Journal of Machine Learning Research (2025) 1-49 Submitted 1/24; Revised 6/25; Published 6/25

Boosting Causal Additive Models

Maximilian Kertel MAXIMILIANKERTEL@GMAIL.COM
Technology Development Battery Cell
BMW Group
Munich, Germany

Nadja Klein NADJA.KLEIN@KIT.EDU
Scientific Computing Center
Karlsruhe Institute of Technology
Karlsruhe, Germany

Editor: Pradeep Ravikumar

Abstract

We present a boosting-based method to learn additive Structural Equation Models (SEMs) from observational data, with a focus on the theoretical aspects of determining the causal order among variables. We introduce a family of score functions based on arbitrary regression techniques, for which we establish sufficient conditions that guarantee consistent identification of the true causal ordering. Our analysis reveals that boosting with early stopping meets these criteria and thus offers a consistent score function for causal orderings. To address the challenges posed by high-dimensional data sets, we adapt our approach through a component-wise gradient descent in the space of additive SEMs. Our simulation study supports the theoretical findings in low-dimensional settings and demonstrates that our high-dimensional adaptation is competitive with state-of-the-art methods. In addition, it exhibits robustness with respect to the choice of hyperparameters, thereby simplifying the tuning process.

Keywords: causal discovery; directed acyclic graph; boosting; high-dimensional data; reproducing kernel Hilbert space

1. Introduction

Causal discovery is the process of deriving causal relationships between variables in a system. These help in improving decisions, predictions and interventions (Kyono et al., 2020). With the rapid growth of large-scale data sets in fields such as healthcare, genetics (Aibar et al., 2017), or manufacturing (Kertel et al., 2023), causal discovery has become increasingly important. Traditionally, the researcher designs an experiment and intervenes on certain variables—for example, by assigning a drug or a placebo—which allows to estimate the causal effect of the manipulated variables. However, it is often more practical, less time- and resource-intensive, or ethically preferable to collect data from the system in a steady state, where no variables are manipulated. For instance, in complex manufacturing domains, the number of input variables might be too large to conduct a design of experiments, and additionally those experiments would lead to production downtime, resulting in high costs.

Thus, although observational data is often complex, noisy, or has confounding variables (Bhattacharya et al., 2021), it is still preferable to derive causal relationships from observational data rather than having to rely on impractical experimental studies.

In this work, we follow the assumption that the causal relationships between variables can be modeled as a Directed Acyclic Graph (DAG). This assumption implies that the impact between two variables flows in at most one direction, and there are no cyclic or self-reinforcing pathways. It is the goal of the present work to identify the DAG from observational data. Existing algorithms for this task can be broadly classified into two categories. First, constraint-based methods rely on statistical tests for conditional independence, but these approaches become computationally expensive and difficult apart from multivariate normal or multinomial distributions (Zhang et al., 2011; Shah and Peters, 2018). Additionally, many of them assume some kind of faithfulness (Peters et al., 2017; Raskutti and Uhler, 2018), and the identified graph is typically not unique (Spirtes et al., 1993; Kalisch and Bühlman, 2007).

Instead, we focus on the second category for identifying DAGs from observational data, namely on Structural Equation Models (SEMs). SEMs rely on the assumption that each variable is a function of other variables in the system and a perturbing noise term. Peters et al. (2014) show that if one restricts the functional relationships and the noise, called the Additive Noise Model (ANM), then the implied distribution is unique. Specifically, this is the case when the functional relationships are non-linear, and the noise term is Gaussian, which we assume throughout this work. Under these conditions, the underlying SEM can be uniquely identified from observational data.

Causal discovery using ANMs is an active field of research (Vowels et al., 2022). Many recent works leverage continuous acyclity characterisations (Lachapelle et al., 2019; Yu et al., 2019; Zheng et al., 2020; Kalainathan et al., 2022; Ng et al., 2022a,b) and find the DAG using gradient-descent. However, in contrast to earlier works (as for example, Shimizu et al., 2011; Bühlmann et al., 2014) the statistical properties of most machine learning-based methods remain poorly understood (Kaiser and Sipos, 2022). In this context, a notable contribution is that of Aibar et al. (2017), which successfully derives the cyclic graphical representation of a gene regulatory network using boosting. Our work is an extension of this approach towards acyclic graphs.

In this paper, we leverage the success of gradient-based methods towards statistical boosting for causal discovery in ANMs and investigate the underlying statistical properties. In particular, we show consistency and robustness of our proposed method.

Our main contributions are the following.

- Many existing consistency results are derived for specific regression techniques—such as maximum likelihood or penalized regression—where consistency of the causal discovery algorithm follows from the statistical properties of the chosen regression method (Bühlmann et al., 2014; Nowzohour and Bühlmann, 2016; van de Geer, 2014). In contrast, we take a more general approach by providing sufficient conditions on generic regression techniques under which the causal discovery algorithm is consistent; see Proposition 5.

- Theorem 15 establishes that L_2 boosting with early stopping satisfies the conditions of Proposition 5. Consequently, L_2 boosting can be employed for consistent causal discovery.

- We show in Theorem 18 that L_2 boosting with early stopping avoids overfitting even in the presence of misspecification.

- We propose a variant of the boosting procedure for high-dimensional settings, where the number of variables p is large.

- We conduct a simulation study demonstrating that our approach is competitive with state-of-the-art methods. We furthermore show that it is robust to the choice of the hyperparameters, making it easy to tune in practice.

The remainder of this paper is organized as follows. Section 2 introduces SEMs and outlines the necessary assumptions. Section 3 reviews L_2 boosting and Reproducing Kernel Hilbert Spaces (RKHSs). In Section 4, we present the proposed novel causal discovery method and show its consistency. Section 5 provides an empirical evaluation of our method and benchmarks its performance to various state-of-the-art algorithms. Finally, Section 6 concludes the paper. Technical details and proofs can be found in the Appendix.

2. Causal Discovery

In this section we present a class of SEMs called Causal Additive Models (CAMs). We summarize the key results from Bühlmann et al. (2014) and Peters et al. (2014) which show how regression estimators can be utilized for causal discovery. Finally, while existing contributions focus on specific regression estimators as maximum likelihood in Bühlmann et al. (2014) or van de Geer (2014), we adopt a more general perspective: Proposition 5 provides sufficient conditions on a generic regression estimator so that the derived causal discovery algorithm is consistent.

Let $\mathbf{X} = (X_1, \ldots, X_p)^\top$ be a p-dimensional random vector. For any $S = \{s_1, \ldots, s_T\} \subset \{1, \ldots, p\}$ and a vector $\mathbf{x} \in \mathbb{R}^p$ we define $\mathbf{x}_S := (x_{s_1}, \ldots, x_{s_T})^\top$. Analogously, for the random vector $\mathbf{X}$ we set $\mathbf{X}_S := (X_{s_1}, \ldots, X_{s_T})^\top$. We assume that the elements of S are ordered to make the representations unique, such that $s_1 < s_2 < \ldots < s_T$. Let G be a Directed Acyclic Graph (DAG) over the variables $X_1, \ldots, X_p$. For each $k \in \{1, \ldots, p\}$ we denote by $\mathbf{pa}_G(k)$ the set of parents of node k in G, i.e., the set of indices $j \in \{1, \ldots, p\}$ for which the edge $X_j \to X_k$ exists in G. We assume that there exists a DAG G such that $\mathbf{X}$ follows a SEM with additive noise, that is

$$X_k = f_k\left(\mathbf{X}_{\mathbf{pa}_G(k)}\right) + \varepsilon_k. \tag{1}$$

Here, ε_k are i.i.d. noise terms for $k = 1, \ldots, p$.

Every DAG has at least one topological ordering π (that is, a permutation) on $\{1, \ldots, p\}$. We denote the nonempty set of topological orderings for G by $\Pi(G)$. With $\pi \in \Pi(G)$ there can only be a directed path in G from X_j to X_k if $\pi(j) < \pi(k)$ but not vice versa; see Figure 1 for an illustrating example.

$$X_1 = f_1(X_2) + \varepsilon_1$$
$$X_2 = \varepsilon_2$$
$$X_3 = f_3(X_2) + \varepsilon_3$$

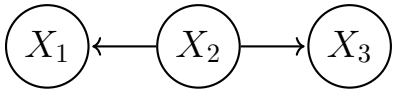

Figure 1: An example of a SEM on the left-hand side and its corresponding DAG G on the right-hand-side for $p = 3$. The set of possible topological orderings is $\{(2,1,3),(2,3,1)\}$. For $\pi^0 = (2,3,1)$ the structural equation for X_1 is given by $X_1 = f_1(X_2) + \varepsilon_1 = f_{12}(X_2) + f_{13}(X_3) + \varepsilon_1$ with $f_{12} = f_1$ and $f_{13} = 0$.

2.1 Identifiability

The goal of our analysis is to identify the DAG G from the distribution of $\mathbf{X}$. In general there is no one-to-one correspondence between the distribution $P(\mathbf{X})$ and the underlying SEM or G. However, if we impose appropriate restrictions on the noise terms $\{\varepsilon_k : k = 1,\dots,p\}$ and the functions $\{f_k : k = 1,\dots,p\}$, then the desired one-to-one correspondence between a distribution and a SEM exists (see Peters et al., 2014, Corollary 31). In this case, we call the SEM identifiable. We consider SEMs that satisfy the following assumptions, which ensure identifiability.

Assumption 1 *For the SEM of Equation* (1) *assume that f_k has the additive decomposition*

$$f_k(\mathbf{x}_{\mathbf{pa}_G(k)}) = \sum_{j \in \mathbf{pa}_G(k)} f_{kj}(x_j), \quad k = 1,\dots,p,$$

where the f_{kj} are three times differentiable, non-linear and non-constant for any $k = 1,\dots,p$ and $j \in \mathbf{pa}_G(k)$. Further, let $(\varepsilon_1,\dots,\varepsilon_p)$ be a random vector of independent components, which are normally distributed with mean zero and standard deviations $\sigma_1,\dots,\sigma_p > 0$. We call SEMs of this form Causal Additive Models (CAMs; Bühlmann et al., 2014).

Define $\varpi_\pi(k) = \{j : \pi(j) < \pi(k)\}$ as the predecessors of k with respect to a permutation π. For example, when $p = 3$ and $\pi = (2,1,3)$, we have: $\varpi_\pi(1) = \{2\}, \varpi_\pi(2) = \emptyset, \varpi_\pi(3) = \{1,2\}$.

Consider a CAM, which is characterized by functions $f_1,\dots,f_p$, standard deviations $\sigma_1,\dots,\sigma_p$, and a DAG G with topological ordering π. The implied density is

$$p(\mathbf{x}) = \prod_{k=1}^{p} p\left(x_k|\mathbf{x}_{\mathbf{pa}_G(k)}\right) = \prod_{k=1}^{p} p\left(x_k|\mathbf{x}_{\varpi_\pi(k)}\right). \tag{2}$$

Here, $p\left(x_k|\mathbf{x}_S\right)$ is the density of the conditional distribution of X_k given $\mathbf{X}_S = \mathbf{x}_S$, which we assume to exist throughout this work. For $S = \emptyset$, we define $p\left(x_k|\mathbf{x}_S\right) = p\left(x_k\right)$. The second equality in Equation (2) holds because X_k is independent of its predecessors in π given its parents (see Peters et al., 2017, Proposition 6.31), that is

$$X_k \perp X_{\varpi_\pi(k)\setminus\mathbf{pa}_G(k)}|\mathbf{X}_{\mathbf{pa}_G(k)}. \tag{3}$$

For any combination of $f_1, \ldots, f_p$, G and any topological ordering π of G, it trivially holds

$$f_k = \sum_{j \in \mathbf{pa}_G(k)} f_{kj}(X_j) = \sum_{j \in \varpi_\pi(k)} \widetilde{f_{kj}}(X_j),$$

where $\widetilde{f_{kj}} = f_{kj}$ if $j \in \mathbf{pa}_G(k)$ and $\widetilde{f_{kj}} = 0$ if $j \notin \mathbf{pa}_G(k)$. Consequently, any CAM characterized by $f_1, \ldots, f_p, G, \sigma_1, \ldots, \sigma_p$ can be reparameterized by $f_1, \ldots, f_p, \pi, \sigma_1, \ldots, \sigma_p$, where π is a topological ordering of G. Note that the latter parametrization is not unique, since π can be chosen arbitrarily from the set of topological orderings of G. However, once the topological ordering π is known, G can be found by identifying the parents of X_k (those j for which $f_{kj} \neq 0$) within $\varpi_\pi(k)$ for any $k = 1, \ldots, p$. This is straightforward using pruning or feature selection methods (Teyssier and Koller, 2005; Shojaie and Michailidis, 2010; Bühlmann et al., 2014). Thus, we simplify our objective and instead of searching for G, we aim to identify its topological ordering π.

Consider a CAM characterized by $f_1, \ldots, f_p, G, \sigma_1, \ldots, \sigma_p$. It can be reparameterized by the parameter tuple $\theta = (f_1, \ldots, f_p, \pi, \sigma_1, \ldots, \sigma_p)$. Using Equation (3), it follows that the implied conditional distribution of $X_k | \mathbf{X}_{\varpi_\pi(k)} = \mathbf{x}_{\varpi_\pi(k)}$ is Gaussian with mean $f_k\left(\mathbf{x}_{\varpi_\pi(k)}\right)$ and standard deviation σ_k. The implied log-density $\log(p_\theta)$ is given by

$$\log(p_\theta(\mathbf{x})) = \sum_{k=1}^{p} \log\left(\frac{1}{\sigma_k}\phi\left(\frac{x_k - f_k(\mathbf{x}_{\varpi_\pi(k)})}{\sigma_k}\right)\right),$$

where ϕ denotes the density function of a univariate standard normal distribution. From now on let $\mathbf{X}$ follow a CAM characterized by $\theta^0 = \left(f_1^0, \ldots, f_p^0, \pi^0, \sigma_1^0, \ldots, \sigma_p^0\right)$. To identify θ^0, we define the population score function

$$\theta \mapsto \mathbb{E}_{p_{\theta^0}}\left[-\log\left(p_\theta(x)\right)\right].$$

It holds

$$\mathbb{E}_{p_{\theta^0}}\left[-\log\left(p_{\theta^0}(x)\right)\right] \leq \mathbb{E}_{p_{\theta^0}}\left[-\log\left(p_\theta(x)\right)\right],$$

with equality if and only if $p_{\theta^0} = p_\theta$ by the properties of the Kullback-Leibler divergence. For such minimizing θ, their ordering π must lie in $\Pi(G^0)$ by the identifiability.

Let us consider the problem of minimizing the score function with respect to θ. Fixing π and $f_1, \ldots, f_p$ in θ, and minimizing with respect $\sigma_1, \ldots, \sigma_p$ leads to the minimizers

$$\sigma^2_{k, p_{\theta^0}, f_k, \pi} := \mathbb{E}_{p_{\theta^0}}\left[\left(X_k - f_k(\mathbf{x}_{\varpi_\pi(k)})\right)^2\right] = \mathbb{E}_{p_{\theta^0}}\left[\left(X_k - \sum_{j \in \varpi_\pi(k)} f_{kj}(X_j)\right)^2\right]$$

for $k = 1, \ldots, p$. Hence, when minimizing the score function we only need to consider the subset of the parameter space $(f_1, \ldots, f_p, \pi, \sigma_1, \ldots, \sigma_p)$, where $\sigma_k = \sigma_{k, p_{\theta^0}, f_k, \pi}, k = 1, \ldots, p$,

that is, the relevant parameter space reduces to $(f_1, \ldots, f_p, \pi)$. Thus, it holds

$$\begin{aligned}\arg\min_{\theta} \mathbb{E}_{p_{\theta^0}}\left[-\log\left(p_\theta(x)\right)\right] &= \arg\min_{\theta=(f_1,\ldots,f_p,\pi,\sigma_1=\sigma_{1,p_{\theta^0},f_1,\pi},\ldots,\sigma_p=\sigma_{p,p_{\theta^0},f_p,\pi})} \mathbb{E}_{p_{\theta^0}}\left[-\log\left(p_\theta(x)\right)\right] \\ &= \arg\min_{(f_1,\ldots,f_p,\pi)} \sum_{k=1}^{p} \log(\sigma^2_{k,p_{\theta^0},f_k,\pi}) + C \\ &= \arg\min_{(f_1,\ldots,f_p,\pi)} \sum_{k=1}^{p} \log(\sigma^2_{k,p_{\theta^0},f_k,\pi}),\end{aligned}$$

with C only depending on p. Denote the functions aligning with π by

$$\vartheta(\pi) = \left\{ (f_1, \ldots, f_p) : f_k = \sum_{\pi(j)<\pi(k)} f_{kj}, f_{kj} : \mathbb{R} \to \mathbb{R}, f_{kj} \text{ is } \begin{array}{l}\text{three times differentiable,}\\ \text{non-linear, and}\\ \text{non-constant.}\end{array} \right\}.$$

We fix π and optimize with respect to $f_1, \ldots, f_p$ to define a population score on the orderings

$$S(\pi) = \min_{(f_1,\ldots,f_p)\in\vartheta(\pi)} \sum_{k=1}^{p} \log(\sigma^2_{k,p_{\theta^0},f_k,\pi}). \tag{4}$$

By identifiability, $S(\pi)$ is minimal if and only if $\pi^0 \in \Pi(G^0)$, that is,

$$\sum_{k=1}^{p} \log\left(\left(\sigma_k^0\right)^2\right) = S(\pi^0) < S(\pi)\, \forall \pi^0 \in \Pi(G^0), \pi \notin \Pi(G^0). \tag{5}$$

Intuitively, the score $S(\pi)$ measures how much variance remains when any X_k is regressed on its predecessors $\mathbf{X}_{\varpi_\pi(k)}$. An example for $p = 2$ is depicted in Figure 2.

2.2 Estimation of the Ordering

In practice the true parameter tuple θ^0 is unknown. Instead, we observe N realizations $\mathbf{x}^N := (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ of $\mathbf{X}$ with density p_{θ^0}, where $\mathbf{x}_\ell \in \mathbb{R}^p, \ell = 1, \ldots, N$ and $\mathbf{x}_\ell = (x_{\ell 1}, \ldots, x_{\ell p})^\top$. It is natural to consider the empirical version of the population score function (4)

$$\widehat{S}(\pi) = \sum_{k=1}^{p} \log(\widehat{\sigma}^2_{k,\widehat{f}_{k,\pi}}), \tag{6}$$

where $\widehat{\sigma}^2_{k,\widehat{f}_{k,\pi}}, k = 1, \ldots, p$ is defined as

$$\widehat{\sigma}^2_{k,\widehat{f}_{k,\pi}} = \frac{1}{N}\sum_{\ell=1}^{N}\left(x_{\ell k} - \widehat{f}_{k,\pi}(\mathbf{x}_{\ell\varpi_\pi(k)})\right)^2 = \frac{1}{N}\sum_{\ell=1}^{N}\left(x_{\ell k} - \sum_{j\in\varpi_\pi(k)} \widehat{f}_{kj,\pi}(x_{\ell j})\right)^2.$$

Here $\widehat{f}_{k,\pi} = \sum_{j\in\varpi_\pi(k)} \widehat{f}_{kj,\pi}$ is a regression function estimate using data $(\mathbf{x}_{\ell\varpi_\pi(k)}, x_{\ell k}), \ell = 1, \ldots, N$ with the convention $\mathbf{x}_{\ell S} := \mathbf{x}_{\ell_S} = (\mathbf{x}_{\ell s_1}, \ldots, \mathbf{x}_{\ell s_T})$ for $S = \{s_1, \ldots, s_T\}$ introduced before. Although the regression estimates depend on the data and thus N, we omit the additional index for better readability.

