Convergence Rates for Nearest Neighbour Regression in Infinite Dimensions

Peter Koepernik[1]

1. Department of Statistics, University of Oxford (peter.koepernik@stcatz.ox.ac.uk)

Nearest neighbour methods are one of the most widely used regression techniques due to the simplicity of their implementation and their wide applicability. They are one of the few approaches viable for regression on non-Euclidean manifolds and infinite dimensional function spaces, which increasingly come up in applications in engineering, data science, and other fields. However, currently, proofs of consistency and rates of convergence are only available for Euclidean domains. In this paper, we prove that nearest neighbour regression on general metric spaces—which includes general manifolds and function spaces—is, under minimal assumptions, universally uniformly L^2-consistent, and present convergence rates in terms of small ball probabilities of the regressor. We use our general framework to derive explicit convergence rates for cases where the regressor is finite-dimensional (in the Hausdorff sense) or a Gaussian random function.

Keywords: Regression, nearest neighbour rules, metric spaces, nonparametric estimation, functional data.

MSC2020: Primary 62G05, secondary 62G30.

Edited by:
Karl Welzel

Reviewed by:
Julian Clausnitzer
Evan Quiney

Copy-edited by:
Arunima Cheruvathoor
Sydelle de Souza

Received: 21 October 2021
Published: 1 March 2023

Cite as: Peter Koepernik. 'Convergence Rates for Nearest Neighbour Regression in Infinite Dimensions'. In: *St. Catherine's Academic Review* 1 (2023), pp. 45–63

1 Introduction

Consider a task where a random quantity Y is to be predicted from a related quantity X, based on a set of training samples $(x_1, y_1), \ldots, (x_n, y_n)$. One approach called the k-nearest neighbour technique takes the average over k of the training samples which are closest to a given observation $X = x$. As an example, consider a streaming platform that collects user ratings for films they have watched in order to predict which other films they are likely to enjoy. The platform needs a way of predicting the rating (in our notation, Y) that a user would give to a new film X based on previous ratings $(x_1, y_1), \ldots, (x_n, y_n)$. The k-nearest neighbour technique in this setting would take the average of the ratings that the user assigned to the k previously watched films that are closest to the new one based on genre, age rating, year of production, etc. In reality, the user's rating will also depend on external factors such as current mood, whether they watch the film together with a group, or the quality of available snacks. The easiest way to model this is as random fluctuations. Hence Y is assumed to be random rather than a fixed function of X. A good estimation procedure would then be expected to predict a user's rating of a film averaged over external factors. If this prediction becomes more accurate with an increasing amount of training data, the estimator is called *consistent*. There are various notions that make this idea of consistency mathematically precise—see [2] for a comprehensive list.

A vital part of theoretical research on nearest neighbour regression is proving that estimation is guaranteed to be consistent under certain assumptions. Preferably, such a result also makes assertions about the speed at which the estimator converges to the true function, called the *rate of convergence*. There are many results available in this direction, but they are all concerned with the case where the domain of X is a subset of the

Euclidean space $\mathbb{R}^d$ [6, 9, 8, 17, 18], or where a more general state space is projected on $\mathbb{R}^d$ by extraction of a finite number of features [13]. Many classical applications can be modelled in such a way, including our earlier example wherein each film could be represented by a vector of features (genre, age rating, production year, etc.). However, many recent applications do not fit this assumption. Fuchs et al. [11] study applications of nearest neighbour techniques on functional data in speech recognition and sensor technologies, in which case the domain of X is an infinite-dimensional space of functions. Lang et al. [14] use regression in the context of 6D object tracking. They use quaternions to describe rotations, which reside on a so-called *manifold*—a space that "locally looks Euclidean" but generally cannot be identified with a subset of $\mathbb{R}^d$. The authors use Gaussian process regression, but nearest neighbour techniques could be applied in the same context.

The most general setting in which nearest neighbour techniques can be applied would be the case where the domain of X is a set with a notion of distance between objects called a *metric space*. More formally, a metric on a set E is a symmetric function $\rho\colon E \times E \to [0, \infty)$, such that $\rho(x, y) = 0$ if and only if $x = y$, and $\rho(x, y) \leq \rho(x, z) + \rho(z, y)$ for all $x, y, z \in E$. This includes $\mathbb{R}^d$ with the Euclidean distance and general manifolds. Function spaces can also be equipped with metrics. Several have been suggested and studied by Fuchs et al. [11] in the context of nearest neighbour regression.[1]

Nearest neighbour *classification* in this setting, that is, the case where Y takes values in $\{0, 1\}$, has been studied in [4, 5, 7]. In this paper, we study *regression*, which is the more general case where Y is real-valued, and establish consistency of nearest neighbour estimation under minimal assumptions. We give a general result that applies to arbitrary metric spaces and show how these results can be used to obtain explicit convergence rates in the case where X is finite-dimensional and in the infinite-dimensional case where X is a Gaussian random function. To our knowledge, this makes the present paper the first to formally prove consistency of nearest neighbour regression in the most general setting of arbitrary metric spaces.

We give a more precise summary of the type of consistency we show. To put our result in context with existing literature, we introduce some important distinctions with regard to the sense in which the estimator converges to the correct function—convergence can hold *pointwise*, i.e. individually on every point of the domain, versus *uniformly* on certain subsets of it; and the convergence can be *weak* (convergence in probability), *strong* (almost sure convergence), or in the L^p-sense, which means that the difference of the estimator and the target function vanishes in the probabilistic L^p-norm. Both strong and L^p-consistency imply weak consistency, but neither implies the other. In the case of Euclidean domains, results on pointwise L^2-consistency were first presented in [17], uniform L^p-consistency is given in [18] and [8]. Uniform strong consistency was established in [9] and [6]. In our setting, where X takes values in a general metric space, we prove uniform L^2-consistency, with convergence rates in terms of small ball probabilities of X, under mild continuity assumptions and sufficient integrability of Y, with no assumptions on the metric domain space or the distribution of X.

In Section 2, we introduce the general framework of the paper, recall some basic probabilistic notions, and define the estimators. In Section 3, we present our main results, which we apply in Section 4 to derive convergence rates in the cases where X is finite-dimensional or a Gaussian random function. The proofs for these results can be found in Section 5.

[1] They also consider *semi-metrics*, that is functions $\rho\colon X \times X \to [0, \infty)$ which are metrics except for the fact that there may be $x_1, x_2 \in X, x_1 \neq x_2$ with $\rho(x_1, x_2) = 0$. Note that in such cases any nearest neighbour estimator will, in general, not be consistent: If $\rho(x_1, x_2) = 0$, then the estimator cannot distinguish between observations at x_1 and x_2, hence will always estimate $\widehat{f}(x_1) = \widehat{f}(x_2)$, which is wrong if the estimated function satisfies $f(x_1) \neq f(x_2)$. Semi-metrics generally perform worse than true metrics for this reason but are often more computationally efficient. To trade off performance and efficiency, one often tries to find a semi-metric which is as simple as possible while still being expected to satisfy $f(x_1) \approx f(x_2)$ whenever $\rho(x_1, x_2) = 0$.

2 Preliminaries and Setup

We first introduce some notation. If $f, g\colon \mathbb{N} \to [0, \infty)$ are two functions, we write $f = O(g)$ if there exists $K > 0$ with $f \leq Kg$, and we write $f = o(g)$ or $f \ll g$ if there is $\alpha_n \downarrow 0$ with $f(n) = \alpha_n g(n)$ for all $n \in \mathbb{N}$. We write $f \sim g$ if there is $\alpha_n \to 1$ with $f(n) = \alpha_n g(n)$ for all $n \in \mathbb{N}$.

Let (X, Y) be a pair of random variables such that Y is real-valued, and X takes values in a metric space (E, ρ). Denote the Borel σ-algebras on E and $\mathbb{R}$ by $\mathcal{E}$ and $\mathcal{B}$ respectively, and the underlying probability space by $(\Omega, \mathcal{A}, \mathbb{P})$. Assume that $\|Y\|_{L^2} < \infty$, where $\|\cdot\|_{L^p} \coloneqq \mathbb{E}\left[|\cdot|^p\right]^{1/p}$ for $p \geq 1$.

2.1 Bayes Estimator

We recall some basic probabilistic notions, details on all of which can be found in [1]. If $Z\colon \Omega \to \mathbb{R}$ is an integrable random variable, then the conditional expectation $\mathbb{E}\left[Z \mid X\right]$ of Z given X is the unique (up to modification on a $\mathbb{P}$-null set) integrable, $X^{-1}(\mathcal{E})$-measurable random variable for which $\mathbb{E}\left[\mathbb{E}\left[Z \mid X\right] \mathbb{1}_{\{X \in A\}}\right] = \mathbb{E}\left[Z \mathbb{1}_{\{X \in A\}}\right]$ for all $A \in \mathcal{E}$. Let $(\mathbb{P}^{Y|X=x})_{x \in E}$ be a conditional distribution of Y given X. That is, $\mathbb{P}^{Y|X=x}(\cdot)\colon \mathcal{B} \to [0, 1]$ is a probability measure on $\mathbb{R}$ for every $x \in E$, and $x \mapsto \mathbb{P}^{Y|X=x}(B)$ is measurable for every $B \in \mathcal{B}$, and

$$\mathbb{P}(X \in A, Y \in B) = \int_A \mathbb{P}^{Y|X=x}(B)\mathbb{P}^X(\mathrm{d}x),$$

for any $A \in \mathcal{E}$ and $B \in \mathcal{B}$, where $\mathbb{P}^X \coloneqq \mathbb{P}(X^{-1}(\cdot))\colon \mathcal{E} \to [0, 1]$ denotes the law of X. Then, for any measurable $g\colon \mathbb{R} \to \mathbb{R}$, let

$$\mathbb{E}\left[g(Y) \mid X = x\right] \coloneqq \int_{\mathbb{R}} g(y)\mathbb{P}^{Y|X=x}(\mathrm{d}y).$$

Denoting the left-hand side above by $f(x)$, it is elementary to confirm that $f(X) = \mathbb{E}\left[g(Y) \mid X\right]$. Other characteristics of Y, such as its variance, can also be expressed conditional on $X = x$ by defining them in terms of $\mathbb{P}^{Y|X=x}$ instead of $\mathbb{P}^Y$.

