Mach Learn (2015) 98:407–433
DOI 10.1007/s10994-014-5437-0

Asymptotic analysis of the learning curve for Gaussian process regression

Loic Le Gratiet · Josselin Garnier

Received: 10 January 2013 / Accepted: 25 February 2014 / Published online: 21 March 2014
© The Author(s) 2014

Abstract This paper deals with the learning curve in a Gaussian process regression framework. The learning curve describes the generalization error of the Gaussian process used for the regression. The main result is the proof of a theorem giving the generalization error for a large class of correlation kernels and for any dimension when the number of observations is large. From this theorem, we can deduce the asymptotic behavior of the generalization error when the observation error is small. The presented proof generalizes previous ones that were limited to special kernels or to small dimensions (one or two). The theoretical results are applied to a nuclear safety problem.

Keywords Gaussian process regression · Asymptotic mean squared error · Learning curves · Generalization error · Convergence rate

1 Introduction

Gaussian process regression is a useful tool to approximate an objective function given some of its observations (Laslett 1994). It has originally been used in geostatistics to interpolate a random field at unobserved locations (Wackernagel 2003; Berger et al. 2001; Gneiting et al. 2010), it has been developed in many areas such as environmental and atmospheric sciences.

Editor: Kristian Kersting.

L. Le Gratiet
Université Paris Diderot, 75205 Paris Cedex 13, France

L. Le Gratiet (✉)
CEA, DAM, DIF, 91297 Arpajon, France
e-mail: loic.legratiet@gmail.com

J. Garnier
Laboratoire de Probabilites et Modeles Aleatoires & Laboratoire Jacques-Louis Lions,
Universite Paris Diderot, 75205 Paris Cedex 13, France

This method has become very popular during the last decades to build surrogate models from noise-free observations. For example, it is widely used in the field of "computer experiments" to build models which surrogate an expensive computer code (Sacks et al. 1989). Then, through the fast approximation of the computer code, uncertainty quantification and sensitivity analysis can be performed with a low computational cost.

Nonetheless, for many realistic cases, we do not have direct access to the function to be approximated but only to noisy versions of it. For example, if the objective function is the result of an experiment, the available responses can be tainted by measurement noise. In that case, we can reduce the noise of the observations by repeating the experiments at the same locations. Another example is Monte-Carlo based simulators—also called stochastic simulators—which use Monte-Carlo or Monte-Carlo Markov Chain methods to solve a system of differential equations through its probabilistic interpretation. For such simulators, the noise level can be tuned by the number of Monte-Carlo particles used in the procedure.

In this paper, we are interested in obtaining learning curves describing the generalization error—defined as the averaged mean squared error—of the Gaussian process regression as a function of the training set size (Rasmussen and Williams 2006). The problem has been addressed in the statistical and numerical analysis areas. For an overview, the reader is referred to (Ritter 2000b) for a numerical analysis point of view and to (Rasmussen and Williams 2006) for a statistical one. In particular, in the numerical analysis literature, the authors are interested in numerical differentiation of functions from noisy data (Ritter 2000a; Bozzini and Rossini 2003). They have found very interesting results for kernels satisfying the Sacks–Ylvisaker conditions of order r (Sacks and Ylvisaker 1981) but only valid for 1-D or 2-D functions.

In the statistical literature Sollich and Hallees (2002) give accurate approximations to the learning curve and Opper and Vivarelli (1999) and Williams and Vivarelli (2000) give upper and lower bounds on it. Their approximations give the asymptotic value of the learning curve (for a very large number of observations). They are based on the Woodbury–Sherman–Morrison matrix inversion lemma (Harville 1997) which holds in finite-dimensional cases which correspond to degenerate covariance kernels in our context. Nonetheless, classical kernels used in Gaussian process regression are non-degenerate and we hence are in an infinite-dimensional case and the Woodbury–Sherman–Morrison formula cannot be used directly.

To deal with asymptotics of Gaussian process learning curves for more general kernels, some authors have used other definitions of the generalization error. For example, Seeger et al (2008) present consistency results and convergence rates for cumulative log loss of Bayesian prediction. Then, their work is revisited by van der Vaart and van Zanten (2011) who suggest and study another risk which is an upper bound for the one presented in (Seeger et al (2008)).

The main result of this paper is the proof of a theorem giving the value of the Gaussian process regression mean squared error (MSE) for a large training set size when the observation noise variance is proportional to the number of observations. This value is given as a function of the eigenvalues and eigenfunctions of the covariance kernel. From this theorem, we can deduce an approximation of the learning curve for non-degenerate and degenerate kernels [which generalizes the proofs given in (Opper and Vivarelli 1999; Sollich and Halees 2002; Picheny 2009)] and for any dimension [which generalizes the proofs given in (Ritter 2000a, b; Bozzini and Rossini 2003)].

The rate of convergence of the best linear unbiased predictor (BLUP) is of practical interest since it provides a powerful tool for decision support. Indeed, from an initial experimental design set, it can predict the additional computational budget (defined as the number of experiments including repetitions) necessary to reach a given desired accuracy.

The paper is organized as follows. First we present the asymptotic framework considered in this paper in Sect. 2. Although the main results of the paper are theoretical contributions, an application is provided in order to emphasize the possible implications for real-world problems. Second, we present in Sect. 3 the main result of the paper which is the theorem giving the MSE of the considered model for a large training size. This theorem is proved in Sect. 4. Third, we study the rate of convergence of the generalization error when the noise variance decreases in Sect. 5. The theoretical asymptotic rates of convergences are compared to the obtained ones in a numerical simulations and academic examples. Furthermore, a study on how large the training set size should be for the asymptotic formulas to agree with the numerical ones is provided for the specific case of the Brownian motion. Finally, an industrial application to the safety assessment of a nuclear system containing fissile materials is considered in Sect. 6. This real case emphasizes the effectiveness of the theoretical rate of convergence of the BLUP since it predicts a very good approximation of the budget needed to reach a prescribed precision.

2 Generalization error for noisy observations

The general framework of the paper is given in this section. First, the mathematical formalism on which the theoretical developments are based is presented. Then, the considered application is introduced. Finally, the bridge between the theoretical developments and the application is given.

2.1 Asymptotic framework for the analysis of the generalization error

Let us suppose that we want to approximate an objective function $x \in \mathbb{R}^d \to f(x) \in \mathbb{R}$ from noisy observations of it at points $(x_i)_{i=1,\dots,n}$ with $x_i \in \mathbb{R}^d$. The points of the experimental design set $(x_i)_{i=1,\dots,n}$ are supposed to be sampled from the probability measure μ over $\mathbb{R}^d$. μ is called the design measure, it can have either a compact support (for a bounded input parameter space domain) or unbounded support (for unbounded input parameter space). We hence have n observations of the form $z_i = f(x_i) + \varepsilon_i$ and we consider that $(\varepsilon_i)_{i=1,\dots,n}$ are independently sampled from the Gaussian distribution with mean zero and variance $n\tau$:

$$\varepsilon \sim \mathcal{N}(0, n\tau). \tag{1}$$

with τ a positive constant. Note that the number of observations and the observation noise variance are both controlled by n. A noise additive in the number of observations is one of the main assumptions of this article. Intuitively, it allows for controlling the convergence of the generalization error when n tends to infinity by increasing the regularization. However, this is also the main limitation of the paper since the noise variance is generally independent of the number of observations. As presented in Sect. 2.3, this assumption is suitable for the particular cases of stochastic simulators or experiments with repetitions when the number of simulations or experiments is fixed. The issue of the convergence of the generalization error for more general cases is still an open problem.

The main idea of the Gaussian process regression is to suppose that the objective function $f(x)$ is a realization of a Gaussian process $Z(x)$ with a known mean and a known covariance kernel $k(x, x')$. The mean can be considered equal to zero without loss of generality. Then, denoting by $z^n = [f(x_i) + \varepsilon_i]_{1 \le i \le n}$ the vector of length n containing the noisy observations, we choose as predictor the BLUP given by the equation:

$$\hat{f}(x) = k(x)^T (K + n\tau I)^{-1} z^n, \tag{2}$$

where $k(x) = [k(x, x_i)]_{1 \leq i \leq n}$ is the n-vector containing the covariances between $Z(x)$ and $Z(x_i)$, $1 \leq i \leq n$, $K = [k(x_i, x_j)]_{1 \leq i,j \leq n}$ is the $n \times n$-matrix containing the covariances between $Z(x_i)$ and $Z(x_j)$, $1 \leq i, j \leq n$ and I the $n \times n$ identity matrix. We note here that the unbiasedness means that $\mathbb{E}[\hat{f}(x)] = \mathbb{E}[Z(x)]$ where $\mathbb{E}$ stands for the expectation with respect to the distribution of the Gaussian process $Z(x)$ and the noise ε. The BLUP minimizes the MSE which equals:

$$\sigma^2(x) = k(x, x) - k(x)^T (K + n\tau I)^{-1} k(x). \tag{3}$$

Indeed, if we consider a linear unbiased predictor (LUP) of the form $a(x)^T z^n$, its MSE is given by:

$$\mathbb{E}\left[(Z(x) - a^T(x) Z^n)^2\right] = k(x, x) - 2a(x)^T k(x) + a(x)^T (K + n\tau I) a(x), \tag{4}$$

where $Z^n = [Z(x_i) + \varepsilon_i]_{1 \leq i \leq n}$. The value of $a(x)$ minimizing (4) is $a_{\text{opt}}(x)^T = k(x)^T (K + n\tau I)^{-1}$. Therefore, the BLUP given by $a_{\text{opt}}(x)^T z^n$ is equal to (2) and by substituting $a(x)$ with $a_{\text{opt}}(x)$ in Eq. (4) we obtain the MSE of the BLUP given by Eq. (3).

The main focus of this paper is the asymptotic value of $\sigma^2(x)$ when $n \to +\infty$. From it, we can deduce the asymptotic value of the integrated mean squared error (IMSE)—also called learning curve or generalization error—when $n \to +\infty$. The IMSE is defined by:

$$\text{IMSE} = \int_{\mathbb{R}^d} \sigma^2(x)\, d\mu(x), \tag{5}$$

where μ is the design measure of the input space parameters.

The obtained asymptotic value has already be mentioned in several works (Rasmussen and Williams 2006; Ritter 2000b, a; Bozzini and Rossini 2003; Opper and Vivarelli 1999; Sollich and Halees 2002; Picheny 2009). The original contribution of this paper is a rigorous proof of this result.

2.2 Introduction to stochastic simulators

We present in this section the industrial application studied in Sect. 6.2. A stochastic simulator is a computer code which solves a system of partial differential equations with Monte-Carlo methods. It has the particularity to provide noisy observations centered on the true solution of the system. Stochastic simulators are widely used in the field of nuclear physics to solve transport equations and model systems containing fissile materials (e.g. nuclear reactors, storages of fissile materials, spacecraft reactors). In this paper, we are interested in a storage of dry Plutonium(IV) oxide (PuO_2) used as fuel for nuclear reactors or several spacecrafts. As the PuO_2 is highly toxic, the safety assessment of such storages is of great importance.

One of the most important factors used to assess the safety of a system containing fissile materials is the neutron multiplication coefficient usually denoted by k_{eff}. It is the average number of neutrons from one fission that cause another fission. This factor models the criticality of a chain nuclear reaction:

- $k_{\text{eff}} > 1$ leads to an uncontrolled chain reaction due to an increasing neutron population.
- $k_{\text{eff}} = 1$ leads to a self-sustained chain reaction with a stable neutron population.
- $k_{\text{eff}} < 1$ leads to a faded chain reaction due to an decreasing neutron population.

The neutron multiplication factor is evaluated using the stochastic simulator called MORET (Fernex et al. 2005). It depends on many parameters. However, we only focus here on the following quantities:

– $d_{PuO_2} \in [0.5, 4]\text{g.cm}^{-3}$, the density of the fissile powder. It is scaled to $[0, 1]$.
– $d_{\text{water}} \in [0, 1]\text{g.cm}^{-3}$, the density of water between storage tubes.