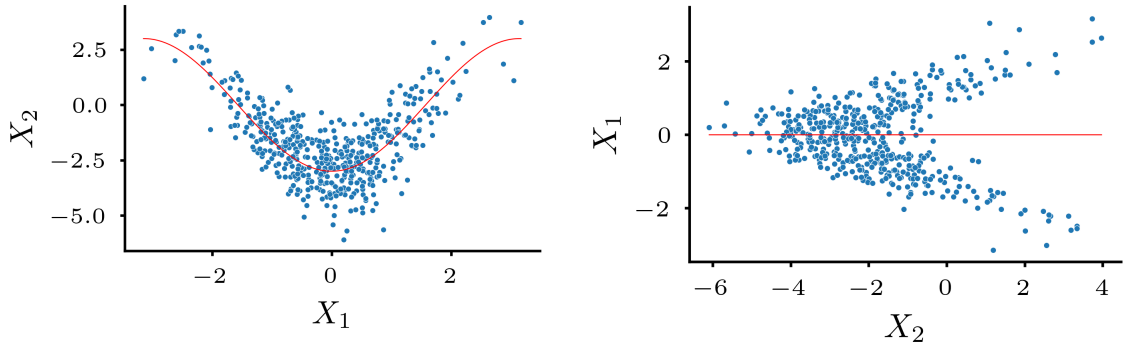

Figure 2: The blue dots represent 500 realizations of a distribution following a SEM with $p = 2$ and $X_1 = \varepsilon_1 \sim \mathcal{N}(0,1)$ and $X_2 = -3\cos(X_1) + \varepsilon_2$ with $\varepsilon_2 \sim \mathcal{N}(0,1)$. In the left-hand plot, X_2 is shown on the y-axis and X_1 on the x-axis, while in the right-hand-plot, the axes are reversed. The red lines give the conditional mean functions. We observe that $\arg\min_{(f_1,f_2)\in\vartheta((1,2))} \sum_{k=1}^{2} \log(\sigma^2_{k,p_{\theta^0},f_k,(1,2)}) = (0, -3\cos(x_1))$ and $\arg\min_{(f_1,f_2)\in\vartheta((2,1))} \sum_{k=1}^{2} \log(\sigma^2_{k,p_{\theta^0},f_k,(2,1)}) = (0,0)$. The distribution $X_1 - \mathbf{E}[X_1|X_2 = x_2]$ becomes bi-modal for larger values of x_2. The unexplained noise (distance of blue dots to red line) is smaller in the left plot, which is the correct ordering, thus $S((1,2)) < S((2,1))$.

Remark 2 *In contrast to the population version* (4)*, it is unclear whether* (6) *is also minimized at* $\pi^0 \in \Pi(G^0)$ *even for* $N \to \infty$*. This is due to the fact, that the regression functions* $\widehat{f}_{k,\pi}$ *and the prediction errors* $\widehat{\sigma}^2_{k,\widehat{f}_{k,\pi}}$ *are estimated from a finite sample.*

Since any regression function estimator induces a score function $\widehat{S}$, the following remark gives some intuition on required properties through two "extreme" examples.

Remark 3

1. Overfitting: *Let the regression estimator interpolate the data, that is,* $\widehat{f}_{k,\pi}(x_{\ell,\varpi_\pi(k)}) = x_{\ell k}$ *for all* $\ell = 1,\dots,N$*,* $k = 1,\dots,p$ *and all* π*. Then the regression estimator is overfitting, and the score diverges to* $-\infty$ *for any permutation* π*..*

2. Underfitting: *Let* $\widehat{f}_{k,\pi}(x_{\ell,\varpi_\pi(k)}) = 0$ *for all* $\ell = 1,\dots,N$*,* $k = 1,\dots,p$ *and all* π*. Then the regression estimator is underfitting, and all permutations have the same score* $\sum_{k=1}^{p} \log\left(\frac{1}{N}\sum_{\ell=1}^{N} x^2_{\ell k}\right)$*.*

In both cases, $\widehat{S}$ *cannot identify a correct ordering* $\pi^0 \in \Pi(G^0)$ *even with infinite data. Intuitively, a suitable regression estimator should:*

1. *avoid overfitting and preserve the unexplained noise, and*

2. *produce estimates that are close to the true regression functions* $f^0_1,\dots,f^0_p$*.*

In this work, we apply L_2 boosting regression in conjunction with early stopping (Bühlmann and Yu, 2003; Raskutti et al., 2014) and show that the resulting score in Equation (6) consistently favors a permutation $\pi^0 \in \Pi(G^0)$. Proposition 5 below formalizes the intuition of Remark 3 and states sufficient conditions on the regression function estimator. In the following Definition 4, Y takes the role of X_k, while $\widetilde{\mathbf{X}}$ takes the role of $\mathbf{X}_{\varpi_\pi(k)}$.

Definition 4 (Non-overfitting) *Let $\widehat{f}$ be a regression estimate based on N i.i.d. samples $(\widetilde{\mathbf{x}}_1, y_1), \ldots, (\widetilde{\mathbf{x}}_N, y_N)$ from $(\widetilde{\mathbf{X}}, Y)$. We say the estimator is not overfitting w.r.t. $(\widetilde{\mathbf{X}}, Y)$, if*

$$\left| \frac{1}{N} \sum_{\ell=1}^{N} \left(y_\ell - \widehat{f}(\widetilde{\mathbf{x}}_\ell) \right)^2 - \mathbb{E}_{\widetilde{\mathbf{X}},Y} \left[\left(Y - \widehat{f}(\widetilde{\mathbf{X}}) \right)^2 \right] \right|$$

converges to 0 in probability for $N \to \infty$.

Proposition 5 *Let Assumption 1 hold. Then, if the regression estimator is such that*

1. *$\widehat{f}_{k,S}$ is not overfitting with respect to $(\mathbf{X}_S, X_k)$ for any combination of $k = 1, \ldots, p$ and $S \subset \{1, \ldots, p\} \setminus \{k\}$ according to Definition 4 and*
2. *$\widehat{\sigma}^2_{k,\widehat{f}_{k,\pi^0}} \xrightarrow{\mathbb{P}} \sigma^2_{k,p_{\theta^0},f^0_k,\pi^0} = \left(\sigma^0_k\right)^2$ for all $\pi^0 \in \Pi(G^0)$ and $k = 1, \ldots, p$, that is*

$$\frac{1}{N} \sum_{\ell=1}^{N} \left(x_{\ell k} - \widehat{f}_{k,\varpi_{\pi^0}(k)}(\mathbf{x}_{\ell \varpi_{\pi^0}(k)}) \right)^2 \xrightarrow{\mathbb{P}} \left(\sigma^0_k\right)^2 = \mathbb{E}\left[\left(X_k - \sum_{j \in \mathbf{pa}_{G^0}(k)} f^0_{kj}(X_j) \right)^2 \right].$$

Then it holds for the derived score function $\widehat{S}$ that

$$\widehat{S}(\pi^0) < \widehat{S}(\pi)$$

for any $\pi^0 \in \Pi(G^0)$ and $\pi \notin \Pi(G^0)$ with probability going to 1 for $N \to \infty$.

Remark 6 *Proposition 5 is more general than existing results such as Theorem 1 in Bühlmann et al. (2014) as it does not specify a particular regression estimator. The challenge in applying Proposition 5 is to find a combination of a regression estimator and corresponding distributional assumptions on $\mathbf{X}$ so that both conditions are fulfilled. For nonparametric maximum likelihood regression, Bühlmann et al. (2014) give these conditions in (A1)–(A4). In contrast, for L_2 boosting we present the required conditions via Assumptions 10, 11, and 13.*

Sketch of the proof of Proposition 5 Our goal is to show that for any $\pi \notin \Pi(G^0)$ and $\pi^0 \in \Pi(G^0)$ it holds asymptotically

$$\sum_{k=1}^{p} \log(\widehat{\sigma}^2_{k,\widehat{f}_{k,\pi^0}}) = \widehat{S}(\pi^0) < \widehat{S}(\pi) = \sum_{k=1}^{p} \log(\widehat{\sigma}^2_{k,\widehat{f}_{k,\pi}}).$$

By inequality (5) this is fulfilled if for $N \to \infty$ and $\pi \notin \Pi(G^0)$

$$\lim_{N \to \infty} \widehat{S}(\pi) \geq S(\pi) \tag{7}$$

and if at the same time for $\pi^0 \in \Pi(G^0)$ it holds that

$$\lim_{N\to\infty} \widehat{S}(\pi^0) = S(\pi^0). \tag{8}$$

Applying the continuous mapping theorem, (8) is ensured by Condition 2.

Contrariwise for relation (7) when $\pi \notin \Pi(G^0)$, the non-overfitting Condition 1. of Proposition 5 ensures that for any $k = 1, \ldots, p$ and for $N \to \infty$

$$\widehat{\sigma}^2_{k,\widehat{f}_{k,\pi}} = \frac{1}{N}\sum_{\ell=1}^{N}\left(x_{\ell k} - \sum_{j\in\varpi_\pi(k)} \widehat{f}_{kj,\pi}\left(x_{\ell j}\right)\right)^2 \geq \mathbb{E}_{p_{\theta^0}}\left[\left(X_k - \sum_{j\in\varpi_\pi(k)} \widehat{f}_{kj,\pi}\left(X_j\right)\right)^2\right].$$

It thus follows that for $N \to \infty$

$$\begin{aligned}
\widehat{S}(\pi) &= \sum_{k=1}^{p} \log\left(\widehat{\sigma}^2_{k,\widehat{f}_{k,\pi}}\right) = \sum_{k=1}^{p} \log\left(\frac{1}{N}\sum_{\ell=1}^{N}\left(x_{\ell k} - \sum_{j\in\varpi_\pi(k)} \widehat{f}_{kj,\pi}\left(x_{\ell j}\right)\right)^2\right) \\
&\geq \sum_{k=1}^{p} \log\left(\mathbb{E}_{p_{\theta^0}}\left[\left(X_k - \sum_{j\in\varpi_\pi(k)} \widehat{f}_{kj,\pi}\left(X_j\right)\right)^2\right]\right) \\
&\geq \min_{(f_1,\ldots,f_p)\in\vartheta(\pi)} \sum_{k=1}^{p} \log(\sigma^2_{k,p_{\theta^0},f_k,\pi}) = S(\pi).
\end{aligned}$$

A detailed proof can be found in the Appendix A.

3. Boosting and Reproducing Kernel Hilbert Spaces

As our main result in Theorem 15 relies on boosting-based Kernel Hilbert space regressions, we briefly review relevant known concepts and results on boosting (Section 3.1) and Reproducing Kernel Hilbert Spaces (RKHSs) (Section 3.2) next. Further details on the two concepts can be found in Bühlmann and Yu (2003); Schapire and Freund (2012) for boosting, and in Wahba (1990); Schölkopf and Smola (2001); Wainwright (2019) for RKHSs.

3.1 Boosting

L_2 boosting addresses the problem of finding a function f in some function space H that minimizes the expected L_2 loss

$$\frac{1}{2}\mathbb{E}_{\mathbf{X},Y}\left[\left(Y - f(\mathbf{X})\right)^2\right]. \tag{9}$$

In practice, (9) is replaced by the empirical minimizer of

$$\frac{1}{2N}\sum_{\ell=1}^{N}\left(y_\ell - f(\mathbf{x}_\ell)\right)^2 \tag{10}$$

based on N i.i.d. samples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$. L_2 boosting employs a functional gradient descent approach using a base learner S that maps $y^N := (y_1, \ldots, y_N)^\top$ to the estimates $\widehat{f}$

and $\widehat{y}^N = \widehat{f}(\mathbf{x}^N)$, where $\widehat{f}(\mathbf{x}^N) := \left(\widehat{f}(\mathbf{x}_1), \ldots, \widehat{f}(\mathbf{x}_N)\right)^\top$. More precisely, after initializing $\widehat{f}^{(0)}$, in each boosting step $m = 1, \ldots, m_{\text{stop}}$, the residuals at the current $\widehat{f}^{(m)}$

$$u_\ell = \frac{\partial}{\partial f} \frac{1}{2N} \left(y_\ell - f(x_\ell)\right)^2 |_{f=\widehat{f}^{(m)}} = \frac{1}{N} \left(y_\ell - \widehat{f}^{(m)}(x_\ell)\right)$$

are computed and $\widehat{f} = S(u) = S(u_1, \ldots, u_N)$ is determined. The solution is then used to update the estimate of the regression function

$$\widehat{f}^{(m+1)} = \widehat{f}^{(m)} + \upsilon \widehat{f},$$

where $0 < \upsilon \leq 1$ is the step size commonly fixed at a small value (Bühlmann and Yu, 2003). For many base learners S this leads to overfitting for fixed N as $m_{\text{stop}} \to \infty$. Hence, stopping earlier is desirable and *early stopping* is often applied (Schapire and Freund, 2012). Following Bühlmann and Yu (2003); Raskutti et al. (2014), we consider linear and symmetric base learners S, that is, $S : y^N \mapsto \widehat{y}^N$ is a linear and symmetric mapping. Spline regression and linear regression, including the generalized ridge regression sense, fall under this definition. The same holds for (additive) kernel ridge regression (Raskutti et al., 2014; Kandasamy and Yu, 2016), which we will consider in the following. In contrast, the popular choice of decision trees is not linear. Proposition 1 of Bühlmann and Yu (2003) shows that the estimate $\widehat{\mathbf{y}}$ after m boosting steps for $\mathbf{y}$ is given by

$$\widehat{f}^{(m)}(\mathbf{x}^N) = \widehat{\mathbf{y}}^N = B^{(m)}\mathbf{y} = (I - (I - S)^m)\mathbf{y}.$$

Since we assume S to be symmetric, there exists an orthogonal $U \in \mathbb{R}^{N \times N}$ containing the eigenvectors of S, such that for a diagonal matrix D with the eigenvalues of S on the diagonal, it holds that $S = UDU^T$. It follows that $B^{(m)} = U(I - (I - D)^m)U^T$.

3.2 Reproducing Kernel Hilbert Spaces

We choose the function estimates $\widehat{f}$ from a Reproducing Kernel Hilbert Space (RKHS) H, while S is a kernel regression estimator. We start by introducing kernel functions.

Definition 7 *We call a symmetric function $K : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ a positive definite (p.d.) kernel on $\mathbb{R}^p$ if*

$$\sum_{k=1}^{N} \sum_{\ell=1}^{N} \alpha_k \alpha_\ell K(\mathbf{x}_k, \mathbf{x}_\ell) \geq 0$$

for any $\{\alpha_1, \ldots, \alpha_N\} \subset \mathbb{R}$ and any $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset \mathbb{R}^p$.

For any p.d. kernel K, there exists a unique Hilbert space H with $K(\cdot, \mathbf{x}) \in H \, \forall \mathbf{x} \in \mathbb{R}^p$ and where further it holds for any $f \in H$ and $\mathbf{x} \in \mathbb{R}^p$

$$f(\mathbf{x}) = < f, K(\cdot, \mathbf{x}) >_H . \tag{11}$$

Equation (11) is called the reproducing property. Consequently H is called a RKHS. By the reproducing property it holds for $f = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} \alpha_k K(\cdot, \mathbf{x}_k)$ and $g = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} \beta_k K(\cdot, \mathbf{x}_k)$, that

$$< f, g >_H = \alpha^T G \beta, \tag{12}$$

where $G \in \mathbb{R}^{N \times N}$ is symmetric with $G_{jk} = \frac{K(\mathbf{x}_j, \mathbf{x}_k)}{N}$. We call G the Gram matrix. By the representation theorem (Wainwright, 2019, Proposition 12.33) the minimizer of

$$\widehat{f} = \underset{f \in H}{\arg\min} \frac{1}{N} \sum_{\ell=1}^{N} (y_\ell - f(\mathbf{x}_\ell))^2 + \gamma ||f||_H^2 \tag{13}$$

can be expressed by

$$\widehat{f} = \frac{1}{\sqrt{N}} \sum_{\ell=1}^{N} \beta_\ell K(\cdot, \mathbf{x}_\ell),$$

where $\beta = \frac{1}{\sqrt{N}}(G + \gamma N I)^{-1} y^N$. By Equation (12), $||\widehat{f}||_H^2 = \beta^T G \beta$ holds. Clearly, the mapping $S : y^N \mapsto G(G + \lambda I)^{-1} y^N = \widehat{f}(\mathbf{x}^N)$ is linear and symmetric. It can be derived that S has the eigenvalues $d_\ell = \frac{\widehat{\mu_\ell}}{\widehat{\mu_\ell} + \gamma N}$, where $\widehat{\mu_1}, \ldots, \widehat{\mu_N}$ are the eigenvalues of G. The regularization parameter $\lambda = \gamma N$ shall be constant in N in this work.

The boosting estimate $\widehat{f}^{(m)}$ is sequentially built by adding small amounts of the current estimates. These current estimates are of the form $\frac{1}{\sqrt{N}} \sum_{\ell=1}^{N} \widehat{\alpha}_\ell K(\cdot, \mathbf{x}_\ell)$. Thus, if S is the base learner used for boosting, it holds by the construction of the boosting estimator, that there exists a $\widehat{\beta} \in \mathbb{R}^N$ with

$$\widehat{f}^{(m)} = \frac{1}{\sqrt{N}} \sum_{\ell=1}^{N} \widehat{\beta}_\ell K(\cdot, \mathbf{x}_\ell). \tag{14}$$

Here, $\widehat{f}^{(m)}$ is the boosting regression estimate after m boosting steps. In this context we define

$$\mathcal{F}_N := \left\{ f \in H : f = \frac{1}{\sqrt{N}} \sum_{\ell=1}^{N} \beta_\ell K(\cdot, \mathbf{x}_\ell), ||f||_H = 1, \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset \mathbb{R}^p \right\}$$

and $\widehat{f}^{(m)} \in h\mathcal{F}_N$ for some $h > 0$. For a continuous kernel $K \in L^2(\mathbb{R} \times \mathbb{R})$, we can define an integral operator $\mathcal{K} : L^2(\mathbb{R}) \to L^2(\mathbb{R})$ by

$$f(\cdot) \mapsto (\mathcal{K}(f))(y) = \int_{\mathbb{R}} K(x, y) f(x) d\mathbb{P}(x).$$

Under assumptions on $\mathbb{P}$ and K, Sun (2005) shows that the operator $\mathcal{K}$ has eigenvalues $\mu_k \geq 0$ and eigenfunctions $\phi_k \in L^2(\mathbb{R})$, so that $\mathcal{K}(\phi_k) = \mu_k \phi_k$. If the eigenvalues are ordered non-increasingly, then μ_k goes to 0. The decay rate of the eigenvalues will be important in our analysis. We close this subsection with two examples.

Example 1 (Kernel functions)

1. ***Gaussian kernels on*** $\mathbb{R}$*:* $K : \mathbb{R} \times \mathbb{R}$ *is defined for some* $\varsigma > 0$ *by*

$$K(x, x') = \exp\left(\frac{|x - x'|^2}{2\varsigma} \right).$$

Its eigenvalues exist under mild assumptions on $P(\mathbf{X})$ (Sun, 2005, Section 4) and follow an exponential decay of the form

$$\mu_k \leq \exp(-Ck)$$

for some $C > 0$. For further details, see Bach and Jordan (2002, Section C) and Example 3 of Cucker and Smale (2002).

2. ***Additive Kernel**: Let $H_1, \ldots, H_p$ be RKHSs with kernels K_k on X_k, $k = 1, \ldots, p$. The space $H_1 \oplus \cdots \oplus H_p := \{f : \mathbb{R}^p \to \mathbb{R} : f(x_1, \ldots, x_p) = \sum_{k=1}^p f_k(x_k), f_k \in H_k\}$ is a RKHS with kernel $K = \sum_{k=1}^p K_k$. Its norm is defined by $||f||_H^2 = \sum_{k=1}^p ||f_k||_{H_k}^2$ (see Wainwright, 2019, Proposition 12.27). For Gaussian kernels $K_1, \ldots, K_p$, we assume that the eigenvalues of K can be upper bounded by*

$$\mu_k \leq p \exp(-Ck) \tag{15}$$

for some $C > 0$. Note that for p fixed, this is of type

$$\mu_k \leq \exp(-C'k)$$

for some $C' > 0$. The solution of (13) can then be written by $\widehat{f}(x_1, \ldots, x_p) = \sum_{k=1}^p \widehat{f_k}(x_k)$, where

$$\widehat{f_k} = \sum_{\ell=1}^N \widehat{\beta}_\ell K_k(\cdot, x_{\ell k})$$

and $\widehat{\beta}$ is shared among the different components, that is, does not depend on k (see Kandasamy and Yu, 2016, for further details).