For $x \in E$ and $k \in \mathbb{N}$, if $\|Y\|_{L^k} < \infty$, define

$$m_k(x) \coloneqq \mathbb{E}\left[Y^k \,\middle|\, X = x\right], \tag{1}$$

and put $m \coloneqq m_1$. Then, $m(X) = \mathbb{E}\left[Y \mid X\right]$, and $m\colon E \to \mathbb{R}$ is the so-called Bayes estimator of Y given X, which minimises the mean squared error in the sense that, for any measurable $\widetilde{m}\colon E \to \mathbb{R}$,

$$\mathbb{E}\left[|m(x) - Y|^2 \,\middle|\, X = x\right] \leq \mathbb{E}\left[|\widetilde{m}(x) - Y|^2 \,\middle|\, X = x\right]$$

holds for $\mathbb{P}^X$-almost all $x \in E$. Even though the Bayes estimator is optimal in the above sense, it is by no means perfect. If we put $\varepsilon \coloneqq Y - m(X)$, then

$$Y = m(X) + \varepsilon,$$

where ε is a noise with $\mathbb{E}\left[\varepsilon \mid X\right] = 0$ that describes the fluctuations of Y around $m(X)$. A measure for the strength of these fluctuations is the variance of the noise,

$$v(x) \coloneqq \mathbb{V}\left(\varepsilon \mid X = x\right) = \mathbb{V}\left(Y \mid X = x\right) = m_2(x) - m(x)^2.$$

By estimating v alongside m, one obtains a sense of the uncertainty associated with the estimation.

2.2 Nearest Neighbour Estimation

Suppose that a sequence (X_i, Y_i), $i \in \mathbb{N}$, of independent copies of (X, Y) is defined on $(\Omega, \mathcal{A}, \mathbb{P})$. Let $k_n \in \{1, \dots, n\}$, $n \in \mathbb{N}$, be a non-decreasing sequence with $k_n \to \infty$ and $k_n/n \to 0$. We now define the k_n-nearest neighbour estimators. For fixed $n \in \mathbb{N}$ and $x \in E$, let $(\sigma_1, \dots, \sigma_n)$ be the (random) permutation of $(1, \dots, n)$ such that

$$\rho(X_{\sigma_1}, x) \leq \rho(X_{\sigma_2}, x) \leq \ldots \leq \rho(X_{\sigma_n}, x),$$

where, if several X_i have the same distance from x, they are ordered uniformly at random, independently of all other quantities and choices. This way, $(\sigma_1, \dots, \sigma_n)$ is any of the $n!$ permutations with equal probability. Denote the inverse permutation by $(\Sigma_1, \dots, \Sigma_n)$, that is, $\Sigma_i = \sum_{j=1}^n j \mathbb{1}_{\{\sigma_j = i\}}$ is the position of X_i in the ordered tuple $(X_{\sigma_1}, \dots, X_{\sigma_n})$. Then, in particular,

$$\mathbb{P}\left(\Sigma_i \leq k\right) = \frac{k}{n}, \qquad \mathbb{P}\left(\Sigma_i \vee \Sigma_j \leq k\right) = \frac{k(k-1)}{n(n-1)}, \tag{2}$$

whenever $i, j, k \in \{1, \dots, n\}$, $i \neq j$, where $x \wedge y \coloneqq \min(x, y)$ and $x \vee y \coloneqq \max(x, y)$ for $x, y \in \mathbb{R}$. Note that σ_i and Σ_i depend on n and x, so strictly speaking we should specify $\sigma_{(i,n)}(x)$ and $\Sigma_{(i,n)}(x)$, but we omit either or both if they are clear from the context. For $n \in \mathbb{N}$ and $k \in \mathbb{N}$, the k_n-nearest neighbour estimators of m_k and v based on the first n observations are defined by

$$\widehat{m}_k^{(n)}(x) \coloneqq \frac{1}{k_n} \sum_{i=1}^{k_n} Y^k_{\sigma_{(i,n)}(x)} = \frac{1}{k_n} \sum_{i=1}^{n} Y_i^k \mathbb{1}_{\left\{\Sigma_{(i,n)}(x) \leq k_n\right\}}, \tag{3}$$

$$\widehat{v}^{(n)}(x) \coloneqq \frac{1}{k_n} \sum_{i=1}^{k_n} \left(Y_{\sigma_i} - \widehat{m}^{(n)}(x)\right)^2 = \widehat{m}_2^{(n)}(x) - \widehat{m}^{(n)}(x)^2, \tag{4}$$

for $x \in E$, where $\widehat{m}^{(n)} \coloneqq \widehat{m}_1^{(n)}$.

2.3 Small Ball Probabilities

We introduce some notation and elementary facts regarding small ball probabilities of X. For $x \in E$, $\varepsilon, \delta > 0$, define

$$p_x(\delta) \coloneqq \mathbb{P}(X \in \overline{B}(x, \delta)), \tag{5}$$

$$p_x^{-1}(\varepsilon) \coloneqq \inf\left\{\delta \geq 0 \colon p_x(\delta) \geq \varepsilon\right\}, \tag{6}$$

where $\overline{B}(x, \delta) = \{y \in E \colon \rho(x, y) \leq \delta\}$. The support of X is the closed set

$$\mathrm{S}(X) \coloneqq \{x \in E \colon \forall \delta > 0 \colon p_x(\delta) > 0\}.$$

Lemma 2.1. *Let $x \in E$.*

(i) $p_x \colon [0, \infty) \to [0, 1]$ *and* $p_x^{-1} \colon [0, 1] \to [0, \infty)$ *are increasing and respectively right- and left-continuous.*

(ii) $p_x(0) = \mathbb{P}(X = x)$, $p_x^{-1}(0) = 0$.

(iii) $p_x(p_x^{-1}(\varepsilon)) \geq \varepsilon$, *for all* $\varepsilon > 0$.

(iv) If $x \in \mathrm{S}(X)$, then $p_x^{-1}(\varepsilon) \downarrow 0$ as $\varepsilon \downarrow 0$.

Proof. (i) Monotonicity is obvious and right-continuity of p_x is just continuity from above of $\mathbb{P}^X$. Let $0 < \varepsilon_n \uparrow \varepsilon_0$, and put $\delta_n := p_x^{-1}(\varepsilon_n)$, $n \in \mathbb{N}_0$. Then, $0 \leq \delta_n \uparrow \delta$ for some $0 \leq \delta \leq \delta_0$ (because $\delta_n \leq \delta_0$ for all $n \in \mathbb{N}$). Hence it suffices to show that $\delta \geq \delta_0$, that is, that $p_x(\delta) \geq \varepsilon_0$. But this is clear because $p_x(\delta) \geq p_x(\delta_n) \geq \varepsilon_n \uparrow \varepsilon_0$.

(ii) Clear from the definitions.

(iii) Follows from Eq. (6) and right-continuity of p_x.

(iv) If $x \in \mathrm{S}(X)$, then $p_x(\delta) > 0$ for all $\delta > 0$. Hence, for every $\delta > 0$ there is $\varepsilon > 0$ with $p_x(\delta) \geq \varepsilon$ and thus $p_x^{-1}(\varepsilon) \leq \delta$. □

3 Main Results

The following two theorems are our main results. Recall the definitions of the moments m_k ($k \in \mathbb{N}$) of Y given X, and of the nearest neighbour estimator $\widehat{m}_k^{(n)}$ of m_k based on the first n observations (see Eqs. (1) and (3)). Also, recall that we assume a sequence $(k_n)_{n\in\mathbb{N}}$, $k_n \in \{1, \ldots, n\}$ with $k_n \to \infty$ and $k_n/n \to 0$ to be given. We call a function $h\colon E \to \mathbb{R}$ γ-Hölder continuous at $x \in E$ for $\gamma \in (0, 1]$ if there are $c, \delta > 0$ such that $|h(y) - h(z)| \leq c\rho(y, z)^\gamma$ for all $y, z \in B(x, \delta)$.

Theorem 3.1. *Let $k \in \mathbb{N}$. Suppose there exists $p > 2k$ such that $\|Y\|_{L^p} < \infty$, that m_k is continuous, and that m_{2k} is locally bounded. In this case, the following statements hold:*

(i) If $C \subset \mathrm{S}(X)$ is compact, then $\widehat{m}_k^{(n)}(\cdot) \xrightarrow{L^2} m_k(\cdot)$ uniformly on C. That is,

$$\sup_{x\in C} \left\| \widehat{m}_k^{(n)}(x) - m_k(x) \right\|_{L^2} \to 0, \quad n \to \infty.$$

(ii) If $x \in \mathrm{S}(X)$ and m_k is γ-Hölder continuous at x for some $\gamma \in (0, 1]$, then, for any sequence $0 < \alpha_n = o(n/k_n)$, there is a $c > 0$ such that

$$\left\| \widehat{m}_k^{(n)}(x) - m_k(x) \right\|_{L^2}^2 \leq c \left(\frac{1}{k_n} + p_x^{-1} \left((1 + \alpha_n) \frac{k_n}{n} \right)^\gamma + n^2 \exp\left(-\frac{k_n}{2q} \frac{\alpha_n^2}{1 + \alpha_n} \right) \right), \tag{7}$$

where $q = \frac{p}{p-k} \in (1, 2)$. In particular,

$$\left\| \widehat{m}_k^{(n)}(x) - m_k(x) \right\|_{L^2}^2 \leq c \left(\frac{1}{k_n} + p_x^{-1} \left(\frac{2k_n}{n} \right)^\gamma + n^2 \mathrm{e}^{-k_n/8} \right).$$

Theorem 3.2. *Suppose there is $p > 4$ such that $\|Y\|_{L^p} < \infty$, and that m, m_2 are continuous and m_4 is locally bounded. Then, $\widehat{v}^{(n)}(\cdot) \xrightarrow{L^1} v(\cdot)$ as $n \to \infty$, uniformly on compact subsets of $\mathrm{S}(X)$.*

4 Examples

Before moving on to the proofs, we demonstrate how Theorem 3.1 can be used to derive explicit convergence rates in the cases where X is finite-dimensional (in the following sense), or a Gaussian random function.

