PADOVA UNIVERSITY PRESS

Dolomites Research Notes on Approximation

Special Issue dedicated to Robert Schaback on the occasion of his 75th birthday, Volume 15 · 2022 · Pages 65–77

Error Bounds and the Asymptotic Setting in Kernel-Based Approximation

Toni Karvonen[1]

Communicated by Gabriele Santin

Abstract

We use ideas from Gaussian process regression to derive computable error bounds that can be used as stopping criteria in kernel-based approximation. The proposed bounds are based on maximum likelihood estimation and cross-validation of a kernel scale parameter and take the form of a product of the scale parameter estimate and the worst-case approximation error in the reproducing kernel Hilbert space induced by the kernel. We also use known results on the so-called asymptotic setting to argue that such worst-case type error bounds are not necessarily conservative.

1 Introduction

Let $K\colon \Omega\times\Omega\to\mathbb{R}$ be a strictly positive-definite kernel on an infinite set Ω and $H(K)$ its reproducing kernel Hilbert space (RKHS) equipped with an inner product $\langle\cdot,\cdot\rangle_K$ and the resulting norm $\|\cdot\|_K$. The RKHS is a Hilbert space consisting of real-valued functions defined on Ω such that

(i) kernel translates are elements of $H(K)$, in that $K(\cdot,x)\in H(K)$ for every $x\in\Omega$, and

(ii) the kernel has the *reproducing property*, which states that $\langle f, K(\cdot,x)\rangle_K = f(x)$ for every $f\in H(K)$ and $x\in\Omega$.

See [15] for a review of RKHSs. Let $\{x_i\}_{i=1}^{\infty}$ be a set of distinct points in Ω and $X_n=\{x_1,\ldots,x_n\}$ the set of first n of them. For every $n\in\mathbb{N}$ and any function $f:\Omega\to\mathbb{R}$ there exists a unique minimum-norm interpolant $I_nf = I_{X_n}f\in H(K)$:

$$I_nf = I_{X_n}f = \operatorname*{arg\,min}_{s\in H(K)}\left\{\|s\|_K : s|_{X_n} = f|_{X_n}\right\} = \mathbf{f}_n^{\mathsf{T}}\mathbf{K}_{n,n}^{-1}\mathbf{K}_n(\cdot), \tag{1}$$

where $\mathbf{f}_n = (f(x_1),\ldots,f(x_n))\in\mathbb{R}^n$, $\mathbf{K}_{n,n}\in\mathbb{R}^{n\times n}$ is the positive-definite kernel matrix with elements $(\mathbf{K}_{n,n})_{ij} = K(x_i,x_j)$ and $\mathbf{K}_n(\cdot) = (K(\cdot,x_1),\ldots,K(\cdot,x_n))\in\mathbb{R}^n$. This interpolant is often called the *kernel interpolant* or, if the kernel K is radial, the *radial basis function interpolant*. The kernel interpolant is the unique function in the span of $K(\cdot,x_1),\ldots,K(\cdot,x_n)$ that interpolates f at X_n. The *power function*

$$P_n(x) = P_{X_n}(x) = \sup_{\|f\|_K\leq 1}|f(x)-(I_nf)(x)| = \sup_{f\in H(K),f\neq 0}\frac{|f(x)-(I_nf)(x)|}{\|f\|_K} = \sqrt{K(x,x)-\mathbf{K}_n(x)^{\mathsf{T}}\mathbf{K}_{n,n}^{-1}\mathbf{K}_n(x)} \tag{2}$$

quantifies the approximation quality of the kernel interpolant in $H(K)$.

Let L be a bounded linear functional on $H(K)$. One can show that the approximation (or quadrature rule) of $L(f)$ obtained by applying L to the kernel interpolant I_nf is worst-case optimal in $H(K)$. That is, the *kernel quadrature rule*

$$Q_{L,n}(f) = L(I_nf) = \mathbf{f}_n^{\mathsf{T}}\mathbf{K}_{n,n}^{-1}L(\mathbf{K}_n) = \sum_{i=1}^n w_{n,i}f(x_i) \tag{3}$$

where $\mathbf{w}_n = (w_{n,1},\ldots,w_{n,n}) = \mathbf{K}_{n,n}^{-1}L(\mathbf{K}_n)\in\mathbb{R}^n$ are the quadrature weights, is the unique worst-case optimal linear approximation in $H(K)$ given standard information at X_n:

$$\mathbf{w}_n = \operatorname*{arg\,min}_{v_1,\ldots,v_n\in\mathbb{R}}\sup_{\|f\|_K\leq 1}\left|L(f)-\sum_{i=1}^n v_if(x_i)\right|.$$

It also follows that the worst-case error of $Q_{L,n}$ in $H(K)$ is

$$E_n(L) = \mathrm{wce}_{H(K)}(Q_{L,n}) = \sup_{\|f\|_K\leq 1}|L(f)-Q_{L,n}(f)| = \sup_{f\in H(K),f\neq 0}\frac{|L(f)-Q_{L,n}(f)|}{\|f\|_K} = \sqrt{K_{L,L}-L(\mathbf{K}_n)^{\mathsf{T}}\mathbf{K}_{n,n}^{-1}L(\mathbf{K}_n)}, \tag{4}$$

[1]Department of Mathematics and Statistics, University of Helsinki, Finland. Email: toni.karvonen@helsinki.fi.

where $K_{L,L} = L(K_L)$ and the function K_L is defined via $K_L(x) = L(K(\cdot, x))$. Note that the kernel interpolant and the power function at $x \in \Omega$ are recovered from (3) and (4) by selecting the point evaluation functional $L = \delta_x$ defined as $\delta_x(f) = f(x)$. That is, $Q_{\delta_x,n}(f) = (I_n f)(x)$ and $E_n(\delta_x) = \mathrm{wce}_K(Q_{\delta_x,n}) = P_n(x)$ for every $x \in \Omega$. For a more thorough review of kernel-based interpolation and approximation, see [13, 26] and [7, Chapter 8] as well as [12, Chapter 10].

From (2) and (4) it immediately follows that

$$|f(x) - (I_n f)(x)| \leq \|f\|_K P_n(x) \quad \text{and} \quad |L(f) - Q_{L,n}(f)| \leq \|f\|_K E_n(L) \tag{EB}$$

for every $f \in H(K)$ and $x \in \Omega$. As is well known [6, Section 5.1], these error bounds can be improved to

$$|f(x) - (I_n f)(x)| \leq \|f - I_n f\|_K P_n(x) \quad \text{and} \quad |L(f) - Q_{L,n}(f)| \leq \|f - I_n f\|_K E_n(L) \tag{I-EB}$$

for every $f \in H(K)$ and $x \in \Omega$. The first bound in (I-EB) is proved by setting $f = f - I_n f$ in (EB) and observing that $I_n(f - I_n f)$ is identically zero because $f - I_n f$ vanishes on X_n while the second one follows from the same argument combined with $L(f) - Q_{L,n}(f) = L(f - I_n f)$. Note that $\|f - I_n f\|_K \leq \|f\|_K$ since $f - I_n f$ and $I_n f$ are $H(K)$-orthogonal, which one can prove by using (1) and the reproducing property. It is a common scenario that, after being provided with an absolute error tolerance $\varepsilon > 0$ and a function f, an approximation algorithm $Q_n(f)$ (e.g., any standard numerical integration routine) proceeds to increase n until its internal error estimation indicates that $|L(f) - Q_n(f)| \leq \varepsilon$ holds. Due to the presence of a worst-case error which is available in closed form as long as K_L and $K_{L,L}$ can be computed, it would be tempting to use the error bounds (EB) or (I-EB) to terminate a kernel-based approximation method. However, two problems arise:

(P1) The bounds are *worst-case* and thus *potentially conservative*. The bounds in (EB) are clearly sub-optimal for fixed f because the rates of decay of the right-hand sides do not depend on f. The improved bounds in (I-EB) are obviously better, but it is still not clear if such worst-case type bounds are optimal in some sense.

(P2) The bounds, even if they are accepted to be useful, are *not computable*. Although the worst-case error has a simple linear-algebraic expression given in (2) or (4), computation or estimation of either of the norms $\|f\|_K$ or $\|f - I_n f\|_K$ is not possible with complete certainty, in the sense that there does not exist a function $c\colon \mathbb{R}^n \to [0, \infty)$ satisfying $\|f\|_K \leq c(f|_{X_n})$ or $\|f - I_n f\|_K \leq c(f|_{X_n})$ for every $f \in H(K)$ (see Proposition 3.1). One therefore has to be content with bounds that fail for some elements of the RKHS. Some discussion of this topic in the context of kernel-based interpolation can be found in [6, Section 5.1].