The idea behind inequality (15) is the following. Each K_k is a self-adjoint and compact operator for any $k = 1, \ldots, p$. Let A, B be linear and self-adjoint operators with non-increasing eigenvalues $\lambda_1, \lambda_2, \ldots$ and $\mu_1, \mu_2, \ldots$, respectively. It holds by Zwahlen (1965/66) for the non-increasing eigenvalues $\gamma_1, \gamma_2, \ldots$ of the self-adjoint and linear operator $A + B$ that for any $1 \leq r, s \leq p$

$$\gamma_{r+s} \leq \gamma_{r+s-1} \leq \lambda_r + \mu_s.$$

Now consider the non-increasing eigenvalues μ_ℓ^k of the operator $A_1 + \ldots + A_k$. Let $\lambda_\ell^j, \ell = 1, 2, \ldots$ be the eigenvalues of A_j. Then it holds that

$$\mu_{\ell p}^p \leq \mu_{(p-1)\ell}^{p-1} + \lambda_\ell^p \leq \ldots \leq \lambda_\ell^1 + \ldots + \lambda_\ell^p.$$

Inequality (15) follows under the assumption that $\lambda_\ell^j \leq \exp(-C\ell)$ for $j = 1, \ldots, p$ for some $C > 0$.

The Gaussian (and its additive counterpart) kernel is bounded, which allows to uniformly upper-bound the supremum norm of the unit ball on H.

Remark 8 *If H is a RKHS with kernel K such that $K(\mathbf{x}, \mathbf{x}) \leq B$ for some $B > 0$, then it holds*

$$\sup_{||f||_H \leq 1} ||f||_\infty \leq B < \infty.$$

4. Boosting DAGs

In this section, we prove Theorem 15, which is our main result. It states that if ~~we choose~~ the regression procedure in Section 2 is chosen as L_2 boosting with early stopping, then the estimator for the topological ordering is consistent. This holds for a uniform and asymptotically increasing number of boosting iterations. We provide the assumptions in Section 4.1 and prove the statement in Section 4.2. In Section 4.3 we propose an adaption of the procedure ~~which~~ that is effective in high dimensions.

4.1 Assumptions

Proposition 5 ~~has shown~~shows that in order to consistently estimate the causal ordering, we ~~have to~~must avoid overfitting whenever we regress X_k onto $\mathbf{X}_S$ for any $k = 1, \ldots, p$ and $S \subset \{1, \ldots, p\} \setminus \{k\}$. This poses the main challenge in applying Proposition 5. We assume that $X_k - \mathbb{E}[X_k | \mathbf{X_S} = \mathbf{x_S}]$ is sub-Gaussian, which later allows one to control the regression estimates.

Definition 9 *We call a random variable ε sub-Gaussian if its Orlicz norm defined by*

$$s(\varepsilon) = \inf \left\{ r \in (0, \infty) : \mathbb{E}\left[\exp\left(\frac{\varepsilon^2}{r^2} \right) \right] \leq 2 \right\}$$

is finite.

Assumption 10 *For any k and $S \subset \{1, \ldots, p\} \setminus \{k\}$ consider the decomposition*

$$X_k = \mu_{k,S}(\mathbf{X}_S) + \varepsilon_{k,S},$$

where

$$\mu_{k,S}(\mathbf{X}_S) = \mathbb{E}[X_k | \mathbf{X}_S] \text{ and } \varepsilon_{k,S} = X_k - \mathbb{E}[X_k | \mathbf{X}_S].$$

We assume that

1. *$\varepsilon_{k,S} | \mathbf{X}_S = \mathbf{x}_S$ is sub-Gaussian with Orlicz norm $s_{k,S}(\mathbf{x}_S)$, $0 < s_{k,S}(\mathbf{x}_S) \leq s_{max}$ for all $\mathbf{x}_S \in \mathbb{R}^{|S|}$, and*
2. *$||\mu_{k,S}||_\infty \leq \mu_{max} < \infty$.*

The constants ~~shall~~are assumed to hold uniformly for ~~any~~all $k \in \{1, \ldots, p\}$ and all $S \subset \{1, \ldots, p\} \setminus \{k\}$.

We will prove that the regression estimate will lie in the function class $h\mathcal{F}_N$ for some radius $h > 0$. We denote the ball of radius h in H by B_h. In Theorem 15 we use the radius $h > 0$ and the function complexity measures Rademacher complexity and covering numbers to derive Condition 1. of Proposition 5. Both measures quantify the richness of a function class. While the Rademacher complexity of $\mathcal{F}_N$ can be upper-bounded by Theorem 39 given in the Appendix, we also need to upper bound the covering number of $\mathcal{F}_N$. The complexity measures for $h\mathcal{F}_N$ can then be upper-bounded using scaling arguments. We thus make the following assumption:

Assumption 11 *For any $k = 1, \dots, p$, let H_k be a RKHS on X_k and $B_1^k := \{f \in H : ||f||_{H_k} \leq 1\}$. Then it shall hold for any $z > 0$ and $k = 1, \dots, p$*

$$\int_0^1 \sqrt{\log\left(\mathcal{N}\left(\frac{uz}{2}, B_1^k\right)\right)} du < \infty.$$

Here, $\mathcal{N}(\cdot, B_1^k)$ is the covering number with respect to $||\cdot||_\infty$, which is defined in Section B.2.

Remark 12 *Let $H_S = H_{s_1} \oplus \dots \oplus H_{s_k}$ and denote the unit ball of H_{s_k} by $B_1^{s_k}$ and the unit ball of H_S by B_1^S. Assume that for any $j = 1, \dots, k$ it holds that*

$$\int_0^1 \sqrt{\log\left(\mathcal{N}\left(\frac{uz}{2}, B_1^j\right)\right)} du \leq C(z) < \infty$$

for some $0 < C(z) < \infty$. Furthermore, one can derive that

$$\mathcal{N}\left(u, B_1^S\right) \leq \prod_{j=1}^k \mathcal{N}\left(\frac{u}{k}, B_1^{s_j}\right).$$

Thus, using Jensen's inequality and the fact that $\mathcal{N}(\cdot, B_1^j)$ is non-increasing, it holds that

$$\int_0^1 \sqrt{\log\left(\mathcal{N}\left(\frac{uz}{2}, B_1^S\right)\right)} du \leq \int_0^1 \sqrt{\sum_{j=1}^k \log\left(\mathcal{N}\left(\frac{uz}{2k}, B_1^{s_j}\right)\right)} du \leq \sqrt{p} C\left(\frac{z}{2p}\right) < \infty.$$

We assume that the eigenvalues of the random Gram matrix G ~~vanish~~decay with the same rate as the eigenvalues of the operator $\mathcal{K}$. For details on the connection between the eigenvalues of G and $\mathcal{K}$, consult Section C of Bach and Jordan (2002).

Assumption 13 *For $S = \{s_1, \dots, s_k\} \subset \{1, \dots, p\}$ let $H = H_{s_1} \oplus \dots \oplus H_{s_k}$ be the RKHS on $\mathbf{X}_S$ generated by the additive kernel $K_{s_1} + \dots + K_{s_k}$. Assume there~~There shall~~ exist events $\mathcal{B}_N$ with $\lim_{N\to\infty} \mathbb{P}(\mathcal{B}_N) = 1$, such ~~so~~ that on $\mathcal{B}_N$, and for $K_0 = \lfloor \frac{1}{2C_d+1} \ln(N) \rfloor$ ~~it holds for~~ the empirical eigenvalues of the Gram matrix G satifsy*

$$\sum_{k=K_0}^N \widehat{\mu_k} \leq \sum_{k=K_0}^N \exp(-C_u k),$$

and additionally

$$\widehat{\mu}_{K_0} \geq \exp(-C_d K_0)$$

for some $C_d > C_u > 0$. The constants hold uniformly for any $S \subset \{1, \dots, p\}$.

Remark 14 *Bach and Jordan (2002, Section C) investigate the eigenvalue decay of the Gram matrix for different tails of the density on $\mathbf{X}_S$. They show that for densities on $\mathbf{X}_S$ with tails in the order of $\exp\left(-\frac{\mathbf{x}_S^T \mathbf{x}_S}{2}\right)$, the eigenvalues of the Gram matrix of a Gaussian kernel decay as in Assumption 13. Furthermore, they show that for the first $k \leq K_0(N)$ eigenvalues μ_k is close to $\exp(-Ak)$, where A depends on the density of $\mathbf{X}_S$ and $K_0(N)$ is in the order of $\log(N)$. This shows that Assumption 13 is fulfilled for ~~appropriate tails of the density on~~ densities $\mathbf{X}_S$ with appropriately decaying tails. An example of such a density decay can be observed when $\mathbf{X}_S$ is Gaussian distributed and the function $f : \mathbb{R} \to \mathbb{R}$ approximately satisfies $f(x) \approx cx$ for $|x| \to \infty$, for some $c \in \mathbb{R}$.*

To give some intuition, recall that for the kernel regression estimator $S : y^N \mapsto \widehat{y}^N$, it holds that $S = UDU^\top$, where U contains the eigenvectors of G, and D is a diagonal matrix with entries $d_\ell = \frac{\widehat{\mu}_\ell}{\lambda+\widehat{\mu}_\ell}$. Clearly, d_ℓ ~~has a decay~~decays at a rate similar to $\widehat{\mu}_\ell$. For simplicity, consider $w^N = U^\top y^N$. Assumption 13 then ensures that only a few entries of w^N largely influence $\widehat{y}^N = UDw^N$, while most entries of w^N contribute little. Thus, $\widehat{y}^N$ is mainly influenced by a~~small~~ low-dimensional subspace of $\mathbb{R}^N$.

4.2 Main Theorem

We now state ~~now~~ the main theorem.

Theorem 15 *Let $H_{k'}$ be a RKHS on $X_{k'}$ with Gaussian kernel $K_{k'}$ for ~~any~~each $k' = 1, \ldots, p$. Assume that $\mathbf{X} = (X_1, \ldots, X_p)$ follows a CAM as described in Assumption 1 for functions $f_1^0 \in H_1, \ldots, f_p^0 \in H_p$, where $H_k = H_{k_1} \oplus \ldots \oplus H_{k_q}$, and $\{k_1, \ldots, k_q\} = \mathbf{pa}(k)$. Given Assumptions 10, 11, and 13 and assuming we estimate $\widehat{f}_{k,\pi}^{(m_{stop})} = \widehat{f}_{k,\pi} = \sum_{j \in \varpi_\pi(k)} \widehat{f}_{kj,\pi}$ using L_2 boosting with X_k as response, $\mathbf{X}_{\varpi_\pi(k)}$ as predictors, and ~~with~~the number of boosting steps chosen as $m_{stop} = N^{\frac{1}{4}\frac{C_u + C_d + 1/2}{C_d + 1}}$, ~~then~~ it holds that*

$$\widehat{S}(\pi^0) < \widehat{S}(\pi)$$

for $N \to \infty$ with probability going to 1 for any $\pi^0 \in \Pi^0$ and $\pi \notin \Pi^0$.

Remark 16 *Theorem 2 of Minh (2010) ensures that $f_1^0, \ldots, f_p^0$ are smooth, non-constant, and non-linear, and ~~thus meet~~therefore satisfy Assumption 1.*

Remark 17 *If the eigenvalues of the Gram matrix decay rapidly, that is, if $C_u \approx C_d$ and C_d is large, then a low-dimensional functional subspace contains a large fraction of the variation of f. This reduces the risk of overfitting. Hence, the number of boosting iterations $m(N)$ can be chosen in the order $N^{\frac{1}{2}}$. On the other hand, if the decay of the eigenvalues is slow, then the risk of overfitting is larger. Thus, $m(N)$ should be chosen in the order $N^{\frac{1}{4}}$.*

To this end, we ~~would like to~~ compare our assumptions required for Theorem 15 with the assumptions of Theorem 1 in Bühlmann et al. (2014). Most importantly, Bühlmann et al. (2014) assume that the number of basis functions grows "sufficiently slowly" and that the choice of basis functions "is deterministic and does not depend on the data". In addition, Bühlmann et al. (2014) require that this choice of functions can approximate the true functions f_k^0 on any compact set in terms of the L^2-norm, see their Assumption (A4); and that the search space of additive functions is closed (Lemma 2 of Bühlmann et al., 2014) in L^2. ~~Conversely~~In contrast, our method is considerably more ~~applicable~~flexible since the regression estimates are directly dependent on the data, eliminating the need to manually restrict the function space's dimensionality.

However, our Assumptions 10 and 11 are stricter than Assumption (A2) in Bühlmann et al. (2014). Specifically, we ~~need to further restrict the~~impose a stronger moment condition ~~of~~than (A2)(ii) of Bühlmann et al. (2014) and require conditional distributions in Assumption 10. Furthermore, Assumption 11 in this work ~~limits~~bounds the covering number entropy integral in supremum norm, while Assumption (A2)(i) of Bühlmann et al. (2014) ~~limits the covering number entropy integral~~does so in L^2-norm, which is a weaker condition.

Finally, Assumption 13 depends on the distribution of $\mathbf{X}_S$ and is therefore related to Assumption (A3) of Bühlmann et al. (2014) although ~~they~~the two are difficult to compare.

Proof We apply Proposition 5 and show the following two conditions.

1. *Boosting under Misspecification:* $\widehat{f}_{k,S}^{(m_{stop})}$ is not overfitting with respect to $(\mathbf{X}_S, X_k)$ for any $k \in \{1, \ldots, p\}$ and $S \subset \{1, \ldots, p\} \setminus \{k\}$, and
2. *Consistency of Variance Estimation:* For any $k = 1, \ldots, p$ and $\pi^0 \in \Pi^0$ it holds

$$\left| \frac{1}{N} \sum_{\ell=1}^{N} \left(\mathbf{x}_{\ell k} - \widehat{f}^{(m_{stop})}(\mathbf{x}_{\ell \varpi_{\pi^0}(k)} \right)^2 - \mathbb{E}\left[\left(X_k - f_k^0(\mathbf{X}_{\mathbf{pa}_{G^0}(k)} \right)^2 \right] \right| \to 0$$

in probability.

We will see that 1. follows from Theorem 18 in Section 4.2.1 and 2. is shown in Theorem 19 in Section 4.2.2. ■

4.2.1 Boosting under Misspecification

In Theorem 18 below, we show Condition 1. of Theorem 15. This is a novel result that specifies conditions so that L_2 boosting is not overfitting even in a misspecified scenario. We fix $S = \{s_1, \ldots, s_d\}$ and k and define $\left(\widetilde{X}_1, \ldots, \widetilde{X}_d\right) = \widetilde{\mathbf{X}} = \mathbf{X}_S$ and $Y = X_k$. Let $\widetilde{\mathbf{X}}^N$ be the random element ~~containing~~consisting of N i.i.d. observations of $\widetilde{\mathbf{X}}$ and denote the realizations of $\widetilde{\mathbf{X}}^N$ by $\widetilde{\mathbf{x}}^N = (\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_N)$. Analogously, we define Y^N and y^N. Our goal is to prove~~,~~ that

$$\left| \frac{1}{N} \sum_{\ell=1}^{N} \left(\widehat{f}^{(m_{stop})}(\widetilde{\mathbf{x}}_\ell) - y_\ell \right)^2 - \mathbb{E}_{\widetilde{\mathbf{X}},Y} \left[\left(\widehat{f}^{(m_{stop})}(\widetilde{\mathbf{X}}) - Y \right)^2 \right] \right| \xrightarrow{\mathbb{P}} 0 \tag{16}$$

for $N \to \infty$ for the boosting estimate $\widehat{f}^{(m_{stop})}$. The left term is an expectation with respect to the empirical distribution P_N (and thus depends on the realizations), whereas the right term is under the population distribution induced by $\left(\widetilde{\mathbf{X}}, Y\right)$ denoted by P. Thus, the l.h.s. of Equation (16) ~~becomes~~can be written as

$$(P_N - P) \left[\left(\widehat{f}^{(m_{stop})}(\widetilde{\mathbf{X}}) - Y \right)^2 \right]. \tag{17}$$

Assumptions 10′, 11′, and 13′ are reformulated versions of Assumptions 10, 11, and 13, respectively, using the notation introduced above.

Assumption 10′ *Let $Y = \mu(\widetilde{\mathbf{X}}) + \varepsilon$, for which we define the random variables*

$$\mu(\widetilde{\mathbf{X}}) = \mathbb{E}\left[Y | \widetilde{\mathbf{X}} \right] \text{ and } \varepsilon = Y - \mathbb{E}\left[Y | \widetilde{\mathbf{X}} \right].$$

We assume that $\varepsilon | \widetilde{\mathbf{X}} = \widetilde{\mathbf{x}}$ is sub-Gaussian with Orlicz norm $s(\widetilde{\mathbf{x}})$ and $0 < s(\widetilde{\mathbf{x}}) \leq s_{max}$ for all $\widetilde{\mathbf{x}} \in \mathbb{R}^d$ and $||\mu||_\infty \leq \mu_{max} < \infty$. Let $\sigma^2_{max} := \max_{\widetilde{\mathbf{x}} \in \mathbb{R}^d} \mathbb{E}\left[\varepsilon^2 | \widetilde{\mathbf{X}} = \widetilde{\mathbf{x}} \right]$.

Assumption 11′ *Let* $B_1 := \{f \in \mathcal{F} : ||f||_H \leq 1\}$, *where* H *is the additive RKHS on* $\widetilde{\mathbf{X}} = \left(\widetilde{X}_1, \ldots, \widetilde{X}_d\right)$. *Then, for any* $z > 0$ *it holds*

$$\int_0^1 \sqrt{\log\left(\mathcal{N}\left(\frac{uz}{2}, B_1\right)\right)} du < \infty.$$

Assumption 13′ *There exist events* $\mathcal{B}_N$ *on* $\widetilde{\mathbf{X}}^N$ *with* $\lim_{N\to\infty} \mathbb{P}(\mathcal{B}_N) = 1$, *so that on* $\mathcal{B}_N$ *and for* $K_0 = \lfloor \frac{1}{2C_d+1} \ln(N) \rfloor$ *it holds for the empirical eigenvalues of the Gram matrix* G ~~*for*~~*computed from* $\widetilde{\mathbf{x}}^N \in \mathcal{B}_N$

$$\sum_{k=K_0}^{N} \widehat{\mu_k} \leq \sum_{k=K_0}^{N} \exp(-C_u k),$$

and additionally

$$\widehat{\mu}_{K_0} \geq \exp(-C_d K_0)$$

for some $C_d > C_u > 0$.

Note that all constants in Assumptions 10′, 11′, 13′ are independent of the choice of k, S. It is the purpose of this subsection to prove the following theorem (and thus Condition 1).

Theorem 18 *Under the Assumptions 10′, 11′ and 13′ it holds for* $m_{stop}(N) = N^{\frac{1}{4}\frac{C_u+C_d+1/2}{C_d+1}}$

$$\left|(P - P_N)\left(Y - \widehat{f}^{(m_{stop})}(\widetilde{\mathbf{X}})\right)^2\right| \xrightarrow{\mathbb{P}} 0. \tag{18}$$

As $\widehat{f}^{(m_{stop})}$ depends on the realizations $\widetilde{\mathbf{x}}^N$ and y^N, the convergence above is not trivial. For simplicity, we drop the dependency of $f, \widehat{f}^{(m_{stop})}$ on $\widetilde{\mathbf{X}}$ in the proof below.