Definition 4.1. Let $s > 0$. We say that X is *at most s-dimensional at $x \in E$* if there exists $c > 0$ and $\delta_0 \in (0, 1)$ such that

$$p_x(\delta) \geq c\delta^s, \quad \delta \in (0, \delta_0). \tag{8}$$

We say that X is *at most s-dimensional* if X is at most s-dimensional at every $x \in \mathrm{S}(X)$.

Remark 4.2. (i) The reason we call X 'at most' s-dimensional in Definition 4.1 is that if Eq. (8) holds for some $s > 0$, then it also holds for all $s' > s$. One might thus be tempted to define the dimension s_X of X as the infimum over all s for which Eq. (8) holds, but in this case Eq. (8) does not in general also hold for s_X.

(ii) Suppose that E has Hausdorff dimension $s > 0$, and that there are measurable sets $E_n \uparrow E$ such that the Hausdorff measure $\mathcal{H}^s : \mathcal{B}(E) \to [0, \infty]$ satisfies $\mathcal{H}^s(E_n) \in (0, \infty)$ for all $n \in \mathbb{N}$. If X has a positive, continuous density with respect to $\mathcal{H}^s$ (which in this case is a sort of uniform measure on E) then X is at most s-dimensional in the sense of Definition 4.1. This includes the case of absolutely continuous random variables X on $\mathbb{R}^d$. See [10] for details.

Theorem 4.3. *Let $k \in \mathbb{N}$, $x \in \mathrm{S}(X)$, suppose that m_k is γ-Hölder continuous at x for some $\gamma \in (0, 1]$, and that X is at most s-dimensional at x. Then, if $k_n \gg \log n$,*

$$\left\| \widehat{m}_k^{(n)}(x) - m_k(x) \right\|_{L^2}^2 = O\left(\frac{1}{k_n} + \left(\frac{k_n}{n} \right)^{\gamma/s} \right). \tag{9}$$

If k_n, $n \in \mathbb{N}$, optimises this bound, then

$$k_n \sim c n^{\frac{1}{1+s/\gamma}},$$

where $c = \left(\frac{s}{\gamma}\right)^{s/(s+\gamma)}$. In that case, and in fact whenever $k_n \sim c' n^{\frac{1}{1+s/\gamma}}$ for some $c' > 0$, then

$$\left\| \widehat{m}_k^{(n)}(x) - m_k(x) \right\|_{L^2} = O\left(n^{-\frac{1}{2+2s/\gamma}} \right). \tag{10}$$

Note that we only considered $k_n \gg \log n$ above, but we know that $k_n = O(\log n)$ cannot transform Eq. (7) into a better bound than Eq. (10), as it would be no smaller than $\frac{1}{k_n} \geq \frac{1}{C \log n} \gg n^{-1/(2+2s/\gamma)}$.

Remark 4.4. (i) The assumptions of Theorem 4.3 include as domains for X manifolds that cannot be embedded into $\mathbb{R}^d$, as well as more general fractal spaces. For example, this includes applications that regress on quaternions, as Lang et al. [14] have previously done in the context of 6D object tracking.

(ii) Theorem 4.3 includes the case where $s = d \in \mathbb{N}$ and $E = \mathbb{R}^d$. In this setting, known results assert a slightly faster convergence rate of $O(n^{-1/(2+s/\gamma)})$, albeit under stronger regularity conditions such as sufficient differentiability of m [2], or with a weaker notion of convergence [12]. Still, it is reasonable to assume that a refinement of the arguments presented here yields the aforementioned stronger rate of convergence also in this more general setting.

We now turn to the case where X is a *centred Gaussian process* on $[0, 1]$. That is, we assume that $X = (X_t)_{t \in [0,1]}$ takes values in $C([0, 1], \mathbb{R})$—which is a complete and separable normed space when equipped with the supremum norm—and that for any $n \in \mathbb{N}$ and $t_1, \ldots, t_n \in [0, 1]$, $(X_{t_1}, \ldots, X_{t_n})$ has an n-dimensional centred Gaussian distribution. In this case, X is fully determined by its *covariance function* or *kernel*

$$K(t, s) := \mathbb{E}\left[X_t X_s\right], \quad t, s \in [0, 1]. \tag{11}$$

We suppose that there is $\beta > 0$ such that for every $t, s \in [0, 1]$,

$$|K(t,s) - K(t,t)| \leq c\,|t-s|^{2\beta}\,, \tag{12}$$

for some constant $c > 0$. For example, if K is Lipschitz continuous, then Eq. (12) is satisfied with $\beta = 1/2$. One can think of β as a measure of the regularity of X.

The *reproducing kernel Hilbert space (RKHS)* or *Cameron–Martin space* $\mathcal{H}$ of X is a dense linear subspace of $\mathrm{S}(X)$ with a scalar product that turns $\mathcal{H}$ into a Hilbert space, see [3] for details.

Remark 4.5. Similar to the universality of the Gaussian distribution in the context of naturally occurring, real-valued data, Gaussian processes are commonly used to model functional data. The following are two of the most commonly used kernels:

(i) The *exponential kernel* is defined by

$$K(t,s) = \mathrm{e}^{-|t-s|/\sigma}, \quad t, s \in [0,1],$$

for some $\sigma > 0$. In this case $\beta = 1/2$ and $\mathcal{H} = C^1([0,1])$.

(ii) The *squared exponential kernel* is defined by

$$K(t,s) = \mathrm{e}^{-(t-s)^2/(2\sigma^2)}, \quad t, s \in [0,1],$$

where $\sigma > 0$. In this case $\beta = 1$, and $\mathcal{H}$ contains $C^\infty([0,1])$.

In both cases, σ determines the length scale of X. The squared exponential kernel is often used if the data is expected to be smooth, while the exponential kernel tends to work well for more ragged functions. A discussion of a variety of kernels in the context of Gaussian process regression can be found in section 4.2.1 of [16].

Theorem 4.6. *Let $k \in \mathbb{N}$, $x \in \mathcal{H}$, and suppose that m_k is γ-Hölder continuous at x for some $\gamma \in (0,1]$. Then,*

$$\left\| \widehat{m}_k^{(n)}(x) - m_k(x) \right\|_{L^2}^2 = O\left(\frac{1}{k_n} + \log\left(\frac{n}{k_n}\right)^{-\gamma\beta} \right). \tag{13}$$

If $k_n \in \mathbb{N}$ optimises this bound, then

$$k_n \sim \tfrac{1}{\gamma\beta}\,(\log n)^{1+\gamma\beta}\,.$$

In that case, and in fact whenever $(\log n)^{\gamma\beta} \ll k_n = O(n^a)$ for some $a \in (0,1)$, then

$$\left\| \widehat{m}_k^{(n)}(x) - m_k(x) \right\|_{L^2} = O\left((\log n)^{-\gamma\beta/2} \right). \tag{14}$$

Remark 4.7. (i) The convergence rate obtained in Theorem 4.6 seems rather slow at first glance, but considering that the optimal rate for nearest neighbour estimation in $\mathbb{R}^r$ is $O(n^{-1/(2+r/\gamma)})$, and that function spaces such as $C([0,1],\mathbb{R})$ are infinite-dimensional, it is not surprising that the convergence rate is slower than n^{-c} for any $c > 0$. It is also noteworthy that although the bound in Eq. (13) is optimised by a logarithmic k_n, its asymptotic form Eq. (14) remains unchanged through a wide range of values for k_n, including $k_n \sim n^a$ for any $a \in (0,1)$. A possible direction for future work could aim for lower bounds on convergence rates in the functional setting which may lead to further insights regarding the optimal choice for k_n.

(ii) These arguments can be generalised to Gaussian processes in $C([0,1]^d,\mathbb{R})$ for $d \in \mathbb{N}$. In that case, the statements of Lemma 5.9 and Theorem 4.6 remain true with β replaced by β/d. In particular, the convergence rates are improved by higher regularity of X and m_k (i.e. large β and γ), and small dimension d.

(iii) The above can readily be extended to non-centred Gaussian processes X by considering the centred process $X - m$, where $m(\cdot) = \mathbb{E}\left[X(\cdot)\right]$ is the mean function of X.

5 Proofs

Recall Section 2.2, in particular the fact that we assume a sequence $k_n \in \{1,\dots,n\}$, $n \in \mathbb{N}$, to be given that satisfies $k_n \to \infty$ and $k_n/n \to 0$. Write $\rho_i := \rho_i(x) := \rho(X_i, x)$ for $x \in E$ and $i \in \mathbb{N}$, where x is omitted if clear from context. For $n \in \mathbb{N}$ and $F \subset E$

$$N_n(F) := \left|\{i \in [n] \colon X_i \in F\}\right|, \quad N_n(x,\delta) := N_n(B(x,\delta)),$$

where $[n] = \{1,\dots,n\}$. The following two lemmas show that, for fixed $\delta > 0$ and $x \in E$, it is exponentially likely that all of the k_n closest X_i's lie inside $B(x,\delta)$.

