Foundations of Computational Mathematics (2020) 20:155–194
https://doi.org/10.1007/s10208-018-09407-7

FOUNDATIONS OF COMPUTATIONAL MATHEMATICS
The Journal of the Society for the Foundations of Computational Mathematics

Convergence Analysis of Deterministic Kernel-Based Quadrature Rules in Misspecified Settings

Motonobu Kanagawa[1] · **Bharath K. Sriperumbudur**[2] · **Kenji Fukumizu**[3]

Published online: 7 January 2019

Abstract

This paper presents convergence analysis of kernel-based quadrature rules in misspecified settings, focusing on deterministic quadrature in Sobolev spaces. In particular, we deal with misspecified settings where a test integrand is less smooth than a Sobolev RKHS based on which a quadrature rule is constructed. We provide convergence guarantees based on two different assumptions on a quadrature rule: one on quadrature weights and the other on design points. More precisely, we show that convergence rates can be derived (i) if the sum of absolute weights remains constant (or does not increase quickly), or (ii) if the minimum distance between design points does not decrease very quickly. As a consequence of the latter result, we derive a rate of convergence for Bayesian quadrature in misspecified settings. We reveal a condition on design points to make Bayesian quadrature robust to misspecification, and show that, under this condition, it may adaptively achieve the optimal rate of convergence in the Sobolev space of a lesser order (i.e., of the unknown smoothness of a test integrand), under a slightly stronger regularity condition on the integrand.

Keywords Kernel-based quadrature rules · Misspecified settings · Sobolev spaces · Reproducing kernel Hilbert spaces · Bayesian quadrature

Mathematics Subject Classification Primary 65D30 · Secondary 65D32 · 65D05 · 46E35 · 46E22

Communicated by Francis Bach.

MK and KF acknowledge support by MEXT Grant-in-Aid for Scientific Research on Innovative Areas (25120012). MK has also been supported in part by MEXT KAKENHI (17K12654) and the European Research Council (StG Project PANAMA). BKS is partly supported by NSF-DMS-1713011.

Extended author information available on the last page of the article

1 Introduction

This paper discusses the problem of numerical integration (or quadrature), which has been a fundamental task in numerical analysis, statistics, computer science including machine learning and other areas. Let P be a (known) Borel probability measure on the Euclidian space $\mathbb{R}^d$ with support contained in an open set $\Omega \subset \mathbb{R}^d$ and f be an integrand on Ω. Suppose that the integral $\int f(x)\mathrm{d}P(x)$ has no closed-form solution. We consider quadrature rules that provide an approximation of the integral, in the form of a weighted sum of function values

$$\sum_{i=1}^{n} w_i f(X_i) \approx \int f(x)\mathrm{d}P(x), \tag{1}$$

where $X_1, \ldots, X_n \in \Omega$ are design points and $w_1, \ldots, w_n \in \mathbb{R}$ are quadrature weights. Throughout this paper, the integral of f and its quadrature estimate are denoted by Pf and $P_n f$, respectively; namely,

$$Pf := \int f(x)\mathrm{d}P(x), \quad P_n f := \sum_{i=1}^{n} w_i f(X_i). \tag{2}$$

Examples of such quadrature rules include Monte Carlo methods, which make use of a random sample from a suitable proposal distribution as $X_1, \ldots, X_n$ and importance weights as $w_1, \ldots, w_n$. A limitation of standard Monte Carlo methods is that a huge number of design points (i.e., large n) may be needed for providing an accurate approximation of the integral; this comes from the fact that the rate of convergence of Monte Carlo methods is typically of the order $\mathbb{E}[|Pf - P_n f|] = O(n^{-1/2})$ as $n \to \infty$, where $\mathbb{E}[\cdot]$ denotes the expectation with respect to the random sample. The need for large n is problematic, when an evaluation of the function value $f(x)$ is expensive for each input x. Such situations appear in modern scientific and engineering problems where the mapping $x \mapsto f(x)$ involves complicated computer simulation. In applications to time-series forecasting, for instance, x may be a parameter of an underlying system, $f(x)$ a certain quantity of interest in future and P a prior distribution on x. Then, the target integral $\int f(x)\mathrm{d}P(x)$ is the predictive value of the future quantity. The evaluation of $f(x)$ for each x may require numerically solving an initial value problem for the differential equation, which results in time-consuming computation [7]. Similar examples can be seen in applications to statistics and machine learning, as mentioned below. In these situations, one can only use a limited number of design points, and thus, it is desirable to have quadrature rules with a faster convergence rate, in order to obtain a reliable solution [46].

