0 \to H$ defined by

$$Vw := \langle w, \Phi_0(\cdot) \rangle_{H_0}, \quad w \in H_0,$$

is a metric surjection, i.e., $VB_{H_0} = B_H$, where B_{H_0} and B_H are the open unit balls of H_0 and H, respectively. Finally, the set

$$H_{pre} := \left\{ \sum_{i=1}^{n} \alpha_i k(\cdot, x^i) : n \in \mathbb{N}, \alpha_1, \ldots, \alpha_n \in \mathbb{K}, x^1, \ldots, x^n \in \mathcal{X} \right\}$$

is dense in H, and for $f := \sum_{i=1}^{n} \alpha_i k(\cdot, x^i) \in H_{pre}$ we have

$$\|f\|_H^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \overline{\alpha_j} k(x^j, x^i).$$

Theorem 6.2.4 can be used to determine the RKHS of a given kernel and its modification such as restrictions. Recall that every $\mathbb{C}$-valued kernel on $\mathcal{X}$ that is actually $\mathbb{R}$-valued has an $\mathbb{R}$-feature space. The following Corollary describes the corresponding $\mathbb{R}$-RKHS.

Corollary 6.2.5. *[SC08, Corollary 4.22] Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ be a kernel and H its corresponding $\mathbb{C}$-RKHS. If we actually have $k(x, x') \in \mathbb{R}$ for all $x, x' \in \mathcal{X}$, then*

$$H_{\mathbb{R}} := \{f : \mathcal{X} \to \mathbb{R} \mid \exists g \in H \text{ with } \mathrm{Re}\, g = f\}$$

equipped with the norm

$$\|f\|_{H_{\mathbb{R}}} := \inf\{\|g\|_H : g \in H \text{ with } \mathrm{Re}\, g = f\}, \quad f \in H_{\mathbb{R}},$$

is the $\mathbb{R}$-RKHS of the $\mathbb{R}$-valued kernel k.

Definition 6.2.6. 1. The supremum norm of a kernel k is defined as $\|k\|_\infty := \sup_{x \in \mathcal{X}} \sqrt{k(x, x)}$.

2. Let k be a kernel on $\mathcal{X}$ with feature map $\Phi : \mathcal{X} \to H$. Define kernel metric d_k on $\mathcal{X}$:
   For $x, x' \in \mathcal{X}$, $d_k(x, x') := \|\Phi(x) - \Phi(x')\|_H$.

Remark 6.2.7. 1. Let H be RKHS of kernel k, note that k is a reproducing kernel of H by Theorem 6.2.4, then $k(\cdot, x') \in H \ \forall x' \in \mathcal{X}$. By reproducing property,

$$k(x, x') = k(\cdot, x')(x) = \langle k(\cdot, x'), k(\cdot, x)\rangle_H.$$

Then

$$|k(x, x')|^2 = |\langle k(\cdot, x'), k(\cdot, x)\rangle_H|^2 \leq \|k(\cdot, x')\|_H^2 \cdot \|k(\cdot, x)\|_H^2 = k(x', x') \cdot k(x, x),$$

where the last inequality is followed by the Cauchy-Schwarz inequality. Therefore, $\sup_{x,x' \in \mathcal{X}} |k(x, x')| = \sup_{x \in \mathcal{X}} k(x, x)$. So k is bounded if and only if

$$\|k\|_\infty := \sup_{x \in \mathcal{X}} \sqrt{k(x, x)} < \infty.$$

2.

$$\begin{aligned} d_k(x, x') &= \sqrt{\langle \Phi(x) - \Phi(x'), \Phi(x) - \Phi(x')\rangle_H} \\ &= \sqrt{\langle \Phi(x), \Phi(x)\rangle_H - 2\langle \Phi(x), \Phi(x')\rangle_H + \langle \Phi(x'), \Phi(x')\rangle_H} \\ &= \sqrt{k(x, x) - 2k(x, x') + k(x', x')}. \end{aligned}$$

So d_k is independent of choice of Φ.

In the following, we introduce some properties of RKHS. (Chapter 4.3 of Steinwart and Christmann [SC08])

Proposition 6.2.8 (RKHSs of bounded kernels)**.** *[SC08, Lemma 4.23] Let $\mathcal{X}$ be a set and k be a kernel on $\mathcal{X}$ with RKHS H. Then k is bounded if and only if every $f \in H$ is bounded. Moreover, in this case the inclusion $id : H \to \ell_\infty(\mathcal{X})$ is continuous and we have $\|id : H \to \ell_\infty(\mathcal{X})\| = \|k\|_\infty$.*

Proposition 6.2.9 (RKHSs of measurable kernels)**.** *[SC08, Lemma 4.24] Let $\mathcal{X}$ be a measurable space and k be a kernel on $\mathcal{X}$ with RKHS H. Then all $f \in H$ are measurable if and only if $k(\cdot, x) : \mathcal{X} \to \mathbb{R}$ is measurable for all $x \in \mathcal{X}$.*

Proposition 6.2.10 (Differentiability of feature maps)**.** *[SC08, Lemma 4.34] Let $\mathcal{X} \subseteq \mathbb{R}^d$ be an open subset, k be a kernel on $\mathcal{X}$, H be a feature space of k, and $\Phi : \mathcal{X} \to H$ be a feature map of k. Let $\partial_i \partial_{i+d} k$ denote the mixed partial derivative of $k(x, x')$ wrt. $i-$th coordinates in x and x' and let $i \in \{1, \ldots, d\}$ be an index such that the mixed partial derivative $\partial_i \partial_{i+d} k$ exists and is continuous. Then the partial derivative $\partial_i \Phi$ of $\Phi : \mathcal{X} \to H$ with respect to the $i-$th coordinate exists, is continuous, and for all $x, x' \in \mathcal{X}$ we have*

$$\langle \partial_i \Phi(x), \partial_i \Phi(x')\rangle_H = \partial_i \partial_{i+d} k(x, x') = \partial_{i+d} \partial_i k(x, x').$$

In other words, $\partial_i \partial_{i+d} k$ is a kernel on $\mathcal{X} \times \mathcal{X}$ with feature map $\partial_i \Phi$.

Remark 6.2.11. In Theorem 7.2.1, we prove the "differentiable reproducing property" which is similar to Proposition 6.2.10.

One of the most commonly used kernels is the Gaussian RBF kernel:

Definition 6.2.12 (Gaussian RBF kernel)**.** [SC08, Proposition 4.10] For $d \in \mathbb{N}$, $\gamma > 0$, $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, and $x' = (x'_1, \ldots, x'_d) \in \mathbb{R}^d$, we define

$$k_\gamma(x, x') := \exp\left(-\frac{\|x - x'\|_2^2}{\gamma^2}\right) = \exp\left(-\gamma^{-2} \sum_{j=1}^{d} (x_j - x'_j)^2\right).$$

k_γ is called the **Gaussian RBF kernel** with width γ.

Proposition 6.2.13. *[SC08, Proposition 4.46] For $0 < \gamma_1 < \gamma_2 < \infty$, and non-empty set $\mathcal{X} \subset \mathbb{R}^d$, let H_{γ_1} and H_{γ_2} denote the RKHS corresponding to the Gaussian RBF kernel with width γ_1 and γ_2 respectively. We obtain $id : H_{\gamma_2}(\mathcal{X}) \to H_{\gamma_1}(\mathcal{X})$ with $\left\|id : H_{\gamma_2}(\mathcal{X}) \to H_{\gamma_1}(\mathcal{X})\right\| \leq (\frac{\gamma_2}{\gamma_1})^{\frac{d}{2}}$.*

6.3 RKHS Representer Theorem

We assume in machine learning approach that we have collected $D := ((x^1, y^1), \ldots, (x^n, y^n))$ of input/output pairs. We use this to learn a function $f : \mathcal{X} \to \mathcal{Y}$ such that $f(x)$ is a good approximation pf possible response y to an arbitrary x. Let $\mathbb{P}$ denote the unknown probability distribution on $\mathcal{X} \times \mathcal{Y}$. More precisely, a pair (x, y) is generated in two steps. First, the input value x is generated according to the marginal distribution $\mathbb{P}_{\mathcal{X}}$. Second, the output value y is generated according to the conditional probability $\mathbb{P}(\cdot|x)$ on $\mathcal{Y}$ given the value of x. Note that by letting x be generated by an unknown distribution $\mathbb{P}_{\mathcal{X}}$, we basically assume that we have no control over how the input values have been and will be observed. Furthermore, assuming that the output value y to a given x is stochastically generated by $\mathbb{P}(\cdot|x)$ accommodates the fact that in general, the information contained in x may not be sufficient to determine a single response in a deterministic manner. In particular, this assumption includes the two extreme cases where all input values determine an almost surely unique output value. Finally, assuming that the conditional probability $\mathbb{P}(\cdot|x)$ is unknown contributes to the fact that we assume that we do not have a reasonable description of the relationship between the input and output values.