Lemma 5.1. *Let $n \in \mathbb{N}$, and $F \subset E$. If $p_F := \mathbb{P}(X \in F) > k_n/n$, then*

$$\mathbb{P}(N_n(F) < k_n) \le \exp\left(-\frac{n}{2p_F}\left(p_F - \frac{k_n}{n}\right)^2\right).$$

If $\delta > 0$ is fixed, and either $A = B(x_0,\delta/2)$ with $x_0 \in \mathrm{S}(X)$, or $A \subset \mathrm{S}(X)$ is compact, then there exists $c > 0$ such that

$$\mathbb{P}(\exists x \in A \colon N_n(x,\delta) < k_n) = O\left(\mathrm{e}^{-cn}\right).$$

Proof. Each X_i falls into F independently with probability p_F, so $N_n(F) \sim \mathrm{Bin}(n,p_F)$. A standard Chernoff bound yields, since $np_F > k_n$,

$$\begin{aligned}
\mathbb{P}(N_n(F) < k_n) &= \mathbb{P}\left(\mathrm{Bin}(n,p_F) < np_F\left(1 - \left(1 - \frac{k_n}{np_F}\right)\right)\right)\\
&\le \exp\left(-\frac{np_F}{2}\left(1 - \frac{k_n}{np_F}\right)^2\right)\\
&= \exp\left(-\frac{n}{2p_F}\left(p_F - \frac{k_n}{n}\right)^2\right).
\end{aligned}$$

If $A = B(x_0,\delta/2)$ for some $x_0 \in \mathrm{S}(X)$, then $B(x_0,\delta/2) \subset B(x,\delta)$ for all $x \in A$, and $p' := \mathbb{P}(X \in B(x_0,\delta/2)) > 0$, so by what we have already shown,

$$\begin{aligned}
\mathbb{P}(\exists x \in A \colon N_n(x,\delta) < k_n) &\le \mathbb{P}\left(N_n(x_0,\delta/2) < k_n\right)\\
&\le \exp\left(-\frac{n}{2p'}\left(p' - \frac{k_n}{n}\right)^2\right)\\
&= O\left(\mathrm{e}^{-np'/8}\right),
\end{aligned}$$

where we used in the last step that $\frac{k_n}{n} \leq \frac{p'}{2}$ for sufficiently large $n \in \mathbb{N}$. The claim for compact $A \subset \mathrm{S}(X)$ follows because A can be covered in finitely many balls of the form $B(x, \delta/2)$ with $x \in A$. $\square$

Lemma 5.2. *If $n \in \mathbb{N}$, $x \in E$, $\delta > 0$, and $I \subset [n]$, $|I| \leq k_n$, then*

$$\mathbb{P}\left(\exists i \in I \colon \rho_i \geq \delta \,\middle|\, \bigvee_{i \in I} \Sigma_i \leq k_n\right) \leq \mathbb{P}(N_n(x,\delta) < k_n).$$

Proof. We assume $I = \{1\}$, the general proof merely requires more notation. Denote by S_n the set of permutations of $[n]$, put $A_\pi := \{(\sigma_1, \ldots, \sigma_n) = \pi\}$ for $\pi \in S_n$, and abbreviate $B := \{N_n(x,\delta) < k_n\}$. Then,

$$\begin{aligned}\mathbb{P}(B \cap A_\pi) &= \mathbb{P}\left(\text{Less than } k_n \text{ of the } (X_i)_{i=1}^n \text{ lie in } B(x,\delta), A_\pi\right) \\ &= \mathbb{P}\left(\text{Less than } k_n \text{ of the } (X_{\pi(i)})_{i=1}^n \text{ lie in } B(x,\delta), A_\pi\right) \\ &= \mathbb{P}\left(\text{Less than } k_n \text{ of the } (X_i)_{i=1}^n \text{ lie in } B(x,\delta), A_{\mathrm{id}}\right) \\ &= \mathbb{P}(B \cap A_{\mathrm{id}}),\end{aligned}$$

where $\mathrm{id} := (1, \ldots, n) \in S_n$, and we used in the third step that $(X_1, \ldots, X_n)$ and $(X_{\pi(1)}, \ldots, X_{\pi(n)})$ are equal in distribution. Hence, for any $\pi \in S_n$,

$$\mathbb{P}(B) = \sum_{\tau \in S_n} \mathbb{P}(B \cap A_\tau) = \sum_{\tau \in S_n} \mathbb{P}(B \cap A_\pi) = n!\mathbb{P}(B \cap A_\pi).$$

Now observe that $\{\Sigma_1 \leq k_n\} = \dot\bigcup_{\substack{\pi \in S_n \\ \pi(1) \leq k_n}} A_\pi$, so

$$\begin{aligned}\mathbb{P}\left(B \cap \{\Sigma_1 \leq k_n\}\right) &= \sum_{\substack{\pi \in S_n \\ \pi(1) \leq k_n}} \mathbb{P}(B \cap A_\pi) = k_n(n-1)! \cdot \frac{1}{n!}\mathbb{P}(B) = \frac{k_n}{n}\mathbb{P}(B) \\ &= \mathbb{P}(\Sigma_1 \leq k_n)\mathbb{P}(B),\end{aligned}$$

where we used Eq. (2) in the last step. Finally, note that $\{\rho_1 \geq \delta\} \cap \{\Sigma_1 \leq k_n\} \subset B \cap \{\Sigma_1 \leq k_n\}$, so

$$\mathbb{P}(B)\mathbb{P}(\Sigma_1 \leq k_n) = \mathbb{P}(B, \Sigma_1 \leq k_n) \geq \mathbb{P}(\rho_1 \geq \delta, \Sigma_1 \leq k_n).$$

Dividing by $\mathbb{P}(\Sigma_1 \leq k_n)$ finishes the proof. $\square$

Theorem 3.2 will turn out to be an immediate corollary from Theorem 3.1. The idea behind the proof of the latter is to show that $\mathbb{E}\left[\widehat{m}_k^{(n)}(\cdot)\right]$ goes to $m_k(\cdot)$ and $\mathbb{V}\left(\widehat{m}^{(n)}(\cdot)\right)$ vanishes as $n \to \infty$. We begin by proving some lemmas.

Lemma 5.3. *If $k \in \mathbb{N}$ and $\|Y\|_{L^k} < \infty$, then*

$$\mathbb{E}\left[\widehat{m}_k^{(n)}(x)\right] = \mathbb{E}\left[m_k(X_1) \,|\, \Sigma_1(x) \leq k_n\right], \quad n \in \mathbb{N}, x \in E.$$

Proof. We may assume $k = 1$ (otherwise consider $\widetilde{Y} := Y^k$). Let $n \in \mathbb{N}$ and $x \in E$. Then, by Eq. (3),

$$\mathbb{E}\left[\widehat{m}^{(n)}(x)\right] = \frac{1}{k_n}\sum_{j=1}^{n}\mathbb{E}\left[Y_j \mathbb{1}_{\{\Sigma_j(x)\le k_n\}}\right] = \frac{n}{k_n}\mathbb{E}\left[Y_1 \mathbb{1}_{\{\Sigma_1(x)\le k_n\}}\right].$$

Using the independence of all the (X_i, Y_i) and basic properties of conditional expectation,

$$\begin{aligned}\mathbb{E}\left[Y_1 \mathbb{1}_{\{\Sigma_1(x)\le k_n\}}\right] &= \mathbb{E}\left[\mathbb{E}\left[Y_1 \mathbb{1}_{\{\Sigma_1(x)\le k_n\}} \,\middle|\, X_1,\dots,X_n\right]\right]\\ &= \mathbb{E}\left[\mathbb{E}\left[Y_1 \mid X_1,\dots,X_n\right]\mathbb{1}_{\{\Sigma_1(x)\le k_n\}}\right]\\ &= \mathbb{E}\left[m(X_1)\mathbb{1}_{\{\Sigma_1(x)\le k_n\}}\right]\\ &= \frac{k_n}{n}\mathbb{E}\left[m(X_1)\mid \Sigma_1(x)\le k_n\right],\end{aligned}$$

where we used Eq. (2) in the last step. □

If $h\colon E \to \mathbb{R}$ and $A \subset E$, we write $\|h\|_A := \sup_{x\in A}|h(x)|$.

Lemma 5.4. *Let $k \in \mathbb{N}$. Suppose that there exists $p > k$ with $\|Y\|_{L^p} < \infty$, and put $q := \frac{p}{p-k}$. Let $\delta > 0$.*

(i) For $x \in E$ and $n \in \mathbb{N}$,

$$\mathbb{E}\left[|m_k(X_1) - m_k(x)|\,\mathbb{1}_{\{\rho_1\ge\delta\}} \,\middle|\, \Sigma_1 \le k_n\right] \le \left(n\|Y\|_{L^p} + |m_k(x)|\right)\mathbb{P}\left(N_n(x,\delta) < k_n\right)^{1/q},$$

(ii) For $x \in E$ and $n \in \mathbb{N}$,

$$\begin{aligned}&\mathbb{E}\left[\left|m_k(X_1)m_k(X_2) - m_k(x)^2\right|\mathbb{1}_{\{\rho_1\vee\rho_2\ge\delta\}} \,\middle|\, \Sigma_1\vee\Sigma_2 \le k_n\right]\\ &\qquad\qquad \le \left(n^2\|Y\|_{L^p}^2 + |m_k(x)|^2\right)\mathbb{P}\left(N_n(x,\delta) < k_n\right)^{1/q}.\end{aligned}$$

For fixed $\delta > 0$, if $A \subset E$ satisfies $\|m_k\|_A < \infty$ and is either a ball of radius $\delta/2$ with centre in $\mathrm{S}(X)$ or $A \subset \mathrm{S}(X)$ is compact, then both bounds vanish uniformly on A as $n \to \infty$.

Proof. We may assume that $k = 1$ (otherwise consider $\widetilde{Y} := Y^k$ and $\widetilde{p} := p/k > 1$). Fix $\delta > 0$, and let $x \in E$ and $n \in \mathbb{N}$.

(i) Abbreviate $A := \{\Sigma_1(x) \le k_n\}$, and observe that

$$\begin{aligned}&\mathbb{E}\left[|m(X_1) - m(x)|\,\mathbb{1}_{\{\rho(X_1,x)\ge\delta\}} \,\middle|\, A\right]\\ &\qquad \le \mathbb{E}\left[|m(X_1)|\,\mathbb{1}_{\{\rho(X_1,x)\ge\delta\}} \,\middle|\, A\right] + |m(x)|\underbrace{\mathbb{P}\left(\rho(X_1,x)\ge\delta \mid A\right)}_{\le \mathbb{P}(\rho(X_1,x)\ge\delta\,|\,A)^{1/q}}.\end{aligned}$$

Since $\frac{1}{q}+\frac{1}{p}=1$, Hölder's inequality gives

$$\begin{aligned}
\mathbb{E}\left[|m(X_1)|\,\mathbb{1}_{\{\rho(X_1,x)\geq\delta\}}\,\middle|\,A\right] &= \frac{\mathbb{E}\left[|m(X_1)|\,\mathbb{1}_{\{\rho(X_1,x)\geq\delta\}\cap A}\right]}{\mathbb{P}(A)} \\
&\leq \mathbb{E}\left[|m(X_1)|^p\right]^{1/p}\frac{\mathbb{P}(\rho(X_1,x)\geq\delta,\,A)^{1/q}}{\mathbb{P}(A)} \\
&= \mathbb{E}\left[|\mathbb{E}\left[Y_1\,|\,X_1\right]|^p\right]^{1/p}\mathbb{P}(A)^{1/q-1}\mathbb{P}\left(\rho(X_1,x)\geq\delta\,|\,A\right)^{1/q} \\
&\leq \mathbb{E}\left[\mathbb{E}\left[|Y_1|^p\,|\,X_1\right]\right]^{1/p}\left(\frac{n}{k_n}\right)^{1-1/q}\mathbb{P}\left(\rho(X_1,x)\geq\delta\,|\,A\right)^{1/q} \\
&\leq n\,\|Y_1\|_{L^p}\,\mathbb{P}\left(\rho(X_1,x)\geq\delta\,|\,A\right)^{1/q},
\end{aligned}$$

where we used Jensen's inequality and Eq. (2) in the penultimate step. Finally, $\|Y_1\|_{L^p}=\|Y\|_{L^p}$, and $\mathbb{P}\left(\rho(X_1,x)\geq\delta\,|\,A\right)\leq\mathbb{P}(N_n(x,\delta)<k_n)$ by Lemma 5.2.