We use the notation $x = (d_{PuO_2}, d_{\text{water}})$ for the input parameters. Let us denote by $(Y_j(x))_{j=1,\ldots,s}$ the output of the MORET code at point x. $(Y_j(x))_{j=1,\ldots,s}$ are realizations of independent and identically distributed random variables centered on $k_{\text{eff}}(x)$. They are themselves obtained by an empirical mean of a Monte-Carlo sample of 4000 particles. From these particles, we can estimate the variance σ_ε^2 of the observation $Y_j(x)$ by a classical empirical estimator.

Finally we can estimate $k_{\text{eff}}(x)$ from the following quantity:

$$k_{\text{eff},s}(x) = \frac{1}{s}\sum_{j=1}^{s} Y_j(x).$$

Therefore, the variance of an observation $k_{\text{eff},s}(x)$ equals σ_ε^2/s.

2.3 Relation between the application and the considered mathematical formalism

Let us consider that we want to approximate the function $x \in \mathbb{R}^d \to f(x) \in \mathbb{R}$ from noisy observations at points $(x_i)_{i=1,\ldots,n}$ sampled from the design measure μ and with s replications at each point. We hence have ns data of the form $z_{i,j} = f(x_i) + \varepsilon_{i,j}$ and we consider that $(\varepsilon_{i,j})_{\substack{i=1,\ldots,n \\ j=1,\ldots,s}}$ are independently distributed from a Gaussian distribution with mean zero and variance σ_ε^2. Then, denoting the vector of observed values by $z^n = (z_i^n)_{i=1,\ldots,n} = (\sum_{j=1}^{s} z_{i,j}/s)_{i=1,\ldots,n}$, the variance of an observation z_i^n is σ_ε^2/s. We recognize here the output form given in Sect. 2.2. Thus, if we consider a fixed budget $T = ns$, we have $\sigma_\varepsilon^2/s = n\tau$ with $\tau = \sigma_\varepsilon^2/T$ and the observation noise variance is proportional to n (as presented in Sect. 2.1). It means that if we increase the number n of observations, we automatically increase the uncertainty on the observations. An observation noise variance proportional to n is natural in the framework of experiments with repetitions or stochastic simulators. Indeed, for a fixed number of experiments (or simulations), the user can decide to perform them in few points with many repetitions (in that case the noise variance will be low) or to perform them in many points with few repetitions (in that case the noise variance will be large).

We note that increasing n with a fixed τ is an idealized asymptotic setting since it would require that the number of replications s could tend to zero while it has to be a positive integer. However, this issue can be tackled in practice since for real applications n is finite and one has just to take a budget T such that $T \geq n$ (i.e. $s \geq 1$). This is a first limitation of the suggested method since it cannot be used for small budget (i.e. when $T < n$). A second one is the assumption that s does not depend on x_i. Indeed, a uniform allocation could not be optimal. In this case, finding the optimal sequence $\{s_1, s_2, \ldots, s_n\}$ leading to the minimal error is of practical interest. However, the corresponding observation noise variance will depend on x_i which means that τ_i will depend on x_i as well. In this case, the presenting results do not hold. Nevertheless, they can be used to provide an upper bound for the convergence of the generalization error by considering the worst case $\tau = \max_i \tau_i$.

The objective of the industrial example presented in Sect. 6.2 is to determine the budget T required to reach a prescribe accuracy $\bar{\varepsilon}$. To deal with this issue, we first build a Gaussian process regression model from an initial budget T_0 and a large number of observations n. Then, from the results on the learning curve, we deduce the budget T such that the IMSE equals $\bar{\varepsilon}$.

3 Convergence of the learning curve for Gaussian process regression

This section deals with the convergence of the BLUP when the number of observations is large. The rate of convergence of the BLUP is evaluated through the generalization error—i.e. the IMSE—defined in (5). The main theorem of this paper follows:

Theorem 1 *Let us consider $Z(x)$ a Gaussian process with zero mean and covariance kernel $k(x,x') \in \mathcal{C}^0(\mathbb{R}^d \times \mathbb{R}^d)$ and $(x_i)_{i=1,\dots,n}$ an experimental design set of n independent random points sampled with the probability measure μ on $\mathbb{R}^d$. We assume that $\sup_{x\in\mathbb{R}^d} k(x,x) < \infty$. According to Mercer's theorem (Mercer 1909), we have the following representation of $k(x,x')$:*

$$k(x,x') = \sum_{p\geq 0} \lambda_p \phi_p(x)\phi_p(x'), \tag{6}$$

where $(\phi_p(x))_p$ is an orthonormal basis of $L^2_\mu(\mathbb{R}^d)$ (denoting the set of square integrable functions) consisting of eigenfunctions of $(T_{\mu,k} f)(x) = \int_{\mathbb{R}^d} k(x,x') f(x') d\mu(x')$ and λ_p is the nonnegative sequence of corresponding eigenvalues sorted in decreasing order. Then, for a non-degenerate kernel—i.e. when $\lambda_p > 0, \forall p > 0$—we have the following convergence in probability for the MSE (3) of the BLUP:

$$\sigma^2(x) \overset{n\to\infty}{\longrightarrow} \sum_{p\geq 0} \frac{\tau\lambda_p}{\tau+\lambda_p}\phi_p(x)^2. \tag{7}$$

For degenerate kernels—i.e. when only a finite number of λ_p are not zero—the convergence is almost sure. We note that we have the convergences with respect to the distribution of the points $(x_i)_{i=1,\dots,n}$ of the experimental design set.

The proof of Theorem 1 is given in Sect. 4.

Remark For non-degenerate kernels such that $||\phi_p(x)||_{L^\infty} < \infty$ uniformly in p, the convergence is almost sure. Some kernels such as the one of the Brownian motion satisfy this property.

The following theorem gives the asymptotic value of the learning curve when n is large.

Theorem 2 *Let us consider $Z(x)$ a Gaussian process with known mean and covariance kernel $k(x,x') \in \mathcal{C}^0(\mathbb{R}^d \times \mathbb{R}^d)$ such that $\sup_{x\in\mathbb{R}^d} k(x,x) < \infty$ and $(x_i)_{i=1,\dots,n}$ an experimental design set of n independent random points sampled with the probability measure μ on $\mathbb{R}^d$. Then, for a non-degenerate kernel, we have the following convergence in probability:*

$$IMSE \overset{n\to\infty}{\longrightarrow} \sum_{p\geq 0} \frac{\tau\lambda_p}{\tau+\lambda_p}. \tag{8}$$

For degenerate kernels, the convergence is almost sure.

Proof From Theorem 1 and the orthonormal property of the basis $(\phi_p(x))_p$ in $L^2_\mu(\mathbb{R})$, the proof of the theorem is straightforward by integration. We note that we can permute the integral and the limit thanks to the dominated convergence theorem since $\sigma^2(x) \leq k(x,x)$. □

A strength of Theorem 2 is that it allows for obtaining the rate of convergence of the learning curve even when the eigenvalues $(\lambda_p)_{p\geq 0}$ are not explicit. Indeed, as presented in Sect. 5.2, this rate can be deduced from the asymptotic behavior of λ_p for large p. Furthermore, this asymptotic behavior is known for usual kernels (fractional Brownian kernel, Matérn covariance kernel, Gaussian covariance kernel, …). However, this is also a limitation since it could be unknown for general covariance kernels.

3.1 Discussion

The limit obtained is identical to the one presented in (Rasmussen and Williams 2006) Sect. 7.3 Eq. (7.26) for a degenerate kernel. Furthermore, the limit in Eq. (8) corresponds to the average bound given for degenerate kernels in (Opper and Vivarelli 1999) in Sect. 6 Eq. (17) with the correspondence $\tau = \sigma^2/n$. In particular, they prove that it is a lower bound for the generalization error and an upper bound for the training error. The training error is defined as the empirical mean $\sum_{i=1}^{n} \sigma^2(x_i)/n$ where $(x_i)_{i=1,\dots,n}$ are the design points. They also note that this bound should be exact for the asymptotic n large since the sum $\sum_{i=1}^{n} \sigma^2(x_i)/n$ approaches to the IMSE asymptotically. Moreover, they numerically observed that this bound is relevant for a Gaussian covariance kernel (Opper and Vivarelli 1999), Eq. (18) which is a non-degenerate kernel. The work of Opper and Vivarelli is also investigated in (Williams and Vivarelli 2000; Sollich and Halees 2002; Picheny 2009). In particular, a proof of Theorem 1 is given for degenerate kernels and the relevance of the bound is illustrated on numerical examples using non-degenerate kernels [e.g. Gaussian covariance kernel and exponential kernel (Rasmussen and Williams 2006)].

We note that the proof of Theorem 1 for non-degenerate kernels is of interest since the usual kernels for Gaussian process regression are non-degenerate and we will exhibit dramatic differences between the learning curves of degenerate and non-degenerate kernels.

4 Proof of Theorem 1

We present in this section the proof of Theorem 1. The aim is to find the asymptotic value of the MSE $\sigma^2(x)$ (3) when n tends to the infinity. The principle of the proof is to find an upper bound and a lower bound for $\sigma^2(x)$ which converge to the same quantity. One of the main ideas of the proof is to use the fact that in a Gaussian process regression framework we consider the BLUP, i.e. the one which minimizes the MSE. Therefore, for a given Gaussian process modeling the function $f(x)$, any LUP has a larger MSE. Furthermore, to provide a lower bound for $\sigma^2(x)$, we use the result presented in Theorem 1 for degenerate kernels. Therefore, we start the proof by presenting the degenerate case.

4.1 The degenerate case

The proof in the degenerate case follows the lines of the ones given by (Opper and Vivarelli 1999; Rasmussen and Williams 2006; Picheny 2009). For a degenerate kernel, the number $\bar{p}$ of non-zero eigenvalues is finite. Let us denote $\Lambda = \mathrm{diag}(\lambda_i)_{1 \le i \le \bar{p}}$, $\phi(x) = (\phi_1(x), \dots, \phi_{\bar{p}}(x))$ and $\Phi = (\phi(x_1)^T, \dots, \phi(x_n)^T)^T$. The MSE of the Gaussian process regression (3) is given by:

$$\sigma^2_{\bar{p}}(x) = \phi(x)\Lambda\phi(x)^T - \phi(x)\Lambda\Phi^T \left(\Phi\Lambda\Phi^T + n\tau I\right)^{-1} \Phi\Lambda\phi(x)^T.$$

Thanks to the Woodbury–Sherman–Morrison formula[1] and according to (Opper and Vivarelli 1999; Picheny 2009) the Gaussian process regression error can be written:

$$\sigma^2_{\bar{p}}(x) = \phi(x) \left(\frac{\Phi^T\Phi}{n\tau} + \Lambda^{-1}\right)^{-1} \phi(x)^T.$$

[1] If B is a non-singular $p \times p$ matrix, C a non-singular $m \times m$ matrix and A a $m \times p$ matrix with m, $p < \infty$, then $(B + AC^{-1}A)^{-1} = B^{-1} - B^{-1}A(A^TB^{-1}A + C)^{-1}A^TB^{-1}$.

Since $\bar{p}$ is finite, by the strong law of large numbers, the $\bar{p} \times \bar{p}$ matrix $\Phi^T \Phi / n$ converges almost surely as $n \to \infty$. Therefore, we have the almost sure convergence:

$$\sigma^2_{\bar{p}}(x) \overset{n\to\infty}{\longrightarrow} \sum_{p \le \bar{p}} \frac{\tau \lambda_p}{\tau + \lambda_p} \phi_p(x)^2. \tag{9}$$

4.2 The lower bound for $\sigma^2(x)$

The objective is to find a lower bound for the MSE $\sigma^2(x)$ (3) for non-degenerate kernels.