The design of termination rules is thus a constant tug of war: a *conservative* rule (i.e., the error bounds are "large") may be *comprehensive*, in that it terminate early for few elements of the function space of interest but runs the risk of terminating far too late for most elements, and thus wasting computational resources; an *optimistic* rule (i.e., the bounds are "small") saves computational resources but may terminate early for many elements and thus result in overconfidence in the quality of approximation.

The purpose of this article is to discuss these two problems, recall certain theoretical results on the relation between the worst-case and asymptotic settings of error analysis and propose a solution, which by the nature of the task is bound to be to some extent unsatisfactory and heuristic:

- Section 2 recalls a result by Trojan from [24, Chapter 10] which, in the language of this article, states that the RKHS contains functions for which $|L(f) - Q_{L,n}(f)|$ decays (as $n \to \infty$) with a rate that is arbitrarily close to the rate of decay of $E_n(L)$. This provides a form of a resolution, albeit weak, to (P1): it is not a problem to make use of bounds based on worst-case analysis because, uniformly over $f \in H(K)$, the rate of decay of the worst-case error is *not* slower than that for individual fixed f. Section 2 owes its existence to the recent work of Owen and Pan [14] who were directed to Trojan's result in [24] by Erich Novak (see also [16, Chapter 6]).
- Section 3, which takes its inspiration from recent work on approximation of deterministic functions with Gaussian processes [8, 25], proposes using maximum likelihood estimation and cross-validation to construct error bounds which, while optimistic, are "likely" to be valid for "many" elements of the RKHS. The proposed error bounds are $c(f, n)P_n(x)$ and $c(f, n)E_n(L)$ with

$$c(f,n) = c_{\mathrm{ML}}(f,n) = \sqrt{\frac{\mathbf{f}_n^{\mathsf{T}} \mathbf{K}_{n,n}^{-1} \mathbf{f}_n}{n}} \quad \text{or} \quad c(f,n) = c_{\mathrm{CV}}(f,n) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \frac{[f(x_i) - (I_{n,i}f)(x_i)]^2}{P_{n,i}(x_i)^2}}, \tag{5}$$

  where $I_{n,i}$ and $P_{n,i}$ stand for the kernel interpolant and the power function based on the points $X_n \setminus \{x_i\}$. The justification for the use of the coefficients in (5) is related to the $d/2$-gap observed by Schaback and Wendland [21] and the supercon-vergence of kernel-based approximation [20]. A numerical example in Section 4 demonstrates that these error bounds are not completely without merit. Stopping criteria derived from maximum likelihood estimation and cross-validation have been recently used in kernel-based integration by Rathinavel and Hickernell [18].

Most theoretical results in this article are well known (and more general, requiring only that the function space be a Hilbert or Banach space) while those that are new are not particularly difficult to prove. Importantly, our arguments for the use of the coefficients in (5) are not mathematically rigorous and it is unclear how, and in what sense, they could be made rigorous. Although we focus on functions in the RKHS, the misspecification results in [1, 8, 11, 28] make some of the discussion applicable also to functions outside the RKHS if the RKHS is norm-equivalent to a Sobolev space.

Let λ and ν be positive parameters. Throughout the article the stationary Matérn kernel

$$K(x,y) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}\|x-y\|}{\lambda}\right)^{\nu} \mathcal{K}_\nu\left(\frac{\sqrt{2\nu}\|x-y\|}{\lambda}\right) \quad \text{for} \quad x, y \in \mathbb{R}^d, \tag{6}$$

where Γ is the Gamma function and $\mathcal{K}_\nu$ the modified Bessel function of the second kind of order ν, is used as an example because its approximation properties are well understood. On any metric space Ω define the fill-distance

$$h_{n,\Omega} = \sup_{x\in\Omega} \min_{i=1,\dots,n} \|x - x_i\|.$$

On any sufficiently regular subset Ω of $\mathbb{R}^d$ (e.g., $\Omega = [0,1]^d$) the RKHS of the Matérn kernel (6) is norm-equivalent to the fractional Sobolev space $H^{\nu+d/2}(\Omega)$. The parameter λ only affects the norm-equivalence constants. Assume that Ω is bounded and let L be the integration functional

$$L(f) = \int_\Omega f(x)p(x)\,\mathrm{d}x \quad \text{for any} \quad p \in L^\infty(\Omega).$$

It is well known that there is a constant $C > 0$ such that

$$\sup_{x\in\Omega} |f(x) - (I_n f)(x)| \leq C h_{n,\Omega}^{\nu} \|f\|_{H^{\nu+d/2}(\Omega)} \quad \text{and} \quad |L(f) - Q_{L,n}(f)| \leq C h_{n,\Omega}^{\nu+1/2} \|f\|_{H^{\nu+d/2}(\Omega)} \tag{7}$$

for every $n \in \mathbb{N}$ and $f \in H^{\nu+d/2}(\Omega)$. If the point set $\{x_i\}_{i=1}^\infty$ is quasi-uniform, in that $h_{n,\Omega} = \Theta(n^{-1/d})$, these bounds become

$$\sup_{x\in\Omega} |f(x) - (I_n f)(x)| \leq C n^{-\nu/d} \|f\|_{H^{\nu+d/2}(\Omega)} \quad \text{and} \quad |L(f) - Q_{L,n}(f)| \leq C n^{-\nu/d-1/2} \|f\|_{H^{\nu+d/2}(\Omega)} \tag{8}$$

for a different constant $C > 0$. In the quasi-uniform case it also holds that

$$\sup_{x\in\Omega} P_n(x) = \Theta(n^{-\nu/d}) \quad \text{and} \quad E_n(L) = \Theta(n^{-\nu/d-1/2}). \tag{9}$$

There is a non-negligible subset of $H^{\nu+d/2}(\Omega)$ for which the algebraic rates in (7) and (8) can be essentially doubled [26, Section 11.5]. Note that the rates above are valid not only for Matérn kernels but for any kernel whose RKHS is norm-equivalent to a Sobolev space.

2 Asymptotic Setting

The worst-case error $E_n(L) = \sup_{\|f\|_K \leq 1} |L(f) - Q_{L,n}(f)|$ in (4) is by its very nature adversarial: for each $n \in \mathbb{N}$ there is a fooling function f_n in the unit ball of $H(K)$ for which $|L(f_n) - Q_{L,n}(f_n)| = E_n(L)$ and, importantly, this function can depend on n. But in the *asymptotic setting* [24, Chapter 10] that this article is concerned with there is a single fixed $f \in H(K)$ for which the error is to be estimated. It is not unreasonable to expect that the the worst-case error decays slower than the error for a fixed element of the RKHS, in that

$$\lim_{n\to\infty} \frac{|L(f) - Q_{L,n}(f)|}{E_n(L)} = 0 \quad \text{for every} \quad f \in H(K), \tag{10}$$

provided that the points $\{x_i\}_{i=1}^\infty$ are suitable. This is indeed the case because from (I-EB) one obtains

$$\frac{|L(f) - Q_{L,n}(f)|}{E_n(L)} \leq \frac{\|f - I_n f\|_K E_n(L)}{E_n(L)} = \|f - I_n f\|_K \tag{11}$$

and, as is well known, $\|f - I_n f\|_K$ tends to zero as $n \to \infty$ if and only if the power function does. We include a proof of this result for completeness (see, e.g., Theorem 8.37 in [7] for the case of a continuous K).

Proposition 2.1. *The following statements are equivalent:*

(i) $\lim_{n\to\infty} P_n(x) = 0$ *for every* $x \in \Omega$.

(ii) $\lim_{n\to\infty} \|f - I_n f\|_K = 0$ *for every* $f \in H(K)$.

Moreover, if Ω is a metric space and $K : \Omega \times \Omega \to \mathbb{R}$ is continuous, then (i) and (ii) are implied by $\{x_i\}_{i=1}^\infty$ being dense in Ω.

Proof. Assume that (i) holds and let $f \in H(K)$. From (EB) it follows that

$$|f(x) - (I_n f)(x)| \leq \|f\|_K P_n(x) \to 0 \quad \text{as} \quad n \to \infty$$

for every $x \in \Omega$, so that $I_n f \to f$ pointwise. Because $I_{n+1} f|_{X_n} = I_n f$, it follows from the minimum-norm interpolation property (1) that (a) the sequence $(\|I_n f\|_K)_{n=1}^\infty$ is increasing and (b) $\|I_n f\|_K \leq \|f\|_K$ for every $n \in \mathbb{N}$. Therefore there is $g \in H(K)$ such that $\lim_{n\to\infty} \|g - I_n f\|_K = 0$ since $H(K)$ is a Hilbert space and thus every Cauchy sequence in $H(K)$ tends to an element of $H(K)$. From the reproducing property and the Cauchy–Schwarz inequality it then follows that

$$|g(x) - (I_n f)(x)| = |\langle g - I_n f, K(\cdot, x)\rangle_K| \leq \|g - I_n f\|_K \sqrt{K(x,x)} \to 0 \quad \text{as} \quad n \to \infty$$

for every $x \in \Omega$, so that $I_n f \to g$ pointwise. Therefore $g = f$ and we conclude that $\lim_{n\to\infty} \|f - I_n f\|_K = 0$. That (ii) implies (i) follows from writing, by using, for example, the reproducing property and the last equality in (2), the power function as $P_n(x) = \|f - I_n f\|_K$ for the function $f = K(\cdot, x) - I_n K(\cdot, x) \in H(K)$.

