Journal of Machine Learning Research 18 (2017) 1-41 Submitted 6/17; Revised 10/17; Published 11/17

Kernel Partial Least Squares for Stationary Data

Marco Singer MSINGER@GWDG.DE
Institute for Mathematical Stochastics
Georg-August-Universität
Göttingen, 37077, Germany

Tatyana Krivobokova TKRIVOB@GWDG.DE
Institute for Mathematical Stochastics
Georg-August-Universität
Göttingen, 37077, Germany

Axel Munk AMUNK1@GWDG.DE
Institute for Mathematical Stochastics
Georg-August-Universität
Göttingen, 37077, Germany

Editor: Sara van de Geer

Abstract

We consider the kernel partial least squares algorithm for non-parametric regression with stationary dependent data. Probabilistic convergence rates of the kernel partial least squares estimator to the true regression function are established under a source and an effective dimensionality condition. It is shown both theoretically and in simulations that long range dependence results in slower convergence rates. A protein dynamics example shows high predictive power of kernel partial least squares.

Keywords: effective dimensionality, long range dependence, nonparametric regression, source condition, protein dynamics

1. Introduction

Partial least squares (PLS) is a regularized regression technique developed by Wold et al. (1984) to deal with collinearities in the regressor matrix. It is an iterative algorithm where the covariance between response and regressor is maximized at each step, see Helland (1988) for a detailed description. Regularization in the PLS algorithm is obtained by stopping the iteration process early.

Several studies showed that partial least squares algorithm is competitive with other regression methods such as ridge regression and principal component regression, needing generally fewer iterations than the latter to achieve comparable estimation and prediction, see, e.g., Frank and Friedman (1993), Krämer and Braun (2007) and Singer et al. (2016). For an overview of further properties of PLS we refer to Rosipall and Krämer (2006).

Reproducing kernel Hilbert spaces (RKHS) have a long history in probability and statistics (see, e.g., Berlinet and Thomas-Agnan, 2004). Here we focus on the supervised kernel based learning approach for the solution of non-parametric regression problems. RKHS methods are both computationally and theoretically attractive, due to the kernel trick

(Schölkopf et al., 1998) and the representer theorem (Wahba, 1999), as well as its generalization (Schölkopf et al., 2001). Within the reproducing kernel Hilbert space framework one can adapt linear regularized regression techniques like ridge regression and principal component regression to a non-parametric setting, see Saunders et al. (1998) and Rosipal et al. (2000), respectively. We refer to Schölkopf and Smola (2001) for more details on the kernel based learning approach.

Kernel PLS (KPLS) was introduced in Rosipal and Trejo (2001) by using the reformulation of the PLS algorithm of Lindgren et al. (1993). The relationship to kernel conjugate gradient (KCG) methods was highlighted in Blanchard and Krämer (2010a). It can be seen in Hanke (1995) that conjugate gradient methods are well suited for handling ill-posed problems, as they arise in kernel learning, see, e.g., De Vito et al. (2006).

Rosipal (2003) investigated the performance of kernel partial least squares for non-linear discriminant analysis. Blanchard and Krämer (2010a) proved the consistency of KPLS when the algorithm is stopped early without giving convergence rates.

Caponnetto and de Vito (2007) showed that kernel ridge regression (KRR) attains optimal probabilistic rates of convergence for independent and identically distributed data, using a source and a polynomial effective dimensionality condition. A generalization of these results to a wider class of effective dimensionality conditions and extension to kernel principal component regression can be found in Dicker et al. (2017).

For independent identically distributed data Blanchard and Krämer (2010b) obtained probabilistic convergence rates for a certain kernel conjugate gradient estimator under early stopping, while Lin and Zhou (2017) considered kernel partial least squares estimators with the cross-validation stopping rule.

In contrast to existing works, we derive probabilistic convergence rates of the kernel partial least squares estimator to the true regression function when the input data are not independent and identically distributed, but rather stationary time series. To the best of our knowledge, none of the kernel regression methods have been considered for the dependent data so far. Our results can be applied to stationary dependence structures, given that certain concentration inequalities for these data hold. The derived convergence rates depend not only on the complexity of the target function and of the data mapped into the kernel space, but also on the persistence of the dependence in the data. For measuring the complexity of the data we consider general effective dimensionality conditions. In a Gaussian setting we prove that the short range dependence still leads to optimal rates, but if the dependence is more persistent, the rates become slower. We illustrate the good predictive performance of KPLS by an application to the molecular dynamics of a bacteriophage protein.

2. Kernel Partial Least Squares

Consider the non-parametric regression problem

$$y_t = f^*(X_t) + \varepsilon_t, \quad t \in \mathbb{Z}. \tag{1}$$

Here $\{X_t\}_{t\in\mathbb{Z}}$ is a d-dimensional, $d \in \mathbb{N}$, stationary time series on a probability space $(\Omega, \mathcal{A}, \mathrm{P})$ and $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is an independent and identically distributed sequence of real valued random variables with expectation zero and variance $\sigma^2 > 0$ that is independent of $\{X_t\}_{t\in\mathbb{Z}}$.

Let X be a random vector that is independent of $\{X_t\}_{t\in\mathbb{Z}}$ and $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ with the same distribution as X_0. The target function we seek to estimate is $f^* \in \mathcal{L}^2\left(\mathrm{P}^X\right)$.

For the purpose of supervised learning assume that we have a training sample $\{(X_t, y_t)\}_{t=1}^n$ for some $n \in \mathbb{N}$. In the following we introduce some basic notation for the kernel based learning approach.

Define with $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ the RKHS of functions on $\mathbb{R}^d$ with reproducing kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, i.e., it holds

$$g(x) = \langle g, k(\cdot, x)\rangle_{\mathcal{H}}, \quad x \in \mathbb{R}^d, g \in \mathcal{H}. \tag{2}$$

The corresponding inner product and norm in $\mathcal{H}$ is denoted by $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ and $\|\cdot\|_{\mathcal{H}}$, respectively. We refer to Berlinet and Thomas-Agnan (2004) for examples of Hilbert spaces and their reproducing kernels. In the following we deal with reproducing kernel Hilbert spaces which fulfill the following, rather standard, conditions:

(K1) $\mathcal{H}$ is separable,

(K2) There exists a $\kappa > 0$ such that $|k(x, y)| \leq \kappa$ for all $x, y \in \mathbb{R}^d$ and k is measurable.

Under (K1) the Hilbert-Schmidt norm $\|\cdot\|_{\mathrm{HS}}$ for operators mapping from $\mathcal{H}$ to $\mathcal{H}$ is well defined. If condition (K2) holds, all functions in $\mathcal{H}$ are bounded, see Berlinet and Thomas-Agnan (2004), chapter 2. The conditions are satisfied for a variety of popular kernels, e.g., Gaussian or triangular.

The main principle of RKHS methods is the mapping of the data X_t into $\mathcal{H}$ via the feature maps $\phi_t = k(\cdot, X_t)$, $t = 1, \ldots, n$. This mapping can be done implicitly by using the kernel trick $\langle\phi_t, \phi_s\rangle_{\mathcal{H}} = k(X_t, X_s)$ and thus only the $n \times n$ dimensional kernel matrix $K_n = n^{-1}[k(X_t, X_s)]_{t,s=1}^n$ is needed in the computations. Then the task for RKHS methods is to find coefficients $\alpha_1, \ldots, \alpha_n$ such that $f_\alpha = \sum_{t=1}^n \alpha_t\phi_t$ is an adequate approximation of f^* in $\mathcal{H}$, measured in the $\mathcal{L}^2\left(\mathrm{P}^X\right)$ norm $\|\cdot\|_2$.

There are a variety of different approaches to estimate the coefficients $\alpha_1, \ldots, \alpha_n$, including kernel ridge regression, kernel principal component regression and, of course, kernel partial least squares. The latter method was introduced by Rosipal and Trejo (2001) and is the focus of the current work.

It was shown by Krämer and Braun (2007) that the KPLS algorithm solves

$$\widehat{\alpha}_i = \arg\min_{v\in\mathcal{K}_i(K_n,y)} n^{-1}\|y - K_n v\|^2, \quad i = 1, \ldots, n, \tag{3}$$

with $y = (y_1, \ldots, y_n)^{\mathrm{T}}$. Here $\mathcal{K}_i(K_n, y) = \operatorname{span}\left\{y, K_n y, K_n^2 y, \ldots, K_n^{i-1} y\right\}$, $i = 1, \ldots, n$, is the ith order Krylov space with respect to K_n and y and $\|\cdot\|$ denotes the Euclidean norm. The dimension i of the Krylov space is the regularization parameter for KPLS.

We will introduce several operators that will be crucial for our further analysis. Define two integral operators: the kernel integral operator $T^* : \mathcal{L}^2\left(\mathrm{P}^X\right) \to \mathcal{H}, g \mapsto \mathrm{E}\{k(\cdot, X)g(X)\}$ and the change of space operator $T : \mathcal{H} \to \mathcal{L}^2\left(\mathrm{P}^X\right), g \mapsto g$, which is well defined if (K2) holds. It is easy to see that T, T^* are adjoint, i.e., for $u \in \mathcal{H}$ and $v \in \mathcal{L}^2\left(\mathrm{P}^X\right)$ it holds $\langle T^* v, u\rangle_{\mathcal{H}} = \langle v, Tu\rangle_2$ with $\langle\cdot,\cdot\rangle_2$ being the inner product in $\mathcal{L}^2\left(\mathrm{P}^X\right)$.

The sample analogues of T, T^* are $T_n : \mathcal{H} \to \mathbb{R}^n, g \mapsto \{g(X_1), \ldots, g(X_n)\}^{\mathrm{T}}$ and $T_n^* : \mathbb{R}^n \to \mathcal{H}, (v_1, \ldots, v_n)^{\mathrm{T}} \mapsto n^{-1}\sum_{t=1}^n v_t k(\cdot, X_t)$, respectively. Both operators are adjoint with respect to the rescaled Euclidean product $n^{-1}u^{\mathrm{T}}v$, $u, v \in \mathbb{R}^d$

Finally, we define the sample kernel covariance operator $S_n = T_n^* T_n : \mathcal{H} \to \mathcal{H}$ and the population kernel covariance operator $S = T^* T : \mathcal{H} \to \mathcal{H}$. Note that it holds $K_n = T_n T_n^*$. Under (K1) and (K2) S is a self-adjoint compact operator with operator norm $\|S\|_{\mathcal{L}} \leq \kappa$, see Caponnetto and de Vito (2007).

With this notation we can restate (3) for the function f_α

$$f_{\widehat{\alpha}_i} = \arg \min_{g \in \mathcal{K}_i(S_n, T_n^* y)} n^{-1} \|y - \{g(X_1), \ldots, g(X_n)\}^{\mathrm{T}}\|^2 = \arg \min_{g \in \mathcal{K}_i(S_n, T_n^* y)} n^{-1} \|y - T_n g\|^2. \quad (4)$$

Hence, we are looking for functions that minimize the squared distance to y constrained to a sequence of Krylov spaces.

In the literature of ill-posed problems it is well known that without further conditions on the target function f^* the convergence rate of the conjugate gradient algorithm can be arbitrarily slow, see Hanke (1995), chapter 3.2. One common a-priori assumption on the regression function f^* is a source condition:

(S) There exist $r \geq 0$, $R > 0$ and $u \in \mathcal{L}^2\left(\mathrm{P}^X\right)$ such that $f^* = (TT^*)^r u$ and $\|u\|_2 \leq R$.

If $r \geq 1/2$, then the target function $f^* \in \mathcal{L}^2\left(\mathrm{P}^X\right)$ coincides almost surely with a function $f \in \mathcal{H}$ and we can write $f^* = Tf$, see Cucker and Smale (2002). With this the kernel partial least squares estimator $f_{\widehat{\alpha}_i}$ estimates the correct target function, not only its best approximation in $\mathcal{H}$. This case is known as the inner case.

The situation with $r < 1/2$ is referred to as the outer case. Under additional assumptions, e.g., the availability of additional unlabeled data, it is still possible that an estimator of f^* converges to the true target function in $\mathcal{L}^2\left(\mathrm{P}^X\right)$ norm with optimal rates (with respect to the number n of labeled data points). See De Vito et al. (2006) for a detailed description of this semi-supervised approach for kernel ridge regression in the independent and identically distributed case. We do not treat the case $r < 1/2$ in this work.

A source condition is often interpreted as an abstract smoothness condition, see, e.g., Bissantz et al. (2007) for several examples. This can be seen as follows. Let $\eta_1 \geq \eta_2 \geq \ldots$ be the eigenvalues and $\psi_1, \psi_2, \ldots$ the corresponding eigenfunctions of the compact operator S. Then it is easy to see that the source condition (S) is equivalent to $f = \sum_{j=1}^{\infty} b_j \psi_j \in \mathcal{L}^2\left(\mathrm{P}^X\right)$ with b_j such that $\sum_{j=1}^{\infty} \eta_j^{-2(r+1/2)} b_j^2 < \infty$. Hence, the higher r is chosen the faster the sequence $\{b_j\}_{j=1}^{\infty}$ must converge to zero. Therefore, the sets of functions for which source conditions hold are nested, i.e., the larger r is the smaller the corresponding set will be. The set with $r = 1/2$ is the largest one and corresponds to a zero smoothness condition, i.e., $\sum_{j=1}^{\infty} \eta_j^{-2} b_j^2 < \infty$, which is equivalent to $f \in \mathcal{H}$. For more details we refer to Dicker et al. (2017).

3. Consistency of Kernel Partial Least Squares

The KCG algorithm as described by Blanchard and Krämer (2010b) is consistent when stopped early and convergence rates can be obtained when a source condition (S) holds. Here we will proof the same property for KPLS. Early stopping in this context means that we stop the algorithm at some $a = a(n) \leq n$ and consider the estimator $f_{\widehat{\alpha}_a}$ for f^*.

The difference between KCG and KPLS is the norm which is optimized. The kernel conjugate gradient algorithm studied in Blanchard and Krämer (2010b) estimates the coefficients $\alpha \in \mathbb{R}^n$ of f_α via $\widehat{\alpha}_i^{CG} = \arg\min_{v \in \mathcal{K}_i(K_n, y)} \langle y - K_n v, K_n(y - K_n v) \rangle$. It is easy to

see that this optimization problem can be rewritten for the function f_α as

$$\min_{g \in \mathcal{K}_i(S_n, T_n^* y)} n^{-1} \|T_n^* y - S_n g\|_{\mathcal{H}}^2 = \min_{g \in \mathcal{K}_i(S_n, T_n^* y)} n^{-1} \|T_n^* \left(y - T_n g\right)\|_{\mathcal{H}}^2,$$

compared to (4) for KPLS. Thus, KCG obtains the least squares approximation g in the $\mathcal{H}$-norm for the normal equation $T_n^* y = T_n^* T_n g$ and KPLS finds a function that minimizes the residual sum of squares. In both methods the solutions are restricted to functions $g \in \mathcal{K}_i(S_n, T_n^* y)$.

An advantage of the kernel conjugate gradient estimator is that concentration inequalities can be established for both $T_n^* y$ and S_n and applied directly as the optimization function contains both quantities. The stopping index for the regularization can be chosen by a discrepancy principle as $a^* = \min\{1 \leq i \leq n : \|S_n f_{\widehat{\alpha}_i^{CG}} - T_n^* y\| \leq \Lambda_n\}$ with Λ_n being a threshold sequence that goes to zero as n increases.

On the other hand, the function to be optimized for KPLS contains only y and $T_n g = \{g(X_1), \ldots, g(X_n)\}^{\mathrm{T}}$ for which statistical properties are not readily available. Thus, we need to find a way to apply the concentration inequalities for $T_n^* y$ and S_n to this slightly different problem. This leads to complications in the proof of consistency and a rather different and more technical stopping rule for choosing the optimal regularization parameter a^* is used, as can be seen in Theorem 1. This stopping rule has its origin in Hanke (1995).

In the following $\|\cdot\|_{\mathcal{L}}$ denotes the operator norm and $\|\cdot\|_{\mathrm{HS}}$ is the Hilbert-Schmidt norm.

Theorem 1 *Assume that conditions (K1), (K2), (S) hold with $r \geq 3/2$ and there are constants $C_\delta(\nu), C_\epsilon(\nu) > 0$ and a sequence $\{\gamma_n\}_{n \in \mathbb{N}} \subset [0, \infty)$, $\gamma_n \to 0$, such that we have for $\nu \in (0, 1]$*

$$\begin{aligned} \mathrm{P}\left(\|S_n - S\|_{\mathcal{L}} \leq C_\delta(\nu)\gamma_n\right) &\geq 1 - \nu/2, \\ \mathrm{P}\left(\|T_n^* y - Sf\|_{\mathcal{H}} \leq C_\epsilon(\nu)\gamma_n\right) &\geq 1 - \nu/2. \end{aligned}$$

Define the stopping index a^* by

$$a^* = \min\left\{1 \leq a \leq n : \sum_{i=0}^{a} \|S_n f_{\widehat{\alpha}_i} - T_n^* y\|_{\mathcal{H}}^{-2} \geq (C\gamma_n)^{-2}\right\}, \tag{5}$$

with $C = C_\epsilon(\nu) + \kappa^{r-1/2}(r + 1/2)R\{1 + C_\delta(\nu)\}$.

Then it holds with probability at least $1 - \nu$ that

$$\begin{aligned} \|f_{\widehat{\alpha}_{a^*}} - f^*\|_2 &= O\left\{\gamma_n^{2r/(2r+1)}\right\}, \\ \|f_{\widehat{\alpha}_{a^*}} - f\|_{\mathcal{H}} &= O\left\{\gamma_n^{(2r-1)/(2r+1)}\right\}, \end{aligned}$$

with $f^* = Tf$.

It can be shown that the stopping rule (5) always determines a finite index, i.e., the set the minimum is taken over is not empty, see Hanke (1995), chapter 4.3.

The theorem yields two convergence results, one in the $\mathcal{H}$-norm and one in the $\mathcal{L}^2\left(\mathrm{P}^X\right)$-norm. It holds that $\|v\|_2 = \|S^{1/2}v\|_{\mathcal{H}}$. These are the endpoints of a continuum of norms $\|v\|_\beta = \|S^\beta v\|_{\mathcal{H}}$, $\beta \in [0, 1/2]$ that were considered in Nemirovskii (1986) for the derivation of convergence rates for KCG algorithms in a deterministic setting.

The convergence rate of the kernel partial least squares estimator depends crucially on the sequence γ_n and the source parameter r. If $\gamma_n = O(n^{-1/2})$, this yields the same convergence rate as Theorem 2.1 of Blanchard and Krämer (2010b) for kernel conjugate gradient or de Vito et al. (2005) for kernel ridge regression with independent and identically distributed data. For stationary Gaussian time series we will derive concentration inequalities in the next section and obtain convergence rates depending on the source parameter r and the range of dependence. Note that Theorem 1 is rather general and it can be applied to any kind of dependence structure, as long as the necessary concentration inequalities can be established.

The next theorem derives faster convergence rates under assumptions on the effective dimensionality of operator S, which is defined as $d_\lambda = \mathrm{tr}\{(S+\lambda)^{-1}S\}$. The concept of effective dimensionality was introduced in Zhang (2003) to get sharp error bounds for general learning problems considered there. If $\mathcal{H}$ is a finite dimensional space, it was shown in Zhang (2003) that $d_\lambda \leq \dim(\mathcal{H})$. For infinite dimensional spaces it describes the complexity of the interactions between data and reproducing kernel.

If $d_\lambda = O(\lambda^{-s})$ for some $s \in (0, 1]$, Caponnetto and de Vito (2007) showed that the order optimal convergence rates $n^{-r/(2r+s)}$ are attained for KRR with independent and identically distributed data.

The effective dimensionality clearly depends on the behaviour of eigenvalues of S. If these converge sufficiently fast to zero, nearly parametric rates of convergence can be achieved for reproducing kernel Hilbert space methods, see, e.g., Dicker et al. (2017). In particular, the behaviour of d_λ around zero is of interest, since it determines how ill-conditioned the operator $(S+\lambda)^{-1}$ becomes. In the following theorem we set $\lambda = \lambda_n$ for a sequence $\{\lambda_n\}_{n\in\mathbb{N}} \subset (0, \infty)$ that converges to zero.

