Reproducing kernel Hilbert spaces, polynomials and the classical moment problem

Holger Dette* and Anatoly Zhigljavsky†

Abstract. We show that polynomials do not belong to the reproducing kernel Hilbert space of infinitely differentiable translation-invariant kernels whose spectral measures have moments corresponding to a determinate moment problem. Our proof is based on relating this question to the problem of the best linear estimation in continuous time one-parameter regression models with a stationary error process defined by the kernel. In particular, we show that the existence of a sequence of estimators with variances converging to 0 implies that the regression function cannot be an element of the reproducing kernel Hilbert space. This question is then related to the determinacy of the Hamburger moment problem for the spectral measure corresponding to the kernel.

AMS Subject Classification: 46E22, 62J05, 44A60
Keywords and Phrases: Reproducing kernel Hilbert spaces, classical moment problem, best linear estimation, continuous time regression model

1. Introduction.

1.1. Main results. Let $X \subseteq \mathbb{R}^d$, $d \geq 1$, $K : X \times X \to \mathbb{R}$ be a positive definite kernel on X and define $H(K)$ as the corresponding Reproducing Kernel Hilbert Space (RKHS). We assume that X has a non-empty interior and K is an infinitely differentiable (on the diagonal) translation-invariant kernel so that $K(x,y) = k(x-y)$, where $k : \mathbb{R}^d \to \mathbb{R}$ is a non-constant positive definite function infinitely differentiable at the point 0. Clearly, the function $k(\cdot)$ is even so that $k(u) = k(-u)$ for all $u \in \mathbb{R}^d$. Without loss of generality, we suppose $k(0) = 1$.
It is well-known, see e.g. Corollary 4.44 in **?**, that in the case of the Gaussian (squared exponential) kernel

$$k(x) = \exp\{-\lambda \|x\|^2\} \;\text{ with } \lambda > 0, \tag{1.1}$$

the constant function does not belong to $H(K)$. This result has been generalized to arbitrary polynomials in **?**. The purpose of this paper is to significantly extend these results (previously only known for the case of the Gaussian kernel) to a substantially larger class of kernels.
In the main part of the paper, we consider the case $X \subset \mathbb{R}$. In this case, by Bochner's theorem (**?**), there exists a measure α, such that the function k can be represented in the form

$$k(x) = \int_{-\infty}^{\infty} e^{itx} \alpha(dt) \quad \text{for all } x \in X. \tag{1.2}$$

*Ruhr-Universität Bochum, Fakultät für Mathematik, 44780 Bochum, Germany
†Cardiff University, UK, email Zhigljavskyaa@cardiff.ac.uk (corresponding author)

The measure α is called *spectral measure*. As function $k(\cdot)$ is even and $k(0) = 1$, $\alpha(dt)$ is a probability measure symmetric around the point 0. The moments of this measure (in the case of their existence) will be denoted by $c_n = \int_{-\infty}^{\infty} t^n \alpha(dt)$. Since we assume $k(\cdot)$ in (**??**) is infinitely differentiable and even, we have

$$c_n = \int_{-\infty}^{\infty} t^n \alpha(dt) = \begin{cases} (-1)^{n/2}\, k^{(n)}(0), & \text{if } n \text{ is even} \\ 0, & \text{if } n \text{ is odd} \end{cases}, \tag{1.3}$$

where $k^{(n)}(0) = \frac{\partial^n}{\partial u^n} k(u)\big|_{u=0}$.
The classical *Hamburger moment problem* is to give necessary and sufficient conditions such that a given real sequence $(c_n)_{n\in\mathbb{N}}$ is in fact a sequence of moments of a distribution α defined on the Borel sets of $\mathbb{R} = (-\infty, \infty)$. In particular, the sequence $(c_n)_{n\in\mathbb{N}}$ is a sequence of moments of some distribution if and only if the *Hankel* matrices $(c_{i+j})_{i,j=0,\ldots,n}$ are positive semidefinite for all $n \in \mathbb{N}$; see e.g. **??** among many others. The Hamburger moment problem is called *determinate* if the sequence of moments $(c_k)_{k\in\mathbb{N}}$ determines the measure $\alpha(dt)$ uniquely.
The main results of this paper are Theorems **??** and **??** formulated below. These theorems provide sufficient conditions ensuring that the polynomials do not belong to the RKHS $H(K)$. The proofs are given in Section **??**.

Theorem 1.1. *Let $X \subset \mathbb{R}$ and assume that the spectral measure $\alpha(dt)$ in* (**??**) *has infinite support and no mass at the point 0. If the Hamburger moment problem for this measure is determinate, then the non-zero constant functions do not belong to the RKHS $H(K)$.*

Theorem 1.2. *Let $X \subset \mathbb{R}$, m be a positive integer and assume that the spectral measure $\alpha(dt)$ in* (**??**) *has infinite support. If the Hamburger moment problem for the measure $\alpha_m(dt) = t^{2m}\alpha(dt)/c_{2m}$ is determinate (here, as in* (**??**)*, $c_{2m} = \int t^{2m}\alpha(dt)$), then the RKHS $H(K)$ does not contain polynomials on X of degree precisely m.*

Theorem **??** can be considered as a corollary of Theorem **??** and therefore the result of Theorem **??** is more fundamental. In the very important particular case (see the sufficient condition for moment determinacy (**??**) and related discussion in Section **??**), when the function $k(\cdot)$ is real analytic and vanishes at infinity while X is bounded, $H(K)$ contains only those analytic functions that vanish at infinity. This is formally proven in **?** and basically follows from the fact that if $f \in H(K)$ then $f(x)$ is a point-wise limit of the sums $\mu_N(x) = \sum_{i=1}^{N} w_i k(x_i - x)$ with $x_i \in X$, which necessarily vanish at infinity.
Combining Theorems **??** and **??** with their variations in the cases when the spectral measure $\alpha(dt)$ has finite support (see Section **??**) and when this measure has positive mass at 0 (see Theorem **??**), we obtain the following corollary.

Corollary 1.3. *Let $X \subset \mathbb{R}$ and the Hamburger moment problem for the spectral measure $\alpha(dt)$ be determinate. Then we have the following:*

(a) the constant functions $f(x) = \text{const} \neq 0$, $\forall x \in X$, belong to $H(K)$ if and only if $\alpha(dt)$ has a positive mass at the point 0;

(b) if additionally $k(\cdot)$ is a real analytic function, then $H(K)$ does not contain non-constant polynomials on X.

Theorems **??**, **??** and Corollary **??** can be easily extended to the multivariate case, see Section **??**.

1.2. Implications and related results. Gaussian process (GP) models deal with an unknown deterministic function assuming that it is a realization of Gaussian process (field) with some mean and covariance kernel, which are perhaps parameterized. The popularity of the GP model comes from its transparency, flexibility and computational tractability; it is used as a general-purpose technique to model, explore and exploit unknown functions. As a result, methods based on the GP models constitute much of the modern statistical toolkit for function approximation, interpolation and prediction (**??**), integration (**?**), machine learning (**??**), space-filling (**?**), signal processing (**?**), probabilistic numerics (**?**) and global optimization (**?**). The practical application areas of the GP model are vast, and we refer the references in the cited papers and to the work of **?**, **?** among others.

Consider the common framework of GP regression (simple kriging), where the a function $f : X \to \mathbb{R}$ to be approximated ($X \subset \mathbb{R}^d$) is considered as a realization of a GP, say $\{Z_x\}_{x\in X}$, with mean zero, covariance

$$\mathbb{E}\{Z_x Z_{x'}\} = \sigma^2 \, K(x, x') \text{ for all } x, x' \in X \subset \mathbb{R}^d, \tag{1.4}$$

and $\sigma^2 > 0$ may be unknown; see **?** for more details. Let the kernel K be strictly positive definite, $X_N = \{x_1, \dots, x_N\}$ be an N-point design consisting of distinct points $x_i \in X$ and $F_N = [f(x_1), \dots, f(x_N)]^\top \in \mathbb{R}^N$ be the vector of exact observations of f at the points of X_N. The conditional process $\{Z_x|(X_N, F_N)\}_{x\in X}$ is again Gaussian with conditional mean

$$\mu_N(x) = F_N^\top K_N^{-1} b_N(x) \tag{1.5}$$

and covariance function

$$C_N(x, y) = K(x, x') - b_N^\top(x) K_N^{-1} b_N(x'), \tag{1.6}$$

where $K_N = (K(x_i, x_j))_{i,j=1}^N$ and $b_N(\cdot) = [K(x_1, \cdot), \dots, K(x_N, \cdot)]^\top$. Straightforward calculation shows that the conditional mean $\mu_N(x)$ is the best linear predictor of $f(x)$ and $\sigma^2 \, C_N(x, x)$ is the corresponding mean squared prediction error at the point x.

GP regression is equivalent to kernel interpolation, see e.g. **?**, Section 3.3 in **?** and Chapter 3 in

?. More precisely, the conditional mean $\mu_N(\cdot)$ is the minimal-norm interpolant to f among all functions in $H(K)$, where we use the norm on $H(K)$ denoted below by $\|\cdot\|_{H(K)}$. This property implies in particular that if $f \in H(K)$ then $\|\mu_N\|_{H(K)} \leq \|f\|_{H(K)}$, where this inequality holds for any $f \in H(K)$ and any set of points X_N. If $f \notin H(K)$ then the conditional mean $\mu_N(\cdot)$ is still an element of $H(K)$ but its norm tends to infinity as N grows and X_N becomes denser. Therefore, there is a fundamental difference between the complexity of the approximation problem of f depending on whether $f \in H(K)$ or $f \notin H(K)$. Correspondingly, properties of all other techniques based on the use of the GP model also heavily depend on whether an unknown function of interest belongs to the corresponding RKHS. We also refer to the work of **?** and to Section 4.4 in **?** for a discussion on the importance of this issue for the learning performance of support vector machines (SVMs) in the case of the Gaussian kernel (**??**) as well as for the difficulty of deciding whether a given function f belongs to the RKHS $H(K)$ for a chosen kernel K. In GP regression, knowing that the non-zero constant functions do not belong to the RKHS $H(K)$ is especially important as it can be used to justify omitting function centering; see, for example, Assumption 2 in **?**.

Assume now that the factor σ^2 in (**??**) is unknown and that the maximum likelihood estimator (MLE) $\widehat{\sigma_N^2}$ of σ^2 is constructed from the observations of the function f at the points $x_i \in X_N$; see Section **??**. If f is indeed a realization of the GP with covariance (**??**), it follows by the well-known results on microergodicity (see Chapter 6 in **?**) that $\widehat{\sigma_N^2} \to \sigma^2$ almost surely (a.s.) as $N \to \infty$ and X_N becomes dense in X. Note, however, that such realizations do not belong to $H(K)$ a.s. and if $f \in H(K)$ then $\widehat{\sigma_N^2} \to 0$ as $N \to \infty$. This observation is a consequence of the useful relation $\widehat{\sigma_N^2} = \frac{1}{N}\|\mu_N\|^2_{H(K)}$ (see equation (3.4) in **?**) and the relation $\|\mu_N\|_{H(K)} \leq \|f\|_{H(K)}$, which has already been mentioned. This is in full agreement with Corollary **??** below, which states that $f \in H(K)$ if and only if $\lim_{N\to\infty} N\,\widehat{\sigma_N^2} < \infty$ (assuming X_N becomes dense in X as $N \to \infty$). Our numerical studies in Section **??** of this paper show, however, that for finite sample sizes the asymptotic behaviour of $\widehat{\sigma_N^2}$ has much less effect on uncertainty quantification in GP regression than the smoothness of the function $k(x)$ and its flatness at $x \simeq 0$.

The problem of deciding whether a given function f belongs to the RKHS $H(K)$ for a chosen kernel K is well-known in literature, see e.g. Section 3.4 in **?**. This problem is well-studied in the case of kernels with finite number of derivatives, (**???**). For infinitely differentiable kernels, only the case of Gaussian kernel (**??**) is well-understood; see **?** for most advanced results. In a preprint citing the present paper, **?** gives further new results for analytic translation-invariant kernels with $\lim_{\|x\|\to\infty} k(x) = 0$.

