Kernel Estimation for Nonlinear Dynamics [*†‡]

Marie-Christine Düker
FAU Erlangen-Nürnberg

Adam Waterbury
Denison University

February 25, 2025

Abstract

Many scientific problems involve data exhibiting both temporal and cross-sectional dependencies. While linear dependencies have been extensively studied, the theoretical analysis of regression estimators under nonlinear dependencies remains scarce. This work studies a kernel-based estimation procedure for nonlinear dynamics within the reproducing kernel Hilbert space framework, focusing on nonlinear vector autoregressive models. We derive nonasymptotic probabilistic bounds on the deviation between a regularized kernel estimator and the nonlinear regression function. A key technical contribution is a concentration bound for quadratic forms of stochastic matrices in the presence of dependent data, which is of independent interest. Additionally, we characterize conditions on multivariate kernels that guarantee optimal convergence rates.

1 Introduction

Many data exhibit nonlinear behaviors that linear models cannot capture. In fields such as finance (Benrhmach, Namir, Namir, and Bouyaghroumni, 2020), economics (Nyberg, 2018), biology and neuroscience (Kato, Taniguchi, and Honda, 2006; Yu, Liu, Heck, Berger, and Song, 2021), relationships among variables often involve thresholds, feedback loops, or sudden changes that linear models fail to address. These data often display temporal and cross-sectional dependencies, making them well-suited for modeling with nonlinear dynamical systems. However, estimation techniques for time series arising from such systems remain far less developed than those for linear models. In contrast, the statistical learning and machine learning literature provide methodology that allow for the estimation of such nonlinear relationships.

We employ the mathematical framework of reproducing kernel Hilbert spaces (RKHSs) to introduce an estimator for nonlinear dynamical systems with temporal and cross-sectional correlations. Kernel-based techniques allow to make inferences in a high-dimensional feature space mapped implicitly by a kernel function. This technique is integral to many statistical learning algorithms, including the support vector machine (SVM) and the kernel principle component analysis (KPCA) algorithms (see Hastie, Tibshirani, Friedman, and Friedman (2009)), but has not been explored in the context of nonlinear time series. Our estimation procedure and theoretical results are phrased for nonlinear vector autoregression models but also cover stochastic regression models. We treat the stochastic regression model as a special case of nonlinear vector autoregression; see Section 3.2.3 for further discussion.

[*] AMS subject classification. Primary: 62J07, 62M10. Secondary: 62G08.

[†] Keywords: kernel regression, reproducing kernel Hilbert space, method of regularization, nonlinear time series, concentration inequalities, Markov chains.

[‡] MD was partially supported by the FAU Emerging Talents Initiative.

Nonlinear vector autoregression: Suppose $\{X_t\}_{t\in\mathbb{Z}}$ follows a causal nonlinear VAR model of order p. To be more precise, we consider an $\mathbb{R}^d$-valued sequence $\{X_t\}_{t\in\mathbb{Z}}$, where $X_t = (X_{1,t},\dots,X_{d,t})'$ is defined according to

$$X_t = g(X_{t-1},\dots,X_{t-p}) + \varepsilon_t = (g_1(X_{t-1},\dots,X_{t-p}),\dots,g_d(X_{t-1},\dots,X_{t-p}))' + \varepsilon_t, \tag{1.1}$$

for a function $g:\mathbb{R}^{dp}\to\mathbb{R}^d$ and a suitable suitable noise sequence $\{\varepsilon_t\}_{t\in\mathbb{Z}}$. The model (1.1) is quite general and allows for nonlinear dependencies across multiple lags (p) and across space (d). The goal is to estimate the function g, which allows for downstream tasks such as forecasting, interpretation of dynamics, and clustering. Note that the model (1.1) also covers additive dynamics of the form

$$X_t = \sum\nolimits_{j=1}^{p} h_j(X_{t-j}) + \varepsilon_t, \tag{1.2}$$

where $h_1,\dots,h_d:\mathbb{R}^d\to\mathbb{R}^d$ are nonlinear regression functions. To estimate the function g in (1.1), we employ a regularized least squares estimator also known as Kernel Ridge Regression (KRR). KRR is a popular technique in supervised learning and has been used to avoid overfitting in regression problems. However, to the best of our knowledge, there is no literature on the use of KRR estimators for temporal models. In particular, we are the first to provide statistical guarantees under temporal and cross-sectional dependencies for such an estimator. On the way, we formalize the use of multivariate RKHSs and corresponding Mercer representations when kernels are evaluated at multivariate data. In contrast to regression problems, the use of temporal models such as nonlinear vector autoregression makes it desirable to allow for unbounded errors, which requires kernels defined on unbounded spaces.

The theoretical difficulties in this work are the development of concentration results for random quadratic forms of a multivariate empirical kernel matrix. We establish nonasymptotic high probability bounds for KRR estimators by proving concentration results for quadratic forms of dependent data. To be more precise, our results establish a Hoeffding-type inequality for quadratic forms involving empirical kernel matrices, and are of independent interest. We provide conditions under which the estimator achieves the optimal convergence rate in the sense that the results align with the rate for quadratic forms of deterministic matrices; see Rudelson and Vershynin (2013).

Literature review: The literature on nonlinear regression models and corresponding statistical guarantees is sparse. The existing literature has considered regression settings (Liu and Li, 2023) and additive models incorporating sparsity assumptions; see Yuan and Zhou (2016), Zhou and Raskutti (2018), who consider models with additive dynamics similar to (1.2). To the best of our knowledge there is no work providing statistical guarantees for regularized estimators of nonlinear vector autoregression models.

The works Smale and Zhou (2005) and Liu and Li (2023) are the closest to our paper and served as inspiration for our estimation procedure. Smale and Zhou (2005) introduce a KRR estimator for a nonlinear regression model and derive conditional concentration results, effectively looking at nonrandom predictors. Liu and Li (2023) provide estimators for higher order derivatives of the regression function based on a KRR estimator. Our main result can recover their results but also allows for much more general settings. In particular, our results allow for nonlinear temporal and cross-sectional dependencies. Furthermore, our class of feasible kernels includes kernel functions on noncompact spaces and allows for multivariate inputs.

Regularization techniques like ridge- and lasso-type penalties have been studied extensively for linear regression problems. Early works study linear regression models with deterministic predictors; see Loh and Wainwright (2012) who provide consistency results for lasso estimators. Later generalizations cover stochastic regression and linear vector autoregression models; see Basu and Michailidis (2015) for lasso and Ballarin (2024) for ridge estimation.

Our theoretical results crucially rely on concentration results for functionals of Markov chains. Recent literature studies nonlinear functionals of Markov chains under a range of boundedness and smoothness conditions; see Adamczak and Bednorz (2015), Adamczak and Wolff (2015), Paulin (2015), Chen and Wu (2017), Alquier, Doukhan, and Fan (2019), Fan, Jiang, and Sun (2021).

Given the kernel-based estimation approach, our work naturally connects to the literature on U-statistics, particularly concentration inequalities for U-statistics. The problem is well studied for kernels evaluated at i.i.d. samples; see Arcones and Gine (1993), Arcones (1995). A survey can be found in Pitcan (2017), and recent works attempt to generalize those results in several directions. For instance, Chakrabortty and Kuchibhotla (2018) consider unbounded kernels and Duchemin, De Castro, and Lacour (2023) study U-statistics for Markov chains and characterize assumptions on the chain that ensure optimal convergence rates. Borisov and Volodko (2015) and Han (2018) established exponential inequalities for U-statistics of order two and larger in time series under mixing conditions. These findings were later refined by Shen, Han, and Witten (2020), which introduced a Hanson-Wright-type inequality for both V- and U-statistics under conditions on the time-dependent process that are more practical to verify.

Outline of the paper: The rest of the paper is organized as follows. In Section 2, we introduce a kernel-based regularized estimator and state our main assumptions. Section 3 formalizes our main results and includes a discussion of these results and associated assumptions. The section is supplemented with several examples that illustrate that our assumptions are satisfied by a wide range of models. In Section 4, we state some more technical tools and auxiliary results. Finally Sections 5 and 6 provide the proofs of our results.

Notation: We write $\mathbb{N} \doteq \{1, 2, \dots\}$ and $\mathbb{N}_0 \doteq \mathbb{N} \cup \{0\}$. For measurable spaces $(\mathcal{X}_1, \mathcal{F}_1)$ and $(\mathcal{X}_2, \mathcal{F}_2)$, we let $\mathcal{B}(\mathcal{X}_1 : \mathcal{X}_2)$ denote the space of measurable functions from $\mathcal{X}_1$ to $\mathcal{X}_2$. Set $\mathcal{X} \subseteq \mathbb{R}^d$ and for a kernel $K : \mathcal{X}^p \times \mathcal{X}^p \to \mathbb{R}$, $x, y \in \mathcal{X}^p$, we use the notation $K(x, y)$ and $K(\mathrm{vec}(x), \mathrm{vec}(y))$ interchangeably, where $\mathrm{vec}(\cdot)$ stacks each column of a matrix. For a separable Hilbert space $\mathbb{H} \subseteq \mathcal{B}(\mathcal{X} : \mathbb{R})$ with inner product $\langle \cdot, \cdot \rangle_{\mathbb{H}}$ and orthonormal basis $\{\phi_k,\ k \in \mathbb{N}\}$, given an $\mathbb{H}$-valued random variable X, we define $\mathbb{E}_{\mathbb{H}}(X) \in \mathbb{H}$ as $\mathbb{E}_{\mathbb{H}}(X) \doteq \sum_{k=1}^{\infty} \mathbb{E}\left(\langle X, \phi_k \rangle_{\mathbb{H}}\right) \phi_k$, whenever the series is convergent in $\mathbb{H}$. If $x \in \mathbb{R}^m$ for some $m \in \mathbb{N}$, we let $\|x\| = (\sum_{i=1}^m x_i^2)^{1/2}$. For $M \in \mathbb{R}^{m \times n}$, we let $\|M\| = (\sum_{i=1}^m \sum_{j=1}^n M_{ij}^2)^{1/2}$ denote the Frobenius norm of M. Other norms are denoted using subscripts. In particular, for a function $f : \mathbb{R}^m \to \mathbb{R}^n$, we set $\|f\|_\infty = \sup_{x \in \mathbb{R}^m} \|f(x)\|_\infty$, where for a vector $y \in \mathbb{R}^n$, we let $\|y\|_\infty = \sup_{i=1,\dots,n} |y_i|$. For a Polish space $\mathcal{S}$, we write $\mathcal{P}(\mathcal{S})$ to denote the space of probability measures on $\mathcal{S}$, and for two finite measures μ, ν on $\mathcal{S}$, we write $(\mu - \nu)$ to denote the signed measure defined as $(\mu - \nu)(A) = \mu(A) - \nu(A)$. Similarly, we write $|\mu - \nu|$ to denote the finite measure defined as $|\mu - \nu|(A) = |\mu(A) - \nu(A)|$. The total variation distance between μ and ν is given by $\|\mu - \nu\|_{\mathrm{TV}} = \sup_{A \in \mathcal{B}(\mathcal{S})} |\mu(A) - \nu(A)|$. For $\pi \in \mathcal{P}(\mathbb{R}^d)$, $\mathbb{P}_\pi(X_1 \in A_1) = \pi(A_1)$ is the law under which $X_1 \sim \pi$. For $x \in \mathcal{X}$, we write $\mathbb{P}_x$ to denote $\mathbb{P}_{\delta_x}$, where δ_x is the Dirac measure at x. For $n \in \mathbb{N}$ and $k_1, \dots, k_j \in \mathbb{N}_0$ such that $\sum_{i=1}^j k_i = n$ and $\boldsymbol{k} = (k_1, \dots, k_j)$, we write $\binom{n}{\boldsymbol{k}} = \frac{n}{\boldsymbol{k}!}$, where $\boldsymbol{k}! = k_1! \cdots k_j!$. For $k \in \mathbb{N}_0$, let $\boldsymbol{k} + k = (k_1 + k, \dots, k_j + k)$. We use Γ to denote the gamma function and B to denote the beta function.

2 Preliminaries

In this section, we establish notation and present preliminary results to introduce a regularized estimator within the RKHS framework. Recalling the model (1.1), it is convenient to introduce the corresponding $\mathcal{X}^p$-valued nonlinear VAR model of order one and treat it as a Markov chain. This $\mathcal{X}^p$-valued process $\{Y_t\}$ is defined by

$$Y_t \doteq \left(X_t', \dots, X_{t-p+1}'\right)', \quad t = p, \dots, T, \tag{2.1}$$

and, for $t = p, \dots, T$, we define

$$\begin{pmatrix} X_t \\ \vdots \\ X_{t-p+1} \end{pmatrix} = \begin{pmatrix} g(X_{t-1}, \dots, X_{t-p}) \\ X_{t-1} \\ \vdots \\ X_{t-p+1} \end{pmatrix} + \begin{pmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \text{or} \quad Y_t = G_Y(Y_{t-1}) + \xi_t.$$

For notational convenience, we sometimes write

$$Y_t = (Y'_{t,1}, \dots, Y'_{t,p})', \quad Y_{t,j} = X_{t+1-j},\ j = 1, \dots, p,$$

so that, in particular, $Y_{t,1} = X_t$. In Lemma 4.10 it is shown that, under some mild assumptions, $\{Y_t\}$ is a $\mathcal{X}^p$-valued geometrically ergodic Markov chain and therefore has a unique stationary distribution $\pi \in \mathcal{P}(\mathcal{X}^p)$. We let

$$L^2(\mathcal{X}^p, \pi) = \left\{ f \in \mathcal{B}(\mathcal{X}^p : \mathbb{R}) : \int_{\mathcal{X}^p} (f(y))^2 \pi(dy) < \infty \right\}$$

denote the associated space of square-integrable functions. An RKHS is defined through a kernel, namely a symmetric and positive definite function $K : \mathcal{X}^p \times \mathcal{X}^p \to \mathbb{R}$. More precisely, K is a kernel if

$$K(x, y) = K(y, x), \quad x, y \in \mathcal{X}^p,$$

and $\sum_{i,j=1}^d c_i c_j K(z_i, z_j) \geqslant 0$ for all $z_1, \dots, z_d \in \mathcal{X}^p$ and $c_1, \dots, c_d \in \mathbb{R}$, with equality if and only if $c_i = 0$ for all $i = 1, \dots, d$. For kernels $K_1, \dots, K_d : \mathcal{X}^p \times \mathcal{X}^p \to \mathbb{R}$, we let $\mathbb{H}_i \subseteq L^2(\mathcal{X}^p, \pi)$ denote the RKHS associated with K_i. That is $(\mathbb{H}_i, \|\cdot\|_{\mathbb{H}_i})$ is the Hilbert space defined as,

$$\mathbb{H}_i = \{ f \in L^2(\mathcal{X}^p, \pi) : \text{ for each } x \in \mathcal{X}^p,\ f(x) = \langle f, K_i(x, \cdot) \rangle_{\mathbb{H}_i} \},$$

where $\|\cdot\|_{\mathbb{H}_i}$ is the RKHS inner product associated with K_i (see Section 4.1). For the multivariate kernel

$$\mathcal{K}(x, y) = \operatorname{diag}(K_1(x, y), \dots, K_d(x, y)), \quad x, y \in \mathcal{X}^p \times \mathcal{X}^p \tag{2.2}$$

we denote the associated RKHS by $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ and note that $\mathbb{H}$ can be identified with a subset of $L^{2,d}(\mathcal{X}^p, \pi) \doteq \times_{i=1}^d L^2(\mathcal{X}^p, \pi)$. Further discussion of RKHS inner products and multivariate RKHS can be found in Section 4.1.

2.1 Kernel ridge regression

In this section, we introduce a regularized kernel-based estimator for the regression function g in (1.1). To estimate this function, we employ a ridge-type objective function of the form

$$\begin{aligned} \widehat{g}_T &= \operatorname{argmin}_{g \in \mathbb{H}} \left\{ \frac{1}{T} \sum_{t=p+1}^{T} (X_t - g(X_{(t-p):(t-1)}))'(X_t - g(X_{(t-p):(t-1)})) + \lambda \|g\|_{\mathbb{H}}^2 \right\} \\ &= \operatorname{argmin}_{g \in \mathbb{H}} \left\{ \frac{1}{T} \sum_{t=p+1}^{T} (Y_{t,1} - g(Y_{t-1}))'(Y_{t,1} - g(Y_{t-1})) + \lambda \|g\|_{\mathbb{H}}^2 \right\}, \end{aligned} \tag{2.3}$$

where $X_{(t-p):(t-1)} \doteq (X_{t-1}, \dots, X_{t-p})$ and $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ is the RKHS induced by $\mathcal{K}$ (see Section 4.1). The second identity in (2.3) uses the notation introduced in (2.1).