Theorem 2 *Assume that conditions (K1), (K2), (S) hold with $r \geq 1/2$ and that the effective dimensionality d_λ is known. Additionally, there are constants $C_\delta(\nu), C_\epsilon(\nu), C_\psi > 0$ and a sequence $\{\gamma_n\}_{n\in\mathbb{N}} \subset [0, \infty)$, $\gamma_n \to 0$, such that for $\nu \in (0, 1]$ and n sufficiently large*

$$\begin{aligned}
\mathrm{P}\left\{\|S_n - S\|_{\mathcal{L}} \leq C_\delta(\nu)\gamma_n\right\} &\geq 1 - \nu/3,\\
\mathrm{P}\left\{\|(S+\lambda_n)^{-1/2}(T_n^* y - Sf)\|_{\mathcal{H}} \leq C_\epsilon(\nu)\sqrt{d_{\lambda_n}}\gamma_n\right\} &\geq 1 - \nu/3,\\
\mathrm{P}\left\{\|(S+\lambda_n)^{1/2}(S_n+\lambda_n)^{-1/2}\|_{\mathcal{L}} \leq C_\psi\right\} &\geq 1 - \nu/3,
\end{aligned}$$

Here $\{\lambda_n\}_{n\in\mathbb{N}} \subset (0, \infty)$ is a sequence converging to zero such that for n large enough

$$\gamma_n \leq \lambda_n^{r-1/2}. \tag{6}$$

Take $\zeta_n = \max\{\sqrt{\lambda_n d_{\lambda_n}}\gamma_n, \lambda_n^{r+1/2}\}$ Define the stopping index a^* by

$$a^* = \min\left\{1 \leq a \leq n : \sum_{i=0}^{a} \|S_n f_{\widehat{\alpha}_i} - T_n^* y\|_{\mathcal{H}}^{-2} \geq (C\zeta_n)^{-2}\right\}, \tag{7}$$

with $C = 4R\max\{1, C_\psi^2, (r-1/2)\kappa^{r-3/2}C_\delta(\nu), 2^{-1/2}R^{-1}C_\psi C_\epsilon(\nu)\}$.

Then it holds with probability at least $1-\nu$ *that*

$$\|f_{\widehat{\alpha}_{a^*}} - f^*\|_2 = O\left\{\lambda_n^{-1/2}\zeta_n\right\},$$
$$\|f_{\widehat{\alpha}_{a^*}} - f\|_{\mathcal{H}} = O\left\{\lambda_n^{-1}\zeta_n\right\},$$

with $f^* = Tf$.

The condition (6) holds trivially for $r = 1/2$ as γ_n converges to zero. For $r > 1/2$ the sequence λ_n must not converge to zero arbitrarily fast.

In its general form Theorem 2 does not give immediate insight in the probabilistic convergence rates of the kernel partial least squares estimator. Therefore, we state two corollaries, where the function d_λ is specified. In both corollaries we explicitly state the choice of the sequence λ_n that yield the corresponding rates.

Corollary 3 *Assume that there exists* $s \in (0,1]$ *such that* $d_\lambda = O(\lambda^{-s})$ *for* $\lambda \to 0$. *Then under conditions of Theorem 2 with* $\lambda_n = \gamma_n^{2/(2r+s)}$ *it holds with probability at least* $1-\nu$ *that*

$$\|f_{\widehat{\alpha}_{a^*}} - f^*\|_2 = O\left\{\gamma_n^{2r/(2r+s)}\right\}.$$

Polynomial decay of the effective dimensionality $d_\lambda = \mathrm{tr}\{(S+\lambda)^{-1}S\}$ occurs if the eigenvalues of S also decay polynomially fast, that is, $\mu_i = c_s i^{-1/s}$ for $s \in (0,1]$, since in this case $d_\lambda = \sum_{i=1}^{\infty}\{1 + \lambda/c_s i^{1/s}\}^{-1} = O(\lambda^{-s})$. This holds, for example, for the Sobolev kernel $k(x,y) = \min(x,y)$, $x, y \in [0,1]$ and data that are uniformly distributed on $[0,1]$, see Raskutti et al. (2014).

If $\gamma_n = n^{-1/2}$, then the KPLS estimator converges in the $\mathcal{L}^2\left(\mathrm{P}^X\right)$-norm with a rate of $n^{-r/(2r+s)}$. This rate is shown to be optimal in Caponnetto and de Vito (2007) for KRR with independent identically distributed data.

Note that the rate obtained in Theorem 1 corresponds to $\gamma_n^{-2r/(2r+s)}$ with $s = 1$, i.e., the worst case rate with respect to the parameter $s \in (0,1]$.

In the next corollary to Theorem 2 we assume that the effective dimensionality behaves in a logarithmic fashion.

Corollary 4 *Let* $d_\lambda = O\{\log(1 + a/\lambda)\}$ *for* $\lambda \to 0$ *and* $a > 0$. *Then under the conditions of Theorem 2 with* $\lambda_n = \gamma_n^2 \log\{\gamma_n^{-2}\}$ *and* $r = 1/2$ *it holds with probability at least* $1-\nu$ *that*

$$\|f_{\widehat{\alpha}_{a^*}} - f^*\|_2 = O\left\{\gamma_n \log(1/2\gamma_n^{-2})\right\}.$$

The effective dimensionality takes the special form considered in this corollary, for example, when the eigenvalues of S decay exponentially fast. This holds, for example, if the data are Gaussian and the Gaussian kernel is used, see Section A. If $\gamma_n = O(n^{-1/2})$, then the convergence rate is of order $O\{n^{-1}\log(n)\}$, which is nearly parametric. It is noteworthy that the source condition only impacts the choice of the sequence λ_n, not the convergence rates of the estimator in the $\mathcal{L}^2\left(\mathrm{P}^X\right)$-norm. Therefore, we stated the corollary for $r = 1/2$,

which is a minimal smoothness condition on f^*, i.e., that $f^* = Tf$ almost surely for an $f \in \mathcal{H}$.

The rates obtained in Corollaries 3 and 4 for $\gamma_n = n^{-1/2}$ were derived in Dicker et al. (2017) for kernel ridge regression and kernel principal component regression under the assumption of independent and identically distributed data.

4. Concentration Inequalities for Gaussian Time Series

Crucial assumptions of Theorem 1 and 2 are the concentration inequalities for S_n and $T_n^* y$ and convergence of the sequence $\{\gamma_n\}_{n\in\mathbb{N}}$. Here we establish such inequalities in a Gaussian setting for stationary time series. At the end of this section we will state explicit convergence rates for $f_{\widehat{\alpha}_{a^*}}$ that depend not only on the source parameter $r \geq 1/2$ and the effective dimensionality d_λ, but also on the persistence of the dependence in the data.

The Gaussian setting is summarized in the following assumptions

(D1) $(X_h, X_0)^\intercal \sim \mathcal{N}_{2d}(0, \Sigma_h)$, $h = 1, \ldots, n-1$, with

$$\Sigma_h = \begin{bmatrix} \tau_0 & \tau_h \\ \tau_h & \tau_0 \end{bmatrix} \otimes \Sigma.$$

Here $\Sigma \in \mathbb{R}^{d\times d}$ and $V = [\tau_{|i-j|}]_{i,j=1}^n \in \mathbb{R}^{n\times n}$ are positive definite, symmetric matrices and $\otimes$ denotes the Kronecker product between matrices. Furthermore $X_0 \sim \mathcal{N}_d(0, \tau_0\Sigma)$.

(D2) For the autocorrelation function $\rho_h = \tau_0^{-1}\tau_h$ there exists a $q > 0$ such that $|\rho_h| \leq (h+1)^{-q}$ for $h = 0, \ldots, n-1$.

Condition (D1) is a separability condition for the covariance matrices Σ_h, $h = 0, \ldots, n-1$. Due to (D1) the effects (on the covariance) over time and between the different variables can be treated separately. Under condition (D2) it is easy to see that from $q > 1$ follows the absolute summability of the autocorrelation function ρ and thus $\{X_t\}_{t\in\mathbb{Z}}$ is a short memory process. Stationary short memory processes keep many of the properties of independent and identically distributed data, see, e.g., Brockwell and Davis (1991).

On the other hand $q \in (0, 1]$ yields a long memory process, see, e.g., Definition 3.1.2 in Giraitis et al. (2012). Examples of long memory processes are the fractional Gaussian noise with an autocorrelation function that behaves like $(h+1)^{-2(1-H)}$, with $H \in [0, 1)$ being the Hurst coefficient. Stationary long memory processes exhibit dependencies between observations that are more persistent, and many statistical results that hold for independent and identically distributed data, turn out to be false. See Samorodnitsky (2007) for more details.

The next theorem gives concentration inequalities for both estimators S_n and $T_n^* y$ in a Gaussian setting with convergence rates depending on the parameter $q > 0$. These inequalities are the ones needed in Theorem 1 and Theorem 2. Recall that $d_\lambda = \mathrm{tr}\{(S + \lambda)^{-1} S\}$ denotes the effective dimensionality of S.

Theorem 5 *(i) Define* $\mathrm{d}\mu_h(x,y) = \mathrm{dP}^{X_h,X_0}(x,y) - \mathrm{dP}^{X_0}(x)\mathrm{dP}^{X_0}(y)$. *Under Assumptions (K1) and (K2) it holds for* $\nu \in (0,1]$ *with probability at least* $1-\nu$ *that*

$$\|S_n - S\|_{\mathcal{L}}^2 \leq \frac{2\nu^{-1}}{n^2}\sum_{h=1}^{n-1}(n-h)\int\limits_{\mathbb{R}^{2d}} k^2(x,y)\mathrm{d}\mu_h(x,y) + \frac{\nu^{-1}}{n}\left[\mathrm{E}\{k^2(X_0,X_0)\} - \|S\|_{\mathrm{HS}}^2\right],$$

$$\|T_n^* y - Sf\|_{\mathcal{H}}^2 \leq \frac{2\nu^{-1}}{n^2}\sum_{h=1}^{n-1}(n-h)\int\limits_{\mathbb{R}^{2d}} k(x,y)f(x)f(y)\mathrm{d}\mu_h(x,y)$$
$$+ \frac{\nu^{-1}}{n}\left[\mathrm{E}\left\{k(X_0,X_0)f^2(X_0)\right\} - \|Sf\|_{\mathcal{H}}^2 + \sigma^2\mathrm{E}\{k^2(X_0,X_0)\}\right].$$

(ii) Assume that additionally to (K1), (K2) also (D1), (D2) for $q > 0$ *are fulfilled. Denote* $M = \sup_{x\in\mathbb{R}^d}|f(x)|$.

Then there exists a constant $C(q) > 0$ *such that*

$$\|S_n - S\|_{\mathcal{L}} \leq \nu^{-1/2}\{\gamma_n^2(q)\kappa C_\gamma + n^{-1}(\kappa^2 - \|S\|_{\mathrm{HS}}^2)\}^{1/2},$$
$$\|T_n^* y - Sf\|_{\mathcal{H}} \leq \nu^{-1/2}\left[\gamma_n^2(q)MC_\gamma + n^{-1}\left\{\kappa(M+\sigma^2) - \|Sf\|_{\mathcal{H}}^2\right\}\right]^{1/2},$$

for $C_\gamma = C(q)\{(2\pi)^d\det(\Sigma)\}^{-1/2}\kappa d^{1/2}(1-4^{-q})^{-1/4(d+2)}$. *The function* $\gamma_n(q)$, $q > 0$, *is defined as*

$$\gamma_n(q) = \begin{cases} n^{-1/2} & , \quad q > 1 \\ n^{-1/2}\log(1/2n) & , \quad q = 1 \\ n^{-q/2} & , \quad q \in (0,1). \end{cases}$$

(iii) Let (K1), (K2) and (S) hold. Let $\gamma_n(q)$ *be the function as defined in (ii). Then there exists a constant* $\tilde{C}_\epsilon > 0$ *such that it holds with probability at least* $1-\nu$ *for* $\lambda > 0$ *that*

$$\|(S+\lambda)^{-1/2}(T_n^* y - S_n f\|_{\mathcal{H}} \leq \nu^{-1/2}\tilde{C}_\epsilon\sigma\sqrt{d_\lambda}\gamma_n(q).$$

(iv) Let (K1), (K2), (S), (D1) and (D2) hold. Let $\lambda_n^{-1/2}d_{\lambda_n}^{1/2}\gamma_n(q) \to 0$ *for a sequence* $\lambda_n \to 0$ *and* $\gamma_n(q)$ *the function defined in (ii). Then there exists an* $n_0 = n_0(\nu,q) \in \mathbb{N}$ *such that with probability at least* $1-\nu$ *we have for all* $n \geq n_0$

$$\|(S+\lambda_n)^{1/2}(S_n+\lambda_n)^{-1/2}\|_{\mathcal{L}} \leq \sqrt{2}.$$

The first part of the theorem is general and can be used to derive concentration inequalities not only in the Gaussian setting and is of interest in itself. The convergence rate is controlled by the sums appearing on the right hand side. If these sums are of $O(n)$, then the mean squared error of both S_n and $T_n^* y$ will converge to zero with a rate of n^{-1}. On the other hand, if the sums are of order $O(n^{2-q})$ for some $q \in (0,1)$, the mean squared errors will converge with the reduced rate n^{-q}.

The second part derives explicit concentration inequalities in the Gaussian setting described by (D1) and (D2) with rates depending on the range of the dependence measured by $q > 0$. These inequalities appear in Theorem 1.

Parts (iii) and (iv) give the additional probabilistic bounds needed to apply Theorem 2. The condition $\lambda_n^{-1/2}d_{\lambda_n}^{1/2}\gamma_n(q) \to 0$ in Theorem 5 (iv) is fulfilled in the settings of Corollary 3 and Corollary 4.

Theorem 1, Corollary 3, Corollary 4 and Theorem 5 together imply

Corollary 6 *Let the conditions of Theorem 2 and (D1), (D2) hold.*

(i) Assume that there exists $s \in (0,1]$ such that $d_\lambda = O(\lambda^{-s})$ for $\lambda \to 0$. Then with probability at least $1-\nu$

$$\|f_{\widehat{\alpha}_{a^*}} - f^*\|_2 = \begin{cases} O\{n^{-r/(2r+s)}\}, & q > 1, \\ O\{n^{-qr/(2r+s)}\}, & q \in (0,1). \end{cases}$$

If instead of conditions of Theorem 2, conditions of Theorem 1 are assumed, then the convergence rates above have $s = 1$.

(ii) Assume that there exists $a > 0$ such that $d_\lambda = O\{\log(1 + a/\lambda)\}$ for $\lambda \to 0$ and $r = 1/2$. Then with probability at least $1-\nu$

$$\|f_{\widehat{\alpha}_{a^*}} - f^*\|_2 = \begin{cases} O\{n^{-1/2}\log(1/2n)\}, & q > 1, \\ O\{n^{-q/2}\log(1/2n^q)\}, & q \in (0,1). \end{cases}$$

Hence, for $q > 1$ the kernel partial least squares algorithm achieves the same rates as if the data were independent and identically distributed. For $q \in (0,1)$ the convergence rates become substantially slower, highlighting that dependence structures that persist over a long time can influence the convergence rates of the algorithm.

5. Simulations

To validate the theoretical results of the previous sections, we conducted a simulation study. The reproducing kernel Hilbert space is chosen to correspond to the Gaussian kernel $k(x,y) = \exp(-l\|x-y\|^2)$, $x, y \in \mathbb{R}^d$, $l = 2$, for $d = 1$. Our data will also be normally distributed. We refer to Proposition 7 in Appendix A for the derivation of functions that fulfill the source condition in this setting. Proposition 8 shows that the the eigenvalues of S decay exponentially fast. Hence the effective dimensionality d_λ behaves as in Corollary 4 and thus we expect convergence rates as given by Corollary 6 (ii).

The source parameter is taken to be $r = 4.5$ and we consider the function

$$f(x) = 4.37^{-1}\{3L_4(x,-4) - 2L_4(x,3) + 1.5L_4(x,9)\}, \quad x \in \mathbb{R}.$$

The normalization constant is chosen such that f takes values in $[-0.35, 0.65]$ and L_4 is the exponential function given in Proposition 7. The function f is shown in Figure 1.

In condition (D1) we set $\sigma_x^2 = \Sigma = 4$ (recall that $d = 1$). For the matrix $V = [\tau_{|i-j|}]_{i,j=1}^n \in \mathbb{R}^{n\times n}$ we choose three different structures for $n \in \{200, 400, 1000\}$. In the first setting $\tau_h = \mathbb{I}(h = 0)$, which corresponds to independent data. The second setting with $\tau_h = 0.9^{-h}$ implies an autoregressive process of order one. Finally, the third setting with $\tau_h = (1+h)^{-q}$, $q = 1/4$, $h = 0, \ldots, n-1$ leads to the long range dependent case.

In a Monte Carlo simulation with $M = 1000$ repetitions the time series $\{X_t^{(j)}\}_{t=1}^n$ are generated via $X^{(j)} = VN^{(j)}$ with $N^{(j)} \sim \mathcal{N}_n(0, \sigma^2 I_n)$, $j = 1, \ldots, M$, where I_n is the $n \times n$-dimensional identity matrix.

The residuals $\varepsilon_1^{(j)}, \ldots, \varepsilon_n^{(j)}$ are generated as independent standard normally distributed random variables and independent of $\{X_t^{(j)}\}_{t=1}^n$. The response is defined as $y_t^{(j)} = f(X_t^{(j)}) + \eta\,\varepsilon_t^{(j)}$, $t = 1, \ldots, n$, $j = 1, \ldots, M$, with $\eta = 1/16$.

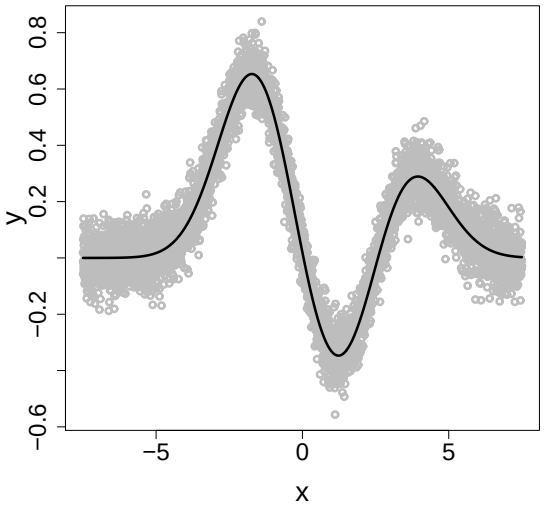

Figure 1: The function f evaluated on $[-7.5, 7.5]$ (black) and one realisation of the noisy data $y = f(x) + \varepsilon$ (grey).

The kernel partial least squares and kernel conjugate gradient algorithms are run for each sample $\{(X_t^{(j)}, y_t^{(j)})^{\mathrm{T}}\}_{t=1}^n$, $j = 1, \ldots, M$, with a maximum of 40 iteration steps. We denote the estimated coefficients with $\widehat{\alpha}_1^{(j,m)}, \ldots, \widehat{\alpha}_{40}^{(j,m)}$, $j = 1, \ldots, M$, with $m = CG$ meaning that the kernel conjugate gradient algorithm was employed and $m = PLS$ that kernel partial least squares was used to estimate $\alpha_1, \ldots, \alpha_n$.

The squared error in the $\mathcal{L}^2\left(\mathrm{P}^X\right)$-norm is calculated via

$$\widehat{e}_{n,\tau}^{(j,m)} = \min_{a=1,\ldots,40} \left[\frac{1}{\sqrt{2\pi\sigma_x^2}} \int_{-\infty}^{\infty} \left\{ f_{\widehat{\alpha}_a^{(j,m)}}(x) - f(x) \right\}^2 \exp\left(-\frac{1}{2\sigma_x^2} x^2 \right) \mathrm{d}x \right],$$

for $j = 1, \ldots, M$, $n = 200, 400, \ldots, 1000$ and $m \in \{CG, PLS\}$.

The results of the Monte-Carlo simulations are depicted in the boxplots of Figure 2. For kernel partial least squares (left panels) one observes that independent and autoregressive dependent data have roughly the same convergence rates, although the latter have a somewhat higher error. In contrast, the long range dependent data show slower convergence with the larger interquartile range, supporting the theoretical results of Corollary 6.

The $\mathcal{L}^2\left(\mathrm{P}^X\right)$-error of kernel conjugate gradient estimators is generally slightly higher than that of kernel partial least squares. Nonetheless, both of them have a similar behaviour.