1.1 Kernel-Based Quadrature Rules

How can we obtain a quadrature rule whose convergence rate is faster than $O(n^{-1/2})$? In practice, one often has prior knowledge or belief on the integrand f, such as

smoothness, periodicity and sparsity. Exploiting such knowledge or assumption in constructing a quadrature rule $\{(w_i, X_i)\}_{i=1}^n$ may achieve faster rates of convergence, and such methods have been extensively studied in the literature for decades; see, e.g., [17] and [9] for review.

This paper deals with quadrature rules using reproducing kernel Hilbert spaces (RKHS) explicitly or implicitly to achieve fast convergence rates; we will refer to such methods as *kernel-based quadrature rules* or simply *kernel quadrature*. As discussed in Sect. 2.4, notable examples include quasi-Monte Carlo methods [17,18,26,42], Bayesian quadrature [9,48] and kernel herding [5,10,11]. These methods have been studied extensively in recent years [4,8,30,45,46,55,62] and have recently found applications in, for instance, machine learning and statistics [3,9,21,31,32,43,50].

In kernel quadrature, we make use of available knowledge on properties of the integrand f by assuming that f belongs to a certain RKHS $\mathcal{H}_k$ that possesses those properties (where k is the reproducing kernel) and then constructing weighted points $\{(w_i, X_i)\}_{i=1}^n$ such that the *worst-case error* in the RKHS

$$e_n(P; \mathcal{H}_k) := \sup_{f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} \leq 1} |Pf - P_n f| \tag{3}$$

is made small, where $\|\cdot\|_{\mathcal{H}_k}$ is the norm of $\mathcal{H}_k$. The use of RKHS is beneficial when compared to other function spaces, as it leads to a closed-form expression of worst-case error (3) in terms of the kernel, and thus, one may explicitly use this expression for designing $\{(w_i, X_i)\}_{i=1}^n$ (see Sect. 2.3).

Note that, in a *well-specified case*, that is, the integrand f satisfies $f \in \mathcal{H}_k$, the quadrature error is bounded as

$$|P_n f - Pf| \leq \|f\|_{\mathcal{H}_k} e_n(P; \mathcal{H}_k).$$

This guarantees that if a quadrature rule satisfies $e_n(P; \mathcal{H}_k) = O(n^{-b})$ as $n \to \infty$ for some $b > 0$, then the quadrature error also satisfies $|P_n f - Pf| = O(n^{-b})$. Take a Sobolev space $H^r(\Omega)$ of order $r > d/2$ on Ω as the RKHS $\mathcal{H}_k$, for example. It is known that optimal quadrature rules achieve $e_n(P; \mathcal{H}_k) = O(n^{-r/d})$ [40], and thus, $|P_n f - Pf| = O(n^{-r/d})$ holds for any $f \in \mathcal{H}_k$. As we have $r/d > 1/2$, this rate is faster than Monte Carlo integration; this is the desideratum that has been discussed.

1.2 Misspecified Settings

This paper focuses on situations where the assumption $f \in \mathcal{H}_k$ is violated, that is, *misspecified settings*. As explained above, convergence guarantees for kernel quadrature rules often assume that $f \in \mathcal{H}_k$. However, in practice one may lack the full knowledge on the properties on the integrand, and therefore, misspecification of the RKHS (via the choice of its reproducing kernel k) may occur, that is, $f \notin \mathcal{H}_k$.