Note that the model (1.1) can be written more compactly as

$$\mathcal{Y}_X \doteq \begin{pmatrix} X_{p+1}' \\ \vdots \\ X_T' \end{pmatrix} = \begin{pmatrix} g(X_p, \ldots, X_1)' \\ \vdots \\ g(X_{T-1}, \ldots, X_{T-p})' \end{pmatrix} + \begin{pmatrix} \varepsilon_{p+1}' \\ \vdots \\ \varepsilon_T' \end{pmatrix} \doteq G_X(g) + \mathcal{E}. \tag{2.4}$$

Furthermore, (2.4) can be written in vectorized form as

$$Y \doteq \operatorname{vec}(\mathcal{Y}_X) = \operatorname{vec}(G_X(g)) + \operatorname{vec}(\mathcal{E}) \doteq G_g(X) + \eta, \tag{2.5}$$

where

$$X \doteq (X_1, \ldots, X_T) \in \mathbb{R}^{d \times T} \quad \text{and} \quad \eta \doteq \operatorname{vec}(\mathcal{E}). \tag{2.6}$$

The goal is to rewrite the objective function (2.3) using a representer theorem in order to obtain an explicit representation for $\hat{g}_T$ akin to the standard ridge regression estimator in linear regression problems. For the kernels $K_1, \ldots, K_d$, there are, by the representer theorem (Schölkopf, Herbrich, and Smola, 2001),

$$\hat{\alpha}_i = (\hat{\alpha}_{i,p+1}, \ldots, \hat{\alpha}_{i,T})' \in \mathbb{R}^{(T-p)\times 1}, \;\; i = 1, \ldots, d,$$

such that, for $z \in \mathcal{X}^p$,

$$\hat{g}_{i,T}(z) = \sum\nolimits_{t=p+1}^{T} \hat{\alpha}_{i,t} K_i(z, X_{(t-p):(t-1)}) = K_{X,i}(z)'\hat{\alpha}_i, \tag{2.7}$$

where, for $z \in \mathcal{X}$, $K_{X,i}(z) \in \mathbb{R}^{(T-p)\times 1}$ is defined by

$$K_{X,i}(z) = \begin{pmatrix} K_i(z, X_{1:p}) \\ \vdots \\ K_i(z, X_{(T-p):(T-1)}) \end{pmatrix}.$$

In particular, using (2.7), the minimizer in (2.3) can be written explicitly as

$$\hat{g}_T(z) = (\hat{g}_{1,T}(z), \ldots, \hat{g}_{d,T}(z))' = \left(K_{X,1}(z)'\hat{\alpha}_1, \ldots, K_{X,d}(z)'\hat{\alpha}_d\right)' = \mathcal{K}_X(z)\hat{\alpha}, \tag{2.8}$$

where, for each $z \in \mathcal{X}^p$, $\mathcal{K}_X(z) \in \mathbb{R}^{d \times d(T-p)}$ is given by

$$\mathcal{K}_X(z) \doteq \operatorname{diag}(K_{X,1}(z)', \ldots, K_{X,d}(z)'), \tag{2.9}$$

and $\hat{\alpha} \doteq \operatorname{vec}(\hat{\alpha}_1, \ldots, \hat{\alpha}_d) \in \mathbb{R}^{d(T-p)\times 1}$. Recall (2.4) and note that, for $i = p+1, \ldots, T$, using (2.8), the estimator for the $(i-p)$-th row of G_X in (2.4) can be written as

$$\begin{aligned} \hat{g}_T(X_{i-1}, \ldots, X_{i-p}) = \Bigg(&\sum\nolimits_{t=p+1}^{T} \hat{\alpha}_{1,t} K_1(X_{(i-p):(i-1)}, X_{(t-p):(t-1)}), \ldots, \\ &\sum\nolimits_{t=p+1}^{T} \hat{\alpha}_{d,t} K_d(X_{(i-p):(i-1)}, X_{(t-p):(t-1)}) \Bigg)'. \end{aligned} \tag{2.10}$$

To rewrite the data matrix $G_X(g)$ in (2.4) as in (2.10), consider the empirical kernel matrices $K_1(X,X), \ldots, K_d(X,X) \in \mathbb{R}^{(T-p)\times(T-p)}$, with $K_i(X,X) \doteq [(K_i(X,X))_{s,t}]_{s,t=p+1,\ldots,T}$ defined as

$$(K_i(X,X))_{s,t} = K_i(X_{(s-p):(s-1)}, X_{(t-p):(t-1)}), \quad s,t = p+1, \ldots, T.$$

We also define an empirical kernel matrix

$$\mathcal{K}(X, X) = \operatorname{diag}(K_1(X, X), \ldots, K_d(X, X)) \in \mathbb{R}^{d(T-p)\times d(T-p)}. \tag{2.11}$$

Then, vectorizing the estimated counterpart of $G_X(g)$ in (2.4), we have

$$\begin{aligned}
\operatorname{vec}\begin{pmatrix} \widehat{g}_T(X_p, \ldots, X_1)' \\ \vdots \\ \widehat{g}_T(X_{T-1}, \ldots, X_{T-p})' \end{pmatrix} &= \operatorname{vec}\left(K_1(X, X)\widehat{\alpha}_1, \ldots, K_d(X, X)\widehat{\alpha}_d\right) \\
&= \operatorname{diag}(K_1(X, X), \ldots, K_d(X, X)) \operatorname{vec}\big((\widehat{\alpha}_1, \ldots, \widehat{\alpha}_d)\big) \\
&= \mathcal{K}(X, X)\widehat{\alpha}.
\end{aligned} \tag{2.12}$$

Using (2.10) and (2.12), we can reduce (2.3) to an equivalent convex objective function in $\mathbb{R}^{dp\times(T-p)}$, of the form

$$\arg\min_{\alpha\in\mathbb{R}^{d(T-p)}} \left\{ \frac{1}{T}\|Y - \mathcal{K}(X, X)\alpha\|_2^2 + \lambda\alpha'\mathcal{K}(X, X)\alpha \right\}. \tag{2.13}$$

Finally, the optimization problem in (2.13) is solved by the estimator

$$\widehat{\alpha} = (\mathcal{K}(X, X)\mathcal{K}(X, X) + \lambda T\mathcal{K}(X, X))^{-1}\mathcal{K}(X, X) \operatorname{vec}(X)$$

so it follows from (2.8) that the minimizer in (2.3) is given by, for $z \in \mathcal{X}^p$,

$$\begin{aligned}
\widehat{g}_T(z) = \mathcal{K}_X(z)\widehat{\alpha} &= \mathcal{K}_X(z) \left(\mathcal{K}(X, X)\mathcal{K}(X, X) + \lambda T\mathcal{K}(X, X)\right)^{-1} \mathcal{K}(X, X)Y \\
&= \mathcal{K}_X(z)(\mathcal{K}(X, X) + \lambda T I_{d(T-p)})^{-1}Y.
\end{aligned} \tag{2.14}$$

On the population level, an objective function analogous to (2.3) can be written as

$$g_\lambda = \arg\min_{\widetilde{g}\in\mathbb{H}} \left\{\|\widetilde{g} - g\|_2^2 + \lambda\|\widetilde{g}\|_{\mathbb{H}}^2\right\}. \tag{2.15}$$

Recalling the definition of $L^{2,d}(\mathcal{X}^p, \pi)$ from Section 2, standard results (see, e.g., Smale and Zhou (2005), Liu and Li (2023)) show that an explicit representation of g_λ is given by

$$g_\lambda(z) = (L_{\mathcal{K}} + \lambda I)^{-1}L_{\mathcal{K}}g(z), \quad z \in \mathcal{X}^p, \tag{2.16}$$

where the integral operator $L_{\mathcal{K}}$ on $L^{2,d}(\mathcal{X}^p, \pi)$ is defined as

$$L_{\mathcal{K}}(f)(z) = \int_{\mathcal{X}^p} \mathcal{K}(z, y)f(y)\pi(dy), \quad z \in \mathcal{X}^p,\ f = (f_1, \ldots, f_d)' \in L^{2,d}(\mathcal{X}^p, \pi), \tag{2.17}$$

and where I is the identity operator in $L^{2,d}(\mathcal{X}^p, \pi)$.

2.2 Assumptions

In this section we collect the assumptions needed to state our main results. The assumptions can be separated into three sets: the first concerns regression function g, the second pertains to the kernel used for estimation, and the third addresses the noise sequence.

Assumption G.1 (Bounded dynamics)**.** There is some $M_g \in (0, \infty)$ such that $\|g\|_\infty \leqslant M_g$.

Assumption G.1 helps ensure that the model in (1.1) is stationary and geometrically ergodic; see Section 4.2. The next set of assumptions is stated for the kernels $K_1, \dots, K_d$.

Assumption K.1 (Bounded kernel)**.** For each $j = 1, \dots, d$ there is some $\kappa_j \in (0, \infty)$ such that

$$\sup_{x \in \mathcal{X}^p} |K_j(x, x)| \leqslant \kappa_j^2.$$

Furthermore, we set $\kappa \doteq (\sum_{j=1}^d \kappa_j^2)^{1/2}$.

Assumption K.2 (Mercer expansion)**.** The kernels $K_1, \dots, K_d$ in (2.2) are separable. That is, for some $M \in \mathbb{N} \cup \{\infty\}$, there is a collection of eigenvalues $\lambda_{i,k}$ and eigenfunctions $\phi_{i,j,k}$, $i = 1, \dots, d$, $k = 1, \dots, M$, $j = 1, \dots, N(k)$, such that $\|\phi_{i,j,k}\|_\infty < \infty$, and

$$K_i(x, y) = \sum_{k=1}^M \lambda_{i,k} \sum_{j=1}^{N(k)} \phi_{i,j,k}(x)\phi_{i,j,k}(y), \tag{2.18}$$

where

$$N(k) = |\mathcal{N}(k)| = \binom{k + dp - 1}{dp - 1}, \tag{2.19}$$

with

$$\mathcal{N}(k) = \left\{(n_{1,1}, \dots, n_{1,p}, \dots, n_{d,1}, \dots, n_{d,p}) \in \mathbb{N}_0^{dp} : \sum_{i=1}^d \sum_{r=1}^p n_{i,r} = k\right\}. \tag{2.20}$$

Remark 2.1. Assumption K.2 gives a general Mercer representation that allows the kernel to have multivariate input variables. When $d = p = 1$, (2.18) reduces to the standard univariate Mercer kernel expansion, namely $K_1(x, y) = \sum_{k=1}^M \lambda_k \phi_k(x)\phi_k(y)$.

Assumption K.3 (Eigenfunction growth)**.** There are $b_1 \in (0, \infty), b_2 \in \mathbb{R}$ such that for each $M \in \mathbb{N}$ and $i = 1, \dots, d$, with

$$\beta_{i,j,k} \doteq \lambda_{i,k}^{1/2} \|\phi_{i,j,k}\|_\infty, \tag{2.21}$$

we have

$$\sum_{k=1}^M \sum_{j=1}^{N(k)} \beta_{i,j,k}^2 \leqslant b_1 M^{b_2}. \tag{2.22}$$

Assumption K.4 (Kernel moment bound)**.** There is an absolutely summable sequence $\{\alpha_k\}$ such that for all $k \in \mathbb{N}$, $t \geqslant p + 1$, and $i = 1, \dots, d$,

$$\lambda_{i,k} \sum_{j=1}^{N(k)} \mathbb{E}((\phi_{i,j,k}(Y_{t-1}))^2) = \lambda_{i,k} \sum_{j=1}^{N(k)} \mathbb{E}((\phi_{i,j,k}(X_{(t-p):(t-1)}))^2) \leqslant \alpha_k. \tag{2.23}$$

Additionally, there are $\beta_1, \beta_2 \in (0, \infty)$ and an increasing sequence $\{M(T)\}_{T \in \mathbb{N}}$ such that for each $T \in \mathbb{N}$,

$$\sum_{k=M(T)}^\infty \alpha_k \leqslant \beta_1 T^{-\beta_2}.$$

Remark 2.2. In Assumption K.2 we only assume that each eigenfunction is bounded. If $M = \infty$, then there may not be a uniform bound over all eigenfunctions. This is addressed through Assumptions K.3 and K.4. Assumption K.3 is trivially satisfied when the kernel has a finite Mercer expansion.

Assumption N.1 (Sub-Gaussian noise)**.** For each $t \in \mathbb{N}$, $i = 1, \dots, d$, $\varepsilon_{i,t}$ is sub-Gaussian with variance proxy $\sigma_i^2 \in (0, \infty)$. Additionally, $\varepsilon_{i,t}$ admits a density ψ such that for each compact set $A \subseteq \operatorname{supp}(\varepsilon_{i,t})$, $\inf_{x \in A} \psi(x) > 0$.

We refer to Proposition 2.5.2 in Vershynin (2018) for different characterizations of sub-Gaussianity. We also state two slightly more restrictive assumptions for the discussion of our convergence rates below. Both are special cases of Assumption N.1.

Assumption N.2 (Gaussian noise)**.** For each $t \in \mathbb{N}$, we have that $\varepsilon_t \sim \mathcal{N}(0, \Sigma)$, where $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$.

Assumption N.3 (Bounded noise)**.** Assumption N.1 is satisfied, and there is some $L_\varepsilon > 0$ such that for each $t \in \mathbb{N}$, $\|\varepsilon_t\|_\infty \leqslant L_\varepsilon$.

Remark 2.3. Under Assumptions G.1 and N.3, we can, without loss of generality, assume that $\mathcal{X}$ is a bounded subset of $\mathbb{R}^d$.

3 Concentration results for kernel functions

In Section 3.1 we present concentration results for kernel regression estimators under temporal dependence. Then, in Section 3.2, we present several applications in which our assumptions are satisfied.

3.1 Main results

Our main result (Theorem 3.1) provides a high probability bound that controls the deviation of the ridge estimator $\widehat{g}_T$ in (2.14) from the true nonlinear regression function g. Recall the function g_λ introduced in (2.16).

Theorem 3.1. *Suppose Assumptions G.1, K.1–K.4, and N.1. There are constants $c_1, c_2 \in (0, \infty)$ and $L, \mathrm{c}_0 \in (0, \infty)$ such that if*

$$\delta \geqslant \frac{1}{\lambda}\sqrt{\frac{\log(T)}{T}} L^{\frac{b_2}{2}} \gamma\, \mathrm{C}_0(g) + \|g_\lambda - g\|_\infty, \quad \mathrm{C}_0(g) = \mathrm{c}_0 \max\{\kappa^2 \|g\|_\infty, 1\} \tag{3.1}$$

and

$$\gamma \geqslant \sqrt{4\sigma^2 \log(dT)},$$

then, with probability at least $1 - c_1 T^{-c_2}$,

$$\|\widehat{g}_T - g\|_\infty \leqslant \delta.$$

If Assumption K.2 is satisfied with $M = \infty$, then we take $L \doteq M(T)$, where $M(T)$ is as in Assumption K.4. If Assumption K.2 is satisfied for some $M < \infty$, then we take $L \doteq M$.

Note that Theorem 3.1 is expressed in terms of $\mathrm{C}_0(g)$ and a constant c_0. Here $\mathrm{C}_0(g)$ depends on g, while c_0 subsumes all constants including d and p and may change throughout the proofs.

In view of Theorem 3.1, it only remains to control the quantity $\|g_\lambda - g\|_\infty$ in (3.1). While this deterministic term is, for general functions g, difficult to control, an estimate is provided in Theorem 4.11 that, under the somewhat restrictive Assumption N.3, bounds this term.

Theorems 3.2 and 3.3 below provide concentration inequalities for certain quadratic forms arising in the proof of Theorem 3.1 and may be of independent interest. To state these results, we introduce functions $c_{1,T}$ and $c_{2,T}$ given by

$$c_{1,T}(\delta, \gamma, M) = \frac{T}{\gamma^2}\left(\delta(db_1 M^{b_2})^{-\frac{1}{2}} - \frac{\sigma}{\sqrt{T}}\right)^2, \quad c_{2,T}(\gamma) = \left(\gamma - \sqrt{2\sigma^2 \log(dT)}\right)^2 / (2\sigma^2). \tag{3.2}$$

Note that Theorems 3.2 and 3.3 address the cases when Assumption K.2 is satisfied with finite ($M < \infty$) and infinite ($M = \infty$) Mercer representations, respectively. Kernels with an infinite Mercer representation require the additional Assumption K.4 on the variance of the eigenfunctions evaluated at the data.

Theorem 3.2. *Suppose Assumptions G.1, K.1–K.3, and N.1. Then, for any $\delta, \gamma \in (0, \infty)$, there is a $c \in (0, \infty)$, such that for all T,*

$$\mathbb{P}\left(\eta' \mathcal{K}(X, X)\eta > \delta^2\right) \leqslant d \exp\left(-c c_{1,T}(\delta, \gamma, M)\right) + d \exp\left(-c_{2,T}(\gamma)\right).$$

There are constants $c_1, c_2 \in (0, \infty)$ and $M, \mathrm{c}_0 \in (0, \infty)$ such that if

$$\delta \geqslant \mathrm{c}_0 \sqrt{\frac{\log(T)}{T}} M^{\frac{b_2}{2}} \gamma, \tag{3.3}$$

and

$$\gamma \geqslant \sqrt{4\sigma^2 \log(dT)}, \tag{3.4}$$

then, with probability at least $1 - c_1 T^{-c_2}$,

$$\eta' \mathcal{K}(X, X)\eta \leqslant \delta^2.$$

While Theorem 3.2 provides a concentration inequality for a quadratic form of an empirical kernel matrix whose associated kernel has a finite Mercer representation, the next theorem assumes an infinite representation.

Theorem 3.3. *Suppose Assumptions G.1, K.1, K.2 with $M = \infty$, K.3, K.4, and N.1. Let $M(T)$ be as in Assumption K.4. Then, for any $\delta, \gamma \in (0, \infty)$, there is a $c \in (0, \infty)$ such that for all T, with $L \doteq M(T)$,*

$$\mathbb{P}\left(\eta' \mathcal{K}(X, X)\eta > \delta^2\right) \leqslant d \exp\left(-c c_{1,T}(\delta, \gamma, L)\right) + d \exp\left(-c_{2,T}(\gamma)\right) + \frac{2d^2\sigma^2}{\delta^2 T} \beta_1 T^{-\beta_2}. \tag{3.5}$$

There are constants $c_1, c_2 \in (0, \infty)$ and $\mathrm{c}_0 \in (0, \infty)$ such that if

$$\delta \geqslant \mathrm{c}_0 \sqrt{\frac{\log(T)}{T}} L^{\frac{b_2}{2}} \gamma, \tag{3.6}$$

and

$$\gamma \geqslant \sqrt{4\sigma^2 \log(dT)} \tag{3.7}$$

then, with probability at least $1 - c_1 T^{-c_2}$,

$$\eta' \mathcal{K}(X, X)\eta \leqslant \delta^2. \tag{3.8}$$

Constants: Note that the constant b_2 in (3.3) and (3.6) stems from Assumption K.4 and is kernel-dependent. Similarly, the last summand in (3.5) depends on β_1, β_2 which are also kernel-dependent and stem from Assumption K.3. One of the key proof ideas of Theorem 3.3 is to separate the infinite Mercer representation into a finite truncation and a remainder term. The remainder term is then bounded by the last summand in (3.5). The choice of $M(T)$ in Theorem 3.3 also depends on the kernel. For instance, the Gaussian kernel requires $M(T) = \log(T)$ while the periodic kernel requires $M(T) = T$; in this case, $b_2 = 0$ in (3.6), so the rate is not effected.

Rate of convergence: We note that Theorems 3.2 and 3.3 achieve the optimal convergence rate $\sqrt{\log(T)/T}$ (potentially up to some logarithmic term). The rate is considered optimal in the sense that it coincides with the rate one would obtain if $\mathcal{K}(X, X)$ were replaced by a deterministic matrix with bounded eigenvalues. Whether or not we get an additional logarithmic term depends on two aspects.

The first aspect relates to the errors in the underlying regression problem (1.1). If the errors in (1.1) are bounded, that is, they satisfy Assumption N.3, we can take $\gamma = L_\varepsilon$ in (3.4) and (3.7) and there is no additional logarithmic term caused by the errors.

The second aspect that can lead to an additional logarithmic term only applies when Assumption K.1 is satisfied with $M = \infty$, namely when the kernels admit an infinite Mercer representation. Then, for instance, the Gaussian kernel, satisfies Assumption K.3 with $M = \log(T)$ and $b_2 = dp/2-1$ (see Lemma 6.7). Whenever $dp > 2$, this results in an additional logarithmic term in the choices of δ in Theorems 3.1 and 3.3.

Connection to U-statistics: We point out that a crucial assumption for Theorem 3.3 is the separability of the kernel (Assumption K.2). In particular, there are certain U-statistics that may not satisfy a Mercer representation. A general result for U-statistics can be found in Duchemin et al. (2023), who consider U-statistics of order two for Markov chains. Their results require a reverse Doeblin condition. While $\eta' \mathcal{K}(X, X)\eta$ in (3.8) can be written as a U-statistic of a Markov chain, their results cannot be applied since the Markov chain does not satisfy the reverse Doeblin condition.