We also investigated the the stopping indices $a = 1, \ldots, 40$ for which the errors $\widehat{e}_{n,\tau}^{(j,m)}$ were attained. These are shown in Figure 3 for independent and identically distributed data. It can be seen that the optimal indices for both algorithms have a rather similar behaviour. Kernel conjugate gradient stops slightly later, but overall the differences seem negligible.

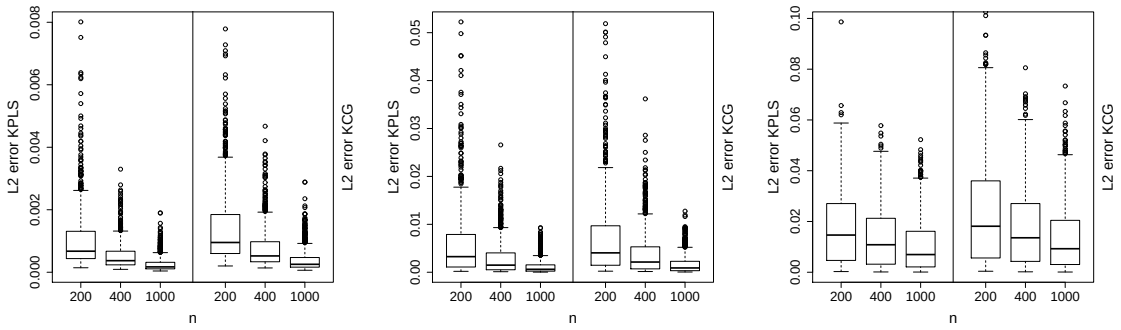

Figure 2: Boxplots of the $\mathcal{L}^2\left(\mathrm{P}^X\right)$-errors $\{\widehat{e}_{n,\tau}^{(j,m)}\}_{j=1}^M$ of kernel partial least squares (left side of each panel) and kernel conjugate gradient (right side of each panel) for different autocovariance functions τ and $n = 200, 400, 1000$. On the left is $\tau_h = \mathbb{I}(h = 0)$, in the middle $\tau_h = 0.9^{-h}$ and on the right $\tau_h = (h+1)^{-1/4}$.

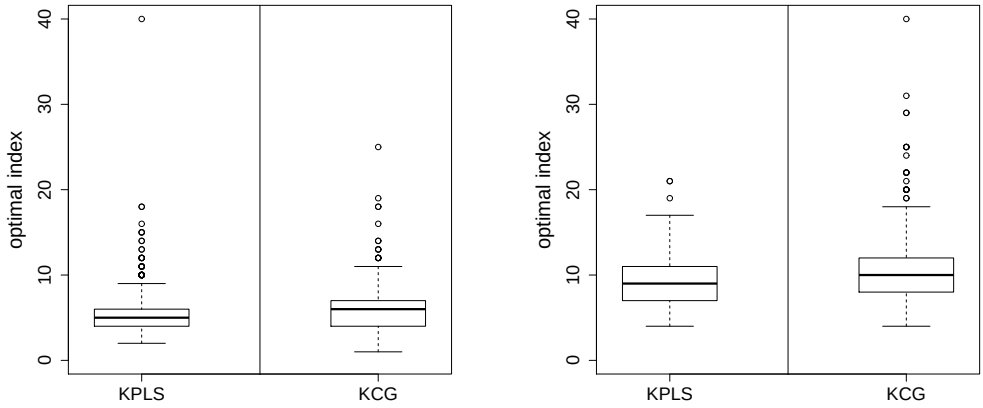

Figure 3: Boxplots of the optimal indices $a \in \{1, \ldots, 40\}$ for which the $\mathcal{L}^2\left(\mathrm{P}^X\right)$-errors $\{\widehat{e}_{n,\tau}^{(j,m)}\}_{j=1}^M$ were attained. Kernel partial least squares is on the left of each panel and kernel conjugate gradient on the right. On the left is $n = 200$, on the right $n = 1000$. The data were assumed to be independent and identically distributed.

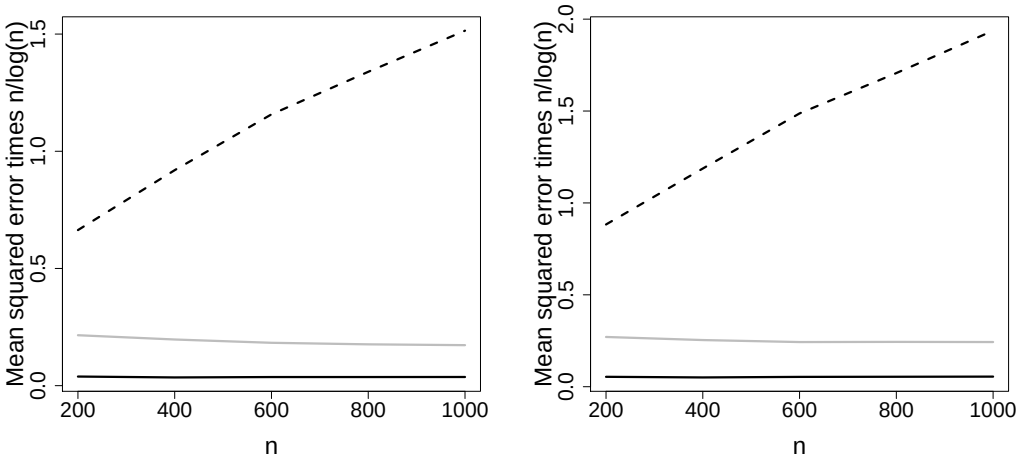

Figure 4: Mean of the $\mathcal{L}^2\left(\mathrm{P}^X\right)$-errors $\{\widehat{e}_{n,\tau}^{(j,m)}\}_{j=1}^M$ of kernel partial least squares (left) and kernel conjugate gradient (right) for $n = 200, 400, \ldots, 1000$ multiplied by $n/\log(n)$. The solid black line is for $\tau_h = \mathbb{I}(h = 0)$, the grey line for $\tau_h = 0.9^{-h}$ and the dashed black line for $\tau_h = (h+1)^{-1/4}$.

Figure 4 shows the mean (over j) of the estimated $\mathcal{L}^2\left(\mathrm{P}^X\right)$ errors $\{\widehat{e}_{n,\tau}^{(j,m)}\}_{j=1}^M$ for different n, τ and $m \in \{CG, PLS\}$. The errors were multiplied by $n/\log(n)$ to illustrate the convergence rates. According to Proposition 8 and Corollary 6 (ii) we expect the rates for the independent and autoregressive cases to be $n^{-1}\log(n)$, which is verified by the fact that the solid black and grey lines are roughly constant. For the long range dependent case we expect worse convergence rates which are also illustrated by the divergence of the dashed black line.

6. Application to Molecular Dynamics Simulations

The collective motions of protein atoms are responsible for its biological function and molecular dynamics simulations is a popular tool to explore this (Henzler-Wildman and Kern, 2007).

Typically, the $p \in \mathbb{N}$ backbone atoms of a protein are considered for the analysis with the relevant dynamics happening in time frames of nanoseconds. Although the dynamics are available exactly, the high dimensionality of the data and large number of observations can be cumbersome for regression analysis, e.g., due to the high collinearity in the columns of the covariates matrix. Many function-dynamic relationships are also non-linear (Hub and de Groot, 2009). A further complication is the fact that the motions of different backbone atoms are highly correlated, making additive non-parametric models for the target function f^* less suitable.

We consider T4 Lysozyme (T4L) of the bacteriophage T4, a protein responsible for the hydrolisis of 1,4-beta-linkages in peptidoglycans and chitodextrins from bacterial cell walls.

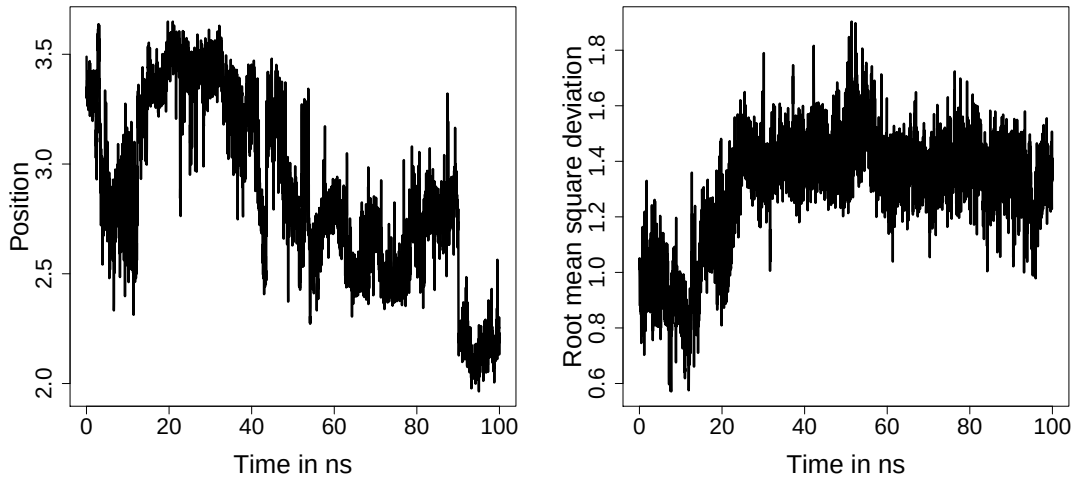

Figure 5: Time series of $X_{t,1}$, i.e., the first coordinate of the first atom T4L consists of (left) and the root mean squared deviation y_t between the protein configuration at time t and the (apo) crystal structure.

The number of available observations is $n = 4601$ and T4L consists of $p = 486$ backbone atoms.

Denote with $A_{t,i} \in \mathbb{R}^3$ the ith atom, $i = 1, \dots, p$, at time $t = 1, \dots, n$ and $c_i \in \mathbb{R}^3$ the ith atom in the (apo) crystal structure of T4L. A usual representation of the protein in a regression setting is the Cartesian one, i.e., we take as the covariate $X_t = (A_{1,t}^{\mathrm{T}}, \dots, A_{p,t}^{\mathrm{T}})^{\mathrm{T}}$, $t = 1, \dots, n$, see Brooks and Karplus (1983). The functional quantity to predict is the root mean square deviation of the protein configuration X_t at time $t = 1, \dots, n$ from the (apo) crystal structure $C = (c_1^{\mathrm{T}}, \dots, c_d^{\mathrm{T}})^{\mathrm{T}}$, i.e.,

$$y_t = \left\{ p^{-1} \sum_{i=1}^{p} \|X_{i,t} - C_i\|^2 \right\}^{1/2}.$$

This nonlinear function was previously considered in Hub and de Groot (2009), where it was established that linear models are insufficient for the prediction.

Figure 5 shows the time series corresponding to $X_{t,1}$ (i.e., the first coordinate of the first atom of T4L) on the left and the functional quantity y_t on the right. These plots reveal certain persistent dependence over time.

Fitting autoregressive moving average models of order $(3, 2)$ ($ARMA(3, 2)$) to y_t and $ARMA(5, 2)$ to $X_{t,1}$ shows that the smallest root of their respective characteristic polynomial is close to one (1.009 for y_t and 1.003 for $X_{t,1}$), highlighting that we are on the border of stationarity, see, e.g., Brockwell and Davis (1991).

Next, we performed an augmented Dickey Fuller test for the null hypothesis of non-stationarity against stationarity. Calculating the test statistics with 16 lags yields a p-value smaller than 0.01 for y_t and 0.0122 for $X_{t,1}$. Hence, both time series are likely stationary.

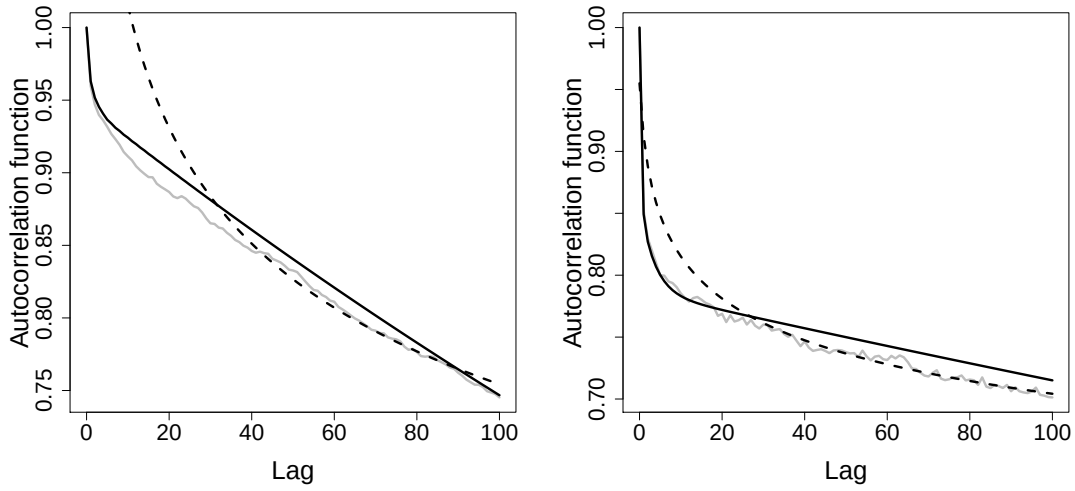

Figure 6: Autocorrelation plots of $X_{t,1}$ (left) and y_t (right). The estimated autocorrelation function is grey, the theoretical one of a fitted $ARMA(3,2)$ process is solid black and $\rho_h \propto (h+1)^{-q}$ for a suitable choice of $q > 0$ is dashed black.

Finally, to test for long range dependence, we employed the rescaled variance test of Giraitis et al. (2003). The null hypothesis of this test is the short range dependence in the data, while the alternative is the long range dependence. Calculating the test statistics with 16 lags gives the p-value for both y_t and $X_{t,1}$ smaller than 0.01, suggesting the long range dependence.

Figure 6 depicts the autocorrelation functions of $X_{t,1}$ and y_t, the theoretical autocorrelation function of the corresponding autoregressive moving average process and $\rho_h \propto (h+1)^{-q}$ for $q = 0.134$ for $X_{t,1}$ and $q = 0.066$ for y_t. The latter, as highlighted in Section 4, is an autocorrelation function for a stationary long range dependent process. These plots together with the above findings suggest that $X_{t,1}$ and y_t are stationary, long range dependent processes.

We apply kernel partial least squares to this data set with the Gaussian kernel $k(x,y) = \exp(-l\|x-y\|^2)$, $x, y \in \mathbb{R}^{3p}$, $l > 0$. The function f we aim to estimate is a distance between protein configurations, so using a distance based kernel seems reasonable. Moreover, we also investigated the impact of other bounded kernels such as triangular and Epanechnikov and obtained similar results. The first 50% of the data form a training set to calculate the kernel partial least squares estimator and the remaining data are used for testing.

The parameter $l > 0$ is calculated via cross validation on the training set. In our evaluation we obtained $l = 10.22$.

Figure 7 compares the observed response in the test set with the prediction on the test set obtained by kernel partial least squares, kernel principal component regression and linear partial least squares. Apparently, kernel partial least squares show the best performance and the kernel principal components algorithm is able to achieve comparable prediction with

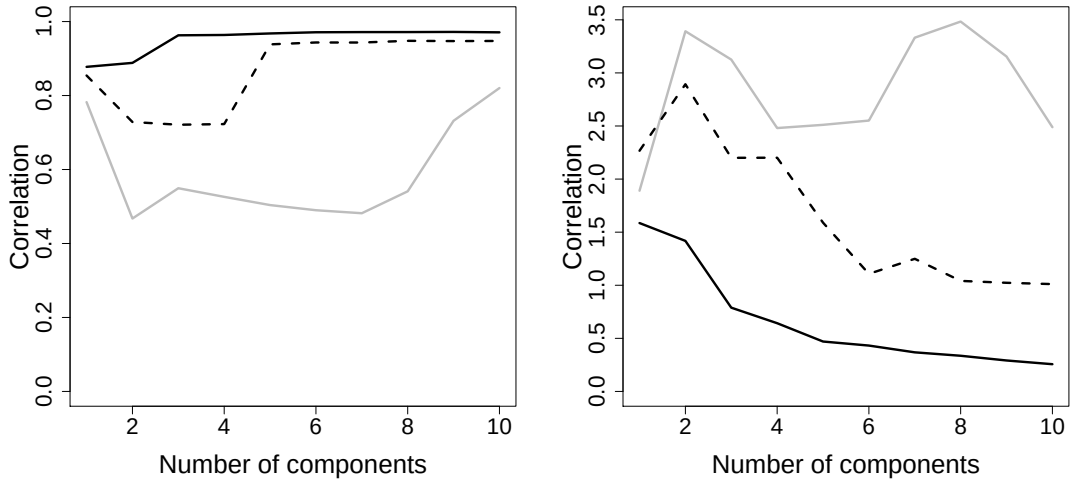

Figure 7: Correlation (left) and residual sum of squares (right) between predicted values and the observed response on the test set depending on the number of used components for kernel partial least squares (solid black), partial least squares (grey) and kernel principal component regression (dashed black).

more components. Obviously, linear partial least squares can not cope with the non-linearity of the problem.

This application highlights that kernel partial least squares still deliver a robust prediction even for long range dependent data, if enough observations are available.

Acknowledgements

The authors are grateful to the action editor and two reviewers for their valuable comments which helped to improve the article. The support of the German Research Foundation via FOR 916 (Project B5) and CRC 803 (Proejct Z2) is gratefully acknowledged.

Appendix A. Source Condition and Effective Dimensionality for Gaussian Kernels

The source condition (S) and the effective dimensionality d_λ are of great importance in the convergence rates derived in Section 3. Here we investigate these conditions for the reproducing kernel Hilbert space corresponding to the Gaussian kernel $k(x,y) = \exp(-l\|x-y\|^2)$, $x, y \in \mathbb{R}^d$, $l > 0$, for $d = 1$. Hence, the space $\mathcal{H}$ is the space of all analytic functions that decay exponentially fast, see Steinwart et al. (2005).

We also impose the normality conditions (D1) and (D2) on $\{X_t\}_{t\in\mathbb{Z}}$, where now $\sigma_x^2 = \Sigma \in \mathbb{R}$ due to $d = 1$. The following proposition derives a more explicit representation for $f \in \mathcal{H}$.

Proposition 7 *Assume that (K1),(K2) and (S) hold for $r \geq 1/2$. Let $d = 1$, $X_0 \sim \mathcal{N}(0, \sigma_x^2), \sigma_x^2 > 0$ and consider the Gaussian kernel $k(x,y) = \exp\{-l(x-y)^2\}$ for $x, y \in \mathbb{R}$, $l > 0$. Then f can be expressed for $\mu = r - 1/2 \in \mathbb{N}$ via $f(x) = \sum_{i=1}^{\infty} c_i L_\mu(x, z_i)$ for fixed $\{z_i\}_{i=1}^{\infty}, \{c_i\}_{i=1}^{\infty} \subset \mathbb{R}$ such that $\sum_{i,j=1}^{\infty} c_i c_j k(z_i, z_j) \leq R^2$, $R > 0$. Here we have for $x, z \in \mathbb{R}$*

$$L_\mu(x,z) = \exp\left[-1/2\left\{\frac{\det(\Lambda)(x^2+z^2) - 2l^{\mu+1}xz}{\det(\Lambda_{1:\mu})}\right\}\right],$$

with $\Lambda \in \mathbb{R}^{(\mu+1)\times(\mu+1)}$ being a tridiagonal matrix with elements

$$\Lambda_{i,j} = \begin{cases} \sigma_x^{-2} + 2l & , \quad i = j < \mu + 1 \\ l & , \quad i = j = \mu + 1 \\ -l & , \quad |i-j| = 1 \\ 0 & , \quad else \end{cases}$$

for $i, j = 1, \ldots, \mu + 1$ and $\Lambda_{1:\mu}$ is the $\mu \times \mu$-dimensional sub-matrix of Λ including the first μ columns and rows.

Conversely any function $f^* = Tf$ with f of the above form fulfills a source condition (S) with $r = \mu + 1/2$, $\mu \in \mathbb{N}$.

Hence if we fix an $r \geq 1/2$ with $r - 1/2 \in \mathbb{N}$ this theorem gives us a way to construct functions $f \in \mathcal{H}$ with $f^* = Tf$ that fulfill (S).