Finally, suppose that (Ω, d_Ω) is a metric space and K is continuous. Since $f - I_n f$ vanishes on X_n and has $H(K)$-norm at most $\|f\|_K$, it follows from the worst-case characterisation of the power function in (2) that[1]

$$P_n(x) = \sup_{\|f\|_K \leq 1} |f(x) - (I_n f)(x)| \leq \sup\{|f(x)| : \|f\|_K \leq 1 \text{ and } f|_{X_n} = 0\} \leq \sup\{|f(x)| : \|f\|_K \leq 1 \text{ and } f(x_m) = 0\} = P_{\{x_m\}}(x)$$

for any $m \leq n$, where $P_{\{x_m\}}(x)$ denotes the power function based on a single point. The explicit linear algebraic expression for the power function in (2) then yields

$$P_n(x) \leq P_{\{x_m\}}(x) = \sqrt{K(x_m, x_m) - \frac{K(x, x_m)^2}{K(x_m, x_m)}}. \tag{12}$$

If $\{x_i\}_{i=1}^\infty$ is dense in Ω, for every $x \in \Omega$ there exists a subsequence $(i_n)_{n=1}^\infty$ such that $i_n \leq n$ and $d_\Omega(x, x_{i_n}) \to 0$ as $n \to \infty$. The continuity of K and (12) with $x_m = x_{i_n}$ yield the last claim. □

However, it is also well known that $\|f - I_n f\|_K$ can tend to zero arbitrarily slowly, in that for every positive sequence $(\delta_n)_{n=1}^\infty$ tending to zero there exists $f \in H(K)$ such that $\|f - I_n f\|_K \geq \delta_n$ for every $n \in \mathbb{N}$ [e.g., 7, Exercise 8.64]. This result (a somewhat roundabout proof of a version of which is given below) and (11) suggest that the error bounds in (EB) are optimal even in the asymptotic setting in the sense that the rates of decay of $|L(f) - Q_{L,n}(f)|$ and $E_n(L)$ can be arbitrarily close. That this is indeed true is confirmed by the following theorem, which is an adaptation to the kernel setting of Theorem 2.1.1 in [24]. This theorem is apparently originally due to Trojan and was brought to my attention by a recent article on quasi-Monte Carlo integration by Owen and Pan [14], who in turn were informed about its existence by Erich Novak.

Theorem 2.2 (Trojan; Theorem 2.1.1 in [24]). *For any positive sequence $(\delta_n)_{n=1}^\infty$ tending to zero the set*

$$A = \left\{ f \in H(K) : \lim_{n\to\infty} \frac{|L(f) - Q_{L,n}(f)|}{\delta_n E_n(L)} = 0 \right\}$$

has empty interior in the norm of $H(K)$. Here we use the convention $0/0 = 1$.

Theorem 2.2 states that there are very few functions in $H(K)$ for which the rate of decay of the error, $|L(f) - Q_{L,n}(f)|$, is faster than that of the worst-case error, $E_n(L)$. Indeed, by Theorem 2.2 the set

$$A^\mathsf{c} = \left\{ f \in H(K) : \limsup_{n\to\infty} \frac{|L(f) - Q_{L,n}(f)|}{\delta_n E_n(L)} > 0 \right\} \tag{13}$$

is dense in $H(K)$—both in the RKHS norm and the supremum norm since the RKHS norm is stronger of the two—for every positive sequence $(\delta_n)_{n=1}^\infty$ tending to zero.

Corollary 2.3. *For any positive sequence $(\delta_n)_{n=1}^\infty$ tending to zero there is $f \in H(K)$ such that*

$$\limsup_{n\to\infty} \frac{|L(f) - Q_{L,n}(f)|}{\delta_n E_n(L)} = \infty,$$

where the convention $0/(\delta_n \times 0) = \delta_n^{-1}$ is used.

Proof. Let $(\delta_n')_{n=1}^\infty$ be any positive sequence tending to zero. By Theorem 2.2 there is $f \in H(K)$ such that

$$\lim_{n\to\infty} r_n(f) = \lim_{n\to\infty} \frac{|L(f) - Q_{L,n}(f)|}{\delta_n' E_n(L)} = 0$$

does *not* hold. That is, there is $f \in H(K)$ such that $\limsup_{n\to\infty} r_n(f) > 0$. Let $(\delta_n)_{n=1}^\infty$ be any positive sequence which tends to zero and set $\delta_n' = \sqrt{\delta_n}$. Then

$$\limsup_{n\to\infty} \frac{|L(f) - Q_{L,n}(f)|}{\delta_n E_n(L)} = \limsup_{n\to\infty} \frac{|L(f) - Q_{L,n}(f)|}{\sqrt{\delta_n}\delta_n' E_n(L)} = \limsup_{n\to\infty} \frac{r_n(f)}{\sqrt{\delta_n}} = \infty,$$

which proves the claim. □

Corollary 2.4. *For any non-negative sequence $(\delta_n)_{n=1}^\infty$ tending to zero there is $f \in H(K)$ such that $\|f - I_n f\|_K \geq \delta_n$ for infinitely many $n \in \mathbb{N}$.*

Proof. It is sufficient to consider positive sequences because $\|f - I_n f\|_K \geq \delta_n$ holds trivially if $\delta_n = 0$. By (I-EB) and Corollary 2.3 for any positive decreasing sequence $(\delta_n)_{n=1}^\infty$ tending to zero there is $f \in H(K)$ such that

$$\frac{1}{\delta_n} \|f - I_n f\|_K \geq \frac{|L(f) - Q_{L,n}(f)|}{\delta_n E_n(L)} \geq 1$$

for infinitely many $n \in \mathbb{N}$. This proves the claim. □

[1] In fact, the first inequality below is an equality [e.g., 10, Satz 2.2.14].

Wenzel et al. [27] have essentially proved the above results in a very explicit way for the Wendland kernel

$$K(x,y) = \max\{1-|x-y|,0\}$$

defined on $\Omega = [0,1]$. For this kernel the supremum of the power function decays as $\Theta(n^{-1/2})$ if the points are quasi-uniform. In Section 6.2 of [27] it is shown that for function $f_\alpha(x) = x^\alpha$ with $\alpha \in (1/2,1)$ it holds that

$$\sup_{x\in[0,1]} |f_\alpha(x) - (I_n f_\alpha)(x)| \geq C_\alpha n^{-\alpha}$$

for a positive constant C_α and all $n \in \mathbb{N}$ if the points are quasi-uniform.

3 Computable Error Bounds

The results in Section 2 demonstrate that there is no gap between the rate of convergence of the worst-case error and the error for fixed elements of the RKHS, which confirms that (EB) and (I-EB) can be used as a basis of computable error bounds. We therefore turn our attention to constructing constants $c(f,n) \geq 0$, which depends on the values of f at X_n, such that

$$|L(f) - Q_{L,n}(f)| \leq c(f,n)\, E_n(L) \quad \text{with "high confidence" for} \quad f \in H(K). \tag{14}$$

The meaning of "high confidence" is, of course, bound to be quite heuristic and no attempt at a rigorous definition will be made in this article. First, one should dispense of any notion that the bound (14) can hold for *all* $f \in H(K)$. We supply a proof in the RKHS setting of this basic fact of numerical analysis that no error bound that holds for all elements of a sufficiently rich functions space can be constructed out of partial information.

Proposition 3.1. *Let $n \in \mathbb{N}$.*

(i) There does not exist a function $c\colon \mathbb{R}^n \to [0,\infty)$ such that $\|f - I_n f\|_K \leq c(f|_{X_n})$ for every $f \in H(K)$.

(ii) Suppose that there is $f \in H(K)$ such that $f|_{X_n} = 0$ and $L(f) \neq 0$. Then there does not exist a function $\varepsilon\colon \mathbb{R}^n \to [0,\infty)$ such that $|L(f) - Q_{L,n}(f)| \leq \varepsilon(f|_{X_n})$ for every $f \in H(K)$.