3.2 Examples

In this section we discuss several scenarios that are covered by our assumptions.

3.2.1 Infinite Mercer representation

In this section, we present several examples of kernels that satisfy Assumption K.2 with $M = \infty$ and for which our results hold.

Gaussian kernel: Suppose the kernels $K_1, \ldots, K_d$ are Gaussian, i.e.,

$$K_i(x, y) = \exp\left(-\|x - y\|^2/\tau^2\right), \quad x, y \in \mathcal{X}^p. \tag{3.9}$$

Note that we use a common smoothing parameter τ for each kernel. Our results also hold if each kernel K_i has a dimension-dependent smoothing parameter τ_i. This change does not affect the proofs, so we, in order to reduce the notational burden, assume that the kernels have a common smoothing parameter.

Theorem 3.4 says that the concentration inequality in Theorem 3.3 holds for the Gaussian kernel when the noise sequence is Gaussian.

Theorem 3.4. *Suppose Assumption N.2. Then, under Assumption G.1, the results from Theorems 3.1 and 3.3 hold for the Gaussian kernel.*

Proof: The result follows from Lemma 6.1, Lemma 6.5, and Lemma 6.7. ■

The proof of the results used in the proof of Theorem 3.4 are quite involved and can be found in Section 6. We present two more examples but, for simplicity, consider the case $d = p = 1$.

Periodic kernel: Suppose the kernel K_1 is the periodic Sobolev kernel, i.e.,

$$\begin{aligned} K(x, y) &= \sigma^2\left(1 + 2\sum_{k=1}^{\infty} \frac{1}{(2\pi k)^{2s}} \cos(2\pi k(x-y))\right) \\ &= \sigma^2\left(1 + \frac{(-1)^{s+1}}{(2s)!} \mathrm{B}_{2s}(|x-y|)\right), \quad x, y \in \mathbb{R}, \end{aligned}$$

where B_{2s} is the Bernoulli polynomial of degree $2s$; see Section 2.1 in Wahba (1990).

Theorem 3.5. *Under Assumptions G.1 and N.1, the result from Theorem 3.3 holds for the periodic Sobolev kernel with $s > 1/2$.*

Proof: It suffices to verify Assumptions K.2–K.4. Assumption K.2 holds by Section 5.3 in Karvonen, Cockayne, Tronarp, and Särkkä (2023). Therein it is shown that, with

$$\phi_k(x) = \sigma\sqrt{2}(2\pi k)^{-s}\cos(2\pi k x), \quad \phi_{-k}(x) = \sigma\sqrt{2}(2\pi k)^{-s}\sin(2\pi k x), \quad k \in \mathbb{N},$$

and $\phi_0 \doteq \sigma$, we have

$$K(x, y) = \sigma^2\left(1 + \frac{(-1)^{s+1}}{(2s)!} \mathrm{B}_{2s}(|x-y|)\right) = \sum_{k\in\mathbb{Z}} \phi_k(x)\phi_k(y).$$

Assumption K.3 is satisfied since

$$\sum\nolimits_{k=1}^{M} \sum\nolimits_{j=1}^{N(k)} \beta_{i,j,k}^2 = \sum\nolimits_{k=1}^{M} \|\phi_k\|_\infty^2 \leqslant \sum\nolimits_{k=1}^{M} \sigma^2 4(2\pi k)^{-2s} \leqslant b_1.$$

Finally, Assumption K.4 can be verified through

$$\sum\nolimits_{j=1}^{N(k)} \mathbb{E}((\phi_{i,j,k}(X_{t-1}))^2) = \mathbb{E}((\sigma\sqrt{2}(2\pi k)^{-s}\cos(2\pi k X_{t-1}))^2) \leqslant ck^{-2s} = \alpha_k.$$

Letting $M(T) = T$, we have

$$\sum\nolimits_{k=M(T)}^{\infty} \alpha_k = \sum\nolimits_{k=T}^{\infty} ck^{-2s} \leqslant \beta_1 T^{-\beta_2},$$

where, due to our choice of s, $\beta_2 = 2s - 1 > 0$. Analogous estimates hold for the terms involving ϕ_{-k}, so the result follows. ■

3.2.2 Finite Mercer representation

In this section we provide two examples of kernels with finite Mercer representations to which our results can be applied.

Polynomial kernel: We assume that Assumption N.3 is satisfied, and therefore that $\mathcal{X}$ is compact. For $x, y \in \mathcal{X}^p$ and $m \in \mathbb{N}$, $c > 0$, define the polynomial kernel

$$K(x, y) = (x'y + c)^m.$$

This kernel satisfies Assumption K.2 with the finite Mercer representation $K(x, y) = \phi(x)'\phi(y)$, where $\phi(x) \in \mathbb{R}^{\binom{dp+m}{m}}$ is given by

$$\phi(x) = \left(\left(\binom{m}{n_1, \ldots, n_{dp+1}}\right)^{1/2} c^{n_{dp+1}/2} \prod_{i=1}^{dp} x_i^{n_i}\right)_{n_i \in \mathbb{N}_0,\, \sum_{i=1}^{dp+1} n_i = m}.$$

Since $\mathcal{X}$ is compact, Assumption K.3 is also satisfied.

Mercer sigmoid kernel: The sigmoid kernel is defined using the sigmoid function

$$K_{\text{sig}}(x, y) = \tanh(ax'y + c), \quad x, y \in \mathbb{R}^{dp}.$$

The sigmoid function is widely used as an activation function in neural networks (Sharma, Sharma, and Athaiya, 2017). However, K_{sig} is not separable in the sense of Assumption K.2. Instead, Carrington, Fieguth, and Chen (2014) introduced the Mercer sigmoid kernel given by

$$K(x, y) = \frac{1}{dp} \sum_{i=1}^{dp} \tanh\left(\frac{x_i - c}{b}\right) \tanh\left(\frac{y_i - c}{b}\right), \quad x, y \in \mathcal{X}^p. \tag{3.10}$$

and demonstrated that it exhibits similar behavior to K_{sig}. In addition, the Mercer sigmoid kernel defined in (3.10) satisfies Assumptions K.2 and K.3.

3.2.3 Stochastic regression model

As a special case of a nonlinear vector autoregression model, our results cover stochastic nonlinear regression models. Such regression models can be written as $y_t = g(X_t) + \varepsilon_t$, where the d-dimensional predictors X_t and the errors ε_t are drawn independently from unknown distributions.

Then, Theorem 3.1 is satisfied under the stated conditions. Those assumptions are satisfied, for instance, under the scenarios lined out in Sections 3.2.1 and 3.2.2. However, we note that for the Gaussian kernel, we need to impose additional assumptions on the distribution of the predictors. For example, our results hold when the predictors X_t are drawn from a Gaussian distribution.

We note that while our results cover the stochastic regression case, this setting allows for an alternative proof of Theorem 3.3. To be more precise, one can replace Theorem 3.3 by a conditional Hanson-Wright inequality.

4 Auxiliary results

For the reader's convenience, we state some of the results used throughout the paper here.

4.1 Reproducing kernel Hilbert spaces

Suppose that $K_1, \ldots, K_d : \mathcal{X}^p \times \mathcal{X}^p \to \mathbb{R}$ are Mercer kernels with associated RKHS $\mathbb{H}_1, \ldots, \mathbb{H}_d$, respectively. Let $\mathbb{H}$ be the Hilbert space defined as

$$\mathbb{H} \doteq \bigtimes_{i=1}^{d} \mathbb{H}_i = \{h \in \mathcal{B}(\mathcal{X}^p : \mathbb{R}^d) : h = (h_1, \ldots, h_d)',\ h_i \in \mathbb{H}_i,\ i = 1, \ldots, d\}, \tag{4.1}$$

with inner product

$$\langle h, \widetilde{h} \rangle_{\mathbb{H}} = \sum_{i=1}^{d} \langle h_i, \widetilde{h}_i \rangle_{\mathbb{H}_i}. \tag{4.2}$$

For a kernel K_i satisfying Assumption K.2, we have, for $f, g \in \mathbb{H}_i$,

$$\langle f, g \rangle_{\mathbb{H}_i} = \sum_{k=1}^{M} \lambda_{i,k}^{-1} \sum_{j=1}^{N(k)} \langle f, \phi_{i,j,k} \rangle_{L^2(\mathcal{X}^p, \pi)} \langle g, \phi_{i,j,k} \rangle_{L^2(\mathcal{X}^p, \pi)},$$

where $\langle \cdot, \cdot \rangle_{L^2(\mathcal{X}^p, \pi)}$ is the standard inner product on $L^2(\mathcal{X}^p, \pi)$.

Definition 4.1. *A vector-valued RKHS is a Hilbert space $\mathbb{H}$ of functions $f : \mathcal{X}^p \to \mathbb{R}^d$, such that for every $z \in \mathbb{R}^d$, and $x \in \mathcal{X}^p$, the map $y \mapsto \mathcal{K}(x, y)z$ belongs to $\mathbb{H}$ and, moreover, $\mathcal{K}$ has the reproducing property*

$$\langle f, \mathcal{K}(x, \cdot)z\rangle_{\mathbb{H}} = f(x)'z.$$

A discussion of vector-valued RKHSs can be found in Alvarez, Rosasco, and Lawrence (2012). The following lemma is an immediate consequence of the fact that each $\mathbb{H}_i$ is an RKHS, (4.1), and (4.2).

Lemma 4.2. *The space $\mathbb{H}$ is an RKHS with associated reproducing kernel (2.2).*

The following lemma shows that $\mathcal{K}$ is uniformly bounded.

Lemma 4.3. *Consider the kernel $\mathcal{K} : \mathcal{X}^p \times \mathcal{X}^p \to \mathbb{R}^{d\times d}$ defined in (2.2). Under Assumption K.1, we have $\sup_{x\in\mathcal{X}^p} \|\mathcal{K}(x, x)\| \leqslant \kappa$.*

Proof: For $x \in \mathcal{X}^p$, $\|\mathcal{K}(x, x)\|^2 = \sum_{i=1}^d |K_i(x, x)|^2 \leqslant \sum_{i=1}^d \kappa_i^2 = \kappa^2$. ■

The following result shows how the RKHS norm can be used to bound the supremum norm of elements of $\mathbb{H}$.

Lemma 4.4. *If $f = (f_1, \ldots, f_d)' \in \mathbb{H}$, then $\|f\|_\infty \leqslant \kappa\|f\|_{\mathbb{H}}$.*

Proof: From Corollary 4.36 of Steinwart and Christmann (2008), we see that for each $i = 1, \ldots, d$, $\|f_i\|_\infty \leqslant \kappa_i\|f_i\|_{\mathbb{H}_i}$, so, using Cauchy-Schwarz inequality,

$$\|f\|_\infty \leqslant \max_{i=1,\ldots,d} \|f_i\|_\infty \leqslant \sum_{i=1}^d \kappa_i\|f_i\|_{\mathbb{H}_i} \leqslant \left(\sum_{i=1}^d \kappa_i^2\right)^{1/2} \left(\sum_{i=1}^d \|f_i\|_{\mathbb{H}_i}^2\right)^{1/2} = \left(\sum_{i=1}^d \kappa_i^2\right)^{1/2} \|f\|_{\mathbb{H}}.$$

■

Lemma 4.5. *Let $\mathbb{H}$ be a separable Hilbert space with inner product $\langle\cdot, \cdot\rangle_{\mathbb{H}}$. Let X and Y be $\mathbb{H}$-valued random variables such that $\mathbb{E}_{\mathbb{H}}(X), \mathbb{E}_{\mathbb{H}}(Y), \mathrm{Var}_{\mathbb{H}}(X)$, and $\mathrm{Var}_{\mathbb{H}}(Y)$ are well defined. If X and Y are independent, then*

$$\mathbb{E}\left[\langle X, Y\rangle_{\mathbb{H}}\right] = \langle\mathbb{E}_{\mathbb{H}}(X), \mathbb{E}_{\mathbb{H}}(Y)\rangle_{\mathbb{H}}.$$

Lemma 4.5 follows from the discussion in Chapter 1.6 of Bosq (2000).

Lemma 4.6. *Recall the definitions of $\mathcal{K}_X$ and $\mathcal{K}(X, X)$ from (2.9) and (2.11), respectively. For $u, v \in \mathbb{R}^{d(T-p)\times 1}$, we have*

$$\langle\mathcal{K}_X(\cdot)u, \mathcal{K}_X(\cdot)v\rangle_{\mathbb{H}} = u'\mathcal{K}(X, X)v.$$

Proof: We write $u' = (u_1', \ldots, u_d')$ and $v' = (v_1', \ldots, v_d')$, where $u_i' = (u_{i,p+1}, \ldots, u_{i,T})$ and $v_i' = (v_{i,p+1}, \ldots, v_{i,T})$. Observe that

$$\langle\mathcal{K}_X(\cdot)u, \mathcal{K}_X(\cdot)v\rangle_{\mathbb{H}} = \sum\nolimits_{i=1}^d \langle(\mathcal{K}_X(\cdot)u)_i, (\mathcal{K}_X(\cdot)v)_i\rangle_{\mathbb{H}_i} \tag{4.3}$$

and

$$\begin{aligned}\langle(\mathcal{K}_X(\cdot)u)_i, (\mathcal{K}_X(\cdot)v)_i\rangle_{\mathbb{H}_i} &= \sum\nolimits_{j=p+1}^T \sum\nolimits_{l=p+1}^T \langle(\mathcal{K}_X(\cdot))_{i,j}u_{i,j}, (\mathcal{K}_X(\cdot))_{i,l}v_{i,l}\rangle_{\mathbb{H}_i} \\ &= \sum\nolimits_{j=p+1}^T \sum\nolimits_{l=p+1}^T u_{i,j}\langle(\mathcal{K}_X(\cdot))_{i,j}, (\mathcal{K}_X(\cdot))_{i,l}\rangle_{\mathbb{H}_i}v_{i,l} \\ &= \sum\nolimits_{j=p+1}^T \sum\nolimits_{l=p+1}^T u_{i,j}K_i(X_{(j-p):(j-1)}, X_{(l-p):(l-1)})v_{i,l} \\ &= u_i'K_i(X, X)v_i.\end{aligned} \tag{4.4}$$

The result follows on combining (2.11), (4.3), and (4.4). ■

Lemma 4.7. *For any bounded $f \in L^{2,d}(\mathcal{X}^p, \pi)$, we have*

$$(L_{\mathcal{K},X} + \lambda I)^{-1} L_{\mathcal{K},X} f(z) = \mathcal{K}_X(z)(\mathcal{K}(X,X) + \lambda T I_{d(T-p)})^{-1} G_f(X), \quad z \in \mathcal{X}^p,$$

where $G_f(X)$ is defined analogously to $G_g(X)$ in (2.5).

Proof: Let $h(z) \doteq (L_{\mathcal{K},X} + \lambda I)^{-1} L_{\mathcal{K},X} f(z)$, so that

$$(L_{\mathcal{K},X} + \lambda I) h(z) = L_{\mathcal{K},X} f(z). \tag{4.5}$$

Then, $h(z) = \lambda^{-1}(-L_{\mathcal{K},X} h(z) + L_{\mathcal{K},X} f(z))$, so there are $\nu_{p+1}, \ldots, \nu_T \in \mathbb{R}^d$ such that with

$$\widehat{\nu} \doteq \begin{pmatrix} \widehat{\nu}_1' & \cdots & \widehat{\nu}_d' \end{pmatrix}' \in \mathbb{R}^{d(T-p)}, \quad \widehat{\nu}_i = \begin{pmatrix} \nu_{i,p+1} & \nu_{i,p+2} & \cdots & \nu_{i,T} \end{pmatrix}' \in \mathbb{R}^{T-p}, \quad i = 1, \ldots, d,$$

we have

$$h(z) = \sum\nolimits_{t=p+1}^T \mathcal{K}(Y_{t-1}, z)\nu_t = \mathcal{K}_X(z)\widehat{\nu}. \tag{4.6}$$

It follows from (4.6) and the definition of $L_{\mathcal{K},X}$ that

$$\begin{aligned} L_{\mathcal{K},X} h(z) &= \frac{1}{T} \sum\nolimits_{t=p+1}^T \mathcal{K}(Y_{t-1}, z) h(Y_{t-1}) \\ &= \frac{1}{T} \sum\nolimits_{t=p+1}^T \mathcal{K}(Y_{t-1}, z) \sum\nolimits_{s=p+1}^T \mathcal{K}(Y_{s-1}, Y_{t-1})\nu_s = \mathcal{K}_X(z) \frac{1}{T} \mathcal{K}(X,X)\widehat{\nu}. \end{aligned} \tag{4.7}$$

We also have that

$$\begin{aligned} L_{\mathcal{K},X}(f)(z) &= \frac{1}{T} \sum\nolimits_{t=p+1}^T \mathcal{K}(X_{(t-p):(t-1)}, z) f(X_{(t-p):(t-1)}) \\ &= \frac{1}{T} \sum\nolimits_{t=p+1}^T \mathcal{K}(Y_{t-1}, z) f(Y_{t-1}) = \frac{1}{T} \mathcal{K}_X(z) G_f(X). \end{aligned} \tag{4.8}$$

Combining (4.5), (4.7), and (4.8), we see that

$$\mathcal{K}_X(z) \left(\left(\mathcal{K}(X,X) + \lambda T I_{d(T-p)} \right) \widehat{\nu} - G_f(X) \right) = 0, \quad z \in \mathcal{X}^p. \tag{4.9}$$

Recalling that $\mathcal{K}$ is positive definite, it follows from the fact that (4.9) holds for all $z \in \mathcal{X}^p$ that

$$\widehat{\nu} = \left(\mathcal{K}(X,X) + \lambda T I_{d(T-p)} \right)^{-1} G_f(X).$$

Together, the previous display and (4.6) imply that

$$h(z) = \mathcal{K}_X(z) \left(\mathcal{K}(X,X) + \lambda T I_{d(T-p)} \right)^{-1} G_f(X).$$

The result follows. ■

4.2 Geometric ergodicity

Our concentration results crucially depend on the geometric ergodicity of the process $\{Y_t\}$ defined in (2.1). We rephrase here a few properties of nonlinear VAR models. We begin by recalling the definition of a geometrically ergodic Markov chain.

Definition 4.8. *An $\mathcal{X}^p$-valued Markov chain $\{Y_t\}$ is said to be geometrically ergodic if there is a $\pi \in \mathcal{P}(\mathcal{X}^p)$, a $\rho \in (0,1)$, and a π-integrable measurable function $J : \mathcal{X}^p \to [0,\infty)$ such that*

$$\|P^n(y,\cdot) - \pi(\cdot)\|_{TV} \leqslant \rho^n J(y), \;\; y \in \mathcal{X}^p, \;\; n \in \mathbb{N}_0,$$

where

$$P^n(y,A) \doteq \mathbb{P}_y(Y_n \in A), \quad A \in \mathcal{B}(\mathcal{X}^p), \; y \in \mathcal{X}^p, \tag{4.10}$$

and $\|\cdot\|_{TV}$ denotes the total variation norm on $\mathcal{P}(\mathcal{X}^p)$.