The next proposition derives the effective dimensionality d_λ in this setting:

Proposition 8 *Let $d = 1$, $X_0 \sim \mathcal{N}(0, \sigma_x^2)$ for some $\sigma_x^2 > 0$ and consider the Gaussian kernel $k(x,y) = \exp\{-l(x-y)^2\}$, $x, y \in \mathbb{R}$, $l > 0$.*

Then there is a constant $D > 0$ such that it holds for any $\lambda \in (0, 1]$

$$d_\lambda = \operatorname{tr}\{(S+\lambda)^{-1}S\} \leq D \log(1 + a/\lambda),$$

with $a = \sqrt{2}(1 + \beta + \sqrt{1+\beta})^{-1/2}$, $\beta = 4l\sigma_x^2$.

With the latter result Corollary 4 is applicable and we expect convergence rates for the kernel partial least squares algorithm of order $O\{\gamma_n \log(1/2\gamma_n^{-2})\}$ for a sequence $\{\gamma_n\}_n$ as in Theorem 2.

Appendix B. Proofs

B.1 Proof of Theorem 1

The proof of Theorem 1 makes use of the connection between the partial least squares and the conjugate gradient algorithm. This section is structured as follows: First we will introduce the link between kernel partial least squares and kernel conjugate gradient. We will state some key facts about orthogonal polynomials and their relationship to the algorithm in Lemma 9. Then the consistency of kernel partial least squares is shown with the help of three error bounds that are obtained in Lemmas 11 — 13.

With a slight abuse of notation we define $f_i = f_{\widehat{\alpha}_i}$ for $i = 1, \ldots, n$.

B.1.1 Orthogonal Polynomials and Some Notation

Denote with $\mathcal{P}_i$ the set of polynomials of degree at most $i = 0, \ldots, n$. For functions $\psi, \phi : \mathbb{R} \to \mathbb{R}$ and $r \in \mathbb{N}_0$ define the inner products $[\psi, \phi]_r = \langle \psi(S_n) T_n^* y, S_n^r \phi(S_n) T_n^* y \rangle_{\mathcal{H}}$. From the definition of the Krylov space it is immediate that every element $v \in \mathcal{K}_i(S_n, T_n^* y)$, $i = 1, \ldots, n$, can be represented by a polynomial $q \in \mathcal{P}_{i-1}$ via $v = q(S_n) T_n^* y$.

The following discussion is based on Hanke (1995), chapter 2. There exist two sequences of polynomials $\{p_i\}_{i=0}^n, \{q_i\}_{i=0}^n \subset \mathcal{P}_i$, such that $f_i = q_{i-1}(S_n) T_n^* y$ with $q_{-1} = 0$ and $T_n^* y - S_n f_i = p_i(S_n) T_n^* y$. Both sequences are connected by the equation $p_i(x) = 1 - x q_{i-1}(x)$, $x \in \mathbb{R}$, and the polynomials $\{p_i\}_{i=0}^n$ are orthogonal with respect to $[\cdot, \cdot]_0$.

We will also consider other sequences of polynomials, namely $\{q_i^{[r]}\}_{i=0}^n, \{p_i^{[r]}\}_{i=0}^n \subset \mathcal{P}_i$, $q_{-1}^{[r]} = 0$, such that $p_i^{[r]}(x) = 1 - x q_{i-1}^{[r]}(x)$, $x \in \mathbb{R}$, and the sequence $\{p_i^{[r]}\}_{i=0}^n$ is orthogonal with respect to $[\cdot, \cdot]_r$. This yields for every $r \in \mathbb{N}_0$ a separate conjugate gradient algorithm with solution $f_i^{[r]} = q_{i-1}^{[r]}(S_n) T_n^* y \in \mathcal{K}_i(S_n, T_n^* y)$ and residuals $T_n^* y - S_n f_i^{[r]} = p_i^{[r]}(S_n) T_n^* y$, $i = 1, \ldots, n$. The $p_i^{[r]}$, $i = 0, \ldots, n$, $r \in \mathbb{N}_0$, are called residual polynomials.

As S_n is self-adjoint, positive semi-definite and the kernel is bounded by κ we know that its spectrum is a subset of $[0, \kappa]$, see Caponnetto and de Vito (2007). This also implies that $\max\{\|S\|_{\mathcal{L}}, \|S_n\|_{\mathcal{L}}\} \leq \kappa$, with $\|\cdot\|_{\mathcal{L}}$ denoting the operator norm. The i distinct roots of $p_i^{[r]}$ will be denoted by $0 < x_{1,i}^{[r]} < \ldots x_{i,i}^{[r]} < \kappa$, $i = 1, \ldots, n$.

We will summarize some key facts about the orthogonal polynomials in the next lemma.

Lemma 9 *Let $r, s \in \mathbb{N}_0$ and $i = 1, \ldots, n$. Then we have:*

(i) The roots of consecutive orthogonal polynomials interlace, i.e., for $j = 1, \ldots, i$ it holds

$$0 < x_{j,i+1}^{[r]} < x_{j,i}^{[r]} < x_{j,i}^{[r+1]} < x_{j+1,i+1}^{[r]} < x_{j+1,i}^{[r]} < \cdots < x_{i,i}^{[r+1]} < x_{i+1,i+1}^{[r]} < \kappa,$$

(ii) the optimality property $[p_i^{[1]}, p_i^{[1]}]_0^{1/2} = \|T_n^* y - S_n f_i^{[1]}\|_{\mathcal{H}} \leq \|T_n^* y - S_n h\|_{\mathcal{H}}$ holds for all $h \in \mathcal{K}_i(S_n, T_n^* y)$,

(iii) on $x \in [0, x_{1,i}^{[r]}]$ it holds $0 \leq p_i^{[r]}(x) \leq 1$ and $q_i^{[r]}(x) \leq \left| \left(p_i^{[r]} \right)'(0) \right|$,

(iv) $p_n^{[r]} = p_n^{[s]}$,

(v) $\left(p_i^{[r]} \right)'(0) = -\sum_{j=1}^i \left(x_{j,i}^{[r]} \right)^{-1}$,

(vi) for $r \geq 1$ define $\phi_i(x) = p_i^{[r]}(x) \left(x_{1,i}^{[r]} \right)^{1/2} \left(x_{1,i}^{[r]} - x \right)^{-1/2}$, $x \in [0, x_{1,i}^{[r]}]$, $i = 1, \ldots, n$. Then it holds for $u \geq 0$ that $x^u \phi_i^2(x) \leq u^u \left| \left(p_i^{[r]} \right)'(0) \right|^{-u}$ with the convention $0^0 = 1$.

Proof: (i) See Hanke (1995), Corollary 2.7.

(ii) See Hanke (1995), Proposition 2.1.

(iii) Due to part (i) we know that all i roots of the polynomial $p_i^{[r]}$ are contained in $(0, \kappa)$. Furthermore $p_i^{[r]}(0) = 1 - 0 q_i^{[r]} = 1$. Thus $p_i^{[r]}$ is convex and falling in $[0, x_{1,i}^{[r]}]$ and the first assertion follows.

Because of the convexity of $p_i^{[r]}$ on $[0, x_{1,i}^{[r]}]$ we get $q_i^{[r]}(x) = x^{-1}\{1-p_i^{[r]}(x)\} \leq \left|\left(p_i^{[r]}\right)'(0)\right|$.

(iv) See the discussion in Hanke (1995) preceding Proposition 2.1 and use the facts that $T_n^* y \in \text{range}(S_n)$ and S_n is an operator of rank n.

(v) Write $p_i^{[r]}(x) = \prod_{j=1}^{i}(1 - x/x_{j,i}^{[r]})$, $x \in [0, \kappa]$, and the result is immediate.

(vi) See equation (3.10) in Hanke (1995). ■

We denote for $x \geq 0$ by P_x the orthogonal projection operator on the eigenspace corresponding to the eigenvalues of S_n that are smaller or equal x and $P_x^\perp = I_\mathcal{H} - P_x$ with $I_\mathcal{H} : \mathcal{H} \to \mathcal{H}$ being the identity operator.

B.1.2 Preparation for the Proof

We consider the kernel partial least squares algorithm as an optimization problem

$$f_i = \arg \min_{g \in \mathcal{K}_i(S_n, T_n^* y)} n^{-1}\|y - T_n g\|^2, \quad i = 1, \ldots, n. \tag{8}$$

This is the conjugate gradient algorithm CGNE discussed in chapter 2.2 of Hanke (1995), to which we refer for more details. Note that, for example,

$$n^{-1}\langle \phi(T_n T_n^*) y, T_n T_n^* \psi(T_n T_n^*) y \rangle = \langle \phi(S_n) T_n^* y, \psi(S_n) T_n^* y \rangle_\mathcal{H} = [\phi, \psi]_0,$$

for polynomials ϕ, ψ.

In the upcoming proofs we will make use of the following operator inequality:

Lemma 10 *Let $B, C : \mathcal{H} \to \mathcal{H}$ be two positive semi-definite, self-adjoint operators with $\max\{\|B\|_\mathcal{L}, \|C\|_\mathcal{L}\} \leq \kappa$. Then it holds for any $r \geq 0$ with $\zeta = \max\{r-1, 0\}$*

$$\|B^r - C^r\|_\mathcal{L} \leq (\zeta + 1)\kappa^\zeta \|B - C\|_\mathcal{L}^{r-\zeta}.$$

Proof: See Blanchard and Krämer (2010b), Lemma A.6. ■

For the remainder of the proof we assume that we are on the set where it holds with probability at least $1-\nu$, $\nu \in (0, 1]$, that $\|S_n - S\|_\mathcal{L} \leq C_\delta(\nu)\gamma_n$ and $\|T_n^* y - Sf\|_\mathcal{H} \leq C_\epsilon(\nu)\gamma_n$ for a sequence $\{\gamma_n\}_n$ converging to zero and constants $C_\delta = C_\delta(\nu), C_\epsilon = C_\epsilon(\nu) > 0$.

With Lemma 2.4 in Hanke (1995) we see that the stopping iteration (5) can also be expressed as

$$a^* = \min\left\{1 \leq a \leq n : \|S_n f_a^{[1]} - T_n^* y\|_\mathcal{H} \leq C\gamma_n\right\}, \tag{9}$$

i.e., we stop the kernel partial least squares algorithm when a discrepancy principle for $f_a^{[1]}$ holds.

It is easy to see that from (S) it follows for $r \geq 1/2$ that

(SH) There exist $\mu \geq 0$, $R > 0$ and $u \in \mathcal{H}$ such that $f = S^\mu u$ and $\|u\|_\mathcal{H} \leq R$.

This condition is known as the Hölder source condition with $\mu = r - 1/2$.

Recall that $\mathcal{H} \subseteq \mathcal{L}^2\left(\mathrm{P}^X\right)$ and $T : \mathcal{H} \to \mathcal{L}^2\left(\mathrm{P}^X\right)$ is the change of space operator. Using the fact that T, T^* are adjoint operators, $f_{a^*} = T f_{a^*}$ and $f^* = Tf$ for $r \geq 1/2$ we see

$$\|f_{a^*} - f^*\|_2 = \|T(f_{a^*} - f)\|_2 = \langle S(f_{a^*} - f), f_{a^*} - f\rangle_\mathcal{H} = \|S^{1/2}(f_{a^*} - f)\|_\mathcal{H}.$$

An application of Lemma 10 yields

$$\begin{aligned}\|f_{a^*} - f^*\|_2 &= \|S^{1/2}(f_{a^*} - f)\|_{\mathcal{H}} \leq \|S^{1/2}(f_{a^*} - f_{a^*}^{[1]})\|_{\mathcal{H}} + \|S^{1/2}(f_{a^*}^{[1]} - f)\|_{\mathcal{H}} \\ &\leq C_\delta^{1/2}\gamma_n^{1/2}\left(\|f_{a^*} - f_{a^*}^{[1]}\|_{\mathcal{H}} + \|f_{a^*}^{[1]} - f\|_{\mathcal{H}}\right) + \|S_n^{1/2}(f_{a^*} - f_{a^*}^{[1]})\|_{\mathcal{H}} + \|S_n^{1/2}(f_{a^*}^{[1]} - f)\|_{\mathcal{H}}.\end{aligned} \tag{10}$$

The following lemmas will deal with bounding the quantities in (10) in terms of the source parameter $r = \mu + 1/2 \geq 1/2$ and the sequence γ_n. First we will derive upper bounds for the quantities containing the difference of the KPLS estimator f_{a^*} and the estimator $f_{a^*}^{[1]}$:

Lemma 11 *Assume $C_x \in (0, 1]$ such that $x_* = (C_x\gamma_n)^{1/(\mu+1)} < x_{1,a^*-1}^{[1]}$ and $C > C_\epsilon + C_xR + C_\delta(\mu+1)\kappa^\mu R$. Under the conditions of the theorem it holds $\mu \geq 0$*

$$\begin{aligned}\|f_{a^*} - f_{a^*}^{[1]}\|_{\mathcal{H}} &\leq \gamma_n^{\mu/(\mu+1)} \frac{C}{C_x^{1/(\mu+1)}\left[1 - C^{-1}\{C_\epsilon + C_xR + C_\delta(\mu+1)\kappa^\mu R\}\right]^2} \\ \|S_n^{1/2}(f_{a^*} - f_{a^*}^{[1]})\|_{\mathcal{H}} &\leq \gamma_n^{(2\mu+1)/(2\mu+2)} \frac{C}{C_x^{1/(2\mu+2)}\left[1 - C^{-1}\{C_\epsilon + C_xR + C_\delta(\mu+1)\kappa^\mu R\}\right]}.\end{aligned}$$

Proof: If the inner products $[\cdot,\cdot]_0$ and $[\cdot,\cdot]_1$ are the same the proof is done because both polynomial sequences are identical.

We now observe that we have for $a^* = n$ due to Lemma 9 (iv) $q_{n-1}(x) - q_{n-1}^{[1]}(x) = x^{-1}\{p_n^{[1]}(x) - p_n(x)\} = 0$, i.e., $\|f_{a^*} - f_{a^*}^{[1]}\|_{\mathcal{H}} = 0$ and $\|S_n^{1/2}(f_{a^*} - f_{a^*}^{[1]})\|_{\mathcal{H}} = 0$ and the proof is done.

If the inner products differ and we have $0 < a^* < n$ it holds $f_{a^*} \neq f_{a^*}^{[1]}$.

Proposition 2.8 in Hanke (1995) can now be applied for $0 < a^* < n$ and yields $q_{a^*-1}(x) - q_{a^*-1}^{[1]}(x) = x^{-1}\{p_{a^*}^{[1]}(x) - p_{a^*}(x)\} = \theta_{a^*}p_{a^*-1}^{[2]}(x)$, $x \geq 0$, with $\theta_{a^*} = (p_{a^*}^{[1]})'(0) - (p_{a^*}^{[0]})'(0) > 0$. We get $f_{a^*} - f_{a^*}^{[1]} = q_{a^*-1}(S_n)T_n^*y - q_{a^*-1}^{[1]}(S_n)T_n^*y = \theta_{a^*}p_{a^*-1}^{[2]}(S_n)T_n^*y$.

Proposition 2.9 in Hanke (1995) yields $\theta_{a^*} = \left[p_{a^*-1}^{[2]}, p_{a^*-1}^{[2]}\right]_1^{-1}\left[p_{a^*}^{[1]}, p_{a^*}^{[1]}\right]_0$. The optimality property of $f_{a^*}^{[1]}$ in Lemma 9 (ii) shows that

$$\|T_n^*y - S_nf_{a^*}^{[1]}\|_{\mathcal{H}} = \|p_{a^*}^{[1]}(S_n)T_n^*y\|_{\mathcal{H}} = \left[p_{a^*}^{[1]}, p_{a^*}^{[1]}\right]_0^{1/2} \leq \left[p_{a^*-1}^{[2]}, p_{a^*-1}^{[2]}\right]_0^{1/2}. \tag{11}$$

Combining these results yields

$$\|f_{a^*} - f_{a^*}^{[1]}\|_{\mathcal{H}} = \frac{\left[p_{a^*}^{[1]}, p_{a^*}^{[1]}\right]_0}{\left[p_{a^*-1}^{[2]}, p_{a^*-1}^{[2]}\right]_1}\left[p_{a^*-1}^{[2]}, p_{a^*-1}^{[2]}\right]_0^{1/2} \leq \frac{\left[p_{a^*-1}^{[2]}, p_{a^*-1}^{[2]}\right]_0}{\left[p_{a^*-1}^{[2]}, p_{a^*-1}^{[2]}\right]_1}\|p_{a^*}^{[1]}(S_n)T_n^*y\|_{\mathcal{H}}. \tag{12}$$

Recall that $x^{[2]}_{1,a^*-1}$ denotes the first root of $p^{[2]}_{a^*-1}$. It holds for any $0 \leq x \leq x^{[2]}_{1,a^*-1}$ that $0 \leq p^{[2]}_{a^*-1}(x) \leq 1$, see Lemma 9 (iii), and thus

$$\begin{aligned}\left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1}\right]_0^{1/2} &\leq \|P_x p^{[2]}_{a^*-1}(S_n)\{T_n^* y - Sf + Sf\}\|_{\mathcal{H}} + \|P_x^{\perp} p^{[2]}_{a^*-1}(S_n) T_n^* y\|_{\mathcal{H}} \\ &\leq C_\epsilon \gamma_n + \|P_x p^{[2]}_{a^*-1}(S_n) S^{\mu+1} u\|_{\mathcal{H}} + x^{-1/2}\|P_x^{\perp} S_n^{1/2} p^{[2]}_{a^*-1}(S_n) T_n^* y\|_{\mathcal{H}} \\ &\leq C_\epsilon \gamma_n + x^{\mu+1} R + \|P_x p^{[2]}_{a^*-1}(S^{\mu+1} - S_n^{\mu+1}) u\|_{\mathcal{H}} + \frac{1}{\sqrt{x}} \left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1}\right]_1^{1/2}.\end{aligned}$$

In the second inequality (SH) with $\mu \geq 0$ was applied.

By assumption $x_* = (C_x \gamma)^{1/(\mu+1)} \leq x^{[1]}_{1,a^*-1} < x^{[2]}_{1,a^*-1}$ due to the interlacing property of the roots of the polynomials $p_i^{[r]}$, $i = 1, \ldots, n$, $r \in \mathbb{N}_0$, see Lemma 9 (i).

Using Lemma 10 we get $\|S^{\mu+1} - S_n^{\mu+1}\|_{\mathcal{L}} \leq (\mu+1)\kappa^\mu C_\delta \gamma_n$ and setting $x = x_*$ we get

$$\begin{aligned}\left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1}\right]_0^{1/2} &\leq C_\epsilon \gamma_n + x_*^{\mu+1} R + C_\delta \gamma_n (\mu+1)\kappa^\mu R + x_*^{-1/2} \left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1}\right]_1^{1/2} \\ &= \gamma_n \left\{C_\epsilon + C_x R + C_\delta(\mu+1)\kappa^\mu R\right\} + x_*^{-1/2} \left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1}\right]_1^{1/2}. \qquad (13)\end{aligned}$$

Due to (9) and (11) we have additionally $C\gamma_n \leq \|S_n f^{[1]}_{a^*-1} - T_n^* y\|_{\mathcal{H}} = \|p^{[1]}_{a^*-1}(S_n) T_n^* y\|_{\mathcal{H}} \leq \left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1}\right]_0^{1/2}$.

Plugging this into (13) yields

$$\left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1}\right]_0^{1/2} \leq \frac{C_\epsilon + C_x R + C_\delta(\mu+1)\kappa^\mu R}{C} \left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1}\right]_0^{1/2} + \frac{1}{\sqrt{x_*}} \left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1}\right]_1^{1/2},$$

or equivalently with $x_* = (C_x \gamma_n)^{1/(\mu+1)}$

$$\left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1}\right]_0^{1/2} \leq \gamma_n^{-1/(2\mu+2)} \frac{\left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1}\right]_1^{1/2}}{C_x^{1/(2\mu+2)} \left[1 - C^{-1}\{C_\epsilon + C_x R + C_\delta(\mu+1)\kappa^\mu R\}\right]}, \qquad (14)$$

where by assumption $C > C_\epsilon + C_x R + C_\delta(\mu+1)\kappa^\mu R$ and $x_* = (C_x \gamma)^{1/(\mu+1)}$.