Proof. Let $f \not\equiv 0$ be any function in $H(K)$ that vanishes at X_n, the existence of which follows from K being strictly positive-definite and Ω being an infinite set. For example, one can take f to be the kernel interpolant $I_{n+1}g$ of any function g such that $g|_{X_n} = 0$ and $g(x_{n+1}) \neq 0$. Set $f_a = af$ for $a > 0$, so that $\|f_a - I_n f_a\|_K = a\|f\|_K > 0$ but $c(f_a|_{X_n}) = c(f|_{X_n}) = c(0)$. Therefore the inequality $\|f_a - I_n f_a\|_K \leq c(f_a|_{X_n})$ is violated for a sufficiently large a, which proves (i). To prove (ii), note that by assumption the function $f \in H(K)$ as above can be selected such that $L(f) \neq 0$. Then $|L(f) - Q_{L,n}(f)| = |L(f)| > 0$, and the rest of the proof is analogous to that of (i). □

The assumption in (ii) of Proposition 3.1 rules out L being a point evaluation functional $\delta_x(f) = f(x)$ for $x \in X_n$. To construct $c(f,n)$ in (14) we use ideas from Gaussian process regression. Sections 3.1 and 3.2 use maximum likelihood estimation and cross-validation, respectively, of kernel hyperparameters to construct this coefficient. One does not however have to think in terms of Gaussian processes because, as we show, the coefficients also arise from non-rigorous approximation theoretic reasoning.

3.1 Maximum Likelihood Estimation

It is well known that Gaussian process regression (or kriging) is equivalent to kernel-based approximation, in that the conditional mean and variance that one obtains after conditioning the Gaussian process prior on point evaluations at X_n equal the kernel quadrature rule and the squared worst-case error [e.g., 23]. In statistics and machine learning, *maximum likelihood estimation* is a popular method to select the parameters θ of a parametric kernel K_θ [17, Section 5.4.1] from a set Π of feasible parameters. For recent examples on the use of maximum likelihood estimation to select the shape parameter $\lambda > 0$, as in the Matérn kernel (6), in radial basis function literature, see [2] and [5, Section 9.4.3].

If f is modelled as a zero-mean Gaussian process with covariance kernel K_θ, the marginal likelihood of the *data* $\mathbf{f}_n \in \mathbb{R}^n$ given the parameter θ is

$$\det(2\pi \mathbf{K}_{\theta,n,n})^{-1/2} \exp\left(-\frac{1}{2}\mathbf{f}_n^{\mathsf{T}} \mathbf{K}_{\theta,n,n}^{-1} \mathbf{f}_n\right),$$

where the subscript θ is used to denote the kernel matrix for the parametric kernel K_θ. Maximisation of the marginal likelihood is equivalent to minimisation of the negative log-likelihood

$$\ell_{\mathrm{ML}}(\theta) = \mathbf{f}_n^{\mathsf{T}} \mathbf{K}_{\theta,n,n}^{-1} \mathbf{f}_n + \log\det \mathbf{K}_{\theta,n,n}. \tag{15}$$

Any maximum likelihood estimate θ_{ML} of θ, which in general need not be unique, therefore satisfies $\theta_{\mathrm{ML}} \in \arg\min_{\theta\in\Pi} \ell_{\mathrm{ML}}(\theta)$. Consider then the parameterisation $\theta = \sigma$ and $K_\sigma(x,y) = \sigma^2 K(x,y)$ for a *scale parameter* $\sigma > 0$. It is straightforward to compute that the unique minimiser of (15) is

$$c_{\mathrm{ML}}(f,n) = \sigma_{\mathrm{ML}} = \arg\min_{\sigma>0} \ell_{\mathrm{ML}}(\sigma) = \sqrt{\frac{\mathbf{f}_n^{\mathsf{T}} \mathbf{K}_{n,n}^{-1} \mathbf{f}_n}{n}} = \frac{\|I_n f\|_K}{\sqrt{n}}, \tag{16}$$

where the equality $\mathbf{f}_n^{\mathsf{T}} \mathbf{K}_n^{-1} \mathbf{f}_n = \|I_n f\|_K^2$ follows from (1) and the reproducing property. Behaviour of $c_{\mathrm{ML}}(f,n)$ for both functions within and without the RKHS has been studied in [8] and [25]. The following proposition collects basic properties of $c_{\mathrm{ML}}(f,n)$.

Proposition 3.2. *If $f \in H(K)$, then*

$$\frac{1}{\sqrt{n}}\left(\frac{|f(x_m)|}{\sqrt{K(x_m,x_m)}}\right) \leq c_{\mathrm{ML}}(f,n) \leq \frac{\|f\|_K}{\sqrt{n}} \quad \text{for every} \quad m \in \mathbb{N} \text{ and } n \geq m. \tag{17}$$

Moreover, if Ω is a metric space, K is continuous and $\{x_i\}_{i=1}^{\infty}$ is dense in Ω, then

$$c_{\mathrm{ML}}(f,n) \sim \frac{\|f\|_K}{\sqrt{n}} \quad \text{as} \quad n \to \infty, \tag{18}$$

where $0/0 = 1$.

Proof. The minimum-norm interpolation property yields

$$\|I_{\{x_m\}} f\|_K \leq \|I_{X_n} f\|_K = \|I_n f\|_K \leq \|f\|_K$$

if $x_m \in X_n$, which is equivalent to $n \geq m$. The upper bound in (17) follows immediately while the lower bound uses

$$\|I_{\{x_m\}} f\|_K^2 = \frac{f(x_m)^2}{K(x_m,x_m)}.$$

The asymptotic equality (18) is a consequence of $\lim_{n\to\infty} \|I_n f\|_K = \|f\|_K$, which follows from Proposition 2.1. □

By inserting $c_{\mathrm{ML}}(f,N)$ for $c(f,n)$ in (14) we obtain the computable error estimate

$$|L(f) - Q_{L,n}(f)| \leq c_{\mathrm{ML}}(f,n)\, E_n(L) = \frac{\|I_n f\|_K}{\sqrt{n}} E_n(L), \tag{19}$$

where the right-hand side is asymptotically $n^{-1/2}\|f\|_K E_n(L)$ when Ω is a metric space, K is continuous and $\{x_i\}_{i=1}^{\infty}$ is dense in Ω. From the results in Section 2 it is clear that (19) fails for a large number of elements of $H(K)$. In fact, the error estimate fails for the dense set A^{c} in (13). Although somewhat disconcerting, this does not have to mean that (19) fails often for functions which are encountered in practice. The following approximation theoretic reasoning provides some non-rigorous justification for this claim. Because $I_n(I_{n-1}f) = I_{n-1}f$, we have from $\|I_n f\|_K^2 = \mathbf{f}_n^{\mathsf{T}} \mathbf{K}_{n,n}^{-1} \mathbf{f}$ that

$$\|I_n f - I_{n-1} f\|_K^2 = \|I_n(f - I_{n-1}f)\|_K^2 = \mathbf{a}^{\mathsf{T}} \mathbf{K}_{n,n}^{-1} \mathbf{a} = (\mathbf{K}_{n,n}^{-1})_{n,n} [f(x_n) - (I_{n-1}f)(x_n)]^2,$$

where $\mathbf{a} = (0,\dots,0,f(x_n) - (I_{n-1}f)(x_n)) \in \mathbb{R}^n$. From the expression for the power function in (2) and the block matrix inversion formula we get $(\mathbf{K}_{n,n}^{-1})_{n,n} = P_{n-1}(x_n)^{-2}$. Thus

$$\|I_n f - I_{n-1} f\|_K^2 = \|I_n(f - I_{n-1}f)\|_K^2 = \left(\frac{f(x_n) - (I_{n-1}f)(x_n)}{P_{n-1}(x_n)}\right)^2.$$

Repeatedly using $\|I_n f\|_K^2 = \|I_n f - I_{n-1} f\|_K^2 + \|I_{n-1} f\|_K^2$ and the above equation then yields the well known expression (e.g., [22, Theorem 6] and [10, Bemerkung 3.1.4])

$$\|I_n f\|_K^2 = \sum_{i=1}^{n} \|I_i f - I_{i-1} f\|_K^2 = \sum_{i=1}^{n} \left(\frac{f(x_i) - (I_{i-1}f)(x_i)}{P_{i-1}(x_i)}\right)^2. \tag{20}$$

Therefore

$$c_{\mathrm{ML}}(f,n)^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{f(x_i) - (I_{i-1}f)(x_i)}{P_{i-1}(x_i)}\right)^2. \tag{21}$$