1.3. Sufficient conditions for moment determinacy. There are many sufficient conditions for moment determinacy of probability measures, see e.g. **?**, **?** and **?**, Chapter 11. The following sufficient condition for moment determinacy of a measure $\alpha(dt)$ with moments c_k in the Hamburger moment problem, the so-called *Carleman condition*, is one of the most commonly used:

$$\text{(A.1)}: \qquad \sum_{n=1}^{\infty} c_{2n}^{-1/(2n)} = \infty\,.$$

Note that if a probability measure $\alpha(dt)$ has finite moments c_j for all $j = 0, 1, \ldots$ and satisfies (A.1), then for any $m = 0, 1, \ldots$ the measure $\alpha_m(dt) = t^{2m}\alpha(dt)/c_{2m}$ of Theorem **??** with moments $\int t^j \alpha_m(dt) = c_{2m+j}/c_{2m}$ also satisfies (A.1).
Less known than (A.1) is the following sufficient condition for the moment determinacy (in the Hamburger moment problem) of a measure $\alpha(dt)$:

$$\text{(A.2)}: \qquad \exists\, \varepsilon > 0 \;\text{ such that } \sum_{n=1}^{\infty} \frac{|c_n| x^n}{n!} < \infty \quad \text{for all } |x| < \varepsilon,$$

see Theorem 30.1 in **?**. This condition is clearly equivalent to the assumption (A.3): the random variable (r.v.) ξ_α with distribution α has moment generating function and, see Theorem 1 in **?**, to (A.4): $\limsup_{n\to\infty} \frac{1}{2n} c_{2n}^{1/(2n)} < \infty$; this is one of the most known and easiest to verify sufficient conditions for moment determinacy. Theorem 2.13 in **?** yields that the assumptions (A.2)–(A.4) are also equivalent to the assumption (A.5): the r.v. ξ_α is sub-exponential. In view of this assumption, the tail behaviour of $\alpha(dt)$ is a natural indicator of the degree of moment-determinacy of α.
The conditions (A.2)–(A.5) are stronger than the Carleman condition (A.1) in the sense that any of them implies (A.1). A technique of constructing measures α satisfying (A.1) but violating $\limsup_{n\to\infty} \frac{1}{2n} c_{2n}^{1/(2n)} < \infty$, and hence all other assumptions in (A.2)–(A.5), is given in **?**, Section 11.9.
In view of (**??**) and symmetry of $k(\cdot)$, $|c_n| = |k^{(n)}(0)|$ for all integers $n = 0, 1, \ldots$ This yields that (A.2) coincides with the assumption that $k(\cdot)$ is a real analytic function at the point 0, see Definition 1.1.5 and Corollary 1.1.16 in **?**. Summarizing, if $k(\cdot)$ is a real analytic at 0, then the corresponding spectral measure $\alpha(dt)$ is moment-determinant in the Hamburger sense. It is not difficult, however, to construct moment-determinant (in the Hamburger sense) spectral measures $\alpha(dt)$ which do not satisfy (A.2); see Sections 11.9 and 11.10 in **?**. For such measures, the corresponding functions $k(\cdot)$ are symmetric, positive definite and infinitely differentiable but not real analytic at 0. Nevertheless, for practical purposes the class of functions which are symmetric, positive definite and real analytic can be considered as the main class of functions $k(\cdot)$, for which the corresponding spectral measures are moment determinant in the Hamburger sense. Note also that if the function $k(\cdot)$ is real analytic then all functions in $H(K)$ have to be analytic too; see **?**.
Let us give four examples of kernels $K(x,y) = \sigma^2 k(x-y)$ whose spectral measures $\alpha(dt)$ satisfy

the Carleman condition (A.1) and hence assumptions of Theorems **??**, **??** and Corollary **??** (we will return to these kernels in Sections **??** and **??**). The spectral density corresponding to the spectral measure $\alpha(\cdot)$ will be denoted by $\varphi(\cdot)$, so that $\alpha(dt) = \varphi(t)dt$. In examples (E.1)–(E.4) below, we also provide asymptotic expressions for $C_N(\alpha) = \sum_{n=1}^{N} c_{2n}^{-1/(2n)}$. In view of (A.1), the rate of divergence of $C_N(\alpha)$ (as $N \to \infty$) can also be used to characterize the degree of moment-determinacy of α (additionally to the tail behaviour of α).

(E.1) Gaussian kernel: $k(x) = \exp\{-\lambda x^2/2\}$, $\varphi_\lambda(t) = \frac{1}{\sqrt{2\pi\lambda}} \exp\{-t^2/(2\lambda)\}$ $(t \in \mathbb{R})$, $c_{2n} = \lambda^n(2n-1)!!$, $C_N(\alpha) = \sum_{n=1}^{N} c_{2n}^{-1/(2n)} = \sqrt{N/\lambda}(1+o(1))$, $N \to \infty$.

(E.2) Cauchy kernel: $k(x) = 1/(1+x^2/\lambda^2)$, $\varphi_\lambda(t) = \frac{\lambda}{2}\exp\{-\lambda|t|\}$ $(t \in \mathbb{R})$, $c_{2n} = \lambda^{-n}(2n)!$, $c_{2n}^{-1/(2n)} = e\sqrt{\lambda}/n + o(1/n)$, so that $C_N(\alpha) = e\sqrt{\lambda}\log N(1+o(1))$ as $N \to \infty$.

(E.3) The kernel whose spectral density is a symmetric Beta-density:
$\varphi_a(t) = 2^{-2a-1}(1-t^2)^a/B(a+1,a+1)$ with $a > -1$ and $t \in [-1,1]$.
Here we have $C_N(\alpha) = \text{const} \cdot N(1+o(1))$ as $N \to \infty$.

(E.4) $k(x) = \cos(\lambda x)$ with $\lambda \neq 0$; here the spectral measure α is concentrated at $\pm\lambda$ with masses 1/2 yielding $c_{2n} = \lambda^{2n}$, $c_{2n}^{-1/(2n)} = 1/\lambda$ and $C_N(\alpha) = N/\lambda$ for all N.

1.4. Main steps in the proofs and the structure of the remaining part of the paper. Section **??** is devoted to proving Theorems **??** and **??**; the proofs are given in several steps. The main idea in our approach is to relate the problem of interest to properties of the best linear unbiased estimate (BLUE) in linear regression models, which will be worked out in Sections **??** and **??**. Sections **??** and **??** provide different characterizations of the moment determinacy of spectral measures and finally the proofs will be completed in Section **??**. We now explain the different steps in more detail.

In Section **??** we consider a one-parameter linear regression model $y(x) = \theta f(x) + \varepsilon(x)$ with $\mathbb{E}\varepsilon(x)\varepsilon(x') = K(x,x')$ and a regression function $f \in H(K)$ and show that in this case $\hat\theta_{BLUE}$, the BLUE of θ, exists and its variance is strictly positive, see Lemma **??**. We also show that in the case $f \notin H(K)$, the BLUE does not exist and establish in Lemma **??** that for proving $f \notin H(K)$, it is sufficient to construct a sequence of linear unbiased estimators $\hat\theta_n$ of the unknown parameter with variances tending to 0. Such a sequence is constructed in Section **??** for the location scale model and an explicit expression for the variance of these estimators in terms of the ratio of determinants

$$\operatorname{var}(\hat\theta_n) = \frac{\det(c_{2(i+j)})_{i,j=0}^{n}}{\det(c_{2(i+j)})_{i,j=1}^{n}} \tag{1.7}$$

of Hankel-type matrices of the moments of the spectral measure is derived in Lemma **??**. In Section **??** we establish several properties of moment-determinant symmetric measures which we use in Section **??** for building up an equivalence between the moment determinacy of the spectral measures and the statement that the sequence (**??**) converges to zero. This is arguably the most important step in the proof of both theorems (see Lemma **??**). Finally, these results are combined in the proofs of Theorem **??** and **??** in Section **??**.

In Section **??** we consider several extensions and interpretations of the main results. In Section **??** we consider spectral measures with finite support, while Section **??** discusses the multivariate case. This discussion is continued in Sections **??**–**??**, where we also consider general metric spaces. In Section **??** we explain a technique of characterizing the inclusion $f \in H(K)$ via suitable discretization of the set X and show that $1/\|f\|_{H(K)}$ is the limit of variances of the related discrete BLUEs. These results are used in Section **??**, where we prove that the constant function belongs to $H(K)$ if and only if the spectral measure has positive mass at 0. In Section **??** we show that the problem of parameter estimation in a one-parameter regression model is equivalent to the problem of estimating the variance of a Gaussian process (field). Thus we are able to relate our findings to the estimation problems considered in **?** and **?**. In Section **??** we return to the one-dimensional case and give an interpretation of Theorem **??** in terms of the L_2-error of the best approximation of a constant function by polynomials of the form $a_1t^2 + a_1t^4 + \ldots + a_nt^{2n}$.

In Section **??**, for two specific classes of kernels we derive explicit results on the rates of convergence to 0 of the ratio of determinants (**??**). In the case of Gaussian kernel (**??**), we detail and improve one of the asymptotic expansions of Theorem 3.3 in **?**. Finally, in Section **??** we discuss results of a numerical study for uncertainty quantification in GP regression in relation to the theoretical results of this paper.

2. Parameter estimation, moment determinacy and proofs of main results.

2.1. BLUE in a one-parameter regression model . Consider a one-parameter regression model with stationary correlated errors:

$$y(x)=\theta f(x)+\varepsilon(x),\ x\in X,\ \mathbb{E}\varepsilon(x)=0,\ \mathbb{E}\varepsilon(x)\varepsilon(x')=k(x-x'). \tag{2.1}$$

Here θ is a scalar parameter, $f: X \to \mathbb{R}$ is a given regression function and $k(\cdot)$ is an infinitely differentiable positive definite function with $k(0) = 1$ making the kernel $K(\cdot,\cdot)$ defined by $K(x,y) = k(k-y)$ an infinitely differentiable correlation kernel. For constructing estimators of the parameter θ, the observations of the process $\{y(x)|x \in X\}$ along with observations of all of its derivatives $\{y^{(k)}(x)|x \in X\}$, $k = 1, 2, \ldots$, can be used.

An estimator $\hat\theta$ for the parameter θ is called linear, if it is a linear function of the observations (in our case of the process and its derivatives). An unbiased estimator satisfies $\mathbb{E}[\hat\theta] = \theta$ for all θ. The best linear unbiased estimator (BLUE) of θ is defined as an unbiased estimator $\hat\theta_{BLUE}$ such that $\text{var}(\hat\theta_{BLUE}) \le \text{var}(\hat\theta)$, where $\hat\theta$ is any linear unbiased estimator of θ. If the kernel K is differentiable and the BLUE exists, then for its computation all available derivatives of $y(x)$ are used, see **?**. In general, the BLUE may not exist but the next lemma shows that it does exist when $f \in H(K)$.

Lemma 2.1. *If $f \in H(K)$, then the BLUE $\hat{\theta}_{BLUE}$ in model (**??**) exists and*

$$\mathrm{var}(\hat{\theta}_{BLUE}) = 1/\|f\|_{H(K)} > 0.$$

The statement of lemma follows from Theorem 6C (p. 975) of **?**. Formally, only the case $X = [0,1]$ is considered in **?**, but Parzen's proof does not use the structure of X and is therefore valid for a general metric space X.