(ii) Abbreviate $A:=\{\Sigma_1(x)\vee\Sigma_2(x)\leq k_n\}$. Then,

$$\begin{aligned}
&\mathbb{E}\left[\left|m(X_1)m(X_2)-m(x)^2\right|\mathbb{1}_{\{\rho_1\vee\rho_2\geq\delta\}}\,\middle|\,A\right] \\
&\qquad\leq \underbrace{\mathbb{E}\left[|m(X_1)m(X_2)|\,\mathbb{1}_{\{\rho_1\vee\rho_2\geq\delta\}}\,\middle|\,A\right]}_{(*)}+|m(x)|^2\underbrace{\mathbb{P}\left(\rho_1\vee\rho_2\geq\delta\,|\,A\right)}_{\leq\mathbb{P}(\rho_1\vee\rho_2\geq\delta\,|\,A)^{1/q}}.
\end{aligned}$$

Hölder's inequality gives

$$\begin{aligned}
(*) &= \frac{1}{\mathbb{P}(A)}\mathbb{E}\left[|m(X_1)m(X_2)|\,\mathbb{1}_{\{\rho_1\vee\rho_2\geq\delta\}\cap A}\right] \\
&\leq \mathbb{E}\left[|m(X_1)m(X_2)|^p\right]^{1/p}\frac{\mathbb{P}\left(\rho_1\vee\rho_2\geq\delta,A\right)^{1/q}}{\mathbb{P}(A)} \\
&= \mathbb{E}\left[|m(X)|^p\right]^{2/p}\mathbb{P}(A)^{1/q-1}\mathbb{P}\left(\rho_1\vee\rho_2\geq\delta\,|\,A\right)^{1/q} \\
&\leq \|Y\|_{L^p}^2\left(\frac{n(n-1)}{k_n(k_n-1)}\right)^{1-1/q}\mathbb{P}\left(\rho_1\vee\rho_2\geq\delta\,|\,A\right)^{1/q} \\
&\leq n^2\,\|Y\|_{L^p}^2\,\mathbb{P}\left(\rho_1\vee\rho_2\geq\delta\,|\,A\right)^{1/q},
\end{aligned}$$

where we used Eq. (2) and that X_1 and X_2 are independent and equal in distribution to X. Lemma 5.2 finishes the argument.

For fixed $\delta>0$, if $A\subset E$ satisfies $\|m\|_A<\infty$ and is either a ball of radius $\delta/2$ with centre in $\mathrm{S}(X)$ or $A\subset\mathrm{S}(X)$ is compact, then Lemma 5.1 implies that both bounds vanish uniformly on A as $n\to\infty$. □

Lemma 5.5. *Let $k\in\mathbb{N}$, and suppose there is $p>k$ with $\|Y\|_{L^p}<\infty$. Then, for any $x_0\in E$ and $\delta>0$,*

$$\varlimsup_{n\to\infty}\left(\sup_{x\in B(x_0,\delta)}\left|\mathbb{E}\left[\widehat{m}_k^{(n)}(x)\right]\right|\right)\leq\|m_k\|_{B(x_0,3\delta)}.$$

In particular, if m_k is locally bounded, then so is $\sup_{n\in\mathbb{N}}\left|\mathbb{E}\left[\widehat{m}_k^{(n)}(\cdot)\right]\right|$.

Proof. We may assume $k = 1$. Let $x_0 \in E$, $\delta > 0$, and abbreviate $A_n(\cdot) := \{\Sigma_{(1,n)}(\cdot) \le k_n\}$. If $\|m\|_{B(x_0,3\delta)} = \infty$, there is nothing to show, hence, we assume otherwise. Then, by Lemma 5.3,

$$\begin{aligned}\sup_{x\in B(x_0,\delta)} \left|\mathbb{E}\left[\widehat{m}^{(n)}(x)\right]\right| &= \sup_{x\in B(x_0,\delta)} \left|\mathbb{E}\left[m(X_1)(\mathbb{1}_{\{\rho_1<2\delta\}} + \mathbb{1}_{\{\rho_1\ge 2\delta\}}) \,\middle|\, A_n(x)\right]\right| \\ &\le \|m\|_{B(x_0,3\delta)} + \sup_{x\in B(x_0,\delta)} \mathbb{E}\left[|m(X_1)|\,\mathbb{1}_{\{\rho_1\ge 2\delta\}} \,\middle|\, A_n(x)\right].\end{aligned}$$

By Lemmas 5.1 and 5.4, the latter summand vanishes as $n \to \infty$. □

Proposition 5.6. *Let $k \in \mathbb{N}$, suppose there exists $p > k$ with $\|Y\|_{L^p} < \infty$, and put $q := \frac{p}{p-k}$.*

(i) If m_k is continuous, then $\mathbb{E}\left[\widehat{m}_k^{(n)}(\cdot)\right] \to m_k(\cdot)$ uniformly on compact subsets of $\mathrm{S}(X)$.

(ii) If $x \in \mathrm{S}(X)$ and m_k is γ-Hölder continuous at x for some $\gamma \in (0,1]$, then, for any sequence $0 < \alpha_n = o(n/k_n)$,

$$\left|\mathbb{E}\left[\widehat{m}_k^{(n)}(x)\right] - m_k(x)\right| = O\left(p_x^{-1}\left((1+\alpha_n)\frac{k_n}{n}\right)^{\gamma} + n\exp\left(-\frac{\alpha_n^2}{1+\alpha_n}\frac{k_n}{2q}\right)\right).$$

Proof. Assume $k = 1$ and $\mathrm{S}(X) = E$ (otherwise consider $\widetilde{Y} := Y^k$, $\widetilde{p} := p/k > 1$, and $\widetilde{E} := \mathrm{S}(X)$). Abbreviate $A := \{\Sigma_1(x) \le k_n\}$ (n and x will be clear from context). If $x \in E$, and $n \in \mathbb{N}$, then Lemma 5.3 implies that, for any $\delta > 0$,

$$\begin{aligned}\left|\mathbb{E}\left[\widehat{m}^{(n)}(x)\right] - m(x)\right| &= \left|\mathbb{E}\left[m(X_1) - m(x) \,|\, A\right]\right| \\ &\le \mathbb{E}\left[|m(X_1) - m(x)|\,\mathbb{1}_{\{\rho(X_1,x)<\delta\}} \,\middle|\, A\right] \\ &\quad + \mathbb{E}\left[|m(X_1) - m(x)|\,\mathbb{1}_{\{\rho(X_1,x)\ge\delta\}} \,\middle|\, A\right].\end{aligned} \tag{15}$$

(i) If $C \subset E$ is compact, then, by continuity of m, for any $\varepsilon > 0$ there exists some $\delta > 0$ such that $|m(x) - m(x')| < \varepsilon$ whenever $x \in C$ and $x' \in E$ with $\rho(x,x') < \delta$, so that, for such δ, the former summand on the right-hand side (RHS) of Eq. (15) is at most ε for any $x \in C$. The latter summand vanishes uniformly on C as $n \to \infty$ by Lemmas 5.1 and 5.4 and since $\|m\|_C < \infty$ by continuity of m. Hence,

$$\varlimsup_{n\to\infty}\left(\sup_{x\in C}\left|\mathbb{E}\left[\widehat{m}^{(n)}(x)\right] - m(x)\right|\right) \le \varepsilon,$$

for any $\varepsilon > 0$.

(ii) Suppose that $x \in E$ and that there exist $\gamma \in (0,1]$ and $c_1, \delta_0 > 0$ such that $|m(y) - m(z)| \le c_1\rho(y,z)^\gamma$ for all $y, z \in B(x,\delta_0)$. Then, for any $n \in \mathbb{N}$ and $0 < \delta < \delta_0$ with $p_x(\delta) > \frac{k_n}{n}$, Eq. (15) and Lemmas 5.1 and 5.4 imply

$$\begin{aligned}\left|\mathbb{E}\left[\widehat{m}^{(n)}(x)\right] - m(x)\right| &\le c_1\delta^\gamma + (|m(x)| + n\,\|Y\|_{L^p})\,\mathbb{P}\left(N_n(x,\delta) < k_n\right)^{1/q} \\ &\le c_1\delta^\gamma + (|m(x)| + n\,\|Y\|_{L^p})\exp\left(-\frac{n}{2qp_x(\delta)}\left(p_x(\delta) - \frac{k_n}{n}\right)^2\right),\end{aligned}$$

where $q = \frac{p}{p-1} \in (1, \infty)$. Let $(\alpha_n) \in (0, \infty)^{\mathbb{N}}$ be such that $\alpha_n = o(n/k_n)$, so that $0 < \varepsilon_n := (1 + \alpha_n)\frac{k_n}{n} \to 0$. Then, by Lemma 2.1, $0 < \delta_n := p_x^{-1}(\varepsilon_n) < \delta_0$ for large enough $n \in \mathbb{N}$, and $p_x(\delta_n) \geq \varepsilon_n > k_n/n$. Thus, for large enough $n \in \mathbb{N}$,

$$\begin{aligned}\left|\mathbb{E}\left[\widehat{m}^{(n)}(x)\right] - m(x)\right| &\leq c_1\delta_n^\gamma + 2n\,\|Y\|_{L^p} \exp\left(-\frac{n}{2q\varepsilon_n}\left(\varepsilon_n - \frac{k_n}{n}\right)^2\right) \\ &= c_1 p_x^{-1}\left((1+\alpha_n)\frac{k_n}{n}\right)^\gamma + 2n\,\|Y\|_{L^p} \exp\left(-\frac{k_n}{2q}\frac{\alpha_n^2}{1+\alpha_n}\right),\end{aligned}$$

which implies (ii). □

Proposition 5.7. *Let $k \in \mathbb{N}$, and suppose there exists $p > 2k$ with $\|Y\|_{L^p} < \infty$, and that m_k is continuous and m_{2k} is locally bounded. Put $q := \frac{p}{p-k}$.*