If we denote by $a_i(x)$ the coefficients of the BLUP $\hat{f}(x)$ associated to $Z(x)$—i.e. $\hat{f}(x) = \sum_{i=1}^n a_i(x)(Z(x_i) + \varepsilon_i)$, the MSE can be written:

$$\sigma^2(x) = \mathbb{E}\left[\left(Z(x) - \sum_{i=1}^n a_i(x)(Z(x_i) + \varepsilon_i)\right)^2\right].$$

Let us consider the Karhunen-Loève decomposition of $Z(x) = \sum_{p\ge 0} Z_p \sqrt{\lambda_p}\phi_p(x)$ where $(Z_p)_p$ is a sequence of independent Gaussian random variables with mean zero and variance 1 and $\lambda_p > 0$ for all $p \in \mathbb{N}^*$. Therefore, we have the equalities $\mathbb{E}\left[Z_p\right] = 0$, $\mathbb{E}\left[Z_p^2\right] = 1$ and $\mathbb{E}\left[Z_p Z_q\right] = 0$ when $p \neq q$. Then, the MSE equals:

$$\begin{aligned}\sigma^2(x) &= \mathbb{E}\left[\left(\sum_{p\ge 0} \sqrt{\lambda_p}\left(\phi_p(x) - \sum_{i=1}^n a_i(x)\phi_p(x_i)\right) Z_p\right)^2\right] + n\tau \sum_{i=1}^n a_i(x)^2 \\ &= \sum_{p\ge 0} \lambda_p \left(\phi_p(x) - \sum_{i=1}^n a_i(x)\phi_p(x_i)\right)^2 + n\tau \sum_{i=1}^n a_i(x)^2.\end{aligned}$$

Then, for a fixed $\bar{p}$, the following inequality holds:

$$\sigma^2(x) \ge \sum_{p\le \bar{p}} \lambda_p \left(\phi_p(x) - \sum_{i=1}^n a_i(x)\phi_p(x_i)\right)^2 + n\tau \sum_{i=1}^n a_i(x)^2 = \sigma^2_{LUP,\bar{p}}(x). \tag{10}$$

$\sigma^2_{LUP,\bar{p}}(x)$ is the MSE of the LUP of coefficients $a_i(x)$ associated to the Gaussian process $Z_{\bar{p}}(x) = \sum_{p\le\bar{p}} Z_p \sqrt{\lambda_p}\phi_p(x)$ and noisy observations with variance $n\tau$. Let us consider $\sigma^2_{\bar{p}}(x)$ the MSE of the BLUP of $Z_{\bar{p}}(x)$, we have the following inequality:

$$\sigma^2_{LUP,\bar{p}}(x) \ge \sigma^2_{\bar{p}}(x). \tag{11}$$

Since $Z_{\bar{p}}(x)$ has a degenerate kernel, the almost sure convergence given in Eq. (9) holds for $\sigma^2_{\bar{p}}(x)$. Then, considering inequalities (10) and (11) and the convergence (9), we obtain: $\liminf_{n\to\infty} \sigma^2(x) \ge \sum_{p\le\bar{p}} \tau\lambda_p/\left(\tau + \lambda_p\right)\phi_p(x)^2$. Taking the limit $\bar{p} \to \infty$ gives the following lower bound:

$$\liminf_{n\to\infty} \sigma^2(x) \ge \sum_{p\ge 0} \frac{\tau\lambda_p}{\tau + \lambda_p} \phi_p(x)^2. \tag{12}$$

4.3 The upper bound for $\sigma^2(x)$

The objective is to find an upper bound for $\sigma^2(x)$. Since $\sigma^2(x)$ is the MSE of the BLUP associated to $Z(x)$, if we consider any other LUP associated to $Z(x)$, then the corresponding MSE denoted by $\sigma^2_{LUP}(x)$ satisfies the following inequality:

$$\sigma^2(x) \leq \sigma^2_{LUP}(x).$$

The idea is to find a LUP so that its MSE is a sharp upper bound of $\sigma^2(x)$. We consider the following LUP:

$$\hat{f}_{LUP}(x) = k(x)^T A z^n, \tag{13}$$

with A the $n \times n$ matrix defined by $A = L^{-1} + \sum_{k=1}^{q} (-1)^k (L^{-1}M)^k L^{-1}$ with $L = n\tau I + \sum_{p<p^*} \lambda_p [\phi_p(x_i)\phi_p(x_j)]_{1\leq i,j\leq n}$, $M = \sum_{p\geq p^*} \lambda_p [\phi_p(x_i)\phi_p(x_j)]_{1\leq i,j\leq n}$, q a finite integer and p^* such that $\lambda_{p^*} < \tau$.

The choice of the LUP (13) is motivated by the fact that the matrix A is an approximation of the inverse of the matrix $(n\tau I + K)$ that is tractable in the calculations. Indeed, we have $(n\tau I + K) = L + M$ and thus $(n\tau I + K)^{-1} = L^{-1}(I + L^{-1}M)^{-1}$. Then, the term $(I + L^{-1}M)^{-1}$ is approximated with the sum $\sum_{k=1}^{q} (-1)^k (L^{-1}M)^k$. We note that the condition p^* such that $\lambda_{p^*} < \tau$ is used to control the convergence of this sum when q tends to the infinity.

The MSE of the LUP (13) is given by:

$$\sigma^2_{LUP}(x) = k(x,x) - k(x)^T \left(2A - A(n\tau I + K)A\right) k(x),$$

and by substituting the expression of A into the previous equation we obtain:

$$\sigma^2_{LUP}(x) = k(x,x) - k(x)^T L^{-1} k(x) - \sum_{i=1}^{2q+1} (-1)^i k(x)^T (L^{-1}M)^i L^{-1} k(x). \tag{14}$$

The rest of the proof consists in finding the asymptotic values of the terms present in the expression of $\sigma^2_{LUP}(x)$.

First, we deal with the term $k(x)^T L^{-1} k(x)$ with the following lemma proved in Appendix.

Lemma 1 *Let us consider the term $k(x)^T L^{-1} k(x)$ in Eq. (14). The following convergence holds:*

$$k(x)^T L^{-1} k(x) \overset{n\to\infty}{\longrightarrow} \sum_{p\leq p^*} \frac{\lambda_p^2}{\lambda_p + \tau} \phi_p(x)^2 + \frac{1}{\tau} \sum_{p>p^*} \lambda_p^2 \phi_p(x)^2. \tag{15}$$

Second, let us consider the term $\sum_{i=1}^{2q+1} (-1)^i k(x)^T (L^{-1}M)^i L^{-1} k(x)$. We have the following equality:

$$\begin{aligned} k(x)^T (L^{-1}M)^i L^{-1} k(x) = {} & \sum_{j=0}^{i} \binom{i}{j} \tfrac{1}{n\tau} k(x)^T \left(\tfrac{M}{n\tau}\right)^j \left(-\tfrac{L'M}{(n\tau)^2}\right)^{i-j} k(x) \\ & - k(x)^T \left(\tfrac{M}{n\tau}\right)^j \left(-\tfrac{L'M}{(n\tau)^2}\right)^{i-j} \tfrac{L'}{(n\tau)^2} k(x), \end{aligned} \tag{16}$$

with $L' = \Phi_{p^*} \left(\frac{\Phi_{p^*}^T \Phi_{p^*}}{n\tau} + \Lambda^{-1} \right)^{-1} \Phi_{p^*}^T = \sum_{p,p'\leq p^*} d^{(n)}_{p,p'} [\phi_p(x_i)\phi_p(x_j)]_{1\leq i,j\leq n}$ and $d^{(n)}_{p,p'} = \left[\left(\frac{\Phi_{p^*}^T \Phi_{p^*}}{n\tau} + \Lambda^{-1} \right)^{-1} \right]_{p,p'}$. Since $q < \infty$, we can obtain the convergence in

probability of $\sum_{i=1}^{2q+1}(-1)^i k(x)^T (L^{-1}M)^i L^{-1}k(x)$ from the ones of:

$$k(x)^T \frac{1}{n}\left(\frac{M}{n}\right)^j \left(\frac{L'M}{n^2}\right)^{i-j} k(x), \tag{17}$$

and:

$$k(x)^T \left(\frac{M}{n}\right)^j \left(\frac{L'M}{n^2}\right)^{i-j} \frac{L'}{n^2} k(x), \tag{18}$$

with $i \leq 2q+1$ and $j \leq i$. We first study the convergence of the term (17) for $i < j$ and the term (18) for $i \leq j$. Then, we study the convergence of (17) for $i = j$. We have the following lemma proved in Appendix:

Lemma 2 *For $i < j$ we have the following convergence when $n \to \infty$:*

$$k(x)^T \frac{1}{n}\left(\frac{M}{n}\right)^j \left(\frac{L'M}{n^2}\right)^{i-j} k(x) \xrightarrow{\mathbb{P}_\mu} 0, \tag{19}$$

and for $i \leq j$ the following one holds:

$$k(x)^T \frac{1}{n}\left(\frac{M}{n}\right)^j \left(\frac{L'M}{n^2}\right)^{i-j} \frac{L'}{n^2} k(x) \xrightarrow{\mathbb{P}_\mu} 0. \tag{20}$$

We note that the convergences presented in Lemma 2 hold in probability. Then, we have the following lemma proved in Appendix:

Lemma 3 *The following convergence holds when $n \to \infty$:*

$$\frac{1}{n} k(x)^T \left(\frac{M}{n}\right)^i k(x) \xrightarrow{\mathbb{P}_\mu} \sum_{p>p^*} \lambda_p^{i+2} \phi_p(x)^2. \tag{21}$$

From the convergences (19) and (20) and thanks to the equality (16), we deduce the following convergence when $n \to \infty$:

$$k(x)^T \left(L^{-1}M\right)^i L^{-1}k(x) - \frac{1}{n\tau^{i+1}} k(x)^T \left(\frac{M}{n}\right)^i k(x) \xrightarrow{\mathbb{P}_\mu} 0.$$

Then, using the convergence (21) we obtain when $n \to \infty$:

$$k(x)^T (L^{-1}M)^i L^{-1}k(x) \xrightarrow{\mathbb{P}_\mu} \left(\frac{1}{\tau}\right)^{i+1} \sum_{p>p^*} \lambda_p^{i+2} \phi_p(x)^2. \tag{22}$$

From the Eq. (14) and the convergences (15) and (22), we obtain the following convergence when $n \to \infty$:

$$\sigma^2_{LUP}(x) \xrightarrow{\mathbb{P}_\mu} \sum_{p \leq p^*} \left(\lambda_p - \frac{\lambda_p^2}{\tau + \lambda_p}\right) \phi_p(x)^2 + \sum_{p>p^*} \lambda_p \phi_p(x)^2 + \sum_{p>p^*} \lambda_p \phi_p(x)^2 \sum_{i=0}^{2q+1} (-1)^{i+1} \left(\frac{1}{\tau}\right)^{i+1} \lambda_p^{i+1}.$$

From classical results about the sum of geometric series, we have:

$$\sigma^2_{LUP}(x) \xrightarrow{\mathbb{P}_\mu} \sum_{p\geq 0}\left(\lambda_p - \frac{\lambda_p^2}{\tau+\lambda_p}\right)\phi_p(x)^2 - \sum_{p>p^*}\lambda_p^2\frac{\left(\frac{\lambda_p}{\tau}\right)^{2q+1}}{\tau+\lambda_p}\phi_p(x)^2. \tag{23}$$

By considering the limit $q \to \infty$ and the inequality $\lambda_{p^*} < \tau$, we obtain the following upper bound for $\sigma^2(x)$:

$$\limsup_{n\to\infty}\sigma^2(x) \leq \sum_{p\geq 0}\frac{\tau\lambda_p}{\tau+\lambda_p}\phi_p(x)^2. \tag{24}$$

The result announced in Theorem 1 is deduced from the lower and upper bounds (12) and (24).