Such misspecification is likely to happen when the integrand is a *black box function*. An illustrative example can be found in applications to computer graphics such as the problem of illumination integration (see, e.g., [9]), where the task is to compute the total amount of light arriving at a camera in a virtual environment. This problem

is solved by quadrature, with integrand $f(x)$ being the intensity of light arriving at the camera from a direction x (angle). However, the value of $f(x)$ is only given by simulation of the environment for each x, so the integrand f is a black box function. Similar situations can be found in application to statistics and machine learning. A representative example is the computation of marginal likelihood for a probabilistic model, which is an important but challenging task required for model selection (see, e.g., [47]). In modern scientific applications where complex phenomena are dealt with (e.g., climate science), we often encounter situations where the evaluation of a likelihood function, which forms the integrand in marginal likelihood computation, involves an expensive simulation model, making the integrand complex and even black box.

If the integrand is a black box function, there is a trade-off between the risk of misspecification and gain in the rate of convergence for kernel-based quadrature rules; for a faster convergence rate, one may want to use a quadrature rule for a narrower $\mathcal{H}_k$ such as of higher-order differentiability, while such a choice may cause misspecification of the function class. Therefore, it is of great importance to elucidate their convergence properties in misspecified situations, in order to make use of such quadrature rules in a safe manner.

1.3 Contributions

This paper provides convergence rates of kernel-based quadrature rules in misspecified settings, focusing on *deterministic* rules (i.e., without randomization). The focus of misspecification is placed on the order of Sobolev spaces: The unknown order s of the integrand f is overestimated as r, that is, $s \leq r$.

Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with a Lipschitz boundary (see Sect. 3 for definition). For $r > d/2$, consider a positive definite kernel k_r on Ω that satisfies the following assumption;

Assumption 1 The kernel k_r on Ω satisfies $k_r(x, y) := \Phi(x - y)$, where $\Phi : \mathbb{R}^d \to \mathbb{R}$ is a positive definite function such that

$$C_1(1 + \|\xi\|^2)^{-r} \leq \hat{\Phi}(\xi) \leq C_2(1 + \|\xi\|^2)^{-r}$$

for some constants $C_1, C_2 > 0$, where $\hat{\Phi}$ is the Fourier transform of Φ. The RKHS $\mathcal{H}_{k_r}(\Omega)$ is the restriction of $\mathcal{H}_{k_r}(\mathbb{R}^d)$ to Ω (see Sect. 2).

The resulting RKHS $\mathcal{H}_{k_r}(\Omega)$ is norm-equivalent to the standard Sobolev space $H^r(\Omega)$. The Matérn and Wendland kernels satisfy Assumption 1 (see Sect. 2).

Consider a quadrature rule $\{(w_i, X_i)\}_{i=1}^n$ with the kernel k_r such that

$$e_n(P; \mathcal{H}_{k_r}(\Omega)) = O(n^{-b}) \quad (n \to \infty). \tag{4}$$

We do not specify how the weighted points are generated, but assume (4) aiming for wide applicability. Suppose that an integrand $f : \Omega \to \mathbb{R}$ has partial derivatives up to order s and they are bounded and uniformly continuous. If $s \leq r$, the integrand may not belong to the assumed RKHS $\mathcal{H}_{k_r}$, in which case a misspecification occurs.