Because

$$\|I_n f\|_K^2 = \sum_{i=1}^{n} \left(\frac{f(x_i) - (I_{i-1}f)(x_i)}{P_{i-1}(x_i)}\right)^2 \leq \|f\|_K^2$$

for every $n \in \mathbb{N}$ if $f \in H(K)$ by the minimum-norm interpolation property, the series

$$\sum_{n=1}^{\infty} a_n^2 = \sum_{n=1}^{\infty} \|I_n f - I_{n-1} f\|_K^2 = \sum_{n=1}^{\infty} \left(\frac{f(x_n) - (I_{n-1}f)(x_n)}{P_{n-1}(x_n)}\right)^2$$

converges. Let a_n be non-negative. Then $|f(x_n) - (I_{n-1}f)(x_n)| = a_n P_{n-1}(x_n)$ for a square-summable sequence $(a_n)_{n=1}^{\infty}$. Supposing that x_n can be replaced with any (or some) $x \in \Omega \backslash X_n$ in this equation—a proposition that we cannot substantiate rigorously—yields

$$|f(x) - (I_n f)(x)| = c_n P_n(x) \tag{22}$$

for a square-summable $(c_n)_{n=1}^{\infty}$, which suggests that at least for interpolation the coefficients $c(f,n)$ in the error bound (14) should form a sequence which either is or, to be on the safe side, "is almost" square-summable. A sequence such that $c(f,n) = \mathcal{O}(n^{-1/2})$, being a prototypical example of a "barely" non-square-summable sequence, is therefore a natural candidate. By Proposition 3.2, the maximum likelihood estimates $c_{\mathrm{ML}}(f,n)$ form such a sequence. Because it is not possible to define any sensible notion of a boundary between convergent and divergent series and the terms of a convergent series can decay arbitrarily slowly [9, § 41], the use of $c(f,n) = \mathcal{O}(n^{-1/2})$ would not make (14) valid for all $f \in H(K)$ even if (22) were true. But one could perhaps argue that square-summable sequences which are not $\mathcal{O}(n^{-1/2})$ are anomalous. In any case, a square-root rate at $x \in \{x_i\}_{i=1}^{\infty}$ can be proved if it is assumed that the sequence $(a_n)_{n=1}^{\infty}$ is decreasing.

Proposition 3.3. *If $f \in H(K)$ and the sequence defined by $a_n = \|I_n f - I_{n-1} f\|_K$ is decreasing, then*

$$\|I_n f - I_{n-1} f\|_K = \frac{|f(x_n) - (I_{n-1} f)(x_n)|}{P_{n-1}(x_n)} = o\left(\frac{1}{\sqrt{n}}\right).$$

Proof. Above we showed that $\sum_{n=1}^{\infty} \|I_n f - I_{n-1} f\|_K^2 \leq \|f\|_K^2 < \infty$ if $f \in H(K)$. Since the non-negative sequence $(a_n)_{n=1}^{\infty}$ is decreasing and square-summable, a general result on series states that $n a_n^2 \to 0$ as $n \to \infty$ [9, p. 124]. This proves the claim. □

By the above reasoning and Proposition 3.2, the maximum likelihood estimate (16) can therefore be interpreted as an approximation of $\|f\|_K$ modulated by a factor $n^{-1/2}$ which ensures that the resulting error bound is not too conservative for those elements of the RKHS which are "well-behaved". From Gaussian process perspective there is a very simple explanation for the presence of $n^{-1/2}$: this factor is needed to make the maximum likelihood estimator unbiased. For suppose that f is a zero-mean Gaussian process with covariance kernel $\sigma_0^2 K$ for some true scaling $\sigma_0 > 0$, so that $\mathrm{Cov}_f[f(x), f(y)] = \mathbb{E}_f[f(x)f(y)] = \sigma_0^2 K(x,y)$. Then

$$\mathbb{E}_f[\sigma_{\mathrm{ML}}^2] = \mathbb{E}_f[c_{\mathrm{ML}}(f,n)^2] = \mathbb{E}_f\left[\frac{\mathbf{f}_n^{\mathsf{T}} \mathbf{K}_{n,n}^{-1} \mathbf{f}_n}{n}\right] = \frac{\mathbb{E}_f[\mathrm{tr}(\mathbf{K}_{n,n}^{-1} \mathbf{f}_n \mathbf{f}_n^{\mathsf{T}})]}{n} = \frac{\mathrm{tr}(\mathbf{K}_{n,n}^{-1} \sigma_0^2 \mathbf{K}_{n,n})}{n} = \sigma_0^2 \frac{\mathrm{tr}(\mathrm{Id}_n)}{n} = \sigma_0^2, \tag{23}$$

which means that σ_{ML}^2 is an unbiased estimator of σ_0^2. Equation (23) implies that $c_{\mathrm{ML}}(f,n)^2$ is on average $\Theta(1)$ for sample paths of the Gaussian process. In other words, $\|I_n f\|_K^2 = \Theta(n)$ on average. This is related to the well known fact that the samples of a Gaussian process are not contained in the RKHS of its covariance kernel [4]. In the Matérn case the samples have, essentially, $d/2$ less smoothness than the RKHS, which leads one to expect that from the results in [11] it should follow that $\|I_n f\|_K^2 = \Theta(h_{n,\Omega}^{-d}) = \Theta(n)$ for quasi-uniform points on a sufficiently regular $\Omega \subset \mathbb{R}^d$ (see [8, Section 4] for details).

It is worth noting that the reasoning above appears to be very closely related to the $d/2$-gap observed by Schaback and Wendland [21] in connection to inverse theorems for kernel-based interpolation. Consider a Matérn kernel (6) of order $\nu > 0$ on a sufficiently regular and bounded $\Omega \subset \mathbb{R}^d$ and recall that

$$\sup_{x \in \Omega} P_n(x) = \mathcal{O}(h_{n,\Omega}^{\nu}) \quad \text{and} \quad \sup_{x \in \Omega} |f(x) - (I_n f)(x)| = \mathcal{O}(h_{n,\Omega}^{\nu}) \tag{24}$$

for every $f \in H(K) = H^{\nu + d/2}(\Omega)$. Schaback and Wendland [21, Theorem 6.1] have proved that if $f : \Omega \to \mathbb{R}$ is any function such that

$$\sup_{x \in \Omega} |f(x) - (I_n f)(x)| = \mathcal{O}(h_{n,\Omega}^{\nu + d/2 + \varepsilon}) \tag{25}$$

for every sequence of distinct points $\{x_i\}_{i=1}^{\infty}$ in Ω and some $\varepsilon > 0$, then $f \in H(K)$. As is evident, there is a gap of $d/2$ between the sufficient and necessary algebraic orders in (24) and (25). When the points are quasi-uniform, the gap is of order $n^{-1/2}$ and thus one can think of the factor $n^{-1/2}$ in the maximum likelihood estimate as a form of a compensation for the lack of this factor in (24). Note that the proof of Theorem 6.1 in [21] uses the expansion (20) that we used to justify the $n^{-1/2}$-factor in $c_{\mathrm{ML}}(f,n)$.

3.2 Cross-Validation

In (probabilistic) *cross-validation* the objective function that the kernel parameters are to minimise is

$$\ell_{\mathrm{CV}}(\theta) = \sum_{i=1}^{n} \frac{[f(x_i) - (I_{n,i} f)(x_i)]^2}{P_{n,i}(x_i)^2} + \sum_{i=1}^{n} \log P_{n,i}(x_i), \tag{26}$$

where the subscripts denote that the interpolant and the power function are computed using evaluations at $X_n \setminus \{x_i\}$; see Section 4.2 in [3] or Section 5.4.1 in [17]. Note that the objective function (26) differs from the one that is typically used in kernel-based approximation literature [e.g., 19], $\tilde{\ell}_{\mathrm{CV}}(\theta) = \sum_{i=1}^{n} (f(x_i) - I_{n,i}(x_i))^2$. Because the kernel interpolant does not depend on scaling of the kernel, $\tilde{\ell}_{\mathrm{CV}}$ cannot be used to select the parameter σ of $K_\sigma(x,y) = \sigma^2 K(x,y)$.

Similarly to maximum likelihood estimation, it is straightforward to compute that the unique minimiser of (26) under the scale parametrisation $\theta = \sigma$ is

$$c_{\mathrm{CV}}(f,n) = \sigma_{\mathrm{CV}} = \operatorname*{arg\,min}_{\sigma > 0} \ell_{\mathrm{CV}}(\sigma) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \frac{[f(x_i) - (I_{n,i} f)(x_i)]^2}{P_{n,i}(x_i)^2}}.$$

Note $c_{\mathrm{CV}}(f,n)$ differs from the form derived for $c_{\mathrm{ML}}(f,n)$ in (21) only in which points the kernel interpolant and power function use in each term of the sum. From the arguments used in the derivation of (21) and the observation that $I_{i-1} f = I_{i,i} f$ we obtain

$$c_{\mathrm{CV}}(f,n)^2 = \frac{1}{n} \sum_{i=1}^{n} \|I_n f - I_{n,i} f\|_K^2 = \frac{1}{n} \sum_{i=1}^{n} \|I_n (f - I_{n,i} f)\|_K^2. \tag{27}$$

The following proposition collects a few basic properties of $c_{\mathrm{CV}}(f,n)$.