5 Examples of rates of convergence for the learning curve

5.1 Numerical study on the assumptions of Theorem 2

Theorem 2 gives the asymptotic value of the IMSE (5) when the number of observations n increases. The aim of this section is to determine when the assumptions of Theorem 2 hold—i.e. to find the critical number of observations n beyond which, for a given τ and a given covariance kernel k, the sum in (8) is a sharp approximation of the IMSE. To perform such a study, we consider a Brownian kernel $k(x, y) = x + y - |x - y|$ with $x, y \in [0, 1]$, $\tau \in \{0.001, 0.01, 0.1\}$ and a uniform measure μ on $[0, 1]$. The eigenvalues of k are the following ones (Bronski 2003):

$$\lambda_p = \frac{1}{(p+1/2)^2\pi^2}, \quad p \in \mathbb{N}.$$

Therefore, for a given τ, we can explicitly obtain the value of the sum presented in (8) and compare it with an empirical estimation of the IMSE. This empirical estimation is obtained by considering the MSE (3) built from n points randomly spread into the interval $[0, 1]$ and by estimating the integral (5) with a numerical integration. Furthermore, for each pair (τ, n) we repeat this procedure 100 times in order to obtain an empirical estimator and confidence intervals for the value of the IMSE. The results of this procedure are presented in Fig. 1.

Figure 1 represents the ratio between the value IMSE for a given n and the asymptotic value IMSE_∞ given by (8). For large n this ratio is close to one. This allows for representing the convergence of the IMSE to its asymptotic value.

We observe in Fig. 1 that the convergence is effective for $n < 100$ for all values of τ. The convergence is robust for small values of τ: the asymptotic value (8) is a good approximation of the IMSE if $n \geq 5$ for $\tau = 10^{-1}$, if $n \geq 20$ for $\tau = 10^{-2}$ and if $n \geq 60$ for $\tau = 10^{-3}$. This corresponds approximately to the threshold values $n\tau = 0.5$ for $\text{IMSE}_\infty = 0.1575$, $n\tau = 0.2$ for $\text{IMSE}_\infty = 0.05$ and $n\tau = 0.06$ for $\text{IMSE}_\infty = 0.0.0158$; or globally to $n\tau \approx 4\text{IMSE}_\infty$.

This highlights the relevance of the asymptotic value of the IMSE given in Theorem 2. However, in general we do not have an explicit expression for the eigenvalues of a covariance kernel. In this case, we can obtain the asymptotic expression of IMSE_∞ for small τ from the asymptotic behavior of the eigenvalues $(\lambda_p)_{p\geq 0}$ for large p. We deal with this issue in the next subsection.

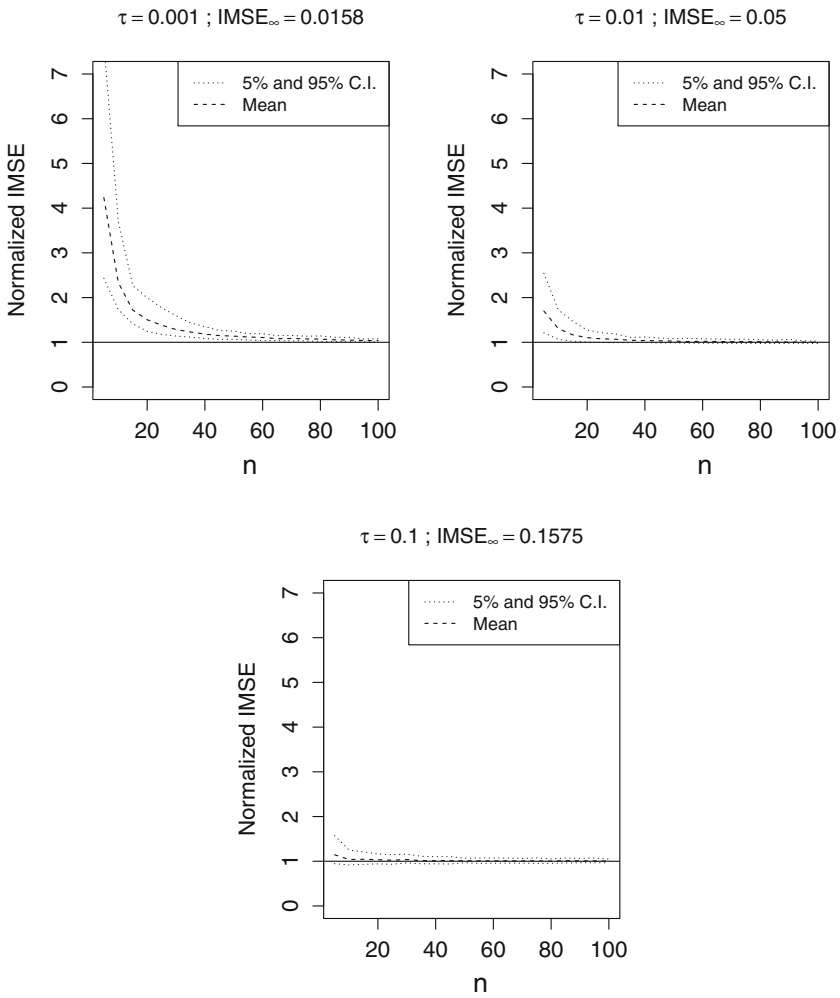

Fig. 1 Comparison between the IMSE for different n and the theoretical asymptotic value IMSE_∞ given by the sum (8). The ratio $\text{IMSE}/\text{IMSE}_\infty$ is plotted as a function of n for three values of τ. For each pair (τ, n) 100 approximations of IMSE are evaluated from design points randomly spread on [0, 1]. From them the empirical mean (represented by the *dashed lines*) and the 5 % and 95 % confidence intervals (represented by the *dotted lines*) of the ratio $\text{IMSE}/\text{IMSE}_\infty$ are evaluated

5.2 Rate of convergence for some usual kernels

Theorem 2 gives the asymptotic value of the generalization error as a function of the eigenvalues of the covariance kernel. However, this asymptotic value is hard to handle since the expression of the eigenvalues is rarely known. To deal with this problem, we introduce in Proposition 1 a quantity B_τ which has the same rate of convergence of the asymptotic value of the generalization error and which is tractable for our purpose.

Proposition 1 *Let us denote* $IMSE_\infty = \lim_{n\to\infty} IMSE$. *The following inequality holds:*

$$\frac{1}{2} B_\tau \leq IMSE_\infty \leq B_\tau, \tag{25}$$

with $B_\tau = \sum_{p \text{ s.t. } \lambda_p \leq \tau} \lambda_p + \tau \# \left\{ p \text{ s.t. } \lambda_p > \tau \right\}$.

Proof The proof is directly deduced from Theorem 2 and the following inequality:

$$\frac{1}{2} h_\tau(x) \leq \frac{x}{x+\tau} \leq h_\tau(x),$$

with:

$$h_\tau(x) = \begin{cases} x/\tau & x \leq \tau \\ 1 & x > \tau \end{cases}.$$

□

Proposition 1 shows that the rate of convergence of the generalization error IMSE_∞ as a function of τ is equivalent to the one of B_τ. In this section, we analyze the rate of convergence of IMSE_∞ (or equivalently B_τ) when τ is small.

In this section, we consider that the design measure μ is uniform on $[0, 1]^d$.

Example 2 (Degenerate kernels) For degenerate kernels we have $\# \left\{ p \text{ s.t. } \lambda_p > 0 \right\} < \infty$. Thus, when $\tau \to 0$, we have:

$$\sum_{p \text{ s.t. } \lambda_p < \tau} \lambda_p = 0,$$

from which:

$$B_\tau \propto \tau. \tag{26}$$

Therefore, the IMSE decreases with τ. We find here a classical result about Monte-Carlo convergence which gives that the variance decay is proportional to the observation noise variance ($n\tau$) divided by the number of observations (n) for any dimension. Nevertheless, for non-degenerate kernels, the number of non-zero eigenvalues is infinite and we are hence in an infinite-dimensional case (contrarily to the degenerate one). We see in the following examples that we do not conserve the usual Monte-Carlo convergence rate in this case which emphasizes the importance of Theorem 1 dealing with non-degenerate kernels.

Example 3 (The fractional Brownian motion) Let us consider the fractional Brownian kernel with Hurst parameter $H \in (0, 1)$:

$$k(x, y) = x^{2H} + y^{2H} - |x - y|^{2H}. \tag{27}$$

The associated Gaussian process—called fractional Brownian motion—is Hölder continuous with exponent $H - \varepsilon, \forall \varepsilon > 0$. According to (Bronski 2003), we have the following result:

Lemma 4 *The eigenvalues of the fractional Brownian motion with Hurst exponent* $H \in (0, 1)$ *satisfy the behavior*

$$\lambda_p = \frac{v_H}{p^{2H+1}} + o\left(p^{-\frac{(2H+2)(4H+3)}{4H+5} + \delta} \right), \quad p \gg 1,$$

where $\delta > 0$ *is arbitrary,* $v_H = \frac{\sin(\pi H)\Gamma(2H+1)}{\pi^{2H+1}}$, *and* Γ *is the Euler Gamma function.*

Therefore, when $\tau \ll 1$, we have:

$$\lambda_p < \tau \quad \text{if} \quad p > \left(\frac{\nu_H}{\tau}\right)^{\frac{1}{2H+1}}.$$

We hence have the following approximation for B_τ:

$$B_\tau \approx \sum_{p>\left(\frac{\nu_H}{\tau}\right)^{\frac{1}{2H+1}}} \frac{\nu_H}{p^{2H+1}} + \tau \left(\frac{\nu_H}{\tau}\right)^{\frac{1}{2H+1}}.$$

Furthermore, we have:

$$\sum_{p>\left(\frac{\nu_H}{\tau}\right)^{\frac{1}{2H+1}}} \frac{\nu_H}{p^{2H+1}} \approx \int_{\left(\frac{\nu_H}{\tau}\right)^{\frac{1}{2H+1}}}^{+\infty} \frac{\nu_H}{x^{2H+1}}\, dx = \frac{\nu_H}{2H\left(\frac{\nu_H}{\tau}\right)^{1-\frac{1}{2H+1}}},$$

from which:

$$B_\tau \approx C_H \tau^{1-\frac{1}{2H+1}}, \quad \tau \ll 1, \tag{28}$$

where C_H is a constant independent of τ.

The rate of convergence for a fractional Brownian motion with Hurst parameter H is $\tau^{1-\frac{1}{2H+1}}$. We note that the case $H = 1/2$ corresponds to the classical Brownian motion. We observe that the larger the Hurst parameter is (i.e. the more regular the Gaussian process is), the faster the convergence is. Furthermore, for $H \to 1$ the convergence rate gets close to $\tau^{2/3}$. Therefore, even for the most regular fractional Brownian motion, we are still far from the classical Monte-Carlo convergence rate.

Example 4 (The 1-D Matérn covariance kernel) In this example we deal with the Matérn kernel with regularity parameter $\nu > 0$ in dimension 1:

$$k_{1D}(x, x'; \nu, l) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}|x - x'|}{l}\right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}|x - x'|}{l}\right), \tag{29}$$

where K_ν is the modified Bessel function (Abramowitz and Stegun 1965). The eigenvalues of this kernel satisfy the following asymptotic behavior (Nazarov and Nikitin 2004):

$$\lambda_p \approx \frac{1}{p^{2(\nu+1/2)}}, \quad p \gg 1.$$

Following the guideline of the Example 3 we deduce the following asymptotic behavior for B_τ:

$$B_\tau \approx C_\nu \tau^{1-\frac{1}{2(\nu+1/2)}}, \quad \tau \ll 1, \tag{30}$$

where C_ν is a constant independent of τ.

This result is in agreement with the one of Ritter (2000a) who proved that for 1-dimensional kernels satisfying the Sacks–Ylvisaker of order r conditions (where r is an integer), the generalization error for the best linear estimator and experimental design set strategy decays as $\tau^{1-\frac{1}{2r+2}}$. Indeed, for such kernels, the eigenvalues satisfy the large-p behavior $\lambda_p \propto 1/p^{2r+2}$ (Rasmussen and Williams 2006) and by following the guideline of the previous examples we find the same convergence rate. We note that the Matérn kernel with parameter $\nu = r + 1/2$ satisfies the Sacks–Ylvisaker of order r conditions.