Moreover, in order to access the quality of a learned function f, it's not sufficient to consider the value of the loss function $\ell(x, y, f(x))$ for a particular choice of (x, y), but in fact we need to consider how small the function $(x, y) \mapsto \ell(x, y, f(x))$ is. Here, we consider the commonly used expected loss of f defined as below:

Definition 6.3.1 (ℓ-risk)**.** [SC08, Definition 2.2] Let $\ell : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a loss function and $\mathbb{P}$ be a probability measure on $\mathcal{X} \times \mathcal{Y}$. Then, for a measurable function $f : \mathcal{X} \to \mathbb{R}$, the **$\ell$-risk** is defined by

$$\mathcal{R}_{\ell,\mathbb{P}}(f) := \int_{\mathcal{X}\times\mathcal{Y}} \ell(x, y, f(x)) d\mathbb{P}(x, y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} \ell(x, y, f(x)) d\mathbb{P}(y|x) d\mathbb{P}_{\mathcal{X}}(x).$$

For a given sequence of data observations $D := ((x^1, y^1), \ldots, (x^n, y^n)) \in (\mathcal{X} \times \mathcal{Y})^n$, we write $\mathcal{D} := \frac{1}{n} \sum_{i=1}^n \delta_{(x^i, y^i)}$, where $\delta_{(x^i, y^i)}$ denotes the Dirac measure at (x^i, y^i). In other words, $\mathcal{D}$ is the empirical measure associated to D. The risk of a function $f : \mathcal{X} \to \mathbb{R}$ with respect to this measure is called the **empirical ℓ-risk**:

$$\mathcal{R}_{\ell,\mathcal{D}}(f) = \frac{1}{n} \sum_{i=1}^n \ell(x^i, y^i, f(x^i)).$$

Lemma 6.3.2 (Convexity of risks)**.** *[SC08, Lemma 2.13] Let $\ell : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a (strictly) convex loss and $\mathbb{P}$ be a distribution on $\mathcal{X} \times \mathcal{Y}$. Then $\mathcal{R}_{\ell,\mathcal{D}} : \mathcal{L}_0(\mathcal{X}) \to [0, \infty]$ is (strictly) convex.*

Let $\ell : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a loss, H be the RKHS of a bounded measurable kernel k on $\mathcal{X}$. We consider the regularized empirical ℓ-risk:

$$\mathcal{R}^{\text{reg}}_{\ell,\mathcal{D},\lambda}(f) := \lambda \|f\|_H^2 + \mathcal{R}_{\ell,\mathcal{D}}(f). \tag{6.2}$$

Let $f_{\mathcal{D},\lambda}$ denote the minimizer of $\mathcal{R}^{\text{reg}}_{\ell,\mathcal{D},\lambda}(\cdot)$ in H. More precisely,

$$\lambda \left\| f_{\mathcal{D},\lambda} \right\|^2 + \mathcal{R}_{\ell,\mathcal{D}}(f_{\mathcal{D},\lambda}) = \inf_{f \in H} \lambda \|f\|_H^2 + \mathcal{R}_{\ell,\mathcal{D}}(f).$$

We call $f_{\mathcal{D},\lambda}$ as empirical SVM solution.

Lemma 6.3.3 (Existence of minimizers)**.** *[SC08, Theorem A.6.9] Let E be a reflexive Banach space and $f : E \to \mathbb{R} \cup \{\infty\}$ be a convex and lower semi-continuous map. If there exists an $M > 0$ such that $\{x \in E : f(x) < M\}$ is non-empty and bounded, then f has a global minimum, i.e., there exists an $x_0 \in E$ with*

$$f(x_0) \leq f(x), \qquad x \in E.$$

Moreover, if f is strictly convex, then x_0 is the only element minimizing f.

Theorem 6.3.4 (RKHS Representer theorem)**.** *[SC08, Theorem 5.5] Let $\ell : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a convex loss and $D := ((x^1, y^1), \dots, (x^n, y^n)) \in (\mathcal{X} \times \mathcal{Y})^n$. Furthermore, let H be an RKHS of a bounded measurable kernel k over $\mathcal{X}$. Then, for all $\lambda > 0$, there exists a unique empirical solution $f_{\mathcal{D},\lambda} \in H$ of the regularized empirical ℓ-risk* (6.2)*. In addition, there exist $(\alpha_1, \dots, \alpha_n) \in \mathbb{R}$ such that*

$$f_{\mathcal{D},\lambda}(x) = \sum_{i=1}^{n} \alpha_i k(x, x^i), \quad x \in \mathcal{X}. \tag{6.3}$$

The above theorem shows that the empirical minimizer of $\mathcal{R}^{\text{reg}}_{L,\mathcal{D},\lambda}(f)$ can be represented by a linear combination of the canonical feature map.

Proof.

- **Uniqueness**: Assume that $\mathcal{R}^{\text{reg}}_{L,\mathcal{D},\lambda}(f) = \lambda\|f\|_H^2 + \mathcal{R}_{\ell,\mathcal{D}}(f)$ has two minimizers $f_1, f_2 \in H$ with $f_1 \neq f_2$. Since we have

$$\frac{1}{4}\|f_1\|_H^2 + \frac{1}{4}\|f_2\|_H^2 - \frac{1}{2}\langle f_1, f_2\rangle_H = \frac{1}{4}(\|f_1\|_H^2 + \|f_2\|_H^2 - 2\langle f_1, f_2\rangle_H) = \frac{1}{4}\|f_1 - f_2\|_H^2 > 0.$$

  Then,

$$\frac{1}{2}\langle f_1, f_2\rangle_H < \frac{1}{4}\|f_1\|_H^2 + \frac{1}{4}\|f_2\|_H^2. \tag{6.4}$$

  Consider $f^* := \frac{1}{2}(f_1 + f_2)$,

$$\|f^*\|_H = \left\|\frac{1}{2}(f_1 + f_2)\right\|_H^2 = \frac{1}{4}\|f_1\|_H^2 + \frac{1}{2}\langle f_1, f_2\rangle_H + \frac{1}{4}\|f_2\|_H^2 \overset{(6.4)}{<} \frac{1}{2}\|f_1\|_H^2 + \frac{1}{2}\|f_2\|_H^2.$$

  Together with the convexity of $f \mapsto \mathcal{R}_{\ell,\mathcal{D}}(f)$ and $\lambda\|f_1\|_H^2 + \mathcal{R}_{\ell,\mathcal{D}}(f_1) = \lambda\|f_2\|_H^2 + \mathcal{R}_{\ell,\mathcal{D}}(f_2)$, it follows that

$$\lambda\|f^*\|_H^2 + \mathcal{R}_{\ell,\mathcal{D}}(f^*) < \lambda\|f_1\|_H^2 + \mathcal{R}_{\ell,\mathcal{D}}(f_1).$$

  Thus f_1 is not a minimizer of $\mathcal{R}^{\text{reg}}_{L,\mathcal{D},\lambda}(f)$ which contradicts to the assumption.

- **Existence**: Since convergence in H implies pointwise convergence, we obtain the continuity of $\mathcal{R}_{\ell,\mathcal{D}} : H \to [0, \infty)$ by the continuity of L. Then Lemma 6.3.2 shows the convexity of this map. Moreover, $f \mapsto \lambda\|f\|_H^2$ is also convex and continuous, and hence so is $f \mapsto \lambda\|f\|_H^2 + \mathcal{R}_{\ell,\mathcal{D}}(f)$. Now consider the set

$$A := \{f \in H : \lambda\|f\|_H^2 + \mathcal{R}_{\ell,\mathcal{D}}(f) \leq M\},$$

  where $M := \mathcal{R}_{\ell,\mathcal{D}}(0)$. Then $0 \in A$ obviously. In addition, $f \in A$ implies $\lambda\|f\|_H^2 \leq M$, and hence $A \subset (M/\lambda)^{\frac{1}{2}} B_H$, where B_H is the closed unit ball of H. In other words, A is a non-empty and bounded subset, and thus Lemma 6.3.3 gives the existence of a minimizer $f_{\mathcal{D},\lambda}$.