Lemma 2.2. *If there exists a sequence of linear unbiased estimators $(\hat{\theta}_n)_{n\in\mathbb{N}}$ of θ in model (**??**) such that $\mathrm{var}(\hat{\theta}_n) \to 0$ as $n \to \infty$, then $f \notin H(K)$.*

Proof. Assume that $f \in H(K)$. By Lemma **??**, the continuous BLUE $\hat{\theta}_{BLUE}$ exists and $\mathrm{var}(\hat{\theta}_{BLUE}) = 1/\|f\|_{H(K)} > 0$. From the definition of the BLUE, $\mathrm{var}(\hat{\theta}_n) \geq \mathrm{var}(\hat{\theta}_{BLUE}) > 0$ for all $n \in \mathbb{N}$. We have arrived at a contradiction and hence $f \notin H(K)$. □

2.2. A family of estimators $\hat{\theta}_n$ in the location scale model.

Consider the location scale model

$$y(x) = \theta + \varepsilon(x)\,,\; x \in X \subset \mathbb{R},\;\; \mathbb{E}\varepsilon(x) = 0,\; \mathbb{E}\varepsilon(x)\varepsilon(x') = k(x - x'), \tag{2.2}$$

where $k(\cdot)$ is an infinitely differentiable at 0 positive definite function. Choose any interior point $x_0 \in X$ and set $\varepsilon_0 = \varepsilon(x_0)$. For construction of the estimator $\hat{\theta}_n$, which we will apply in Lemma **??**, we use the following $n+1$ observations: the observation $y(x_0) = \theta + \varepsilon_0$ at the point x_0 and n mean-square derivatives of the process y at the point x_0:

$$\varepsilon_j = y^{(j)}(x_0) = \left.\frac{d^j y(x)}{dx^j}\right|_{x=x_0} = \left.\frac{d^j \varepsilon(x)}{dx^j}\right|_{x=x_0}, \quad j = 2, 4, \ldots, 2n\,. \tag{2.3}$$

As discussed in Section **??**, for the main class of kernels of interest, the RKHS $H(K)$ is a subset of analytic functions. If we observe $y(x)$ everywhere on [0,1], then, since $f \in H(K)$ and $\varepsilon(x)$ are analytic, we also know all $y^{(k)}(x)$ for all $x \in [0,1]$ and any integer $k \geq 0$. Again, because of the analyticity, observing $y(x)$ everywhere on $X = [0,1]$ is the same as observing $y^{(k)}(x)$ for any $x \in [0,1]$ and any integer $k \geq 0$. This yields that in practice we do not need to directly observe $y^{(k)}(\cdot)$ for constructing estimators of θ.
The following result provides a necessary and sufficient condition for the existence of the derivatives. For a proof, see page 164 (Section 12) in **?**.

Lemma 2.3. *Let x_0 be an interior point of X. The mean-square derivative $\varepsilon_j = d^j\varepsilon(x)/dx^j\big|_{x=x_0}$*

of the stationary process $\{\varepsilon(x)|x \in X\}$ *in* (**??**) *at the point* x_0 *exists if and only if* $c_{2j} < \infty$, *where*

$$c_{2j} = \int_{-\infty}^{\infty} t^{2j}\alpha(dt) = (-1)^j \frac{\partial^{2j}}{\partial u^{2j}} k(u)\Big|_{u=0} \tag{2.4}$$

is the $2j$*-th moment of the spectral measure* α *corresponding to the kernel* k *in Bochner's theorem.*

As we have assumed that the kernel $k(\cdot)$ is infinitely differentiable at 0, all moments c_j ($j = 0, 1, \ldots$) exist. As an immediate consequence of the existence of all moments and the representation (**??**), for the random variables ε_j defined in (**??**), we obtain by Lemma **??**

$$\mathbb{E}\varepsilon_i\varepsilon_j = \frac{\partial^{i+j}}{\partial x^i y^j} k(x-y)\Big|_{x,y=x_0} = (-1)^{i+j} \frac{\partial^{i+j}}{\partial u^{i+j}} k(u)\Big|_{u=0} = c_{i+j} \, . \tag{2.5}$$

for all $i, j = 0, 1, \ldots$ Note that all derivatives $\partial^m/\partial u^m k(u)\big|_{u=0}$ of odd order m vanish as the function $k(\cdot)$ is symmetric around the point 0.
Next, we introduce the random variables $\delta_i = (-1)^i \varepsilon_{2i}$, $i = 0, 1, \ldots$. The observations (**??**) used for constructing the discrete BLUE in model (**??**) can then be rewritten as

$$y_0 = y(x_0) = \theta + \varepsilon_0, \; y_1 = y^{(2)}(x_0) = \delta_1, \; \ldots, \; y_n = y^{(2n)}(x_0) = \delta_n.$$

Moreover, the covariance matrix of the vector $(\delta_0, \delta_1, \ldots, \delta_n)^\top$ is the Hankel matrix

$$C_n = (\mathbb{E}\delta_i\delta_j)_{i,j=0}^n = (c_{2(i+j)})_{i,j=0}^n \, , \tag{2.6}$$

where $c_2, \ldots, c_{2n}$ are the moments defined in (**??**).
Assume that the spectral measure $\alpha(dt)$ has infinite support. In this case, the matrices C_n are positive definite for all $n = 0, 1, \ldots$ (see, for example, Proposition 3.11 in **?**) and the discrete BLUE is obtained as

$$\hat{\theta}_n = \frac{e_{0,n}^\top C_n^{-1} Y_n}{e_{0,n}^\top C_n^{-1} e_{0,n}}, \tag{2.7}$$

where $Y_n = (y_0, y_1, \ldots, y_n)^\top$ and $e_{0,n} = (1, 0, \ldots, 0)^\top \in \mathbb{R}^{n+1}$ denotes the first coordinate vector in $\mathbb{R}^{n+1}$.

Lemma 2.4. *The variance of the estimator* (**??**) *is*

$$\mathrm{var}(\hat{\theta}_n) = \frac{1}{e_{0,n}^\top C_n^{-1} e_{0,n}} = \frac{H_n}{G_n} \, , \tag{2.8}$$

where H_n and G_n are the determinants

$$H_n = \det(C_n) = \det\left[\left(c_{2(i+j)}\right)_{i,j=0}^{n}\right], \quad G_n = \det\left[\left(c_{2(i+j)}\right)_{i,j=1}^{n}\right]. \tag{2.9}$$

Proof. The expression (**??**) follows from the standard formula $\mathrm{var}(\hat{\theta}_n) = 1/(e_{0,n}^{\top} C_n^{-1} e_{0,n})$ for the variance of the BLUE and Cramér's rule for computing elements of a matrix inverse; in our case, $e_{0,n}^{\top} C_n^{-1} e_{0,n}$ coincides with the top-left element of the matrix C_n^{-1}. □

Observing Lemma **??** we conclude that a non-vanishing constant function does not belong to $H(K)$ if $\lim_{n\to\infty} H_n/G_n = 0$. In the following sections we relate this condition to the moment determinacy of the spectral measure.

Let us now briefly consider the case where the spectral measure has a positive mass at the point 0. Consider the location scale model (**??**) and let

$$\alpha_\gamma(dt) = (1-\gamma)\alpha(dt) + \gamma\delta_0(dt) \tag{2.10}$$

denote the spectral measure corresponding to a nonnegative definite and symmetric kernel k_γ, where $0 < \gamma < 1$, δ_0 is the Dirac measure at the point 0 and $\alpha(dt)$ is a symmetric probability measure on $\mathbb{R}$ with no mass at 0. The measure $\alpha_\gamma(dt)$ is symmetric around the point 0 with even moments $\tilde{c}_0 = 1$ and

$$\tilde{c}_{2j} = (1-\gamma)c_{2j}, \quad j = 1, 2, \ldots$$

Recall the definition of the matrix C_n in (**??**) and define the matrices

$$\tilde{C}_n = (\tilde{c}_{2(i+j)})_{i,j=0}^{n} = \gamma e_{0,n} e_{0,n}^{\top} + (1-\gamma)C_n$$

and the corresponding determinants

$$\tilde{H}_n = \det \tilde{C}_n, \quad \tilde{G}_n = \det\left[\left(\tilde{c}_{2(i+j)}\right)_{i,j=1}^{n}\right] = (1-\gamma)^n G_n,$$

where G_n is defined in (**??**). Using standard formulas of linear algebra we obtain

$$\tilde{H}_n = \det\left[\gamma e_{0,n} e_{0,n}^{\top} + (1-\gamma)C_n\right] = (1-\gamma)^n \left[(1-\gamma) + \gamma e_{0,n}^{\top} C_n^{-1} e_{0,n}\right] H_n.$$

In accordance with (**??**), the variance of $\tilde{\theta}_n$, the BLUE of θ constructed similarly to $\hat{\theta}_n$ but for the spectral measure $\alpha_\gamma(dt)$, is given by

$$\begin{aligned}\operatorname{var}(\tilde{\theta}_n) &= \frac{1}{e_{0,n}^\top \tilde{C}_n^{-1} e_{0,n}} = \frac{\tilde{H}_n}{\tilde{G}_n} = \frac{H_n}{G_n}\left[(1-\gamma) + \gamma e_{0,n}^\top C_n^{-1} e_{0,n}\right] \\ &= \operatorname{var}(\hat{\theta}_n)[(1-\gamma) + \gamma/\operatorname{var}(\hat{\theta}_n)] = (1-\gamma)\operatorname{var}(\hat{\theta}_n) + \gamma > 0\,.\end{aligned}$$

This implies that $\operatorname{var}(\tilde{\theta}_n)$ cannot converge to 0 and Lemma **??** is not applicable if the spectral measure has a positive mass at the point 0.
In Theorem **??** of Section **??** we will prove that for any compact set $X \subset \mathbb{R}^d$ the constant functions indeed belong to $H(K)$, if the spectral measure has a positive mass at the point 0.

2.3. Moment-determinacy of the spectral measure. Consider the spectral measure α introduced in equation (**??**). As a spectral measure, α is a symmetric measure (around 0) on the real line and we have assumed that α does not have a positive mass at the point 0. Moreover, we have assumed $k(0) = 1$ making α a probability distribution. In the following we relate α to a (unique) measure on the nonnegative axis $[0, \infty)$. Loosely speaking, if a real valued random variable ξ has distribution $\alpha(dt)$, then $\alpha_+(dt)$ is the distribution of the random variable ξ^2. In the opposite direction, if the nonnegative random variable η has distribution $\alpha_+(dt)$, then $\pm\sqrt{\eta}$ has distribution $\alpha(dt)$, where $\pm$ denotes a random sign.
For a more formal construction we follow the arguments in Section 3.3 of **?** and denote by $\mathcal{B}$ the Borel sigma field on $\mathbb{R}$, define $\tau : \mathbb{R} \to [0, \infty); \tau(x) = x^2$ and $\kappa : [0, \infty) \to \mathbb{R}$, $\kappa(x) = \sqrt{x}$. Then for any symmetric (Radon) measure α on $\mathcal{B}$, the measure α_+ defined by

$$\alpha_+(B) = \alpha(\tau^{-1}(B)) \qquad B \in \mathcal{B} \cap [0, \infty) \tag{2.11}$$

defines a measure on $\mathcal{B} \cap [0, \infty)$. Conversely, if α_+ is a measure on $\mathcal{B} \cap [0, \infty)$, then

$$\alpha(B) = \frac{1}{2}\big(\alpha_+(\kappa^{-1}(B)) + \alpha_+((-\kappa)^{-1}(B))\big) \tag{2.12}$$

defines a symmetric measure on $\mathcal{B}$. It now follows from Theorem 3.17 in **?** that the relations (**??**) and (**??**) define a bijection from the set of all symmetric measures on $\mathbb{R}$ onto the set of all measures on $[0, \infty)$.
The even moments of a symmetric probability measure α on $\mathcal{B}$ are related to the moments of the measure α_+ from (**??**) by

$$c_{2j} = \int_{-\infty}^{\infty} t^{2j} \alpha(dt) = 2\int_0^\infty t^{2j} \alpha(dt) = \int_0^\infty t^j \, d\alpha_+(t) = b_j, \;\; j \in \mathbb{N}\,, \tag{2.13}$$

and as a consequence the determinants H_n and G_n in (**??**) can be represented as

$$H_n = \det\left[(b_{i+j})_{i,j=0}^n\right], \quad G_n = \det\left[(b_{i+j})_{i,j=1}^n\right]. \tag{2.14}$$

Similarly to the case of the Hamburger moment problem, the *Stieltjes moment problem* is to give necessary and sufficient conditions such that a real sequence $(b_j)_{j\in\mathbb{N}}$ is in fact a sequence of moments of a measure $\alpha_+(dt)$ on the Borel sets of $[0,\infty)$; that is $b_j = \int_0^\infty t^j d\alpha_+(t)$ for all $j \in \mathbb{N}_0$. The Stieltjes moment problem is determinate if the sequence of moments $(b_j)_{j\in\mathbb{N}}$ determines the measure $\alpha_+(dt)$ uniquely. For a proof of the following result, which relates the Hamburger and Stieltjes moment problem, see **?**, Lemma 1, **?**, Proposition 3.19 and **?**, Sect. 11.10.