Example 5 (The d-D tensorial Matérn covariance kernel) We focus here on the d-dimensional tensorial Matérn kernel with isotropic regularity parameter $\nu > \frac{1}{2}$. According to Pusev (2011) the eigenvalues of this kernel satisfy the asymptotics:

$$\lambda_p \approx \phi(p), \quad p \gg 1,$$

where the function ϕ is defined by:

$$\phi(p) = \frac{\log(1+p)^{2(d-1)(\nu+1/2)}}{p^{2(\nu+1/2)}}.$$

Its inverse ϕ^{-1} satisfies:

$$\phi^{-1}(\varepsilon) = \varepsilon^{-\frac{1}{2(\nu+1/2)}} \left(\log\left(\varepsilon^{-\frac{1}{2(\nu+1/2)}}\right)\right)^{d-1} (1+o(1)), \quad \varepsilon \ll 1.$$

We hence have the approximation:

$$B_\tau \approx \frac{2(\nu+1/2)-1}{\phi^{-1}(\tau)^{2(\nu+1/2)-1}} \log\left(1+\phi^{-1}(\tau)\right)^{2(d-1)(\nu+1/2)} + \tau\phi^{-1}(\tau).$$

We can deduce the following rate of convergence for B_τ:

$$B_\tau \approx C_{(\nu+1/2),d}\, \tau^{1-\frac{1}{2(\nu+1/2)}} \log(1/\tau)^{d-1}, \quad \tau \ll 1, \tag{31}$$

with $C_{\nu,d}$ a constant independent of τ.

Example 6 (The d-D Gaussian covariance kernel) According to Todor, 2006 the asymptotic behavior of the eigenvalues for a Gaussian kernel is:

$$\lambda_p \lesssim \exp\left(-p^{\frac{1}{d}}\right).$$

Applying the procedure presented in the previous examples, it can be shown than the rate of convergence of the IMSE is bounded by:

$$C_d \tau \log(1/\tau)^d, \quad \tau \ll 1, \tag{32}$$

with C_d a constant independent of τ.

Remark We can see from the previous examples that for smooth kernels, the convergence rate is close to τ, i.e. the classical Monte-Carlo rate.

5.3 Numerical examples

We compare the previous theoretical results on the rate of convergence of the generalization error with full numerical simulations. In order to observe the asymptotic convergence, we fix $n = 200$ and we consider $1/\tau$ varying from 50 to 1000. The experimental design sets are sampled from a uniform measure on $[0, 1]$ and the observation noise is $n\tau$. To estimate the IMSE (5) we use a trapezoidal numerical integration with 4000 quadrature points over $[0, 1]$. Furthermore, to build the convergence curves in Figs. 2 and 3 we use a linear regression with the first value of the IMSE, an intercept fixed to zero (since the IMSE tends to 0 when τ tends to 0) and a unique explanatory variable corresponding to the tested convergence (e.g. $\tau^{0.1}$, $\tau\log(1/\tau)$, ...).

First, we deal with the 1-D fractional Brownian kernel (27) with Hurst parameter H. We have proved that for large n, the IMSE decays as $\tau^{1-\frac{1}{2H+1}}$. Figure 2 compares the numerically estimated convergences to the theoretical ones.

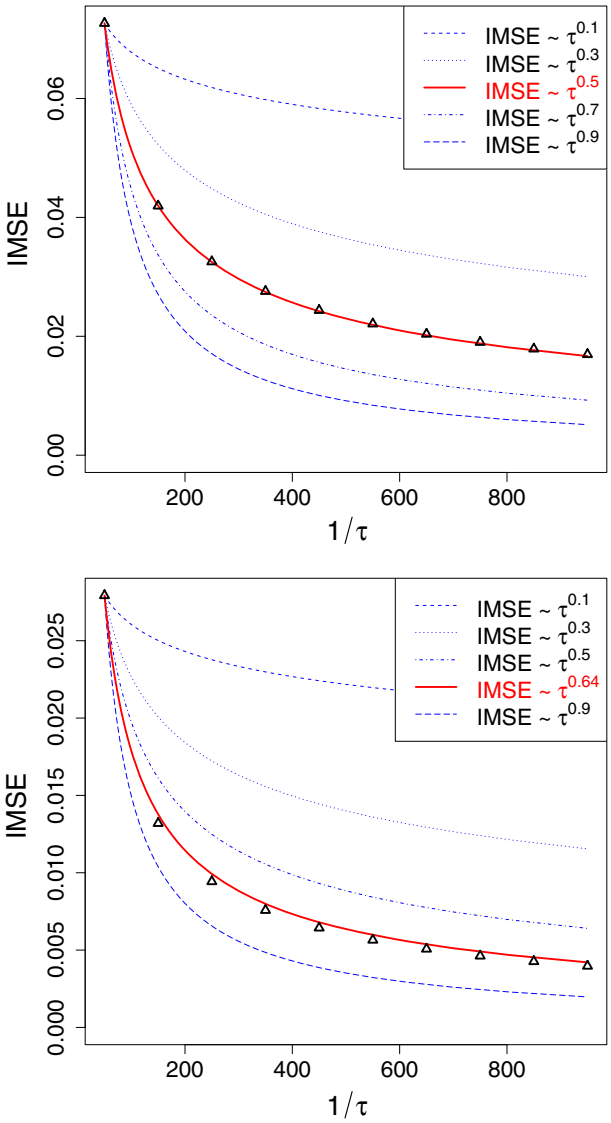

Fig. 2 Rate of convergence of the IMSE when the level of observation noise decreases for a fractional Brownian motion with Hurst parameter $H = 0.5$ (*left*) and $H = 0.9$ (*right*). The number of observations is $n = 200$ and the observation noise variance is $n\tau$ with $1/\tau$ varying from 50 to 1000. The triangles represent the numerically estimated IMSE, the *solid line* represents the theoretical convergence, and the other *non-solid lines* represent various convergence rates

We see in Fig. 2 that the observed rate of convergence is perfectly fitted by the theoretical one. We note that we are far from the classical Monte-Carlo rate since we are not in a non-degenerate case.

Finally, we deal with the 2-D tensorial Matérn-5/2 kernel and the 1-D Gaussian kernel. The 1-dimensional Matérn-ν class of covariance functions $k_{1D}(t, t'; \nu, \theta)$ is given by (29) and the 2-D tensorial Matérn-ν covariance function is given by:

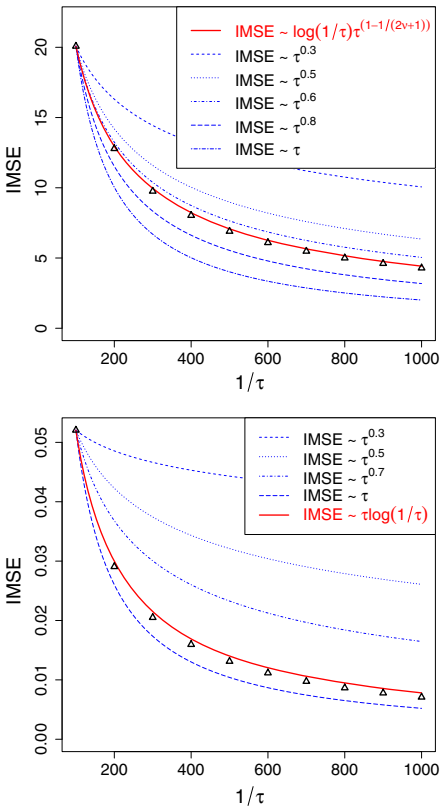

Fig. 3 Rate of convergence of the IMSE when the level of observation noise decreases for a 2-D tensorial Matérn-5/2 kernel on the left hand side and for a 1-D Gaussian kernel on the right hand side. The number of observations is $n = 200$ and the observation noise variance is $n\tau$ with $1/\tau$ varying from 100 to 1000. The *triangles* represent the numerically estimated IMSE, the *solid line* represents the theoretical convergence, and the other *non-solid lines* represent various convergences

$$k(x, x'; \nu, \theta) = k_{1D}(x_1, x_1'; \nu, \theta_1) k_{1D}(x_2, x_2'; \nu, \theta_2). \tag{33}$$

Furthermore, the 1-D Gaussian kernel is defined by:

$$k(x, x'; \theta) = \exp\left(-\frac{1}{2}\frac{(x - x')^2}{\theta^2}\right).$$

Figure 3 compares the numerically observed convergence of the IMSE to the theoretical one when $\theta_1 = \theta_2 = 0.2$ for the Matérn-5/2 kernel and when $\theta = 0.2$ for the Gaussian kernel. We see in Fig. 3 that the theoretical rate of convergence is a sharp approximation of the observed one.

6 Applications of the learning curve

Let us consider that we want to approximate the function $x \in \mathbb{R}^d \to f(x)$ from noisy observations at fixed points $(x_i)_{i=1,\ldots,n}$, with $n \gg 1$, sampled from the design measure μ and with s replications at each point x_i. In Sect. 6.1 we present how to determine the needed budget $T = ns$ to achieve a prescribed precision. Then, in Sect. 6.2, we illustrate this method on an industrial example.

6.1 Estimation of the budget required to reach a prescribed precision

Let us consider a prescribed generalization error denoted by $\bar{\varepsilon}$. The purpose of this subsection is to determine from an initial budget T_0 the budget T for which the generalization error reaches the value $\bar{\varepsilon}$.

First, we build an initial experimental design set $(x_i^{\text{train}})_{i=1,\dots,n}$ sampled with respect to the design measure μ and with s^* replications at each point such that $T_0 = ns^*$. From the s^* replications $(z_{i,j})_{j=1,\dots,s^*}$, we can estimate the observation noise variances σ_ε^2 with a classical empirical estimator: $\bar{\sigma}_\varepsilon^2 = \sum_{i=1}^n \sum_{j=1}^{s^*} (z_{i,j} - z_i^n)^2/(n(s^*-1))$, $z_i^n = \sum_{j=1}^{s^*} z_{i,j}/s^*$.

Second, we use the observations $z_i^n = (\sum_{j=1}^{s^*} z_{i,j})/s^*$ to estimate the covariance kernel $k(x, x')$. In practice, we consider a parametrized family of covariance kernels and we select the parameters which maximize the likelihood (Stein 1999).

Third, from Theorem 2 we can get the expression of the generalization error decay with respect to T (denoted by IMSE_T). Therefore, we just have to determine the budget T such that $\text{IMSE}_T = \bar{\boldsymbol{\varepsilon}}$. In practice, we will not use Theorem 2 but the asymptotic results described in Sect. 5.2.

This strategy is applied to an industrial case in Sect. 6.2. We note that in the application presented in Sect. 6.2, we have $s^* = 1$. In fact, in this example the observations are themselves obtained by an empirical mean of a Monte-Carlo sample and thus the noise variance can be estimated without processing replications.

6.2 Industrial case: MORET code

We illustrate in this section an industrial application of our results about the rate of convergence of the IMSE.

6.2.1 Data presentation

We use in this section the notation presented in Sect. 2.2. The outputs of the MORET code at point x_i are denoted by $Y_j(x_i)$ where $j = 1, \dots, s_i$ and $i = 1, \dots, n$.

A large data base $(Y_j(x_i))_{i=1,\dots,5625,\, j=1,\dots,200}$ is available to us. We divide it into a training set and a test set. The 5625 points x_i of the data base come from a 75×75 grid over $[0, 1]^2$. The training set consists of $n = 100$ points $(x_i^{\text{train}})_{i=1,\dots,n}$ extracted from the complete data base using a Latin Hypercube Sample (Fang et al. 2006) optimized with respect to the maximin criterion and of the first observations $(Y_1(x_i^{\text{train}}))_{i=1,\dots,100}$. We note that the maximin criterion aims to maximize the minimal distance (with respect to the L_2-norm) between the points of the design. We will use the other 5525 points as a test set.