Combining (12), (14) and $\|p^{[1]}_{a^*}(S_n) T_n^* y\|_{\mathcal{H}} \leq C\gamma_n$ due to the stopping index (9) yields

$$\begin{aligned}\|f_{a^*} - f^{[1]}_{a^*}\|_{\mathcal{H}} &\leq \gamma_n^{-1/(\mu+1)} \frac{\|p^{[1]}_{a^*}(S_n) T_n^* y\|_{\mathcal{H}}}{C_x^{1/(\mu+1)} \left[1 - C^{-1}\{C_\epsilon + C_x R + C_\delta(\mu+1)\kappa^\mu R\}\right]^2} \\ &\leq \gamma_n^{\mu/(\mu+1)} \frac{C}{C_x^{1/(\mu+1)} \left[1 - C^{-1}\{C_\epsilon + C_x R + C_\delta(\mu+1)\kappa^\mu R\}\right]^2}.\end{aligned}$$

For the second part of the proof we derive in the same way as (12)

$$\|S_n^{1/2}(f_{a^*} - f^{[1]}_{a^*})\|_{\mathcal{H}} \leq \frac{\left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1}\right]_0^{1/2}}{\left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1}\right]_1^{1/2}} \|p^{[1]}_{a^*}(S_n) T_n^* y\|_{\mathcal{H}}.$$

Using (14) and $\|p^{[1]}_{a^*}(S_n)T_n^* y\|_{\mathcal{H}} \leq C\gamma_n$ gives

$$\|S_n^{1/2}(f_{a^*} - f^{[1]}_{a^*})\|_{\mathcal{H}} \leq \gamma_n^{(2\mu+1)/(2\mu+2)} \frac{C}{C_x^{1/(2\mu+2)}\left[1 - C^{-1}\{C_\epsilon + C_x R + C_\delta(\mu+1)\kappa^\mu R\}\right]},$$

finishing the proof. ■

We now derive an upper bound on the quantities that contain $f^{[1]}_{a^*}$ and f. These contain $\|S_n f_i^{[1]} - T_n^* y\|_{\mathcal{H}}$, which can be controlled for $i = a^*$ by the discrepancy principle (9), and $(p_i^{[1]})'(0)$, for which we have to derive a separate bound later on.

Lemma 12 *For any $i = 1, \ldots, n$ and any $0 < x \leq x^{[1]}_{1,i}$ we have under the conditions of the theorem for $\mu \geq 1$*

$$\begin{aligned}
\|f - f_i^{[1]}\|_{\mathcal{H}} &\leq R\left\{x^\mu + C_\delta \mu \kappa^{\mu-1}\gamma_n\right\} + x^{-1}\left\{\|S_n f_i^{[1]} - T_n^* y\|_{\mathcal{H}} + (C_\epsilon + C_\delta \kappa^\mu R)\gamma_n\right\} \\
&\quad + (C_\epsilon + C_\delta \kappa^\mu R)\gamma_n |(p_i^{[1]})'(0)|, \\
\|S_n^{1/2}(f - f_i^{[1]})\|_{\mathcal{H}} &\leq R\left\{x^{\mu+1/2} + x^{1/2} C_\delta \mu \kappa^{\mu-1}\gamma_n\right\} \\
&\quad + x^{-1/2}\left\{\|S_n f_i^{[1]} - T_n^* y\|_{\mathcal{H}} + (C_\epsilon + C_\delta \kappa^\mu R)\gamma_n\right\} \\
&\quad + x^{1/2}(C_\epsilon + C_\delta \kappa^\mu R)\gamma_n |(p_i^{[1]})'(0)|.
\end{aligned}$$

Proof: Denote $\bar{f}_i = q^{[1]}_{i-1}(S_n) S_n f$ and consider

$$\|f - f_i^{[1]}\|_{\mathcal{H}} \leq \|P_x(f - \bar{f}_i)\|_{\mathcal{H}} + \|P_x(\bar{f}_i - f_i^{[1]})\|_{\mathcal{H}} + \|P_x^\perp(f - f_i^{[1]})\|_{\mathcal{H}}. \tag{15}$$

The first term of (15) can be bound by an application of Lemma 10 and (SH) with $\mu \geq 1$

$$\begin{aligned}
\|P_x(f - \bar{f}_i)\|_{\mathcal{H}} &= \|P_x\{I - q^{[1]}_{i-1}(S_n)S_n\}f\|_{\mathcal{H}} = \|P_x p_i^{[1]}(S_n) f\|_{\mathcal{H}} = \|P_x p_i^{[1]}(S_n) S^\mu u\|_{\mathcal{H}} \\
&\leq \|P_x p_i^{[1]}(S_n) S_n^\mu u\|_{\mathcal{H}} + \|P_x p_i^{[1]}(S_n)(S^\mu - S_n^\mu)u\|_{\mathcal{H}} \\
&\leq R\left\{x^\mu + C_\delta \mu \kappa^{\mu-1}\gamma_n\right\}.
\end{aligned}$$

In the last inequality we used that on $0 \leq x \leq x^{[1]}_{1,i}$ we have $0 \leq p_i^{[1]}(x) \leq 1$.

For the second term of (15) we use Lemma 9 (iii) $q_i^{[1]}(x) \leq |(p_i^{[1]})'(0)|$ on $x \in [0, x^{[1]}_{1,i}]$. This yields

$$\begin{aligned}
\|P_x(f_i^{[1]} - \bar{f}_i)\|_{\mathcal{H}} &= \|P_x q_i^{[1]}(S_n)(S_n f - T_n^* y)\|_{\mathcal{H}} \\
&\leq \|P_x q_i^{[1]}(S_n)(Sf - T_n^* y)\|_{\mathcal{H}} + \|P_x q_i^{[1]}(S_n)(S_n - S)f\|_{\mathcal{H}} \\
&\leq (C_\epsilon + C_\delta \kappa^\mu R)\gamma_n \left|\left(p_i^{[1]}\right)'(0)\right|.
\end{aligned}$$

Finally, we have

$$\begin{aligned}
\|P_x^\perp(f - f_i^{[1]})\|_{\mathcal{H}} &\leq x^{-1}\|P_x^\perp S_n(f - f_i^{[1]})\|_{\mathcal{H}} \\
&\leq x^{-1}\left\{\|S_n f_i^{[1]} - T_n^* y\|_{\mathcal{H}} + \|P_x(T_n^* y - S_n f)\|_{\mathcal{H}}\right\} \\
&\leq x^{-1}\left\{\|S_n f_i^{[1]} - T_n^* y\|_{\mathcal{H}} + (C_\epsilon + C_\delta \kappa^\mu R)\gamma_n\right\}
\end{aligned}$$

and thus the first inequality is proven.

For the second inequality we use

$$\|S_n^{1/2}(f - f_i^{[1]})\|_{\mathcal{H}} \leq \|P_x S_n^{1/2}(f - \bar{f}_i)\|_{\mathcal{H}} + \|P_x S_n^{1/2}(\bar{f}_i - f_i^{[1]})\|_{\mathcal{H}} + \|P_x^{\perp} S_n^{1/2}(f - f_i^{[1]})\|_{\mathcal{H}}.$$

In the same way as before we derive bounds for the three terms:

$$\|P_x S_n^{1/2}(f - \bar{f}_i)\|_{\mathcal{H}} \leq x^{1/2} C_\delta \mu \kappa^{\mu-1} R \gamma_n + x^{\mu+1/2} R,$$

$$\|P_x S_n^{1/2}(\bar{f}_i - f_i^{[1]})\|_{\mathcal{H}} \leq x^{1/2}(C_\epsilon + C_\delta R \kappa^\mu)\gamma_n \left| \left(p_i^{[1]}\right)'(0) \right|,$$

$$\|P_x^{\perp} S_n^{1/2}(f - f_i^{[1]})\|_{\mathcal{H}} \leq x^{-1/2}\{\|S_n f_i^{[1]} - T_n^* y\|_{\mathcal{H}} + (C_\epsilon + C_\delta \kappa^\mu R)\gamma_n\},$$

completing the proof. ■

Lemma 12 depends on $(p_i^{[1]})'(0)$ and hence an upper bound for this term is needed:

Lemma 13 *Assume that $C_x \in (0, 1]$ is such that $x_* = (C_x \gamma)^{1/(\mu+1)} < x_{1,a^*-1}^{[1]}$ and $C > C_\epsilon + C_x R + C_\delta(\mu+1)\kappa^\mu R$. Under the conditions of the theorem it holds for $\mu \geq 0$*

$$\left| \left(p_{a^*}^{[1]}\right)'(0) \right| \leq \gamma_n^{-1/(\mu+1)} \left[C_x^{-1/(\mu+1)} \left\{ 1 - \frac{C_\epsilon + C_x R + C_\delta(\mu+1)\kappa^\mu R}{C} \right\}^{-2} + \left\{ \frac{(2\mu+2)^{\mu+1} R}{C - C_\delta(\mu+1)\kappa^\mu R + C_\epsilon} \right\}^{1/(\mu+1)} \right]$$

Proof: The proof is done in two steps by using the inequality $\left| \left(p_{a^*}^{[1]}\right)'(0) \right| \leq \left| \left(p_{a^*-1}^{[1]}\right)'(0) \right| + \left| \left(p_{a^*}^{[1]}\right)'(0) - \left(p_{a^*-1}^{[1]}\right)'(0) \right|$.

Consider first $a^* > 1$.

We will bound $\|S_n f_{a^*-1}^{[1]} - T_n^* y\|_{\mathcal{H}}$ from above. Define $z = x_{1,a^*-1}^{[1]}$ and $\phi_i(x) = p_i^{[1]}(x)(z-x)^{-1/2} z^{1/2}$, $0 \leq x \leq z$. Due to Lemma 9 (vi) it holds that $\sup_{0 \leq x \leq z} x^\nu \phi_{a^*-1}^2(x) \leq \nu^\nu |(p_{a^*-1}^{[1]})'(0)|^{-\nu}$, $\nu \geq 0$. The proof of Lemma 3.7 in Hanke (1995) shows that

$$\left[p_{a^*-1}^{[1]}, p_{a^*-1}^{[1]} \right]_0^{1/2} \leq \|P_z \phi_{a^*-1}(S_n) T_n^* y\|_{\mathcal{H}}.$$

This yields with (SH)

$$\begin{aligned}
&\|S_n f_{a^*-1}^{[1]} - T_n^* y\|_{\mathcal{H}} = \left[p_{a^*-1}^{[1]}, p_{a^*-1}^{[1]} \right]_0^{1/2} \leq \|P_z \phi_{a^*-1}(S_n) T_n^* y\|_{\mathcal{H}} \\
&\leq \|P_z \phi_{a^*-1}(S_n) S f\|_{\mathcal{H}} + \|P_z \phi_{a^*-1}(S_n)(T_n^* y - S f)\|_{\mathcal{H}} \\
&\leq \|P_z \phi_{a^*-1}(S_n) S f\|_{\mathcal{H}} + C_\epsilon \gamma_n \left(\sup_{0 \leq x \leq z} \phi_{a^*-1}^2 \right)^{1/2} \\
&\leq \|P_z \phi_{a^*-1}(S_n) S_n^{\mu+1} u\|_{\mathcal{H}} + \|P_z \phi_{a^*-1}(S_n)(S_n^{\mu+1} - S^{\mu+1}) u\|_{\mathcal{H}} + C_\epsilon \gamma_n \\
&\leq R \left\{ \left(\sup_{0 \leq x \leq z} x^{2\mu+2} \phi_{a^*-1}^2 \right)^{1/2} + C_\delta(\mu+1)\kappa^\mu \gamma_n \left(\sup_{0 \leq x \leq z} \phi_{a^*-1}^2 \right)^{1/2} \right\} + C_\epsilon \gamma_n \\
&\leq \left| \left(p_{a^*-1}^{[1]}\right)'(0) \right|^{-\mu-1} (2\mu+2)^{\mu+1} R + \{C_\delta(\mu+1)\kappa^\mu R + C_\epsilon\}\gamma_n.
\end{aligned}$$

This gives together with $C\gamma_n \leq \|S_n f^{[1]}_{a^*-1} - T_n^* y\|_{\mathcal{H}}$

$$C\gamma_n \leq \left| \left(p^{[1]}_{a^*-1} \right)' (0) \right|^{-\mu-1} (2\mu+2)^{\mu+1} R + \left\{ C_\delta (\mu+1) \kappa^\mu R + C_\epsilon \right\} \gamma_n .$$

If $C > C_\delta(\mu+1)\kappa^\mu R + C_\epsilon$ we finally have

$$\left| \left(p^{[1]}_{a^*-1} \right)' (0) \right| \leq \gamma_n^{-1/(\mu+1)} \left\{ \frac{(2\mu+2)^{\mu+1} R}{C - C_\delta(\mu+1)\kappa^\mu R + C_\epsilon} \right\}^{1/(\mu+1)} . \tag{16}$$

If $a^* = 1$ it holds $p^{[1]}_{a^*-1} = 1$ and thus $\left| \left(p^{[1]}_{a^*-1} \right)' (0) \right| = 0$ and the inequality (16) is true as well.

We will derive an upper bound on $\left| \left(p^{[1]}_{a^*} \right)' (0) - \left(p^{[1]}_{a^*-1} \right)' (0) \right|$. Due to Corollary 2.6 of Hanke (1995) we have

$$\left| \left(p^{[1]}_{a^*-1} \right)' (0) - \left(p^{[1]}_{a^*} \right)' (0) \right| \leq \frac{\left[p^{[1]}_{a^*-1}, p^{[1]}_{a^*-1} \right]_0}{\left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1} \right]_1} . \tag{17}$$

We have $0 \leq x \leq x^{[1]}_{1,a^*-1} < x^{[2]}_{1,a^*-1}$ due to the interlacing property of the roots in Lemma 9 (i) and thus $0 \leq p^{[2]}_{a^*-1}(x) \leq 1$ for $0 \leq x \leq x^{[2]}_{1,a^*-1}$. With that we get with (SH)

$$\begin{aligned}
&\|p^{[1]}_{a^*-1}(S_n) T_n^* y\|_{\mathcal{H}} \leq \left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1} \right]_0^{1/2} \\
&\leq \|P_x p^{[2]}_{a^*-1}(S_n) T_n^* y\|_{\mathcal{H}} + x^{-1/2} \|P_x^\perp S_n^{1/2} p^{[2]}_{a^*-1}(S_n) T_n^* y\|_{\mathcal{H}} \\
&\leq \|P_x p^{[2]}_{a^*-1}(S_n)(T_n^* y - Sf)\|_{\mathcal{H}} + \|P_x p^{[2]}_{a^*-1}(S_n) S^{\mu+1} u\|_{\mathcal{H}} + x^{-1/2} \left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1} \right]_1^{1/2} \\
&\leq C_\epsilon \gamma_n + R \left\{ C_\delta (\mu+1) \kappa^\mu \gamma_n + x^{\mu+1} \right\} + x^{-1/2} \left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1} \right]_1^{1/2} .
\end{aligned}$$

For the choice $x_* = (C_x \gamma)^{1/(\mu+1)}$ we get

$$\left[p^{[1]}_{a^*-1}, p^{[1]}_{a^*-1} \right]_0^{1/2} \leq \gamma_n \left\{ C_\epsilon + C_\delta(\mu+1)\kappa^\mu R + C_x \right\} + x_*^{-1/2} \left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1} \right]_1^{1/2} .$$

It holds $\left[p^{[1]}_{a^*-1}, p^{[1]}_{a^*-1} \right]_0^{1/2} = \|S_n f^{[1]}_{a^*-1} - T_n^* y\|_{\mathcal{H}} \geq C\gamma_n$. This yields with $C > C_\epsilon + C_x R + C_\delta(\mu+1)\kappa^\mu R$

$$\left[p^{[1]}_{a^*-1}, p^{[1]}_{a^*-1} \right]_0 \leq \gamma_n^{-1/(\mu+1)} C_x^{-1/(\mu+1)} \left\{ 1 - \frac{C_\epsilon + C_x R + C_\delta(\mu+1)\kappa^\mu R}{C} \right\}^{-2} \left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1} \right]_1 .$$

Together with (17) we have

$$\left| \left(p^{[1]}_{a^*-1} \right)' (0) - \left(p^{[1]}_{a^*} \right)' (0) \right| \leq \gamma_n^{-1/(\mu+1)} C_x^{-1/(\mu+1)} \left\{ 1 - \frac{C_\epsilon + C_x R + C_\delta(\mu+1)\kappa^\mu R}{C} \right\}^{-2} .$$

Combining this with (16) completes the proof. ■

B.1.3 Proof of Theorem 1

The proof is an application of Lemmas 11 — 13 to (10). First note that $r \geq 3/2$ implies $\mu \geq 1$ and thus this condition in Lemma 12 holds.

Let us choose $x_* = (C_x\gamma_n)^{1/(\mu+1)}$. Lemma 9 (v) shows that $\left|\left(p_i^{[r]}\right)'(0)\right| = \sum_{j=1}^{i}(x_{j,i}^{[r]})^{-1}$ for $i = 1, \ldots, n$, $r \in \mathbb{N}_0$. Thus it holds $\left|\left(p_i^{[1]}\right)'(0)\right|^{-1} \leq x_{1,i}^{[1]}$.

Equation (16) thus shows that C_x can be chosen small enough such that

$$x_* \leq \left|\left(p_{a^*-1}^{[1]}\right)'(0)\right|^{-1} \leq x_{1,a^*-1}^{[1]}$$

and $C_x < 1$, which makes the first condition in Lemma 11 and 13 hold true. The choice $C = C_\epsilon + (\mu+1)\kappa^\mu R(1 + C_\delta)$ gives the second condition.

Now we need to check the remaining condition of Lemma 12, namely that a C_z can be chosen such that $(C_z\gamma_n)^{1/(\mu+1)} \leq x_{1,a^*}^{[1]}$ is true. Lemma 13 yields a $C_z > 0$ such that $C_z\gamma_n^{1/(\mu+1)} \leq \left|\left(p_{a^*}^{[1]}\right)'(0)\right|^{-1} \leq x_{1,a^*}^{[1]}$. Denote $z_* = (C_z\gamma_n)^{1/(\mu+1)}$ and Lemma 12 can be applied.

To ease notation we will denote everything in the derived bounds that does not depend on γ_n as a constant c_j, $j \in \mathbb{N}$. Thus we get by combining Lemmas 12 and 13 that with probability at least $1 - \nu$

$$\begin{aligned}\|f - f_{a^*}^{[1]}\|_{\mathcal{H}}^2 &\leq c_1\gamma_n^{\mu/(\mu+1)} + c_2\gamma_n + c_3\gamma_n^{1-1/(\mu+1)} + c_4\gamma_n\left|\left(p_{a^*}^{[1]}\right)'(0)\right| \\ &\leq c_1\gamma_n^{\mu/(\mu+1)} + c_2\gamma_n + c_3\gamma_n^{\mu/(\mu+1)} + c_5\gamma_n^{1-1/(\mu+1)} = O\{\gamma_n^{\mu/(\mu+1)}\}\end{aligned}$$

and

$$\begin{aligned}&\|S_n^{1/2}(f - f_{a^*}^{[1]})\|_{\mathcal{H}}^2 \\ &\leq c_6\gamma_n^{(\mu+1/2)/(\mu+1)} + c_7\gamma_n^{1/(2\mu+2)}\gamma_n + c_8\gamma_n^{-1/(2\mu+2)}\gamma_n + c_9\gamma_n^{1/(2\mu+1)}\gamma_n\left|\left(p_{a^*}^{[1]}\right)'(0)\right| \\ &\leq c_6\gamma_n^{(\mu+1/2)/(\mu+1)} + c_7\gamma_n^{(2\mu+3)/(2\mu+2)} + c_8\gamma_n^{(2\mu+1)/(2\mu+2)} + c_10\gamma_n^{1+1/(2\mu+2)-1/(\mu+1)} \\ &= O\{\gamma_n^{(2\mu+1)/(2\mu+2)}\}.\end{aligned}$$

Finally Lemma 11 gives

$$\|f_{a^*} - f_{a^*}^{[1]}\|_{\mathcal{H}}^2 = O\{\gamma_n^{\mu/(\mu+1)}\}, \quad \|S_n^{1/2}(f_{a^*} - f_{a^*}^{[1]})\|_{\mathcal{H}} = O\{\gamma_n^{(2\mu+1)/(2\mu+2)}\}.$$

Combining the above with (10) yields

$$\begin{aligned}&\|f - f_{a^*}\|_{\mathcal{H}}^2 = O\{\gamma_n^{\mu/(\mu+1)}\}, \\ &\|f^* - f_{a^*}\|_2^2 = O\{\gamma^{1/2}\gamma_n^{\mu/(\mu+1)}\} + O\{\gamma_n^{(2\mu+1)/(2\mu+2)}\} = O\{\gamma_n^{(2\mu+1)/(2\mu+2)}\},\end{aligned}$$

completing the proof with $\mu = r - 1/2$. ■

B.2 Proof of Theorem 2

The overall design of this proof is similar to the one of Theorem 1 and makes heavy use of results obtained in Blanchard and Krämer (2010b).