Remark 4.9. We require the somewhat nonstandard assumption that J is integrable. In view of (1.1) and Assumption G.1, this is satisfied, for instance, when Assumption N.1 holds. Then, one can infer that the chain is aperiodic and that Doeblin condition is satisfied, which implies the uniform ergodicity of the chain; see Theorem 16.0.2 in Meyn and Tweedie (2012).

The following lemma says that the lagged process $\{Y_t\}$ is geometrically ergodic on $\mathcal{X}^p$. This follows immediately from Theorem 1 of Lu and Jiang (2001) (see also Remark 4.1 therein).

Lemma 4.10. *Under Assumptions G.1 and N.1, the Markov chain $\{Y_t\}$ defined in (2.1) is geometrically ergodic.*

There have been several works discussing geometric ergodicity for nonlinear VAR models, including Cline and Pu (1999) and Lu and Jiang (2001). Throughout this work, we take advantage of this property.

4.3 Abstract estimates

The convergence rates for the KRR estimator $\widehat{g}_T$ established in Theorem 3.1 depend on the deterministic quantity $\|g_\lambda - g\|_\infty$. Theorem 4.11 presents one estimate that can be used to control this term.

Theorem 4.11. *Suppose the function $g = (g_1, \ldots, g_d)'$ is continuous and Assumption N.3. If $L_{\mathcal{K}}^{-r} g \in L^{2,d}(\mathcal{X}^p, \pi)$ for some $r \in (1/2, 1]$, then*

$$\|g_\lambda - g\|_\infty \leqslant \kappa \lambda^{r-1/2} \|L_{\mathcal{K}}^{-r} g\|_{L^{2,d}(\mathcal{X}^p,\pi)}.$$

The condition $L_{\mathcal{K}}^{-r} g \in L^{2,d}(\mathcal{X}^p, \pi)$ originates from Smale and Zhou (2005), where r is interpreted as a smoothness parameter for g. When $r = 1/2$, the condition $L_{\mathcal{K}}^{-1/2} g \in L^{2,d}(\mathcal{X}^p, \pi)$ is equivalent to the assumption that $g \in \mathbb{H}$, as shown by the identity

$$\begin{aligned}
\|L_{\mathcal{K}}^{-1/2} g\|^2_{L^{2,d}(\mathcal{X}^p,\pi)} &= \sum_{i=1}^d \left\| \sum_{k=1}^\infty \sum_{j=1}^{N(k)} \langle g_i, \phi_{i,j,k} \rangle_{L^2(\mathcal{X}^p,\pi)} \phi_{i,j,k} / \sqrt{\lambda_{i,k}} \right\|^2_{L^2(\mathcal{X}^p,\pi)} \\
&= \sum_{i=1}^d \sum_{k=1}^\infty \sum_{j=1}^{N(k)} \langle g_i, \phi_{i,j,k} \rangle^2_{L^2(\mathcal{X}^p,\pi)} / \lambda_{i,k} = \|g\|^2_{\mathbb{H}}.
\end{aligned}$$

Together with Theorem 3.1, Theorem 4.11 provides a convergence rate for our estimator. A proof of Theorem 4.11 can be found in Liu and Li (2023). We also refer the reader to Theorems 6(b) and 7 in Liu and Li (2023), where the authors assume that the respective kernel admits an eigendecomposition with respect to the Fourier basis, resulting in estimates similar to those in Theorem 4.11.

5 Proofs of main results

5.1 Proofs of results in Section 3.1

Recall the definition of the integral operator $L_{\mathcal{K}}$ in (2.17). We consider a sample analogue of $L_{\mathcal{K}}$ defined by

$$
\begin{aligned}
L_{\mathcal{K},X}(f)(z) &= \frac{1}{T}\sum_{t=p+1}^{T} \mathcal{K}(X_{(t-p):(t-1)}, z) f(X_{(t-p):(t-1)}) \\
&= \frac{1}{T}\sum_{t=p+1}^{T} \mathcal{K}(Y_{t-1}, z) f(Y_{t-1}) \\
&= \frac{1}{T}\mathcal{K}_X(z) G_f(X), \quad z \in \mathcal{X}^p,
\end{aligned} \tag{5.1}
$$

where $G_f(X)$ is defined analogously to $G_g(X)$ in (2.5). Similarly, recall the function g_λ defined in (2.16). We consider a sample analogue $g_{X,\lambda} : \mathcal{X}^p \to \mathbb{R}^d$ given by

$$
g_{X,\lambda} \doteq (L_{\mathcal{K},X} + \lambda I)^{-1} L_{\mathcal{K},X} g. \tag{5.2}
$$

Proof of Theorem 3.1 Let δ satisfy (3.1) and note that

$$
\mathbb{P}\left(\|\widehat{g}_T - g\|_\infty > \delta\right) \leqslant \mathbb{P}\left(\|\widehat{g}_T - g_{X,\lambda}\|_\infty > \delta/2\right) + \mathbb{P}\left(\|g_{X,\lambda} - g_\lambda\|_\infty + \|g_\lambda - g\|_\infty > \delta/2\right), \tag{5.3}
$$

where g_λ and $g_{X,\lambda}$ are defined in (2.16) and (5.2), respectively. We consider the two summands in (5.3) separately.

Recall $L_{\mathcal{K},X}$ from (5.1). Beginning with the second summand in (5.3), let

$$
\mathcal{G}(X) \doteq (L_{\mathcal{K},X} + \lambda I)^{-1} L_{\mathcal{K},X}(g) - (L_{\mathcal{K}} + \lambda I)^{-1} L_{\mathcal{K}}(g) = g_{X,\lambda} - g_\lambda.
$$

With δ as in (3.1), we get

$$
\begin{aligned}
\mathbb{P}\left(\|g_{X,\lambda} - g_\lambda\|_{\mathbb{H}} + \|g_\lambda - g\|_\infty > \delta/2\right) &\leqslant \mathbb{P}\left(\|g_{X,\lambda} - g_\lambda\|_{\mathbb{H}} > \frac{1}{\lambda}\sqrt{\frac{\log(T)}{T}} L^{\frac{b_2}{2}} \gamma\, \mathrm{C}_0(g)\right) \\
&\leqslant \mathbb{P}\left(\|g_{X,\lambda} - g_\lambda\|_{\mathbb{H}} > \mathrm{C}_0(g)\frac{1}{\lambda}\sqrt{\frac{\log(T)}{T}}\right) \\
&= \mathbb{P}\left(\|\mathcal{G}(X)\|_{\mathbb{H}} > \mathrm{C}_0(g)\frac{1}{\lambda}\sqrt{\frac{\log(T)}{T}}\right) \leqslant c_1 T^{-c_2},
\end{aligned} \tag{5.4}
$$

where the final inequality is due to Lemma 5.1.

We now consider the first summand in (5.3). Using (2.5), (2.14), (5.2), and Lemma 4.7, we get

$$
\begin{aligned}
\widehat{g}_T(\cdot) - g_{X,\lambda}(\cdot) &= \mathcal{K}_X(\cdot)(\mathcal{K}(X,X) + \lambda T I_{d(T-p)})^{-1} Y - (L_{\mathcal{K},X} + \lambda I)^{-1} L_{\mathcal{K},X}(g)(\cdot) \\
&= \mathcal{K}_X(\cdot)(\mathcal{K}(X,X) + \lambda T I_{d(T-p)})^{-1} Y - \mathcal{K}_X(\cdot)(\mathcal{K}(X,X) + \lambda T I_{d(T-p)})^{-1} G_g(X) \\
&= \frac{1}{T}\mathcal{K}_X(\cdot)(\mathcal{K}(X,X)/T + \lambda I_{d(T-p)})^{-1} \eta.
\end{aligned} \tag{5.5}
$$

Let

$$
\begin{aligned}
Z_{X,T} &\doteq (\mathcal{K}(X,X)/T + \lambda I_{d(T-p)})^{-1} \\
&= \mathrm{diag}((K_1(X,X)/T + \lambda I_{d(T-p)})^{-1}, \ldots, (K_d(X,X)/T + \lambda I_{d(T-p)})^{-1}),
\end{aligned}
$$

and observe that

$$
\begin{aligned}
\|\widehat{g}_T - g_{X,\lambda}\|_{\mathbb{H}}^2 &= \frac{1}{T^2}\eta'(\mathcal{K}(X,X)/T + \lambda I_{d(T-p)})^{-1}\mathcal{K}(X,X)(\mathcal{K}(X,X)/T + \lambda I_{d(T-p)})^{-1}\eta \\
&= \frac{1}{T^2}\mathrm{tr}(\eta' Z_{X,T}\mathcal{K}(X,X)Z_{X,T}\eta) = \frac{1}{T^2}\mathrm{tr}(\eta\eta' Z_{X,T}\mathcal{K}(X,X)Z_{X,T}) \\
&\leqslant (\lambda_{\max}(Z_{X,T}))^2 \frac{1}{T^2}\mathrm{tr}(\eta'\mathcal{K}(X,X)\eta) \\
&\leqslant \lambda^{-2}\frac{1}{T^2}\eta'\mathcal{K}(X,X)\eta,
\end{aligned} \tag{5.6}
$$

where the first identity follows from (5.5) and Lemma 4.6. For the first inequality, note that the matrices $\eta\eta'$, $Z_{X,T}$ and $\mathcal{K}(X,X)$ are all positive semidefinite, so the inequality follows from Theorem 1 of Fang, Loparo, and Feng (1994). The final inequality follows upon observing that

$$
\begin{aligned}
\lambda_{\max}(Z_{X,T}) &= \lambda_{\max}((\mathcal{K}(X,X)/T + \lambda I_{d(T-p)})^{-1}) \\
&= (\lambda_{\min}(\mathcal{K}(X,X)/T + \lambda I_{d(T-p)}))^{-1} \leqslant \lambda^{-1}.
\end{aligned}
$$

It remains to show a high probability bound for the quantity in (5.6). We distinguish between the cases where Assumption K.2 is satisfied with $M < \infty$ or $M = \infty$.

- $M < \infty$: In this case, δ satisfies (3.1) with $L \doteq M$. Note that $\lambda\delta$ satisfies the inequality in (3.3). Thus, we can apply Theorem 3.2, which, under Assumptions K.2 and K.3, ensures that there are constants $c_1, c_2 \in (0,\infty)$ such that for all T, with probability at least $1 - c_1 T^{-c_2}$,

$$
\frac{1}{T^2}\eta'\mathcal{K}(X,X)\eta \leqslant (\lambda\delta)^2. \tag{5.7}
$$

  Then, the result follows on combining (5.3), (5.4), (5.6), and the probability bound in (5.7).

- $M = \infty$: In this case, δ satisfies (3.1) with $L = M(T)$ as given in Assumption K.4. Note that $\lambda\delta$ satisfies the inequality in (3.6). Thus, we can apply Theorem 3.3, which, under Assumptions K.2–K.4, ensures that there are constants $c_1, c_2 \in (0,\infty)$ such that for all T, with probability at least $1 - c_1 T^{-c_2}$,

$$
\frac{1}{T^2}\eta'\mathcal{K}(X,X)\eta \leqslant (\lambda\delta)^2. \tag{5.8}
$$

  Then, the result follows on combining (5.3), (5.4), (5.6), and the probability bound in (5.8). ■

Proof of Theorem 3.2 Note that

$$
\begin{aligned}
&\mathbb{P}\left(\eta'\mathcal{K}(X,X)\eta > \delta^2\right) \\
&= \mathbb{P}\left(\frac{1}{T^2}\sum_{i=1}^{d}\sum_{t=p+1}^{T}\sum_{s=p+1}^{T}\varepsilon_{i,t}K_i(Y_{t-1},Y_{s-1})\varepsilon_{i,s} > \delta^2\right) \\
&= \mathbb{P}\left(\frac{1}{T^2}\sum_{i=1}^{d}\sum_{k=1}^{M}\lambda_{i,k}\sum_{j=1}^{N(k)}\sum_{s=p+1}^{T}\sum_{t=p+1}^{T}\varepsilon_{i,s}\varepsilon_{i,t}\phi_{i,j,k}(Y_{t-1})\phi_{i,j,k}(Y_{s-1}) > \delta^2\right) \\
&= \mathbb{P}\left(\sum_{i=1}^{d}\sum_{k=1}^{M}\lambda_{i,k}\sum_{j=1}^{N(k)}\left(\frac{1}{T}\sum_{t=p+1}^{T}\varepsilon_{i,t}\phi_{i,j,k}(Y_{t-1})\right)^2 > \delta^2\right)
\end{aligned}
$$

$$\leqslant \sum_{i=1}^{d} \mathbb{P}\left(\sum_{k=1}^{M} \lambda_{i,k} \sum_{j=1}^{N(k)} \left(\frac{1}{T}\sum_{t=p+1}^{T} \varepsilon_{i,t}\phi_{i,j,k}(Y_{t-1})\right)^2 > \frac{\delta^2}{d}\right). \tag{5.9}$$

For $i = 1, \ldots, d$, define

$$A_{i,M} \doteq \sum_{k=1}^{M} \lambda_{i,k} \sum_{j=1}^{N(k)} \left(\frac{1}{T}\sum_{t=p+1}^{T} \varepsilon_{i,t}\phi_{i,j,k}(Y_{t-1})\right)^2,$$

and observe that

$$\begin{aligned}
&\mathbb{P}\left(A_{i,M} > \frac{\delta^2}{d}\right)\\
&= \mathbb{P}\left(\sum_{k=1}^{M}\sum_{j=1}^{N(k)} \left(\frac{1}{T}\sum_{t=p+1}^{T} \varepsilon_{i,t}\lambda_{i,k}^{1/2}\phi_{i,j,k}(Y_{t-1})\right)^2 > \frac{\delta^2}{d}\right)\\
&\leqslant \mathbb{P}\left(\max_{k=1,\ldots,M}\max_{j=1,\ldots,N(k)} \left(\frac{1}{T}\sum_{t=p+1}^{T} \varepsilon_{i,t}\beta_{i,j,k}^{-1}\lambda_{i,k}^{1/2}\phi_{i,j,k}(Y_{t-1})\right)^2 \sum_{k=1}^{M-1}\sum_{j=1}^{N(k)} \beta_{i,j,k}^2 > \frac{\delta^2}{d}\right)\\
&\leqslant \mathbb{P}\left(\max_{k=1,\ldots,M}\max_{j=1,\ldots,N(k)} \left(\frac{1}{T}\sum_{t=p+1}^{T} \varepsilon_{i,t}\beta_{i,j,k}^{-1}\lambda_{i,k}^{1/2}\phi_{i,j,k}(Y_{t-1})\right)^2 > \delta_1^2\right),
\end{aligned} \tag{5.10}$$

where (5.10) follows by (2.22) in Assumption K.3 with

$$\delta_1 \doteq \delta\left(db_1M^{b_2}\right)^{-1/2}. \tag{5.11}$$

Then, from (5.10), we see that, for $\gamma > 0$,

$$\begin{aligned}
&\mathbb{P}\left(A_{i,M} > \frac{\delta^2}{d}\right)\\
&\leqslant \mathbb{P}\left(\max_{k=1,\ldots,M-1}\max_{j=1,\ldots,N(k)} \left|\frac{1}{T}\sum_{t=p+1}^{T} \varepsilon_{i,t}\beta_{i,j,k}^{-1}\lambda_{i,k}^{1/2}\phi_{i,j,k}(Y_{t-1})\right| > \delta_1\right)\\
&\leqslant \mathbb{P}\left(\max_{k=1,\ldots,M}\max_{j=1,\ldots,N(k)} \mathbf{1}_{\{\|\eta\|_\infty\leqslant\gamma\}}\left|\frac{1}{T}\sum_{t=p+1}^{T} \varepsilon_{i,t}\beta_{i,j,k}^{-1}\lambda_{i,k}^{1/2}\phi_{i,j,k}(Y_{t-1})\right| > \delta_1\right)\\
&\qquad + \mathbb{P}(\|\eta\|_\infty > \gamma) \qquad (5.12)\\
&\leqslant \mathbb{P}\left(\max_{k=1,\ldots,M}\max_{j=1,\ldots,N(k)} \left|\frac{1}{T}\sum_{t=p+1}^{T} \varepsilon_{i,t}\beta_{i,j,k}^{-1}\lambda_{i,k}^{1/2}\phi_{i,j,k}(Y_{t-1})\mathbf{1}_{\{|\varepsilon_{i,t}|\leqslant\gamma\}}\right| > \delta_1\right)\\
&\qquad + \mathbb{P}(\|\eta\|_\infty > \gamma), \qquad (5.13)
\end{aligned}$$

where (5.12) is due to a truncation argument by intersecting with the event $\{\|\eta\|_\infty \leqslant \gamma\}$, and (5.13) follows on observing that for $a_1, \ldots, a_n, b_1, \ldots, b_n, c \in \mathbb{R}$,

$$\mathbf{1}_{\{\max\{a_1,\ldots,a_n\}\leqslant c\}}\left|\sum_{i=1}^{n} b_i\right| \leqslant \left|\sum_{i=1}^{n} \mathbf{1}_{\{a_i\leqslant c\}} b_i\right|.$$

Letting, for $i = 1, \dots, d$,

$$P_{1,i,M} \doteq \mathbb{P}\left(\max_{k=1,\dots,M} \max_{j=1,\dots,N(k)} \left| \frac{1}{T} \sum_{t=p+1}^{T} \varepsilon_{i,t} \beta_{i,j,k}^{-1} \lambda_{i,k}^{1/2} \phi_{i,j,k}(Y_{t-1}) \mathbf{1}_{\{|\varepsilon_{i,t}| \leqslant \gamma\}} \right| > \delta_1 \right), \tag{5.14}$$

and

$$P_{2,i,M} \doteq \mathbb{P}(\|\eta\|_\infty > \gamma), \tag{5.15}$$

it follows from (5.13) that

$$\mathbb{P}\left(A_{i,M} > \frac{\delta^2}{d} \right) \leqslant P_{1,i,M} + P_{2,i,M}. \tag{5.16}$$

We consider the probabilities $P_{1,i,M}$ and $P_{2,i,M}$ separately. For $P_{1,i,M}$ in (5.14), consider the function $F_{i,M} : (\mathcal{X}^p)^{T-p+1} \to [0, \infty)$ defined by

$$\begin{aligned} &F_{i,M}(y_p, \dots, y_T) \\ &= \max_{k=1,\dots,M} \max_{j=1,\dots,N(k)} \frac{1}{T} \left| \sum_{t=p+1}^{T} \left((y_{t,1} - f(y_{t-1}))_i \beta_{i,j,k}^{-1} \lambda_{i,k}^{\frac{1}{2}} \phi_{i,j,k}(y_{t-1}) \mathbf{1}_{\{|(y_{t,1} - f(y_{t-1}))_i| \leqslant \gamma\}} \right) \right|. \end{aligned}$$