The aim of the study is—given the training set—to predict the budget needed to achieve a prescribed precision for the surrogate model.

Furthermore, the observation noise variance σ_ε^2 is estimated by $\bar{\sigma}_\varepsilon^2 = 3.3 \times 10^{-3}$ (see Sect. 6.1).

6.2.2 Model selection

To build the model, we consider the training set plotted in Fig. 4. It is composed of the $n = 100$ points $(x_i^{\text{train}})_{i=1,\dots,n}$ which are uniformly spread on $Q = [0, 1]^2$.

Let us suppose that the response is a realization of a Gaussian process with a tensorial Matérn-ν covariance function. The 2-D tensorial Matérn-ν covariance function $k(x, x'; \nu, \theta)$

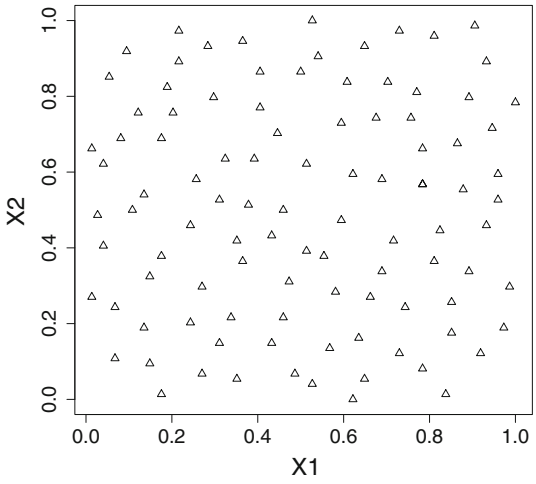

Fig. 4 Initial experimental design set with $n = 100$

is given in (33). The hyper-parameters are estimated by maximizing the concentrated Maximum Likelihood (Stein 1999):

$$-\frac{1}{2}(z-m)^T(\sigma^2 K + \sigma_\varepsilon^2 I)^{-1}(z-m) - \frac{1}{2}\det(\sigma^2 K + \bar{\sigma}_\varepsilon^2 I),$$

where $K = [k(x_i^{\text{train}}, x_j^{\text{train}}; \nu, \theta)]_{i,j=1,\ldots,n}$, I is the identity matrix, σ^2 the variance parameter, m the mean of $k_{\text{eff},s}(x)$ and $z = (Y_1(x_1^{\text{train}}), \ldots, Y_1(x_n^{\text{train}}))$ the observations at points in the training set. The mean of $k_{\text{eff},s}(x)$ is estimated by $m = \frac{1}{100}\sum_{i=1}^{100} Y_1(x_i^{\text{train}}) = 0.65$.

Due to the fact that the convergence rate is strongly dependent of the regularity parameter ν, we have to perform a good estimation of this hyper-parameter to evaluate the model error decay accurately. Note that we cannot have a closed form expression for the estimator of σ^2, it hence has to be estimated jointly with θ and ν.

Let us consider the vector of parameters $\phi = (\nu, \theta_1, \theta_2, \sigma^2)$. In order to perform the maximization, we have first randomly generated a set of 10,000 parameters $(\phi_k)_{k=1,\ldots,10^4}$ on the domain $[0.5, 3] \times [0.01, 2] \times [0.01, 2] \times [0.01, 1]$. We have then selected the 150 best parameters (i.e. the ones maximizing the concentrated Maximum Likelihood) and we have started a quasi-Newton based maximization from these parameters. More specifically, we have used the BFGS method (Shanno 1970). Finally, from the results of the 150 maximization procedures, we have selected the best parameter. We note that the quasi-Newton based maximizations have all converged to two parameter values, around 30 % to the actual maximum and 70 % to another local maximum.

The estimation of the hyper-parameters are $\nu = 1.31$, $\theta_1 = 0.67$, $\theta_2 = 0.45$ and $\sigma^2 = 0.24$. This means that we have a rough surrogate model which is not differentiable and α-Hölder continuous with exponent $\alpha = 0.81$. The variance of the observations is $\bar{\sigma}_\varepsilon^2 = 3.3 \times 10^{-3}$, using the same notations as Sect. 2.3, we have $\tau = \bar{\sigma}_\varepsilon^2 / T_0$ with $T_0 = n$ (it corresponds to $s = 1$).

The IMSE of the Gaussian process regression is $\text{IMSE}_{T_0} = 1.0 \times 10^{-3}$ and its empirical mean squared error is $\text{EMSE}_{T_0} = 1.2 \times 10^{-3}$. To compute the empirical mean squared error (EMSE), we use the observations $(Y_j(x_i))_{i=1,\ldots,5525,\ j=1\ldots,200}$ with $x_i \neq x_k^{\text{train}}\ \forall k = 1, \ldots, 100$, $i = 1, \ldots, 5525$ and to compute the IMSE (5) (that depends only on the positions of the training set and on the selected hyper-parameters) we use a trapezoidal numerical

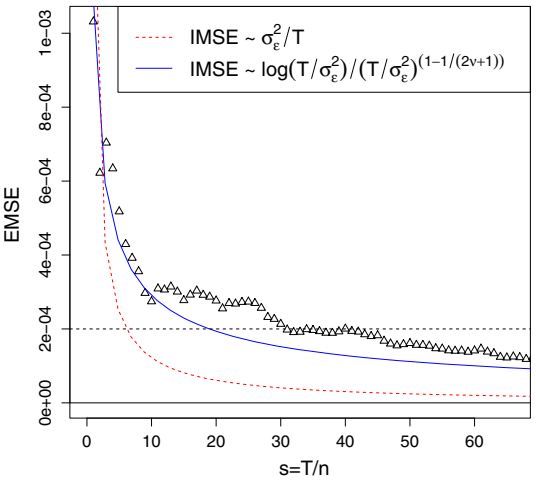

Fig. 5 Comparison between empirical mean squared error (EMSE) decay and theoretical IMSE decay for $n = 100$ when the total budget $T = ns$ increases. The *triangles* represent the EMSE, the *solid line* represents the theoretical decay, the *horizontal dashed line* represents the desired accuracy and the *dashed line* the classical M-C convergence. We see that Monte-Carlo decay does not match the empirical MSE and it is too fast

integration into a 75×75 grid over $[0, 1]^2$. For $s = 200$, the observation variance of the output $k_{\text{eff},s}(x)$ equals $\bar{\sigma}_\varepsilon^2/200 = 1.64 \times 10^{-5}$ and is neglected for the estimation of the empirical error. We can see that the IMSE is close to the empirical MSE which means that our model describes the observations accurately.

6.2.3 Convergence of the IMSE

According to (31), we have the following convergence rate for the IMSE:

$$\text{IMSE} \sim \log(1/\tau)\tau^{1-\frac{1}{2(\nu+1/2)}} = \frac{\log(T/\bar{\sigma}_\varepsilon^2)}{(T/\bar{\sigma}_\varepsilon^2)^{1-\frac{1}{2(\nu+1/2)}}}, \tag{34}$$

where the model parameter ν plays a crucial role. We can therefore expect that the IMSE decays as (see Sect. 6.1):

$$\text{IMSE}_T = \text{IMSE}_{T_0} \frac{\log(T/\bar{\sigma}_\varepsilon^2)}{(T/\bar{\sigma}_\varepsilon^2)^{1-\frac{1}{2(\nu+1/2)}}} \Big/ \frac{\log(T_0/\bar{\sigma}_\varepsilon^2)}{(T_0/\bar{\sigma}_\varepsilon^2)^{1-\frac{1}{2(\nu+1/2)}}}. \tag{35}$$

Let us assume that we want to reach an IMSE of $\bar{\boldsymbol{\varepsilon}} = 2.0 \times 10^{-4}$. According to the IMSE decay and the fact that the IMSE for the budget T_0 has been estimated to be equal to 1.0×10^{-3}, the total budget required is $T = ns = 2000$, i.e. $s = 20$. Figure 5 compares the empirical mean squared error convergence and the predicted convergence (35) of the IMSE.

We see empirically that the EMSE of $\bar{\boldsymbol{\varepsilon}} = 2.0 \times 10^{-4}$ is achieved for $s = 31$. This shows that the predicted IMSE and the empirical MSE are close and that the selected kernel captures the regularity of the response accurately.

Let us consider the classical Monte-Carlo convergence rate $\bar{\sigma}_\varepsilon^2/T$, which corresponds to the convergence rate of degenerate kernels, i.e. in the finite-dimensional case. Figure 5 compares the theoretical rate of convergence of the IMSE with the classical Monte-Carlo one. We see that the Monte-Carlo decay is too fast and does not represent correctly the empirical MSE decay. If we had considered the rate of convergence IMSE $\sim \bar{\sigma}_\varepsilon^2/T$, we would have reached an IMSE of $\bar{\boldsymbol{\varepsilon}} = 2.0 \times 10^{-4}$ for $s = 6$ (which is very far from the observed value $s = 31$).

7 Conclusion

The main result of this paper is the proof of a theorem giving the Gaussian process regression MSE when the number of observations is large and the observation noise variance is proportional to the number of observations. The proof generalizes previous ones which prove this result in dimension one or two or for a restricted class of covariance kernels (for degenerate ones).

A first limitation of the presented results is that the noise variance generally does not depend on the number of observations. The additive dependence of the noise variance in the number of observations is a technical assumption which allows for controlling the convergence of the learning curve. However, it is natural in the framework of experiments with replications or Monte-Carlo simulators. Deriving the presented results for the case of constant noise is still an open problem and is of great practical interest.

The asymptotic value of the MSE is derived in terms of the eigenvalues and eigenfunctions of the covariance function and holds for degenerate and non-degenerate kernels and for any dimension. From this theorem, we can deduce the asymptotic behavior of the generalization error—defined in this paper as the IMSE—as a function of the reduced observation noise variance (it corresponds to the noise variance when the number of observations equals one). A strength of this theorem is that the rate of convergence of the generalization error can be deduced from the one of the eigenvalues which is known for usual covariance kernels. The relevance of this rate of convergence is emphasized on a numerical study for different kernels. However, this leads to another limitation since the presented results cannot be used for general covariance kernels for which the eigenvalue decay rate is unknown.

The significant differences between the rate of convergence of degenerate and non-degenerate kernels highlight the importance to prove this result for non-degenerate kernels. This is especially important as usual kernels for Gaussian process regression are non-degenerate.

Finally, for practical perspectives, the presented method allows for evaluating the computational budget required to reach a given accuracy. It has been successfully applied to a real-word problem about the safety assessment of a nuclear system. However, it is efficient for specific applications (e.g. stochastic simulators with a constant observation noise variance) and when the computational budget is important. More investigations have to be performed to deal with the cases of heterogeneous noise, noise-free simulators or for very limited computational budget.

Acknowledgments The authors are grateful to Dr. Yann Richet of the IRSN—Institute for Radiological Protection and Nuclear Safety—for providing the data for the industrial case through the reDICE project.