Proof To prove (18), ~~W~~we decompose

$$\begin{aligned}
&|(P_N - P)\left(Y - \widehat{f}^{(m_{stop})}\right)^2| \\
&\leq \underbrace{|(P_N - P)Y^2|}_{\text{I}} + \underbrace{2|(P_N - P)\widehat{f}^{(m_{stop})}Y|}_{\text{II}} + \underbrace{|(P_N - P)\left(\widehat{f}^{(m_{stop})}\right)^2|}_{\text{III}},
\end{aligned}$$

and show the convergence in probability for I–III separately. Term I goes to 0 in probability since Y has a finite fourth moment. For term II it holds for any $\xi > 0$ that

$$\begin{aligned}
&\mathbb{P}\left(|(P_N - P)Y\widehat{f}^{(m_{stop})}| \geq \xi\right) \\
&= \mathbb{P}\left(\left(|(P_N - P)Y\widehat{f}^{(m_{stop})}| \geq \xi\right) \cap \left(||\widehat{f}^{(m_{stop})}||_H \in h(N)\mathcal{F}_N\right)\right) \\
&\quad + \mathbb{P}\left(\left(|(P_N - P)Y\widehat{f}^{(m_{stop})}| \geq \xi\right) \cap \left(||\widehat{f}^{(m_{stop})}||_H \notin h(N)\mathcal{F}_N\right) \cap \left\{\widetilde{\mathbf{X}}^N \in \mathcal{B}_N\right\}\right) \\
&\quad + \mathbb{P}\left(\left(|(P_N - P)Y\widehat{f}^{(m_{stop})}| \geq \xi\right) \cap \left(||\widehat{f}^{(m_{stop})}||_H \notin h(N)\mathcal{F}_N\right) \cap \left\{\widetilde{\mathbf{X}}^N \notin \mathcal{B}_N\right\}\right) \\
&\leq \mathbb{P}\left(\left(|(P_N - P)Y\widehat{f}^{(m_{stop})}| \geq \xi\right) \cap \left(||\widehat{f}^{(m_{stop})}||_H \in h(N)\mathcal{F}_N\right)\right) \\
&\quad + \mathbb{P}\left(\left(||\widehat{f}^{(m_{stop})}||_H \notin h(N)\mathcal{F}_N\right) \cap \left\{\widetilde{\mathbf{X}}^N \in \mathcal{B}_N\right\}\right) \\
&\quad + \mathbb{P}\left(\left\{\widetilde{\mathbf{X}}^N \notin \mathcal{B}_N\right\}\right) \\
&\leq \mathbb{P}\left(\sup_{f \in h(N)\mathcal{F}_N} |(P_N - P)Yf| \geq \xi\right) \\
&\quad + \mathbb{P}\left(\left(||\widehat{f}^{(m_{stop})}||_H \notin h(N)\mathcal{F}_N\right) \cap \left\{\widetilde{\mathbf{X}}^N \in \mathcal{B}_N\right\}\right) \\
&\quad + \mathbb{P}\left(\left\{\widetilde{\mathbf{X}}^N \notin \mathcal{B}_N\right\}\right).
\end{aligned}$$

For $h(N) \in o\left(N^{1/4}\right)$, the first line on the r.h.s. goes to 0 by Corollary 34 and the second line vanishes by Lemma 27. The third line converges to 0 due to Assumption 13′. This shows the convergence in probability of term II.
Similarly, for term III it holds for any $\xi > 0$ that

$$\begin{aligned}
&\mathbb{P}\left(|(P_N - P)\left(\widehat{f}^{(m_{stop})}\right)^2| \geq \xi\right) \\
&= \mathbb{P}\left(\left(|(P_N - P)\left(\widehat{f}^{(m_{stop})}\right)^2| \geq \xi\right) \cap \left(||\widehat{f}^{(m_{stop})}||_H \in h(N)\mathcal{F}_N\right)\right) \\
&\quad + \mathbb{P}\left(\left(|(P_N - P)\left(\widehat{f}^{(m_{stop})}\right)^2| \geq \xi\right) \cap \left(||\widehat{f}^{(m_{stop})}||_H \notin h(N)\mathcal{F}_N\right) \cap \{\widetilde{\mathbf{X}}^N \in \mathcal{B}_N\}\right) \\
&\quad + \mathbb{P}\left(\left(|(P_N - P)\left(\widehat{f}^{(m_{stop})}\right)^2| \geq \xi\right) \cap \left(||\widehat{f}^{(m_{stop})}||_H \notin h(N)\mathcal{F}_N\right) \cap \{\widetilde{\mathbf{X}}^N \notin \mathcal{B}_N\}\right) \\
&\leq \mathbb{P}\left(\sup_{f \in h(N)\mathcal{F}_N} |(P_N - P)f^2| \geq \xi\right) \\
&\quad + \mathbb{P}\left(\left(||\widehat{f}^{(m_{stop})}||_H \notin h(N)\mathcal{F}_N\right) \cap \{\widetilde{\mathbf{X}}^N \in \mathcal{B}_N\}\right) \\
&\quad + \mathbb{P}\left(\{\widetilde{\mathbf{X}}^N \notin \mathcal{B}_N\}\right).
\end{aligned}$$

For $h(N) \in o\left(N^{1/4}\right)$, the first line on the r.h.s. converges to 0 due to Corollary 40 and the other terms behave as described for term II. This shows the convergence in probability for term III.

Overall, this proves Condition 1. ■

4.2.2 Consistency of Variance Estimation

In Theorem 19 below, we proof Condition 2. of Theorem 15 using the techniques in Bühlmann and Yu (2003); Raskutti et al. (2014). We fix again k and assume that $\varpi_{\pi^0}(k)$ has size d, that is, $\pi^0(k) = d+1$. We set $\widetilde{\mathbf{X}} = \left(\widetilde{X}_1, \ldots, \widetilde{X}_d\right) = \mathbf{X}_{\varpi_{\pi^0}(k)}$ and $Y = X_k$.

Theorem 19 *Let*

$$Y = f^0(\widetilde{\mathbf{X}}) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2),$$

where f^0 lies in a RKHS H for which Assumption 13′ holds with $||f^0||_H = R$. Then, it holds with the number of boosting steps chosen as $m_{stop} = m(N) = N^{\frac{1}{4}\frac{C_u + C_d + 1/2}{C_d + 1}}$ for $N \to \infty$

$$\left| \frac{1}{N} \sum_{\ell=1}^{N} (y_\ell - \widehat{f}^{(m_{stop})}(\widetilde{\mathbf{x}}_\ell))^2 - \mathbb{E}\left[\left(Y - f^0(\widetilde{\mathbf{X}}) \right)^2 \right] \right| = \left| \widehat{\sigma}^2 - \sigma^2 \right| \to 0$$

in probability.

Proof We define the semi-norm

$$||g||_{2,N}^2 := \frac{1}{N} \sum_{\ell=1}^{N} g(y_\ell, \widetilde{\mathbf{x}}_\ell)^2.$$

By the triangle inequality it holds that

$$||Y - \widehat{f}^{(m_{stop})}||_{2,N} \leq ||Y - f^0||_{2,N} + ||f^0 - \widehat{f}^{(m_{stop})}||_{2,N}. \tag{19}$$

Next, we show how to asymptotically lower and upper bound $||Y - \widehat{f}^{(m_{stop})}||_{2,N}$ by σ.

Lower bound: As $||f^0||_H = R$, $||f^0||_\infty$ is bounded by Remark 8~~and as $||f^0||_H = R$~~. Further, $f^0(\widetilde{\mathbf{x}}) = \mathbb{E}\left[Y | \widetilde{\mathbf{X}} = \widetilde{\mathbf{x}} \right]$ and the noise $Y - f^0(\widetilde{\mathbf{x}}) = \varepsilon$ is Gaussian and thus sub-Gaussian with a uniformly bounded Orlicz norm ~~that is uniformly bounded~~. Hence, by Theorem 18 and the continuous mapping theorem the l.h.s. of (19) converges and it holds

$$||Y - \widehat{f}^{(m_{stop})}||_{2,N} = \sqrt{||Y - \widehat{f}^{(m_{stop})}||_{2,N}^2} \to \sqrt{\mathbb{E}\left[||Y - \widehat{f}^{(m_{stop})}||_2^2 \right]} \geq \sqrt{\mathbb{E}\left[||Y - f^0||_2^2 \right]} = \sigma$$

in probability.

Upper bound: The term $||Y - f^0||_{2,N} = \sqrt{\frac{1}{N} \sum_{\ell=1}^{N} \varepsilon_\ell^2}$ converges to σ in probability. Thus, it remains to show that $||f^0 - \widehat{f}^{(m_{stop})}||_{2,N} \xrightarrow{\mathbb{P}} 0$ for $N \to \infty$, which follows from Lemma 20 below. ■

Lemma 20 *For $Y = f^0(\widetilde{\mathbf{X}}) + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, $f^0 \in H$, and if $\widehat{f}$ is the boosting estimate with $m_{stop} = m(N) = N^{\frac{1}{4}\frac{C_u + C_d + 1/2}{C_d + 1}}$ boosting steps, ~~then~~ $\widehat{f}^{(m_{stop})}$ converges to f^0 in a fixed design, that is,*

$$||f^0 - \widehat{f}^{(m_{stop})}||_{2,N} = \left(\frac{1}{N} \sum_{\ell=1}^{N} \left(f^0(\widetilde{\mathbf{x}}_\ell) - \widehat{f}^{(m_{stop})}(\widetilde{\mathbf{x}}_\ell) \right)^2 \right)^{1/2} \xrightarrow{\mathbb{P}} 0.$$

Proof The proof is in the Appendix C. ■

Remark 21 *Lemma 20 also holds for heteroscedastic and sub-Gaussian noise.*

Remark 22 *The combination of Theorem 18 and 19 is insightful. If we are unsure whether $f^0 \in H$ and the noise is independent, then limiting appropriately the number of boosting iterations leads to:*

1. *a consistent estimator for f^0 if the assumptions hold, and*
2. *an estimator, ~~so~~such that the prediction error on the samples $\frac{1}{N} \sum_{\ell=1}^{N} \left(y_\ell - \widehat{f}^{(m_{stop})}(\widetilde{\mathbf{x}}_\ell) \right)^2$ is asymptotically close to the L^2-error $\mathbb{E}\left[\left(Y - \widehat{f}^{(m_{stop})}(\widetilde{\mathbf{X}}) \right)^2 \right]$ for yet unobserved realizations of $(\widetilde{\mathbf{X}}, Y)$. Here, $\frac{1}{N} \sum_{\ell=1}^{N} \left(y_\ell - \widehat{f}^{(m_{stop})}(\widetilde{\mathbf{x}}_\ell) \right)^2$ depends only on the ~~observations used for learning~~training data used to learn $\widehat{f}^{(m_{stop})}$ and thus no hold-out set is necessary.*

We emphasize that although the results are stated for the Gaussian kernel they can be ~~adapted~~extended to kernels with other eigenvalue decay rates.

4.3 Boosting DAGs for Large Dimensions

Theorem 15 shows that we can asymptotically identify the true causal ordering using boosting regressions. However, there are $p!$ possible permutations on $\{1, \ldots, p\}$ that constitute the search space. Thus, beyond a very small p or without extensive prior knowledge on the topological order the computational costs are prohibitive. We address this issue through component-wise boosting in an additive noise model.

Additive Noise Model The true graph G^0 and the true additive structural equations $f_1^0, \ldots, f_p^0$ with

$$f_k^0(\mathbf{x}_{\mathbf{pa}(k)}) = \sum_{j \in \mathbf{pa}(k)} f_{kj}^0(x_j)$$

can be represented by a function $F^0 : \mathbb{R}^p \to \mathbb{R}^p$, where

$$F^0(\mathbf{x}) = \left(f_1^0(\mathbf{x}), \ldots, f_p^0(\mathbf{x}) \right)^\top = \left(f_1^0(\mathbf{x}_{\mathbf{pa}(1)}), \ldots, f_p^0(\mathbf{x}_{\mathbf{pa}(p)}) \right)^\top.$$

Thus, the graph G^0 has an edge from X_j to X_k if and only if the function f_k^0 is not constant in its $j-$th component. Note that the set of functions $F : \mathbb{R}^p \to \mathbb{R}^p$ corresponding to a DAG is non-convex (Zheng et al., 2020).

Component-wise Boosting Instead of applying L_2 boosting to estimate f_k^0, $k = 1, \ldots, p$ one-by-one as in Theorem 15, we employ component-wise boosting to estimate F^0. This means, we define the loss function on the functions $F : \mathbb{R}^p \to \mathbb{R}^p$ as the log-likelihood function given $\mathbf{x}^N$

$$L(F, \mathbf{x}^N) = L\left((f_1, \ldots, f_p), \mathbf{x}^N\right) = \sum_{k=1}^{p} \log\left(\sum_{\ell=1}^{N} (\mathbf{x}_{\ell k} - f_k(\mathbf{x}_\ell))^2\right)$$

and proceed as follows. Choose a step size $0 < \nu \leq 1$ and let $F^{(1)} = 0$ be the starting value. Then, for $m = 1, 2, \ldots$ we set $f_k^{(m+1)}(\mathbf{x}) = f_k^{(m)}(\mathbf{x}) + \nu \widehat{f}_{kj}(x_j)$, where $(j, k) \in \{1, \ldots, p\} \times \{1, \ldots, p\}$ is the solution of

$$\underset{(j,k) \notin N^m}{\arg\min}\, S(j, k; F^{(m)}),$$

and $S(j, k; F^{(m)})$ is the score on the edges defined by

$$S(j, k; F^{(m)}) := \log\left(\sum_{\ell=1}^{N} \left(\widehat{f}_{kj}(\mathbf{x}_{\ell j}) - \left(\mathbf{x}_{\ell k} - f_k^{(m)}(\mathbf{x}_\ell)\right)\right)^2\right). \tag{20}$$

Here, the candidate functions $\widehat{f}_{kj}$ are determined by solving the kernel ridge regression

$$\widehat{f}_{kj} = \underset{g_{kj} \in H_j}{\arg\min} \sum_{\ell=1}^{N} \left(g_{kj}(\mathbf{x}_{\ell j}) - \left(\mathbf{x}_{\ell k} - f_k^{(m)}(\mathbf{x}_\ell)\right)\right)^2 + \lambda ||g_{kj}||_{H_j}^2. \tag{21}$$

That is, we run a RKHS regression on the marginal residuals. Hence, the resulting functions $f_{kj}, j, k = 1, \ldots, p$ belong to an RKHS.

In the set N^m we track the edges (j, k) that would cause a cycle when added to $F^{(m)}$. Hence, $F^{(m)}$ corresponds to a DAG for any $m = 1, 2, \ldots$.

Note that if the edge (j, k) is chosen, then we only need to update $S(j', k; F^{(m+1)})$ for $(j', k) \notin N^{m+1}$, while this score remains unchanged for $k' \neq k$, that is, $S(j, k'; F^{(m+1)}) = S(j, k'; F^{(m)})$. This reduces the computational burden. We stop the procedure after m_{stop} steps, which is the crucial tuning parameter of (component-wise) boosting. As the RKHS regression scales poorly with N, it could be approximated by some spline regression instead.

Choosing m_{stop} Inspired by results from boosting for regression (Tutz and Binder, 2006; Bühlmann and Hothorn, 2007) we use the Akaike Information Criterion (AIC) to select m_{stop}. For any $f_k^{(m)}$ we calculate the trace of the mapping $B_k^{(m)} : (x_{1k}, \ldots, x_{Nk}) \mapsto \left(f_k^{(m)}(\mathbf{x}_1), \ldots, f_k^{(m)}(\mathbf{x}_N)\right)$. Then we define the AIC score by

$$AIC\left(F^{(m)}, \mathbf{x}^N\right) = \sum_{k=1}^{p} AIC_k\left(f_k^{(m)}, \mathbf{x}^N\right) = \sum_{k=1}^{p} \left(\sum_{\ell=1}^{N} \left(x_{\ell k} - f_k^{(m)}(\mathbf{x}_\ell)\right)^2 + tr\left(B_k^{(m)}\right)\right). \tag{22}$$

We stop the procedure and set $m_{stop} = m$ if $AIC(F^{(m)}, \mathbf{x}^N)$ increases with m. We emphasize that this ~~is merely~~corresponds only to a local minimum w.r.t the AIC score and the global

optimum is hard to find due to the non-convexity of the search space. The algorithm using the AIC is outlined in Algorithm 1 in Appendix D. It can be understood as a component-wise functional gradient descent in the space of those additive functions $\mathbb{R}^p \to \mathbb{R}^p$ that imply a DAG.

Pruning We prune the estimated graph $\widehat{G}$ by ~~running~~performing an additive model regression of ~~every~~each node on its parents. Finally, we keep only those nodes as parents whose p-value is below 0.001. For more details on pruning, see Bühlmann et al. (2014).

5. Simulation Study

In this section we empirically investigate the proposed algorithms. To this end, we first describe the data-generating processes in Section 5.1. In Section 5.2 we verify Theorem 15 for data sets of small dimensions. In Section 5.3 we benchmark our algorithm of Section 4.3 denoted by DAGBoost on high-dimensional data sets against state-of-the-art methods, and conduct a sensitivity analysis on the effect of the step size ν, penalty parameter λ, and the stopping criterion m_{stop}.

Every result presented is based on 100 randomly generated data sets unless otherwise stated. Throughout, we set the step size $\nu = 0.3$ and the penalty parameter $\lambda = 0.01$. While it is known that boosting is commonly robust with respect to the step size (as long as it is small enough), we find in Section 5.3.2, that our method DAGBoost is also robust against the specific choice of λ. We therefore refrain from further tuning. The code and the data-generation procedure are publicly available at `https://github.com/mkrtl/BoostingDAGs`.

5.1 Sampling SEMs and Data Generation

The generation of synthetic data for causal discovery can lead to data sets leaking information on the underlying graph. For example, Reisach et al. (2021) show that for many simulated data sets, the causal ordering can be obtained by the marginal variance of the nodes. Further, the signal-to-noise ratio for downstream nodes typically increases (Agrawal et al., 2023). To address these issues, we adapt the generation of synthetic data sets of Bühlmann et al. (2014); Lachapelle et al. (2019); Zheng et al. (2020); Ng et al. (2022b) to our aims. To this end, we apply a normalization step as described in Agrawal et al. (2023) to generate data sets as follows.

1. Generate **underlying graph** G^0 with one of the following two methods.
   (a) Generate a DAG according to the Erdös-Renyi (ER) model (Erdös and Rényi, 1959) as follows. We first generate a random DAG with the maximal number of edges and keep every edge with a constant probability. Every node has the same distribution for the number of its neighbors.
   (b) Generate a scale-free (SF) graph using the model of Barabási and Albert (1999). There exist hubs of nodes with large degree, while other nodes have a smaller degree. This graph structure is observed in various applications (Jeong et al., 2000; Wille et al., 2004; Kertel et al., 2023).
2. Generate additive or non-additive **structural equations** (SEs).

(a) **Additive:** For any edge (j,k) in the graph, sample f^0_{kj} from a Gaussian process with zero mean and covariance function $cov(f^0(x_{\ell j}), f^0(x_{\ell' j})) = \exp\left(-\frac{(x_{\ell j}-x_{\ell' j})^2}{2}\right)$, and set $f^0_k(\mathbf{x}_{\ell \mathbf{pa}_{G^0}(k)}) = \sum_{j\in\mathbf{pa}_{G^0}(k)} f^0_{kj}(x_{\ell j})$ for $\ell, \ell' = 1, \ldots, N$.

(b) **Non-additive:** For every node k consider its parents $\mathbf{pa}_{G^0}(k)$, and sample f^0_k from a Gaussian Process with zero mean and covariance function

$$cov(f^0(x_{\ell \mathbf{pa}_{G^0}(k)}), f^0(x_{\ell' \mathbf{pa}_{G^0}(k)})) = \exp\left(-\frac{||x_{\ell \mathbf{pa}_{G^0}(k)} - x_{\ell' \mathbf{pa}_{G^0}(k)}||_2^2}{2}\right),$$

for $\ell, \ell' = 1, \ldots, N$.

3. **Normalization:** Scale $f^0_k(\mathbf{X}_{\ell \mathbf{pa}_{G^0}(k)})$, such that

$$\mathbb{E}\left[f^0_k(\mathbf{X}_{\mathbf{pa}_{G^0}(k)})^2\right] = 1.$$

4. Generate the **standard deviations** σ^0_k for all $k = 1, \ldots, p$ from a uniform distribution on $\left[\sqrt{2}/5, \sqrt{2}\right]$.