Proposition 3.4. *The following statements hold:*

(i) $c_{\mathrm{CV}}(f,n) > 0$ *if and only if* $f(x_i) \neq 0$ *for some* $i \leq n$.

(ii) If $f \in H(K)$, *then* $c_{\mathrm{CV}}(f,n) \leq \|f\|_K$.

Proof. Let us first prove (i). It is clear that $c_{\mathrm{CV}}(f,n)=0$ if $f|_{X_n}=0$ because in this case $f(x_i)=(I_{n,i}f)(x_i)=0$ for every $i\leq n$. Suppose that $f(x_i)\neq 0$ for some $i\leq n$. The positivity of $c_{\mathrm{CV}}(f,n)$ is equivalent to $f(x_j)\neq(I_{n,j}f)(x_j)$ for at least one $j\leq n$. Assume to the contrary that $f(x_j)=(I_{n,j}f)(x_j)$ for every $j\leq n$, which means that each $I_{n,j}f$ interpolates f at X_n. Therefore $I_nf=I_{n,j}f$ and consequently $I_nf\in\operatorname{span}\{K(\cdot,x_i)\}_{i=1,i\neq j}^n$ for every $j\leq n$. But because the kernel translates are linearly independent, this implies that $I_nf\equiv 0$, which contradicts the assumption that $f(x_i)\neq 0$ for some $i\leq n$. Hence $c_{\mathrm{CV}}(f,n)>0$.

Let us then prove (ii). From (27) and the minimum-norm interpolation property it follows that

$$c_{\mathrm{CV}}(f,n)^2=\frac{1}{n}\sum_{i=1}^n\|I_n(f-I_{n,i}f)\|_K^2\leq\frac{1}{n}\sum_{i=1}^n\|f-I_{n,i}f\|_K^2\leq\frac{1}{n}\sum_{i=1}^n\|f\|_K^2=\|f\|_K^2,$$

which proves the claim. □

The upper bound in (ii) of Proposition 3.4 is clearly extremely conservative. Indeed, precisely the same chain of inequalities could have been used to show that $c_{\mathrm{ML}}(f,n)\leq\|f\|_K$, which is conservative by a factor of $n^{1/2}$. Although we are unable to prove this, we believe that in most cases it should be expected that

$$c_{\mathrm{CV}}(f,n)\leq c_{\mathrm{ML}}(f,n)=\frac{\|I_nf\|_K}{\sqrt{n}}. \tag{28}$$

For example, if Ω is a subset of $\mathbb{R}^d$ which equals the closure of its interior and $\{x_i\}_{i=1}^\infty$ is dense in Ω, then

$$\lim_{n\to\infty}\|I_nf\|_K=\lim_{n\to\infty}\|I_{n,i}f\|_K=\|f\|_K$$

for every $i\in\mathbb{N}$ if $f\in H(K)$ by Proposition 2.1. Therefore

$$\lim_{n\to\infty}\|I_n(f-I_{n,i}f)\|_K^2=0\quad\text{for every}\quad i\in\mathbb{N}, \tag{29}$$

which in combination with (21) and (27) suggests that (28) ought to hold when n is sufficiently large because each term in the expansion for $c_{\mathrm{CV}}(f,n)^2$ tends to zero as n increases while the terms in the expansion of $c_{\mathrm{ML}}(f,n)^2$ are fixed. However, to make this argument rigorous the convergence (29) would need to be assumed or proved to be uniform over $i\in\mathbb{N}$. What we can prove is limited to the following much weaker result, which is a generalisation of the claim (ii) in Proposition 3.4,

Proposition 3.5. *Suppose that $f\in H(K)$ and let $(\delta_n)_{n=1}^\infty$ be any non-negative sequence such that $\max_{i=1,\dots,n}\|f-I_{n,i}f\|_K\leq\delta_n$ for every $n\in\mathbb{N}$. Then*

$$c_{\mathrm{CV}}(f,n)\leq\delta_n\quad\text{for every } n\in\mathbb{N}. \tag{30}$$

Proof. By the minimum-norm interpolation property, $\|I_n(f-I_{n,i}f)\|_K\leq\|f-I_{n,i}f\|_K\leq\delta_n$ for every $i\leq n$. From (27) we thus obtain

$$c_{\mathrm{CV}}(f,n)^2=\frac{1}{n}\sum_{i=1}^n\|I_n(f-I_{n,i}f)\|_K^2\leq\frac{1}{n}\sum_{i=1}^n\delta_n^2=\delta_n^2.$$

□

In certain cases it is possible to obtain the sequence $(\delta_n)_{n=1}^\infty$ in Proposition 3.5 explicitly. Assume that Ω is a compact subset of $\mathbb{R}^d$ and the kernel K is continuous. Then the integral operator $T\colon L^2(\Omega)\to L^2(\Omega)$ defined via

$$(Tf)(y)=\int_\Omega f(y)K(x,y)\,\mathrm{d}y$$

is compact and self-adjoint. Moreover, the range $T(L^2(\Omega))$ of T is contained in $H(K)$. One can then show that [26, Section 11.5]

$$\|f-I_nf\|_K\leq\|T^{-1}f\|_{L^2(\Omega)}\|P_n\|_{L^2(\Omega)}\quad\text{if}\quad f\in T(L^2(\Omega)).$$

In particular, if K is a Matérn kernel of order ν, Ω is sufficiently regular and the points are quasi-uniform,

$$\|f-I_{n,i}f\|_K=\mathcal{O}(\|f-I_nf\|_K)=\mathcal{O}(\|P_n\|_{L^2(\Omega)})=\mathcal{O}(n^{-\nu/d})$$

by (9). That is, in this case Proposition 3.5 gives $c_{\mathrm{CV}}(f,n)=\mathcal{O}(n^{-\nu/d})$ if $f\in T(L^2(\Omega))$. Although the bound (30) is likely to be somewhat conservative, this nevertheless demonstrates that for certain elements of the RKHS cross-validation may yield less conservative error bounds than maximum likelihood estimation.

As argued in Section 3.1, the maximum likelihood estimate $c_{\mathrm{ML}}(f,n)$ equals an approximation $\|I_nf\|_K$ of $\|f\|_K$ modulated by a factor of $n^{-1/2}$. Given (I-EB) and $\|f-I_nf\|_K\leq\|f\|_K$, one should obtain a better error bound by approximating $\|f-I_nf\|_K$ instead of $\|f\|_K$. However, $\|f-I_nf\|_K$ cannot be approximated directly as

$$\|I_n(f-I_nf)\|_K^2\approx\|f-I_nf\|_K^2$$

because the left-hand side is always zero due to $f-I_nf$ vanishing on X_n. But for each $i\leq n$, we can use the approximation

$$\|I_n(f-I_{n,i}f)\|_K^2\approx\|f-I_{n,i}f\|_K^2\approx\|f-I_nf\|_K^2.$$

Because not every $\|I_n(f-I_{n,i}f)\|_K$ can be zero by by Proposition 3.4 unless f vanishes on X_n, the average of $\|I_n(f-I_{n,i}f)\|_K^2$ would seem to make for a good approximation of $\|f-I_nf\|_K^2$. As we have seen in (27), this average is precisely $c_{\mathrm{CV}}(f,n)^2$:

$$c_{\mathrm{CV}}(f,n)^2=\frac{1}{n}\sum_{i=1}^n\|I_n(f-I_{n,i}f)\|_K^2.$$

3.3 Summary

Let us briefly summarise the findings of this section. The coefficients that we suggest using in the computable error bound (14) are

$$c_{\mathrm{ML}}(f,n)=\sqrt{\frac{\mathbf{f}_n^{\mathsf{T}}\mathbf{K}_{n,n}^{-1}\mathbf{f}_n}{n}}=\frac{\|I_nf\|_K}{\sqrt{n}}=\sqrt{\frac{1}{n}\sum_{i=1}^n\left(\frac{f(x_i)-(I_{i-1}f)(x_i)}{P_{i-1}(x_i)}\right)^2}=\sqrt{\frac{1}{n}\sum_{i=1}^n\|I_if-I_{i-1}f\|_K^2}=\sqrt{\frac{1}{n}\sum_{i=1}^n\|I_i(f-I_{i-1}f)\|_K^2}$$

and

$$c_{\mathrm{CV}}(f,n)=\sqrt{\frac{1}{n}\sum_{i=1}^n\frac{[f(x_i)-(I_{n,i}f)(x_i)]^2}{P_{n,i}(x_i)^2}}=\sqrt{\frac{1}{n}\sum_{i=1}^n\|I_nf-I_{n,i}f\|_K^2}=\sqrt{\frac{1}{n}\sum_{i=1}^n\|I_n(f-I_{n,i}f)\|_K^2}.$$

The former, $c_{\mathrm{ML}}(f,n)$, can be considered a modulated approximation to $\|f\|_K$ while the latter, $c_{\mathrm{CV}}(f,n)$, approximates $\|f-I_nf\|_K$. It is therefore to be expected that using $c_{\mathrm{CV}}(f,n)$ results in tighter error bounds.