B.2.1 Preparation for the Proof

The stopping index (7) can be reformulated with $\mu = r - 1/2$ as

$$a^* = \min\{1 \leq a \leq n : \|S_n f_a^{[1]} - T_n^* y\|_{\mathcal{H}} \leq C\zeta_n\}, \tag{18}$$

with $\zeta_n = \max\{\sqrt{\lambda_n d_\lambda}\gamma_n, \lambda_n^{\mu+1}\}$.

We will derive the result in a similar way to Theorem 1. First it holds due to (10)

$$\begin{aligned}\|f_{a^*} - f^*\|_2 &= \|S^{1/2}(f_{a^*} - f)\|_{\mathcal{H}} \leq \|S^{1/2}(f_{a^*} - f_{a^*}^{[1]})\|_{\mathcal{H}} + \|S^{1/2}(f_{a^*}^{[1]} - f)\|_{\mathcal{H}} \\ &\leq C_\psi \lambda^{1/2}\|f_{a^*} - f_{a^*}^{[1]}\|_{\mathcal{H}} + C_\psi \|S_n^{1/2}(f_{a^*} - f_{a^*}^{[1]})\|_{\mathcal{H}} + \|S^{1/2}(f_{a^*}^{[1]} - f)\|_{\mathcal{H}}. \end{aligned} \tag{19}$$

Now we prove the analogue versions of Lemma 11 — 13. The comments regarding those lemmas also hold here.

Lemma 14 *Let $x = C_x\lambda_n$, $C_x > 0$ such that $0 < x < x_{1,a^*-1}^{[2]}$. Choose $C > \tilde{c}_2$, with $\tilde{c}_1 = R\max\{1, C_\psi^2, \mu\kappa^{\mu-1}C_\delta\}$ and $\tilde{c}_2 = 2\max\{C_\psi C_\epsilon\sqrt{C_x+1}, \tilde{c}_1 C_x(C_x^\mu + 1)\}$. Then it holds*

$$\begin{aligned}\|f_{a^*} - f_{a^*}^{[1]}\|_{\mathcal{H}} &\leq \frac{C^3}{C_x(C-\tilde{c}_2)^2}\lambda_n^{-1}\zeta_n, \\ \|S_n^{1/2}(f_{a^*} - f_{a^*}^{[1]})\|_{\mathcal{H}} &\leq \frac{C^2}{C_x^{1/2}(C-\tilde{c}_2)}\lambda_n^{-1/2}\zeta_n.\end{aligned}$$

Proof: According to the proof of Lemma 14 we can focus on the case $0 < a^* < n$. Furthermore we have due to (12)

$$\|f_{a^*} - f_{a^*}^{[1]}\|_{\mathcal{H}} \leq \frac{\left[p_{a^*-1}^{[2]}, p_{a^*-1}^{[2]}\right]_0}{\left[p_{a^*-1}^{[2]}, p_{a^*-1}^{[2]}\right]_1}\|p_{a^*}^{[1]}(S_n)T_n^* y\|_{\mathcal{H}}. \tag{20}$$

Using Lemma A.3 in Blanchard and Krämer (2010b) (and the first line of its proof) we have for $0 < x < x_{1,a^*-1}^{[2]}$ with $\tilde{c}_1 = R\max\{1, C_\psi^2, \mu\kappa^{\mu-1}C_\delta\}$

$$\left[p_{a^*-1}^{[2]}, p_{a^*-1}^{[2]}\right]_0^{1/2} \leq C_\psi C_\epsilon\sqrt{x+\lambda_n}\sqrt{d_{\lambda_n}}\gamma_n + \tilde{c}_1 x\{x^\mu + Z_\mu(\lambda_n)\} + x^{-1/2}\left[p_{a^*-1}^{[2]}, p_{a^*-1}^{[2]}\right]_1^{1/2}. \tag{21}$$

Here we define $Z_\mu(\lambda) = \lambda^\mu \mathbb{I}(\mu \leq 1) + \gamma_n \mathbb{I}(\mu > 1)$. Note that under the assumptions of the theorem it holds $Z_\mu(\lambda_n) \leq \lambda_n^\mu$.

Choosing $x = C_x\lambda_n$ yields in (21)

$$\begin{aligned}
&\left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1}\right]_0^{1/2} \\
&\leq C_\psi C_\epsilon \sqrt{C_x+1}\sqrt{\lambda_n d_{\lambda_n}}\gamma_n + \tilde{c}_1 C_x (C_x^\mu + 1)\lambda_n^{\mu+1} + C_x^{-1/2}\lambda_n^{-1/2}\left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1}\right]_1^{1/2} \\
&\leq \tilde{c}_2 \max\{\sqrt{\lambda_n d_{\lambda_n}}\gamma_n, \lambda_n^{\mu+1}\} + C_x^{-1/2}\lambda_n^{-1/2}\left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1}\right]_1^{1/2} \\
&= \tilde{c}_2\zeta_n + C_x^{-1/2}\lambda_n^{-1/2}\left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1}\right]_1^{1/2},
\end{aligned}$$

with $\tilde{c}_2 = 2\max\{C_\psi C_\epsilon\sqrt{C_x+1}, \tilde{c}_1 C_x(C_x^\mu+1)\}$. Due to the stopping condition (18) we know that

$$\left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1}\right]_0^{1/2} \geq \left[p^{[1]}_{a^*-1}, p^{[1]}_{a^*-1}\right]_0^{1/2} = \|S_n f_a^{[1]} - T_n^* y\|_{\mathcal{H}} \geq C\zeta_n.$$

This gives

$$\left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1}\right]_0^{1/2} \leq \frac{C}{\sqrt{C_x}(C-\tilde{c}_2)}\lambda_n^{-1/2}\left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1}\right]_1^{1/2}. \tag{22}$$

Plugging this into (20) yields together with the definition of the stopping index a^*

$$\|f_{a^*} - f^{[1]}_{a^*}\|_{\mathcal{H}} \leq \frac{C^3}{C_x(C-\tilde{c}_2)^2}\lambda_n^{-1}\zeta_n.$$

In a similar way we derive for the second case

$$\|S_n^{1/2}(f_{a^*} - f^{[1]}_{a^*})\|_{\mathcal{H}} = \frac{\left[p^{[1]}_{a^*}, p^{[1]}_{a^*}\right]_0}{\left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1}\right]_1^{1/2}} \leq \frac{\left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1}\right]_0^{1/2}}{\left[p^{[2]}_{a^*-1}, p^{[2]}_{a^*-1}\right]_1^{1/2}}\left[p^{[1]}_{a^*}, p^{[1]}_{a^*}\right]_0.$$

An application of (22) yields

$$\|S_n^{1/2}(f_{a^*} - f^{[1]}_{a^*})\|_{\mathcal{H}} \leq \frac{C^2}{\sqrt{C_x}(C-\tilde{c}_2)}\lambda_n^{-1/2}\zeta_n.$$

■

Lemma 15 *Denote* $\tilde{c}_1 = R\max\{1, C_\psi^2, \mu\kappa^{\mu-1}C_\delta\}$. *For any* $i = 1, \ldots, n$ *and* $0 < x < x^{[1]}_{1,i}$ *we have under the conditions of the theorem*

$$\begin{aligned}
\|S^{1/2}(f_i^{[1]} - f)\|_{\mathcal{H}} \leq C_\psi &\left[C_\delta + \sqrt{2}C_\epsilon + \lambda_n\left\{C_\delta\left|\left(p_i^{[1]}\right)'(0)\right| + \sqrt{2}C_\epsilon x^{-1}\right\}\right]\sqrt{d_{\lambda_n}}\gamma_n \\
&+ \tilde{c}_1(\sqrt{x}+\sqrt{\lambda_n})(x^\mu + \lambda_n^\mu) + (1+\sqrt{x^{-1}\lambda_n})x^{-1/2}\|S_n f_i^{[1]} - T_n^* y\|_{\mathcal{H}}.
\end{aligned}$$

Proof: Follow the proof of Lemma A.2 in Blanchard and Krämer (2010b). Note that $Z_\mu(\lambda_n) \leq \lambda_n^\mu$. ■

Lemma 16 *Let $C > \max\{C_\psi C_\epsilon, \tilde{c}_2\}$, where $\tilde{c}_2$ is given in Lemma 14. Choose $x = C_x\lambda_n$ such that $0 < x \leq x^{[1]}_{1,a^*-1}$. Then there exists a constant $c^* > 0$ such that*

$$\left|\left(p^{[1]}_{a^*}\right)'(0)\right| \leq c^*\lambda_n^{-1}. \tag{23}$$

Proof: In analogue to Lemma 13 we will first derive an upper bound on $\left|\left(p^{[1]}_{a^*-1}\right)'(0)\right|$. Lemma A.1 in Blanchard and Krämer (2010b) yields

$$\begin{aligned}
\|S_n f^{[1]}_{a^*-1} - T_n^* y\|_{\mathcal{H}} \leq\ & R(2\mu+2)^{\mu+1}\max\{1, C_\psi^{2\mu}\}\left|\left(p^{[1]}_{a^*-1}\right)'(0)\right|^{-\mu-1} \\
& + 2R\mu\kappa^{\mu-1}\max\{1, C_\delta, C_\psi^{2\mu}, C_\psi^{2\mu}C_\delta\}\left|\left(p^{[1]}_{a^*-1}\right)'(0)\right|^{-1} Z_\mu(\lambda_n) \\
& + C_\epsilon C_\psi\left\{\left|\left(p^{[1]}_{a^*-1}\right)'(0)\right|^{-1/2} + \sqrt{\lambda_n}\right\}\sqrt{d_{\lambda_n}}\gamma_n.
\end{aligned}$$

Denote $\tilde{c}_3 = R(2\mu+2)^{\mu+1}\max\{1, C_\psi^{2\mu}\}$ and $\tilde{c}_4 = 2R\mu\kappa^{\mu-1}\max\{1, C_\delta, C_\psi^{2\mu}, C_\psi^{2\mu}C_\delta\}$.

The definition of a^* gives $C\zeta_n \leq \|S_n f^{[1]}_{a^*-1} - T_n^* y\|_{\mathcal{H}}$. Combining both inequalities, setting $x = C_x\lambda_n$ and keeping $\sqrt{\lambda_n d_{\lambda_n}}\gamma_n \leq \zeta_n$ in mind gives

$$\begin{aligned}
(C - C_\psi C_\epsilon C_\lambda)\zeta_n \leq\ & \tilde{c}_3\left|\left(p^{[1]}_{a^*-1}\right)'(0)\right|^{-\mu-1} + \tilde{c}_4\left|\left(p^{[1]}_{a^*-1}\right)'(0)\right|^{-1}\lambda_n^\mu \\
& + C_\epsilon C_\psi\left|\left(p^{[1]}_{a^*-1}\right)'(0)\right|^{-1/2}\sqrt{d_{\lambda_n}}\gamma_n \\
\leq\ & 3\max\left\{\tilde{c}_3\left|\left(p^{[1]}_{a^*-1}\right)'(0)\right|^{-\mu-1}, \tilde{c}_4\left|\left(p^{[1]}_{a^*-1}\right)'(0)\right|^{-1}\lambda_n^\mu,\right. \\
& \left. C_\epsilon C_\psi\left|\left(p^{[1]}_{a^*-1}\right)'(0)\right|^{-1/2}\sqrt{d_{\lambda_n}}\gamma_n\right\}.
\end{aligned}$$

Now we assume that the maximum on the right hand side is attained in each of the three possible cases

$$\begin{aligned}
\left|\left(p^{[1]}_{a^*-1}\right)'(0)\right| &\leq \{3(C - C_\epsilon C_\psi)^{-1}\tilde{c}_3\}^{1/(\mu+1)}\zeta_n^{-1/(\mu+1)}, \\
\left|\left(p^{[1]}_{a^*-1}\right)'(0)\right| &\leq 3(C - C_\epsilon C_\psi)^{-1}\tilde{c}_4\zeta_n^{-1}\lambda_n^\mu, \\
\left|\left(p^{[1]}_{a^*-1}\right)'(0)\right| &\leq 9(C - C_\epsilon C_\psi)^{-2}C_\epsilon^2 C_\psi^2\zeta_n^{-2}d_{\lambda_n}\gamma_n^2.
\end{aligned}$$

Take $\tilde{c}_5 = \max[\{3(C - C_\epsilon C_\psi)^{-1}\tilde{c}_3\}^{1/(\mu+1)}, 3(C - C_\epsilon C_\psi)^{-1}\tilde{c}_4, 9(C - C_\epsilon C_\psi)^{-2}C_\epsilon^2 C_\psi^2]$.

It is easy to see that $\zeta_n^{-1/(\mu+1)}, \zeta_n^{-1}\lambda_n^{\mu}$ and $\zeta_n^{-2} d_{\lambda_n}\gamma_n^2$ are all bound from above by λ_n^{-1}. Hence we get

$$\left|\left(p_{a^*-1}^{[1]}\right)'(0)\right| \leq \tilde{c}_5 \lambda_n^{-1}. \tag{24}$$

For the final step in the proof we have due to (17)

$$\left|\left(p_{a^*-1}^{[1]}\right)'(0) - \left(p_{a^*}^{[1]}\right)'(0)\right| \leq \frac{\left[p_{a^*-1}^{[1]}, p_{a^*-1}^{[1]}\right]_0}{\left[p_{a^*-1}^{[2]}, p_{a^*-1}^{[2]}\right]_1}.$$

It holds $\left[p_{a^*-1}^{[1]}, p_{a^*-1}^{[1]}\right]_0 \leq \left[p_{a^*-1}^{[2]}, p_{a^*-1}^{[2]}\right]_0$ and hence (22) yields

$$\left|\left(p_{a^*-1}^{[1]}\right)'(0) - \left(p_{a^*}^{[1]}\right)'(0)\right| \leq \frac{C^2}{C_x(C-\tilde{c}_2)^2}\lambda_n^{-1}. \tag{25}$$

The proof is complete by combining (24) and (25). ■

B.2.2 Proof of Theorem 2

We first restrict ourselves to the set where all concentration inequalities stated in the theorem hold simultaneously with probability at least $1-\nu$, $\nu \in (0,1]$. We only proof the convergence rates in the $\mathcal{L}^2$-norm, the corresponding rates in the $\mathcal{H}$-norm are done in the same way.

The theorem is proven by an application of Lemmas 14—16. To that end we need to check the conditions of those. Equation (24) and the proof of Theorem 1 show that we can take $C_x = \min\{1/2, \tilde{c}_5\}$ to fulfill $0 < x \leq x_{1,a^*-1}^{[1]}$. Furthermore we can take $C = 4R\max\{1, C_\psi^2(\nu), (r-1/2)\kappa^{r-3/2}C_\delta(\nu), 2^{-1/2}R^{-1}C_\psi(\nu)C_\epsilon(\nu)\}$ and the conditions of Lemma 14 and 16 hold. Note that $x_{1,a^*-1}^{[1]} \leq x_{1,a^*-1}^{[2]}$ due to the interlacing property of the roots, see Lemma 9 (i).

For Lemma 15 we need to find a $0 < z < x_{1,a^*}^{[1]}$. By Lemma 16 there exists a constant $c^* > 0$ such that

$$(c_*)^{-1}\lambda_n \leq \left|\left(p_{a^*}^{[1]}\right)'(0)\right|^{-1} \leq x_{1,a^*}^{[1]},$$

hence we choose $C_z = \min\{1/2, 1/c^*\}$. Now, applying Lemmas 14—16 to (19) gives the result (we again denote any constant that does not depend on n with C_i, $i \in \mathbb{N}$)

$$\begin{aligned}
&\|f_{a^*} - f^*\|_2 \leq C_\psi \lambda_n^{1/2}\|f_{a^*} - f_{a^*}^{[1]}\|_{\mathcal{H}} + C_\psi \|S_n^{1/2}(f_{a^*} - f_{a^*}^{[1]})\|_{\mathcal{H}} + \|S^{1/2}(f_{a^*}^{[1]} - f)\|_{\mathcal{H}} \\
\leq& C_1\lambda_n^{-1/2}\zeta_n + C_2\lambda_n\left|\left(p_i^{[1]}\right)'(0)\right|\sqrt{d_{\lambda_n}}\gamma_n + C_3\lambda_n^{\mu+1/2} + C_4\lambda_n^{-1/2}\|S_n f_{a^*}^{[1]} - T_n^* y\|_{\mathcal{H}}. \\
\leq& C_5\lambda_n^{-1/2}\zeta_n + C_6\sqrt{d_{\lambda_n}}\gamma_n + C_4\lambda_n^{\mu+1/2} \leq \max\{C_4, C_5, C_6\}\lambda_n^{-1/2}\zeta_n.
\end{aligned}$$

The error bound in the $\mathcal{H}$-norm is proven in an analogue fashion. ■

B.3 Proof of Corollary 3

Take $\lambda_n = \gamma_n^{2/(2r+s)}$. It is immediate that $\lambda_n^{r-1/2} = \gamma_n^{(2r-1)/(2r+s)} \geq \gamma_n$ for n sufficiently large, hence the inequality (6) holds as soon as $\gamma_n \leq 1$. Then we have by Theorem 2 that

$$\|f_{\widehat{\alpha}_{a^*}} - f^*\|_2 = O\left\{\lambda_n^{-1/2}\zeta_n(\lambda_n)\right\} = O\left\{\gamma_n^{2r/(2r+s)}\right\}.$$

■

B.4 Proof of Corollary 4

Set $\lambda_n = \gamma_n^{1/r}\log\{1/(2r)\gamma_n^{-2}\}$. It is immediate that $\lambda_n \to 0$ as γ_n converges to zero. For $r = 1/2$ condition (6) holds trivially. Let $r > 1/2$, then we have

$$\lambda_n^{r-1/2} = \gamma_n^{(r-1)/r}\log\{(r-1/2)/(2r)\gamma_n^{-2}\} \geq \gamma_n,$$

This is equivalent to $2r - 1 \geq 2\exp(\gamma_n)\gamma_n^2$, which holds for n sufficiently large and $r > 1/2$.

For the convergence rate we first show that $d_{\lambda_n}\gamma_n^2 \leq \lambda_n^{2r}$. We have

$$d_{\lambda_n}\gamma_n^2 = \log\left\{1 + \frac{a}{\gamma_n^{1/r}\log^{1/r}(1/2\gamma_n^{-1})}\right\}\gamma_n^2 \leq \log\left(\gamma_n^{-2}\right)\gamma_n^2.$$

Equivalently we need $a^r \leq \gamma_n(\gamma_n^{-2}-1)^r\log(1/2\gamma_n^{-2})$. As γ_n converges to zero $\gamma_n(\gamma_n^{-2}-1)^r$ goes to infinity for any $r > 1/2$. Hence for suitably large n it holds $\lambda_n^r \geq \sqrt{d_{\lambda_n}}\gamma_n$. Then the convergence rate is $\lambda_n^{-1/2}\zeta_n(\lambda_n) = \lambda_n^r = \gamma_n\log(1/2\gamma^{-2})$.

Because the convergence rate does not depend on $r \geq 1/2$ we can set $r = 1/2$. ■

B.5 Proof of Theorem 5

B.5.1 Preparation for the Proof

We denote with $\mathrm{tr}(A)$ the trace of a trace class operator $A : \mathcal{H} \to \mathcal{H}$ and the tensor product $(f_1 \otimes f_2)h = \langle f_1, h\rangle_{\mathcal{H}} f_2$ for functions $f_1, f_2, h \in \mathcal{H}$. We use the notation $k_t = k(\cdot, X_t)$. Note that it holds $\|A\|_{\mathrm{HS}}^2 = \mathrm{tr}(A^*A)$ for a Hilbert-Schmidt operator A.