Finally, we generate the variables with no incoming edges from a centered Gaussian distribution with standard deviation as chosen in Step 4. Following the topological order of G^0, we generate the data sets recursively according to the SEM. Since the scaling constant for the normalization step is difficult to attain, we generate 1000 observations that are used for the normalization step only. For a more detailed description, see Agrawal et al. (2023, Appendix E).

We emphasize, that non-additive SEs conflict with Assumption 1. Moreover, neither the signal-to-noise ratio nor the variance of the nodes do not reveal the causal order.

5.2 Low-Dimensional Data

In this low-dimensional simulation study we generate data by setting $p = 5$ from an ER graph with on average five edges. Then, we calculate the score for any of the $p!$ permutations π as described in Section 2. We choose the number of boosting iterations by applying the regression AIC score of Bühlmann and Hothorn (2007). While the AIC score of Section 4.3 penalizes the complexity of the full SEM, this score only penalizes the complexity of one regression. Recall that $f^{(m)}_k$ is a function on $\mathbf{x}_{\varpi_\pi(k)}$ and we use regular (not component-wise) boosting.

We generate data sets with $N = 10, 20, 50, 100, 200$ observations with either additive or non-additive SEs. To evaluate the estimated permutations, we use the transposition distance

$$d_{trans}(\pi_1, \pi_2) := \min\left|\{\text{transpositions } \sigma_1, \ldots, \sigma_J : \sigma_1 \circ \ldots \circ \sigma_J \circ \pi_1 = \pi_2\}\right|.$$

For an estimated permutation $\widehat{\pi}$ we then set

$$d_{trans}\left(\widehat{\pi}, \Pi^0\right) := \min_{\pi^0 \in \Pi^0} d_{trans}(\widehat{\pi}, \pi^0).$$

	Additve SEs		Non-additve SEs	
N	$\overline{(d_{trans}(\widehat{\pi}, \Pi^0))}$	$\mathrm{SD}\big(d_{trans}(\widehat{\pi}, \Pi^0)\big)$	$\overline{(d_{trans}(\widehat{\pi}, \Pi^0))}$	$\mathrm{SD}\big(d_{trans}(\widehat{\pi}, \Pi^0)\big)$
10	1.59	1.69	1.57	1.82
20	0.80	1.39	0.73	1.21
50	0.07	0.29	0.36	0.76
100	0.01	0.10	0.31	0.74
200	0.01	0.10	0.30	0.66

Table 1: Mean ($\overline{(d_{trans}(\widehat{\pi}, \Pi^0))}$) and standard deviation (SD) of the transposition distances (d_{trans}) between the estimated permutation $\widehat{\pi}$ and the set of true permutations Π^0 for SEMs with additive and non-additive SEs. The underlying graphs are of ER type with on average five edges.

Thus, we calculate the minimal number of adjacent swaps so that the estimated permutation aligns with the underlying topological order.

The results are shown in Table 1 and, as expected, increasing the sample size decreases the transposition distances. This supports Theorem 15. Further, the results indicate that convergence even holds under non-additive SEs and thus that the algorithm is robust towards misspecifcation via non-additve SEs.

5.3 High-Dimensional Data

5.3.1 Comparison with Existing Methods

In the following we compare DAGBoost with Bühlmann et al. (2014) (denoted by CAM) and the non-parametric extension of NOTEARS by Zheng et al. (2020) (denoted by DAGSNOTEARS). For CAM we employ the default configuration and for DAGSNOTEARS we set $\lambda = 0.03$ and the cutoff to 0.3. For a comparison with further benchmark methods, we refer the reader to Bühlmann et al. (2014).

As performance measures, we calculate the Structural Hamming Distance (SHD) and the Structural Interventional Distance (SID; Peters and Bühlmann, 2015) between the true and estimated graphs for data sets containing $N = 200$ observations of $p = 100$ variables. Their means and standard deviations (SDs) are given in Tables 2 and 3. The SHD counts the number of edge additions, deletions, and reversals to transpose one graph into the other. Conversely, the SID describes the quality of the causal inferential statements derived from the estimated graph. For both, a lower value indicates a better result. Finally, Table 4 summarizes the means of precision and recall. Precision is the ratio between the correctly identified edges and all identified edges. Recall is the share of the correctly identified edges among all true edges.

From Table 2 we make the following observations. In terms of the SHD, DAGSNOTEARS does not provide satisfying results, while CAM and DAGBoost perform notably better in all simulation scenarios. Generally, all methods suffer from an uneven edge distribution (SF graphs). Further, DAGBoost achieves better results across all simulation scenarios. The advantage is largest in the most complex scenario of non-additve SEs and a non-even distribution of the edges among the nodes (SF graphs). In this case, DAGBoost outper-

		CAM		DAGBoost		DAGSNOTEARS	
Additive	Graph	$\overline{\text{SHD}}$	SD(SHD)	$\overline{\text{SHD}}$	SD(SHD)	$\overline{\text{SHD}}$	SD(SHD)
True	SF	28.82	9.10	**25.57**	8.98	87.47	3.61
True	ER	9.62	12.18	**3.33**	7.26	93.63	11.02
False	SF	74.04	7.77	**63.37**	5.88	87.79	4.10
False	ER	34.84	7.28	**28.30**	7.12	87.24	9.95

Table 2: Mean ($\overline{\text{SHD}}$) and standard deviation (SD) of SHDs between the true graph and the graphs estimated by the three presented algorithms.

		CAM		DAGBoost		DAGSNOTEARS	
Additive	Graph	$\overline{\text{SID}}$	SD(SID)	$\overline{\text{SID}}$	SD(SID)	$\overline{\text{SID}}$	SD(SID)
True	SF	**97.24**	67.45	106.23	46.43	442.32	118.69
True	ER	**8.96**	12.18	16.71	52.08	791.91	257.79
False	SF	236.13	66.71	**224.02**	49.38	294.78	58.73
False	ER	**191.97**	98.67	239.31	100.40	655.74	227.55

Table 3: Mean ($\overline{\text{SID}}$) and standard deviation (SD) of SIDs between the true graph and the graphs estimated by the three presented algorithms.

forms CAM by more than one standard deviation. In contrast to the SHD, Table 3 shows

		CAM		DAGBoost		DAGSNOTEARS	
Additive	Graph	Precision	Recall	Precision	Recall	Precision	Recall
True	SF	0.872	**0.825**	**0.984**	0.751	0.424	0.135
True	ER	0.919	**0.990**	**0.985**	0.980	0.252	0.395
False	SF	0.689	**0.421**	**0.925**	0.373	0.587	0.120
False	ER	0.814	**0.812**	**0.932**	0.748	0.506	0.242

Table 4: Mean of precision and recall for the three presented algorithms. Precision is the ratio between the correctly identified edges and all identified edges. Recall is the share of the correctly identified edges among all true edges.

that the SID values come with high SDs. Ignoring these, the table provides some evidence that CAM outperforms DAGBoost slightly in terms of the SID. A notable difference is the most complex scenario of non-additive functions and an SF graph, where DAGBoost is the best method. Together with the results for the SHD, this is an important insight for real-world applications, which often follow SF graphs. This is a strong argument in favor of our DAGBoost over CAM for scenarios with locally non-sparse graphs. Table 4 reveals that precision, that is, the ratio of the correctly identified edges to all identified edges, is higher for DAGBoost. However, the recall, which is the share of the identified edges among all true edges, is larger for CAM. Hence, the edges of DAGBoost are more reliable, while CAM misses a lower number of the underlying relationships. This observation also explains the

different observations for the SID and SHD. As SID does not penalize superfluous edges, the ability of DAGBoost to identify edges more reliably is not rewarded.

In summary, DAGBoost is a strong competitor to CAM, both of which outperform DAGSNOTEARS. In particular, DAGBoost tends to estimate more reliable edges and performs superior compared to CAM in complex non-additive and SF settings, which are commonly observed in real scenarios. Additionally, DAGBoost requires fewer tuning parameters. While CAM has additional tuning parameters for the preliminary neighborhood selection (PNS, see Bühlmann et al., 2014), DAGBoost only requires tuning the step size ν and the penalty parameter λ which we found to be rather robust (see the sensitivity analysis below), thus relatively easy to tune. Pruning the resulting graph improves the SHD between the estimated and the true graph for DAGBoost. However, this effect is much larger for CAM as the graph before pruning contains many more edges. Thus, DAGBoost is less reliant on the hyperparameters of the pruning step compared to CAM.

5.3.2 Sensitivity Analysis

We investigate the performance of DAGBoost with respect to a variation of the step size ν and the penalty parameter λ. When varying λ, we set $\nu = 0.3$. When varying ν, we fix $\lambda = 0.01$. We conduct our analysis with ER graphs with $p = 100$ nodes and additive SEs. The means and SDs of the SHD between the estimated and true graph are reported in Tables 5 and 6.

ν	mean(SHD)	SD(SHD)	mean(runtime in s)	SD(runtime in s)
0.3	3.33	7.26	89.34	13.20
0.5	2.40	2.20	44.74	10.36
0.7	2.37	2.29	27.61	3.94
0.9	2.34	2.09	22.41	2.97

Table 5: Mean and standard deviation (SD) of SHDs between estimated and true graph for a varying step size ν. The penalty parameter is fixed at $\lambda = 0.01$. The graphs are of ER type and the SEs are additive. The runtime statistics are based on 10 experiments.

λ	mean(SHD)	SD(SHD)	mean(runtime in s)	SD(runtime in s)
0.001	2.59	2.51	64.05	13.61
0.01	3.33	7.26	89.34	13.20
0.1	3.87	2.77	255.81	49.20

Table 6: Mean and standard deviation (SD) of SHDs between estimated and true graph for a varying penalty parameter λ. The step size is fixed at $\nu = 0.3$. The graphs are of ER type and the SEs are additive. The runtime statistics are based on 10 experiments.

From these tables, one can see that the influence of the hyperparameters on the SHDs is minor. The AIC score controls the number of boosting steps m_{stop} very efficiently. Thus, it accounts well for the different base learners, which depend on the step size ν and the penalty parameter λ. An increase in the step size or a reduction in the penalty parameter leads to a smaller number of boosting iterations which, in turn, reduces runtime. At the same time, the quality of the estimated graph is largely unaffected. We thus recommend to use DAGBoost with a large step size or a low penalty parameter if the computational resources or time are limited.

Although the impact of the hyperparameters is shown for one specific setting, based on our observations, they similarly hold in a wide range of data-generating processes.

5.3.3 Early stopping based on hold-out set

Our default choice for determining m_{stop} is the AIC. However, motivated by a referee's suggestion, we also consider an alternative early stopping strategy using a hold-out set for determining m_{stop}. Hereby, we split the data into a train and a hold-out set of the same size and monitor the mean squared error (MSE) on the hold-out set. The algorithm stops if the MSE calculated on the hold-out set increases. All numbers presented below are based on ER graphs with additive SEs.

Small p The results of this approach are worse than early stopping with the AIC score. For example, for $N = 50$, the mean permutation distance with early stopping via the MSE on the hold-out set is 0.73 with standard deviation 1.11. With the AIC score, the mean permutation distance is 0.07 with a standard deviation of 0.29. The permutation distance for the hold-out set uses a training set of size $N = 25$ and the mean and standard deviation are close to early stopping with the AIC score with $N = 20$. Hence, stopping based on a hold-out cannot compensate for the reduced training set size.

Large p In the example of ER graphs with additive SEs, we see a very high mean precision of 0.996. However, the mean recall is low at 0.261. We conclude that the hold-out set usually stops too early and leads to inferior results.

Although both results could potentially be improved by tuning the stopping criterion or adapting the train-to-test-set ratio, we observe that the AIC score requires no such tuning, and provides stronger results.

6. Conclusion

In this work we investigated boosting for causal discovery and for the estimation of causal ordering. We proposed a generic score function on the orderings that depends on a regression estimator. We presented two sufficient conditions on the regression estimator, so that the score function can consistently distinguish between compatible and incompatible orderings:

1. The regression estimator must consistently find the true regression function in a correctly specified scenario with homoscedastic noise.

2. The mean squared prediction error on the samples must converge to the expected L_2 prediction error for yet unseen observations even under model misspecification.

Together, these conditions provide a theoretical safety net for the regression estimator, which is interesting on its own. On the one hand, in a misspecified setting, the fit of the regression function to the observed samples gives a good estimate for the expected squared prediction error for yet unobserved realizations. In a correctly specified setting on the other hand, the regression estimator still identifies the underlying functional relationship.
We showed that boosting with appropriate early stopping. This provides new insights into the generalization ability of boosting methods, especially in real-world scenarios where model assumptions may not hold.

To use a score function on the orderings for the identification of the topological order, one needs to score every possible permutation. This is infeasible for large p and insufficient prior knowledge on the causal structure. Thus, we proposed a greedy boosting algorithm in the space of functions of $\mathbb{R}^p \to \mathbb{R}^p$ which correspond to a DAG. The algorithm can be understood as a functional gradient descent in the space of additive SEMs, also known as component-wise boosting.

Our simulation study underlined that the score function on the permutations consistently prefers the correct causal order and the convergence manifests already for small N. These findings were even robust to non-additive SEs. For small p or in the case of extensive prior knowledge that drastically reduces the search space of permutations, the combination of the score and a feature selection procedure can be used to derive the causal graph.
The second part of our simulation study showed that the gradient descent is highly competitive with state-of-the-art algorithms. In particular, for complex data-generating processes, the algorithm provides a noticeable benefit. Moreoever, the exact choice of the hyperparameters are efficiently and automatically balanced by the number of boosting iterations and the AIC score. Thus, the procedure is easy to tune and ready to be applied to a variety of data sets.

In the future, the generalization of Theorem 15 for high dimensions would be highly desirable, and the approach of Section 4 could be enhanced by a Preliminary Neighborhood Selection (PNS) step as suggested in Bühlmann et al. (2014). However, PNS would require an additional screening algorithm with additional hyperparameters. Without PNS we would need to show, among others, Condition 1 of Proposition 5, that is, the consistency of the nonparametric L_2 boosting regression for high dimensions. To the best of our knowledge, such a result does not yet exist and would be of great interest alone. Recently, Stankewitz (2024) considered early stopping L_2 boosting in high-dimensional linear models and it could be a promising starting point.

Finally, many aspects of our analysis are generic. The RKHS regression could easily be replaced by other regression estimators, such as spline regression or neural networks. Thus, one could further investigate which regression estimators lead to consistent estimators for the causal order, and their empirical performance and theoretical properties could be explored. A combination of continuous optimization methods as in Zheng et al. (2018) and component-wise boosting may offer further insights into scalable and robust causal discovery.

Acknowledgments

This work has been supported by the German Federal Ministry of Education and Research within the project "CausalNet", grant no. 01IS24082 and the BMW AG.

Appendix A. Proof of Proposition 5

Proof Recall that $\varpi_\pi(k) := \{j : \pi(j) < \pi(k)\}$.

$$
\begin{aligned}
\widehat{S}(\pi) &= \sum_{k=1}^{p} \log(\widehat{\sigma}^2_{k,\widehat{f}_{k,\pi}}) \\
&\geq \min_{(f_1,\ldots,f_p)\in\vartheta(\pi)} \sum_{k=1}^{p} \log\left((P_N - P)\left(X_k - \widehat{f}_{k,\pi}(\mathbf{X}_{\varpi_\pi(k)})\right)^2 + \sigma^2_{k,p_{\theta^0},f_k,\pi} \right) \\
&\geq \min_{(f_1,\ldots,f_p)\in\vartheta(\pi)} \sum_{k=1}^{p} \log(\sigma^2_{k,p_{\theta^0},f_k,\pi}) \\
&\quad + \underbrace{\max\left\{0, \frac{-(P_N - P)\left(X_k - \widehat{f}_{k,\pi}(\mathbf{X}_{\varpi_\pi(k)})\right)^2}{\sigma^2_{k,p_{\theta^0},f_k,\pi} + (P_N - P)\left(X_k - \widehat{f}_{k,\pi}(\mathbf{X}_{\varpi_\pi(k)})\right)^2}\right\}}_{=:\Delta_{N,k}} \\
&= \min_{(f_1,\ldots,f_p)\in\vartheta(\pi)} \sum_{k=1}^{p} \log(\sigma^2_{k,p_{\theta^0},f_k,\pi}) + \sum_{k=1}^{p} \Delta_{N,k} \\
&= S(\pi) + \sum_{k=1}^{p} \Delta_{N,k} \\
&= S(\pi^0) + \sum_{k=1}^{p} \Delta_{N,k} + \underbrace{S(\pi) - S(\pi^0)}_{\xi_{\pi,\pi^0}>0} \\
&= \sum_{k=1}^{p} \log\left(\left(\sigma^0_k\right)^2 - \widehat{\sigma}^2_{k,\widehat{f}_{k,\pi^0}} + \widehat{\sigma}^2_{k,\widehat{f}_{k,\pi^0}} \right) + \Delta_{N,k} + \xi_{\pi,\pi^0} \\
&\geq \sum_{k=1}^{p} \log\left(\widehat{\sigma}^2_{k,\widehat{f}_{k,\pi^0}} \right) + \underbrace{\max\left\{0, -\frac{\left(\sigma^0_k\right)^2 - \widehat{\sigma}^2_{k,\widehat{f}_{k,\pi^0}}}{\left(\sigma^0_k\right)^2}\right\}}_{=:\gamma_{N,k}} + \Delta_{N,k} + \xi_{\pi,\pi^0} \\
&= \widehat{S}(\pi^0) + \sum_{k=1}^{p} (\gamma_{N,k} + \Delta_{N,k}) + \xi_{\pi,\pi^0}
\end{aligned}
$$

In the first inequality, we used that for any $k = 1, \ldots, p$

$$
\begin{aligned}
P\left((X_k - \widehat{f}_{k,\pi}(\mathbf{X}_{\varpi_\pi(k)})^2\right) &= \mathbb{E}_{p_{\theta^0}}\left[(X_k - \widehat{f}_{k,\pi}(\mathbf{X}_{\varpi_\pi(k)})^2\right] \\
&\geq \min_{(f_1,\ldots,f_p)\in\vartheta(\pi)} \mathbb{E}_{p_{\theta^0}}\left[(X_k - f_k(\mathbf{X}_{\varpi_\pi(k)})^2\right] \\
&= \min_{(f_1,\ldots,f_p)\in\vartheta(\pi)} \sigma^2_{k,p_{\theta^0},f_k,\pi}.
\end{aligned}
$$

In the second and last inequality Lemma 23 below was used. Further, $\xi_{\pi,\pi^0} > 0$, which does not depend on N, by the identifiability of the model. By the assumptions and the

continuous mapping theorem it holds

$$\mathbb{P}\left(|\Delta_{N,k}| \geq \frac{\xi_{\pi,\pi^0}}{2p}\right) \to 0 \text{ and } \mathbb{P}\left(|\gamma_{N,k}| \geq \frac{\xi_{\pi,\pi^0}}{2p}\right) \to 0 \text{ for all } k = 1, \ldots, p \text{ and } N \to \infty,$$

from which we derive that

$$\xi_N = |\sum_{k=1}^{p} \Delta_{N,k} + \gamma_{N,k}| \leq \sum_{k=1}^{p} |\Delta_{N,k}| + \sum_{k=1}^{p} |\gamma_{N,k}| < \xi_{\pi,\pi^0}$$

with probability going to 1 for $N \to \infty$. ■

Lemma 23 *For $x > 0$ and $x + \delta > 0$ it holds that*

$$\log(x + \delta) \geq \log(x) + \max\left\{0, -\frac{\delta}{x + \delta}\right\}$$

Proof The statement is true for $\delta \geq 0$. For $\delta < 0$ it holds that

$$\begin{aligned}\log(x) &= \log(x - |\delta| + |\delta|) \leq \log(x - |\delta|) + \frac{|\delta|}{x - |\delta|} = \log(x + \delta) - \frac{\delta}{x + \delta} \\ &= \log(x + \delta) - \max\left\{0, -\frac{\delta}{x + \delta}\right\}\end{aligned}$$

from which the result follows. ■

Appendix B. Proof of Theorem 18

The proof of Theorem 18 decomposes $(P_N - P)\left(Y - \widehat{f}^{(m_{stop})}\right)^2$ into $(P_N - P)\left(Y\widehat{f}^{(m_{stop})}\right)$ and $(P_N - P)\left(\widehat{f}^{(m_{stop})}\right)^2$. We show both convergences using the theory of empirical processes.