- **Representation** (6.3): We denote $X' := \{x^i : i = 1, \dots, n\}$ and $H|_{X'} := \text{span}\{k(\cdot, x^i) : i = 1, \dots, n\}$. Let $H|_{X'}^{\perp}$ be the orthogonal complement of $H|_{X'}$ in H. Then by the Hilbert Projection Theorem it follows that 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 $f_j^{\perp} \in H_{|X'}^{\perp}$. By reproducing property, it holds

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

  Thus,

$$\mathcal{R}_{\ell,\mathcal{D}}(f^{\|}) = \mathcal{R}_{\ell,\mathcal{D}}(f).$$

  Note that $\langle f^{\|}, f^{\perp}\rangle_H = 0$, hence

$$\|f\|_H = \left\|f^{\|}\right\|_H + \left\|f^{\perp}\right\|_H \geq \left\|f^{\|}\right\|_H.$$

We conclude that

$$\begin{aligned}\inf_{f\in H} \lambda\|f\|_H^2 + \mathcal{R}_{\ell,\mathcal{D}}(f) &\leq \inf_{f\in H|_{X'}} \lambda\|f\|_H^2 + \mathcal{R}_{\ell,\mathcal{D}}(f) \\ &= \inf_{f\in H} \lambda\left\|f^{\|}\right\|_H^2 + \mathcal{R}_{\ell,\mathcal{D}}(f^{\|}) \\ &\leq \inf_{f\in H} \lambda\|f\|_H^2 + \mathcal{R}_{\ell,\mathcal{D}}(f).\end{aligned}$$

Therefore, we obtain that

$$\inf_{f\in H} \lambda\|f\|_H^2 + \mathcal{R}_{\ell,\mathcal{D}}(f) = \inf_{f\in H|_{X'}} \lambda\|f\|_H^2 + \mathcal{R}_{\ell,\mathcal{D}}(f),$$

which finalizes the proof.

□

7 Structure Learning in General Non-Parametric Structural Equation Models based on RKHS

The general non-parametric NOTEARS algorithms introduced in chapter 5 approximate the non-linear relationships between random variables either by MLPs or basis expansions. However, the MLP approximation is sensitive to the size of hidden units. Although increasing the size of the hidden layers increases the flexibility of MLP functions, larger networks require more samples to estimate the parameters [Zhe+20]. Moreover, the current MLP approximation relies on random initialization for the weights which causes obvious randomness in results [see the illustration from WBD24, Figure 2 therein]. Fine-tuning the architecture of a neural network is, thus, a non-trivial task. On the other hand, basis expansion approximation is restricted through its focus on additive models. Motivated by these, we introduce a novel approximation family by kernel methods in this chapter. In section 7.1, we introduce a model-agnostic sparsity regularizer based on partial derivatives. Similar to Rosasco et al. [Ros+13], we establish a version of an RKHS Representer Theorem for an empirical acyclicity constrained optimization problem. We approximate the non-parametric relationships with the help of an RKHS given by a differentiable kernel. Finally, we assemble the DAGMA optimization scheme (section 5.3) and develop an algorithm called "RKHS-DAGMA" in section 7.3.

7.1 Sparsity Regularizer

In causal inference, we usually expect that a random variable depends only on a few other random variables. In other words, we favor functions for which each partial derivative is small at different points. Recall that we assume functions $f_j \colon \mathcal{X} \to \mathbb{R}, j \in [d]$ are in the class $C^1(\mathcal{X})$, and the functions f_j and their derivatives are both square-integrable. Denote $\mathbb{P}_X$ to be the joint distribution of the random vector X, we consider the following sparsity regularizer (see Rosasco et al. [Ros+13] for applications in statistical learning and asymptotic optimality of the obtained minimizer on the classes of functions which contain RKHS H).

$$\Omega_1(f_j) = \sum_{k=1}^{d} \left\| \frac{\partial f_j(\cdot)}{\partial x_k} \right\|_{L_2} = \sum_{k=1}^{d} \sqrt{\int_{\mathcal{X}} \left(\frac{\partial f_j(x)}{\partial x_k} \right)^2 \mathbb{P}_X(dx)}. \tag{7.1}$$

To develop a version of the data-based decision rule we need to consider an empirical counterpart for $\|\cdot\|_{L^2}$ of the derivative. Thus, we "mimic" it by plugging in the empirical measure $\mathcal{D} := \frac{1}{n} \sum_{i=1}^{n} \delta_{(x^i)}$. We set

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

Then the empirical estimate of (7.1) 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

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

7.2 Constrained Empirical Optimization Problem Solved by Kernel Methods

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

$$\sum_{j=1}^{d}\left\{\frac{1}{2n}\sum_{i=1}^{n}(x_j^i - f_j(x^i))^2 + \tau(2\Omega_1^{\mathcal{D}}(f_j) + \lambda\left\|f_j\right\|_H^2)\right\} \text{ s.t. } h^s_{\text{ldet}}(W^{\mathcal{D}}) = 0, \tag{7.2}$$

where τ, λ are positive numbers and $s > 0$ being some fixed number (typically set to 1). Following ideas as in Rosasco et al. [Ros+13], we show a version of RKHS Representer Theorem for the optimization criteria (7.2) 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 (7.2) over the reproducing Hilbert space H admits a solution of the form (for every $j \in [d]$):

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

Theorem 7.2.1. *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* (7.2) *can be written as*

$$\widehat{f}_j^{\tau}(x) = \sum_{i=1}^{n}\alpha_i^j k(x, x^i) + \sum_{i=1}^{n}\sum_{a=1}^{d}\beta_{ai}^j\frac{\partial k(x, s)}{\partial s_a}\bigg|_{s=x^i}, \quad x \in \mathcal{X}, \tag{7.3}$$

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

$$\left\|\widehat{f}_j^{\tau}\right\|_H^2 = \sum_{i,l=1}^{n}\alpha_i^j\alpha_l^j k(x^i, x^l) + 2\sum_{i,l=1}^{n}\sum_{a=1}^{d}\alpha_i^j\beta_{al}^j\frac{\partial k(x^i, x^l)}{\partial x_a^l} + \sum_{i,l=1}^{n}\sum_{a,b=1}^{d}\beta_{ai}^j\beta_{bl}^j\frac{\partial k(x^i, x^l)}{\partial x_a^i \partial x_b^l}. \tag{7.4}$$

Proof. (7.3): Consider arbitrary elements $f, g \in H$; since $g \in H$ and H is complete, there exists a sequence of elements $(g)_{n\geq 1}$, $g_n \in H$ such that it converges to g in the norm of Hilbert space H. Furthermore, for every $n \in \mathbb{N}$ we have:

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

Note that since k is bounded it implies that f is bounded [see ex. SC08, 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 scalar product is continuous in the first coordinate.

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

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

Namely, we refine the proof (see Theorem 1 point a), b) in Zhou [Zho08]) in case $\alpha = 0$ to show that the "derivative element" exists in H. We deviate from the proof of Theorem 1 in the part to establish that differential reproducing property holds for all $f \in H$. The latter part is different from that of Zhou [Zho08] as it uses the completeness of H and the fact that weak convergence and pointwise convergence are equivalent in RKHS H.