Lemma 2.5. *Let α be a symmetric probability measure on $\mathcal{B}$. The Hamburger moment problem for α is determinate if and only if the Stieltjes moment problems for the measure α_+ defined by* (**??**) *is determinate.*

Note that for the equivalence in Lemma **??** to hold, the assumption that α does not have mass at 0 is not required. This assumption, however, is needed in the next lemma.

Lemma 2.6. *Let α be a symmetric probability measure on $\mathcal{B}$ with no mass at the point* 0*. The Hamburger moment problem for α is determinate if and only if the Hamburger moment problem for the measure α_+ defined by* (**??**) *is determinate.*

Proof. Using the result of Theorem A in **?** (see also (**?**, p.113) and (**?**, Remark 2.12)), if the Stieltjes moment problems for the measure α_+ is determinate and the measure α_+ has no mass at 0, then the Hamburger moment problems for this measure is also determinate. From Lemma **??**, the required equivalence follows. □

2.4. Relating moment-determinacy of the measure α_+ to $\operatorname{var}(\hat{\theta}_n)$.

Lemma 2.7. *Let α be a symmetric probability measure on $\mathcal{B}$ with infinite support and no mass at the point* 0*. The Hamburger moment problem for the measure α_+ defined by* (**??**) *is determinate if and only if $H_n/G_n \to 0$ as $n \to \infty$, where the determinants H_n and G_n are defined in* (**??**)*.*

Proof. (i) Assume that the moment problem for the measure α_+ is determinate. Let $\mathcal{P}_n$ denote the class of all polynomials of degree n and define

$$\rho_n(t_0) = \min\left\{\int_{\mathbb{R}} |P_n(t)|^2 \alpha_+(dt) \mid P_n \in \mathcal{P}_n, P_n(t_0) = 1\right\}$$

for any $t_0 \in \mathbb{R}$, which is not a root of the nth orthogonal polynomial with respect to the measure α_+ (see equation (2.26) in Lemma 2.11 of **?**). Then

$$\lim_{n\to\infty} \rho_n(t_0) =: \rho(t_0)$$

exists, by Theorem 2.6 in **?**. As the point 0 is not a support point of the measure α_+ and all roots of the orthogonal polynomials with respect to the measure α_+ are located in $\mathrm{supp}(\alpha_+) \subset (0,\infty)$ we have from Corollary 2.6 in **?** that

$$\rho(0) = \lim_{n\to\infty} \rho_n(0) = 0.$$

Moreover, by the discussion on p. 72 (middle of the page) in **?** it follows that $\rho_n(0)$ is exactly the ratio H_n/G_n, where H_n and G_n are the determinants in (**??**). Hence the moment determinacy for the measure α_+ implies $H_n/G_n \to 0$ as $n \to \infty$.

(ii) To prove the converse, assume that $H_n/G_n \to 0$ as $n \to \infty$. Let λ_n be the smallest eigenvalue of the matrix C_n. Theorem 1.1 in **?** states that the condition

$$\lim_{n\to\infty} \lambda_n = 0$$

is necessary and sufficient for the moment-determinacy of the measure α_+.

From the definition of λ_n as the smallest eigenvalue of the matrix C_n and the representation (**??**) it follows

$$\lambda_n \le \frac{1}{e_{0,n}^\top C_n^{-1} e_{0,n}} = \frac{H_n}{G_n} = \rho_n(0)$$

for all $n \in \mathbb{N}$ (see also a related discussion in **?**). Therefore, $H_n/G_n \to 0$ as $n \to \infty$ implies $\lambda_n \to 0$ as $n \to \infty$ and this yields the moment determinacy of α_+. □

2.5. Proof of Theorem ?? and ??. *Proof of Theorem* **??**. Use Lemma **??** with the estimator defined in (**??**). By Lemma **??** the variance of this estimator is given by (**??**). From Lemma **??**, the determinacy of the measure α_+ is equivalent to $\mathrm{var}(\hat\theta_n) \to 0$ as $n \to \infty$. By Lemma **??**, this is also equivalent to the moment determinacy of the spectral measure α. □

Proof of Theorem **??**. Assume that the function f in (**??**) is a polynomial of degree $m \ge 1$. Take m derivatives of both sides in (**??**). The model (**??**) thus reduces to $\tilde y(x) = \tilde\theta + \tilde\varepsilon(x)$, $x \in X$, where $\tilde\theta$ is the new parameter, $\tilde y(x) = y^{(m)}(x)$ are new observations and $\tilde\varepsilon = \varepsilon^{(m)}$ is the new error process. From (2.178) in **?**, the autocovariance function of the process $\{\varepsilon^{(m)}(x) | x \in X\}$ is given by

$$\mathbb{E}\varepsilon^{(m)}(x)\varepsilon^{(m)}(x') = k_m(x-x') \;\text{ with }\; k_m(x) = (-1)^m k^{(2m)}(x).$$

From (**??**), the spectral measure associated with the kernel $k_m(x-x')$ is $\alpha_m(dt) = t^{2m}\alpha(dt)/c_{2m}$. Hence, the statement for the case when f is a polynomial of degree $m \ge 1$ is reduced to the case of the constant function proved in Theorem **??**; this theorem is applicable as the measure $\alpha_m(dt)$ does not have mass at 0 for any $m \ge 1$. □

3. Extensions of Theorems ?? and ?? and further discussion. In this section we discuss several extensions of the results derived in Sections **??** and **??**. In particular, we consider spectral measures with positive mass at the point 0 and extends the results to the multivariate case. Moreover, we briefly indicate a relation of our results to the optimal approximation of a constant function by polynomials with no intercept.

3.1. Spectral measures with finite support . If the spectral measure $\alpha(dt)$ in (**??**) has finite support, say $\mathcal{T} = \{\pm t_1, \ldots, \pm t_m\}$ with $m \geq 1$ and $0 < t_1 < \ldots < t_m$, then the matrices C_n in (**??**) are invertible for $n \leq m-1$ but

$$\det(C_n) = \det(c_{2(i+j)})_{i,j=0}^{n} = 0 \quad \text{for} \quad n \geq m. \tag{3.1}$$

Consequently, observing Lemma **??** we have in this case

$$\mathrm{var}(\hat{\theta}_n) = 0 \qquad \text{for } n = m, m+1, \ldots$$

Therefore, by Lemma **??** a non-vanishing constant function does not belong to $H(K)$ if the corresponding spectral measure has finite support.
The relation (**??**) follows, observing the representation

$$C_n = 2\sum_{i=1}^{m} w_i g(t_i) g^\top(t_i) \in \mathbb{R}^{(n+1)\times(n+1)}$$

where $g(t) = (1, t^2, \ldots, t^{2n})^\top$ and $w_1, \ldots, w_m$ are the masses of the measure α at the points $t_1, \ldots, t_m$. As C_n is a sum of rank one matrices, it is singular whenever $n > m-1$. On the other hand, in the case $m = n+1$ we have by the Vandermond determinant formula

$$\det C_n = \prod_{i=1}^{n+l} (2w_i) \prod_{1\leq i<j\leq n+1} (t_i^2 - t_j^2)^2 > 0 \ ,$$

which shows that C_n is nonsingular. Finally, if $m \geq n+1$ we have (in the Loewner ordering)

$$C_n \geq 2\sum_{i=1}^{n+1} w_i g(t_i) g^\top(t_i)$$

where the matrix on the right-hand side is positive definite.

3.2. Multivariate case . Consider the location scale model (**??**) but assume that X is a subset of $\mathbb{R}^d$ with non-empty interior. Extensions of Theorems **??** and **??** to the multivariate case, when $d > 1$, essentially follow from the one-dimensional results because it is sufficient to use derivatives of the process $\{y(x); x \in X\}$ with respect to one variable for construction of estimators $(\hat{\theta}_n)_{n\in\mathbb{N}}$ and subsequent application of Lemma **??**. In the following discussion we consider two cases for the kernel $K(x,x')$ using the notation $x = (x_1,\dots,x_d)^\top$, $x' = (x'_1,\dots,x'_d)^\top$ and $t = (t_1,\dots,t_d)^\top$. We also denote by $x_{(i)}, x'_{(i)}$ and $t_{(i)} \in \mathbb{R}^{d-1}$ the vectors x, x' and t with i-th component removed, respectively.

Case 1: Assume that K is a product kernel, that is

$$K(x,x') = \prod_{j=1}^{d} K_i(x_j, x'_j)\,, \tag{3.2}$$

where for all $j = 1,\dots,d$ the kernel K_j (defined on a subset of $\mathbb{R}^2$) satisfies $K_j(x_j, x'_j) = k_j(x_j - x'_j)$ and k_j is a non-constant positive definite function infinitely differentiable at the point 0. Denote by $\alpha_j(dt_j)$ the spectral measure for k_j and define $\alpha(dt) = \alpha_1(dt_1)\cdots\alpha_d(dt_d)$. To construct the sequence of estimators $(\hat{\theta}_n)_{n\in\mathbb{N}}$ for the application of Lemma **??**, we can use the derivatives with respect to the i-th coordinate for any i. Therefore, Corollary **??** can be generalized as follows.

Corollary 3.1. *Assume that $X \subset \mathbb{R}^d$ and the kernel K has the form (**??**). Then we have the following:*

(a) If the measure α has a positive mass at the point 0, then the constant functions belong to $H(K)$.

(b) If for at least one $i \in \{1,\dots,d\}$ the Hamburger moment problem for the measure $\alpha_i(\cdot)$ is determinate and the measure α_i does not have a positive mass at the point 0, then any non-vanishing constant function does not belong to $H(K)$.

(c) If for at least one $i \in \{1,\dots,d\}$ the Hamburger moment problem for the measures $t_i^{2m}\alpha_i(dt_i)/c_{2m}$ is determinate for all $m = 0,1,\dots,$ then $H(K)$ does not contain non-constant polynomials on X.

Note that the set X in Corollary **??** does not have to be a product of one-dimensional sets. Moreover, we also point out that the assumption (**??**) can be generalized to kernels of the form

$$K(x,x') = k_i(x_i - x'_i) K_{(d-1)}(x_{(i)}, x'_{(i)}),$$

where $K_{(d-1)}(\cdot,\cdot)$ is a positive definite and suitably differentiable kernel on $\mathbb{R}^{d-1}\times\mathbb{R}^{d-1}$ and k_i is a non-constant positive definite function infinitely differentiable at the point 0.

Case 2: The kernel K satisfies

$$K(x,x') = k(x - x'),$$

where k is a positive definite function on $\mathbb{R}^d$. Consider the spectral measure $\alpha(dt)$ corresponding to k by Bochner's theorem, that is

$$k(x) = \int_{\mathbb{R}^d} e^{i(t_1x_1+\ldots+t_dx_d)}\alpha(dt) \ , \tag{3.3}$$

and denote by

$$\alpha_i(B) = \int_{\mathbb{R}^d} I_B(t_i)\alpha(dt) \ , \quad B \in \mathcal{B},$$

the ith the marginal distribution of the measure α ($i = 1, \ldots, d$), where I_B denotes the indicator function of the set B. In this case, we can generalize Corollary **??** as follows.

Corollary 3.2. *If the spectral measure $\alpha(dt)$ does not have a positive mass at the point 0 and if for at least one $i \in \{1, \ldots, d\}$ the Hamburger moment problems for the measures proportional to $t^{2m}\alpha_i(dt_i)$ are determinate for all $m = 0, 1, \ldots$, then $H(K)$ does not contain non-vanishing polynomials.*

The case when the spectral measure has positive mass at the point 0 is treated similarly in one-dimensional and multi-dimensional cases, see Section **??**.

3.3. Discretization of space and the limit of discrete BLUEs. In Section **??** below we will prove that constant functions belong to $H(K)$ if the spectral measure has positive mass at the point 0. The proof requires an auxiliary result which is of own interest and shows that in the case $f \in H(K)$ the variance of the continuous BLUE is the limit of the variances of discrete BLUEs, after a suitable discretization of X has been performed.