For this purpose, we show in Section B.1 that $||\widehat{f}^{(m_{stop})}||_H$ can be upper-bounded. Section B.2 then derives the convergence of $(P_N - P)\left(Y\widehat{f}^{(m_{stop})}\right)$ using *Covering Numbers*. The convergence of $(P_N - P)\left(\widehat{f}^{(m_{stop})}\right)^2$ is then shown using *Rademacher Complexities*.

B.1 Upper-Bound of the RKHS Norm of Regression Function Estimate

The main result of this section is Lemma 27, which upper-bounds the Hilbert space norm of the boosting estimate $||\widehat{f}^{(m)}||_H$ with a probability going to 1 for $N \to \infty$ for a suitably chosen number of boosting iterations m. The analysis is based on the fact that $||\widehat{f}^{(m)}||_H^2$ can be expressed as a quadratic form.

Lemma 24 *Let S be the kernel regression learner with penalty parameter λ. Assume that the Gram matrix G is invertible. Then it holds for the boosting estimate after m boosting steps that*

$$||\widehat{f}^{(m)}||_H^2 = \frac{1}{N}(y^N)^T U(I-(I-D)^m)^2\Lambda^{-1}U^T y^N,$$

where $U, D, \Lambda \in \mathbb{R}^{N\times N}$ and D, Λ are diagonal matrices. Λ has the eigenvalues $\widehat{\mu}_1, \ldots, \widehat{\mu}_N$ of G and D has $\frac{\widehat{\mu}_1}{\widehat{\mu}_1+\lambda}, \ldots, \frac{\widehat{\mu}_N}{\widehat{\mu}_N+\lambda}$ on the diagonal. U contains the corresponding eigenvectors of G.

Proof It holds for the linear base learner S mapping y^N to $\widehat{f}(\mathbf{x}^N)$, that

$$S = G(G+\lambda I)^{-1}.$$

The matrix S is symmetric and has the eigenvalues $d_1 = \frac{\widehat{\mu}_1}{\widehat{\mu}_1+\lambda}, \ldots, d_N = \frac{\widehat{\mu}_N}{\widehat{\mu}_N+\lambda}$. Thus, for the orthogonal matrix U containing the eigenvectors of S and the diagonal matrix D containing the eigenvalues $d_1, \ldots, d_N$ of S it holds that

$$S = UDU^T.$$

By Equation (14), the estimate $\widehat{f}^{(m)} = B^{(m)}y^N = (I-(I-S)^m)y^N$ can be expressed by

$$\widehat{f}^{(m)}(\widetilde{\mathbf{x}}) = \frac{1}{\sqrt{N}}\sum_{k=1}^N \widehat{\beta}_k K(\widetilde{\mathbf{x}}, \widetilde{\mathbf{x}}_k)$$

for some $\widehat{\beta} \in \mathbb{R}^N$. Using the results of Section 3.2 we obtain

$$\begin{aligned}
||\widehat{f}^{(m)}||_H^2 &= \widehat{\beta}^\top G \widehat{\beta} \\
&= \widehat{f}^{(m)}(\widetilde{x}^N)^\top \frac{G^{-1}}{\sqrt{N}} G \frac{G^{-1}}{\sqrt{N}} \widehat{f}^{(m)}(\widetilde{x}^N) \\
&= \frac{1}{N}(y^N)^\top B^{(m)} G^{-1} B^{(m)} y^N.
\end{aligned}$$

Note that G, S and $B^{(m)}$ have the same eigenvectors. Besides, Λ and $I-(I-D)^m$ are diagonal matrices, so they commute. Hence, $B^{(m)}G^{-1}B^{(m)} = U(I-(I-D)^m)U^\top U\Lambda^{-1}U^\top U(I-(I-D)^m)U^\top = U(I-(I-D)^m)^2\Lambda^{-1}U^T$. Thus,

$$||\widehat{f}^{(m)}||_H^2 = \frac{1}{N}(y^N)^\top B^{(m)} G^{-1} B^{(m)} y^N = \frac{1}{N}(y^N)^\top U(I-(I-D)^m)^2\Lambda^{-1}U^\top y^N.$$

■

The following Hanson-Wright-inequality gives probabilistic upper bounds for quadratic forms as derived in Lemma 24.

Theorem 25 (Hanson-Wright-Inequality) *(Rudelson and Vershynin, 2013, Theorem 1.1) Consider a random vector $\mathbf{Z} = (Z_1, \ldots, Z_N) \in \mathbb{R}^N$ with independent components, for which $\mathbb{E}(Z_\ell) = 0, \ell = 1, \ldots, N$ and for which the Orlicz norm of $Z_1, \ldots, Z_N$ is uniformly bounded by s_{max}. For any $M \in \mathbb{R}^{N\times N}$ it holds for every $t > 0$*

$$\mathbb{P}\left(\left|||M\mathbf{Z}||^2 - \mathbb{E}\left[||M\mathbf{Z}||^2\right]\right| > t\right) \leq 2\exp\left(-c\min\left\{\frac{t^2}{s_{max}^4||M^2||_F^2}, \frac{t}{s_{max}^2||M^2||}\right\}\right).$$

Lemma 24 shows that the RKHS norm can be expressed as a quadratic form. In the next steps, we upper bound the quadratic form, that is, the RKHS norm of $\widehat{f}^{(m)}$ using Theorem 25. We see, that we can choose M in Theorem 25 as the matrix square root of $U(I-(I-D)^m)^2\Lambda^{-1}U^T$ so that $||MY||^2 = ||\widehat{f}^{(m)}||_H^2$. Thus, these bounds depend on the number of boosting steps m and D and Λ (which are functions of G), where the latter are probabilistically depending on $\widetilde{\mathbf{X}}^N$. Controlling D, this allows us to vary m with N such that the growth of the upper bound for $||\widehat{f}^{(m)}||_H^2$ can be controlled with a high probability. Observe that Theorem 25 requires centered random variables $Z_1, \ldots, Z_N$.

Lemma 26 *Decompose* $Y = \mu(\widetilde{\mathbf{X}}) + \varepsilon(\widetilde{\mathbf{X}})$, *where* $\mu(\widetilde{\mathbf{X}}) = \mathbb{E}\left[Y|\widetilde{\mathbf{X}}\right]$ *and* $\varepsilon(\widetilde{\mathbf{X}}) = Y - \mathbb{E}\left[Y|\widetilde{\mathbf{X}}\right]$. *Assume that* $||\mu||_\infty < \mu_{max}$ *and the Orlicz norm and variance of the conditional distribution of* $\varepsilon(\widetilde{\mathbf{X}})$ *given some realization* $\widetilde{\mathbf{x}}$ *of* $\widetilde{\mathbf{X}}$ *is uniformly bounded by* s_{max} *and* σ^2_{max}, *respectively. Let* $\widehat{f}^{(m)}$ *be the boosting estimate after* m *boosting steps and* $\Lambda, D, U \in \mathbb{R}^{N\times N}$ *as in Lemma 24. We can upper bound*

$$\begin{aligned}\mathbb{P}\left(||\widehat{f}^{(m)}||_H^2 > 1 + 2N(\mu^2_{max}+\sigma^2_{max})||M_m^{1/2}(\widetilde{\mathbf{x}}^N)||_F^2|\widetilde{\mathbf{X}}^N = \widetilde{\mathbf{x}}^N\right)\\ \leq 2\exp\left(-C\min\left\{\frac{1}{4s^4_{max}||M_m||_F^2}, \frac{1}{2s^2_{max}||M_m||}\right\}\right).\end{aligned}$$

where

$$M_m(\widetilde{\mathbf{x}}^N) := M_m := \frac{1}{N}U(I-(I-D)^m)^2\Lambda^{-1}U^T.$$

Proof By Lemma 24 it holds

$$||\widehat{f}^{(m)}||_H^2 = \frac{1}{N}(y^N)^T U(I-(I-D)^m)^2\Lambda^{-1}U^T y^N = \left(y^N\right)^\top M_m(\widetilde{\mathbf{x}}^N)y^N.$$

We emphasize that G and thus D, U and S are functions of $\widetilde{\mathbf{X}}^N$. We calculate for $\mu^N = (\mu(\widetilde{\mathbf{x}}_1), \ldots, \mu(\widetilde{\mathbf{x}}_N)) \in \mathbb{R}^N$ and $\varepsilon^N = (\varepsilon_1(\widetilde{\mathbf{x}}_1), \ldots, \varepsilon_N(\widetilde{\mathbf{x}}_N)) \in \mathbb{R}^N$:

$$\begin{aligned}&\mathbb{P}\left(||\widehat{f}^{(m)}||_H^2 - 2N(\mu^2_{max}+\sigma^2_{max})||M_m^{1/2}(\widetilde{\mathbf{X}}^N)||_F^2 > 1|\widetilde{\mathbf{X}}^N = \widetilde{\mathbf{x}}^N\right)\\ &= \mathbb{P}\left(||M_m^{1/2}(\widetilde{\mathbf{X}}^N)Y^N||^2 - 2N(\mu^2_{max}+\sigma^2_{max})||M_m^{1/2}(\widetilde{\mathbf{X}}^N)||_F^2 > 1|\widetilde{\mathbf{X}}^N = \widetilde{\mathbf{x}}^N\right)\\ &\leq \mathbb{P}\Big(2||M_m^{1/2}(\widetilde{\mathbf{X}}^N)\mu^N||^2 + 2||M_m^{1/2}(\widetilde{\mathbf{X}}^N)\varepsilon^N||^2\\ &\qquad - 2N(\mu^2_{max}+\sigma^2_{max})||M_m^{1/2}(\widetilde{\mathbf{X}}^N)||_F^2 > 1|\widetilde{\mathbf{X}}^N = \widetilde{\mathbf{x}}^N\Big)\end{aligned}$$

In the third line we used the decomposition $y^N = \mu^N + \varepsilon^N$ and the inequality $||a+b||^2 \leq 2||a||^2 + 2||b||^2$. In the following we upper-bound the terms $||M_m^{1/2}(\widetilde{\mathbf{X}}^N)\mu^N||$ and

$-N||M_m^{1/2}(\widetilde{\mathbf{X}}^N)||_F^2\sigma_{max}^2$. For the first term observe that

$$\begin{aligned}||M_m^{1/2}(\widetilde{\mathbf{X}}^N)\mu^N||^2 &= tr\left(\left(\left(\mu^N\right)^\top M_m(\widetilde{\mathbf{X}}^N)\mu^N\right)\right.\\ &= tr\left(M_m(\widetilde{\mathbf{X}}^N)\mu^N\left(\mu^N\right)^\top\right)\\ &\leq tr\left(M_m(\widetilde{\mathbf{X}}^N)\right)tr\left(\mu^N\left(\mu^N\right)^\top\right)\\ &\leq N||M_m^{1/2}\left(\widetilde{\mathbf{X}}^N\right)||_F^2\mu_{max}^2.\end{aligned}$$

For the latter term using that $\mathbf{E}\left[\varepsilon_k^2|\widetilde{\mathbf{X}}^N\right] \leq \sigma_{max}^2, k = 1,\ldots,N$ it holds analogously

$$\begin{aligned}\mathbb{E}\left[||M_m^{1/2}(\widetilde{\mathbf{X}}^N)\varepsilon^N||^2|\widetilde{\mathbf{X}}^N\right] &= \mathbb{E}\left[tr\left(\left(\varepsilon^N\right)^\top M_m(\widetilde{\mathbf{X}}^N)\varepsilon^N\right)|\widetilde{\mathbf{X}}^N\right]\\ &\leq \mathbb{E}\left[tr\left(M_m(\widetilde{\mathbf{X}}^N)\right)tr\left(\varepsilon^N\left(\varepsilon^N\right)^\top\right)|\widetilde{\mathbf{X}}^N\right]\\ &\leq tr\left(M_m(\widetilde{\mathbf{X}}^N)\right)\mathbb{E}\left[\left(\varepsilon^N\right)^\top\varepsilon^N|\widetilde{\mathbf{X}}^N\right]\\ &\leq N||M_m^{1/2}(\widetilde{\mathbf{X}}^N)||_F^2\sigma_{max}^2\end{aligned}$$

and hence $-N||M_m^{1/2}(\widetilde{\mathbf{X}}^N)||_F^2\sigma_{max}^2 \leq -\mathbb{E}\left[||M_m^{1/2}(\widetilde{\mathbf{X}}^N)\varepsilon^N||^2|\widetilde{\mathbf{X}}^N\right]$. Using these results and plugging in the condition $\widetilde{\mathbf{X}}^N = \widetilde{\mathbf{x}}^N$ we can upper-bound

$$\begin{aligned}&\mathbb{P}\left(2||M_m^{1/2}\mu^N||^2 + 2||M_m^{1/2}\varepsilon^N||^2 - 2N(\mu_{max}^2+\sigma_{max}^2)||M_m^{1/2}(\widetilde{\mathbf{X}}^N)||_F^2 > 1|\widetilde{\mathbf{X}}^N = \widetilde{\mathbf{x}}^N\right)\\ &\leq \mathbb{P}\left(||M_m^{1/2}\varepsilon^N||^2 - \mathbb{E}\left(||M_m^{1/2}(\widetilde{\mathbf{X}}^N)\varepsilon^N||^2|\widetilde{\mathbf{X}}^N\right) > \frac{1}{2}|\widetilde{\mathbf{X}}^N = \widetilde{\mathbf{x}}^N\right)\\ &= \mathbb{P}\left(||M_m^{1/2}\varepsilon^N||^2 - \mathbb{E}\left(||M_m^{1/2}\varepsilon^N||^2\right) > \frac{1}{2}|\widetilde{\mathbf{X}}^N = \widetilde{\mathbf{x}}^N\right)\\ &\leq 2\exp\left(-C\min\left\{\frac{1}{4s_{max}^4||M_m||_F^2}, \frac{1}{2s_{max}^2||M_m||}\right\}\right)\end{aligned}$$

by Theorem 25 (Hanson-Wright-inequality) and the fact that $\varepsilon^N|\widetilde{\mathbf{X}}^N = \widetilde{\mathbf{x}}^N$ is a centered, sub-Gaussian random vector with independent components with Orlicz norm bounded by s_{max}. ■

We now show that if we choose $m = m(N) = N^{\frac{1}{4}\frac{C_u+C_d+1/2}{C_d+1}}$ then the growth of $||\widehat{f}^{(m)}||_H$ with N is of lower order as $N^{1/4}$ with a probability going to 1 if $\widetilde{\mathbf{x}}^N \in \mathcal{B}_N$.

Lemma 27 (Upper bound $||\widehat{f}^{(m_{stop})}||_H$) *Under Assumption 13′ and for $m_{stop} = m(N) = N^{\frac{1}{4}\frac{C_u+C_d+1/2}{C_d+1}}$ there exists a $\delta > 0$ and a $h(N) \in o(N^{1/4-\delta})$ so that*

$$\mathbb{P}\left(\widehat{f}^{(m_{stop})} \notin h(N)\mathcal{F}_N \cap \{\widetilde{\mathbf{X}}^N \in \mathcal{B}_N\}\right) \to 0 \tag{23}$$

for $N \to \infty$.

Proof As outlined in relation (14), by the representation theorem it holds that $\widehat{f}^{(m)} \in h(N)\mathcal{F}_N$ for some $h(N) \in \mathbb{R}$. Thus we need to show that $||\widehat{f}^{(m_{stop})}||_H \leq h(N)$. We prove the statement by showing the convergence

$$\mathbb{P}\left(||\widehat{f}^{(m_{stop})}||_H^2 > h(N)^2 | \widetilde{\mathbf{X}}^N = \widetilde{\mathbf{x}}^N\right) \to 0$$

uniformly for any $\widetilde{\mathbf{x}}^N \in \mathcal{B}_N$ and some $h(N) \in o(N^{1/4-\delta})$. The statement then follows by integrating out with respect to $\widetilde{\mathbf{x}}^N$. Applying Lemma 26, we need to prove the following two statements.

1. $N||M_m^{1/2}||_F^2 \in o\left(N^{1/2-2\delta}\right)$. This implies, $||M_m||_F^2 \leq ||M_m^{1/2}||_F^4 \in o(1)$.

2. $||M_m|| \in o(1)$.

It is emphasized that M_m is a function of the random sample $\widetilde{\mathbf{x}}^N$.
The statement is proven in Lemma 28. ■

Lemma 28 *Under Assumption 13′ it holds for*

$$M_m(\widetilde{\mathbf{x}}^N) := M_m := \frac{1}{N} U(I - (I-D)^m)^2 \Lambda^{-1} U^T$$

that there exists a $\delta > 0$ so that the following statements hold for $m = m(N) = N^{\frac{1}{4}\frac{C_u + C_d + 1/2}{C_d+1}}$ uniformly in $\widetilde{\mathbf{x}}^N \in \mathcal{B}_N$, where $\mathcal{B}_N$ is defined in Assumption 13′.

1. $N^{1/2+2\delta}||M_m^{1/2}||_F^2 = \frac{1}{N^{1/2-2\delta}} tr\left((I - (I-D)^m)^2 \Lambda^{-1}\right) \to 0$,

2. $||M_m||_F^2 \to 0$, *and*

3. $||M_m|| \to 0$.