4 Numerical Example

This section contains a simple numerical study of the computable error bounds proposed in Section 3. Related examples for the maximum likelihood estimate can be found in [8, Section 5.2]. Stopping criteria derived from maximum likelihood estimation and cross-validation have been used in numerical integration at lattice points by Rathinavel and Hickernell [18]. We consider the integration functional

$$L(f)=\int_0^1 f(x)\,\mathrm{d}x$$

and the Matérn kernel (6) with $\nu=3/2$ and $\lambda=1$. For this kernel and integration functional we have

$$K_L(x)=\int_0^1 K(x,y)\,\mathrm{d}y=\frac{4}{\sqrt{3}}-\frac{1}{3}\exp\big(\sqrt{3}(x-1)\big)\big(3+2\sqrt{3}-3x\big)-\frac{1}{3}\exp(-\sqrt{3}x)\big(3x+2\sqrt{3}\big)$$

and

$$K_{L,L}=\int_0^1 K_L(x)\,\mathrm{d}x=\frac{2}{3}\Big[2\sqrt{3}-3+\exp(-\sqrt{3})\big(\sqrt{3}+3\big)\Big],$$

so that the worst-case error is computable in closed from. We consider the following six test functions of varying smoothness:

$$\begin{aligned}
&f_1(x)=K_{\nu=1,\lambda=0.5}(x-0.6)+K_{\nu=1,\lambda=0.5}(x-0.2), && f_2(x)=K_{\nu=1.6,\lambda=0.5}(x-0.6),\\
&f_3(x)=K_{\nu=3.1,\lambda=0.9}(x-0.6), && f_4(x)=\exp(-(x-0.5)^2),\\
&f_5(x)=1+0.5x^2, && f_6(x)=K_{\nu=1,\lambda=0.5}(x-0.6)+K_{\nu=2,\lambda=1.2}(x-0.2).
\end{aligned}\tag{31}$$

The subscripts ν and λ for K denote that the smoothness and scale parameters of a Matérn kernel (6). We use four different sequences of point sets:

1. **uniform — endpoint included**: Uniform points on $[0,1]$ with both 0 and 1 included:
$$X_n=\left\{0,\frac{1}{n-1},\cdots,1-\frac{1}{n-1},1\right\}.$$
These point sets are not nested.
2. **uniform — endpoint not included**: Uniform points on $[0,1]$ with only 0 included:
$$X_n=\left\{0,\frac{1}{n},\cdots,1-\frac{1}{n}\right\}.$$
These point sets are not nested.
3. **van der Corput — endpoint included**: The first n elements of the van der Corput sequence $\{1,0,0.5,0.25,0.75,\ldots\}$ with 0 and 1 included. These point sets are nested.
4. **van der Corput — endpoint not included**: The first n elements of the van der Corput sequence $\{0,0.5,0.25,0.75,\ldots\}$ with only 0 included. These point sets are nested.

Figures 1 and 2 show the behaviour of the ratios

$$r_{\mathrm{ML}}(n)=\frac{|L(f)-Q_{L,n}(f)|}{c_{\mathrm{ML}}(f,n)E_n(L)}\quad\text{and}\quad r_{\mathrm{CV}}(n)=\frac{|L(f)-Q_{L,n}(f)|}{c_{\mathrm{CV}}(f,n)E_n(L)}\tag{32}$$

for $f=f_1,\ldots,f_6$ and $n=1,\ldots,200$. It is desirable that the ratios be as close to one (marked by the blue line) as possible. If the ratios exceed one, the error bounds are optimistic; if not, the bounds are conservative. The following observations can be made from the figures:

- Except for very small n, maximum likelihood estimation always yields error bounds that are not optimistic. For uniform points the bounds appear to become increasingly conservative as n increases.
- When the endpoint is included, cross-validation appears to yield error bounds that are off only by a constant factor.
- However, when the endpoint is not included, cross-validation is optimistic for the test functions f_2, f_3, f_4 and f_5. We are unable to explain this phenomenon, which may be related to these four functions being the smoothest of our six test functions.

Note the oscillation of the ratios for the least smooth test functions, f_1, f_2 and f_6, when uniform points are used. This is likely caused by the non-nestedness of the uniform point sets due to which the point configurations around the points $x = 0.2$ and $x = 0.6$, at which these functions are not infinitely differentiable, keep changing from one n to another.

Acknowledgements

The author was supported by the Academy of Finland postdoctoral researcher grant #338567 “Scalable, adaptive and reliable probabilistic integration”. Comments by the reviewers helped to simplify the proof of Proposition 3.1 and made Section 4 much more interesting than it originally was.

References

[1] Arcangéli, R., de Silanes, M. C. L., and Torrens, J. J. (2007). An extension of a bound for functions in Sobolev spaces, with applications to (m,s)-spline interpolation and smoothing. *Numerische Mathematik*, 107(2):181–211.

[2] Cavoretto, R. (2021). Adaptive radial basis function partition of unity interpolation: A bivariate algorithm for unstructured data. *Journal of Scientific Computing*, 87(41).

[3] Currin, C., Mitchell, T., Morris, M., and Ylvisaker, D. (1988). A Bayesian approach to the design and analysis of computer experiments. ORNL-6498, Oak Ridge National Laboratory.

[4] Driscoll, M. F. (1973). The reproducing kernel Hilbert space structure of the sample paths of a Gaussian process. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 26(4):309–316.

[5] Fasshauer, G. and McCourt, M. (2015). *Kernel-based Approximation Methods Using MATLAB*. Number 19 in Interdisciplinary Mathematical Sciences. World Scientific Publishing.

[6] Fasshauer, G. E. (2011). Positive definite kernels: Past, present and future. *Dolomites Research Notes on Approximation*, 4:21–63.

[7] Iske, A. (2018). *Approximation Theory and Algorithms for Data Analysis*. Springer.

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

[9] Knopp, K. (1951). *Theory and Application of Infinite Series*. Blackie & Son, 2nd edition.

[10] Müller, S. (2008). *Komplexität und Stabilität von kernbasierten Rekonstruktionsmethoden*. PhD thesis, University of Göttingen.

[11] Narcowich, F. J., Ward, J. D., and Wendland, H. (2006). Sobolev error estimates and a Bernstein inequality for scattered data interpolation via radial basis functions. *Constructive Approximation*, 24(2):175–186.

[12] Novak, E. and Woźniakowski, H. (2010). *Tractability of Multivariate Problems. Volume II: Standard Information for Functionals*, volume 12 of *EMS Tracts in Mathematics*. European Mathematical Society.

[13] Oettershagen, J. (2017). *Construction of Optimal Cubature Algorithms with Applications to Econometrics and Uncertainty Quantification*. PhD thesis, Faculty of Mathematics and Natural Sciences, University of Bonn.

[14] Owen, A. B. and Pan, Z. (2022). Where are the logs? *arXiv:2110.06420v2*.

[15] Paulsen, V. I. and Raghupathi, M. (2016). *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*. Number 152 in Cambridge Studies in Advanced Mathematics. Cambridge University Press.

[16] Plaskota, L. (1996). *Noisy Information and Computational Complexity*. Cambridge University Press.

[17] Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. MIT Press.

[18] Rathinavel, J. and Hickernell, F. J. (2019). Fast automatic Bayesian cubature using lattice sampling. *Statistics and Computing*, 29(6):1215–1229.

[19] Rippa, S. (1999). An algorithm for selecting a good value for the parameter c in radial basis function interpolation. *Advances in Computational Mathematics*, 11(2):193–210.

[20] Schaback, R. (2018). Superconvergence of kernel-based interpolation. *Journal of Approximation Theory*, 235:1–19.

[21] Schaback, R. and Wendland, H. (2002). Inverse and saturation theorems for radial basis function interpolation. *Mathematics of Computation*, 71(238):669–681.

[22] Schaback, R. and Werner, J. (2006). Linearly constrained reconstruction of functions by kernels with applications to machine learning. *Advances in Computational Mathematics*, 25:237.

[23] Scheuerer, M., Schaback, R., and Schlather, M. (2013). Interpolation of spatial data – A stochastic or a deterministic problem? *European Journal of Applied Mathematics*, 24(4):601–629.

[24] Traub, J. F., Wasilkowski, G. W., and Woźniakowski, H. (1988). *Information-Based Complexity*. Computer Science and Scientific Computing. Academic Press.

[25] Wang, W. (2021). On the inference of applying Gaussian process modeling to a deterministic function. *Electronic Journal of Statistics*, 15(2):5014–5066.