Then,

$$P_{1,i,M} = \mathbb{P}\left(F_{i,M}(Y_{p:T}) > \delta_1\right) = \mathbb{P}\left(F_{i,M}(Y_{p:T}) - \mathbb{E}\left(F_{i,M}(Y_{p:T})\right) > \delta_1 - \mathbb{E}\left(F_{i,M}(Y_{p:T})\right)\right). \tag{5.17}$$

To establish an upper bound for the probability in (5.17), we first estimate the expected value in that display. Observe that

$$\begin{aligned} &\left(\mathbb{E}\left(F_{i,M}(Y_{p:T})\right)\right)^2 \leqslant \mathbb{E}\left(\left(F_{i,M}(Y_{p:T})\right)^2 \right) \\ &= \mathbb{E}\left(\max_{k=1,\dots,M} \max_{j=1,\dots,N(k)} \left| \frac{1}{T} \sum_{t=p+1}^{T} \varepsilon_{i,t} \beta_{i,j,k}^{-1} \lambda_{i,k}^{\frac{1}{2}} \phi_{i,j,k}(Y_{t-1}) \mathbf{1}_{\{|\varepsilon_{i,t}| \leqslant \gamma\}} \right|^2 \right) \\ &= \mathbb{E}\left(\max_{k=1,\dots,M} \max_{j=1,\dots,N(k)} \frac{1}{T^2} \sum_{s,t=p+1}^{T} \varepsilon_{i,s}\varepsilon_{i,t} \mathbf{1}_{\{|\varepsilon_{i,s}|,|\varepsilon_{i,t}| \leqslant \gamma\}} \beta_{i,j,k}^{-2} \lambda_{i,k} \phi_{i,j,k}(Y_{t-1}) \phi_{i,j,k}(Y_{s-1}) \right) \\ &\leqslant \frac{1}{T^2} \sum_{s,t=p+1}^{T} \mathbb{E}\left(\varepsilon_{i,s}\varepsilon_{i,t} \mathbf{1}_{\{|\varepsilon_{i,s}|,|\varepsilon_{i,t}| \leqslant \gamma\}} \max_{k=1,\dots,M} \max_{j=1,\dots,N(k)} \beta_{i,j,k}^{-2} \lambda_{i,k} \phi_{i,j,k}(Y_{t-1}) \phi_{i,j,k}(Y_{s-1}) \right), \end{aligned} \tag{5.18}$$

where the first inequality uses Jensen's inequality and the second inequality uses that the maximum of a sum is bounded above by the sum of the maxima. We consider the diagonal and cross terms of the sum in the last line of (5.18) separately. Define

$$\Phi(Y_{t-1}, Y_{s-1}) \doteq \max_{k=1,\dots,M} \max_{j=1,\dots,N(k)} \beta_{i,j,k}^{-2} \lambda_{i,k} \phi_{i,j,k}(Y_{t-1}) \phi_{i,j,k}(Y_{s-1}),$$

and note, due to the definition of $\beta_{i,j,k}$ in (2.21), that $\|\Phi\|_\infty = 1$. For the cross terms in the last line of (5.18), if $t > s$, then

$$\begin{aligned} &\mathbb{E}\left(\varepsilon_{i,s}\varepsilon_{i,t} \mathbf{1}_{\{|\varepsilon_{i,s}|,|\varepsilon_{i,t}| \leqslant \gamma\}} \Phi(Y_{t-1}, Y_{s-1}) \right) \\ &= \mathbb{E}\left(\mathbb{E}\left[\varepsilon_{i,s}\varepsilon_{i,t} \mathbf{1}_{\{|\varepsilon_{i,s}|,|\varepsilon_{i,t}| \leqslant \gamma\}} \Phi(Y_{t-1}, Y_{s-1}) \mid X_{1:(r-1)},\ r = \max\{s,t\} \right] \right) \end{aligned} \tag{5.19}$$

$$= \mathbb{E}\bigg(\Phi(Y_{t-1}, Y_{s-1})\varepsilon_{i,s}\mathbf{1}_{\{|\varepsilon_{i,s}|\leqslant\gamma\}}\mathbb{E}\Big[\varepsilon_{i,t}\mathbf{1}_{\{|\varepsilon_{i,t}|\leqslant\gamma\}} \mid X_{1:(t-1)}\Big]\bigg)$$

$$= \mathbb{E}\bigg(\Phi(Y_{t-1}, Y_{s-1})\varepsilon_{i,s}\mathbf{1}_{\{|\varepsilon_{i,s}|\leqslant\gamma\}}\bigg)\mathbb{E}\big[\varepsilon_{i,t}\mathbf{1}_{\{|\varepsilon_{i,t}|\leqslant\gamma\}}\big] \tag{5.20}$$

$$\leqslant \Big(\mathbb{E}(\varepsilon_{i,s}^2)\mathbb{E}\big[\mathbf{1}_{\{|\varepsilon_{i,s}|\leqslant\gamma\}}\big]\Big)^{\frac{1}{2}}\mathbb{E}\big[\varepsilon_{i,t}\mathbf{1}_{\{|\varepsilon_{i,t}|\leqslant\gamma\}}\big] \leqslant 2\sigma^2(dT)^{-1}, \tag{5.21}$$

where (5.19) uses the law of total expectation and (5.20) uses that ε_t is independent of $Y_1, \ldots, Y_{t-1}$ and $\varepsilon_1, \ldots, \varepsilon_{t-1}$. The first inequality in (5.21) uses the fact that $\|\Phi\|_\infty = 1$ and the Cauchy-Schwarz inequality. The second inequality in (5.21) follows on observing that, since $\mathbb{E}(\varepsilon_{i,t}) = 0$, with γ as in (3.4),

$$\begin{aligned}\left(\mathbb{E}\left(\varepsilon_{i,t}\mathbf{1}_{\{|\varepsilon_{i,t}|\leqslant\gamma\}}\right)\right)^2 &= \left(-\mathbb{E}\left(\varepsilon_{i,t}\mathbf{1}_{\{|\varepsilon_{i,t}|>\gamma\}}\right)\right)^2 \leqslant \mathbb{E}\left(\varepsilon_{i,t}^2\right)\mathbb{E}\left(\mathbf{1}_{\{|\varepsilon_{i,t}|>\gamma\}}\right) \\ &\leqslant \sigma^2\mathbb{P}\left(|\varepsilon_{i,t}| > \gamma\right) \leqslant 2\sigma^2\exp\left(-\gamma^2/(2\sigma^2)\right) \leqslant 2\sigma^2(dT)^{-2},\end{aligned} \tag{5.22}$$

where the third inequality in (5.22) follows from Assumption N.1.

For the diagonal terms of the sum in the last line of (5.18), we have

$$\begin{aligned}&\mathbb{E}\bigg(\mathbb{E}\Big[\varepsilon_{i,t}^2\mathbf{1}_{\{|\varepsilon_{i,t}|\leqslant\gamma\}}\Phi(Y_{t-1}, Y_{t-1}) \mid X_{1:(t-1)}\Big]\bigg) \\ &= \mathbb{E}\left(\varepsilon_{i,t}^2\mathbf{1}_{\{|\varepsilon_{i,t}|\leqslant\gamma\}}\right)\mathbb{E}\bigg(\Phi(Y_{t-1}, Y_{t-1})\bigg) \leqslant \sigma^2,\end{aligned} \tag{5.23}$$

where we once more used that $\|\Phi\|_\infty = 1$. Combining (5.18), (5.21), (5.22) and (5.23), we see that

$$(\mathbb{E}(F_{i,M}(Y_{p:T})))^2 \leqslant 3\frac{\sigma^2}{T}. \tag{5.24}$$

From (5.17) and (5.24), we see that

$$P_{1,i,M} \leqslant \mathbb{P}\left(F_{i,M}(Y_{p:T}) - \mathbb{E}\left(F_{i,M}(Y_{p:T})\right) > \delta_2\right), \tag{5.25}$$

where, with δ_1 as in (5.11),

$$\delta_2 \doteq \delta_1 - \frac{2\sigma}{\sqrt{T}}. \tag{5.26}$$

Note that Lemma 5.5 says that $F_{i,M}$ is separately bounded by $2\gamma/T$. Then, applying Theorem 0.2 in Dedecker and Gouëzel (2015) with $L = 2\gamma/T$, we can see that there is a $C_{\mathrm{mc}} \in (0, \infty)$ such that

$$P_{1,i,M} \leqslant 2\exp\left(-\frac{C_{\mathrm{mc}}\delta_2^2}{T\left(2\gamma/T\right)^2}\right) = 2\exp\left(-\frac{C_{\mathrm{mc}}T\delta_2^2}{4\gamma^2}\right).$$

It remains to bound the probability $P_{2,i,M}$ defined in (5.15). Using Lemma 5.3, we see that

$$P_{2,i,M} \leqslant \exp\left(-\left(\gamma - \sqrt{2\sigma^2\log(dT)}\right)^2/(2\sigma^2)\right). \tag{5.27}$$

Together, (5.16), (5.25), (5.26), and (5.27) show that

$$
\begin{aligned}
&\mathbb{P}\left(A_{i,M} > \frac{\delta^2}{d}\right) \\
&\leqslant 2\exp\left(-\frac{C_{\mathrm{mc}} T \delta_2^2}{4\gamma^2}\right) + \exp\left(-\left(\gamma - \sqrt{2\sigma^2\log(dT)}\right)^2/(2\sigma^2)\right), \\
&\leqslant 2\exp\left(-C_{\mathrm{mc}}\frac{T}{\gamma^2}\left(\delta\left(db_1 M^{b_2}\right)^{-1/2} - \frac{\sigma}{\sqrt{T}}\right)^2\right) + \exp\left(-\left(\gamma - \sqrt{2\sigma^2\log(dT)}\right)^2/(2\sigma^2)\right) \\
&\leqslant \exp\left(-C_{\mathrm{mc}} c_{1,T}(\delta,\gamma,M)\right) + \exp\left(-c_{2,T}(\gamma)\right),
\end{aligned}
\tag{5.28}
$$

where $c_{1,T}, c_{2,T}$ are as in (3.2). Together, (5.9) and (5.28) complete the proof. ■

Proof of Theorem 3.3 In the following, we write $\eta'\mathcal{K}(X,X)\eta$ as a sum of a kernel with a finite series expansion and a remainder term. We then apply Theorem 3.2 to the part of the decomposition corresponding to a kernel with a finite expansion. The remainder term is handled using suitable upper bounds for the second moments of the eigenfunctions in the tail of the series expansion. Using arguments similar to those in (5.9), we can decompose the probability of interest as

$$
\begin{aligned}
\mathbb{P}\left(\eta'\mathcal{K}(X,X)\eta > \delta^2\right) &= \mathbb{P}\left(\sum_{i=1}^{d}\sum_{k=1}^{\infty}\lambda_{i,k}\sum_{j=1}^{N(k)}\left(\frac{1}{T}\sum_{t=p+1}^{T}\varepsilon_{i,t}\phi_{i,j,k}(Y_{t-1})\right)^2 > \delta^2\right) \\
&\leqslant \sum_{i=1}^{d}\mathbb{P}\left(\sum_{k=1}^{\infty}\lambda_{i,k}\sum_{j=1}^{N(k)}\left(\frac{1}{T}\sum_{t=p+1}^{T}\varepsilon_{i,t}\phi_{i,j,k}(Y_{t-1})\right)^2 > \frac{\delta^2}{d}\right) \\
&\leqslant \sum_{i=1}^{d}\left(\mathbb{P}\left(A_{i,M} > \frac{\delta^2}{2d}\right) + \mathbb{P}\left(B_{i,M} > \frac{\delta^2}{2d}\right)\right),
\end{aligned}
\tag{5.29}
$$

where

$$
A_{i,M} = \sum\nolimits_{k=1}^{M-1}\lambda_{i,k}\sum\nolimits_{j=1}^{N(k)}\left(\frac{1}{T}\sum\nolimits_{t=p+1}^{T}\varepsilon_{i,t}\phi_{i,j,k}(Y_{t-1})\right)^2,
$$

and

$$
B_{i,M} = \sum\nolimits_{k=M}^{\infty}\lambda_{i,k}\sum\nolimits_{j=1}^{N(k)}\left(\frac{1}{T}\sum\nolimits_{t=p+1}^{T}\varepsilon_{i,t}\phi_{i,j,k}(Y_{t-1})\right)^2.
$$

We consider the two probabilities in (5.29) separately. For the first summand, using Theorem 3.2, we see that

$$
\sum\nolimits_{i=1}^{d}\mathbb{P}\left(A_{i,M} > \frac{\delta^2}{2d}\right) \leqslant d\exp\left(-C_{\mathrm{mc}} c_{1,T}(\delta,\gamma,M)\right) + d\exp\left(-c_{2,T}(\gamma)\right). \tag{5.30}
$$

For the second summand in (5.29), since $B_{i,M} \geqslant 0$, using Markov's inequality gives

$$
\mathbb{P}\left(B_{i,M} > \frac{\delta^2}{2d}\right) \leqslant \frac{2d}{\delta^2}\mathbb{E}(B_{i,M}). \tag{5.31}
$$

Applying the Fubini-Tonelli Theorem, we have

$$
\mathbb{E}(B_{i,M}) = \mathbb{E}\left(\sum_{k=M}^{\infty}\lambda_{i,k}\sum_{j=1}^{N(k)}\left(\frac{1}{T}\sum_{t=p+1}^{T}\varepsilon_{i,t}\phi_{i,j,k}(Y_{t-1})\right)^2\right)
$$

$$\leqslant \frac{\sigma^2}{T^2} \sum_{k=M}^{\infty} \lambda_{i,k} \sum_{j=1}^{N(k)} \sum_{t=p+1}^{T} \mathbb{E}((\phi_{i,j,k}(Y_{t-1}))^2), \tag{5.32}$$

where (5.32) follows upon using the law of total expectation and noting that ε_t is independent of Y_{t-1}, that $\mathbb{E}(\varepsilon_{i,t}) = 0$, that $\varepsilon_{i,t}$ and $\varepsilon_{i,s}$ are independent whenever $s \neq t$, and that $\mathbb{E}(\varepsilon_{i,t}^2) \leqslant \sigma^2$, where $\sigma^2 \doteq \max_{i=1,\ldots,d} \sigma_i^2$. Then, using Assumption K.4, it follows from (5.32) that

$$\mathbb{E}(B_{i,M}) \leqslant \frac{\sigma^2}{T^2} \sum_{k=M}^{\infty} \lambda_{i,k} \sum_{j=1}^{N(k)} \sum_{t=p+1}^{T} \mathbb{E}((\phi_{i,j,k}(Y_{t-1}))^2) \leqslant \frac{\sigma^2}{T} \sum_{k=M(T)}^{\infty} \alpha_k \leqslant \frac{\sigma^2}{T} \beta_1 T^{-\beta_2}. \tag{5.33}$$

From (5.31) and (5.33), we see that

$$\mathbb{P}\left(B_{i,M} > \frac{\delta^2}{2d}\right) \leqslant \frac{2d\sigma^2}{\delta^2 T} \beta_1 T^{-\beta_2}. \tag{5.34}$$

Finally, on combining (5.29), (5.30), and (5.34), we see that

$$\mathbb{P}\left(\eta' \mathcal{K}(X,X)\eta > \delta^2\right) \leqslant d\exp\left(-C_{\mathrm{mc}} c_{1,T}(\delta,\gamma,M)\right) + d\exp\left(-c_{2,T}(\gamma)\right) + \frac{2d^2\sigma^2}{\delta^2 T}\beta_1 T^{-\beta_2},$$

where $c_{1,T}$ and $c_{2,T}$ are as in (3.2). ■

5.2 Auxiliary concentration results and their proofs

Lemma 5.1. *For the regression function $g : \mathcal{X}^p \to \mathbb{R}^d$ satisfying Assumption G.1, let*

$$\mathcal{G}(X) \doteq (L_{\mathcal{K},X} + \lambda I)^{-1} L_{\mathcal{K},X}\, g - (L_{\mathcal{K}} + \lambda I)^{-1} L_{\mathcal{K}}\, g.$$

There are constants $c_1, c_2 \in (0,\infty)$ and $\mathrm{c}_0 \in (0,\infty)$ such that, if $\delta \geqslant \mathrm{C}_0(g)\frac{1}{\lambda}\sqrt{\log(T)/T}$ where $\mathrm{C}_0(g) \doteq \mathrm{c}_0\sqrt{d}\kappa^2\|g\|_\infty$, then, with probability at least $1 - c_1 T^{-c_2}$, we have $\|\mathcal{G}(X)\|_{\mathbb{H}} \leqslant \delta$.