Proof 1.: Uniformly in $\widetilde{\mathbf{x}}^N \in \mathcal{B}_N$ it holds that:

$$\begin{aligned}
N^{1/2+2\delta}||M_m^{1/2}||_F^2 &= N^{1/2+2\delta} tr(M_m) \\
&= \frac{1}{N^{1/2-2\delta}} \sum_{k=1}^{N} (1-(1-d_k)^m)^2 \widehat{\mu_k}^{-1} \\
&= \frac{1}{N^{1/2-2\delta}} \sum_{k=1}^{N} \left(1-\left(1-\frac{\widehat{\mu_k}}{\widehat{\mu_k}+\lambda}\right)^m\right)^2 \widehat{\mu_k}^{-1} \\
&\overset{(i)}{\leq} \frac{1}{N^{1/2-2\delta}} \sum_{k=1}^{N} \min\left\{1, m^2\left(\frac{\widehat{\mu_k}}{\widehat{\mu_k}+\lambda}\right)^2\right\} \widehat{\mu_k}^{-1} \\
&\leq \frac{1}{\lambda^2 N^{1/2-2\delta}} \sum_{k=1}^{N} \min\left\{\lambda^2 \widehat{\mu_k}^{-1}, m^2 \widehat{\mu_k}\right\} \\
&\leq \frac{1}{\lambda^2 N^{1/2-2\delta}} \left(\sum_{k=1}^{\lfloor K_0 \rfloor} \min\left\{\lambda^2 \widehat{\mu_k}^{-1}, m^2 \widehat{\mu_k}\right\} + \sum_{k=\lceil K_0 \rceil}^{N} \min\left\{\lambda^2 \widehat{\mu_k}^{-1}, m^2 \widehat{\mu_k}\right\}\right) \\
&\leq \frac{1}{\lambda^2 N^{1/2-2\delta}} \left(K_0 \widehat{\mu_{K_0}}^{-1} \lambda^2 + \sum_{k=\lceil K_0 \rceil}^{N} m^2 \widehat{\mu_k}\right),
\end{aligned}$$

which holds for any $K_0 \in \mathbb{N}$. The inequality (i) is shown in Lemma 30. The last inequality is due to the fact that $\widehat{\mu}_k^{-1}$ is monotonically increasing. We choose $K_0 = K_0(N) = \lfloor \frac{1}{2C_d+1} \ln(N) \rfloor$. Uniformly in $\widetilde{\mathbf{x}}^N \in \mathcal{B}_N$ it holds by Assumption 13′ for a small $\delta > 0$

$$\frac{K_0 \widehat{\mu_{K_0}}^{-1}}{N^{1/2-2\delta}} \leq \frac{\exp(C_d K_0)}{N^{1/2-2\delta}} \to 0.$$

For the latter part observe that it holds for any $\widetilde{\mathbf{x}}^N \in \mathcal{B}_N$ and by Assumption 13′

$$\begin{aligned}
\frac{1}{N^{1/2-2\delta}} \sum_{k=\lceil K_0 \rceil}^{N} m^2 \widehat{\mu_k} &\leq \frac{m^2}{N^{1/2-2\delta}} \sum_{k=\lceil K_0 \rceil}^{N} \exp(-C_u k) \\
&\leq \frac{m^2}{N^{1/2-2\delta}} \int_{K_0}^{\infty} \exp(-C_u(z-1)) dz \\
&= \frac{m^2 \exp(C_u)}{N^{1/2-2\delta} C_u} \exp(-C_u K_0) \\
&= \frac{m^2 \exp(C_u)}{N^{1/2-2\delta} C_u} \exp(-C_u(\underbrace{K_0+1}_{\geq \frac{1}{2C_d+1}\ln(N)} - 1)) \\
&\leq \frac{m^2 \exp(2C_u)}{N^{1/2-2\delta} C_u} N^{-\frac{C_u}{2C_d+1}} \\
&= \frac{m^2 \exp(2C_u)}{C_u} N^{-\frac{C_u + C_d + 1/2}{2C_d+1}} N^{2\delta}.
\end{aligned}$$

Observe that for $m = m(N) = N^{\frac{1}{4}\frac{C_u+C_d+1/2}{C_d+1}}$ where the constants are chosen independently of $\widetilde{\mathbf{x}}^N$, it holds

$$m^2 N^{-\frac{C_u+C_d+1/2}{2C_d+1}} = N^{\frac{C_u+C_d+1/2}{2C_d+2} - \frac{C_u+C_d+1/2}{2C_d+1}} = N^{-\xi},$$

where $\xi := \frac{C_u+C_d+1/2}{2C_d+1} - \frac{C_u+C_d+1/2}{2C_d+2} > 0$. For $\delta < \frac{\xi}{2}$ it holds that $\frac{1}{N^{1/2-2\delta}} \sum_{k=\lceil K_0 \rceil}^{N} m^2 \widehat{\mu_k} \to 0$.
2. follows by $||M_m||_F^2 \leq ||M_m^{1/2}||_F^4 \to 0$.
3. follows by

$$\begin{aligned}
||M_m|| &= \frac{1}{N} \max_{k=1,\ldots,N} \left(1 - \left(1 - \frac{\widehat{\mu}_k}{\widehat{\mu}_k + \lambda}\right)^m\right)^2 \widehat{\mu}_k^{-1} \\
&\overset{(i)}{\leq} \frac{1}{N} \max_{k=1,\ldots,N} \min\left\{1, m^2 \left(\frac{\widehat{\mu_k}}{\widehat{\mu_k} + \lambda}\right)^2\right\} \widehat{\mu_k}^{-1} \\
&\leq \frac{1}{\lambda^2 N} \max_{k=1,\ldots,N} \min\left\{\widehat{\mu_k}^{-1}, m^2 \widehat{\mu_k}\right\} \leq \frac{m^2 \widehat{\mu}_1}{\lambda^2 N} \to 0,
\end{aligned}$$

where inequality (i) follows again from Lemma 30 as $\widehat{\mu}_1 \leq 1$.

■

The following Lemma immediately follows from the proof of Lemma 28.

Lemma 29 *Under Assumption 13′ it holds for* $m = m(N) = N^{\frac{1}{4}\frac{C_u+C_d+1/2}{C_d+1}}$ *on* $\mathcal{B}_N$, *that*

$$\frac{1}{\sqrt{N}} \sum_{\ell=1}^{N} \left(1 - \left(1 - \frac{\widehat{\mu}_\ell}{\widehat{\mu}_\ell + \lambda}\right)^m\right)^2 \to 0.$$

Lemma 30 *It holds for* $0 \leq \widehat{\mu_k} \leq 1$, *that*

$$1 - (1 - \widehat{\mu_k})^m \leq 1 - \max\{0, 1 - m\widehat{\mu_k}\} = \min\{1, m\widehat{\mu_k}\}$$

Proof It is equivalent to show that

$$(1 - \widehat{\mu_k})^m \geq \max\{0, 1 - m\widehat{\mu_k}\}.$$

The l.h.s. and the r.h.s. are equal for $\widehat{\mu_k} = 0$. On the interval $[0, \frac{1}{m})$ it holds that

$$\frac{\partial(1 - \widehat{\mu_k})^m}{\partial \widehat{\mu_k}} = -m(1 - \widehat{\mu_k})^{m-1} \geq -m = \frac{\partial(\max\{0, 1 - m\widehat{\mu_k}\}}{\partial \widehat{\mu_k}}.$$

Hence,

$$(1 - \widehat{\mu_k})^m \geq \max\{0, 1 - m\widehat{\mu_k}\} \text{ for } \widehat{\mu_k} \in [0, \frac{1}{m}).$$

Clearly, for $\mu_k \in [\frac{1}{m}, 1]$

$$(1 - \widehat{\mu_k})^m \geq \max\{0, 1 - m\widehat{\mu_k}\}.$$

■

B.2 Results on Covering Numbers

Lemma 27 has shown that for $m_{stop} = N^{\frac{1}{4}\frac{C_u+C_d+1/2}{C_d+1}}$, it holds $\widehat{f}^{(m_{stop})} \in h(N)\mathcal{F}_N$ for some $h(N) \in o(N^{1/4})$ with probability going to 1 for $N \to \infty$. In this section we use the covering numbers from empirical process theory to show the convergence of the inner product

$$|(P - P_N)Y\widehat{f}^{(m_{stop})}|. \tag{24}$$

The covering numbers measure the complexity of a function class.

Definition 31 *For a function class $\mathcal{F}$ and a semi-metric d on $\mathcal{F}$ the covering number $\mathcal{N}(\varepsilon, \mathcal{F}, d)$ is the minimal size of a subset $S \subset \mathcal{F}$, such that for every $f \in \mathcal{F}$ there is an $s \in S$ so that $d(f, s) < \varepsilon$. More precisely,*

$$\mathcal{N}(\varepsilon, \mathcal{F}, d) = \min_{\{S \subset \mathcal{F} | \forall f \in \mathcal{F}\, \exists s \in S : d(f,s) < \varepsilon\}} |S|.$$

In this work, we choose $d(f, g) = ||f - g||_\infty$ and thus write $\mathcal{N}(u, \mathcal{F}, ||\cdot||_\infty) = \mathcal{N}(u, \mathcal{F})$. For a suitable C_0 (chosen as in Dudley's Theorem, see Theorem 8.4 of van de Geer, 2014) not depending on $\mathcal{F}$ we define the covering number entropy integral by

$$\mathcal{J}(z, \mathcal{F}) := C_0 z \int_0^1 \sqrt{\log \mathcal{N}\left(\frac{uz}{2}, \mathcal{F}\right)} du.$$

For a constant $C > 0$, let $\mathcal{H} = \{Ch | h \in \mathcal{G}\}$ be a scaled version of some function class $\mathcal{G}$. The following remark shows that the entropy integral $\mathcal{J}(z, \mathcal{H})$ of $\mathcal{H}$ can be upper bounded by an expression depending on C and the entropy integral $\mathcal{J}(z, \mathcal{G})$ of $\mathcal{G}$.

Remark 32 *For $C > 0$ the identity*

$$\mathcal{N}(Cz, C\mathcal{G}, ||\cdot||) \leq \mathcal{N}(z, \mathcal{G}, ||\cdot||)$$

holds and thus $\mathcal{J}(Cz, C\mathcal{G}) \leq C\mathcal{J}(z, \mathcal{G})$, where $C\mathcal{G} = \{Cg | g \in \mathcal{G}\}$. This upper bound is not optimal (see Cucker and Smale, 2002) but sufficient for our purposes.

Eventually, we will show in Corollary 34 that if $\widehat{f}^{(m_{stop})}$ is in the ball of radius $h(N) \in o(N^{1/4})$, then this growth rate $h(N)$ is slow enough to ensure the convergence of (24). We rely on the following theorem.

Theorem 33 (van de Geer (2014, Theorem 3.2)) *Let $K = \sup_{f \in \mathcal{F}} ||f||_\infty$ and assume that Y is sub-Gaussian with Orlicz norm smaller than s. Then for t, N such that* [1] $\sqrt{\frac{2t}{N}} + \frac{t}{N} \leq 1$ *it holds with probability* $1 - 8\exp(-t)$

$$\sup_{f \in \mathcal{F}} |(P_N - P)Yf|/C \leq \frac{2\mathcal{J}(Ks, \mathcal{F}) + Ks\sqrt{t}}{\sqrt{N}}. \tag{25}$$

1. There is a typo in van de Geer (2014), where it says $J_0(K\sigma, \mathcal{F})$ instead of $J_\infty(K\sigma, \mathcal{F})$.

Corollary 34 *Let $f_1, f_2, \ldots$ be a sequence in H such that*

$$f_N \in h(N)\mathcal{F}_N,$$

where $h(N) \in o(N^{1/4})$. If $\mathcal{J}(z, \mathcal{F}_N) \leq \mathcal{J}(z, B_1) = C_0 z \int_0^1 \sqrt{\log\left(\mathcal{N}(\frac{uz}{2}, B_1)\right)} < \infty$ for all $z > 0$ as in Assumption 11′, then

$$|(P_N - P)Y f_N| \leq \xi.$$

converges to 0 in probability.

Proof For $h(N)\mathcal{F}_N$ it holds that

$$K := \sup_{f \in h(N)\mathcal{F}_N} ||f||_\infty \leq \sup_{f \in h(N)B_1} ||f||_\infty \leq Bh(N).$$

Recalling the definition of $\mathcal{J}(u, \mathcal{F})$ and applying Remark 32, we obtain

$$\mathcal{J}(Ku, h(N)\mathcal{F}_N) \leq \mathcal{J}(Bh(N)u, h(N)B_1) \leq h(N)\mathcal{J}(Bu, B_1)\, \forall u \in \mathbb{R}_+.$$

As the Orlicz norm fulfills the triangle inequality and as the Orlicz norm of the bounded random variable $\mu(\widetilde{\mathbf{X}})$ is finite, the Orlicz norm of $Y = \mu(\widetilde{\mathbf{X}}) + \varepsilon$, denoted by s, is bounded and Y is thus sub-Gaussian. We now apply Theorem 33 and set $t = N^{1/2}$ (the condition $\sqrt{\frac{2t}{N}} + \frac{t}{N} \leq 1$ is then fulfilled for $N > \frac{1}{2}\left(7 + 3\sqrt{5}\right)$).

It holds with probability $1 - 8\exp(-N^{1/2})$

$$\begin{aligned}
|(P_N - P)Y f_N| &\leq \sup_{f \in h(N)\mathcal{F}_N} |(P_N - P)Y f| \\
&\leq \frac{\mathcal{J}(Bh(N)s, h(N)\mathcal{F}_N) + Bh(N)s_{max}N^{1/4}}{\sqrt{N}} \\
&\leq \frac{h(N)\mathcal{J}(Bs, \mathcal{F}_N) + Bh(N)s_{max}N^{1/4}}{\sqrt{N}} \\
&\leq \frac{h(N)\mathcal{J}(Bs, B_1) + Bh(N)s_{max}N^{1/4}}{\sqrt{N}}.
\end{aligned}$$

$\mathcal{J}(Bs, B_1)$ is finite and constant in N by Assumption 11′. Let $\xi > 0$ be arbitrary. As $h(N) \in o(N^{1/4})$, there exists N^0 so that

$$\frac{h(N)\mathcal{J}(Bs, B_1)}{\sqrt{N}} \leq \frac{\xi}{2}\, \forall N > N^0.$$

For the second term observe that as $h(N) \in o(N^{1/4})$

$$\frac{Bh(N)sN^{1/4}}{\sqrt{N}} = \frac{Bh(N)s}{N^{1/4}} \leq \frac{\xi}{2}$$

for all $N > N^1$ for some $N^1 \in \mathbb{N}$. Thus for $N > \max\{N^0, N^1\}$

$$|(P_N - P)Y f_N| \leq \xi$$

with probability $1 - 8\exp(-8N^{1/2}) \to 1$. This proves the convergence in probability. ■

B.3 Results on Rademacher Complexities

In this section we use concept of the Rademacher complexity to show the convergence of

$$\left|(P - P_N)\left(\widehat{f}^{(m_{stop})}\right)^2\right|.$$

It is again based on Lemma 27, which ensures that for $m_{stop} = N^{\frac{1}{4}\frac{C_u+C_d+1/2}{C_d+1}}$, it holds $\widehat{f}^{(m_{stop})} \in h(N)\mathcal{F}_N$ for some $h(N) \in o(N^{1/4-\delta})$ for some $\delta > 0$ with probability going to 1 for $N \to \infty$.

Definition 35 (Rademacher complexity) *Let $\mathcal{F}$ be a function class on $\mathbf{X}$. Let $\sigma_1, \ldots, \sigma_N$ be i.i.d. realizations of Rademacher random variables, which are independent of $\mathbf{X}$. Further let $\mathbf{x}_1, \ldots, \mathbf{x}_N$ be i.i.d realizations of $\mathbf{X}$. We define the Rademacher complexity $R_N(\mathcal{F})$ by*

$$R_N(\mathcal{F}) = \frac{1}{N}\mathbb{E}_{\mathbf{X},\sigma}\left[\sup_{f\in\mathcal{F}}\left|\sum_{\ell=1}^{N}\sigma_\ell f(\mathbf{x}_\ell)\right|\right].$$

Note that $R_N(\mathcal{F})$ is deterministic. Given fixed observations $\mathbf{x}_1, \ldots, \mathbf{x}_N$, the empirical Rademacher complexity is given by

$$\widehat{R}_N(\mathcal{F}|\mathbf{x}_1, \ldots, \mathbf{x}_N) = \widehat{R}_N(\mathcal{F}) = \frac{1}{N}\mathbb{E}_{\sigma}\left[\sup_{f\in\mathcal{F}}\left|\sum_{\ell=1}^{N}\sigma_\ell f(\mathbf{x}_\ell)\right|\right].$$

The Rademacher complexity is a tool to upper bound $\sup_{f\in\mathcal{F}}|(P_N - P)f|$.

Theorem 36 *(Wainwright, 2019, Theorem 4.10) Let $\mathcal{F}$ be uniformly bounded with constant B. Then it holds for any $N \in \mathbb{N}$ and with probability $1 - \exp(-\frac{\varepsilon^2 N}{2B^2})$, that*

$$\sup_{f\in\mathcal{F}}|(P_N - P)f| \leq 2R_N(\mathcal{F}) + \varepsilon.$$

The Rademacher complexity is linked to its empirical counterpart with high probability.

Theorem 37 (Bartlett and Mendelson (2002, Theorem 11)) *Let $\mathcal{F}$ be uniformly bounded by B. Then it holds with probability $1 - 2\exp\left(-\frac{\varepsilon^2 N}{B^2}\right)$*

$$\left|\widehat{R}_N(\mathcal{F}) - R_N(\mathcal{F})\right| < \varepsilon.$$

We collect some helpful relationships for the Rademacher complexity.

Theorem 38 (Bartlett and Mendelson (2002, Theorem 12)) *Let $\mathcal{F}, \mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_p$ be function classes and let $\mathcal{F}$ be uniformly bounded by B. Then*

1. *for $c \in \mathbb{R}$: $R_N(c\mathcal{F}) = |c|R_N(\mathcal{F})$,*
2. *$R_N\left(\sum_{j=1}^{p}\mathcal{F}_j\right) \leq \sum_{j=1}^{p} R_N\left(\mathcal{F}_j\right)$,*
3. *for $\mathcal{F}^2 := \{|f|^2 : f \in \mathcal{F}\}$, it follows $R_N(\mathcal{F}^2) \leq 4BR_N(\mathcal{F})$.*

The Rademacher complexity can be upper bounded for kernel functions.

Theorem 39 (Bartlett and Mendelson (2002, Lemma 22)) *The empirical Rademacher complexity of $\mathcal{F}_N$ is upper bounded by*

$$\widehat{R}_N(\mathcal{F}_N) \leq \left(\frac{2}{N} \sum_{k=1}^{N} \widehat{\mu_k} \right)^{\frac{1}{2}},$$

while the population Rademacher complexity can be upper bounded by

$$R_N(\mathcal{F}_N) \leq \left(\frac{2}{N} \sum_{k=1}^{N} \mu_k \right)^{\frac{1}{2}}.$$

Note that if the sequence $\mu_1, \mu_2, \ldots$ is summable, then $R_N(\mathcal{F}_N) \in O(N^{-1/2})$. For this case we connect the results above in a corollary.

Corollary 40 *For any sequence $f_1, f_2, \ldots$ in H for which*

$$f_N \in h(N)\mathcal{F}_N,$$

where $h(N) \in o(N^{1/4-\delta})$ for some $\delta > 0$ and $\widehat{R}_N(\mathcal{F}_N) \in O(N^{-1/2})$, it holds

$$|(P_N - P)f_N^2| \xrightarrow{\mathbb{P}} 0 \text{ for } N \to \infty.$$

Proof Let $\xi > 0$ be arbitrary and fixed. By Theorem 38 it holds that

$$R_N(h(N)\mathcal{F}_N) = h(N)R_N(\mathcal{F}_N),$$

and as $\mathcal{F}_N$ is uniformly bounded by $Bh(N)$ it holds by Theorem 38

$$R_N\left((h(N)\mathcal{F}_N)^2\right) = 4Bh(N)R_N(\mathcal{F}_N).$$

From Theorem 36 we obtain that for any $\xi > 0$

$$|(P_N - P)f_N^2| \leq \sup_{f \in h(N)\mathcal{F}_N} \left|(P_N - P)f^2\right| \leq 2R_N\left((h(N)\mathcal{F}_N)^2\right) + \frac{\xi}{2} = 8Bh(N)^2 R_N(\mathcal{F}_N) + \frac{\xi}{3}$$

with probability $1 - 2\exp\left(\frac{\xi^2 N}{4(Bh(N))^2}\right)$, which converges to 1 as $h(N) \in o(N^{1/4})$ for $N \to \infty$. Similarly, as $\mathcal{F}_N$ is uniformly bounded by B and by setting $\varepsilon = N^{-1/2+2\delta}$ in Theorem 37, we observe that

$$|\widehat{R}_N(\mathcal{F}_N) - R_N(\mathcal{F}_N)| \leq N^{-1/2+2\delta}$$

with probability going to 1 for any $\delta > 0$. We conclude that

$$\begin{aligned}
|(P_N - P)f_N^2| &\leq \sup_{f \in h(N)\mathcal{F}_N} \left|(P_N - P)f^2\right| \\
&\leq 8Bh(N)^2 R_N(\mathcal{F}_N) + \frac{\xi}{3} \\
&\leq 8Bh(N)^2 |R_N(\mathcal{F}_N) - \widehat{R}_N(\mathcal{F}_N)| + 8Bh(N)^2 \widehat{R}_N(\mathcal{F}_N) + \frac{\xi}{3} \\
&\leq 8Bh(N)^2 N^{-1/2+2\delta} + 8Bh(N)^2 \widehat{R}_N(\mathcal{F}_N) + \frac{\xi}{3}
\end{aligned}$$

with probability going to 1 for $N \to \infty$. As $h(N) \in o(N^{1/4})$, it holds $h(N)^2 \widehat{R}_N(\mathcal{F}_N) < \frac{\xi}{3}$ by Assumption 13′ and Theorem 39 for N chosen large enough. Similarly, $h(N)^2 N^{-1/2+2\delta} < \frac{\xi}{3}$ for N chosen large enough, as $h(N) \in o\left(N^{1/4-\delta}\right)$. This proves the statement. ■