Consider arbitrary $x \in \mathcal{X}$; since $\mathcal{X}$ is open there exists $r > 0$ such that $X_r := \{x + y, y \in \mathbb{R}^d, \|y\|_2 \leq r\} \subset \mathcal{X}$. For $a \in [d]$, denote e_a to be $a-$th orthonormal vector in the standard Euclidean basis in $\mathbb{R}^d$. For arbitrary $x \in \mathcal{X}$ denote $\tilde{h}_{x,a}$ to be the function $\tilde{h}_{x,a} : \mathcal{X} \mapsto \mathbb{R}$ such that $\tilde{h}_{x,a}(y) = \frac{\partial}{\partial s_a} k(s, y)|_{s=x}$ for all $y \in \mathcal{X}$, i.e., for all x, $y \mapsto \frac{\partial}{\partial s_a} k(s, y)|_{s=x}$. Since, from the definition of the RKHS, $k(\cdot, x) \in H$ we have that set of functions $H \mapsto \mathbb{R}$

$$\left\{ \frac{1}{t}(k(\cdot, x + te_a) - k(\cdot, x)), |t| \leq r \right\} \tag{7.5}$$

is such that it holds 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 application of the Mean Value Theorem (twice, once in every coordinate). The latter inequality implies that set (7.5) 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 ex. [Rud91], Chapter 3). Thus, there exist 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 for arbitrary $f \in H$ it holds:

$$\lim_{n\to\infty} \left\langle \frac{1}{t_n}(k(\cdot, x + t_n e_a) - k(\cdot, x)), f \right\rangle_H = \langle h_x, f \rangle_H. \tag{7.6}$$

Consider $f = k(\cdot, y)$, where $y \in \mathcal{X}$ arbitrary. Since $k(\cdot, y)$ is differentiable (as a function of first coordinate) we have that for the LHS of the last equality it holds:

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

But from the other side, it holds that

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

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

Now we show that "differentiable reproducing property" holds, i.e., that 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 the 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 $g \in H$. The latter means we have for all $f \in H$ it holds that $\lim_{n\to\infty} \langle g_n, f \rangle = \langle g, f \rangle$. Then, 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 $f \in span\{k(x, \cdot)\}$ (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 Equation (7.6) 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 for every $f \in H$ holds:

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

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

Denote $X' := \{x^i, i = 1, \dots, n\}$ and $H|_{X'} := \text{span}\{k(\cdot, x^i), \frac{\partial k(\cdot,s)}{\partial s_a}|_{s=x^i} : i = 1, \dots, n, a = 1, \dots, d\}$, let $H^{\perp}_{|X'}$ be the orthogonal complement of $H|_{X'}$ in H (notice that it exists and well-defined as every element in the span exist and well-defined). Then, from the Hilbert Projection Theorem it follows that every $f_j \in H$, $j \in [d]$ can be uniquely decomposed as $f_j = f_j^{\parallel} + f_j^{\perp}$, where $f_j^{\parallel} \in H|_{X'}$ and $f_j^{\perp} \in H^{\perp}_{|X'}$.

Let $e_a \in \mathbb{R}^d$ being a-th vector of the standard Euclidean basis in $\mathbb{R}^d$; by reproducing property and definition of $H^{\perp}_{|X'}$ in H, it holds

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

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

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

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

$$\frac{1}{2n} \sum_{i=1}^{n} (x_j^i - f_j(x^i))^2 = \frac{1}{2n} \sum_{i=1}^{n} (x_j^i - f_j^{\parallel}(x^i))^2,$$

whereas the differential reproducing property implies that it holds

$$W_{aj}^{\mathcal{D}}(f) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial f_j^2(x^i)}{\partial x_a^i} = \frac{1}{n} \sum_{i=1}^{n} \frac{\left(\partial f_j^{\parallel}(x^i)\right)^2}{\partial x_a^i} = W_{aj}^{\mathcal{D}}(f^{\parallel}).$$

Notice furthermore, that

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

which in turn implies that $\Omega_1^{\mathcal{D}}(f_j) = \Omega_1^{\mathcal{D}}(f_j^{\parallel})$.

Thus, by denoting

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

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

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

Thus, we showed that it holds:

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

Thus, the first claim of the Theorem holds.
To prove (7.4), we first note that by the differential reproducing property it holds that

$$\langle \frac{\partial k(\cdot,x)}{\partial x_a}, \frac{\partial k(\cdot,y)}{\partial y_b} \rangle_H = \frac{\partial^2 k(x,y)}{\partial x_a \partial y_b}. \tag{7.7}$$

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

$$\begin{aligned}
\left\| \widehat{f}_j^{\,\tau} \right\|_H^2 &= \left\langle \sum_{i=1}^n \alpha_i^j k\big(\cdot, x^i\big) + \sum_{i=1}^n \sum_{a=1}^d \beta_{ai}^j \frac{\partial k(\cdot, x^i)}{\partial x_a^i}, \sum_{i=1}^n \alpha_i^j k\big(\cdot, x^i\big) + \sum_{i=1}^n \sum_{a=1}^d \beta_{ai}^j \frac{\partial k(\cdot, x^i)}{\partial x_a^i} \right\rangle_H \\
&= \sum_{i,l=1}^n \alpha_i^j \alpha_l^j k\Big(x^i, x^l\Big) + 2 \sum_{i,l=1}^n \sum_{a=1}^d \alpha_i^j \beta_{al}^j \left\langle \frac{\partial k\big(\cdot, x^l\big)}{\partial x_a^l}, k\big(\cdot, x^i\big) \right\rangle_H \\
&+ \sum_{i,l=1}^n \sum_{a,b=1}^d \beta_{ai}^j \beta_{bl}^j \left\langle \frac{\partial k\big(\cdot, x^i\big)}{\partial x_a^i}, \frac{\partial k\big(\cdot, x^l\big)}{\partial x_b^l} \right\rangle_H \\
&= \sum_{i,l=1}^n \alpha_i^j \alpha_l^j k\Big(x^i, x^l\Big) + 2 \sum_{i,l=1}^n \sum_{a=1}^d \alpha_i^j \beta_{al}^j \frac{\partial k\big(x^i, x^l\big)}{\partial x_a^l} + \sum_{i,l=1}^n \sum_{a,b=1}^d \beta_{ai}^j \beta_{bl}^j \frac{\partial k\big(x^i, x^l\big)}{\partial x_a^i \partial x_b^l},
\end{aligned} \tag{7.8}$$

which finalizes the proof. □

Motivated by Theorem 7.2.1, we estimate every function f_j by their kernel estimators as above. We provide our algorithm with the common-used Gaussian kernel which is based on the Euclidean distance of two input data points. For $\gamma > 0$, $j \in [d]$ let $k_\gamma^{-j} : \mathbb{R}^{d-1} \mapsto \mathbb{R}$ be a *Gaussian kernel* defined as

$$k_\gamma^{-j}(x, x') := \exp\left(-\gamma^{-2} \sum_{i \neq j}^d (x_i - x_i')^2 \right), \tag{7.9}$$

i.e., it corresponds to the Gaussian kernel $k_\gamma\big(x, x'\big) = \exp\Big(- \frac{\left\| x - x' \right\|_2^2}{\gamma^2} \Big)$ evaluated as if the $j-$th coordinate was set to a constant.

Note that each random variable can't be the cause of itself, f_j shouldn't have x_j as its input, thus we replace k in (7.3) with k^{-j} to approximate f_j, i.e.

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

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

$$\sum_{j=1}^d \left\{ \frac{1}{2n} \sum_{i=1}^n (x_j^i - \widehat{f}_j^{\,\theta}(x^i))^2 + \tau[2\Omega_1^{\mathcal{D}}(\widehat{f}_j^{\,\theta}) + \lambda \left\| \widehat{f}_j^{\,\theta} \right\|_H^2] \right\}. \tag{7.10}$$

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

$$\begin{aligned}
W_{kj}^{\mathcal{D}}(\widehat{f}_j^{\,\theta}) = W^{\mathcal{D}}(\theta)_{kj} &= \left\| \frac{\partial \widehat{f}_j^{\,\theta}(x)}{\partial x_k} \right\|_n \\
&= \left\{ \frac{1}{n} \sum_{i=1}^n \left[\sum_{l=1}^n \alpha_l^j \frac{\partial k^{-j}(x^i, x^l)}{\partial x_k^i} + \sum_{l=1}^n \sum_{a=1}^d \beta_{al}^j \frac{\partial k^{-j}(x^i, x^l)}{\partial x_k^i \partial x_a^l} \right]^2 \right\}^{\frac{1}{2}},
\end{aligned} \tag{7.11}$$

and consequently

$$\Omega_1^{\mathcal{D}}(\widehat{f}_j^{\,\theta}) = \sum_{k=1}^d \sqrt{\frac{1}{n} \sum_{i=1}^n (\frac{\partial \widehat{f}_j^{\,\theta}(x^i)}{\partial x_k})^2} = \sum_{k=1}^d W^{\mathcal{D}}(\theta)_{kj}. \tag{7.12}$$