Lemma 3.3. *Let X be a compact in $\mathbb{R}^d$, $(x_N)_{N\in\mathbb{N}}$ be a sequence of distinct points in X such that $f(x_1) \neq 0$ and*

$$\sup_{x\in X} \min_{1\le i\le N} \|x - x_i\| \to 0 \ \text{ as } N \to \infty \, . \tag{3.4}$$

*Let $\widehat{\theta}_{BLUE,N}$ be the BLUE of θ in model (**??**) from the observations of $y(x_1), \ldots, y(x_N)$. Then $f \in H(K)$ if and only if $var(\widehat{\theta}_{BLUE,N}) \to c > 0$ as $N \to \infty$.*
*Moreover, if $f \in H(K)$, the continuous BLUE $\hat{\theta}_{BLUE}$ of θ in model (**??**) exists and*

$$c = 1/\|f\|_{H(K)} = \text{var}(\hat{\theta}_{BLUE}).$$

Proof. Let $X_N = \{x_1, \ldots, x_N\}$, K_N denote the restriction of K on X_N, and define $H_N = H(K_N)$ as the RKHS corresponding to the kernel K_N. By Theorem 6 in Section 1.4.2 of **?** we have for $f_N = f\big|_{X_N}$, the restriction of f on X_N, that $f_N \in H_N$ and $\|f_N\|_{H_N} \le \|f_{N+1}\|_{H_{N+1}} \le \|f\|_{H(K)}$.
Consequently, the sequence of $(\text{var}(\widehat{\theta}_{BLUE,N}))_{N\in\mathbb{N}} = (1/\|f_N\|_{H_N})_{N\in\mathbb{N}}$ is monotonously decreasing so that the limit $c = \lim_{N\to\infty} \text{var}(\widehat{\theta}_{BLUE,N}) \ge 0$ exists for any f. Moreover, $\text{var}(\widehat{\theta}_{BLUE,N}) \ge c$

for all $N \in \mathbb{N}$. If $f \in H(K)$ we have by Proposition 3.9 in **?** that $\lim_{N\to\infty} \mathrm{var}(\widehat{\theta}_{BLUE,N}) = c = 1/\|f\|_{H(K)}$. Conversely, if $\mathrm{var}(\widehat{\theta}_{BLUE,N}) \to c$ as $N \to \infty$ for some $c > 0$, we can use the equivalence between (1) and (2) in Theorem 3.11 of **?** to deduce that $f \in H(K)$. □

Recall that the explicit expression for the variance of the discrete BLUE $\widehat{\theta}_{BLUE,N}$ of Lemma **??** is given by

$$\mathrm{var}(\widehat{\theta}_{BLUE,N}) = 1/F_N^\top K_N^{-1} F_N \,, \tag{3.5}$$

where

$$F_N = (f(x_1), \dots, f(x_N))^\top \,, \;\; K_N = (K(x_i, x_j))_{i,j=1}^N \,. \tag{3.6}$$

Since the kernel $K(\cdot,\cdot)$ is assumed to be strictly positive definite, the matrix K_N is invertible for all $N = 1, 2, \dots$

3.4. Spectral measures with positive mass at the point 0. In this section we investigate the case, where the spectral measure has a positive mass at the point 0 and hence generalize the result obtained at the end of Section **??**. We assume that the covariance kernel of the error process has the form

$$K_\gamma(x, x') = \gamma + (1-\gamma) K(x, x') \,, \tag{3.7}$$

where $0 \le \gamma < 1$ and $K(x, x')$ is a strictly positive definite kernel on a compact set $X \subset \mathbb{R}^d$. Note that in the particular case $d = 1$ and $K(x, x') = k(x - x')$ with k having the spectral measure $\alpha(dt)$, we obtain the representation (**??**) for the spectral measure α_γ.

Theorem 3.4. *Let $X \subset \mathbb{R}^d$ be a compact set and assume the kernel K_γ has the form (**??**) with $0 < \gamma < 1$. Then then the constant functions belong to $H(K_\gamma)$.*

Proof. Consider the location scale model

$$y(x) = \theta + \varepsilon(x) \,, \;\; x \in X, \;\; \mathbb{E}\varepsilon(x) = 0, \;\; \mathbb{E}\varepsilon(x)\varepsilon(x') = K_\gamma(x, x') \,. \tag{3.8}$$

and let $(x_n)_{n\in\mathbb{N}}$ denote a sequence of distinct points in X such that (**??**) is satisfied. Let $\hat{\theta}_{m,\gamma}$ be the BLUE of θ in the model (**??**), constructed on the observations of $y(x_1), \dots, y(x_m)$. Define $W_{m,\gamma} = (K_\gamma(x_i, x_j))_{i,j=1}^m$, $Y_m = (y(x_1), \dots, y(x_m))^\top$ and $\mathbf{1}_m = (1, \dots, 1)^\top \in \mathbb{R}^m$. As the covariance kernel $K(x, x')$ is strictly positive definite, the matrix $W_{m,\gamma}$ is invertible for all $m \ge 1$, $0 \le \gamma < 1$. Therefore, the BLUE is unique and given by

$$\hat{\theta}_{m,\gamma} = \mathbf{1}_m^\top W_{m,\gamma}^{-1} Y_m / \mathbf{1}_m^\top W_{m,\gamma}^{-1} \mathbf{1}_m \,.$$

Its variance is

$$\mathrm{var}(\hat\theta_{m,\gamma}) = 1/\mathbf{1}_m^\top W_{m,\gamma}^{-1}\mathbf{1}_m\,.$$

For simplicity of notation, denote $\kappa_{m,\gamma} = \mathbf{1}_m^\top W_{m,\gamma}^{-1}\mathbf{1}_m = 1/\mathrm{var}(\hat\theta_{m,\gamma})$. The same arguments as given in the proof of Lemma **??** show that for any $0 \le \gamma < 1$, the sequence $(\kappa_{m,\gamma})_{m\in\mathbb{N}}$ is monotonously increasing with some limit $c_\gamma = \lim_{m\to\infty}\kappa_{m,\gamma} \in (0,\infty]$. Observing the representation

$$W_{m,\gamma} = (1-\gamma)W_{m,0} + \gamma\mathbf{1}_m\mathbf{1}_m^\top$$

(for all $m = 1,2,\dots$ and $0<\gamma<1$), we have

$$W_{m,\gamma}^{-1} = \frac{1}{1-\gamma}\Big[W_{m,0}^{-1} - \frac{\gamma}{1-\gamma+\gamma\mathbf{1}_m^\top W_{m,0}^{-1}\mathbf{1}_m} W_{m,0}^{-1}\mathbf{1}_m\mathbf{1}_m^\top W_{m,0}^{-1}\Big]\,.$$

This implies

$$\kappa_{m,\gamma} = \frac{\kappa_{m,0}}{1-\gamma}\Big[1 - \frac{\gamma\kappa_{m,0}}{1-\gamma+\gamma\kappa_{m,0}}\Big] = \frac{\kappa_{m,0}}{1-\gamma+\gamma\kappa_{m,0}},$$

and therefore it follows that

$$\mathrm{var}(\hat\theta_{m,\gamma}) = 1/\kappa_{m,\gamma} = \gamma + (1-\gamma)\mathrm{var}(\hat\theta_{m,0})\,. \tag{3.9}$$

Taking the limit (as $m\to\infty$) in (**??**) we obtain for all $0<\gamma<1$:

$$\lim_{m\to\infty}\mathrm{var}(\hat\theta_{m,\gamma}) = \gamma + (1-\gamma)/c_0 \ge \gamma > 0\,.$$

Lemma **??** now yields that the constant functions belong to $H(K_\gamma)$. □

3.5. Estimation of the variance of a Gaussian random field. Let $X \subset \mathbb{R}^d$ be a compact set, and let f denote of a Gaussian random process (field) on X with a strictly positive definite covariance kernel $R(x,x') = \sigma^2 K(x,x')$ on $X\times X$, where the kernel $K(x,x')$ is known but σ^2 is unknown. For estimating σ^2 we assume that one can observe f at N distinct points $x_1,\dots,x_N \in X$. Then it is easy to deduce (see, for example, p.140 in **?**) that the corresponding log-likelihood function is

$$LL(\sigma^2) = \frac{1}{2}\left[-N\log(2\pi) - N\log(\sigma^2) - \log(\det(K_N)) - \frac{1}{\sigma^2}F_N^\top K_N^{-1}F_N\right], \tag{3.10}$$

where F_N and K_N are defined by (**??**). Moreover, a simple calculation shows that the maximum likelihood estimator (MLE) of σ^2 is given by

$$\widehat{\sigma_N^2} = \frac{1}{N} F_N^T K_N^{-1} F_N \,. \tag{3.11}$$

Comparing (**??**) with (**??**) we get

$$\widehat{\sigma_N^2} = \frac{1}{N\,\mathrm{var}(\widehat{\theta}_{BLUE,N})}\,, \tag{3.12}$$

and by Lemma **??** we obtain the following corollary.

Corollary 3.5. *Let X be a compact set in $\mathbb{R}^d$, K a strictly positive definite kernel on $X \times X$ and f a function on X. If $x_1, x_2, \ldots$ is a sequence of distinct points in X satisfying* (**??**) *and $\widehat{\sigma_N^2}$ is the MLE of σ^2 constructed from the observations $f(x_1), \ldots, f(x_N)$ under the assumption that f is a realization of a GP with zero mean and covariance* (**??**)*, then $f \in H(K)$ if and only if $\lim_{N\to\infty} N\,\widehat{\sigma_N^2} < \infty$.*

3.6. Best polynomial approximation . Let $L_2(\alpha)$ denote the space of square integrable functions with respect to the measure $\alpha(dt)$ on the real line and define $\mathcal{P}_{n-1}$ to be the space of of polynomials of degree $n-1$. For $p \in \mathcal{P}_{n-1}$ we consider the $L_2(\alpha)$-distance

$$V(p) = \int_{-\infty}^{-\infty} (1 - t^2 p(t^2))^2 \alpha(dt)$$

between the constant function $g(t) \equiv 1$ and the even polynomial $t^2 p(t^2)$ of degree $2n$ with no intercept. A well know result in approximation theory (see, for example, **?**, p. 15-16) shows that

$$\mathrm{min}_{p_n \in \mathcal{P}_{n-1}} V(p) = \frac{\det(c_{2(i+j)})_{i,j=0}^n}{\det(c_{2(i+j)})_{i,j=1}^n} = \mathrm{var}(\hat{\theta}_n)\,, \tag{3.13}$$

where $c_0, c_2, c_4, \ldots$ are the (even) moments of the spectral measure α defined in (**??**) and the last equality is a consequence of Lemma **??**.
From this representation it follows that $\mathrm{var}(\hat{\theta}_n) \to 0$ as $n \to \infty$ if and only if non-zero constant functions can be approximated by polynomials of the form $\tilde{p}_n(t) = t^2 p_n(t^2)$ with arbitrary small error. Moreover, for any polynomial p on $(-\infty, \infty)$, we have

$$V(p) = \int_0^\infty (1 - tp(t))^2 \alpha_+(dt) = b_2 \int_0^\infty (1/t - p(t))^2 \alpha_{2,+}(dt),$$

where the measure $\alpha_+(dt)$ is defined by (**??**), $b_2 = \int_0^\infty t^2 d\alpha_+(t)$ and $\alpha_{2,+}(dt) = t^2 \tilde{\alpha}_+(dt)/b_2$. From Corollary 2.3.3 in **?**, it therefore follows that the set of all polynomials $\mathcal{P}_\infty = \cup_{n=0}^\infty \mathcal{P}_n$

is dense in the space $L_2([0,\infty),\nu)$ if the measure ν on $[0,\infty)$ is the (unique) solution of a determinate Hamburger moment problem. As the function $f(t)=1/t$ belongs to $L_2((0,\infty),\alpha_{2,+})$ we thus obtain from (**??**) another proof of the fact that if $\alpha(dt)$ has no mass at 0 and $\alpha_{2,+}(dt)$ is moment-determinate in the Hamburger sense then $\mathrm{var}(\hat{\theta}_n)\to 0$. Note that this is almost equivalent to the 'if' statement in the important Lemma **??**.

4. Rates of convergence.

In this section, we derive for several specific classes of correlation kernels explicit results on the rate of convergence of the ratio $\mathrm{var}(\hat{\theta}_n)=H_n/G_n$, see (**??**), where H_n and G_n are the determinants defined in (**??**).