Under this misspecified setting, two types of assumptions on the quadrature rule $\{(w_i, X_i)\}_{i=1}^n$ will be considered: one on the quadrature weights $w_1, \ldots, w_n$ (Sect. 4.1) and the other on the design points $X_1, \ldots, X_n$ (Sect. 4.2). In both cases, a rate of convergence of the form

$$|P_n f - Pf| = O(n^{-bs/r}) \quad (n \to \infty) \tag{5}$$

will be derived under some additional conditions. The results guarantee the convergence in the misspecified setting, and the rate is determined by the ratio s/r between the true smoothness s and the assumed smoothness r. As discussed in Sect. 2, the optimal rate of deterministic quadrature rules for the Sobolev space $H^r(\Omega)$ is $O(n^{-r/d})$ [40]. If a quadrature rule satisfies this optimal rate (i.e., $b = r/d$), then rate (5) becomes $O(n^{-s/d})$ for an integrand $f \in H^s(\Omega)$ $(s < r)$, which matches the optimal rate for $H^s(\Omega)$.

The specific results are summarized as follows:

- In Sect. 4.1, it is assumed that $\sum_{i=1}^n |w_i| = O(n^c)$ as $n \to \infty$ for some constant $c \geq 0$. Note that $c = 0$ is taken if the weights satisfy $\max_{i=1,\ldots,n} |w_i| = O(n^{-1})$, an example of which is the equal weights $w_1 = \cdots = w_n = 1/n$. Under this assumption and other suitable conditions, Corollary 2 shows

  $$|P_n f - Pf| = O(n^{-bs/r + c(r-s)/r}) \quad (n \to \infty).$$

  The rate $O(n^{-bs/r})$ in (5) holds if $c = 0$. Therefore, this result provides convergence guarantees in particular for equal-weight quadrature rules, such as quasi-Monte Carlo methods and kernel herding, in the misspecified setting.
- Section 4.2 uses an assumption on design points $X^n := \{X_1, \ldots, X_n\}$ in terms of *separation radius* q_{X_n}, which is defined by

  $$q_{X_n} := \frac{1}{2} \min_{i \neq j} \|X_i - X_j\|. \tag{6}$$

  Corollary 3 shows that, if $q_{X^n} = \Theta(n^{-a})$ as $n \to \infty$ for some $a > 0$, under other regularity conditions,

  $$|P_n f - Pf| = O(n^{-\min(b - a(r-s), as)}) \quad (n \to \infty). \tag{7}$$

  The best possible rate is $O(n^{-bs/r})$ when $a = b/r$. This result provides a convergence guarantee for quadrature rules that obtain the weights $w_1, \ldots, w_n$ to give $O(n^{-b})$ for the worst-case error with $X_1, \ldots, X_n$ fixed beforehand. We demonstrate this result by applying it to Bayesian quadrature, as explained below. Our result may also provide the following guideline for practitioners: in order to make a kernel quadrature rule robust to misspecification, one should specify the design points so that the spacing is not too small.
- Section 5 discusses a convergence rate for Bayesian quadrature under the misspecified setting, demonstrating the results of Sect. 4.2. Given design points

$X^n = \{X_1, \dots, X_n\}$, Bayesian quadrature defines weights $w_1, \dots, w_n$ as the minimizer of worst-case error (3), which can be obtained by solving a linear equation (see Sect. 2.4 for more detail). For points $X^n = \{X_1, \dots, X_n\}$ in Ω, the *fill distance* $h_{X^n,\Omega}$ is defined by

$$h_{X^n,\Omega} := \sup_{x\in\Omega} \min_{i=1,\dots,n} \|x - X_i\|. \tag{8}$$

Assume that there exists a constant $c_q > 0$ independent of X^n such that

$$h_{X^n,\Omega} \leq c_q q_{X^n}, \tag{9}$$

and that $h_{X^n,\Omega} = O(n^{-1/d})$ as $n \to \infty$. Then, Corollary 4 shows that with Bayesian quadrature weights based on the kernel k_r we have

$$|P_n f - Pf| = O(n^{-s/d}) \quad (n \to \infty).$$

Note that the rate $O(n^{-s/d})$ matches the minimax optimal rate for deterministic quadrature rules in the Sobolev space of order s [40], which implies that Bayesian quadrature can be *adaptive* to the unknown smoothness s of the integrand f. The adaptivity means that it can achieve the rate $O(n^{-s/d})$ without the knowledge of s; it only requires the knowledge of the upper bound of the true smoothness $s \leq r$.