Appendix C. Proof of Theorem 19

Proof [Proof of Lemma 20] Using

$$
\begin{aligned}
||f^0 - \widehat{f}^{(m)}||_{2,N}^2 &= \frac{1}{N}||f^0(\widetilde{\mathbf{x}}^N) - B^{(m)} y^N||_2^2 \\
&= \frac{1}{N}||f^0(\widetilde{\mathbf{x}}^N) - U\left(I - (I-D)^m\right) U^\top y^N||_2^2 \\
&= \frac{1}{N}||f^0(\widetilde{\mathbf{x}}^N) - U\left(I - (I-D)^m\right) U^\top \left(f^0(\widetilde{\mathbf{x}}^N) + \varepsilon^N\right)||_2^2 \\
&\leq \frac{2}{N}||f^0(\widetilde{\mathbf{x}}^N) - U\left(I - (I-D)^m\right) U^\top f^0(\widetilde{\mathbf{x}}^N)||_2^2 \\
&\quad + \frac{2}{N}||U\left(I - (I-D)^m\right) U^\top \varepsilon^N)||_2^2 \\
&\leq \underbrace{\frac{2}{N}||U\left(I-D\right)^m U^\top f^0(\widetilde{\mathbf{x}}^N)||_2^2}_{\text{I}} + \underbrace{\frac{2}{N}||U\left(I - (I-D)^m\right) U^\top \varepsilon^N||_2^2}_{\text{II}},
\end{aligned}
$$

where the first inequality holds due to $||a+b||^2 \leq 2||a||^2 + ||b||^2$, we show the convergence of I and II to 0 in probability for $N \to \infty$. Recall that $S = UDU^T$ with U containing the orthonormal eigenvalues of S and D being a diagonal matrix with diagonal entries $D_{\ell\ell} = d_\ell = \frac{\widehat{\mu}_\ell}{\widehat{\mu}_\ell + \lambda}$, where $\widehat{\mu}_\ell, \ell = 1, \ldots, N$ are the eigenvalues of G.
Convergence of I:

$$
\begin{aligned}
\frac{2}{N}||U\left(I-D\right)^m U^\top f^0(\widetilde{\mathbf{x}}^N)||_2^2 &= \frac{2}{N} f^0(\widetilde{\mathbf{x}}^N)^\top U\left(I-D\right)^m U^\top U\left(I-D\right)^m U^\top f^0(\widetilde{\mathbf{x}}^N) \\
&= \frac{2}{N}\sum_{\ell=1}^N (1-d_\ell)^{2m}\left(U^\top f^0(\widetilde{\mathbf{x}}^N)\right)_\ell^2
\end{aligned}
$$

As G has full rank, there exists a $\beta \in \mathbb{R}^N$ such that $f^0(\widetilde{\mathbf{x}}^{\mathbf{N}}) = \sqrt{N} G\beta$. Define $\widetilde{f} := \frac{1}{\sqrt{N}}\sum_{\ell=1}^N \beta_\ell K(\cdot, \widetilde{\mathbf{x}}_\ell)$ for which holds $\widetilde{f}(\widetilde{\mathbf{x}}^N) = f^0(\widetilde{\mathbf{x}}^N)$. By the representation theorem it further holds $\beta^\top G\beta = ||\widetilde{f}||_H^2 \leq ||f^0||_H^2 = R^2$. Let $D_{\widehat{\mu}} \in \mathbb{R}^{N\times N}$ be the diagonal matrix with

diagonal entries $\widehat{\mu}_1, \ldots, \widehat{\mu}_N$. Then,

$$\begin{aligned}
\frac{2}{N}\sum_{\ell=1}^{N}(1-d_\ell)^{2m}\left(U^\top f^0(\widetilde{\mathbf{x}}^N)\right)_\ell^2 &\leq \frac{2}{N}\sum_{\ell=1}^{N}\frac{\left(U^\top f^0(\widetilde{\mathbf{x}}^N)\right)_\ell^2}{2e\,m\,d_\ell} \\
&= \frac{1}{N}\sum_{\ell=1}^{N}\frac{\left(U^\top\sqrt{N}G\beta\right)_\ell^2}{e\,m\,d_\ell} \\
&= \sum_{\ell=1}^{N}\frac{\left(U^\top G\beta\right)_\ell^2}{e\,m\,d_\ell} \\
&\leq (1+\lambda)\sum_{\ell=1}^{N}\frac{\left(U^\top G\beta\right)_\ell^2}{e\,m\,\widehat{\mu}_\ell} \\
&= \frac{1+\lambda}{e\,m}tr\left(D_{\widehat{\mu}}^{-1}U^\top G\beta\beta^\top GU\right) \\
&= \frac{1+\lambda}{e\,m}tr\left(\underbrace{UD_{\widehat{\mu}}^{-1}U^\top}_{G^{-1}}G\beta\beta^\top G\right) \\
&= \frac{1+\lambda}{e\,m}\beta^\top G\beta = \frac{1+\lambda}{e\,m}||\widetilde{f}||_H^2 \leq \frac{1+\lambda}{e\,m}R^2,
\end{aligned}$$

which goes to 0 as $m(N) \to \infty$ for $N \to \infty$. In the first inequality we have used the fact that $(1-x)^{2m} \leq \exp(-x)^{2m} = \exp(-2mx) \leq \frac{1}{2e\,m\,x}$ for all $x \in \mathbb{R}$, where e is Euler's number. In the second inequality we have used $\frac{1}{d_\ell} = \frac{\widehat{\mu}_\ell+\lambda}{\widehat{\mu}_\ell} \leq \frac{1+\lambda}{\widehat{\mu}_\ell}$, as $0 < \widehat{\mu}_\ell \leq 1$ for all $\ell = 1, 2, \ldots, N$.
Convergence of II:

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{N}||U\left(I-(I-D)^m\right)U^\top\varepsilon^N||_2^2 > \xi\right) \\
&\leq \mathbb{P}\left(\left(\frac{1}{N}||U\left(I-(I-D)^m\right)U^\top\varepsilon^N||_2^2 > \xi\right) \cap \left\{\widetilde{\mathbf{X}}^N \in \mathcal{B}_N\right\}\right) + \mathbb{P}\left(\left\{\widetilde{\mathbf{X}}^N \notin \mathcal{B}_N\right\}\right)
\end{aligned}$$

The latter term goes again to 0 by Assumption 13′. For any $\widetilde{\mathbf{x}}^N \in \mathcal{B}_N$ we show that

$$\mathbb{P}\left(\frac{1}{N}||U\left(I-(I-D)^m\right)U^\top\varepsilon^N||_2^2 > \xi \mid \widetilde{\mathbf{X}}^N = \widetilde{\mathbf{x}}^N\right) \to 0$$

for any $\xi > 0$ and uniformly in $\widetilde{\mathbf{x}}^N$. Recall that $\varepsilon^N \mid \widetilde{\mathbf{X}}^N = \widetilde{\mathbf{x}}^N$ is a sub-Gaussian vector with mean 0 and independent components. Thus, we can apply the Hanson-Wright inequality of Theorem 25. Remember that D is a function of $\widetilde{\mathbf{x}}^N$. We need to show for any $\widetilde{\mathbf{x}}^N \in \mathcal{B}_N$, $A_m = \frac{1}{\sqrt{N}}U\left(I-(I-D)^m\right)U^\top$ that $\mathbb{E}\left[\left|\left|A_m\varepsilon^N\right|\right|_2^2 \mid \widetilde{\mathbf{X}}^{\mathbf{N}} = \widetilde{\mathbf{x}}^N\right] \to 0$, $||A_m^2||_F^2 \to 0$ and

$||A_m^2|| \to 0$. It holds by Lemma 30 uniformly for $\widetilde{\mathbf{x}}^N \in \mathcal{B}_N$

$$\begin{aligned}
\mathbb{E}\left[\left|\left|A_m\varepsilon^N\right|\right|_2^2 | \widetilde{\mathbf{X}} = \widetilde{\mathbf{x}}^N\right] &= \mathbb{E}_{Y^N|\widetilde{\mathbf{X}}^N=\widetilde{\mathbf{x}}^N}\left[\left|\left|\frac{1}{\sqrt{N}}U\left(I-(I-D)^m\right)U^\top\varepsilon^N\right|\right|_2^2\right] \\
&= \mathbb{E}_{Y^N|\widetilde{\mathbf{X}}^N=\widetilde{\mathbf{x}}^N}\left[\frac{1}{N}tr\left(\left(\varepsilon^N\right)^\top U\left(I-(I-D)^m\right)^2U^\top\varepsilon^N\right)\right] \\
&\leq \mathbb{E}_{Y^N|\widetilde{\mathbf{X}}^N=\widetilde{\mathbf{x}}^N}\left[\frac{1}{N}tr\left(\varepsilon^N\left(\varepsilon^N\right)^\top\right)tr\left(U\left(I-(I-D)^m\right)^2U^\top\right)\right] \\
&\leq \frac{\sigma^2}{N}tr\left(\left(I-(I-D)^m\right)^2\right) \\
&= \frac{\sigma^2}{N}\sum_{\ell=1}^N\left(1-(1-d_\ell)^m\right)^2 \\
&\leq \frac{\sigma^2}{N}\sum_{\ell=1}^N\left(1-\left(1-\frac{\widehat{\mu}_\ell}{\lambda+\widehat{\mu}_\ell}\right)^m\right)^2 \to 0
\end{aligned}$$

for $N \to \infty$. Further,

$$||A_m^2||_F^2 = \frac{1}{N^2}tr\left(\left(I-(I-D)^m\right)^4\right) \leq \left(\frac{1}{N}tr\left(\left(I-(I-D)^m\right)^2\right)\right)^2,$$

which goes to 0 with the same calculation as above. Finally,

$$||A_m^2|| \leq \frac{1}{N} \to 0.$$

The statement follows by integrating out $\mathcal{B}_N$ with respect to $\widetilde{\mathbf{x}}^N$.

■

Appendix D. Algorithm

Data: $\mathbf{x}^N = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^\top \in \mathbb{R}^{N \times p}$
Input: Kernels $K_1, \ldots, K_p$ implying RKHSs $H_1, \ldots, H_p$, penalty λ, step size ν
Output: $\widehat{G}, \widehat{F}$
$F^{(1)} \leftarrow 0$
$\widehat{G} \leftarrow \emptyset$
$N \leftarrow \emptyset$
$\widehat{f}_{kj} = \arg\min_{g_{kj} \in H_j} \sum_{\ell=1}^N \left(g_{kj}(\mathbf{x}_{\ell j}) - \mathbf{x}_{\ell k}\right)^2 + \lambda ||g_{kj}||^2_{H_j}, j, k = 1, \ldots, p, j \neq k$
$S(j,k) \leftarrow \log\left(\sum_{\ell=1}^N \left(\widehat{f}_{kj}(\mathbf{x}_{\ell j}) - \mathbf{x}_{\ell k}\right)^2\right)$
for $m \leftarrow 2$ **to** m_{stop} **do**
  // Find the next edge and update graph
  $(j^0, k^0) \leftarrow \arg\min_{(j,k) \notin N} S(j, k, F^{(m-1)})$; $\widehat{G} \leftarrow \widehat{G} + (j^0, k^0)$
  $f_{k^0}^{(m)} \leftarrow f_{k^0}^{(m-1)} + \nu \widehat{f}_{k^0 j^0}$
  $f_k^{(m)} \leftarrow f_k^{(m-1)}, k = 1, \ldots, k^0 - 1, k^0 + 1, \ldots, p$
  $F^{(m)} \leftarrow \left(f_1^{(m)}, \ldots, f_p^{(m)}\right)$
  // Check if AIC score has increased
  if $AIC(F^{(m)}, \mathbf{x}^N) > AIC(F^{(m-1)}, \mathbf{x}^N)$ **then** break
  // Update edges that cause cycle and ensure they are not chosen anymore
  $N \leftarrow getForbiddenEdges(\widehat{G})$
  // Update S
  $\widehat{f}_{k^0 j} \leftarrow$ Equation (21) , $j = 1, \ldots, p$
  $S(j, k^0, F^{(m)}) \leftarrow$ Equation 20, $j = 1, \ldots, p$
end
return $\widehat{G}, F^{(m)}$

References

Raj Agrawal, Chandler Squires, Neha Prasad, and Caroline Uhler. The decamfounder: nonlinear causal discovery in the presence of hidden variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(5):1639–1658, 07 2023. ISSN 1369-7412. doi: 10.1093/jrsssb/qkad071. URL `https://doi.org/10.1093/jrsssb/qkad071`.

Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman, Vân Anh Huynh-Thu, Hana Imrichova, Gert Hulselmans, Florian Rambow, Jean-Christophe Marine, Pierre Geurts, Jan Aerts, et al. Scenic: single-cell regulatory network inference and clustering. *Nature methods*, 14(11):1083–1086, 2017.

Francis R. Bach and Michael I. Jordan. Kernel independent component analysis. *The Journal of Machine Learning Research*, 3(Jul):1–48, 2002.

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999. doi: 10.1126/science.286.5439.509.

Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Rohit Bhattacharya, Tushar Nagarajan, Daniel Malinsky, and Ilya Shpitser. Differentiable causal discovery under unmeasured confounding. In Arindam Banerjee and Kenji Fukumizu, editors, *International Conference on Artificial Intelligence and Statistics*, volume 130, pages 2314–2322. The Proceedings of Machine Learning Research, 13–15 Apr 2021.

Peter Bühlmann and Torsten Hothorn. Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, 22(4):477–505, 2007. doi: 10.1214/07-STS242.

Peter Bühlmann and Bin Yu. Boosting with the l2 loss. *Journal of the American Statistical Association*, 98(462):324–339, 2003. doi: 10.1198/016214503000125. URL `https://doi.org/10.1198/016214503000125`.

Peter Bühlmann, Jonas Peters, and Jan Ernest. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014. doi: 10.1214/14-AOS1260.

Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1):1–49, 2002.

Paul Erdös and Alfréd Rényi. On random graphs i. *Publicationes Mathematicae Debrecen*, 6:290, 1959.

Hawoong Jeong, Bálint Tombor, Réka Albert, Zoltan N Oltvai, and A-L Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000.

Marcus Kaiser and Maksim Sipos. Unsuitability of NOTEARS for causal graph discovery when dealing with dimensional quantities. *Neural Processing Letters*, 54(3):1587–1595, 2022.

Diviyan Kalainathan, Olivier Goudet, Isabelle Guyon, David Lopez-Paz, and Michale Sebag. Structural agnostic modeling: Adversarial learning of causal graphs. *The Journal of Machine Learning Research*, 23(219):1–62, 2022.

Markus Kalisch and Peter Bühlman. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *The Journal of Machine Learning Research*, 8(3):613–636, 2007.

Kirthevasan Kandasamy and Yaoliang Yu. Additive approximations in high dimensional nonparametric regression via the salsa. In *International Conference on Machine Learning*, pages 69–78. The Proceedings of Machine Learning Research, 2016.

Maximilian Kertel, Stefan Harmeling, Markus Pauly, and Nadja Klein. Learning causal graphs in manufacturing domains using structural equation models. *International Journal of Semantic Computing*, 17(04):511–528, 2023. doi: 10.1142/S1793351X23630023.

Trent Kyono, Yao Zhang, and Mihaela van der Schaar. CASTLE: Regularization via auxiliary causal graph discovery. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1501–1512. Curran Associates, Inc., 2020.

Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural DAG learning. In *International Conference on Learning Representations*, 2019.

Ha Quang Minh. Some properties of Gaussian reproducing kernel Hilbert spaces and their implications for function approximation and learning theory. *Constructive Approximation*, 32:307–338, 2010.

Ignavier Ng, Sébastien Lachapelle, Nan Rosemary Ke, Simon Lacoste-Julien, and Kun Zhang. On the convergence of continuous constrained optimization for structure learning. In *International Conference on Artificial Intelligence and Statistics*, pages 8176–8198. The Proceedings of Machine Learning Research, 2022a.

Ignavier Ng, Shengyu Zhu, Zhuangyan Fang, Haoyang Li, Zhitang Chen, and Jun Wang. Masked gradient-based causal structure learning. In *International Conference on Data Mining (SDM)*, pages 424–432. SIAM, 2022b.

Christopher Nowzohour and Peter Bühlmann. Score-based causal learning in additive noise models. *Statistics*, 50(3):471–485, 2016. doi: 10.1080/02331888.2015.1060237. URL `https://doi.org/10.1080/02331888.2015.1060237`.

Jonas Peters and Peter Bühlmann. Structural intervention distance for evaluating causal graphs. *Neural Comput.*, 27(3):771–799, March 2015. ISSN 0899-7667. doi: 10.1162/NECO_a_00708. URL `https://doi.org/10.1162/NECO_a_00708`.

Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15 (1):2009–2053, 01 2014. ISSN 1532–4435.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, USA, 2017.

Garvesh Raskutti and Caroline Uhler. Learning directed acyclic graph models based on sparsest permutations. *Stat*, 7(1):e183, 2018. doi: https://doi.org/10.1002/sta4.183. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/sta4.183`. e183 sta4.183.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *The Journal of Machine Learning Research*, 15(1):335–366, 2014.

Alexander G. Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated dag! causal discovery benchmarks may be easy to game. *Advances in Neural Information Processing Systems*, 34, 2021.

Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-Gaussian concentration. *Electronic Communications in Probability*, 18:1–9, 2013. doi: 10.1214/ECP.v18-2865.

Robert E. Schapire and Yoav Freund. *Boosting: Foundations and Algorithms*. The MIT Press, 2012. ISBN 0262017180.

Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001. ISBN 0262194759.

Rajen Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *Annals of Statistics*, 48(3):1514–1538, 2018.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, and Kenneth Bollen. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *The Journal of Machine Learning Research*, 12:1225–1248, 2011.

Ali Shojaie and George Michailidis. Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika*, 97(3):519–538, 07 2010.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*, volume 81. MIT Press, 01 1993. ISBN 978-1-4612-7650-0. doi: https://doi.org/10.7551/mitpress/1754.001.0001.

Bernhard Stankewitz. Early stopping for L2-boosting in high-dimensional linear models. *Annals of Statistics*, 52(2):491–518, 2024. doi: 10.1214/24-AOS2356.

Hongwei Sun. Mercer theorem for RKHS on noncompact sets. *Journal of Complexity*, 21 (3):337–349, 2005. ISSN 0885-064X.

Marc Teyssier and Daphne Koller. Ordering-based search: A simple and effective algorithm for learning bayesian networks. In *Conference on Uncertainty in Artificial Intelligence*, pages 584–590, Arlington, Virginia, USA, 2005. AUAI Press. ISBN 0974903914.

Gerhard Tutz and Harald Binder. Generalized additive modeling with implicit variable selection by likelihood-based boosting. *Biometrics*, 62(4):961–971, 2006.

Sara van de Geer. On the uniform convergence of empirical norms and inner products, with application to causal inference. *Electronic Journal of Statistics*, 8(1):543–574, 2014.

Matthew J. Vowels, Necati Cihan Camgoz, and Richard Bowden. D'ya like DAGs? a survey on structure learning and causal discovery. *ACM Computing Surveys*, 55(4), 11 2022.

Grace Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, 1990. doi: 10.1137/1.9781611970128.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge, 2019.

Anja Wille, Philip Zimmermann, Eva Vranová, Andreas Fürholz, Oliver Laule, Stefan Bleuler, Lars Hennig, Amela Prelić, Peter von Rohr, Lothar Thiele, et al. Sparse graphical gaussian modeling of the isoprenoid gene network in arabidopsis thaliana. *Genome biology*, 5(11):1–13, 2004.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. DAG-GNN: DAG structure learning with graph neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *International Conference on Machine Learning*, volume 97 of *The Proceedings of Machine Learning Research*, pages 7154–7163. PMLR, 09–15 Jun 2019.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Conference on Uncertainty in Artificial Intelligence*, pages 804–813, Arlington, Virginia, USA, 2011. AUAI Press. ISBN 9780974903972.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P Xing. DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, volume 32, 2018.

Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Learning sparse nonparametric dags. In *International Conference on Artificial Intelligence and Statistics*, pages 3414–3425. The Proceedings of Machine Learning Research, 2020.

Bruno Peter Zwahlen. Über die Eigenwerte der Summe zweier selbstadjungierter Operatoren. *Commentarii Mathematici Helvetici*, 40:81–116, 1965/66.