4.1. Gaussian kernel.

Let $K(x,x')=\exp\{-\lambda(x-x')^2\}$ with $X\subset\mathbb{R}$ and $\lambda>0$. Assuming for simplicity $\lambda=1/4$, we obtain that the spectral measure is absolutely continuous with density

$$\varphi(t)=\frac{1}{\sqrt{\pi}}e^{-t^2},\quad -\infty<t<\infty\,.$$

The moments of even order of the measure α are given by

$$c_{2j}=\int_{-\infty}^{\infty}t^{2j}\varphi(t)dx=\int_0^{\infty}t^j g(t)dt=b_j=2^j(2j-1)!!\quad j=0,1,\ldots,$$

where $g(y)=\frac{1}{\sqrt{\pi}}y^{-1/2}e^{-y},\ y>0$. Using any of the sufficient conditions (A.1)–(A.5) of Section **??**, the corresponding Hamburger moment problem is determinate and therefore non-vanishing constant functions (and all polynomials) do not belong to the corresponding RKHS. We now investigate the variance of the discrete BLUE defined in (**??**), which is given by the ratio of the determinants H_n and G_n.

It follows from results in **?** that the determinant of the Hankel matrix defined in (**??**) has the representation

$$H_n=\left|c_{2(i+j)}\right|_{i,j=0}^{n}=\left|b_{i+j}\right|_{i,j=0}^{n}=\prod_{i=1}^{n}\left(\tilde{d}_{2i-1}\tilde{d}_{2i}\right)^{n-i+1}, \tag{4.1}$$

where $\tilde{d}_j$ are the coefficients of the three-term recurrence relation

$$P_{\ell+1}(t)=(t-\tilde{d}_{2\ell}-\tilde{d}_{2\ell+1})P_\ell(t)-\tilde{d}_{2\ell-1}\tilde{d}_{2\ell}P_{\ell-1}(t),\qquad \ell=0,1,\ldots \tag{4.2}$$

of the monic orthogonal polynomials with respect to measure $g(y)dy$ ($\tilde{d}_0=0$, $P_0(t)=1$, $P_{-1}(t)=0$). Observing the three-term recurrence relation

$$(\ell+1)L_{\ell+1}^{(\alpha)}(t)=(-t+2\ell+\alpha+1)L_\ell^{(\alpha)}(t)-(\ell+\alpha)L_{\ell-1}^{(\alpha)}(t)$$

for the Laguerre polynomials $L_n^{(\alpha)}(t)$ (orthogonal with respect to $e^{-y}y^\alpha dy$, $y > 0$) we can identify the coefficients in (**??**). More precisely, the monic polynomials

$$\overline{L}_{\ell+1}^{(\alpha)}(t) = (-1)^{\ell+1}(\ell+1)!L_{\ell+1}^{(\alpha)}(t)$$

satisfy a three-term recurrence relation of the form (**??**) with $\tilde{d}_{2k} = k, \tilde{d}_{2k-1} = k + \alpha$, see **?**, Lemma 2.2 (b). As $P_\ell(t) = \overline{L}_\ell^{(-1/2)}(t)$ we have $\tilde{d}_{2k} = k,\ \ \tilde{d}_{2k-1} = k - 1/2$, and therefore obtain

$$H_n = \prod_{k=1}^n (k(2k-1))^{n-k+1} \prod_{k=1}^n \Big(\frac{1}{2}\Big)^{n-k+1} = \Big(\frac{1}{2}\Big)^{n(n+1)/2} \prod_{k=1}^n (k(2k-1))^{n-k+1} . \tag{4.3}$$

Now we move on to the determinant $G_n = \left|b_{i+j}\right|_{i,j=1}^n$. Note that we have

$$b_j = \frac{1}{\sqrt{\pi}} \int_0^\infty y^{j-2} y^{3/2} e^{-y} dy = \frac{3}{4} a_{j-2}$$

for $j \geq 2$, where $a_k = \int_0^\infty y^k \tilde{g}(y)dy$ and the density $\tilde{g}_k$ is defined by $\tilde{g}(y) = \frac{4}{3\sqrt{\pi}} y^{3/2} e^{-y}$, $y > 0$. Therefore,

$$G_n = \Big(\frac{3}{4}\Big)^n \left|a_{i+j}\right|_{i,j=0}^{n-1} = \Big(\frac{3}{4}\Big)^n \prod_{l=1}^{n-1} \big(\overline{d}_{2l-1}\overline{d}_{2l}\big)^{n-l} ,$$

where $\overline{d}_{2l-1} = l + 3/2$, $\overline{d}_{2l} = l$. Consequently,

$$G_n = \Big(\frac{3}{4}\Big)^n \Big(\frac{1}{2}\Big)^{n(n-1)/2} \prod_{k=1}^{n-1} (k(2k+3))^{n-k}$$

and it follows

$$\begin{aligned} \frac{H_n}{G_n} &= \Big(\frac{4}{3}\Big)^n \Big(\frac{1}{2}\Big)^n \Big[\prod_{k=1}^{n-1} \frac{(k(2k-1))^{n-k+1}}{(k(2k+3))^{n-k}}\Big] n(2n-1) \\ &= \Big(\frac{2}{3}\Big)^n n!(2n-1) \prod_{k=1}^{n-1} \frac{(2k-1)^{n-k+1}}{(2k+3)^{n-k}} . \end{aligned}$$

Since

$$\prod_{k=1}^{n-1} \frac{(2k-1)^{n-k+1}}{(2k+3)^{n-k}} = \frac{3^n}{(2n-1)(2n+1)!!}$$

we obtain

$$\frac{H_n}{G_n} = \frac{2^n\, n!}{(2n+1)!!} = \frac{\sqrt{\pi}}{2\sqrt{n}} \Big[1 - \frac{3}{8n} + \frac{25}{128n^2} + O\Big(\frac{1}{n^3}\Big)\Big],\ n \to \infty. \tag{4.4}$$

The expansion (**??**) details the asymptotic relation formulated as Theorem 3.3 in **?** in the case $p = 0$. Note that formula (**??**) also corrects a minor mistake in this reference, which gives $\frac{\sqrt{\pi}}{\sqrt{2n}}$ as the leading term.

4.2. Spectral measure with Beta distribution . For measures with a compact support the determinants H_n and G_n can be conveniently evaluated using the theory of canonical moments, see e.g. **?**. Exemplarily, we consider the symmetric Beta (α, α) distribution on the interval $[-1, 1]$ with density

$$\psi'_\alpha(t) = \frac{1}{2^{2\alpha+1}B(\alpha+1, \alpha+1)}(1-t^2)^\alpha, \qquad -1 < t < 1, \tag{4.5}$$

where $\alpha > -1$ and $B(\alpha, \beta)$ denotes the Beta-function. For later purposes we also introduce the Beta(α, β) distribution on the interval $[0, 1]$ with density

$$\phi_{\alpha,\beta}(t) = \frac{1}{B(\beta+1, \alpha+1)} t^\beta (1-t)^\alpha, \quad 0 < t < 1, \tag{4.6}$$

where the $\alpha, \beta > -1$. The canonical moments of the Beta-distribution with density (**??**) are given by

$$p_{2j} = \frac{j}{2j+1+\alpha+\beta}, \quad p_{2j-1} = \frac{\beta+j}{2j+\alpha+\beta}\,; \tag{4.7}$$

see e.g. formula (1.3.11) in **?**. It is easy to see that the distribution on the interval $[0, 1]$ related to the distribution ψ_α in (**??**) by the transformation (**??**) is a Beta $(\alpha, -\frac{1}{2})$ distribution. Therefore, it follows from (**??**) that the corresponding canonical moments are given by

$$p_{2j} = \frac{j}{2j+1/2+\alpha}, \quad p_{2j-1} = \frac{j-1/2}{2j-1/2+\alpha}\ . \tag{4.8}$$

Now Theorem 1.4.10 in **?** gives

$$H_n = |(b_{i+j})_{i,j=0}^n| = \prod_{i=1}^n (q_{2i-2}p_{2i-1}q_{2i-1}p_{2i})^{n+1-i}\,, \tag{4.9}$$

where $q_0 = 1, q_j = 1 - p_j$ $(j \geq 1)$ and (observing (**??**))

$$q_{2i-2}p_{2i-1}q_{2i-1}p_{2i} = \frac{4i\,(i+\alpha)\,(2\,i-1+2\,\alpha)\,(2\,i-1)}{(4\,i+1+2\,\alpha)\,(4\,i-1+2\,\alpha)^2\,(4\,i-3+2\,\alpha)}\,, \quad i = 1, 2\ldots \tag{4.10}$$

For the calculation of the determinant $G_n = |(b_{i+j})_{i,j=1}^n|$ we note the relation

$$b_i = \frac{B(\frac{5}{2}, \alpha+1)}{B(\frac{1}{2}, \alpha+1)} \tilde{b}_{i-2} \qquad i = 2, 3, \dots \tag{4.11}$$

where $\tilde{b}_0, \tilde{b}_1, \dots$ are the moments of the Beta$(\alpha, 3/2)$ distribution. Consequently, we obtain from Theorem 1.4.10 in **?** that

$$G_n = |(b_{i+j})_{i,j=1}^n| = |(\tilde{b}_{i+j})_{i,j=0}^{n-1}| = \left[\frac{B(\frac{5}{2}, \alpha+1)}{B(\frac{1}{2}, \alpha+1)}\right]^n \times \prod_{i=1}^{n-1} (\tilde{q}_{2i-2}\tilde{p}_{2i-1}\tilde{q}_{2i-1}\tilde{p}_{2i})^{n-i}$$

$$= \left[\frac{3}{(2\alpha+3)(2\alpha+5)}\right]^n \times \prod_{i=2}^{n} (\tilde{q}_{2i-4}\tilde{p}_{2i-3}\tilde{q}_{2i-3}\tilde{p}_{2i-2})^{n+1-i} , \tag{4.12}$$

where $\tilde{p}_1, \tilde{p}_2$ are the canonical moments of Beta$(\alpha, 3/2)$ distribution, that is

$$\tilde{p}_{2i} = \frac{j}{2i + 5/2 + \alpha} , \quad \tilde{p}_{2i-1} = \frac{3/2 + i}{2i + 3/2 + \alpha} ,$$

and

$$\tilde{q}_{2i-2}\tilde{p}_{2i-1}\tilde{q}_{2i-1}\tilde{p}_{2i} = \frac{4i\,(i+\alpha)\,(2\,i+3+2\,\alpha)\,(2\,i+3)}{(4\,i+5+2\,\alpha)\,(4\,i+3+2\,\alpha)^2\,(4\,i+1+2\,\alpha)} , \quad i = 1, 2 \dots \tag{4.13}$$

Consequently, it follows from (**??**), (**??**) and (**??**)

$$\frac{H_n}{G_n} = \Big[\frac{(2\alpha+3)(2\alpha+5)}{3}\Big]^n (q_0p_1q_1p_2)^n \prod_{i=2}^{n} \Big[\frac{q_{2i-2}p_{2i-1}q_{2i-1}p_{2i}}{\tilde{q}_{2i-4}\tilde{p}_{2i-3}\tilde{q}_{2i-3}\tilde{p}_{2i-2}}\Big]^{n+1-i}$$

$$= \Big[\frac{4(1+\alpha)}{3\,(3+2\,\alpha)^2}\Big]^n \prod_{i=2}^{n} \Big[\frac{i\,(i+\alpha)\,(i-1/2)\,(i+\alpha-1/2)}{(i-1)\,(i-1+\alpha)\,(i+1/2)\,(i+\alpha+1/2)}\Big]^{n+1-i} .$$

Observing the relations

$$\prod_{i=2}^{n}\Big[\frac{i}{i-1}\Big]^{n+1-i} = n!\,,$$
$$\prod_{i=2}^{n}\Big[\frac{i-1/2}{i+1/2}\Big]^{n+1-i} = \frac{3^n}{(2n+1)!!}\,,$$
$$\prod_{i=2}^{n}\Big[\frac{i+\alpha}{i-1+\alpha}\Big]^{n+1-i} = \frac{\Gamma(n+1+\alpha)}{(1+\alpha)^n\Gamma(1+\alpha)}\,,$$
$$\prod_{i=2}^{n}\Big[\frac{i+\alpha-1/2}{i+\alpha+1/2}\Big]^{n+1-i} = \frac{(3+2\alpha)^n\Gamma(3/2+\alpha)}{2^n\Gamma(n+3/2+\alpha)}\,.$$

we obtain

$$\begin{aligned}\frac{H_n}{G_n} &= \Big[\frac{4(1+\alpha)}{3\,(3+2\,\alpha)}\Big]^n \frac{n!3^n\Gamma(n+1+\alpha)(3+2\alpha)^n\Gamma(3/2+\alpha)}{(2n+1)!!(1+\alpha)^n\Gamma(1+\alpha)2^n\Gamma(n+3/2+\alpha)}\\ &= \frac{\sqrt{\pi}}{2^{2\alpha+1}B(\alpha+1,\alpha+1)} \times \frac{(2n)!!}{(2n+1)!!}\cdot\frac{\Gamma(n+1+\alpha)}{\Gamma(n+3/2+\alpha)}\\ &= \frac{\pi}{2^{2\alpha+2}B(\alpha+1,\alpha+1)} \times \frac{1}{n}\Big(1+O\Big(\frac{1}{n}\Big)\Big),\ n\to\infty \end{aligned} \tag{4.14}$$

where the expansion in the last line follows by straightforward but tedious calculation using Stirling's formula.