(i) $\mathbb{V}\left(\widehat{m}_k^{(n)}(\cdot)\right) \to 0$ *uniformly on compact subsets of* $\mathrm{S}(X)$,

(ii) If $x \in \mathrm{S}(X)$ and m_k is γ-Hölder continuous at x for some $\gamma \in (0,1]$, then, for any sequence $0 < \alpha_n = o(n/k_n)$,

$$\mathbb{V}\left(\widehat{m}_k^{(n)}(x)\right) = O\left(\frac{1}{k_n} + p_x^{-1}\left((1+\alpha_n)\frac{k_n}{n}\right)^\gamma + n^2\exp\left(-\frac{k_n}{2q}\frac{\alpha_n^2}{1+\alpha_n}\right)\right).$$

Proof. We may again assume $k = 1$ and $\mathrm{S}(X) = E$. Let $x \in E$ and $n \in \mathbb{N}$. Then,

$$\begin{aligned}\mathbb{E}\left[\widehat{m}^{(n)}(x)^2\right] &= \frac{1}{k_n^2}\sum_{i,j=1}^n \mathbb{E}\left[Y_iY_j\mathbb{1}_{\{\Sigma_i(x)\leq k_n, \Sigma_j(x)\leq k_n\}}\right] \\ &= \frac{1}{k_n^2}\left(\sum_{i=1}^n \mathbb{E}\left[Y_i^2\mathbb{1}_{\{\Sigma_i(x)\leq k_n\}}\right] + n(n-1)\mathbb{E}\left[Y_1Y_2\mathbb{1}_{\{\Sigma_1(x)\leq k_n, \Sigma_2(x)\leq k_n\}}\right]\right) \\ &= \frac{1}{k_n}\mathbb{E}\left[\widehat{m}_2^{(n)}(x)\right] + \frac{n(n-1)}{k_n^2}\mathbb{E}\left[Y_1Y_2\mathbb{1}_{\{\Sigma_1(x)\leq k_n, \Sigma_2(x)\leq k_n\}}\right].\end{aligned} \tag{16}$$

Conditioning on X_1, X_2 in the second summand, and using Eq. (2) yields

$$\begin{aligned}\mathbb{E}\left[Y_1Y_2\mathbb{1}_{\{\Sigma_1(x)\vee\Sigma_2(x)\leq k_n\}}\right] &= \mathbb{E}\left[m(X_1)m(X_2)\mathbb{1}_{\{\Sigma_1(x)\vee\Sigma_2(x)\leq k_n\}}\right] \\ &= \frac{k_n(k_n-1)}{n(n-1)}\mathbb{E}\left[m(X_1)m(X_2)\,|\,\Sigma_1(x)\vee\Sigma_2(x)\leq k_n\right],\end{aligned}$$

which we combine with Eq. (16) to obtain, abbreviating $A := A_n(x) := \{\Sigma_1(x)\vee\Sigma_2(x)\leq k_n\}$,

$$\mathbb{E}\left[\widehat{m}^{(n)}(x)^2\right] = \frac{1}{k_n}\mathbb{E}\left[\widehat{m}_2^{(n)}(x)\right] + \left(1 - \frac{1}{k_n}\right)\mathbb{E}\left[m(X_1)m(X_2)\,|\,A\right].$$

We conclude that

$$
\begin{aligned}
\mathbb{V}\left(\widehat{m}^{(n)}(x)\right) = \mathbb{E}\left[\widehat{m}^{(n)}(x)^2\right] - \mathbb{E}\left[\widehat{m}^{(n)}(x)\right]^2 &\leq \overbrace{\left|\mathbb{E}\left[\widehat{m}^{(n)}(x)\right]^2 - m(x)^2\right|}^{(*_1)} \\
+ \underbrace{\left|\mathbb{E}\left[m(X_1)m(X_2) \mid A\right] - m(x)^2\right|}_{(*_2)} &+ \underbrace{\frac{1}{k_n}\left(\left|\mathbb{E}\left[\widehat{m}_2^{(n)}(x)\right]\right| + \left|\mathbb{E}\left[m(X_1)m(X_2) \mid A\right]\right|\right)}_{(*_3)},
\end{aligned} \tag{17}
$$

where we introduced $m(x)^2$ using the triangle inequality. Put $\omega(m, A) := \sup_{y,z\in A} |m(y) - m(z)|$ for $A \subset E$. Then, whenever $\delta > 0$ and $x_1, x_2 \in B(x, \delta)$,

$$
\begin{aligned}
\left|m(x_1)m(x_2) - m(x)^2\right| &= \left|m(x_1)m(x_2) - m(x_1)m(x) + m(x_1)m(x) - m(x)^2\right| \\
&\leq |m(x_1)|\,|m(x_2) - m(x)| + |m(x)|\,|m(x_1) - m(x)| \\
&\leq 2\,\|m\|_{B(x,\delta)}\,\omega(m, B(x, \delta)).
\end{aligned} \tag{18}
$$

Hence, and by Lemma 5.4, for any $\delta > 0$,

$$
\begin{aligned}
(*_2) &\leq \mathbb{E}\left[\left|m(X_1)m(X_2) - m(x)^2\right|\left(\mathbb{1}_{\{\rho_1\vee\rho_2\geq\delta\}} + \mathbb{1}_{\{\rho_1\vee\rho_2<\delta\}}\right) \mid A\right] \\
&\leq \left(n^2\,\|Y\|_{L^p}^2 + |m(x)|^2\right)\mathbb{P}\left(N_n(x,\delta) < k_n\right)^{1/q} \\
&\quad + 2\,\|m\|_{B(x,\delta)}\,\omega(m, B(x, \delta)),
\end{aligned} \tag{19}
$$

where $q = \frac{p}{p-1} \in (1, 2)$.

(i) Fix a compact set $C \subset \mathrm{S}(X)$, and let $\varepsilon > 0$. Since C is compact, there is $\delta > 0$ such that $\|m\|_{B(x,\delta)} \leq \|m\|_C + 1$ and $\omega(m, B(x, \delta)) \leq \varepsilon$ for all $x \in C$. Together with Eq. (19) and Lemma 5.1, this implies

$$
\varlimsup_{n\to\infty}\left(\sup_{x\in C}(*_2)\right) \leq 2(\|m\|_C + 1)\varepsilon
$$

for all $\varepsilon > 0$, so $(*_2)$ vanishes uniformly on C. In particular,

$$
\sup_{n\in\mathbb{N},x\in C}\left|\mathbb{E}\left[m(X_1)m(X_2) \mid A_n(x)\right]\right| < \infty,
$$

which together with Lemma 5.5 implies that $(*_3)$ also vanishes uniformly on C, and $(*_1)$ does the same by Proposition 5.6.

(ii) Suppose that $x \in E$ and $\gamma \in (0, 1]$, $\delta_0, c_1 > 0$ are such that $\omega(m, B(x, \delta)) \leq c_1\delta^\gamma$ for all $\delta \in (0, \delta_0)$, so by Eq. (19) and Lemma 5.1, for any $\delta \in (0, \delta_0)$,

$$
(*_2) \leq 2c_1\,\|m\|_{B(x,\delta)}\,\delta^\gamma + \left(n^2\,\|Y\|_{L^p}^2 + |m(x)|^2\right)\exp\left(-\frac{n}{2qp_x(\delta)}\left(p_x(\delta) - \frac{k_n}{n}\right)^2\right).
$$

Now let $(\alpha_n) \in (0, \infty)^{\mathbb{N}}$ be such that $\alpha_n = o(n/k_n)$, so that $0 < \varepsilon_n := (1 + \alpha_n)\frac{k_n}{n} \to 0$. Then, by Lemma 2.1, $0 < \delta_n := p_x^{-1}(\varepsilon_n) < \delta_0$ for large enough $n \in \mathbb{N}$, and $p_x(\delta_n) \geq \varepsilon_n > k_n/n$. Thus, for large enough $n \in \mathbb{N}$,

$$\begin{aligned}(*_2) &\leq 2c_1 \|m\|_{B(x,\delta_n)} \delta_n^\gamma + \left(n^2 \|Y\|_{L^p}^2 + |m(x)|^2\right) \exp\left(-\frac{n}{2q\varepsilon_n}\left(\varepsilon_n - \frac{k_n}{n}\right)^2\right) \\ &= O\left(p_x^{-1}\left((1+\alpha_n)\frac{k_n}{n}\right)^\gamma + n^2 \exp\left(-\frac{k_n}{2q}\frac{\alpha_n^2}{1+\alpha_n}\right)\right).\end{aligned}$$

We have shown in (i) that $\mathbb{E}\left[m(X_1)m(X_2) \mid A_n(x)\right] \to m(x)^2$ as $n \to \infty$, and, by Lemma 5.5, $\overline{\lim}_{n\to\infty} \left|\mathbb{E}\left[\widehat{m}_2^{(n)}(x)\right]\right| \leq m_2(x)$, so

$$(*_3) \leq (1+o(1))\frac{m_2(x) + m(x)^2}{k_n} = O\left(\frac{1}{k_n}\right).$$

Finally, Proposition 5.6 and Lemma 5.5 imply

$$\begin{aligned}(*_1) &= \left|\left(\mathbb{E}\left[\widehat{m}^{(n)}(x)\right] + m(x)\right)\left(\mathbb{E}\left[\widehat{m}^{(n)}(x)\right] - m(x)\right)\right| \\ &= O\left(p_x^{-1}\left((1+\alpha_n)\frac{k_n}{n}\right)^\gamma + n \exp\left(-\frac{k_n}{2q}\frac{\alpha_n^2}{1+\alpha_n}\right)\right).\end{aligned}$$

□

Theorem 3.1 is now an easy consequence of Propositions 5.6 and 5.7.