Remark 7.2.2. Since we exploit the Gaussian kernel, we want that the decision rule (and, generally speaking, considered classes of prediction functions) returned by the Algorithm to belong to the space of continuously-differentiable functions (and thus implying the condition that if partial derivative exists and equal to zero then the function does not depend on that coordinate). Below we justify this fact by showing that any function that belongs to Gaussian RKHS $H_\gamma(\mathcal{X})$ also belongs to $C^1(\mathcal{X})$. Let H_γ be the RKHS of the real-valued Gaussian RBF kernel k_γ for $\gamma > 0$. Notice, from Theorem 10.45 in Wendland [Wen04] it follows that, since $k_\gamma(\cdot,\cdot)$ is (infinitely many times) differentiable, so the associated RKHS $H_\gamma(\mathcal{X})$ is a subset of the space of continuously differentiable functions. Therefore, the inclusion $H_\gamma(\mathcal{X}) \subset C^1(\mathcal{X})$ holds, which implies that for every $f \in H_\gamma$ it holds $\|f\|_{C^1(\mathcal{X})} < \infty$.

7.3 Optimization

Combining Equations (7.10), (7.11), (7.12) together with log-determinant acyclicity constraint, we obtain the following constrained empirical optimization problem:

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

As introduced in DAGMA, we use a central path optimization method to solve the constrained optimization problem. We give our method in Algorithm 6 and call it RKHS-DAGMA. Note that the subproblem (7.13) of Algorithm 6 means that starting at $\theta = \theta^{(t)}$, $\theta^{(t+1)}$ is obtained by the ADAM optimizer [KB14].

Algorithm 6 RKHS-DAGMA

1: **Input:** Data matrix $\mathbf{X}$, 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) and threshold ω (e.g., 0.1).
2: **Output:** $\mathbf{W}$, the estimated weighted adjacency matrix.
3: Initialize $\theta^{(0)}$ so that $W^{\mathcal{D}}(\theta^{(0)}) \in \mathbb{W}^s$.
4: **for** $t \leftarrow 0$ to $T-1$ **do**
5: Starting at $\theta^{(t)}$, solve

$$\theta^{(t+1)} = \arg\min_\theta \mu^{(t)} \sum_{j=1}^{d} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (x_j^i - \widehat{f}_j^{\,\theta}(x^i))^2 + \tau[2\Omega_1^{\mathcal{D}}(\widehat{f}_j^{\,\theta}) + \lambda \|\widehat{f}_j^{\,\theta}\|_H^2] \right\}) + h^s_{\text{ldet}}(W^{\mathcal{D}}(\theta)). \quad (7.13)$$

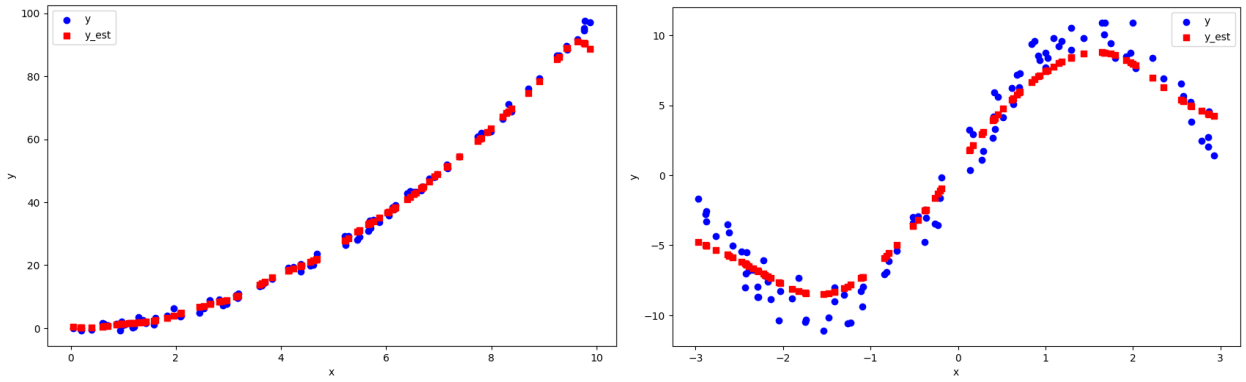
6: Set $\mu^{(t+1)} = \alpha\mu^{(t)}$.
7: Threshold matrix $\widehat{W} = W^{\mathcal{D}}(\theta^{(T)}) \cdot \mathbb{1}(W^{\mathcal{D}}(\theta^{(T)}) > \omega)$.

8 Experiments

The contents of this section are divided into three parts. First, we analyze the performance of the RKHS-DAGMA model in a rather simplistic setting of the bivariate prediction (distinguishing cause-effect) within artificially constructed toy models. Second, we evaluate and compare the properties of the RKHS-DAGMA solution with non-parametric NOTEARS algorithms like NOTEARS-MLP and NOTEARS-SOB on the sampled directed Erdős-Rényi graphs of growing dimension. Finally, we compare the performance of RKHS-DAGMA with NOTEARS-MLP and NOTEARS-SOB on the real-world bivariate datasets [Moo+16].

8.1 Toy Example

We illustrate the performance of RKHS-DAGMA by two simple simulations with two nodes and 100 data points. Plots show the ground truth data points of corresponding simulations (blue) and estimated function values (red) obtained by RKHS-DAGMA in the bivariate causal models $Y = X^2 + \varepsilon, X \sim \mathcal{U}[0, 10]$ (left) and $Y = 10\sin(X) + \varepsilon, X \sim \mathcal{U}[-3, 3]$ (right). Denote W_{est} as the estimated weighted adjacency matrix without any thresholding. Results of figure 8.1(a) correspond to the estimated matrix $W_{\text{est}} = \begin{pmatrix} 0 & 10.35 \\ 6.22 \times 10^{-4} & 0 \end{pmatrix}$. Thus, results of 8.1(b) give $W_{\text{est}} = \begin{pmatrix} 0 & 4.91 \\ 8.49 \times 10^{-4} & 0 \end{pmatrix}$. In both cases, we observe W_{12} is significantly large, and W_{21} is sufficiently small to be ignored after thresholding, indicating that the RKHS-DAGMA finds correct causal relationships. (see 8.1 and explanations therein).



(a) quadratic relationship **(b)** sinus relationship

Figure 8.1 Toy examples

8.2 Structure Learning

Next, we examine the structure recovery of RKHS-DAGMA compared to the baselines nonparametric NOTEARS methods by comparing the estimated DAG with the ground truth generated from Erdős-Rényi directed graph with the given topological ordering of the vertices. Noting that in Zheng et al. [Zhe+20], several graph models are considered such as ER1, ER2, ER4, SF1, SF2, SF4 where ERm denotes Erdős-Rényi graph with $m \times d$ edges; same for scale-free (SF) graph. Here we focus on one of the hardest settings, the ER4 graphs where the enhancements of NOTEARS algorithms are less competitive compared to other algorithms like fast greedy equivalence search [Ram+17], DAG-GNN [Yu+19], greedy equivalence search with generalized scores [Hua+18] (see Figure 4 in Zheng et al. [Zhe+20]).