– Section 3 establishes a rate of convergence for Bayesian quadrature in the *well-specified* case, which serves as a basis for the results in the misspecified case (Sect. 5). Corollary 1 asserts that if the design points satisfy $h_{X^n,\Omega} = O(n^{-1/d})$ as $n \to \infty$, then

$$e_n(P; \mathcal{H}_{k_r}(\Omega)) = O(n^{-r/d}) \quad (n \to \infty).$$

This rate $O(n^{-r/d})$ is minimax optimal for deterministic quadrature rules in Sobolev spaces. To the best of our knowledge, this optimality of Bayesian quadrature has not been established before, while recently there has been extensive theoretical analysis on Bayesian quadrature [4,8,9,44].

This paper is organized as follows. Section 2 provides various definitions, notation and preliminaries including reviews on kernel-based quadrature rules. Section 3 then establishes a rate of convergence for the worst-case error of Bayesian quadrature in a Sobolev space. Section 4 presents the main contributions on the convergence analysis in misspecified settings, and Sect. 5 demonstrates these results by applying them to Bayesian quadrature. We illustrate the obtained theoretical results with simulation experiments in Sect. 6. Finally Sect. 7 concludes the paper with possible future directions.

Preliminary results This paper expands on preliminary results reported in a conference paper by the authors [29]. Specifically, this paper is a complete version of the results presented in Section 5 of [29]. The current paper contains significantly new topics mainly in the following points: (i) We establish the rate of convergence for

Bayesian quadrature with deterministic design points and show that it can achieve minimax optimal rates in Sobolev spaces (Sect. 3); (ii) we apply our general convergence guarantees in misspecified settings to the specific case of Bayesian quadrature and reveal the conditions required for Bayesian quadrature to be robust to misspecification (Sect. 5); to make the contribution (ii) possible, we derive finite sample bounds on quadrature error in misspecified settings (Sect. 4). These results are not included in the conference paper.

We also mention that this paper does not contain the results presented in Section 4 of the conference paper [29], which deal with *randomized* design points. For randomized design points, theoretical analysis can be done based on an approximation theory developed in the statical learning theory literature [12]. On the other hand, the analysis in the deterministic case makes use of the approximation theory developed by [37], which is based on Calderón's decomposition formula in harmonic analysis [19]. This paper focuses on the deterministic case, and we will report a complete version of the randomized case in a forthcoming paper.

Related work The setting of this paper is complementary to that of [45], in which the integrand is *smoother* than assumed. That paper proposes to apply the control functional method by [46] to quasi-Monte Carlo integration, in order to make it adaptable to the (unknown) greater smoothness of the integrand.

Another related line of research is the proposals of quadrature rules that are adaptive to less smooth integrands [14–16,20,23]. For instance, [20] proposed a kernel-based quadrature rule on a finite-dimensional *sphere*. Their method is essentially a Bayesian quadrature using a specific kernel designed for spheres. They derive convergence rates for this method in both well-specified and misspecified settings and obtain results similar to ours. The current work differs from [20] in mainly two aspects: (i) Quadrature problems are considered in standard Euclidean spaces, as opposed to spheres; (ii) a generic framework is presented, as opposed to the analysis of a specific quadrature rule. See also a recent work by [62], in which Bayesian quadrature for vector-valued numerical integration is proposed and its adaptability to the less smooth integrands is discussed.