The general structure for the proof is as follows: To derive the concentration inequalities we need $\mathrm{E}\{\|S_n - S\|_{\mathcal{L}}^2\}$ as well as $\mathrm{E}\{\|T_n^* y - Sf\|_{\mathcal{H}}^2\}$ for Markov's inequality. With the help of Lemma 17 we derive representations of these expectations as sums by using the stationarity of X_t. Under normality assumptions Lemma 18 gives upper bounds for the summands in terms of the autocorrelation function of X_t using (D1). Assuming the special structure implied by (D2) Corollary 19 shows how these upper bound depend on the long range dependence coefficient $q > 0$. To get explicit convergence rates Lemma 20 gives a simple result stating how sums of these autocorrelation functions depend on q. An analogue result is given in Lemma 21 for the expectation of a warped norm.

The next technical lemma derives some simple identities for the operator S under the Hilbert-Schmidt norm, its trace and the kernel k. These will be used to find alternate representations of $\mathrm{E}\{\|S_n - S\|_{\mathcal{L}}^2\}$ and $\mathrm{E}\{\|T_n^* y - Sf\|_{\mathcal{H}}^2\}$ under stationarity.

Lemma 17 *Under the assumptions (K1) and (K2) the following hold*

(i) $\mathrm{tr}\{(k_t \otimes k_t)(k_s \otimes k_s)\} = k^2(X_t, X_s)$,

(ii) $\|S\|_{\mathrm{HS}}^2 = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} k^2(x,y)\mathrm{dP}^{X_0}(x)\mathrm{dP}^{X_0}(y)$,

(iii) $\mathrm{E}[\mathrm{tr}\{(k_0 \otimes k_0)S\}] = \|S\|_{\mathrm{HS}}^2$.

(iv) Let X' and X'' be independent and identically distributed and denote $k' = k(\cdot, X')$, $k'' = k(\cdot, X'')$. It holds for $\nu = 1, 2$ and $\lambda > 0$

$$\mathrm{E}\left[\mathrm{tr}^\nu\left\{(S+\lambda)^{-1}k' \otimes k'k'' \otimes k''\right\}\right] = \mathrm{tr}^\nu\{(S+\lambda)^{-1}S^2\}.$$

Proof: (i) Let $\{v_i\}_{i\in\mathbb{N}}$ denote an orthonormal base of $\mathcal{H}$. Then it holds due to the reproducing property (2)

$$\mathrm{tr}\left\{(k_t \otimes k_t)(k_s \otimes k_s)\right\} = \sum_{i=1}^\infty \langle v_i, k_t\rangle_{\mathcal{H}}\langle v_i, k_s\rangle_{\mathcal{H}} k(X_t, X_s) = \left\langle \sum_{i=1}^\infty \langle v_i, k_s\rangle_{\mathcal{H}} v_i, k_t \right\rangle_{\mathcal{H}} k(X_t, X_s).$$

(ii)

$$\begin{aligned}\|S\|_{\mathrm{HS}}^2 &= \sum_{i=1}^\infty \langle Sv_i, Sv_i\rangle_{\mathcal{H}} = \sum_{i=1}^\infty \int_{\mathbb{R}^d} \langle Sv_i, k(\cdot,x)\rangle_{\mathcal{H}}\langle v_i, k(\cdot,x)\rangle_{\mathcal{H}}\mathrm{dP}^X(x) \\ &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \left\langle \sum_{i=1}^\infty \langle v_i, k(\cdot,x)\rangle_{\mathcal{H}} v_i, k(\cdot,y)\right\rangle_{\mathcal{H}} k(x,y)\mathrm{dP}^X(x)\mathrm{dP}^X(y).\end{aligned}$$

The assertion follows because $\mathrm{P}^X = \mathrm{P}^{X_0}$.

(iii)

$$\begin{aligned}\mathrm{E}[\mathrm{tr}\{(k_0 \otimes k_0)S\}] = \mathrm{E}(\langle Sk_0, k_0\rangle_{\mathcal{H}}) &= \mathrm{E}\left(\int_{\mathbb{R}^d} \langle k_0, k(\cdot,x)\rangle_{\mathcal{H}}^2 \mathrm{dP}^X(x)\right) \\ &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} k^2(x,y)\mathrm{dP}^X(x)\mathrm{dP}^{X_0}(y) = \|S\|_{\mathrm{HS}}^2.\end{aligned}$$

(iv) Because S is a compact operator the spectral decomposition $S = \sum_{i=1}^\infty \mu_i \psi_i \otimes \psi_i$ holds (recall that $\{\mu_i, \psi_i\}_{i=1}^\infty$ is the eigensystem of S). Let $k(\cdot, x) = \sum_{i=1}^\infty \alpha_i(x)\psi_i$, $x \in \mathbb{R}^d$. For $\nu = 1$ we have

$$\begin{aligned}\mathrm{E}\left[\mathrm{tr}\{(S+\lambda)^{-1}k' \otimes k'k'' \otimes k''\}\right] &= \mathrm{E}\left\{k(X', X'')\sum_{i=1}^\infty \langle \psi_i, k''\rangle_{\mathcal{H}}\langle (S+\lambda)^{-1}k', \psi_i\rangle_{\mathcal{H}}\right\} \\ &= \sum_{i=1}^\infty \frac{1}{\mu_i + \lambda}\mathrm{E}\left\{k(X', X'')\langle \psi_i, k''\rangle_{\mathcal{H}}\langle \psi_i, k'\rangle_{\mathcal{H}}\right\} \\ &= \sum_{i=1}^\infty \frac{1}{\mu_i + \lambda}\sum_{j=1}^\infty \mathrm{E}\left\{\alpha_j(X')\alpha_i(X')\alpha_j(X'')\alpha_i(X'')\right\}\end{aligned}$$

On the other hand

$$\begin{aligned}\operatorname{tr}\left\{(S+\lambda)^{-1}S^2\right\} &= \operatorname{tr}\left\{(S+\lambda)^{-1}\mathrm{E}(k'\otimes k'k''\otimes k'')\right\}\\ &= \sum_{i=1}^{\infty}\langle(S+\lambda)^{-1}\mathrm{E}\{k(X',X'')\langle\psi_i,k''\rangle_{\mathcal{H}}k'\},\psi_i\rangle_{\mathcal{H}}\\ &= \sum_{i=1}^{\infty}\frac{1}{\mu_i+\lambda}\langle\mathrm{E}\{k(X',X'')\langle\psi_i,k''\rangle_{\mathcal{H}}k'\},\psi_i\rangle_{\mathcal{H}}\\ &= \sum_{i=1}^{\infty}\frac{1}{\mu_i+\lambda}\sum_{j=1}^{\infty}\mathrm{E}\{\alpha_j(X')\alpha_i(X')\alpha_j(X'')\alpha_i(X'')\}\end{aligned}$$

and we are done. The proof for $\nu = 2$ is along the same lines. ■

Denote with g_h the common density of $(X_h, X_0)^{\mathrm{T}}$ and g_0 the density of X_0. The next lemma and the subsequent corollary will be used to show that the summands appearing in Theorem 5 (i) can be linked to the autocorrelation function ρ under the assumptions of normality (D1):

Lemma 18 *Under the assumptions (K1), (K2) and (D1) it holds for $h > 0$ with $\rho_h = \tau_0^{-1}\tau_h$*

$$\int_{\mathbb{R}^{2d}} k^2(x,y)\{g_h(x,y) - g_0(x)g_0(y)\}\mathrm{d}(x,y) \leq \frac{\kappa^2}{\{(4\pi\tau_0)^d\det(\Sigma)\}^{1/2}}\theta^{1/2}(\rho_h),$$

$$\int_{\mathbb{R}^{2d}} k(x,y)f(x)f(y)\{g_h(x,y) - g_0(x)g_0(y)\}\mathrm{d}(x,y) \leq \frac{\kappa M}{\{(4\pi\tau_0)^d\det(\Sigma)\}^{1/2}}\theta^{1/2}(\rho_h),$$

with $\theta(\rho) = 1 + (1-\rho^2)^{-d/2} - 2^{d+1}(4-\rho^2)^{-d/2}$, $\rho \in [0,1)$.

Proof: We will only proof the first inequality, the second one follows in the same way.

By Jensen's inequality and (K2) we know

$$\begin{aligned}&\int_{\mathbb{R}^{2d}} k^2(x,y)\{g_h(x,y) - g_0(x)g_0(y)\}\mathrm{d}(x,y)\\ &\leq \kappa^2\left[\int_{\mathbb{R}^{2d}}\left\{g_h^2(x,y) - 2g_h(x,y)g_0(x)g_0(y) + g_0^2(x)g_0^2(y)\right\}\mathrm{d}(x,y)\right]^{1/2}.\end{aligned}$$

The first and third integral term can readily be calculated as

$$\begin{aligned}\int_{\mathbb{R}^{2d}} g_h^2(x,y)\mathrm{d}(x,y) &= [(4\pi)^d(\tau_0^2-\tau_h^2)^{d/2}\det(\Sigma)]^{-1}\\ \left\{\int_{\mathbb{R}^d} g_0^2(x)\mathrm{d}x\right\}^2 &= \{(4\pi)^d\tau_0^d\det(\Sigma)\}^{-1}.\end{aligned}$$

For the first equality we use $\det(A \otimes \Sigma) = \det(A)^d \det(\Sigma)^2$ for $A \in \mathbb{R}^{2\times 2}$ and thus

$$\int_{\mathbb{R}^{2d}} g_h(x,y) g_0(x) g_0(y) \mathrm{d}(x,y) = \frac{\int_{\mathbb{R}^{2d}} \exp\left(-1/2 z^{\mathrm{T}} G^{-1} z\right) \mathrm{d}z}{(2\pi)^{2d} \det(\Sigma)^2 \tau_0^d (\tau_0^2 - \tau_h^2)^{d/2}}, \tag{26}$$

with

$$G^{-1} = \left\{ \begin{pmatrix} \tau_0 & \tau_h \\ \tau_h & \tau_0 \end{pmatrix}^{-1} + \begin{pmatrix} \tau_0^{-1} & 0 \\ 0 & \tau_0^{-1} \end{pmatrix} \right\} \otimes \Sigma^{-1}.$$

It holds $\det(G) = (4\tau_0^2 - \tau_h^2)^{-d} (\tau_0^4 - \tau_0^2 \tau_h^2)^d \det(\Sigma)^2$. Thus we get with (26)

$$\begin{aligned} \int_{\mathbb{R}^{2d}} g_h(x,y) g_0(x) g_0(y) \mathrm{d}(x,y) &= \frac{(2\pi)^d \tau_0^d (\tau_0^2 - \tau_h^2)^{d/2} \det(\Sigma)}{(2\pi)^{2d} \det(\Sigma)^2 (4\tau_0^2 - \tau_h^2)^{d/2} \tau_0^d (\tau_0^2 - \tau_h^2)^{d/2}} \\ &= \left\{ (2\pi)^d (4\tau_0^2 - \tau_h^2)^{d/2} \det(\Sigma) \right\}^{-1}, \end{aligned}$$

completing the proof by multiplying all terms with $\tau_0^{-d} \tau_0^d$. ■

Using the special structure imposed by (D2) we can link the bounds derived in Lemma 18 to the long range coefficient $q > 0$:

Corollary 19 *Under the assumptions (K1), (K2), (D1) and (D2) it holds for all $h > 0$ and $q > 0$*

$$\int_{\mathbb{R}^{2d}} k^2(x,y) \{ g_h(x,y) - g_0(x) g_0(y) \} \mathrm{d}(x,y) \leq \frac{\kappa^2 d^{1/2}}{\sqrt{(2\pi)^d \det(\Sigma)}} (1 - 4^{-q})^{-1/4(d-2)} |\rho_h|$$

$$\int_{\mathbb{R}^{2d}} k(x,y) f(x) f(y) \{ g_h(x,y) - g_0(x) g_0(y) \} \mathrm{d}(x,y) \leq \frac{\kappa M d^{1/2}}{\sqrt{(2\pi)^d \det(\Sigma)}} (1 - 4^{-q})^{-1/4(d-2)} |\rho_h|.$$

Proof: Recall that $\theta(\rho) = 1 + \{1 - \rho^2\}^{-d/2} - 2^{d+1} \{4 - \rho^2\}^{-d/2}$ for $\rho \in [0, 1)$. We seek to find bounds on θ and the corollary can be proven by an application of Lemma 18.

By assumption (D2) we know there is a ρ_* such that $\rho_h^2 \leq \rho_*^2 < 1$ for all $h > 0$. Thus consider $\rho \in [0, \rho_*]$. We start by finding a constant $C > 0$ with

$$\theta'(\rho) = \rho \left\{ (1 - \rho^2)^{-d/2-1} - 2^{d+1} (4 - \rho^2)^{-d/2-1} \right\} d \leq C \rho^2.$$

Thus C can be taken as $C = d \left\{ (1 - \rho_*^2)^{-d/2-1} - 2^{d+1} (4 - \rho_*^2)^{-d/2-1} \right\}$.

Thus we know that the slope of θ is always less than that of $C\rho^2$. Finally it holds that $\theta(0) = 0$ and thus $0 \leq \theta(\rho) \leq C\rho^2$, $\rho \in [0, \rho_*]$.

Under condition (D2) it holds $\{1 - \rho_*^2\}^{-d/2} \leq \{1 - 2^{-2q}\}^{-d/2}$, completing the proof by using Lemma 18. ■

Having derived bounds for the summands in Theorem 5 (i) in terms of ρ the overall behavior of the sums is still open. The next simple lemma gives insight into this and shows how the convergence rate of the sums crucially depends on $q > 0$:

Lemma 20 *Assume that condition (D2) holds. Then we have*

$$n^{-2}\sum_{h=1}^{n-1}(n-h)|\rho_h| \leq C(q)\begin{cases} n^{-1} & , \quad q>1 \\ n^{-1}\log(n) & , \quad q=1 \\ n^{-q} & , \quad q\in(0,1). \end{cases}, \tag{27}$$

with $C(q)=\zeta(q)\mathbb{I}(q>1)+\{5-\log(4)\}\mathbb{I}(q=1)+\{2(1-q)^{-1}-(2-q)^{-1}+(2-q)^{-1}2^{2-q}\}\mathbb{I}\{q\in(0,1)\}$. *Here* ζ *denotes the Riemann zeta function.*

Proof: Recall that by condition (D2) we have $|\rho_h|\leq(h+1)^{-q}$, $h=0,\ldots,n-1$ for some $q>0$.

First assume $q\in(0,1]$. The integral test for series convergence gives lower and upper bounds for the hyperharmonic series as

$$(1-q)^{-1}\{(n+1)^{1-q}-2^{1-q}\}\leq\sum_{h=2}^{n}h^{-q}\leq 2^{-q}+(1-q)^{-1}\{n^{1-q}-2^{1-q}\}.$$

This yields

$$\begin{aligned} &n^{-2}\sum_{h=1}^{n-1}(n-h)(h+1)^{-q}=n^{-2}\sum_{h=2}^{n}(n+1-h)h^{-q}=n^{-2}\left\{(n+1)\sum_{h=2}^{n}h^{-q}-\sum_{h=2}^{n}h^{-(q-1)}\right\} \\ &\leq n^{-2}\left[(n+1)\left\{2^{-q}+(1-q)^{-1}(n^{1-q}-2^{1-q}\right\}-(2-q)^{-1}\left\{(n+1)^{2-q}-2^{2-q}\right\}\right]. \end{aligned} \tag{28}$$

Now let $q\in(0,1)$, then it holds from (28) and the fact that $n^{-2}\leq n^{-1}\leq n^{-q}$

$$\begin{aligned} &n^{-2}\sum_{h=1}^{n-1}(n-h)(h+1)^{-q} \\ \leq&\frac{n+1}{n^2}\left\{\frac{2^{-q}(1-q)-2^{1-q}}{1-q}\right\}+\frac{n+1}{n^{1+q}}(1-q)^{-1}-\frac{(n+1)^{2-q}}{n^2}(2-q)^{-1}+\frac{1}{n^2}\frac{2^{2-q}}{2-q} \\ \leq&n^{-q}[\{2(1-q)^{-1}-(2-q)^{-1}\}+(2-q)^{-1}2^{2-q}], \end{aligned}$$

due to $2^{-q}(1-q)-2^{1-q}<0$.

For $q=1$ we evaluate the limit

$$\begin{aligned} &\lim_{q\to1\pm}n^{-2}\left[(n+1)\left\{2^{-q}+(1-q)^{-1}(n^{1-q}-2^{1-q}\right\}-(2-q)^{-1}\left\{(n+1)^{2-q}-2^{2-q}\right\}\right] \\ &=(2n^2)^{-1}[3-\log(4)-n\{1+\log(4)\}]+n^{-2}(n+1)\log(n) \\ &\leq\frac{\log(n)}{n}\left[5-\log(4)\right]. \end{aligned}$$

The case $q>1$ is clear because the zeta-function $\zeta(q)$ is defined as the hyperharmonic series with coefficient q. ■

The final preparatory result is used to derive the probablistic bound in Theorem 5 (iv) and is similar to Corollary 19:

Lemma 21 *Under the assumptions (K1), (K2), (D1) and (D2) it holds for $\lambda > 0$*

$$\int\limits_{\mathbb{R}^{2d}} k(x,y)\langle (S+\lambda)^{-1}k(\cdot,x), k(\cdot,y)\rangle_{\mathcal{H}} \left\{f_h(x,y) - f_0(x)f_0(y)\right\} \mathrm{d}(x,y) \leq \tilde{c}|\rho_h| d_\lambda,$$

with $\tilde{c} = \sqrt{d\{1-4^{-q}\}^{-d-1}}\kappa$.

Proof: Denote $\beta(x,y) = \langle (S+\lambda)^{-1}k(\cdot,x), k(\cdot,y)\rangle_{\mathcal{H}}$. By the Cauchy-Schwarz inequality

$$\begin{aligned}\phi_h &= \int\limits_{\mathbb{R}^{2d}} k(x,y)\beta(x,y)\left\{f_h(x,y) - f_0(x)f_0(y)\right\}\mathrm{d}(x,y) \\ &\leq \left[\int k^2(x,y)\beta^2(x,y)f_0(x)f_0(y)\mathrm{d}(x,y)\int\left\{\frac{f_h(x,y)}{\sqrt{f_0(x)f_0(y)}} - \sqrt{f_0(x)f_0(y)}\right\}^2 \mathrm{d}(x,y)\right]^{1/2}. \end{aligned} \tag{29}$$

Denote by X' and X'' two independent copies of X_0 and $k' = k(\cdot, X')$, $k'' = k(\cdot, X'')$. We start by bounding the first integral term in the product:

$$\begin{aligned}\int\limits_{\mathbb{R}^{2d}} k^2(x,y)\beta^2(x,y)f_0(x)f_0(y)\mathrm{d}(x,y) &= \mathrm{E}\left\{k^2(X',X'')\langle (S+\lambda)^{-1}k', k''\rangle_{\mathcal{H}}^2\right\} \\ &= \mathrm{E}\left[\mathrm{tr}^2\left\{(S+\lambda)^{-1}k'\otimes k'k''\otimes k''\right\}\right] \\ &\leq \kappa^2\mathrm{tr}^2\{(S+\lambda)^{-1}S\} = \kappa^2 d_\lambda^2.\end{aligned}$$

In the second to last inequality we used Lemma 17 (iv) and the definition of d_λ.

The second integral in the product in (29) is

$$\int\limits_{\mathbb{R}^{2d}}\left\{\frac{f_h(x,y)}{\sqrt{f_0(x)f_0(y)}} - \sqrt{f_0(x)f_0(y)}\right\}^2\mathrm{d}(x,y) = \int\limits_{\mathbb{R}^{2d}}\frac{f_h(x,y)}{\sqrt{f_0(x)f_0(y)}}\mathrm{d}(x,y) - 1.$$

We proceed in the same way as in the proof of Lemma 18 by using properties of the Gaussian distributions at hand.