Simulation Given ground truth DAG, we simulate ER4 DAGs according to the procedure of Zheng et al. [Zhe+20] with the functional relationship in the following three ways: The first way is by the Gaussian process:

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

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 $p \leq d$, $g_j(X_{\mathrm{pa}(j)}) \sim \mathcal{N}(0, K(\mathbb{X}_{\mathrm{pa}(j)}, \mathbb{X}_{\mathrm{pa}(j)}))$ with

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

where $\vec{x}^a, \vec{x}^b$ denotes the $a-$th and $b-$th row of $\mathbb{X}_{\mathrm{pa}(j)}$ respectively and lengthscale $l = 1$. Moreover, $\varepsilon_j \sim \mathcal{N}(0, I_n)$ is a standard Gaussian noise.
We call the second way additive GP:

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

where each g_{kj} is sampled from RBF GP with lengthscale 1.
The third way is by a MLP network with hidden size 100 and a sigmoid activation function where all weights are sampled from $\mathcal{U}((-2.0, -0.5) \cup (0.5, 2.0))$. Moreover, we add 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:

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

As mentioned in Bello, Aragam, and Ravikumar [BAR22], the initial point $W(\theta^{(0)})$ is required to be inside $\mathbb{W}^s$. Zero matrix is always inside $\mathbb{W}^s$ for any $s > 0$, thus we set parameters $\theta^{(0)}$ be 0. Since our approximation method and sparsity regularizer fundamentally differ from NOTEARS algorithms, the hyperparameters λ and τ are tuned by 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 for $T = 6$, 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 Remark 8.2.1 for the intuitive explanation of such choice, a proper parameter choice should also be done by grid search). Due to the explicit computation of derivatives and Hessian of the kernel function, we set the maximum number of iterations of the ADAM optimizer to 10% of corresponding values in DAGMA to compensate for the additional cost. For NOTEARS algorithms, we choose the default hyperparameters as described in Waxman, Butler, and Djurić [WBD24] and Zheng et al. [Zhe+20]. We choose the structural Hamming distance (SHD) which is the total number of edge additions, deletions, and reversals needed to convert the estimated graph into the true graph, to evaluate the model performance. Thus, the lower the SHD is, the better the model performs. To avoid the scaling impact, we plot the model comparison separately.

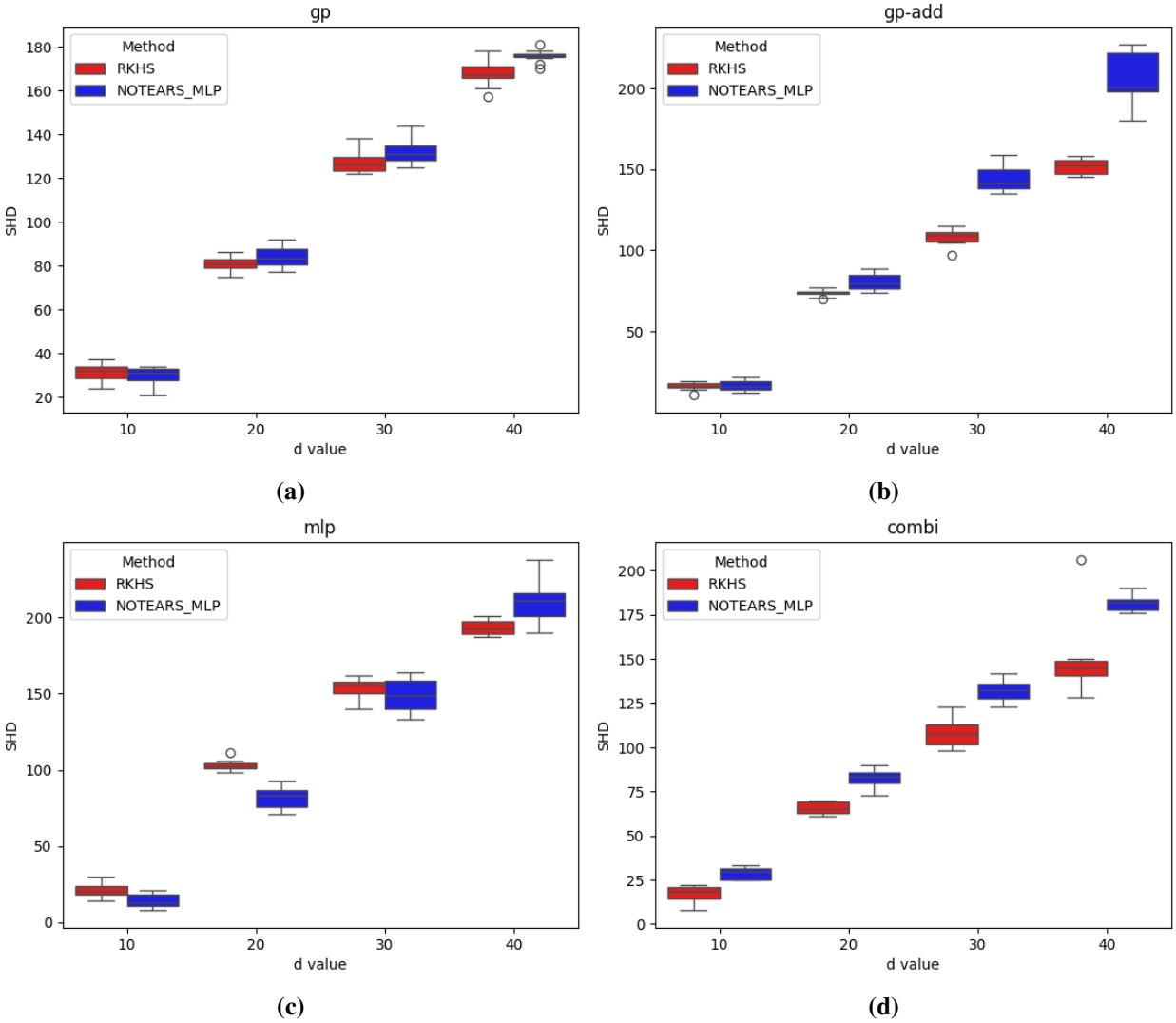


Figure 8.2 Comparison between RKHS-DAGMA and NOTEARS-MLP by SHD (lower is better) for random data generated from 8.2(a) the ER-4 GP model, 8.2(b) the ER-4 GP-additive model, 8.2(c) the ER-4 MLP model, 8.2(d) the model with combination of functions. Boxplots show the median and quartiles across 10 different simulations for each simulation model.