Proof: Observe that

$$\begin{aligned}&\mathcal{G}(X)\\ &= (L_{\mathcal{K},X} + \lambda I)^{-1} L_{\mathcal{K},X}\, g - (L_{\mathcal{K},X} + \lambda I)^{-1} L_{\mathcal{K}}\, g + (L_{\mathcal{K},X} + \lambda I)^{-1} L_{\mathcal{K}}\, g - (L_{\mathcal{K}} + \lambda I)^{-1} L_{\mathcal{K}}\, g,\end{aligned}$$

so

$$\|\mathcal{G}(X)\|_{\mathbb{H}} \leqslant \|(L_{\mathcal{K},X} + \lambda I)^{-1}\left(L_{\mathcal{K},X}\, g - L_{\mathcal{K}}\, g\right)\|_{\mathbb{H}} + \|(L_{\mathcal{K},X} + \lambda I)^{-1} L_{\mathcal{K}}\, g - (L_{\mathcal{K}} + \lambda I)^{-1} L_{\mathcal{K}}\, g\|_{\mathbb{H}}. \tag{5.35}$$

We consider the two summands in (5.35) separately. First, since $L_{\mathcal{K},X}$ is self adjoint and positive semidefinite, we have

$$\|(L_{\mathcal{K},X} + \lambda I)^{-1}\left(L_{\mathcal{K},X}\, g - L_{\mathcal{K}}\, g\right)\|_{\mathbb{H}} \leqslant \lambda^{-1}\|L_{\mathcal{K},X}\, g - L_{\mathcal{K}}\, g\|_{\mathbb{H}} \leqslant \delta,$$

where the last inequality holds for $\delta \geqslant \mathrm{c}_0\,\lambda^{-1}\sqrt{d}\kappa\|g\|_\infty\sqrt{\log(T)/T}$ with probability at least $1 - c_1 T^{-c_2}$, which follows from Lemma 5.2 since g is bounded. For the second summand in (5.35), we can once more use the fact that $L_{\mathcal{K},X}$ is positive semidefinite to see that

$$\begin{aligned}&\|(L_{\mathcal{K},X} + \lambda I)^{-1} L_{\mathcal{K}}\, g - (L_{\mathcal{K}} + \lambda I)^{-1} L_{\mathcal{K}}\, g\|_{\mathbb{H}}\\ &= \|(L_{\mathcal{K},X} + \lambda I)^{-1}(L_{\mathcal{K}} + \lambda I)(L_{\mathcal{K}} + \lambda I)^{-1} L_{\mathcal{K}}\, g - (L_{\mathcal{K},X} + \lambda I)^{-1}(L_{\mathcal{K},X} + \lambda I)(L_{\mathcal{K}} + \lambda I)^{-1} L_{\mathcal{K}}\, g\|_{\mathbb{H}}\\ &= \|(L_{\mathcal{K},X} + \lambda I)^{-1}\left((L_{\mathcal{K}} + \lambda I)(L_{\mathcal{K}} + \lambda I)^{-1} L_{\mathcal{K}}\, g - (L_{\mathcal{K},X} + \lambda I)(L_{\mathcal{K}} + \lambda I)^{-1} L_{\mathcal{K}}\, g\right)\|_{\mathbb{H}}\\ &\leqslant \lambda^{-1}\left\|L_{\mathcal{K}}\left((L_{\mathcal{K}} + \lambda I)^{-1} L_{\mathcal{K}}\, g\right) - L_{\mathcal{K},X}\left((L_{\mathcal{K}} + \lambda I)^{-1} L_{\mathcal{K}}\, g\right)\right\|_{\mathbb{H}}\\ &= \lambda^{-1}\|L_{\mathcal{K}}\, g_\lambda - L_{\mathcal{K},X}\, g_\lambda\|_{\mathbb{H}},\end{aligned}$$

where g_λ is as in (2.16). To apply Lemma 5.2, we need to show that g_λ is bounded. Taking $\widetilde{g} \doteq 0$ in (2.15), we see that

$$\|g_\lambda - g\|_2^2 + \lambda\|g_\lambda\|_{\mathbb{H}}^2 \leqslant \|\widetilde{g} - g\|_2^2 + \lambda\|\widetilde{g}\|_{\mathbb{H}}^2 = \|g\|_2^2,$$

which ensures that

$$\|g_\lambda\|_2^2 \leqslant \sqrt{2}\|g\|_2, \quad \|g_\lambda\|_{\mathbb{H}} \leqslant \lambda^{-1/2}\|g\|_2.$$

Using Lemma 4.4, we see that

$$\|g_\lambda\|_\infty \leqslant \kappa\|g_\lambda\|_{\mathbb{H}} \leqslant \kappa\lambda^{-1/2}\|g\|_2 \leqslant \kappa\lambda^{-1/2}\|g\|_\infty,$$

which shows that g_λ is bounded. Since g_λ is bounded, it follows from Lemma 5.2 that there is some $\mathrm{c}_0 > 0$ such that with $\delta \geqslant \mathrm{c}_0\sqrt{d}\kappa^2\lambda^{-1/2}\|g\|_\infty\sqrt{\log(T)/T}$ there are constants $c_1, c_2 \in (0, \infty)$ such that, with probability at least $1 - c_1 T^{-c_2}$, $\|L_{\mathcal{K},X} g_\lambda - L_{\mathcal{K}} g_\lambda\|_{\mathbb{H}} \leqslant \delta$. ■

Lemma 5.2. *Let $f = (f_1, \ldots, f_d)' : \mathcal{X}^p \to \mathbb{R}^d$ be a bounded map. There are constants $c_1, c_2 \in (0, \infty)$ and $\mathrm{c}_0 > 0$ such that, if $\delta \geqslant \mathrm{c}_0\sqrt{d}\kappa\|f\|_\infty\sqrt{\log(T)/T}$, then, with probability at least $1 - c_1 T^{-c_2}$,*

$$\|L_{\mathcal{K},X} f - L_{\mathcal{K}} f\|_{\mathbb{H}} \leqslant \delta.$$

Proof: Let $F : (\mathcal{X}^p)^{T-p} \to \mathbb{R}$ be given by

$$F(y_p, \ldots, y_{T-1}) = \left\| \frac{1}{T} \sum_{t=p+1}^{T} f(y_{t-1})\mathcal{K}(y_{t-1}, \cdot) - L_{\mathcal{K}} f \right\|_{\mathbb{H}}, \quad y_p, \ldots, y_{T-1} \in \mathcal{X}^p,$$

and observe that

$$F(Y_p, \ldots, Y_{T-1}) = \|L_{\mathcal{K},X}(f) - L_{\mathcal{K}} f\|_{\mathbb{H}}.$$

We show that there is a constant $C_F \in (0, \infty)$ such that, for all $T \in \mathbb{N}$ and

$$\widetilde{y} = (y_p, \ldots, y_{i-1}, \widetilde{z}, y_{i+1}, \ldots y_{T-1}), \quad \bar{y} = (y_p, \ldots, y_{i-1}, \bar{z}, y_{i+1}, \ldots, y_{T-1}) \in (\mathcal{X}^p)^{T-p},$$

we have

$$|F(\widetilde{y}) - F(\bar{y})| \leqslant C_F T^{-1} \tag{5.36}$$

with $C_F \doteq 4\sqrt{d}\kappa\|f\|_\infty$. The proof of (5.36) can be found below. If (5.36) holds, then, with further explanations given below,

$$\begin{aligned}
&\mathbb{P}(F(Y_p, \ldots, Y_{T-1}) > \delta) \\
&= \mathbb{P}(F(Y_p, \ldots, Y_{T-1}) - \mathbb{E}(F(Y_p, \ldots, Y_{T-1})) > \delta - \mathbb{E}(F(Y_p, \ldots, Y_{T-1}))) \\
&\leqslant \mathbb{P}(F(Y_p, \ldots, Y_{T-1}) - \mathbb{E}(F(Y_p, \ldots, Y_{T-1})) > \delta - C_f T^{-1/2}) \qquad (5.37) \\
&\leqslant 2\exp\left(-\frac{C_{\mathrm{mc}}}{C_F^2} T(\delta - C_f T^{-1/2})^2 \right) \leqslant c_1 T^{-c_2}, \qquad (5.38)
\end{aligned}$$

where (5.37) follows by Lemma 5.4. Due to Lemma 4.10, $\{Y_t\}$ is a geometrically ergodic Markov chain, so we can apply Theorem 0.2 of Dedecker and Gouëzel (2015). This result, together with (5.36), ensures that there is a $C_{\mathrm{mc}} \in (0, \infty)$ such (5.38) holds. The final inequality then follows for some constants $c_1, c_2 \in (0, \infty)$, since $\delta \geqslant (C_F/\sqrt{C_{\mathrm{mc}}})\sqrt{\log(T)/T}$.

Thus, it suffices to prove (5.36). Observe that

$$
\begin{aligned}
&|F(\widetilde{y}) - F(\bar{y})| \\
&= \left| \left\| \frac{1}{T} \sum\nolimits_{t=p+1}^{T} f(\widetilde{y}_{t-1}) \mathcal{K}(\widetilde{y}_{t-1}, \cdot) - L_{\mathcal{K}} f \right\|_{\mathbb{H}} - \left\| \frac{1}{T} \sum\nolimits_{t=p+1}^{T} f(\bar{y}_{t-1}) \mathcal{K}(\bar{y}_{t-1}, \cdot) - L_{\mathcal{K}} f \right\|_{\mathbb{H}} \right| \\
&\leqslant \left\| \frac{1}{T} \sum\nolimits_{t=p+1}^{T} f(\widetilde{y}_{t-1}) \mathcal{K}(\widetilde{y}_{t-1}, \cdot) - \frac{1}{T} \sum\nolimits_{t=p+1}^{T} f(\bar{y}_{t-1}) \mathcal{K}(\bar{y}_{t-1}, \cdot) \right\|_{\mathbb{H}} \\
&= \frac{1}{T} \| f(\widetilde{z}) \mathcal{K}(\widetilde{z}, \cdot) - f(\bar{z}) \mathcal{K}(\bar{z}, \cdot) \|_{\mathbb{H}} \\
&\leqslant \frac{1}{T} \sup_{w, \widetilde{w} \in \mathcal{X}^p} \| f(w) \mathcal{K}(w, \cdot) - f(\widetilde{w}) \mathcal{K}(\widetilde{w}, \cdot) \|_{\mathbb{H}} \leqslant \frac{1}{T} C_F.
\end{aligned} \tag{5.39}
$$

The final inequality in (5.39) is verified in the calculations below.

For $w, \widetilde{w} \in \mathcal{X}^p$ we have

$$
\begin{aligned}
&\| f(w) \mathcal{K}(w, \cdot) - f(\widetilde{w}) \mathcal{K}(\widetilde{w}, \cdot) \|_{\mathbb{H}} \\
&\leqslant \| f(w) \mathcal{K}(w, \cdot) - f(w) \mathcal{K}(\widetilde{w}, \cdot) \|_{\mathbb{H}} + \| f(w) \mathcal{K}(\widetilde{w}, \cdot) - f(\widetilde{w}) \mathcal{K}(\widetilde{w}, \cdot) \|_{\mathbb{H}} \\
&= \sum\nolimits_{i=1}^{d} \left(\| f_i(w) (K_i(w, \cdot) - K_i(\widetilde{w}, \cdot)) \|_{\mathbb{H}_i} + \| f_i(w) K_i(\widetilde{w}, \cdot) - f_i(\widetilde{w}) K_i(\widetilde{w}, \cdot) \|_{\mathbb{H}_i} \right) \\
&\leqslant (2 + \sqrt{2}) \sum\nolimits_{i=1}^{d} \kappa_i \| f \|_\infty \leqslant 4 \sqrt{d} \kappa \| f \|_\infty,
\end{aligned} \tag{5.40}
$$

where the second inequality follows on observing that, due to Assumption K.1,

$$
\| f_i(w) (K_i(w, \cdot) - K_i(\widetilde{w}, \cdot)) \|_{\mathbb{H}_i} \leqslant \| f \|_\infty \left(\| K_i(w, \cdot) \|_{\mathbb{H}_i} + \| K_i(\widetilde{w}, \cdot) \|_{\mathbb{H}_i} \right) \leqslant 2 \| f \|_\infty \kappa_i
$$

and

$$
\| f_i(w) K_i(\widetilde{w}, \cdot) - f_i(\widetilde{w}) K_i(\widetilde{w}, \cdot) \|_{\mathbb{H}_i} \leqslant 2 \| f_i \|_\infty \| K_i(\widetilde{w}, \cdot) \|_{\mathbb{H}_i} \leqslant 2 \| f \|_\infty \kappa_i.
$$

Combining (5.39) and (5.40), we see that (5.36) holds with our choice of C_F. The result follows. ■

Lemma 5.3 establishes a concentration inequality for the maximum of the noise sequence, which is used in Theorem 3.2.

Lemma 5.3. *Recall η from (2.6) and suppose Assumption N.1. Then, for any $\gamma > \sqrt{2\sigma^2 \log(dT)}$,*

$$
\mathbb{P}(\| \eta \|_\infty > \gamma) \leqslant \exp\left(- \left(\gamma - \sqrt{2\sigma^2 \log(dT)} \right)^2 / (2\sigma^2) \right).
$$

Proof: Note that

$$
\begin{aligned}
\mathbb{P}\left(\| \eta \|_\infty > \gamma \right) &= \mathbb{P}\left(\max_{t=1,\ldots,T} \| \varepsilon_t \|_\infty - \mathbb{E}\left(\max_{t=1,\ldots,T} \| \varepsilon_t \|_\infty \right) > \gamma - \mathbb{E}\left(\max_{t=1,\ldots,T} \| \varepsilon_t \|_\infty \right) \right) \\
&\leqslant \exp\left(- \left(\gamma - \mathbb{E}\left(\max_{t=1,\ldots,T} \| \varepsilon_t \|_\infty \right) \right)^2 / (2\sigma^2) \right) \qquad (5.41) \\
&\leqslant \exp\left(- \left(\gamma - \sqrt{2\sigma^2 \log(dT)} \right)^2 \right) / (2\sigma^2) \right), \qquad (5.42)
\end{aligned}
$$

where we applied Example 2.29 of Wainwright (2019) with $\sigma^2 = \max_{i=1,\ldots,d} \sigma_i^2$ in (5.41) and assuming that $\gamma > \mathbb{E}(\max_{t=1,\ldots,T} \|\varepsilon_t\|_\infty)$. In (5.42), we use that

$$\mathbb{E}(\max_{t=1,\ldots,T} \|\varepsilon_t\|_\infty) \leqslant \sqrt{2\sigma^2 \log(dT)},$$

which is due to Exercise 2.5.10 in Vershynin (2018). ■

Lemma 5.4 is used in the proof of Lemma 5.1.

Lemma 5.4. *Let $f = (f_1, \ldots, f_d)' : \mathcal{X}^p \to \mathbb{R}^d$ be a bounded map. For each $T \in \mathbb{N}$, let $F_T : (\mathcal{X}^p)^{T-p} \to \mathbb{R}$ be given by*

$$F_T(y_p, \ldots, y_{T-1}) \doteq \left\| \frac{1}{T} \sum_{t=p+1}^{T} f(y_{t-1}) \mathcal{K}(y_{t-1}, \cdot) - L_{\mathcal{K}}(f) \right\|_{\mathbb{H}}.$$

Then, there is a constant $C_f \in (0, \infty)$ such that for all $T \in \mathbb{N}$,

$$\mathbb{E}(F_T(Y_p, \ldots, Y_{T-1})) \leqslant T^{-1/2} C_f.$$

Proof: Let $h : \mathcal{X}^p \to \mathbb{H}$ be given by

$$h(x) = f(x)\mathcal{K}(x, \cdot), \quad x \in \mathcal{X}^p, \tag{5.43}$$

and consider the map $H : \mathcal{X}^p \times \mathcal{X}^p \to \mathbb{R}$ given by

$$H(x, y) = \langle h(x), h(y) \rangle_{\mathbb{H}}, \quad x, y \in \mathcal{X}^p.$$

Then, by the Cauchy-Schwarz inequality, and with further explanations given below,

$$\sup_{x,y \in \mathcal{X}^p} |H(x, y)| \leqslant \sup_{x,y \in \mathcal{X}^p} \|h(x)\|_{\mathbb{H}} \|h(y)\|_{\mathbb{H}} \leqslant \widehat{c}_1 \doteq \|f\|_\infty^2 \kappa \sqrt{d}. \tag{5.44}$$

The last inequality in (5.44) follows, since, by Lemma 4.3,

$$\begin{aligned} \|h(x)\|_{\mathbb{H}}^2 &= \langle f(x)\mathcal{K}(x, \cdot), f(x)\mathcal{K}(x, \cdot) \rangle_{\mathbb{H}} = \sum_{i=1}^{d} \langle f_i(x) K_i(x, \cdot), f_i(x) K_i(x, \cdot) \rangle_{\mathbb{H}} \\ &\leqslant \sum_{i=1}^{d} \|f_i\|_\infty^2 K_i(x, x) \leqslant \|f\|_\infty^2 \sum_{i=1}^{d} K_i(x, x) \leqslant \kappa \|f\|_\infty^2 \sqrt{d}. \end{aligned} \tag{5.45}$$

Additionally, Lemma 4.10 tells us that there are $\pi \in \mathcal{P}(\mathcal{X}^p)$, $\rho \in (0, 1)$, and a π-integrable function $J : \mathcal{X}^p \to [0, \infty)$ such that, for each $s, t \in \mathbb{N}$,

$$\|P^{|s-t|}(x, \cdot) - \pi(\cdot)\|_{\mathrm{TV}} \leqslant \rho^{|s-t|} J(x), \quad x \in \mathcal{X}^p, \tag{5.46}$$

where P is the transition kernel of the process $\{Y_t\}$ defined in (4.10). Then, for $s, t \in \mathbb{N}$, (5.44) and (5.46) tell us that

$$\begin{aligned} &\left| \mathbb{E}_\pi(H(Y_{t-1}, Y_{s-1})) - \int_{\mathcal{X}^p \times \mathcal{X}^p} H(x, y) \pi(dy) \pi(dx) \right| \\ &= \left| \int_{\mathcal{X}^p \times \mathcal{X}^p} H(x, y) P^{|s-t|}(x, dy) \pi(dx) - \int_{\mathcal{X}^p \times \mathcal{X}^p} H(x, y) \pi(dy) \pi(dx) \right| \\ &= \left| \int_{\mathcal{X}^p \times \mathcal{X}^p} H(x, y) \left(P^{|s-t|}(x, dy) - \pi(dy) \right) \pi(dx) \right| \\ &\leqslant \widehat{c}_1 \int_{\mathcal{X}^p \times \mathcal{X}^p} \left| P^{|s-t|}(x, dy) - \pi(dy) \right| \pi(dx) \\ &\leqslant 2\widehat{c}_1 \int_{\mathcal{X}^p} \rho^{|s-t|} J(x) \pi(dx) = \widehat{c}_2 \rho^{|s-t|}, \end{aligned} \tag{5.47}$$

where $\widehat{c}_2 \in [0, \infty)$ is defined as $\widehat{c}_2 = 2\widehat{c}_1 \int_{\mathcal{X}^p} J(x)\pi(dx)$. If $\Pi_1, \Pi_2 \overset{iid}{\sim} \pi$, then, using Lemma 4.5, we see that

$$\begin{aligned}\int_{\mathcal{X}^p \times \mathcal{X}^p} &H(x,y)\pi(dy)\pi(dx) = \mathbb{E}[\langle h(\Pi_1), h(\Pi_2)\rangle_{\mathbb{H}}] \\ &= \langle \mathbb{E}_{\mathbb{H}}(h(\Pi_1)), \mathbb{E}_{\mathbb{H}}(h(\Pi_2))\rangle_{\mathbb{H}} = \langle \pi(h), \pi(h)\rangle_{\mathbb{H}} = \|\pi(h)\|^2_{\mathbb{H}},\end{aligned} \tag{5.48}$$

where

$$\pi(h)(y) = \int_{\mathcal{X}^p} f(x)\mathcal{K}(x,y)\pi(dx).$$

Above we have used the fact that $\pi(h) \in \mathbb{H}$; this can be shown using Lemma 4.5 and the fact that $h(y) \in \mathbb{H}$ for each $y \in \mathcal{X}^p$. Together, (5.47) and (5.48) imply that

$$|\mathbb{E}_\pi(H(Y_{t-1}, Y_{s-1})) - \|\pi(h)\|^2_{\mathbb{H}}| \leqslant \widehat{c}_2 \rho^{|s-t|}. \tag{5.49}$$