Proof of Theorem 3.1. As always we assume $k = 1$ and $E = \mathrm{S}(X)$. If $x \in E$, then

$$\begin{aligned}\left\|\widehat{m}^{(n)}(x) - m(x)\right\|_{L^2} &\leq \left\|\widehat{m}^{(n)}(x) - \mathbb{E}\left[\widehat{m}^{(n)}(x)\right]\right\|_{L^2} + \left\|\mathbb{E}\left[\widehat{m}^{(n)}(x)\right] - m(x)\right\|_{L^2} \\ &= \sqrt{\mathbb{E}\left[\left(\widehat{m}^{(n)}(x) - \mathbb{E}\left[\widehat{m}^{(n)}(x)\right]\right)^2\right]} + \left|\mathbb{E}\left[\widehat{m}^{(n)}(x)\right] - m(x)\right| \\ &= \sqrt{\mathbb{V}\left(\widehat{m}^{(n)}(x)\right)} + \left|\mathbb{E}\left[\widehat{m}^{(n)}(x)\right] - m(x)\right|.\end{aligned}$$

This bound vanishes uniformly on compact sets by Propositions 5.6 and 5.7, which proves (i). Now suppose $x \in E$ and $\gamma \in (0,1]$, $c_1, \delta_0 > 0$, such that $|m(y) - m(z)| \leq c_1 \rho(y,z)^\gamma$ for all $y, z \in B(x,\delta)$, and fix a sequence $0 < \alpha_n = o(n/k_n)$. Then, Proposition 5.7(ii) and Proposition 5.6(ii) respectively give asymptotic bounds $\mathbb{V}(\widehat{m}^{(n)}(x)) = O((\mathrm{v}))$ and $\left|\mathbb{E}\left[\widehat{m}^{(n)}(x)\right] - m(x)\right| = O((\mathrm{e}))$. A comparison yields $(\mathrm{e}) = O((\mathrm{v}))$, and since both go to zero as $n \to \infty$, $(\mathrm{e}) = o(\sqrt{(\mathrm{v})})$. Thus,

$$\left\|\widehat{m}^{(n)}(x) - m(x)\right\|_{L^2} = O\left(\sqrt{(\mathrm{v})} + (\mathrm{e})\right) = O(\sqrt{(\mathrm{v})}).$$

The final assertion follows by choosing $\alpha_n = 1,\ n \in \mathbb{N}$, and because $q < 2$ (since $\frac{1}{q} + \frac{1}{p} = 1$ and $p > 2$). □

Proof of Theorem 3.2. Let $C \subset \mathrm{S}(X)$ be compact. Recall from Eq. (4) that $\widehat{v}^{(n)}(\cdot) = \widehat{m}_2^{(n)}(\cdot) - \widehat{m}^{(n)}(\cdot)^2$, so, for $x \in E$,

$$\begin{aligned}\left\|\widehat{v}^{(n)}(x) - v(x)\right\|_{L^1} &\leq \left\|\widehat{m}_2^{(n)}(x) - m_2(x)\right\|_{L^1} + \left\|\widehat{m}^{(n)}(x)^2 - m(x)^2\right\|_{L^1} \\ &\leq \left\|\widehat{m}_2^{(n)}(x) - m_2(x)\right\|_{L^2} + \left\|\widehat{m}^{(n)}(x) + m(x)\right\|_{L^2}\left\|\widehat{m}^{(n)}(x) - m(x)\right\|_{L^2}.\end{aligned}$$

where we used the Hölder inequality on both summands in the second step. Applying Theorem 3.1 with $k = 1$ and $k = 2$ gives $\widehat{m}^{(n)}(\cdot) \xrightarrow{L^2} m(\cdot)$ and $\widehat{m}_2^{(n)}(\cdot) \xrightarrow{L^2} m_2(\cdot)$ uniformly on C. In particular, $\sup_{n\in\mathbb{N},x\in C} \left\|\widehat{m}^{(n)}(x)\right\|_{L^2} < \infty$, so the bound above vanishes uniformly on C. $\square$

We close with the proofs of Theorems 4.3 and 4.6.

Proof of Theorem 4.3. Assume $k = 1$ and $\mathrm{S}(X) = E$. By assumption, there are $c_1, \delta_0 > 0$ such that $p_x(\delta) \geq c_1\delta^s$ for all $\delta \in (0, \delta_0)$, and thus $p_x^{-1}(\varepsilon) \leq \left(\frac{\varepsilon}{c_1}\right)^{1/s}$ for $\varepsilon \in (0, c_1\delta_0^s)$. Hence, and by Theorem 3.1, for any $0 < \alpha_n = o(n/k_n)$,

$$\left\|\widehat{m}^{(n)}(x) - m(x)\right\|_{L^2}^2 = O\left(\frac{1}{k_n} + \left((1+\alpha_n)\frac{k_n}{n}\right)^{\gamma/s} + n^2 \exp\left(-\frac{k_n}{4}\frac{\alpha_n^2}{1+\alpha_n}\right)\right). \quad (20)$$

Choose $\alpha_n > 0$ such that $\frac{k_n}{4}\frac{\alpha_n^2}{1+\alpha_n} = (2 + \gamma/s)\log n$. That is,

$$\alpha_n = 2(2+\gamma/s)\frac{\log n}{k_n}\left(1 + \sqrt{1 + \frac{1}{2+\gamma/s}\frac{k_n}{\log n}}\right) \sim 2\sqrt{(2+\gamma/s)\frac{\log n}{k_n}},$$

since $\log n/k_n \to 0$. Then, $\alpha_n = o(1) = o(n/k_n)$, so Eq. (20) is valid and turns into

$$\left\|\widehat{m}^{(n)}(x) - m(x)\right\|_{L^2}^2 = O\left(\frac{1}{k_n} + \left((1+\alpha_n)\frac{k_n}{n}\right)^{\gamma/s} + n^{-\gamma/s}\right) = O\left(\frac{1}{k_n} + \left(\frac{k_n}{n}\right)^{\gamma/s}\right),$$

where we used that $1 + \alpha_n = O(1)$ and $n^{-\gamma/s} = o((\frac{k_n}{n})^{\gamma/s})$ since $k_n \to \infty$.

Claim. *If $a, b, \eta > 0$, then $f\colon (0,\infty) \to (0,\infty); x \mapsto \frac{a}{x} + bx^\eta$ has a unique minimum at $x_0 = \left(\frac{a}{\eta b}\right)^{1/(\eta+1)}$. Furthermore, if $k \in \mathbb{N}$ minimises f on $\mathbb{N}$, then $k \in \{\lfloor x_0\rfloor, \lceil x_0\rceil\}$.*

Proof of Claim. Clearly $f \in C^1((0,\infty))$, and if $\bowtie \in \{<, >, =\}$, then

$$f'(x) = -\frac{a}{x^2} + \eta b x^{\eta-1} \bowtie 0 \iff x \bowtie x_0 := \left(\frac{a}{\eta b}\right)^{1/(\eta+1)}.$$

Hence f has a unique minimum at x_0, is decreasing on $(0, x_0)$, and increasing on (x_0, ∞). In particular, if $k \in \mathbb{N}$ minimises f on $\mathbb{N}$, then $k \in \{\lfloor x_0\rfloor, \lceil x_0\rceil\}$. $\diamond$

Applying this with $a = 1, b = n^{-\gamma/s}, \eta = \gamma/s$ yields that the bound in Eq. (9) is optimised over $(0,\infty)$ by $\kappa_n := cn^{1/(1+s/\gamma)}, n \in \mathbb{N}$, where $c = \left(\frac{s}{\gamma}\right)^{s/(s+\gamma)}$, and that an optimal $k_n \in \mathbb{N}$ satisfies $k_n \in \{\lfloor\kappa_n\rfloor, \lceil\kappa_n\rceil\}$. Now suppose that $k_n \sim c'n^{1/(1+s/\gamma)}$ for some $c' > 0$. Then, $\log n \ll k_n$, and $k_n = o(n)$, so Eq. (9) is valid and turns into

$$\left\|\widehat{m}^{(n)}(x) - m(x)\right\|_{L^2}^2 = O\left(\left(1/c' + c'^{\gamma/s}\right) n^{-\frac{1}{1+s/\gamma}}\right) = O\left(n^{-\frac{1}{1+s/\gamma}}\right),$$

which implies Eq. (10). $\square$

Remark 5.8. The choice for α_n we made in the proof above was optimal in the sense that the achieved asymptotic bound Eq. (9) is a strict lower bound on the RHS of Eq. (20) for any choice of $\alpha_n > 0$.

We turn to the proof of Theorem 4.6. Recall that we assume X to be a centred Gaussian process taking values in $E = C([0,1])$, which satisfies Eq. (12). We again denote by $\mathcal{H} \subset E$ the reproducing kernel Hilbert space of X.

Lemma 5.9. *There is a $c > 0$ such that, for any $x \in \mathcal{H}$,*

$$p_x(\delta) \geq \mathrm{e}^{-\|x\|_{\mathcal{H}}^2/2} \exp\left(-c\delta^{-1/\beta}\right), \quad \delta > 0,$$

where $\|\cdot\|_{\mathcal{H}}$ denotes the norm induced by the scalar product on $\mathcal{H}$.