We finally mention the special cases $\alpha = 0$ (the spectral measure is a uniform spectral density on the interval $[-1, 1]$ with corresponding kernel function $k(x) = \sin(x)/x$ and $\alpha = -1/2$ (the spectral measure is the arcsine distribution on $[-1, 1]$ and the corresponding kernel is $k(x) = 2J_1(x)/x$, where $J_\alpha(\cdot)$ is the Bessel function of the first kind) for which the expansions are given, respectively, by

$$\begin{aligned}\frac{H_n}{G_n} &= \Big[\frac{(2n)!!}{(2n+1)!!}\Big]^2 = \frac{\pi}{4n} + O\Big(\frac{1}{n^2}\Big),\\ \frac{H_n}{G_n} &= \Big(\frac{8}{3}\Big)^n\Big[\frac{1}{8}\Big]^n n!\frac{3^n}{(2n+1)!!}\frac{2^n\Gamma(n+1/2)}{\Gamma(1/2)}\frac{1}{n!} = \frac{1}{2n+1} = \frac{1}{2n} + O\Big(\frac{1}{n^2}\Big)\end{aligned} \tag{4.15}$$

as $n\to\infty$. Interestingly, the ratio H_n/G_n in (**??**) is the squared ratio H_n/G_n of (**??**).

5. Some results of numerical studies and discussions. We have made extensive numerical studies to assess the uncertainty quantification in GP regression, as introduced in Section **??** for functions $f \in H(K)$ and $f \notin H(K)$; some of our results are illustrated in the figures below. At the end of this section, we summarize our conclusions. Different kernels K (including Matérn kernels and the kernels discussed at the end of Section ??) have been investigated as well. In the figures below, we use $X = [0, 1]$, Gaussian and Cauchy kernels (see (E.1) and (E.2) at the end of Section ??) and the two functions

$$f_1(x) = \exp\{-2(x - 1/3)^2\} \quad \text{and} \quad f_2(x) = 1 - 2(x - 1/3)^2.$$

These two functions look similar but we note that $f_1 \in H(K)$ and $f_2 \notin H(K)$ for both kernels with the correlation lengths considered below. Visually, the chosen kernels also look similar but it turns out that they exhibit completely different behaviour.

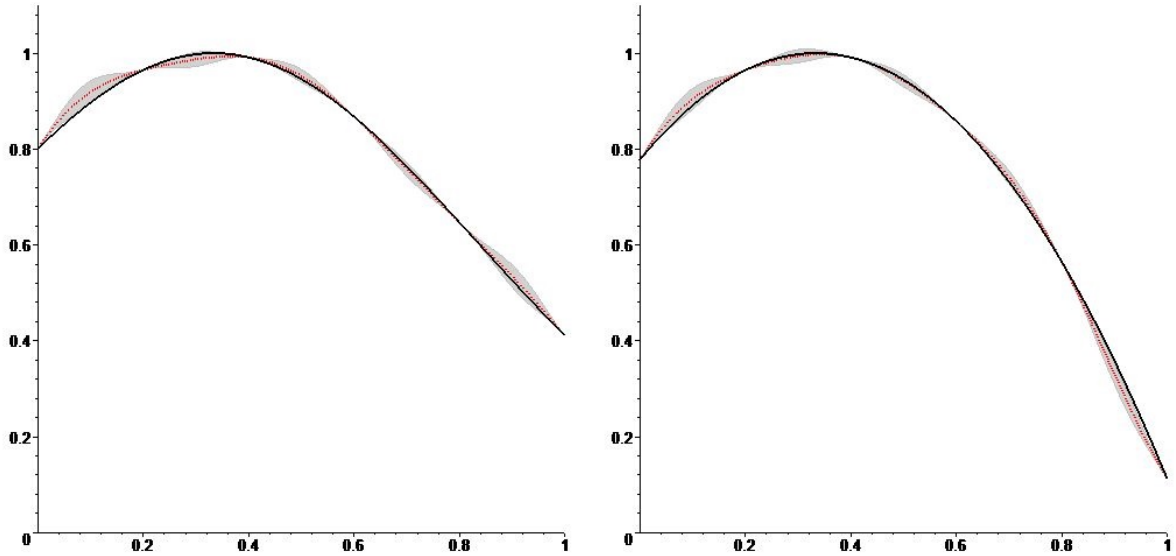

Figure 1. *Kriging confidence regions for kernel approximation of f_1 (left) and f_2 (right): Gaussian kernel, $\lambda = 15$, $N = 6$.*

In Figs. ?? and ?? we plot either f_1 or f_2 in solid black, the kernel approximation $\mu_N(x)$ computed by (??) in dotted red and the so-called kriging confidence regions $\mu_N(x) \pm 3\widehat{\sigma_N^2}\, C_N(x, x)$ in grey, where $C_N(x, x)$ is the kernel variance computed by (??) and $\widehat{\sigma_N^2}$ is the MLE of σ^2 computed by (??). The main reason for providing Figs. ?? and ?? is the demonstration of the big difference in the width of the confidence regions for the Gaussian and Cauchy kernels. In Figs. ??, ?? and ??, we plot the deviation $f(x) - \mu_N(x)$ in brown and confidence bounds $\mu_N(x) - f(x) \pm 3\widehat{\sigma_N^2}\, C_N(x, x)$ in filled grey. Again, the left and right panels show the results for the functions f_1 and f_2, respectively. The points x_j, where observations of f are taken, are equally spaced on the interval $[0, 1]$ with $x_j = (j - 1)/(N - 1)$, $j = 1, \ldots, N$.
The results for the Gaussian kernel $K(x, y) = \sigma^2 \exp\{-15(x - y)^2\}$ are depicted on Figs. ?? and ??. The corresponding results for the Cauchy kernel $K(x, y) = \sigma^2/(1 + 20(x - y)^2)$ can

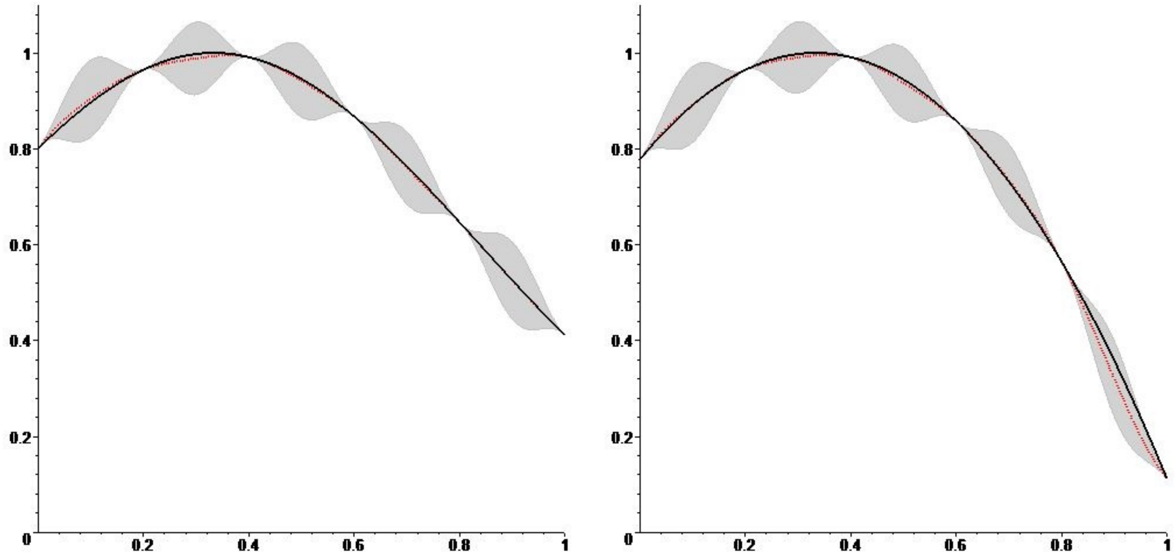

Figure 2. *Kriging confidence regions for kernel approximation of f_1 (left) and f_2 (right): Cauchy kernel, $\lambda = 20$, $N = 6$.*

be found Figs. **??** and **??**. It is clear from comparing left and right panels in Figs. **??**–**??** that kernel approximations for $f_1 \in H(K)$ are significantly more accurate than for $f_2 \notin H(K)$. The two chosen kernels (Gaussian with $\lambda = 15$ and Cauchy with $\lambda = 20$) look very similar but have different tail behaviour of the corresponding spectral density: the tail of the spectral density of the Cauchy kernel has a heavier tail. The confidence regions for the regression function constructed by the Cauchy kernel are rather wide and resemble the regions for the Matérn kernels with shape parameters 3/2 and 5/2 having similar correlation lengths. The confidence regions in the case of Gaussian kernel are much narrower (in fact, far too narrow) and the confidence regions for the kernels in (E.3) and (E.4) of Section **??** are even narrower; the spectral measures for these kernels have finite support.

In Figure **??** we plot the deviations and confidence regions for kernel approximation with Gaussian kernel $K(x, y) = \sigma^2 \exp\{-2(x - y)^2\}$; the Gaussian kernel with $\lambda = 2$ is perfectly suited for function f_1. A naive visual inspection of the two functions might suggest the Gaussian kernel should also work well for f_2 but, as we can observe from Figure **??** (right), it is not so for $f_2 \notin H(K)$. The confidence region on Figure **??** (right) cannot be seen as the deviations $|\mu_N(x) - f_2(x)|$ are on average 10^5 times larger than $3\widehat{\sigma_N^2}\, C_N(x, x)$.