Quasi-Monte Carlo rules based on a certain digit interlacing algorithm [14–16,23] are also shown to be adaptive to the (unknown) lower smoothness of an integrand. These papers assume that an integrand is in an *anisotropic* function class in which every function possesses (square-integrable) partial mixed derivatives of order $\alpha \in \mathbb{N}$ in *each* variable. Examples of such spaces include Korobov spaces, Walsh spaces and Sobolev spaces of dominating mixed smoothness (see, e.g., [17,42]). In their notation, an integer d, which is a parameter called an interlacing factor, can be regarded as an assumed smoothness. Then, if an integrand belongs to an anisotropic function class with smoothness $\alpha \in \mathbb{N}$ such that $\alpha \leq d$, the rate of the form $O(n^{-\alpha+\varepsilon})$ (or $O(n^{-\alpha-1/2+\varepsilon})$ in a randomized setting) is guaranteed for the quadrature error for arbitrary $\varepsilon > 0$. The present work differs from these works in that (i) isotropic Sobolev spaces are discussed, where the order of differentiability is identical in all directions of variables, and that (ii) theoretical guarantees are provided for generic quadrature rules, as opposed to analysis of specific quadrature methods.

2 Preliminaries

2.1 Basic Definitions and Notation

We will use the following notation throughout the paper. The set of positive integers is denoted by $\mathbb{N}$, and $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. For $\alpha := (\alpha_1, \ldots, \alpha_d)^T \in \mathbb{N}_0^d$, we write $|\alpha| := \sum_{i=1}^d \alpha_i$. The d-dimensional Euclidean space is denoted by $\mathbb{R}^d$, and the closed ball of radius $R > 0$ centered at $z \in \mathbb{R}^d$ by $B(z, R)$. For $a \in \mathbb{R}$, $\lfloor a \rfloor$ is the greatest integer that is less than a. For a set $\Omega \subset \mathbb{R}^d$, $\mathrm{diam}(\Omega) := \sup_{x,y\in\Omega} \|x - y\|$ is the diameter of Ω.

Let $p > 0$ and μ be a Borel measure on a Borel set Ω in $\mathbb{R}^d$. The Banach space $L_p(\mu)$ of p-integrable functions is defined in the standard way with norm $\|f\|_{L_p(\mu)} = (\int |f(x)|^p d\mu(x))^{1/p}$, and $L_\infty(\Omega)$ is the class of essentially bounded measurable functions on Ω with norm $\|f\|_{L_\infty(\Omega)} := \operatorname{ess\,sup}_{x\in\Omega} |f(x)|$. If μ is the Lebesgue measure on $\Omega \subset \mathbb{R}^d$, we write $L_p(\Omega) := L_p(\mu)$ and further $L_p := L_p(\mathbb{R}^d)$ for $p \in \mathbb{N} \cup \{\infty\}$. For $f \in L_1(\mathbb{R}^d)$, its Fourier transform $\hat{f}$ is defined by

$$\hat{f}(\xi) := \int_{\mathbb{R}^d} f(x) e^{-i\xi^T x} \mathrm{d}x, \quad \xi \in \mathbb{R}^d,$$

where $i := \sqrt{-1}$.

For $s \in \mathbb{N}$ and an open set Ω in $\mathbb{R}^d$, $C^s(\Omega)$ denotes the vector space of all functions on Ω that are continuously differentiable up to order s, and $C_B^s(\Omega) \subset C^s(\Omega)$ the Banach space of all functions whose partial derivatives up to order s are bounded and uniformly continuous. The norm of $C_B^s(\Omega)$ is given by $\|f\|_{C_B^s(\Omega)} := \sum_{\alpha\in\mathbb{N}_0^d:|\alpha|\le s} \sup_{x\in\Omega} |\partial^\alpha f(x)|$, where ∂^α is the partial derivative with multi-index $\alpha \in \mathbb{N}_0^d$. The Banach space of the continuous functions that vanish at infinity is denoted by $C_0 := C_0(\mathbb{R}^d)$ with sup norm. Let $C_0^s := C_0^s(\mathbb{R}^d) := C_0(\mathbb{R}^d) \cap C_B^s(\mathbb{R}^d)$ be a Banach space with the norm $\|f\|_{C_0^s(\mathbb{R}^d)} := \|f\|_{C_B^s(\mathbb{R}^d)}$.