Appendix: Proofs of the technical lemmas

Proof of Lemma 1

Let us consider the term $k(x)^T L^{-1} k(x)$. Since $p^* < \infty$, the matrix L can be written:

$$L = n\tau I + \Phi_{p^*} \Lambda \Phi_{p^*}^T, \tag{36}$$

where $\Lambda = \mathrm{diag}(\lambda_i)_{1 \le i \le p^*}$, $\Phi_{p^*} = \left(\phi(x_1)^T \ \ldots \ \phi(x_n)^T \right)^T$ and $\phi(x) = (\phi_1(x), \ldots, \phi_{p^*}(x))$. Thanks to the Woodbury–Sherman–Morrison formula, the matrix L^{-1} is given by:

$$L^{-1} = \frac{I}{n\tau} - \frac{\Phi_{p^*}}{n\tau}\left(\frac{\Phi_{p^*}^T\Phi_{p^*}}{n\tau} + \Lambda^{-1}\right)^{-1}\frac{\Phi_{p^*}^T}{n\tau}. \tag{37}$$

From the continuity of the inverse operator for invertible $p^* \times p^*$ matrices and by applying the strong law of large numbers, we obtain the following almost sure convergence:

$$\begin{aligned} k(x)^T L^{-1} k(x) = {} & \frac{1}{n\tau}\sum_{i=1}^{n} k(x,x_i)^2 - \frac{1}{\tau^2}\sum_{p,q=0}^{p^*}\left[\left(\frac{\Phi_{p^*}^T\Phi_{p^*}}{n\tau} + \Lambda^{-1}\right)^{-1}\right]_{p,q} \\ & \times \left[\frac{1}{n}\sum_{i=1}^{n} k(x,x_i)\phi_p(x_i)\right]\left[\frac{1}{n}\sum_{j=1}^{n} k(x,x_j)\phi_q(x_j)\right] \\ & \xrightarrow{n\to\infty} \frac{1}{\tau}\mathbb{E}_\mu[k(x,X)^2] - \frac{1}{\tau^2}\sum_{p,q=0}^{p^*}\left[\left(\frac{I}{\tau} + \Lambda^{-1}\right)^{-1}\right]_{p,q} \\ & \times \mathbb{E}_\mu[k(x,X)\phi_p(X)]\mathbb{E}_\mu[k(x,X)\phi_q(X)], \end{aligned}$$

where $\mathbb{E}_\mu$ is the expectation with respect to the design measure μ. We note that we can use the Woodbury–Sherman–Morrison formula and the strong law of large numbers since p^* is finite and independent of n. Then, the orthonormal property of the basis $(\phi_p(x))_{p\geq 0}$ implies:

$$\mathbb{E}_\mu[k(x,X)^2] = \sum_{p\geq 0}\lambda_p^2\phi_p(x)^2, \quad \mathbb{E}_\mu[k(x,X)\phi_p(X)] = \lambda_p\phi_p(x).$$

Therefore, we have the following almost sure convergence:

$$k(x)^T L^{-1} k(x) \xrightarrow{n\to\infty} \sum_{p\leq p^*}\frac{\lambda_p^2}{\lambda_p+\tau}\phi_p(x)^2 + \frac{1}{\tau}\sum_{p>p^*}\lambda_p^2\phi_p(x)^2.$$

Proof of Lemma 2

Let us consider $k(x)^T\frac{1}{n}\left(\frac{M}{n}\right)^j\left(\frac{L'M}{n^2}\right)^{i-j}k(x)$ and $i > j$, we have:

$$k(x)^T\frac{1}{n}\left(\frac{M}{n}\right)^j\left(\frac{L'M}{n^2}\right)^{i-j}k(x) = \sum_{\substack{p_1,\ldots,p_{i-j}\leq p^* \\ p_1',\ldots,p_{i-j}'\leq p^*}} d^{(n)}_{p_1,p_1'}\ldots d^{(n)}_{p_{i-j},p_{i-j}'}\sum_{\substack{q_1,\ldots,q_{i-j}>p^* \\ m_1,\ldots,m_j>p^*}} S^{(n)}_{q,m}, \tag{38}$$

with:

$$\begin{aligned} S^{(n)}_{q,m} = {} & \left(\frac{\sqrt{\lambda_{m_1}}}{n}\sum_{r=1}^{n} k(x,x_r)\phi_{m_1}(x_r)\right)\left(\frac{\sqrt{\lambda_{m_j}}}{n}\sum_{r=1}^{n}\phi_{m_j}(x_r)\phi_{p_1'}(x_r)\right) \\ & \times\left(\frac{\lambda_{q_{i-j}}}{n}\sum_{r=1}^{n} k(x,x_r)\phi_{q_{i-j}}(x_r)\sum_{r=1}^{n}\phi_{p_{i-j}}(x_r)\phi_{q_{i-j}}(x_r)\right) \\ & \times\prod_{l=1}^{j-1}\frac{\sqrt{\lambda_{m_l}\lambda_{m_{l+1}}}}{n}\sum_{r=1}^{n}\phi_{m_l}(x_r)\phi_{m_{l+1}}(x_r) \end{aligned}$$

$$\times \prod_{l=1}^{i-j-1} \frac{\lambda_{q_l}}{n} \sum_{r=1}^{n} \phi_{q_l}(x_r)\phi_{p_{l+1}}(x_r) \sum_{r=1}^{n} \phi_{q_l}(x_r)\phi_{p'_l}(x_r).$$

We consider now the term:

$$a^{(n)}_{q,p,p'} = \frac{\lambda_q}{n} \sum_{r=1}^{n} \phi_q(x_r)\phi_p(x_r) \frac{1}{n} \sum_{r=1}^{n} \phi_{p'}(x_r)\phi_q(x_r), \tag{39}$$

with $p, p' \leq p^*$. From Cauchy Schwarz inequality and thanks to the following inequality:

$$|\phi_p(x)|^2 \leq \frac{1}{\lambda_p} \sum_{p' \geq 0} \lambda_{p'} |\phi_{p'}(x)|^2 = \lambda_p^{-1} k(x,x),$$

we obtain (using $\lambda_p \geq \lambda_{p^*}, \forall p \leq p^*$ and $[\sum_{r=1}^n |\phi_q(x_r)|]^2 \leq n \sum_{r=1}^n \phi_q(x_r)^2$):

$$\left| a^{(n)}_{q,p,p'} \right| \leq \sigma^2 \lambda_{p^*}^{-1} \frac{\lambda_q}{n} \sum_{r=1}^{n} \phi_q(x_r)^2, \quad \forall p, p' \leq p^*,$$

with $\sigma^2 = \sup_x k(x,x)$. Considering the expectation with respect to the distribution of points x_r, we obtain $\forall \bar{p} < \infty$:

$$\mathbb{E}_\mu \left[\sum_{q > \bar{p}} \left| a^{(n)}_{q,p,p'} \right| \right] \leq \sigma^2 \lambda_{p^*}^{-1} \sum_{q > \bar{p}} \lambda_q.$$

From Markov inequality, $\forall \delta > 0$, we have:

$$\mathbb{P}_\mu \left(\left| \sum_{q > \bar{p}} a^{(n)}_{q,p,p'} \right| > \delta \right) \leq \frac{\mathbb{E}_\mu \left[\left| \sum_{q > \bar{p}} a^{(n)}_{q,p,p'} \right| \right]}{\delta} \leq \frac{\sigma^2 \lambda_{p^*}^{-1} \sum_{q > \bar{p}} \lambda_q}{\delta}. \tag{40}$$

Furthermore, $\forall \delta > 0, \forall \bar{p} > p^*$:

$$\mathbb{P}_\mu \left(\left| \sum_{q > p^*} a^{(n)}_{q,p,p'} \right| > 2\delta \right) \leq \mathbb{P}_\mu \left(\left| \sum_{p^* < q \leq \bar{p}} a^{(n)}_{q,p,p'} \right| > \delta \right) + \mathbb{P}_\mu \left(\left| \sum_{q > \bar{p}} a^{(n)}_{q,p,p'} \right| > \delta \right).$$

We have for all $q \in (p^*, \bar{p}]$: $a^{(n)}_{q,p,p'} \to a_{q,p,p'} = \lambda_q \delta_{q=p} \delta_{q=p'} = 0$ (with δ the Kronecker product), as $n \to \infty$, therefore:

$$\limsup_{n \to \infty} \mathbb{P}_\mu \left(\left| \sum_{q > p*} a^{(n)}_{q,p,p'} \right| > 2\delta \right) \leq \frac{\sigma^2 \lambda_{p^*}^{-1} \sum_{q > \bar{p}} \lambda_q}{\delta}.$$

Taking the limit $\bar{p} \to \infty$ on the right hand side, we obtain the convergence in probability of $\sum_{q > p^*} a^{(n)}_{q,p,p'}$ when $n \to \infty$:

$$\sum_{q > p^*} \frac{\lambda_q}{n} \sum_{r=1}^{n} \phi_q(x_r)\phi_p(x_r) \frac{1}{n} \sum_{r=1}^{n} \phi_{p'}(x_r)\phi_q(x_r) \xrightarrow{\mathbb{P}_\mu} 0, \quad \forall p, p' \leq p^*. \tag{41}$$

Following the same method, we obtain the convergence:

$$\sum_{q > p^*} \frac{\lambda_q}{n} \sum_{r=1}^{n} k(x, x_r)\phi_q(x_r) \sum_{r=1}^{n} \phi_p(x_r)\phi_q(x_r) \xrightarrow{\mathbb{P}_\mu} 0, \quad \forall p \leq p^*. \tag{42}$$

Let us return to $S_{q,m}^{(n)}$. By using Cauchy Schwarz inequality and bounding by the constant M all the terms independent of q_i and m_i, we obtain:

$$\left|\sum_{q_1,\ldots,q_{i-j}>p^*} S_{q,m}^{(n)}\right| \leq M \prod_{l=1}^{j} \lambda_{m_l} \frac{1}{n} \sum_{r=1}^{n} \phi_{m_l}(x_r)^2 \times \left|\sum_{q_{i-j}>p^*} \left(\frac{\lambda_{q_{i-j}}}{n} \sum_{r=1}^{n} k(x,x_r)\phi_{q_{i-j}}(x_r) \sum_{r=1}^{n} \phi_{p_{i-j}}(x_r)\phi_{q_{i-j}}(x_r)\right)\right| \times \left|\sum_{q_1,\ldots,q_{i-j-1}>p^*} \prod_{l=1}^{i-j-1} \frac{\lambda_{q_l}}{n} \sum_{r=1}^{n} \phi_{q_l}(x_r)\phi_{p_{l+1}}(x_r) \sum_{r=1}^{n} \phi_{q_l}(x_r)\phi_{p'_l}(x_r)\right|.$$

Since $\sum_{p\geq 0} \lambda_p \phi_p(x)^2 = k(x,x) \leq \sigma^2$, we have the inequalities:

$$0 \leq \sum_{m_1,\ldots,m_j} \prod_{l=1}^{j} \lambda_{m_l} \frac{1}{n} \sum_{r=1}^{n} \phi_{m_l}(x_r)^2 \leq (\sigma^2)^j.$$

Thus, for $i > j$ and from (41) and (42) we obtain the following convergence in probability when $n \to \infty$:

$$\sum_{\substack{q_1,\ldots,q_{i-j}>p^* \\ m_1,\ldots,m_j>p^*}} S_{q,m}^{(n)} \xrightarrow{\mathbb{P}_\mu} 0.$$

Therefore, from (38) we obtain the following convergence when $n \to \infty$:

$$k(x)^T \frac{1}{n} \left(\frac{M}{n}\right)^j \left(\frac{L'M}{n^2}\right)^{i-j} k(x) \xrightarrow{\mathbb{P}_\mu} 0, \quad \forall i < j.$$

Following the same guideline as previously, it can be shown that when $n \to \infty$:

$$k(x)^T \frac{1}{n} \left(\frac{M}{n}\right)^j \left(\frac{L'M}{n^2}\right)^{i-j} \frac{L'}{n^2} k(x) \xrightarrow{\mathbb{P}_\mu} 0, \quad \forall i \leq j.$$

Proof of Lemma 3

Let us consider for a fixed $j \geq 1$:

$$\frac{1}{n} k(x)^T \left(\frac{M}{n}\right)^j k(x) = \sum_{m_1,\ldots,m_j>p^*} a_m^{(n)}(x),$$

with $m = (m_1, \ldots, m_j)$ and:

$$a_m^{(n)}(x) = \left(\frac{1}{n}\sum_{r=1}^{n} k(x,x_r)\phi_{m_1}(x_r)\right)\left(\frac{1}{n}\sum_{r=1}^{n} k(x,x_r)\phi_{m_j}(x_r)\right) \times \prod_{l=1}^{j-1} \frac{1}{n} \sum_{r=1}^{n} \phi_{m_l}(x_r)\phi_{m_{l+1}}(x_r) \prod_{i=1}^{j} \lambda_{m_i}.$$