From the numerical studies partially illustrated in these figures we make the following conclusions concerning uncertainty quantification in GP regression models with infinitely differentiable translation-invariant kernels:

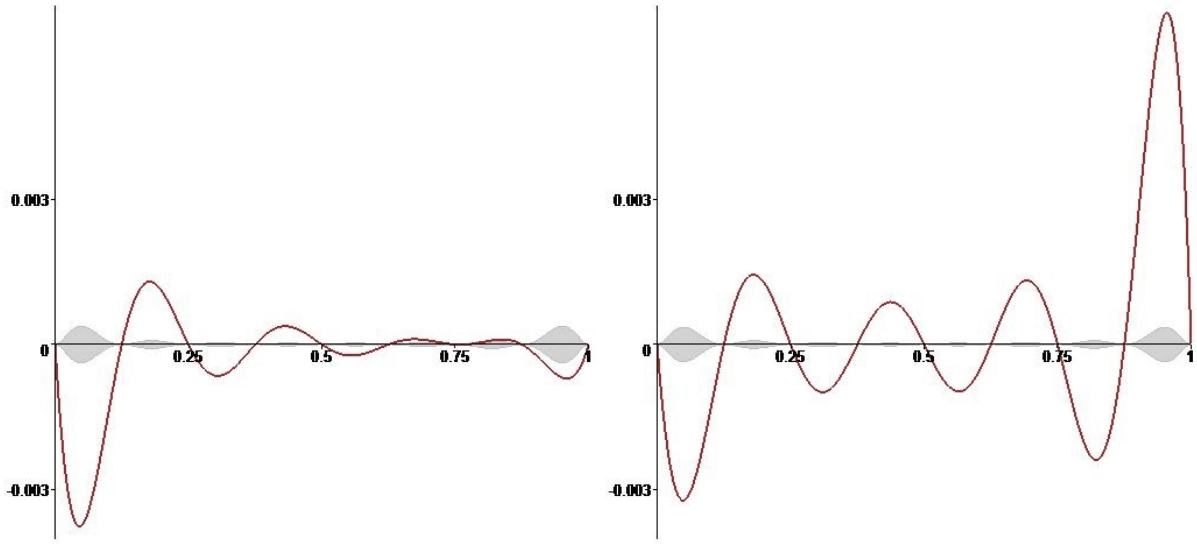

Figure 3. *Deviation and confidence bounds in kernel approximation: Gaussian kernel, $\lambda = 15$, $N = 9$; f_1 (left) and f_2 (right)*

- if $f \notin H(K)$, then the kriging confidence regions for f are always inaccurate;
- the heavier are the tails of the spectral measure of the kernel, the wider are the confidence regions;
- if the tail of the spectral measure of the kernel is light and the function f does not belong to the respective RKHS, then the kernel approximation of f appears to be rather inaccurate and the confidence regions seem to be missing f almost entirely;
- if the function f does not belong to the respective RKHS, then $\widehat{\sigma_N^2} \to \infty$ as $N \to \infty$, but this has little effect on the size of the confidence intervals, at least for small N;
- for kernels with light tails of the respective spectral measures, the kernel approximation is accurate and confidence regions are adequate only if the shape of f precisely matches the shape of the kernel functions $K(x, \cdot)$, as in Figure **??** (left).

Acknowledgements This research was partially supported by the research grant DFG DE 502/27-1 of the German Research Foundation (DFG). The authors are grateful to Timo Karvonen (Alan Turing Institute) for intelligent discussions, spotting an essential typo and pointing out several important references. We would also like to thank both referees and especially the associate editor for their constructive and very valuable comments on earlier versions of this paper.

References.

Achieser, N. I. (1956). *Theory of Approximation.* Frederick Ungar Publishing Co., N.Y.

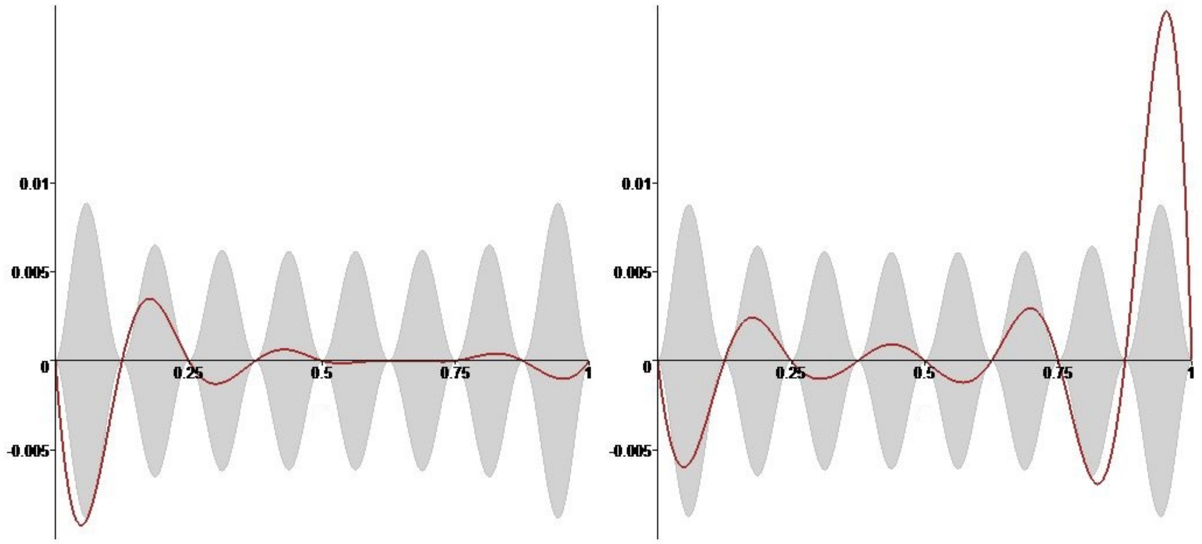

Figure 4. *Deviation and confidence bounds in kernel approximation: Cauchy kernel, $\lambda = 20$, $N = 9$; f_1 (left) and f_2 (right)*

Akhiezer, N. (1965). *The Classical Moment Problem and Some Related Questions in Analysis*. Oliver & Boyd.

Archetti, F. and Candelieri, A. (2019). *Bayesian Optimization and Data Science*. Springer.

Berg, C., Chen, Y., and Ismail, M. E. (2002). Small eigenvalues of large Hankel matrices: The indeterminate case. *Mathematica Scandinavica*, 90:67–81.

Berlinet, A. and Thomas-Agnan, C. (2011). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer.

Billingsley, P. (2008). *Probability and Measure*. John Wiley & Sons, forth edition.

Bochner, S. and Chandrasekharan, K. (1949). *Fourier Transforms. (AM-19)*. University Press, Princeton, N.J.

Briol, F.-X., Oates, C., Girolami, M., Osborne, M., and Sejdinovic, D. (2019). Probabilistic integration: A role in statistical computation? *Statistical Science*, 34(1):1–22.

Cambanis, S. and Masry, E. (1983). Sampling designs for the detection of signals in noise. *IEEE Transactions on Information Theory*, 29(1):83–104.

Dette, H., Pepelyshev, A., and Zhigljavsky, A. (2019). The BLUE in continuous-time regression models with correlated errors. *Annals of Statistics*, 47(4):1928–1959.

Dette, H. and Studden, W. (1992). On a new characterization of the classical orthogonal polynomials. *Journal of Approximation Theory*, 71(1):3–17.

Dette, H. and Studden, W. (1997). *The Theory of Canonical Moments with Applications in*

Statistics, Probability and Analysis. John Wiley & Sons.

Hennig, P., Osborne, M., and Girolami, M. (2015). Probabilistic numerics and uncertainty in computations. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2179):20150142.

Heyde, C. (1963). Some remarks on the moment problem (i). *The Quarterly Journal of Mathematics*, 14(1):91–96.

Kanagawa, M., Hennig, P., Sejdinovic, D., and Sriperumbudur, B. (2018). Gaussian processes and kernel methods: A review on connections and equivalences. arxiv e-prints, art. *arXiv preprint arXiv:1807.02582*.

Karvonen, T. (2021). On non-inclusion of certain functions in reproducing kernel Hilbert spaces. *arXiv preprint arXiv:2102.10628*.

Karvonen, T., Wynne, G., Tronarp, F., Oates, C., and Särkkä, S. (2020). Maximum likelihood estimation and uncertainty quantification for Gaussian process approximation of deterministic functions. *SIAM/ASA J. on Uncertainty Quantification*, 8(3):926–958.

Krantz, S. and Parks, H. (2002). *A Primer of Real Analytic Functions*. Springer.

Lau, T.-S. and Studden, W. (1988). On an extremal problem of Fejér. *Journal of Approximation Theory*, 53(2):184–194.

Lee, K.-Y., Li, B., and Zhao, H. (2016). Variable selection via additive conditional independence. *Journal of the Royal Statistical Society B*, 78(5):1037–1055.

Lin, G. D. (2017). Recent developments on the moment problem. *Journal of Statistical Distributions and Applications*, 4(1):1–17.

Minh, H. Q. (2010). Some properties of Gaussian reproducing kernel Hilbert spaces and their implications for function approximation and learning theory. *Constructive Approximation*, 32(2):307–338.

Parzen, E. (1961). An approach to time series analysis. *The Annals of Mathematical Statistics*, 32(4):951–989.

Paulsen, V. and Raghupathi, M. (2016). *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*. Cambridge University Press.

Pronzato, L. and Zhigljavsky, A. (2020). Bayesian quadrature, energy minimization, and space-filling design. *SIAM/ASA Journal on Uncertainty Quantification*, 8(3):959–1011.

Rasmussen, C. and Williams, C. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Ritter, K. (2007). *Average-Case Analysis of Numerical Problems*. Springer.

Ritter, K., Wasilkowski, G., and Woźniakowski, H. (1995). Multivariate integration and approximation for random fields satisfying Sacks-Ylvisaker conditions. *The Annals of Applied Probability*, 5(2):518–540.

Scheuerer, M., Schaback, R., and Schlather, M. (2013). Interpolation of spatial data–a stochastic or a deterministic problem? *European Journal of Applied Mathematics*, 24(04):601–629.

Schmüdgen, K. (2017). *The Moment Problem*. Springer.

Schulz, E., Speekenbrink, M., and Krause, A. (2018). A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology*, 85:1–16.

Shohat, J. and Tamarkin, J. (1943). *The Problem of Moments*. American Math. Soc.

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer.

Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Springer.

Steinwart, I., Hush, D., and Scovel, C. (2006). An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels. *IEEE Trans. on Inform. Theory*, 52(10):4635–4643.

Stoyanov, J. M. (2013). *Counterexamples in Probability*. Courier Corporation.

Sun, H.-W. and Zhou, D.-X. (2008). Reproducing kernel Hilbert spaces associated with analytic translation-invariant Mercer kernels. *J. Fourier Analysis and Applications*, 14(1):89–101.

Wainwright, M. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press.

Wendland, H. (2004). *Scattered Data Approximation*. Cambridge University Press.

Xu, W. and Stein, M. L. (2017). Maximum likelihood estimation for a smooth Gaussian random field model. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):138–175.

Yaglom, A. (1986). *Correlation Theory of Stationary and Related Random Functions. Volume I: Basic Results*. Springer.

Zhigljavsky, A. and Žilinskas, A. (2021). *Bayesian and High-Dimensional Global Optimization*. Springer.

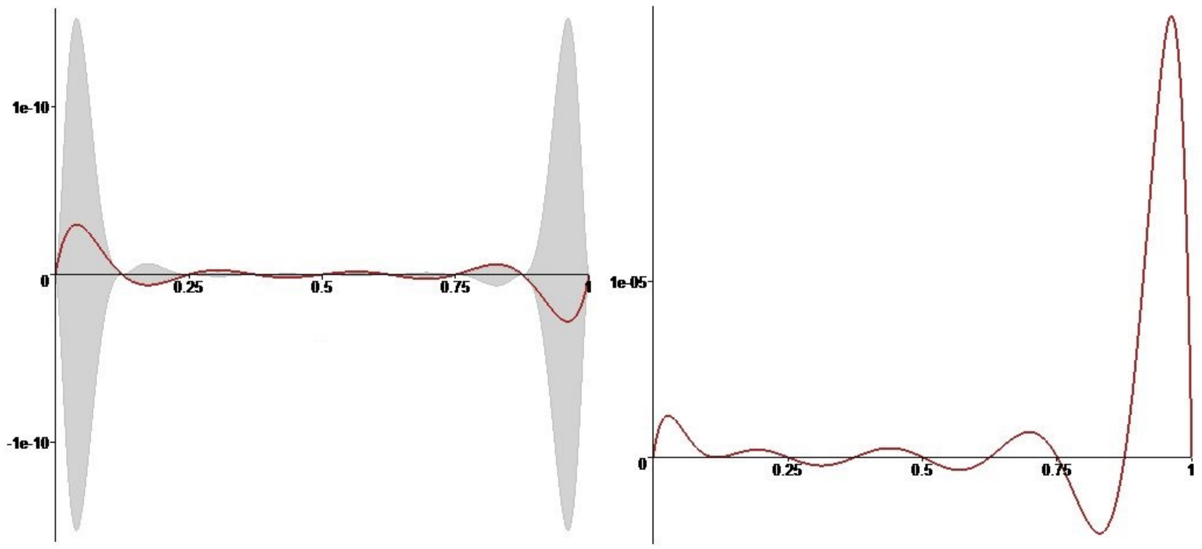

Figure 5. *Deviation and confidence bounds in kernel approximation: Gaussian kernel,* $\lambda = 2$, $N = 9$; f_1 *(left) and* f_2 *(right).*