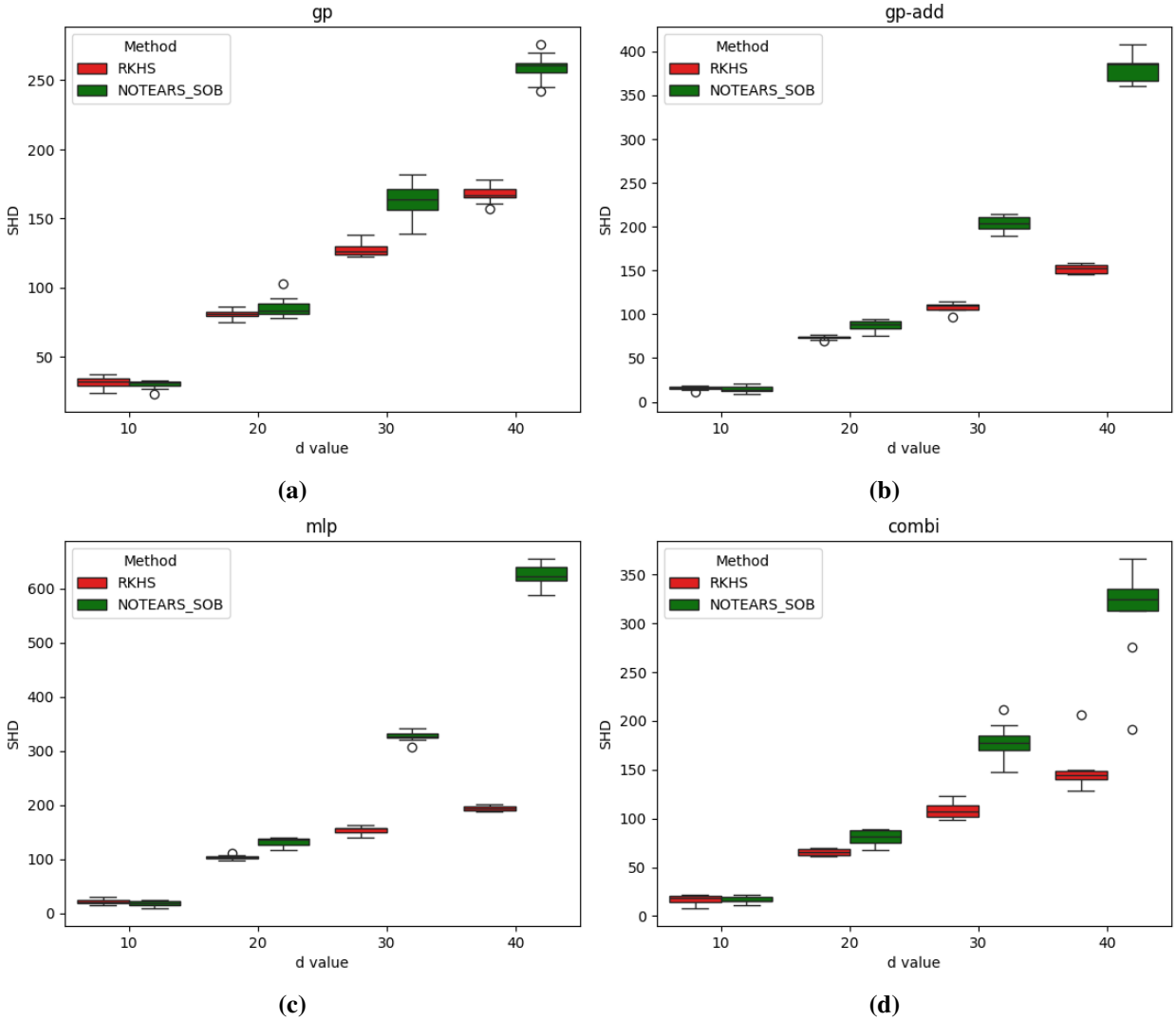


Figure 8.3 Comparison between RKHS-DAGMA and NOTEARS-SOB by SHD (lower is better) for random data generated from 8.3(a) the ER-4 GP model, 8.3(b) the ER-4 GP-additive model, 8.3(c) the ER-4 MLP model, 8.3(d) the model with combination of functions. Boxplots show the median and quartiles across 10 different simulations for each simulation model.

Our results indicate that RKHS-DAGMA consistently outperforms NOTEARS-SOB among all types of simulations in structured Hamming distance (SHD). Additionally, compared to NOTEARS-MLP, RKHS-DAGMA demonstrates superior performance (in terms of the SHD) in simulations based on GP, additive GP, and combinatorial models, while maintaining competitive results in MLP experiments (see figure 8.2 and figure 8.3).

Remark 8.2.1. While considering the experimental setting with different dimensions of the underlying Erdős-Rényi graphs, we notice that the complexity of the decision rule increases as the dimension grows. Thus, one observes a "classical" phenomenon of the curse of dimensionality [see for example Gir21; Gy02]. 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, one employs an observation that in order to reduce the estimation error of the signal one uses decision rules of higher regularity.

In our case, since we are employing the machinery of RKHS rules based on the Gaussian kernel (7.9), and since we have that for the RKHS it holds $H_{\gamma_2} \subset H_{\gamma_1}$ for $\gamma_2 > \gamma_1 > 0$ and H_γ being the Gaussian RKHS with reproducing kernel $k(x,\cdot) := k_\gamma(x,\cdot) = \exp\left(-\frac{\|x-\cdot\|^2}{\gamma^2}\right)$ (see Proposition 6.2.13), it would result with the larger choice of γ when the dimensionality of the problem ($ER(4d)$) is large. Notice that for the bounded domains $\mathcal{X}$ the same effect (i.e., restricting to the spaces of larger smoothness), when estimating

the unknown functions with Gaussian kernel, can be ensured by considering the re-scaled domain with parameter $\frac{1}{\gamma}$.

8.3 Real Data

Finally, we compared the model performance of RKHS-DAGMA to that of NOTEARS-MLP, and NOTEARS-SOB on a benchmark collection of different datasets with cause-effect pairs from Mooij et al. [Moo+16]. These are the bivariate datasets, each consisting of one pair of statistically dependent variables. We remove 6 datasets that contain multi-dimensional random variables. First, we standardize all remaining datasets. If the sample size exceeds 400, the dataset is divided into 300 grids based on the first covariate to reduce computational costs. The median of each grid is then calculated and used for model evaluation. RKHS-DAGMA achieves the best accuracy of 55.88% among the remaining 102 different datasets, while NOTEARS-SOB and NOTEARS-MLP achieve an accuracy of 45.10% and 0.98% correspondingly. We suppose the bad performance of NOTEARS-MLP is due to the small sample size with a relatively large number of hidden units compared to the number of nodes and the specific definition of the weighted adjacency matrix depends on the weight of the first hidden layer which may be quite different than those defined by derivatives [WBD24].

9 Conclusion

In this work, we addressed the non-parametric DAG learning problem utilizing a procedure that exploits the machinery of infinite-dimensional (Gaussian) RKHS. Namely, we showed in Theorem 7.2.1 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 the ones used in the DAGMA algorithm. We compared the efficiency of the RKHS-DAGMA algorithm with the known baselines in the setting of nonparametric structural equation modeling such as NOTEARS-MLP and NOTEARS-SOB (see ex. Zheng et al. [Zhe+20]) in the settings comparable to those of Zheng et al. [Zhe+20]. The versatility of the non-linear RKHS-DAGMA algorithm appears to be especially useful on the datasets where the non-linear nature of dependency (see experiments on the cause-effect data in Subsection 8.3) comes into play. The code is available at `https://github.com/yurou-liang/RKHS-DAGMA`.

A Appendix

A.1 Proximal Quasi-Newton (PQN) Method

Consider the problem (P): $\min_{\boldsymbol{\omega}} f(\boldsymbol{\omega}) := l(\boldsymbol{\omega}) + \lambda\|\boldsymbol{\omega}\|_1$, where $l(\boldsymbol{\omega})$ is a convex differentiable loss function, then $\boldsymbol{\omega}$ can be updated by the following algorithm:

Algorithm 7 PQN for unconstrained problem [Zho+14]

Input: $\boldsymbol{\omega}_0, g_0 = \nabla l(w_0)$, activate set $\mathcal{A} = [p] = 1, \ldots, p$, L-BFGS memory size m, termination criterion ε
For $k = 0, 1, 2 \ldots$:

(a) Shrink $\mathcal{A}$ to rule out j with $w_j = 0$ or small subgradient $|\partial_j L(w)|$.

(b) If shrinking stopping criteria is satisfied,
   (i) Reset $\mathcal{A} = [p]$ and L-BFGS memory.
   (ii) Update shrinking stopping criteria and continue.

(c) Solve following equation for descent direction d_k using coordinate update below on active set:

$$\boldsymbol{d}_k = \arg\min_{\boldsymbol{d}\in\mathbb{R}^p} \boldsymbol{g}_k^\top \boldsymbol{d} + \frac{1}{2}\boldsymbol{d}^\top B_k \boldsymbol{d} + \lambda\|\boldsymbol{\omega}_k + \boldsymbol{d}\|_1,$$

where $\boldsymbol{g}_k$ is the gradient of $f(w)$ and B_k is the L-BFGS approximation of the Hessian of $l(\boldsymbol{\omega})$ which is computed by followings:

(d)

$$\begin{gathered}
B_t = B_0 - QRQ^\top = B_0 - Q\widehat{Q}, \\
\text{where } \boldsymbol{s}_t = \boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}_t \text{ and } \boldsymbol{y}_t = \boldsymbol{g}_{t+1} - \boldsymbol{g}_t, \\
Q := [B_0 S_t \quad Y_t], R := \begin{bmatrix} S_t^\top B_0 S_t & L_t \\ L_t^\top & -D_t \end{bmatrix}^{-1}, \widehat{Q} := RQ^\top, \\
S_t = [\boldsymbol{s}_0, \ldots, \boldsymbol{s}_{t-1}], Y_t = [\boldsymbol{y}_0, \ldots, \boldsymbol{y}_{t-1}], \\
D_t = \operatorname{diag}[\boldsymbol{s}_0^\top \boldsymbol{y}_0, \ldots, \boldsymbol{s}_{t-1}^\top \boldsymbol{y}_{t-1}] \text{ and } (L_t)_{i.j} = \begin{cases} \boldsymbol{s}_{i-1}^\top \boldsymbol{y}_{j-1} & \text{if } i > j, \\ 0 & \text{otherwise,} \end{cases} \\
B_0 = \boldsymbol{y}_{t-1}^\top \boldsymbol{s}_{t-1} / \boldsymbol{s}_{t-1}^\top \boldsymbol{s}_{t-1} I.
\end{gathered}$$

In practice, the memory of BFGS is limited to m.