Thus, using the fact that $\mathbb{E}_\pi(h(Y_t)) = \pi(h)$ for all t, we have

$$\begin{aligned}&\mathrm{Cov}_{\mathbb{H}}(h(Y_t) - \pi(h), h(Y_s) - \pi(h)) = \mathbb{E}_\pi[\langle h(Y_t) - \pi(h), h(Y_s) - \pi(h)\rangle_{\mathbb{H}}] \\ &= \mathbb{E}_\pi[\langle h(Y_t), h(Y_s)\rangle_{\mathbb{H}}] - \langle \pi(h), \pi(h)\rangle_{\mathbb{H}} - \langle \pi(h), \pi(h)\rangle_{\mathbb{H}} + \langle \pi(h), \pi(h)\rangle_{\mathbb{H}} \\ &= \mathbb{E}_\pi[\langle h(Y_t), h(Y_s)\rangle_{\mathbb{H}}] - \|\pi(h)\|^2_{\mathbb{H}} = \mathbb{E}_\pi(H(Y_t, Y_s)) - \|\pi(h)\|^2_{\mathbb{H}},\end{aligned} \tag{5.50}$$

so (5.49) and (5.50) imply that

$$|\mathrm{Cov}_{\mathbb{H}}(h(Y_t), h(Y_s))| = |\mathrm{Cov}_{\mathbb{H}}(h(Y_t) - \pi(h), h(Y_s) - \pi(h))| \leqslant \widehat{c}_2 \rho^{|s-t|}. \tag{5.51}$$

We also note that, due to (5.45),

$$\sup_{t \in \mathbb{N}} \mathrm{Var}_{\mathbb{H}}(h(Y_{t-1})) \leqslant \widehat{c}_3 \doteq \sup_{x \in \mathcal{X}^p} \|h(x)\|^2_{\mathbb{H}} < \infty. \tag{5.52}$$

Define $\mathcal{I}_{p:T} = \{(i,j) : i, j \in \{p+1, \ldots, T\} \text{ and } i \neq j\}$, then,

$$(\mathbb{E}_\pi (F(Y_p, \ldots, Y_{T-1})))^2 \leqslant \mathbb{E}_\pi (F(Y_p, \ldots, Y_{T-1}))^2 \tag{5.53}$$

$$= \mathbb{E}_\pi \left[\left\| \frac{1}{T} \sum\nolimits_{t=p+1}^{T} f(Y_{t-1})\mathcal{K}(Y_{t-1}, \cdot) - L_{\mathcal{K}}(f) \right\|^2_{\mathbb{H}} \right]$$

$$= \frac{1}{T^2} \mathbb{E}_\pi \left[\left\| \sum\nolimits_{t=p+1}^{T} (h(Y_{t-1}) - \pi(h)) \right\|^2_{\mathbb{H}} \right] \tag{5.54}$$

$$= \frac{1}{T^2} \sum\nolimits_{t=p+1}^{T} \mathrm{Var}_{\mathbb{H}}(h(Y_{t-1})) + \frac{1}{T^2} \sum\nolimits_{(s,t) \in \mathcal{I}_{p:T}} \mathrm{Cov}_{\mathbb{H}}(h(Y_{t-1}), h(Y_{s-1}))$$

$$\leqslant \frac{\widehat{c}_3}{T} + \frac{\widehat{c}_2}{T^2} \sum\nolimits_{s,t=1}^{T} \rho^{|s-t|} \tag{5.55}$$

$$\leqslant \frac{1}{T}\left(\widehat{c}_3 + \widehat{c}_2 \left(1 + \frac{2\rho}{1-\rho}\right)\right) \doteq \frac{1}{T} C_f^2,$$

where (5.53) uses Jensen's inequality, (5.54) uses (5.43) and the fact that $\pi(h) = L_{\mathcal{K}}(f)$, and (5.55) uses (5.51) and (5.52). ■

The next lemma is also used in the proof of Theorem 3.2.

Lemma 5.5. *Let $F_{i,M} : (\mathcal{X}^p)^{T-p+1} \to [0,\infty)$ defined by*

$$\begin{aligned} &F_{i,M}(y_p, \ldots, y_T) \\ &= \max_{k=1,\ldots,M} \max_{j=1,\ldots,N(k)} \frac{1}{T} \left| \sum_{t=p+1}^{T} \left((y_{t,1} - f(y_{t-1}))_i \beta_{i,j,k}^{-1} \lambda_{i,k}^{\frac{1}{2}} \phi_{i,j,k}(y_{t-1}) \mathbf{1}_{\{|(y_{t,1}-f(y_{t-1}))_i| \leqslant \gamma\}} \right) \right|, \end{aligned}$$

where $\beta_{i,j,k}$ is defined as in (2.21). Then, $F_{i,M}$ satisfies a separately bounded condition in the sense of Theorem 0.2. of Dedecker and Gouëzel (2015). Namely, for all $T \in \mathbb{N}$, all $M \leqslant T$ and

$$\widetilde{y} = (y_p, \ldots, y_{j-1}, \widetilde{z}, y_{j+1}, \ldots y_T), \quad \bar{y} = (y_p, \ldots, y_{j-1}, \bar{z}, y_{j+1}, \ldots, y_T) \in (\mathcal{X}^p)^{T-p+1},$$

we have

$$|F_{i,M}(\widetilde{y}) - F_{i,M}(\bar{y})| \leqslant \frac{2\gamma}{T}.$$

Proof: Then,

$$\begin{aligned} &|F_{i,M}(\widetilde{y}) - F_{i,M}(\bar{y})| \\ &\leqslant \max_{k=1,\ldots,M} \max_{j=1,\ldots,N(k)} \frac{1}{T} \Big| \sum_{t=p+1}^{T} ((\widetilde{y}_{t,1} - f(\widetilde{y}_{t-1}))_i \beta_{i,j,k}^{-1} \lambda_{i,k}^{\frac{1}{2}} \phi_{i,j,k}(\widetilde{y}_{t-1}) \mathbf{1}_{\{|(\widetilde{y}_{t,1}-f(\widetilde{y}_{t-1}))_i| \leqslant \gamma\}} \\ &\qquad - (\bar{y}_{t,1} - f(\bar{y}_{t-1}))_i \beta_{i,j,k}^{-1} \lambda_{i,k}^{\frac{1}{2}} \phi_{i,j,k}(\bar{y}_{t-1}) \mathbf{1}_{\{|(\bar{y}_{t,1}-f(\bar{y}_{t-1}))_i| \leqslant \gamma\}} \Big| \\ &= \max_{k=1,\ldots,M} \max_{j=1,\ldots,N(k)} \frac{1}{T} \Big| \sum_{t=j}^{j+1} ((\widetilde{y}_{t,1} - f(\widetilde{y}_{t-1}))_i \beta_{i,j,k}^{-1} \lambda_{i,k}^{\frac{1}{2}} \phi_{i,j,k}(\widetilde{y}_{t-1}) \mathbf{1}_{\{|(\widetilde{y}_{t,1}-f(\widetilde{y}_{t-1}))_i| \leqslant \gamma\}} \\ &\qquad - (\bar{y}_{t,1} - f(\bar{y}_{t-1}))_i \beta_{i,j,k}^{-1} \lambda_{i,k}^{\frac{1}{2}} \phi_{i,j,k}(\bar{y}_{t-1}) \mathbf{1}_{\{|(\bar{y}_{t,1}-f(\bar{y}_{t-1}))_i| \leqslant \gamma\}} \Big| \leqslant \frac{2\gamma}{T}, \end{aligned}$$

where the final inequality uses that, by definition of $\beta_{i,j,k}$ in (2.21), $\beta_{i,j,k}^{-1} \lambda_{i,k}^{1/2} \|\phi_{i,j,k}\|_\infty \leqslant 1$, and

$$|(y_{t,1} - f(y_{t-1}))_i \mathbf{1}_{\{|(y_{t,1}-f(y_{t-1}))_i| \leqslant \gamma\}}| \leqslant \gamma, \quad y_t, y_{t-1} \in \mathcal{X}^p.$$

∎

6 Results for Gaussian kernels and their proofs

Recall the definition of a Gaussian kernel (3.9). In this section, we show that the Gaussian kernel satisfies Assumptions K.1–K.4.

Lemma 6.1 provides an eigenfunction expansion of the Gaussian kernel and states that it satisfies Assumption K.1.

Lemma 6.1. *Recall the set $\mathcal{N}(k)$ from (2.20). Enumerate $\mathcal{N}(k)$ as $\mathcal{N}(k) = \{\boldsymbol{n}_{k,1}, \ldots, \boldsymbol{n}_{k,N(k)}\}$, where, for each $j = 1, \ldots, N(k)$,*

$$\boldsymbol{n}_{k,j} = (n_{k,j,1,1}, \ldots, n_{k,j,1,p}, \ldots, n_{k,j,d,1}, \ldots, n_{k,j,d,p}). \tag{6.1}$$

For $x \in \mathcal{X}^p$ and $\boldsymbol{n}_{k,j} \in \mathcal{N}(k)$, let $x^{\boldsymbol{n}_{k,j}} \doteq \prod_{i=1}^d \prod_{r=1}^p x_{i,r}^{n_{k,j,i,r}}$. The Gaussian kernel in (3.9) satisfies Assumptions K.1 and K.2 with $\phi_{i,j,k} : \mathcal{X}^p \to \mathbb{R}$, $i = 1, \ldots, d$, $k \in \mathbb{N}$, $j = 1, \ldots, N(k)$, given by

$$\phi_{i,j,k}(x) = \exp\left(-\frac{\|x\|^2}{\tau^2}\right) \binom{k}{\boldsymbol{n}_{k,j}}^{1/2} x^{\boldsymbol{n}_{k,j}}, \quad x \in \mathcal{X}^p, \tag{6.2}$$

and

$$\lambda_{i,k} = \frac{1}{k!}\left(\frac{2}{\tau^2}\right)^k. \tag{6.3}$$

Proof: The Gaussian kernel satisfies Assumption K.1 with each $\kappa_i = 1$. Finally, the fact that the Gaussian kernel satisfies Assumption K.2 follows from the Taylor expansion of the map $u \mapsto \exp(u)$. ■

Remark 6.2. For the Gaussian kernel, the convergence in (2.18) (with $M = \infty$) holds pointwise but not uniformly over $\mathcal{X}^p \times \mathcal{X}^p$.

The following lemma establishes an upper bound for exponentially-weighted second moments of Gaussian random variables, and is used in the proof of Lemma 6.4 to compute second moments of the eigenfunctions (6.2).

Lemma 6.3. *Suppose that $X_\alpha \sim \mathcal{N}(\alpha, \sigma^2)$, for $|\alpha| \leqslant L$. Then, there is a constant $C_L \in (0, \infty)$ such that for each $k \in \mathbb{N}$,*

$$\sup_{|\alpha|\leqslant L} \mathbb{E}\left[X_\alpha^{2k} \exp\left(-2\frac{X_\alpha^2}{\tau^2}\right)\right] \leqslant C_L \left(\frac{\sigma^2\tau^2}{4\sigma^2+\tau^2}\right)^k 2^k (k+1)!.$$

Proof: Observe that

$$\begin{aligned}
&\mathbb{E}\left(X_\alpha^{2k} \exp\left(-2\left(\frac{X_\alpha}{\tau}\right)^2\right)\right) \\
&= \int_{-\infty}^{\infty} x^{2k} \exp\left(-2\left(\frac{x}{\tau}\right)^2\right) \exp\left(-\frac{(x-\alpha)^2}{2\sigma^2}\right) dx \\
&= \exp\left(-\frac{\alpha^2}{2} + \frac{\alpha^2\tau^4}{4\sigma^2+\tau^2}\right) \int_{-\infty}^{\infty} x^{2k} \exp\left(-\frac{-(4\sigma^2+\tau^2)x^2 + 2\alpha x\tau^2 - \alpha^2\tau^4/(4\sigma^2+\tau^2))}{2\sigma^2\tau^2}\right) dx \\
&= \left(\frac{2\pi\tau^2}{4\sigma^2+\tau^2}\right)^{1/2} \exp\left(\alpha^2\left(\frac{\tau^4}{4\sigma^2+\tau^2} - \frac{1}{2\sigma^2}\right)\right) \mathbb{E}(Y_\alpha^{2k}),
\end{aligned} \tag{6.4}$$

where $Y_\alpha \sim \mathcal{N}(\mu_\alpha, \sigma_Y^2)$, $\mu_\alpha = \frac{\tau^2\alpha}{4\sigma^2+\tau^2}$, and $\sigma_Y^2 = \frac{\sigma^2\tau^2}{4\sigma^2+\tau^2}$. Let ${}_1F_1$ denote the confluent hypergeometric function of the first kind (see, e.g., Luke (1972), Joshi and Bissu (1996)). Then, by the moment formulas for Gaussian random variables in Winkelbauer (2012), we have

$$\begin{aligned}
\mathbb{E}Y_\alpha^{2k} &= \pi^{-1/2}\sigma_Y^{2k}2^k\Gamma\left(\frac{1+2k}{2}\right) {}_1F_1\left(-k, \frac{1}{2}, -\frac{1}{2}\left(\frac{\mu_\alpha}{\sigma_Y}\right)^2\right) \\
&\leqslant \pi^{-1/2}\sigma_Y^{2k}2^k\Gamma\left(\frac{1+2k}{2}\right) \exp\left(-\frac{1}{2}\left(\frac{\mu_\alpha}{\sigma_Y}\right)^2\right) {}_1F_1\left(\frac{1}{2}+k, \frac{1}{2}, \frac{1}{2}\left(\frac{\mu_\alpha}{\sigma_Y}\right)^2\right) && (6.5) \\
&\leqslant \pi^{-1/2}\sigma_Y^{2k}2^k\Gamma\left(\frac{1+2k}{2}\right) \exp\left(-\frac{1}{2}\left(\frac{\mu_\alpha}{\sigma_Y}\right)^2\right) \left(1 + (1+2k)\left(\frac{\mu_\alpha}{\sigma_Y}\right)^2\right) && (6.6) \\
&\leqslant \sigma_Y^{2k}2^k k!\left(1 + (1+2k)\left(\frac{\mu_\alpha}{\sigma_Y}\right)^2\right), && (6.7)
\end{aligned}$$

where (6.5) follows from (1.18) in Luke (1972) and (6.6) is due to (4.3) in Joshi and Bissu (1996). Combining (6.4) and (6.7), we see that the result holds with

$$C_L \doteq 2\left(\frac{2\pi\tau^2}{4\sigma^2+\tau^2}\right)^{1/2}\left(\sup_{|\alpha|\leqslant L}\left(\exp\left(\alpha^2\left(\frac{\tau^4}{4\sigma^2+\tau^2}-\frac{1}{2\sigma^2}\right)\right)\right)\right)\left(\max\left\{\sup_{|\alpha|\leqslant L}\left(\frac{\mu_\alpha}{\sigma_Y}\right)^2,1\right\}\right).$$

■

Lemma 6.4. *For $k\in\mathbb{N}$ and $j=1,\dots,N(k)$, recall from (6.2) the eigenfunctions of the Gaussian kernel. Then, there is some $b\in(0,\infty)$ such that for all $k\in\mathbb{N}$, $j=1,\dots,N(k)$, $i=1,\dots,d$, and $t\in\mathbb{N}$, $t\geqslant p+1$,*

$$\mathbb{E}((\phi_{i,j,k}(X_{(t-p):(t-1)}))^2)\leqslant b\left(\frac{2\sigma^2\tau^2}{4\sigma^2+\tau^2}\right)^k k!(k+1)^{pd}.$$

Proof: Recall the quantity $\boldsymbol{n}_{k,j}$ introduced in Lemma 6.1. Since k and j are fixed, we, with a slide abuse of notation, write $\boldsymbol{n}$ in place of $\boldsymbol{n}_{k,j}$, where $\boldsymbol{n}\doteq(n_{1,1},\dots,n_{1,p},\dots,n_{d,1},\dots,n_{d,p})$. Let

$$Z_{\alpha,i,r}\sim\mathcal{N}(\alpha_i,\sigma_i^2),\quad \alpha=(\alpha_1,\dots,\alpha_d)'\in\mathbb{R}^d,\ i=1,\dots,d,\ r=1,\dots,p,$$

and note that Lemma 6.3 ensures that there is some $b\in(0,\infty)$ such that

$$\sup_{\|\alpha\|\leqslant\|g\|_\infty}\mathbb{E}\left(\exp\left(-\frac{2Z_{\alpha,i,r}^2}{\tau^2}\right)Z_{\alpha,i,r}^{2n_{i,r}}\right)\leqslant C\left(\frac{\sigma_i^2\tau^2}{4\sigma_i^2+\tau^2}\right)^{n_{i,r}}2^{n_{i,r}}(n_{i,r}+1)!. \tag{6.8}$$

With further explanations given below, the second moments of the eigenfunctions can then be bounded as

$$\begin{aligned}\mathbb{E}((\phi_{i,j,k}(X_{(t-p):(t-1)}))^2)&=\binom{k}{\boldsymbol{n}}\mathbb{E}\left[\left(\exp\left(-\frac{\|X_{(t-p):(t-1)}\|_2^2}{\tau^2}\right)\prod_{r=1}^p\prod_{i=1}^d X_{i,t-r}^{n_{i,r}}\right)^2\right]\\&=\binom{k}{\boldsymbol{n}}\mathbb{E}\left(\prod_{r=1}^p\prod_{i=1}^d\exp\left(-\frac{2X_{i,t-r}^2}{\tau^2}\right)X_{i,t-r}^{2n_{i,r}}\right)\\&\leqslant\binom{k}{\boldsymbol{n}}\prod_{r=1}^p\prod_{i=1}^d\sup_{\|\alpha\|\leqslant\|g\|_\infty}\mathbb{E}\left(\exp\left(-\frac{2Z_{\alpha,i,r}^2}{\tau^2}\right)Z_{\alpha,i,r}^{2n_{i,r}}\right) &(6.9)\\&\leqslant b\binom{k}{\boldsymbol{n}}\prod_{r=1}^p\prod_{i=1}^d\left(\frac{\sigma_i^2\tau^2}{4\sigma_i^2+\tau^2}\right)^{n_{i,r}}2^{n_{i,r}}(n_{i,r}+1)! &(6.10)\\&\leqslant b\binom{k}{\boldsymbol{n}}\left(\frac{\sigma^2\tau^2}{4\sigma^2+\tau^2}\right)^k 2^k\prod_{r=1}^p\prod_{i=1}^d(n_{i,r}+1)!, &(6.11)\end{aligned}$$

where (6.9) is proved below and (6.10) follows from (6.8). For (6.9), we recursively apply the following estimate:

$$\begin{aligned}&\mathbb{E}\left(\prod_{r=1}^p\prod_{i=1}^d\exp\left(-\frac{2X_{i,t-r}^2}{\tau^2}\right)X_{i,t-r}^{2n_{i,r}}\right)\\&=\mathbb{E}\left(\prod_{r=2}^p\prod_{i=1}^d\exp\left(-\frac{2X_{i,t-r}^2}{\tau^2}\right)X_{i,t-r}^{2n_{i,r}}\mathbb{E}\left[\prod_{i=1}^d\exp\left(-\frac{2X_{i,t-1}^2}{\tau^2}\right)X_{i,t-1}^{2n_{i,1}}\mid X_{(t-p-1):(t-2)}\right]\right)\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\left(\prod_{r=2}^{p}\prod_{i=1}^{d}\exp\left(-\frac{2X_{i,t-r}^2}{\tau^2}\right)X_{i,t-r}^{2n_{i,r}}\prod_{i=1}^{d}\mathbb{E}\left[\exp\left(-\frac{2X_{i,t-1}^2}{\tau^2}\right)X_{i,t-1}^{2n_{i,1}}\mid X_{(t-p-1):(t-2)}\right]\right)\\
&\leqslant \mathbb{E}\left(\prod_{r=2}^{p}\prod_{i=1}^{d}\exp\left(-\frac{2X_{i,t-r}^2}{\tau^2}\right)X_{i,t-r}^{2n_{i,r}}\right)\prod_{i=1}^{d}\sup_{\|\alpha\|\leqslant\|g\|_\infty}\mathbb{E}\left(\exp\left(-\frac{2Z_{\alpha,i,t-1}^2}{\tau^2}\right)Z_{\alpha,i,t-1}^{2n_{i,1}}\right),
\end{aligned}
$$

where the first identity uses the law of total expectation and the second identity uses that $\Sigma = \operatorname{diag}(\sigma_1^2,\dots,\sigma_d^2)$ is a diagonal matrix. For the inequality in the final line, note that $X_{t-1} \mid X_{(t-p-1):(t-2)} \sim \mathcal{N}(g(X_{t-2},\dots,X_{t-2-p}),\Sigma)$. To further bound (6.11), note that $n_{i,r} \leqslant k$, so

$$
\binom{k}{\boldsymbol{n}}\prod_{r=1}^{p}\prod_{i=1}^{d}(n_{i,r}+1)! = \binom{k}{\boldsymbol{n}}\prod_{r=1}^{p}\prod_{i=1}^{d}(n_{i,r}+1)n_{i,r}! = k!\prod_{r=1}^{p}\prod_{i=1}^{d}(n_{i,r}+1) \leqslant k!(k+1)^{pd} \tag{6.12}
$$

Combining (6.11) and (6.12), we see that

$$
\mathbb{E}((\phi_{i,j,k}(X_{(t-p):(t-1)}))^2) \leqslant b\left(\frac{2\sigma^2\tau^2}{4\sigma^2+\tau^2}\right)^k (k+1)^{pd}k!,
$$

as claimed. ■

Lemma 6.5. *There are $b_3 \in (0,\infty)$ and*

$$
\rho_0 \in \left(\frac{4\sigma^2}{4\sigma^2+\tau^2}, 1\right), \tag{6.13}
$$

such that Assumption K.4 holds for the Gaussian kernel.