Proof. For $t, s \in [0,1]$, by Eq. (12),

$$\begin{aligned}\mathbb{E}\left[|X_t - X_s|^2\right] = K(t,t) - 2K(t,s) + K(s,s) &\leq |K(t,t) - K(t,s)| + |K(s,t) - K(s,s)| \\ &\leq 2c\,|t-s|^{2\beta}.\end{aligned}$$

Now the claim follows from Theorems 3.2 and 5.2 in [15]. □

Proof of Theorem 4.6. Assume $k = 1$ and $\mathrm{S}(X) = E$. By Lemma 5.9, there are $c_1, c_2 > 0$ such that $p_x(\delta) \geq c_1 \exp\left(-c_2\delta^{-1/\beta}\right)$ for all $\delta > 0$, and thus $p_x^{-1}(\varepsilon) \leq \left(\frac{1}{c_2}\log\frac{c_1}{\varepsilon}\right)^{-\beta}$ for $\varepsilon \in (0,1)$. Hence, and by Theorem 3.1, for any $0 < \alpha_n = o(n/k_n)$,

$$\left\|\widehat{m}^{(n)}(x) - m(x)\right\|_{L^2}^2 = O\left(\frac{1}{k_n} + \log\left(\frac{n}{(1+\alpha_n)k_n}\right)^{-\gamma\beta} + n^2\exp\left(-\frac{k_n}{4}\frac{\alpha_n^2}{1+\alpha_n}\right)\right). \quad (21)$$

Choose $\alpha_n > 0$ such that $\frac{k_n}{4}\frac{\alpha_n^2}{1+\alpha_n} = 3\log n$, that is,

$$\begin{aligned}\alpha_n = 6\frac{\log n}{k_n}\left(1 + \sqrt{1 + \frac{3k_n}{\log n}}\right) &\leq 12\frac{\log n}{k_n}\sqrt{1 + \frac{3k_n}{\log n}} \\ &\leq 12\frac{\log n}{k_n}\begin{cases}2, & k_n \leq \log n \\ \sqrt{\frac{4k_n}{\log n}}, & k_n > \log n\end{cases} \\ &\leq 24\begin{cases}\frac{\log n}{k_n}, & k_n \leq \log n \\ 1, & k_n > \log n.\end{cases}\end{aligned}$$

In either case, $\alpha_n = o(n/k_n)$, so Eq. (21) is valid. If $k_n > \log n$, then

$$\log\left(\frac{n}{(1+\alpha_n)k_n}\right) \geq \log\left(\frac{n}{25k_n}\right) = \log\frac{n}{k_n} - \log 25 \sim \log\frac{n}{k_n}.$$

If $k_n \leq \log n$, then

$$\begin{aligned}\log\left(\frac{n}{(1+\alpha_n)k_n}\right) &\geq \log\left(\frac{n}{k_n + 24\log n}\right) \geq \log\left(\frac{n}{(k_n+24)\log n}\right)\\ &= \log n - \log(k_n+24) - \log\log n \sim \log n - \log k_n\\ &= \log\frac{n}{k_n}.\end{aligned}$$

Hence, for any $n \in \mathbb{N}$, Eq. (21) turns into

$$\begin{aligned}\left\|\widehat{m}^{(n)}(x) - m(x)\right\|_{L^2}^2 &= O\left(\frac{1}{k_n} + \log\left(\frac{n}{(1+\alpha_n)k_n}\right)^{-\gamma\beta} + \frac{1}{n}\right)\\ &= O\left(\frac{1}{k_n} + \log\left(\frac{n}{k_n}\right)^{-\gamma\beta}\right),\end{aligned}$$

where we used that $\frac{1}{n} = o(\frac{1}{k_n})$ since $k_n = o(n)$.

Claim. *If $\eta > 0$, $n \in \mathbb{N}$, then $f\colon (0,n) \to (0,\infty) : x \mapsto \frac{1}{x} + \left(\log\frac{n}{x}\right)^{-\eta}$ has a unique minimum at $x_0 = x_0(n) \in (0,n)$ with*

$$x_0 \sim \frac{1}{\eta}\left(\log n\right)^{\eta+1}.$$

If $k \in \mathbb{N} \cap (0,n)$ minimises f, then $k \in \{\lfloor x_0 \rfloor, \lceil x_0 \rceil\}$.

Proof of Claim. Clearly, $f \in C^1((0,n))$, and $\lim_{x\downarrow 0} f(x) = \lim_{x\uparrow n} f(x) = \infty$, so f has at least one local minimum, all of which must be zeros of f'. For $x \in (0,n)$,

$$\begin{aligned}f'(x) = 0 &\iff -\frac{1}{x^2} + \frac{\eta}{x}\left(\log\frac{n}{x}\right)^{-(\eta+1)} = 0\\ &\iff (\eta x)^{1/(\eta+1)} = \log\frac{n}{x}\\ &\iff n = x\exp\left((\eta x)^{1/(\eta+1)}\right)\\ &\iff x_0 = aW\left((n/a)^{1/(1+\eta)}\right)^{1+\eta},\end{aligned}$$

where $a = \frac{1}{\eta}(1+\eta)^{(1+\eta)}$, and W denotes the Lambert-W function. For $z > 0$, $w = W(z)$ is the unique positive number for which $we^w = z$. It is well-known and easy to show that $W(z) \sim \log z$, so

$$x_0 \sim a\left(\log\left((n/a)^{1/(1+\eta)}\right)\right)^{1+\eta} = \frac{1}{\eta}\left(\log n - \log a\right)^{1+\eta} \sim \frac{1}{\eta}(\log n)^{1+\eta}.$$

Now suppose that $k \in \mathbb{N} \cap (0,n)$ minimises f, and assume for contradiction that $k \notin \{\lfloor x_0 \rfloor, \lceil x_0 \rceil\}$, say $k \in (0, \lfloor x_0 \rfloor)$. Then $f(k) < f(\lfloor x_0 \rfloor)$, but since $\lim_{x\downarrow 0} f(x) = \infty$, this would imply that f has a local minimum in $(0, \lfloor x_0 \rfloor)$, which contradicts the fact that x_0 is the only local minimum. ◇

Applying this with $\eta = \gamma\beta$ yields that the bound in Eq. (13) is minimised over $(0,n)$ by a unique κ_n that satisfies $\kappa_n \sim \frac{1}{\gamma\beta}\left(\log n\right)^{1+\gamma\beta}$, and that an optimal $k_n \in \mathbb{N}$ satisfies $k_n \in \{\lfloor \kappa_n \rfloor, \lceil \kappa_n \rceil\}$.

Finally, suppose that there is $a \in (0, 1)$ and $K > 0$ with $(\log n)^{\gamma\beta} \ll k_n \leq K n^a$. Then, $k_n = o(n)$ and $\log \frac{n}{k_n} = \log n - \log k_n \geq \log n - a \log n - \log K \sim (1 - a) \log n$, so Eq. (13) turns into

$$\left\|\widehat{m}^{(n)}(x) - m(x)\right\|_{L^2}^2 = O\left(\frac{1}{k_n} + (\log n)^{-\gamma\beta}\right) = O\left((\log n)^{-\gamma\beta}\right). \qquad \square$$

Again, note that the choice for α_n we made in the proof above was optimal in the sense that the achieved asymptotic bound Eq. (13) is a lower bound on the RHS of Eq. (21) for any choice of $\alpha_n > 0$.

References

[1] Heinz Bauer. *Wahrscheinlichkeitstheorie und Grundzüge der Maßtheorie*. De Gruyter, 1968. ISBN: 9783111655260. DOI: 10.1515/9783111655260.

[2] Gérard Biau and Luc Devroye. *Lectures on the Nearest Neighbor Method*. Vol. 246. Springer, 2015. DOI: 10.1007/978-3-319-25388-6.

[3] Vladimir I. Bogachev. *Gaussian Measures*. Mathematical Surveys and Monographs. American Mathematical Society, 1998. ISBN: 9781470418694. DOI: 10.1090/surv/062.

[4] Frédéric Cérou and Arnaud Guyader. 'Nearest neighbor classification in infinite dimension'. In: *ESAIM: Probability and Statistics* 10 (2006), pp. 340–355. DOI: 10.1051/ps:2006014.

[5] Kamalika Chaudhuri and Sanjoy Dasgupta. 'Rates of Convergence for Nearest Neighbor Classification'. In: *Advances in Neural Information Processing Systems*. Vol. 27. 2014. URL: https://proceedings.neurips.cc/paper/2014/file/db957c626a8cd7a27231adfbf51e20eb-Paper.pdf.

[6] Philip E. Cheng. 'Strong Consistency of Nearest Neighbor Regression Function Estimators'. In: *Journal of Multivariate Analysis* 15.1 (1984), pp. 63–72. DOI: 10.1016/0047-259x(84)90067-8.

[7] Thomas Cover and Peter Hart. 'Nearest Neighbor Pattern Classification'. In: *IEEE Transactions on Information Theory* 13.1 (1967), pp. 21–27. DOI: 10.1109/tit.1967.1053964.

[8] Luc Devroye. 'On the Almost Everywhere Convergence of Nonparametric Regression Function Estimates'. In: *The Annals of Statistics* 9.6 (1981), pp. 1310–1319. DOI: 10.1214/aos/1176345647.

[9] Luc Devroye. 'The Uniform Convergence of Nearest Neighbor Regression Function Estimators and Their Application in Optimization'. In: *IEEE Transactions on Information Theory* 24.2 (1978), pp. 142–151. DOI: 10.1109/tit.1978.1055865.

[10] Kenneth Falconer. *Fractal Geometry: Mathematical Foundations and Applications*. John Wiley & Sons, 2003. ISBN: 9780470013854. DOI: 10.1002/0470013850.

[11] Karen Fuchs, Wolfgang Pößnecker and Gerhard Tutz. 'Classification of Functional Data with k-Nearest-Neighbor Ensembles by Fitting Constrained Multinomial Logit Models'. In: *arXiv e-prints* (2016). DOI: 10.48550/ARXIV.1612.04710.

[12] Laszlo Györfi. 'The rate of convergence of k-NN regression estimates and classification rules'. eng. In: *IEEE transactions on information theory* 27.3 (1981), pp. 362–364. DOI: 10.1109/tit.1981.1056344.

[13] Thomas Laloë. 'A k-nearest neighbor approach for functional regression'. In: *Statistics & Probability Letters* 78.10 (2008), pp. 1189–1193. DOI: 10.1016/j.spl.2007.11.014.

[14] Muriel Lang, Oliver Dunkley and Sandra Hirche. 'Gaussian Process Kernels for Rotations and 6D Rigid Body Motions'. In: *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2014, pp. 5165–5170. DOI: 10.1109/icra.2014.6907617.

[15] Wenbo V. Li and Qiman Shao. 'Gaussian Processes: Inequalities, Small Ball Probabilities and Applications'. In: *Stochastic Processes: Theory and Methods*. Vol. 19. Handbook of Statistics. Elsevier, 2001, pp. 533–597. DOI: 10.1016/S0169-7161(01)19019-X.

[16] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2005. ISBN: 9780262256834. DOI: 10.7551/mitpress/3206.001.0001.

[17] Richard M. Royall. *A Class of Non-parametric Estimates of a Smooth Regression Function*. Tech. rep. Stanford University, 1966. URL: https://purl.stanford.edu/vq604wd0459.

[18] Charles J. Stone. 'Consistent Nonparametric Regression'. In: *The Annals of Statistics* 5.4 (1977), pp. 595–620. DOI: 10.1214/aos/1176343886.