From Cauchy–Schwarz inequality, we have:

$$\left|a_m^{(n)}(x)\right| \leq \left(\frac{1}{n}\sum_{r=1}^{n} k(x,x_r)^2\right)\prod_{i=1}^{j}\frac{1}{n}\sum_{r=1}^{n}\lambda_{m_i}\phi_{m_i}(x_r)^2 \tag{43}$$

$$\leq \sigma^4 \prod_{i=1}^{j}\frac{1}{n}\sum_{r=1}^{n}\lambda_{m_i}\phi_{m_i}(x_r)^2. \tag{44}$$

Therefore, considering the expectation with respect to the distribution of the points $(x_r)_{r=1,\ldots,n}$, we have:

$$\mathbb{E}_\mu\left[\left|a_m^{(n)}(x)\right|\right] \leq \sigma^4\left(\prod_{i=1}^{j}\lambda_{m_i}\right)\frac{1}{n^j}\sum_{t_1,\ldots,t_j=1}^{n}\mathbb{E}_\mu\left[\phi_{m_1}(X_{t_1})^2\ldots\phi_{m_j}(X_{t_j})^2\right], \quad \forall x\in\mathbb{R}^d.$$

The following inequality holds uniformly in $t_1,\ldots,t_j = 1,\ldots,n$:

$$\mathbb{E}_\mu\left[\prod_{i=1}^{j}\phi_{m_i}(X_{t_i})^2\right] \leq b_m,$$

where $b_m = \sum_{\substack{\mathcal{P}\in\Pi(\{1,\ldots,j\})\\ \mathcal{P}=\cup_{r=1}^{l} I_r}}\prod_{r=1}^{l}\mathbb{E}_\mu\left[\prod_{i\in I_r}\phi_{m_i}(X)^2\right]$ because the term of left hand side of the inequality is equal to one of the terms in the sum on the right hand side. Here $\Pi(\{1,\ldots,j\})$ is the collection of all partitions of $\{1,\ldots,j\}$ and $I_r\cap I_{r'}=\emptyset, \forall r\neq r'$. We hence have:

$$\mathbb{E}_\mu\left[\left|a_m^{(n)}(x)\right|\right] \leq \sigma^4\prod_{i=1}^{j}\lambda_{m_i}b_m.$$

Since $\sum_{p\geq 0}\lambda_p\phi_p(x)^2\leq\sigma^2$, we have:

$$\begin{aligned}\sum_{m_1,\ldots,m_j>p^*}\prod_{i=1}^{j}\lambda_{m_i}b_m &= \sum_{m_1,\ldots,m_j>p^*}\prod_{l=1}^{j}\lambda_{m_l}\sum_{\substack{\mathcal{P}\in\Pi(\{1,\ldots,j\})\\ \mathcal{P}=\cup_{r=1}^{l}I_r}}\prod_{r=1}^{l}\mathbb{E}_\mu\left[\prod_{i\in I_r}\phi_{m_i}(X)^2\right]\\ &= \sum_{\substack{\mathcal{P}\in\Pi(\{1,\ldots,j\})\\ \mathcal{P}=\cup_{r=1}^{l}I_r}}\prod_{r=1}^{l}\mathbb{E}_\mu\left[\prod_{i\in I_r}\sum_{m_i>p^*}\lambda_{m_i}\phi_{m_i}(X)^2\right]\\ &\leq \sigma^{2j}\#\{\Pi(\{1,\ldots,j\})\}.\end{aligned}$$

Since the cardinality of the collection $\Pi(\{1,\ldots,j\})$ of partitions of $\{1,\ldots,j\}$ is finite, the series $\sum_{m_1,\ldots,m_j>p^*}\prod_{i=1}^{j}\lambda_{m_i}b_m$ converges. Furthermore, as it is a series with non-negative terms, $\forall\varepsilon>0$, $\exists\bar{p}>p^*$ such that :

$$\sigma^4\sum_{m\in M_{\bar{p}}^C}\prod_{i=1}^{j}\lambda_{m_i}b_m\leq\varepsilon,$$

where $M_{\bar{p}}^C$ designs the complement of $M_{\bar{p}}$ defined by the collection of $m = (m_1, \ldots, m_j)$ such that:

$$\begin{aligned} M &= \{m = (m_1, \ldots, m_j) \quad \text{such that} \quad m_i > p^*, \quad i = 1, \ldots, j\}, \\ M_{\bar{p}} &= \{m = (m_1, \ldots, m_j) \quad \text{such that} \quad p^* < m_i \leq \bar{p}, \quad i = 1, \ldots, j\}, \\ M_{\bar{p}}^C &= M \backslash M_{\bar{p}}. \end{aligned}$$

Therefore, we have $\forall \delta > 0, \forall \varepsilon > 0 \exists \bar{p} > 0$ such that uniformly in n:

$$\sum_{m \in M_{\bar{p}}^C} \mathbb{E}_\mu \left[\left| a_m^{(n)}(x) \right| \right] \leq \frac{\varepsilon \delta}{2}.$$

Applying the Markov inequality, we obtain:

$$\mathbb{P}\left(\sum_{m \in M_{\bar{p}}^C} \left| a_m^{(n)}(x) \right| > \frac{\delta}{2} \right) \leq \varepsilon. \tag{45}$$

Furthermore, by denoting $a_m(x) = \lim_{n \to \infty} a_m^{(n)}(x)$, we have:

$$a_m(x) = \lambda_{m_1} \lambda_{m_j} \phi_{m_1}(x) \phi_{m_j}(x) \prod_{i=1}^{j} \lambda_{m_i} \prod_{i=1}^{j-1} \delta_{m_i = m_{i+1}}, \tag{46}$$

and from Cauchy–Schwarz inequality [see Eq. (44)], we have:

$$|a_m(x)| \leq \sigma^4 \prod_{i=1}^{j} \lambda_{m_i}.$$

We hence can deduce the inequality:

$$\sum_{m \in M_{\bar{p}}^C} |a_m(x)| \leq \sigma^4 \sum_{m \in M_{\bar{p}}^C} \prod_{i=1}^{j} \lambda_{m_i}. \tag{47}$$

Thus, $\exists \bar{p}$ such that $\sum_{m \in M_{\bar{p}}^C} |a_m(x)| \leq \frac{\delta}{2}$ for all $x \in \mathbb{R}^d$. From the inequalities (45) and (47), we find that $\exists \bar{p}$ such that:

$$\mathbb{P}_\mu \left(\left| \sum_{m \in M} a_m^{(n)}(x) - \sum_{m \in M} a_m(x) \right| > 2\delta \right) \leq \varepsilon + \mathbb{P}_\mu \left(\left| \sum_{m \in M_{\bar{p}}} a_m^{(n)}(x) - \sum_{m \in M_{\bar{p}}} a_m(x) \right| > \delta \right).$$

Since $M_{\bar{p}}$ is a finite set:

$$\limsup_{n \to \infty} \mathbb{P}_\mu \left(\left| \sum_{m \in M_{\bar{p}}} a_m^{(n)}(x) - \sum_{m \in M_{\bar{p}}} a_m(x) \right| > \delta \right) = 0,$$

therefore:

$$\limsup_{n \to \infty} \mathbb{P}_\mu \left(\left| \sum_{m \in M} a_m^{(n)}(x) - \sum_{m \in M} a_m(x) \right| > 2\delta \right) \leq \varepsilon.$$

The previous inequality holds $\forall \varepsilon > 0$, thus we have the convergence in probability of $\sum_{m\in M} a_m^{(n)}(x)$ to $\sum_{m\in M} a_m(x)$ with [by using the limit in the Eq. (46)]:

$$\sum_{m\in M} a_m(x) = \sum_{p>p^*} \lambda_p^{j+2} \phi_p(x)^2.$$

References

Abramowitz, M., & Stegun, I. A. (1965). *Handbook of mathematical functions*. New York: Dover.

Berger, J. O., De Oliveira, V., & Sans, B. (2001). Objective bayesian analysis of spatially correlated data objective bayesian analysis of spatially correlated data. *Journal of the American Statistical Association*, *96*, 1361–1374.

Bozzini, M., & Rossini, M. (2003). Numerical differentiation of 2d functions from noisy data. *Computer and Mathematics with Applications*, *45*, 309–327.

Bronski, J. C. (2003). Asymptotics of Karhunen–Loève eigenvalues and tight constants for probability distributions of passive scalar transport. *Communications in Mathematical Physics*, *238*, 563–582.

Fang, K. T., Li, R., & Sudjianto, A. (2006). *Design and modeling for computer experiments. Computer science and data analysis series*. London: Chapman & Hall.

Fernex, F., Heulers, L., Jacquet, O., Miss, J., Richet, Y. (2005). The Moret 4b monte carlo code new features to treat complex criticality systems. In: MandC International Conference on Mathematics and Computation Supercomputing, Reactor and Nuclear and Biological Application, Avignon, France.

Gneiting, T., Kleiber, W., & Schlater, M. (2010). Matérn cross-covariance functions for multivariate random fields. *Journal of the American Statistical Association*, *105*, 1167–1177.

Harville, D. A. (1997). *Matrix algebra from statistician's perspective*. New York: Springer-Verlag.

Laslett, G. M. (1994). Kriging and splines: An empirical comparison of their predictive performance in some applications kriging and splines: An empirical comparison of their predictive performance in some applications. *Journal of the American Statistical Association*, *89*, 391–400.

Mercer, J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society A*, *209*, 441–458.

Nazarov, A. I., & Nikitin, Y. Y. (2004). Exact l_2-small ball behaviour of integrated Gaussian processes and spectral asymptotics of boundary value problems. *Probability Theory and Related Fields*, *129*, 469–494.

Opper, M., & Vivarelli, F. (1999). General bounds on Bayes errors for regression with Gaussian processes. *Advances in Neural Information Processing Systems*, *11*, 302–308.

Picheny, V. (2009). *Improving accuracy and compensating for uncertainty in surrogate modeling*. PhD thesis, Ecole Nationale Supérieure des Mines de Saint Etienne.

Pusev, R. S. (2011). Small deviation asymptotics for Matérn processes and fields under weighted quadratic norm. *Theory of Probability and its Applications*, *55*, 164–172.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. Cambridge: MIT Press.

Ritter, K. (2000a). Almost optimal differentiation using noisy data. *Journal of Approximation Theory*, *86*, 293–309.

Ritter, K. (2000b). *Average-case analysis of numerical problems*. Berlin: Springer Verlag.

Sacks, J., & Ylvisaker, D. (1981). Variance estimation for approximately linear models. *Series Statistics*, *12*, 147–162.

Sacks, J., Welch, W. J., Mitchell, T. J., & Wynn, H. P. (1989). Design and analysis of computer experiments. *Statistical Science*, *4*, 409–423.

Seeger, M. W., Kakade, S. M., & Foster, D. P. (2008). Information consistency of nonparametric Gaussian process methods. *IEEE Transactions on Information Theory*, *54*(5), 2376–2382.

Shanno, D. F. (1970). Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation*, *24*, 647–656.

Sollich, P., & Halees, A. (2002). Learning curves for Gaussian process regression: Approximations and bounds. *Neural Computation*, *14*, 1393–1428.

Stein, M. L. (1999). *Interpolation of spatial data. Series in statistics*. New York: Springer.

van der Vaart, A., & van Zanten, H. (2011). Information rates of nonparametric Gaussian process methods. *The Journal of Machine Learning Research*, *12*, 2095–2119.

Wackernagel, H. (2003). *Multivariate geostatistics*. Berlin: Springer-Verlag.

Williams, C. K. I., & Vivarelli, F. (2000). Upper and lower bounds on the learning curve for Gaussian processes. *Machine Learning*, *40*, 77–102.