First we have

$$F_h(x,y) = \frac{f_h^2(x,y)}{f_0(x)f_0(y)} = \frac{(2\pi)^d\tau_0^d\det(\Sigma)}{(2\pi)^{2d}\det(\Sigma_h)}\exp\left[-\frac{1}{2}(x^{\mathrm{T}},y^{\mathrm{T}})\left\{2\Sigma_h^{-1} - \frac{1}{\tau_0}\begin{pmatrix}\Sigma & 0\\ 0 & \Sigma\end{pmatrix}\right\}^{-1}\begin{pmatrix}x\\ y\end{pmatrix}\right].$$

Denote $G^{-1} = 2\Sigma_h^{-1} - \tau_0^{-1}\begin{pmatrix}\Sigma & 0\\ 0 & \Sigma\end{pmatrix}$. It follows in a similar way as in the proof of Lemma 18 that $\det(G) = \tau_0^{2d}\det^2(\Sigma)$ and $\det(\Sigma_h) = (\tau_0^2 - \tau_h^2)^d\det^2(\Sigma)$. Hence we have

$$\begin{aligned}\int\limits_{\mathbb{R}^{2d}} F_h(x,y)\mathrm{d}(x,y) &= \frac{(2\pi)^d\tau_0^d\det(\Sigma)}{(2\pi)^{2d}(\tau_0^2-\tau_h^2)^d\det^2(\Sigma)}(2\pi)^d\tau_0^d\det(\Sigma) \\ &= \frac{\tau_0^{2d}}{(\tau_0^2-\tau_h^2)^d} = \frac{1}{(1-\rho_h^2)^d}.\end{aligned}$$

Under (D2) we have $\rho_h < 1$ for all $h > 0$ and there is a $\tilde{\rho} = \max_h|\rho_h| \leq 2^{-q} < 1$ and hence it holds $(1-\rho_h^2)^{-d} \leq d(1-4^{-q})^{-d-1}\rho_h^2$ and we are done. ■

B.5.2 Proof of the Theorem

First note that the the operator norm is dominated by the Hilbert-Schmidt norm. By Markov's inequality we have for $\nu \in (0,1]$

$$\mathrm{P}\left(\|S_n - S\|_{\mathrm{HS}}^2 \leq \nu^{-1}\mathrm{E}\|S_n - S\|_{\mathrm{HS}}^2\right) \geq 1-\nu,$$
$$\mathrm{P}\left(\|T_n^* y - Sf\|_{\mathcal{H}}^2 \leq \nu^{-1}\mathrm{E}\|T_n^* y - Sf\|_{\mathcal{H}}^2\right) \geq 1-\nu.$$

(i) It holds due to $S_n = n^{-1}\sum_{t=1}^n k_t \otimes k_t$

$$\mathrm{E}\left(\|S_n - S\|_{\mathrm{HS}}^2\right) = \frac{1}{n^2}\sum_{t,s=1}^n \left(\mathrm{E}[\mathrm{tr}\{(k_t \otimes k_t)(k_s \otimes k_s)\}] - 2\mathrm{E}[\mathrm{tr}\{(k_0 \otimes k_0)S\}] + \|S\|_{\mathrm{HS}}^2\right).$$

For the first summand we get $\mathrm{E}[\mathrm{tr}\{(k_t \otimes k_t)(k_s \otimes k_s)\}] = \mathrm{E}\{k^2(X_t, X_s)\}$, due to Lemma 17(i). Using the stationarity of $\{X_t\}_{t=1}^n$ and Lemma 17(iii) we get

$$\mathrm{E}\left(\|S_n - S\|_{\mathrm{HS}}^2\right) = \frac{1}{n}\left\{\mathrm{E}\{k^2(X_0, X_0)\} - \|S\|_{\mathrm{HS}}^2\right\} + 2\sum_{h=1}^{n-1}\frac{n-h}{n^2}\left[\mathrm{E}\{k^2(X_h, X_0)\} - \|S\|_{\mathrm{HS}}^2\right],$$

yielding the first result by an application of Lemma 17 (ii).
For the second equation we see due to the independence of $\{X_t\}_{t=1}^n$ and $\{\varepsilon_t\}_{t=1}^n$ that

$$\|T_n^* y - Sf\|_{\mathcal{H}}^2 = \sigma^2 n^{-1}\mathrm{E}\{k(X_0, X_0)\} + \mathrm{E}\left(\|S_n f - Sf\|_{\mathcal{H}}^2\right).$$

The rest follows along the same lines as the first part of the proof.

(ii) An application of part (i) of this theorem, Corollary 19 and Lemma 20 yields this result.

(iii) Because the $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ are independent and identically distributed and $\{X_t\}_{t\in\mathbb{Z}}$ is stationary it holds

$$\begin{aligned}
&\mathrm{E}\left\{\left\|(S+\lambda)^{-1/2}(S_n f - T_n^* y)\right\|_{\mathcal{H}}^2\right\} \\
&= \mathrm{E}\left\{\left\|(S+\lambda)^{-1/2}T_n^*\varepsilon\right\|_{\mathcal{H}}^2\right\} = n^{-2}\sum_{t,s=1}^n \mathrm{E}\left\{\left\langle\varepsilon_t(S+\lambda)^{-1}k_t, \varepsilon_s k_s\right\rangle_{\mathcal{H}}\right\} \\
&= n^{-1}\sigma^2\mathrm{E}\left\{\left\langle(S+\lambda)^{-1}k_0, k_0\right\rangle_{\mathcal{H}}\right\} = n^{-1}\sigma^2\mathrm{E}\left\{\left\|(S+\lambda)^{-1/2}k_0\right\|_{\mathcal{H}}^2\right\}.
\end{aligned}$$

By the definition of d_λ we get

$$\mathrm{E}\{\|(S+\lambda)^{-1/2}k_0\|_{\mathcal{H}}^2\} = \mathrm{E}[\mathrm{tr}\{(S+\lambda)^{-1}k_0 \otimes k_0\}] = \mathrm{tr}\{(S+\lambda)^{-1}S\} = d_\lambda.$$

Using $n^{-1/2} \leq \gamma_n(q)$ proves the result.

(iv) Consider first

$$\begin{aligned}
\mathrm{E}\left\{\|(S+\lambda)^{-1/2}(S_n - S)\|_{\mathrm{HS}}^2\right\} &= n^{-2}\sum_{t,s=1}^n \mathrm{E}\left[\mathrm{tr}\{(S+\lambda)^{-1}(k_t \otimes k_t - S)(k_s \otimes k_s - S)\}\right] \\
&= n^{-1}\mathrm{E}\|(S+\lambda)^{-1/2}(k_0 \otimes k_0 - S)\|_{\mathrm{HS}}^2 \\
&\quad + \sum_{h=1}^{n-1}\mathrm{E}\left[\mathrm{tr}\{(S+\lambda)^{-1}(k_0 \otimes k_0 - S)(k_h \otimes k_h - S)\}\right].
\end{aligned}$$

Continuing with the expression inside the sums we expand

$$\begin{aligned}\phi_h &= \mathrm{E}\left[\mathrm{tr}\{(S+\lambda)^{-1}(k_0 \otimes k_0 - S)(k_h \otimes k_h - S)\}\right] \\ &= \mathrm{E}\left[\mathrm{tr}\{(S+\lambda)^{-1}k_0 \otimes k_0 k_h \otimes k_h\}\right] - \mathrm{tr}\{(S+\lambda)^{-1}S^2\} \\ &= \mathrm{E}\left\{k(X_0, X_h)\langle (S+\lambda)^{-1}k_0, k_h\rangle_{\mathcal{H}}\right\} - \mathrm{tr}\{(S+\lambda)^{-1}S^2\}.\end{aligned}$$

Using Lemma 17 (iv) we see that

$$\mathrm{tr}\{(S+\lambda)^{-1}S^2\} = \int\limits_{\mathbb{R}^{2d}} k(x,y)\langle (S+\lambda)^{-1}k(\cdot,x), k(\cdot,y)\rangle_{\mathcal{H}} \mathrm{dP}^X(x)\mathrm{dP}^X(y).$$

Hence we have

$$\phi_h = \int\limits_{\mathbb{R}^{2d}} k(x,y)\langle (S+\lambda)^{-1}k(\cdot,x), k(\cdot,y)\rangle_{\mathcal{H}} \left\{\mathrm{dP}^{X_0,X_h}(x,y) - \mathrm{dP}^{X_0}(x)\mathrm{dP}^{X_0}(y)\right\}.$$

This can be bound by the results of Lemma 21 and together with Lemma 20 there exists a constant $C(q) > 0$ such that with probability at least $1-\nu$

$$\|(S+\lambda)^{-1/2}(S_n - S)\|_{\mathcal{L}}^2 \leq \nu^{-1}C^2(q)\gamma_n^2(q)d_\lambda,$$

with $\gamma_n^2(q) = \begin{cases} n^{-1}, & q > 1 \\ n^{-1}\log(n), & q = 1 \\ n^{-q}, & q \in (0,1). \end{cases}$

This implies $\|(S+\lambda)^{-1/2}(S_n - S)(S+\lambda)^{-1/2}\|_{\mathcal{L}} \leq \nu^{-1/2}C(q)\lambda^{-1/2}\sqrt{d_\lambda}\gamma_n(q)$. Let $\lambda = \lambda_n$ be a sequence converging to zero such that $\lambda_n^{-1/2}\sqrt{d_{\lambda_n}}\gamma_n(q) \to 0$. Let n be large enough such that $\nu^{-1/2}C(q)\lambda_n^{-1/2}\sqrt{d_{\lambda_n}}\gamma_n(q) < 1$. Using Lemma A.5 in Blanchard and Krämer (2010b) we obtain

$$\|(S+\lambda)^{1/2}(S_n+\lambda)^{-1/2}\| \leq [1 - \nu^{-1/2}C(q)\lambda_n^{-1/2}\sqrt{d_{\lambda_n}}\gamma_n(q)]^{-1/2} \leq \sqrt{2}.$$

The latter inequality can be fulfilled for n large enough such that $\nu^{-1/2}C(q)\lambda_n^{-1/2}\sqrt{d_{\lambda_n}}\gamma_n(q) \leq 1/2$. ■

B.6 Proof of Proposition 7

Recall that $Su = \mathrm{E}\{u(X_0)k(\cdot, X_0)\}$ for $u \in \mathcal{H}$. Define the independent random variables $Y_1, \ldots, Y_\mu$ that are all distributed as X_0.

First consider the following observation for $\mu \in \mathbb{N}$:

$$\begin{aligned}S^\mu u = S(S^{\mu-1}u) &= \mathrm{E}_{Y_1}\{(S^{\mu-1}u)(Y_1)k(\cdot, Y_1)\} = \mathrm{E}_{Y_2}\mathrm{E}_{Y_1}\{(S^{\mu-2}u)(Y_2)k(Y_1,Y_2)k(\cdot,Y_1)\} \\ &= \mathrm{E}_{Y_\mu}\cdots\mathrm{E}_{Y_1}\left\{\prod_{\nu=1}^{\mu-1} k(Y_\nu, Y_{\nu+1})u(Y_\mu)k(\cdot,Y_1)\right\}. \qquad (30)\end{aligned}$$

We take $u = \sum_{i=1}^\infty c_i k(\cdot, z_i)$ for $\{z_i\}_{i\in\mathbb{N}}, \{c_i\}_{i\in\mathbb{N}} \subset \mathbb{R}$ such that $\|u\|_{\mathcal{H}}^2 = \sum_{i,j=1}^\infty c_i c_j k(z_i, z_j) \leq R^2$. The fact that a function $u \in \mathcal{H}$ can be represented as a linear combination of kernel functions is due to the Moore-Aronszajn Theorem, see Berlinet and Thomas-Agnan (2004).

Define the matrix $\Gamma = [\Gamma_{i,j}]_{i,j=1}^{\mu+2} \in \mathbb{R}^{(\mu+2)\times(\mu+2)}$ via

$$\Gamma_{i,j} = \begin{cases} \sigma_x^{-2} + 2l & , \quad i = j = 2, \ldots, \mu+1 \\ l & , \quad i = j = 1, \mu+2 \\ -l & , \quad |i-j| = 1 \\ 0 & , \quad else \end{cases}.$$

Then we have via the integration of Gaussian functions and (30)

$$\begin{aligned} f(x) &= \frac{1}{(2\pi\sigma_x^2)^{\mu/2}} \sum_{i=1}^{\infty} c_i \int_{\mathbb{R}^\mu} \exp\left\{-1/2 (x, x_1, \ldots, x_\mu, z_i)\Gamma(x, x_1, \ldots, x_\mu, z_i)^{\mathrm{T}}\right\} \mathrm{d}(x_1, \ldots, x_\mu) \\ &= \frac{1}{\sigma_x^\mu \det(\Gamma_{2:\mu+1})^{1/2}} \sum_{i=1}^{\infty} c_i \exp\left[-1/2 \frac{\det(\Lambda_{1:\mu+1})(x^2 + z_i^2) - 2l^{\mu+1} x z_i}{\det(\Gamma_{2:\mu+1})}\right]. \end{aligned}$$

Here we used the symmetry property $\det(\Gamma_{2:\mu+2}) = \det(\Gamma_{1:\mu+1})$ as the first and last rows and columns of Γ are identical. This concludes the proof. ■

B.7 Proof of Proposition 8

In Shi et al. (2008) it was shown that the eigenvalues of S have the form $\mu_i = ab^{i-1}$, $i = 1, 2, \ldots$ with

$$a = \sqrt{2}(1 + \beta + \sqrt{1+\beta})^{-1/2}, \quad b = (1 + \beta + \sqrt{1+2\beta})^{-1}\beta.$$

and $\beta = 4l\sigma_x^2$. It is clear that $0 < b < 1$ and hence $0 < \mu_i \leq a$. We have $d_\lambda = \sum_{i=0}^{\infty}\{1 + a^{-1}b^{-i}\lambda\}^{-1}$. Denote $f(x) = \{1 + a^{-1}b^{-x}\lambda\}^{-1}$. We want to apply the integral test to the sum. We have $\int_0^\infty f(x)\mathrm{d}x = \log^{-1}(b^{-1})\log(1 + a\lambda^{-1})$. This yields the bounds

$$\frac{\log(1 + a/\lambda)}{\log(b^{-1})} \leq d_\lambda \leq \frac{1}{1 + \lambda/a} + \frac{\log(1 + a/\lambda)}{\log(b^{-1})}.$$

On $\lambda \in (0, 1]$ we get $d_\lambda \leq D\log(1 + a/\lambda)$ for a constant $D > 0$. This can be seen as follows: The function $g_1(\lambda) = (1 + \lambda/a)^{-1}$ is bounded from above by $C_1 = 1$ and the function $g_2(\lambda) = (b^{-1})\log(1 + a/\lambda)$ is lower bounded by $c_2 = \log(1 + a)$ and has no upper bound.

Hence on the set $I = \{\lambda \in (0, 1] : g_2(\lambda) \geq C_1\}$ we can choose $C = 2$. On the set I^c we have on the other hand $Cg_2(x) \geq c_2 C \geq g_1(x) + g_2(x)$, hence we need $C = 2c_2^{-1}C_1 = 2\log^{-1}(1 + a)$. The choice $D = 2\log^{-1}(b^{-1})\max\{1, \log^{-1}(1 + a)\}$ is sufficient and we have $d_\lambda \leq D\log(1 + a/\lambda)$, $\lambda \in (0, 1]$. ■

References

A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic, Boston, 2004.

N. Bissantz, T. Hohage, A. Munk, and F. Ruymgaart. Convergence rates of general regularization methods for statistical inverse problems. *SIAM J. Numer. Anal.*, 45:2610–2636, 2007.

G. Blanchard and N. Krämer. Kernel partial least squares is universally consistent. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings*, volume 9, pages 57–64. JMLR, 2010a.

G. Blanchard and N. Krämer. Optimal learning rates for kernel conjugate gradient regression. *Adv. Neural Inf. Process. Syst.*, 23:226–234, 2010b.

P.J. Brockwell and R.A. Davis. *Time Series: Theory and Methods*. Springer, New York, 2 edition, 1991.

B. Brooks and M. Karplus. Harmonic dynamics of proteins: Normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proc. Natl. Acad. Sci.*, 80:6571–6575, 1983.

A. Caponnetto and E. de Vito. Optimal rates for regularized least-squares algorithm. *Found. Comp. Math.*, 7:331–368, 2007.

F. Cucker and S. Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc.*, 39:1–49, 2002.

E. de Vito, L. Rosasco, A. Caponnetto, U. de Giovanni, and F. Odone. Learning from examples as an inverse problem. *J. Mach. Learn. Res.*, 6:883–904, 2005.

E. De Vito, A. Caponnetto, and L. Rosasco. Discretization error analysis for Tikhonov regularization in learning theory. *Anal. Appl.*, 4:81–99, 2006.

L.H. Dicker, D.P. Foster, and D. Hsu. Kernel ridge vs. principal component regression: Minimax bounds and the qualification of regularization operators. *Electron. J. Stat.*, 11: 1022–1047, 2017.

I. Frank and J.H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

L. Giraitis, P. Kokoszka, R. Leipus, and G. Teyssière. Rescaled variance and related tests for long memory in volatility and levels. *J. Econometr.*, 112:265–294, 2003.

L. Giraitis, L.K. Hira, and D. Surgailis. *Large Sample Inference for Long Memory Processes*. Imperial College Press, London, 1 edition, 2012.

M. Hanke. *Conjugate Gradient Type Methods for Ill-posed Problems*. Wiley, New York, 1 edition, 1995.

I.S. Helland. On the structure of partial least squares regression. *Commun. Stat. Simul. Comput.*, 17(2):581–607, 1988.

K. Henzler-Wildman and D. Kern. Dynamic personalities of proteins. *Nature*, 450:964–972, 2007.

J.S. Hub and B.L. de Groot. Detection of functional modes in protein dynamics. *PLoS Comput. Biol.*, 5:1029–1044, 2009.

N. Krämer and M. L. Braun. Kernelizing PLS, degrees of freedom, and efficient model selection. In *Proceedings of the 24th International Conference on Machine Learning*, pages 441–448. ACM, 2007.

SB. Lin and DX. Zhou. Optimal learning rates for kernel partial least squares. *J. Fourier Anal. Appl.*, 2017.

F. Lindgren, P. Geladi, and S. Wold. The kernel algorithm for PLS. *J. Chemometr.*, 7: 45–59, 1993.

A.S. Nemirovskii. The regularizing properties of the adjoint gradient method in ill-posed problems. *Comput. Math. Math. Phys.*, 26:7–16, 1986.

G. Raskutti, M.J. Wainwright, and B. Yu. Early stopping and non-parametric regression: An optimal data-dependent stopping rule. *J. Mach. Learn. Res.*, 15:335–366, 2014.

R. Rosipal. Kernel partial least squares for nonlinear regression and discrimination. *Neural Netw. World*, 13:291–300, 2003.

R. Rosipal and L.J. Trejo. Kernel partial least squares regression in reproducing kernel Hilbert space. *J. Mach. Learn. Res.*, 2:97–123, 2001.

R. Rosipal, M. Girolami, and L.J. Trejo. Kernel PCA for feature extraction of event-related potentials for human signal detection performance. In *Proceedings of ANNIMAB-1 Conference*, pages 321–326. Springer, 2000.

R. Rosipall and N. Krämer. Overview and recent advances in partial least squares. In *Lecture Notes in Computer Science*, volume 3940, pages 34–51. Springer, 2006.

G. Samorodnitsky. *Long Range Dependence*. now Publisher, Hanover, 1 edition, 2007.

C. Saunders, A. Gammerman, and V. Vovk. Ridge regression learning algorithm in dual variables. In *Proceedings of the 15th International Conference on Machine Learning*, pages 515–521. Morgan Kaufmann Publishers, 1998.

B. Schölkopf and A.J. Smola. *Learning with Kernels*. MIT Press, Cambridge, 1 edition, 2001.

B. Schölkopf, A. J. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, 10:1299–1319, 1998.

B. Schölkopf, R. Herbrich, and A.J. Smola. A generalized representer theorem. In *International Conference on Computational Learning Theory*, pages 416–426. Springer, 2001.

T. Shi, M. Belkin, and B. Yu. Data spectroscopy: Learning mixture models using eigenspaces of convolution operators. In *Proceedings of the 25th International Conference on Machine Learning*, pages 936–943. Omnipress, 2008.

M. Singer, T. Krivobokova, B.L. de Groot, and A. Munk. Partial least squares for dependent data. *Biometrika*, 103:351–362, 2016.

I. Steinwart, D. Hush, and C. Scovel. An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels. Technical report, IEEE Trans. Inform. Theory, 2005.

G. Wahba. Support vector machines, reproducing kernel Hilbert spaces and randomized GACV. In *Advances in Kernel Methods - Support Vector Learning*, pages 69–88. MIT Press, 1999.

S. Wold, A. Ruhe, H. Wold, and W.J. Dunn. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM J. Sci. Comput.*, 5:735–743, 1984.

T. Zhang. Effective dimension and generalization of kernel learning. In *Advances in Neural Information Processing Systems 15*, pages 471–478. MIT Press, 2003.