(e) Note that for each coordinate j, problem (c) has a closed form update $\boldsymbol{d} \leftarrow \boldsymbol{d} + z^* \boldsymbol{e}_j$ given by

$$z^* = \arg\min_z \frac{1}{2} \underbrace{B_{jj}}_{a} z^2 + \underbrace{(\boldsymbol{g}_j + (B\boldsymbol{d})_j)}_{b} z + \lambda |\underbrace{(\boldsymbol{\omega}_t)_j + d_j}_{c} + z| = -c + S(c - \frac{b}{a}, \frac{\lambda}{a}),$$

where the soft threshold function $S(x, a) := \operatorname{sign}(x)\max(|x| - a, 0)$.

(f) Line search for step size $\rho \in (0, 1]$ until the following "Armijo rule" is satisfied:

$$f(\boldsymbol{\omega}_k + \rho \boldsymbol{d}_k) \leq f(\boldsymbol{\omega}_k) + \rho c_1 (\lambda\|\boldsymbol{\omega}_k + \boldsymbol{d}_k\|_1 - \lambda\|\boldsymbol{\omega}_k\|_1 + \boldsymbol{g}_k^\top \boldsymbol{d}_k),$$

where c_1 is some small constant, typically set to 10^{-3} or 10^{-4}.

(g) Generate new iterate $\boldsymbol{\omega}_{k+1} \leftarrow \boldsymbol{\omega}_k + \rho \boldsymbol{d}_k$.

(h) Update $\boldsymbol{g}, \boldsymbol{s}, \boldsymbol{y}, Q, R, \widehat{Q}$ restricted to $\mathcal{A}$.

Bibliography

[BAR22] K. Bello, B. Aragam, and P. Ravikumar. "DAGMA: Learning DAGs via m-matrices and a log-determinant acyclicity characterization". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 8226–8239.

[BP94] A. Berman and R. J. Plemmons. *Nonnegative matrices in the mathematical sciences*. SIAM, 1994.

[BV04] S. P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[Chi02] D. M. Chickering. "Optimal structure identification with greedy search". In: *Journal of machine learning research* 3.Nov (2002), pp. 507–554.

[CHM04] M. Chickering, D. Heckerman, and C. Meek. "Large-sample learning of Bayesian networks is NP-hard". In: *Journal of Machine Learning Research* 5 (2004), pp. 1287–1330.

[DM17] M. Drton and M. H. Maathuis. "Structure learning in graphical modeling". In: *Annu. Rev. Stat. Appl.* 4 (2017), pp. 365–393.

[Efr08] S. Efromovich. *Nonparametric curve estimation: methods, theory, and applications*. Springer Science & Business Media, 2008.

[Gir21] C. Giraud. *Introduction to high-dimensional statistics*. Chapman and Hall/CRC, 2021.

[G̈24] K. Göbler et al. "`causalAssembly`: Generating Realistic Production Data for Benchmarking Causal Discovery". In: *Proceedings of the Third Conference on Causal Learning and Reasoning*. Vol. 236. Proceedings of Machine Learning Research. PMLR, 2024, pp. 609–642.

[Gy02] L. Györfi et al. *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer-Verlag, New York, 2002, pp. xvi+647.

[HGC95] D. Heckerman, D. Geiger, and D. M. Chickering. "Learning Bayesian networks: The combination of knowledge and statistical data". In: *Machine Learning* 20 (1995), pp. 197–243.

[HJ12] R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge university press, 2012.

[HSW89] K. Hornik, M. Stinchcombe, and H. White. "Multilayer feedforward networks are universal approximators". In: *Neural networks* 2.5 (1989), pp. 359–366.

[Hua+18] B. Huang et al. "Generalized score functions for causal discovery". In: *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 2018, pp. 1551–1560.

[HS12] Z. Huseynov and A. Shykhammedov. "On bases of sines and cosines in Sobolev spaces". In: *Applied Mathematics Letters* 25.3 (2012), pp. 275–278. ISSN: 0893-9659.

[Ji+18] Q. Ji et al. "Network causality structures among Bitcoin and other financial assets: A directed acyclic graph approach". In: *The Quarterly Review of Economics and Finance* 70 (2018), pp. 203–213.

[KB14] D. P. Kingma and J. Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[Lau96] S. L. Lauritzen. *Graphical models*. Vol. 17. Clarendon Press, 1996.

[Maa+19] M. Maathuis et al., eds. *Handbook of graphical models*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, FL, 2019, pp. xviii+536.

[Mag85] J. R. Magnus. "On differentiating eigenvalues and eigenvectors". In: *Econometric theory* 1.2 (1985), pp. 179–191.

[MT99] D. Margaritis and S. Thrun. "Bayesian network induction via local neighborhoods". In: *Advances in Neural Information Processing Systems* 12 (1999).

[Moo+16] J. M. Mooij et al. “Distinguishing Cause from Effect Using Observational Data: Methods and Benchmarks”. In: *Journal of Machine Learning Research* 17.32 (2016), pp. 1–102.

[Naz+23] A. Nazaret et al. “Stable Differentiable Causal Discovery”. In: *arXiv preprint arXiv:2311.10263* (2023).

[Nem99] A. Nemirovsky. “Optimization II. Numerical methods for nonlinear continuous optimization”. In: (1999).

[NGZ20] I. Ng, A. Ghassami, and K. Zhang. “On the role of sparsity and DAG constraints for learning linear DAGs”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 17943–17954.

[Ram+17] J. Ramsey et al. “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”. In: *International Journal of Data Science and Analytics* 3 (2017), pp. 121–129.

[Rav+09] P. Ravikumar et al. “Sparse additive models”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 71.5 (2009), pp. 1009–1030.

[Ros+13] L. A. Rosasco et al. “Nonparametric sparsity and regularization”. In: *JLMR* (2013).

[Rud91] W. Rudin. *Functional analysis*. Vol. 2. McGraw-Hil, 1991.

[Sar] V. de Sartenejas. “TAYLOR SERIES FOR MULTI-VARIABLE FUNCTIONS”. In: ().

[Sch67] S. C. Schwartz. “Estimation of probability density by an orthogonal series”. In: *The Annals of Mathematical Statistics* (1967), pp. 1261–1265.

[SG91] P. Spirtes and C. Glymour. “An algorithm for fast recovery of sparse causal graphs”. In: *Social Science Computer Review* 9.1 (1991), pp. 62–72.

[SZ19] P. Spirtes and K. Zhang. “Search for causal models”. In: *Handbook of graphical models*. Chapman & Hall/CRC Handb. Mod. Stat. Methods. CRC Press, Boca Raton, FL, 2019, pp. 439–469.

[SC08] I. Steinwart and A. Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[Tsa+03] I. Tsamardinos et al. “Algorithms for large scale Markov blanket discovery.” In: *FLAIRS*. Vol. 2. 2003, pp. 376–81.

[Vaa00] A. W. Van der Vaart. *Asymptotic statistics*. Vol. 3. Cambridge university press, 2000.

[WBD24] D. Waxman, K. Butler, and P. M. Djurić. “Dagma-DCE: Interpretable, Non-Parametric Differentiable Causal Discovery”. In: *IEEE Open Journal of Signal Processing* (2024).

[Wen04] H. Wendland. *Scattered data approximation*. Vol. 17. Cambridge university press, 2004.

[Xu+22] D. Xu et al. “On the sparse DAG structure learning based on adaptive Lasso”. In: *arXiv preprint arXiv:2209.02946* (2022).

[Yu+19] Y. Yu et al. “DAG-GNN: DAG structure learning with graph neural networks”. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 7154–7163.

[Zha+23] J. Zhang et al. “Active learning for optimal intervention design in causal models”. In: *Nature Machine Intelligence* 5.10 (2023), pp. 1066–1075.

[Zhe+18] X. Zheng et al. “DAGs with NO TEARS: Continuous optimization for structure learning”. In: *Advances in Neural Information Processing Systems* 31 (2018).

[Zhe+20] X. Zheng et al. “Learning sparse nonparametric DAGs”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 3414–3425.

[Zho+14] K. Zhong et al. “Proximal quasi-newton for computationally intensive l1-regularized m-estimators”. In: *Advances in Neural Information Processing Systems* 27 (2014).

[Zho08] D.-X. Zhou. “Derivative reproducing properties for kernel methods in learning theory”. In: *Journal of computational and Applied Mathematics* 220.1-2 (2008), pp. 456–463.