Proof: Letting b be as in the statement of Lemma 6.4 and fixing ρ_0 as in (6.13), we can find $k_0 \in \mathbb{N}$ such that

$$
\left(\frac{4\sigma^2\tau^2}{4\sigma^2+\tau^2}\right)^k \frac{((k+1)(k+dp-1))^{dp}}{(dp-1)!} \leqslant \rho_0^k, \quad k \geqslant k_0. \tag{6.14}
$$

Then, for each $k \geqslant k_0$,

$$
\begin{aligned}
\lambda_{i,k}\sum_{j=1}^{N(k)}\mathbb{E}((\phi_{i,j,k}(X_{(t-p):(t-1)}))^2) &\leqslant b\frac{1}{k!}\left(\frac{2}{\tau^2}\right)^k\sum_{j=1}^{N(k)}\left(\frac{2\sigma^2\tau^2}{4\sigma^2+\tau^2}\right)^k k!(k+1)^{pd}\\
&= b\sum_{j=1}^{N(k)}\left(\frac{4\sigma^2}{4\sigma^2+\tau^2}\right)^k (k+1)^{pd}\\
&\leqslant b\left(\frac{4\sigma^2}{4\sigma^2+\tau^2}\right)^k \frac{((k+1)(k+dp-1))^{dp}}{(dp-1)!} \leqslant b\rho_0^k,
\end{aligned}
$$

where the first inequality uses (6.3) and Lemma 6.4, the second inequality uses the fact that

$$
N(k) = \binom{k+dp-1}{dp-1} \leqslant \frac{(k+dp-1)^{dp}}{(dp-1)!},
$$

and the third inequality uses (6.14). It follows that there is some $b_3 \in (0,\infty)$ such that (2.23) holds for all $k \in \mathbb{N}$. ■

Lemma 6.6 establishes an upper bound for a quantity that arises when computing (suitably rescaled) suprema of the Gaussian kernel's eigenfunctions, and is used in the proof of Lemma 6.7.

Lemma 6.6. *Recall $N(k)$ from (2.19). Then, for each $k \in \mathbb{N}$,*

$$\sum\nolimits_{j=1}^{N(k)} \prod\nolimits_{\{(i,r)\in\{1,\ldots,d\}\times\{1,\ldots,p\}\ |\ n_{i,r}>0\}} n_{k,j,i,r}^{-1/2} \leqslant \pi^{\frac{dp}{2}} k^{\frac{dp}{2}-1}. \tag{6.15}$$

Proof: For notational convenience, we assume that $p = 1$, we enumerate the elements of $\mathcal{N}(k)$ in (2.20) by $\{\boldsymbol{n}_1, \ldots, \boldsymbol{n}_{N(k)}\}$, and we write $\boldsymbol{n}_j = (n_{j,1}, \ldots, n_{j,d})$, $j = 1, \ldots, N(k)$.

We employ an induction principle across d. As base case, we show (6.15) for $d = 2$ (it holds trivially when $d = 1$). First, note that the map $u \mapsto (k-u)^{-1/2}u^{-1/2}$ is decreasing on the interval $[0, k/2]$ and is increasing on the interval $[k/2, k]$, so

$$\begin{aligned}
\sum_{j=1}^{N(k)} \prod_{i\in\{1,2\ |\ n_{j,i}>0\}} n_{j,i}^{-1/2} &= \sum_{j=1}^{k-1} \frac{1}{\sqrt{(k-j)j}} \sum_{j=1}^{\lfloor k/2\rfloor} \frac{1}{\sqrt{(k-j)j}} + \sum_{j=\lfloor k/2\rfloor+1}^{k-1} \frac{1}{\sqrt{(k-j)j}} \\
&\leqslant \int_0^{\lfloor k/2\rfloor} \frac{1}{\sqrt{(k-x)x}} dx + \int_{\lfloor k/2\rfloor+1}^{k} \frac{1}{\sqrt{(k-x)x}} dx \\
&\leqslant \int_0^{k} \frac{1}{\sqrt{(k-x)x}} dx = \pi,
\end{aligned}$$

which shows that (6.15) holds for $d = 2$. Now, suppose that (6.15) holds for some $d \geqslant 2$. Using the inductive hypothesis, we see that

$$\begin{aligned}
\sum_{n_1+\cdots+n_{d+1}=k} \prod_{i=1}^{d+1} n_i^{-1/2} &= \sum_{n_{d+1}=1}^{k-d} \sum_{n_1+\cdots+n_d=k-n_{d+1}} n_{d+1}^{-1/2} \prod_{i=1}^{d} n_i^{-1/2} \\
&= \sum_{n_{d+1}=1}^{k-d} n_{d+1}^{-1/2} \left(\sum_{n_1+\cdots+n_d=k-n_{d+1}} \prod_{i=1}^{d} n_i^{-1/2} \right) \\
&\leqslant \pi^{d/2} \sum_{n_{d+1}=1}^{k-d} n_{d+1}^{-1/2} (k-n_{d+1})^{d/2-1}.
\end{aligned} \tag{6.16}$$

Using the fact that, since $d \geqslant 2$, the map $u \mapsto u^{-1/2}(k-u)^{d/2-1}$ is decreasing on $[0, k]$, we have

$$\begin{aligned}
\sum_{n_{d+1}=1}^{k-d} n_{d+1}^{-1/2} (k-n_{d+1})^{d/2-1} &\leqslant \int_0^k x^{-1/2}(k-x)^{d/2-1} dx \\
&= k^{d/2-1/2} \int_0^1 u^{-1/2}(1-u)^{d/2-1} du \\
&= k^{\frac{d+1}{2}-1} \mathrm{B}(1/2, d/2),
\end{aligned} \tag{6.17}$$

where the third line uses an integral representation of the beta function. Noting that Γ is monotonically increasing, we obtain

$$\mathrm{B}(1/2, d/2) = \frac{\Gamma(1/2)\Gamma(d/2)}{\Gamma(d/2+1/2)} \leqslant \Gamma(1/2) = \sqrt{\pi}.$$

The result follows on combining the estimate in the previous display with (6.16) and (6.17). ∎

Lemma 6.7 provides upper bounds for the suprema of (suitably normalized) eigenfunctions of the Gaussian kernel.

Lemma 6.7. *The Gaussian kernel satisfies Assumption K.3 with*

$$b_1 \doteq \pi^{\frac{dp}{2}}, \quad b_2 \doteq \frac{dp}{2} - 1.$$

Proof: As in the proof of Lemma 6.4, since i and k are fixed, we write $\boldsymbol{n}$ in place of $\boldsymbol{n}_{k,j}$, where $\boldsymbol{n} \doteq (n_{1,1}, \ldots, n_{1,p}, \ldots, n_{d,1}, \ldots, n_{d,p})$, in place of the notation used in (6.1). We first observe that $\phi^2_{i,j,k}$ is maximized at

$$\widetilde{x}(j) = (\widetilde{x}_{1,1}(j), \ldots, \widetilde{x}_{1,p}(j), \ldots, \widetilde{x}_{d,1}(j), \widetilde{x}_{d,p}(j)),$$

where $\widetilde{x}_{i,r}(j) = \left(\frac{\tau^2 n_{i,r}}{2}\right)^{1/2}$. Using Stirling's approximation, we see that

$$\begin{aligned}
\sum_{j=1}^{N(k)} \beta^2_{i,j,k} &= \sum_{j=1}^{N(k)} \sup_{x \in \mathcal{X}^p} |\lambda_{i,k} \phi^2_{i,j,k}(x)| = \sum_{j=1}^{N(k)} \frac{1}{k!} \left(\frac{2}{\tau^2}\right)^k \sup_{x \in \mathcal{X}^p} \left(\exp\left(-\frac{2\|x\|^2}{\tau^2}\right) \binom{k}{\boldsymbol{n}} x^{2\boldsymbol{n}} \right) \\
&= \sum_{j=1}^{N(k)} \frac{1}{k!} \left(\frac{2}{\tau^2}\right)^k \exp\left(-\frac{2}{\tau^2} \sum_{i=1}^{d} \sum_{r=1}^{p} (\widetilde{x}_{i,r}(j))^2\right) \frac{k!}{\boldsymbol{n}!} \prod_{i=1}^{d} \prod_{r=1}^{p} \left(\left(\frac{\tau^2 n_{i,r}}{2}\right)^{1/2} \right)^{2n_{i,r}} \\
&= \sum_{j=1}^{N(k)} \exp(-k) \prod_{i=1}^{d} \prod_{r=1}^{p} \frac{n_{i,r}^{n_{i,r}}}{n_{i,r}!} = \sum_{j=1}^{N(k)} \exp(-k) \prod_{\{(i,r) \in \{1,\ldots,d\} \times \{1,\ldots,p\} \mid n_{i,r} > 0\}} \frac{n_{i,r}^{n_{i,r}}}{n_{i,r}!} \\
&\leqslant \sum_{j=1}^{N(k)} \exp(-k) \prod_{\{(i,r) \in \{1,\ldots,d\} \times \{1,\ldots,p\} \mid n_{i,r} > 0\}} n_{i,r}^{n_{i,r}} n_{i,r}^{-1/2} \left(\frac{e}{n_{i,r}}\right)^{n_{i,r}} \\
&= \sum_{j=1}^{N(k)} \exp(-k) \prod_{\{(i,r) \in \{1,\ldots,d\} \times \{1,\ldots,p\} \mid n_{i,r} > 0\}} n_{i,r}^{-1/2} \exp(n_{i,r}) \\
&= \sum_{j=1}^{N(k)} \prod_{\{(i,r) \in \{1,\ldots,d\} \times \{1,\ldots,p\} \mid n_{i,r} > 0\}} n_{i,r}^{-1/2}.
\end{aligned}$$

The result follows from the previous display and Lemma 6.6. ∎

References

Adamczak, R. and Bednorz, W. Exponential concentration inequalities for additive functionals of Markov chains. *ESAIM: Probability and Statistics*, 19:440–481, 2015.

Adamczak, R. and Wolff, P. Concentration inequalities for non-Lipschitz functions with bounded derivatives of higher order. *Probability Theory and Related Fields*, 162(3):531–586, 2015.

Alquier, P., Doukhan, P., and Fan, X. Exponential inequalities for nonstationary Markov chains. *Dependence Modeling*, 7(1):150–168, 2019.

Alvarez, M. A., Rosasco, L., and Lawrence, N. D. Kernels for vector-valued functions: A review. *Foundations and Trends® in Machine Learning*, 4(3):195–266, 2012.

Arcones, M. A. A Bernstein-type inequality for U-statistics and U-processes. *Statistics & Probability Letters*, 22(3):239–247, 1995.

Arcones, M. A. and Gine, E. Limit theorems for U-processes. *The Annals of Probability*, 21(3):1494–1542, 1993.

Ballarin, G. Ridge regularized estimation of VAR models for inference. *Journal of Time Series Analysis*, (Special Issue):1–23, 2024.

Basu, S. and Michailidis, G. Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43(4):1535–1567, 2015.

Benrhmach, G., Namir, K., Namir, A., and Bouyaghroumni, J. Nonlinear autoregressive neural network and extended kalman filters for prediction of financial time series. *Journal of Applied Mathematics*, 2020(1): 5057801, 2020.

Borisov, I. and Volodko, N. A note on exponential inequalities for the distribution tails of canonical von Mises' statistics of dependent observations. *Statistics & Probability Letters*, 96(C):287–291, 2015.

Bosq, D. *Linear processes in function spaces: theory and applications*, volume 149. Springer Science & Business Media, 2000.

Carrington, A. M., Fieguth, P. W., and Chen, H. H. A new Mercer sigmoid kernel for clinical data classification. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 6397–6401. IEEE, 2014.

Chakrabortty, A. and Kuchibhotla, A. K. Tail bounds for canonical U-statistics and U-processes with unbounded kernels. Technical report, Working paper, Wharton School, University of Pennsylvania, 2018.

Chen, L. and Wu, W. B. Concentration inequalities for empirical processes of linear time series. *Journal of Machine Learning Research*, 18:231–1, 2017.

Cline, D. B. and Pu, H.-M. H. Geometric ergodicity of nonlinear time series. *Statistica Sinica*, 9:1103–1118, 1999.

Dedecker, J. and Gouëzel, S. Subgaussian concentration inequalities for geometrically ergodic Markov chains. *Electronic Communications in Probability*, 20(64):1–12, 2015.

Duchemin, Q., De Castro, Y., and Lacour, C. Concentration inequality for U-statistics of order two for uniformly ergodic Markov chains. *Bernoulli*, 29(2):929–956, 2023.

Fan, J., Jiang, B., and Sun, Q. Hoeffding's inequality for general Markov chains and its applications to statistical learning. *Journal of Machine Learning Research*, 22(139):1–35, 2021.

Fang, Y., Loparo, K. A., and Feng, X. Inequalities for the trace of matrix product. *IEEE Transactions on Automatic Control*, 39(12):2489–2490, 1994.

Han, F. An exponential inequality for U-statistics under mixing conditions. *Journal of Theoretical Probability*, 31(1):556–578, 2018.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Joshi, C. and Bissu, S. Inequalities for some special functions. *Journal of Computational and Applied Mathematics*, 69(2):251–259, 1996.

Karvonen, T., Cockayne, J., Tronarp, F., and Särkkä, S. A probabilistic Taylor expansion with Gaussian processes. *Transactions on Machine Learning Research*, 2023(8), 2023.

Kato, H., Taniguchi, M., and Honda, M. Statistical analysis for multiplicatively modulated nonlinear autoregressive model and its applications to electrophysiological signal analysis in humans. *IEEE Transactions on Signal Processing*, 54(9):3414–3425, 2006.

Liu, Z. and Li, M. On the estimation of derivatives using plug-in kernel ridge regression estimators. *Journal of Machine Learning Research*, 24(266):1–37, 2023.

Loh, P.-L. and Wainwright, M. J. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664, 2012.

Lu, Z. and Jiang, Z. L1 geometric ergodicity of a multivariate nonlinear AR model with an ARCH term. *Statistics & Probability Letters*, 51(2):121–130, 2001.

Luke, Y. L. Inequalities for generalized hypergeometric functions. *Journal of Approximation Theory*, 5(1): 41–65, 1972.

Meyn, S. P. and Tweedie, R. L. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

Nyberg, H. Forecasting US interest rates and business cycle with a nonlinear regime switching VAR model. *Journal of Forecasting*, 37(1):1–15, 2018.

Paulin, D. Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electronic Journal of Probability*, 20(79):1–32, 2015.

Pitcan, Y. A note on concentration inequalities for U-statistics. *arXiv preprint arXiv:1712.06160*, 2017.

Rudelson, M. and Vershynin, R. Hanson-Wright inequality and sub-Gaussian concentration. *Electronic Communications in Probability*, 18:1–9, 2013.

Schölkopf, B., Herbrich, R., and Smola, A. J. A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer, 2001.

Sharma, S., Sharma, S., and Athaiya, A. Activation functions in neural networks. *Towards Data Sci*, 6(12): 310–316, 2017.

Shen, Y., Han, F., and Witten, D. Exponential inequalities for dependent V-statistics via random Fourier features. *Electronic Journal Probability*, 25(7):1–18, 2020.

Smale, S. and Zhou, D.-X. Shannon sampling II: Connections to learning theory. *Applied and Computational Harmonic Analysis*, 19(3):285–302, 2005.

Steinwart, I. and Christmann, A. *Support Vector Machines*. Information Science and Statistics. Springer New York, 2008.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Wahba, G. Spline models for observational data. In *CBMS-NSF Regional Conference Series in Applied Mathematics*, volume 59, 1990.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Winkelbauer, A. Moments and absolute moments of the normal distribution. *arXiv preprint arXiv:1209.4340*, 2012.

Yu, P.-N., Liu, C. Y., Heck, C. N., Berger, T. W., and Song, D. A sparse multiscale nonlinear autoregressive model for seizure prediction. *Journal of Neural Engineering*, 18(2):026012, 2021.

Yuan, M. and Zhou, D.-X. Minimax optimal rates of estimation in high dimensional additive models. *The Annals of Statistics*, 44(6):2564–2593, 2016.

Zhou, H. H. and Raskutti, G. Non-parametric sparse additive auto-regressive network models. *IEEE Transactions on Information Theory*, 65(3):1473–1492, 2018.