[26] Wendland, H. (2005). *Scattered Data Approximation*. Number 17 in Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press.

[27] Wenzel, T., Santin, G., and Haasdonk, B. (2021). Analysis of target data-dependent greedy kernel algorithms: Convergence rates for f-, $f \cdot P$- and f/P-greedy. *arXiv:2105.07411v1*.

[28] Wynne, G., Briol, F.-X., and Girolami, M. (2021). Convergence guarantees for Gaussian process means with misspecified likelihoods and smoothness. *Journal of Machine Learning Research*, 22(123):1–40.

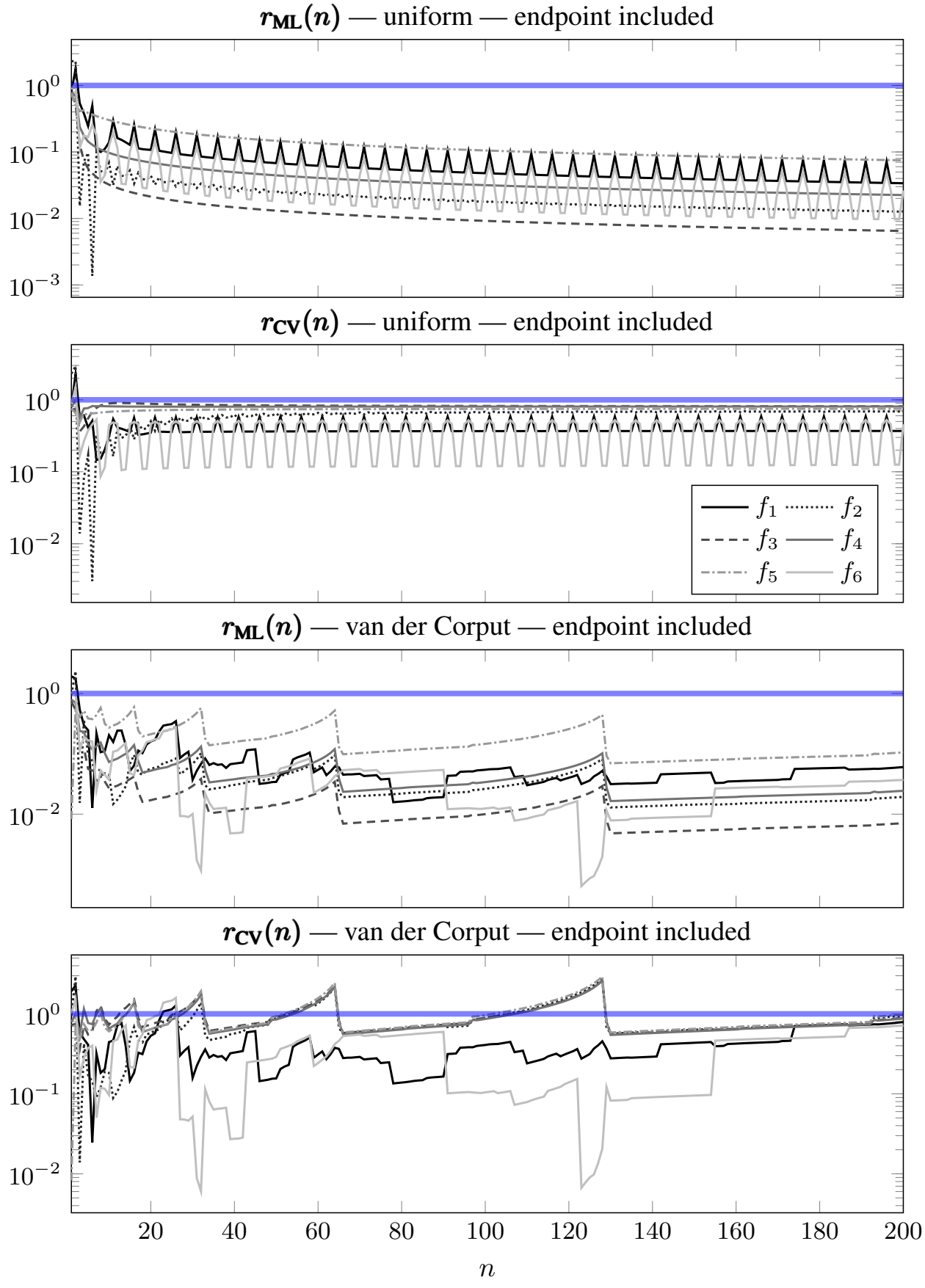

Figure 1: The ratios $r_{\mathrm{ML}}(n)$ and $r_{\mathrm{CV}}(n)$ in (32) for the functions $f = f_1, \dots, f_6$ in (31) and $n = 1, \dots, 200$ when the point sets X_n include the endpoint.

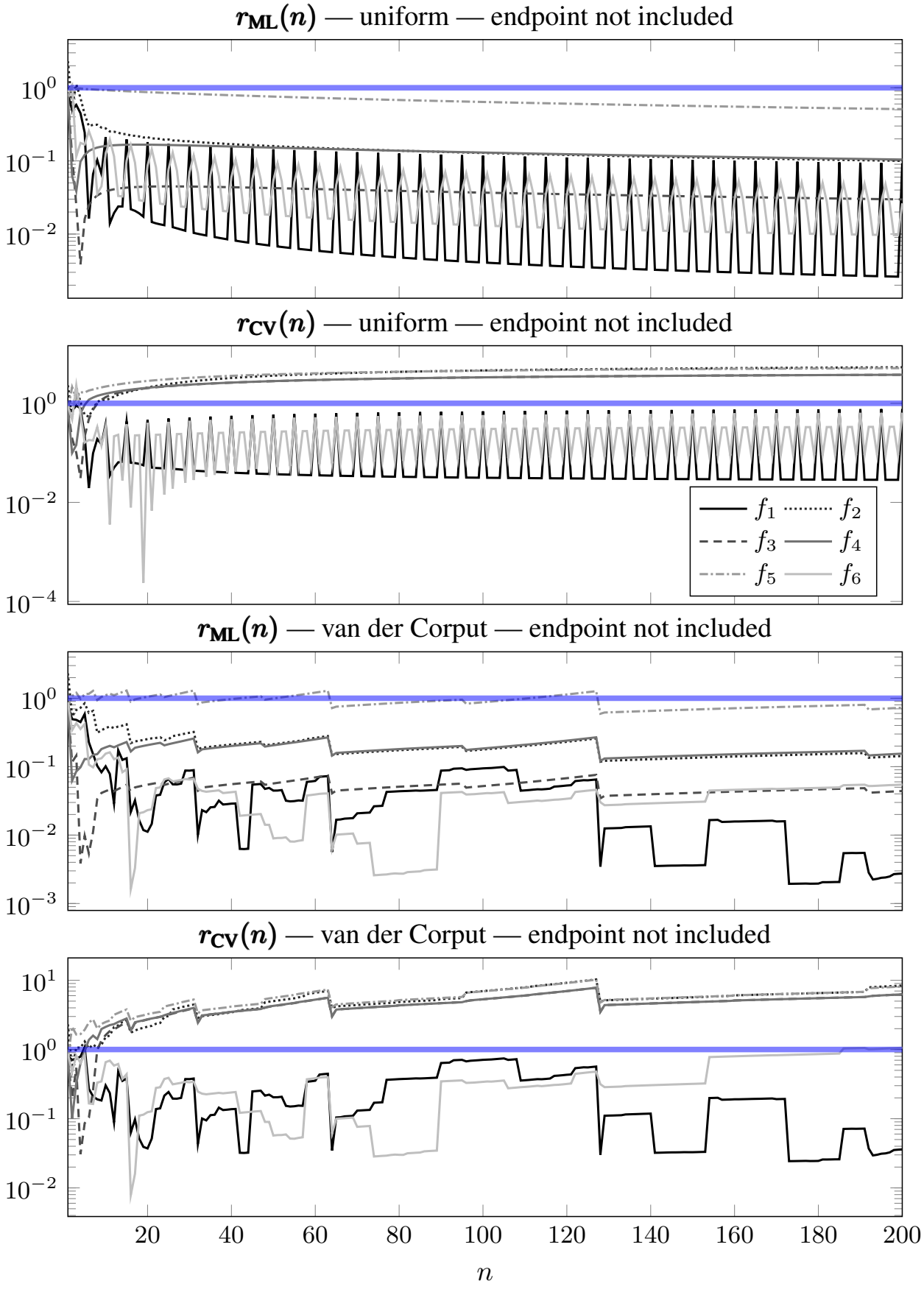

Figure 2: The ratios $r_{\mathrm{ML}}(n)$ and $r_{\mathrm{CV}}(n)$ in (32) for the functions $f = f_1, \dots, f_6$ in (31) and $n = 1, \dots, 200$ when the point sets X_n do not include the endpoint.