For function f and a measure μ on $\mathbb{R}^d$, the support of f and μ is denoted by $\mathrm{supp}(f)$ and $\mathrm{supp}(\mu)$, respectively. The restriction of f to a subset $\Omega \in \mathbb{R}^d$ is denoted by $f|_\Omega$.

Let F and F^* be normed vector spaces with norms $\|\cdot\|_F$ and $\|\cdot\|_{F^*}$, respectively. Then, F and F^* are said to be *norm-equivalent*, if $F = F^*$ as a set, and there exist constants $C_1, C_2 > 0$ such that $C_1\|f\|_{F^*} \le \|f\|_F \le C_2\|f\|_{F^*}$ for all $f \in F$. For a Hilbert space $\mathcal{H}$ with inner product $\langle\cdot,\cdot\rangle_\mathcal{H}$, the norm of $f \in \mathcal{H}$ is denoted by $\|f\|_\mathcal{H}$.

2.2 Sobolev Spaces and Reproducing Kernel Hilbert Spaces

Here we briefly review key facts regarding Sobolev spaces necessary for stating and proving our contributions; for details, we refer to [1,6,59]. We first introduce reproducing kernel Hilbert spaces. For details, see, e.g., [58, Section 4] and [61, Section 10].

Let Ω be a set. A Hilbert space $\mathcal{H}$ of real-valued functions on Ω is a reproducing kernel Hilbert space (RKHS) if the functional $f \mapsto f(x)$ is continuous for any $x \in \Omega$. Let $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ be the inner product of $\mathcal{H}$. Then, there is a unique function $k_x \in \mathcal{H}$ such that $f(x) = \langle f, k_x \rangle_{\mathcal{H}}$. The kernel defined by $k(x, y) := k_x(y)$ is positive definite and called reproducing kernel of $\mathcal{H}$. It is known (Moore–Aronszajn theorem [2]) that for every positive definite kernel $k : \Omega \times \Omega \to \mathbb{R}$ there exists a unique RKHS $\mathcal{H}$ with k as the reproducing kernel. Therefore, the notation $\mathcal{H}_k$ is used to the RKHS associated with k.

In the following, we will introduce two definitions of Sobolev spaces, i.e., (10) and (11), as both will be used throughout our analysis.

For a measurable set $\Omega \subset \mathbb{R}^d$ and $r \in \mathbb{N}$, a Sobolev space $W_2^r(\Omega)$ of order r on Ω is defined by

$$W_2^r(\Omega) := \{f \in L_2(\Omega) : D^\alpha f \in L_2(\Omega) \text{ exists for all } \alpha \in \mathbb{N}_0^d \text{ with } |\alpha| \le r\}, \quad (10)$$

where $D^\alpha f$ denotes the α-th weak derivative of f. This is a Hilbert space with inner product

$$\langle f, g \rangle_{W_2^r(\Omega)} = \sum_{|\alpha| \le r} \langle D^\alpha f, D^\alpha g \rangle_{L_2(\Omega)}, \quad f, g \in W_2^r(\Omega).$$

For a positive real $r > 0$, another definition of Sobolev space of order r on $\mathbb{R}^d$ is given by

$$H^r(\mathbb{R}^d) := \left\{ f \in L_2(\mathbb{R}^d) : \int |\hat{f}(\xi)|^2 \hat{\Phi}(\xi)^{-1} d\xi < \infty \right\}, \quad (11)$$

where the function $\hat{\Phi} : \mathbb{R}^d \to \mathbb{R}$ is defined by

$$\hat{\Phi}(\xi) := (1 + \|\xi\|^2)^{-r}, \quad \xi \in \mathbb{R}^d.$$

The inner product of $H^r(\mathbb{R}^d)$ is defined by

$$\langle f, g \rangle_{H^r(\mathbb{R}^d)} := \int \hat{f}(\xi) \overline{\hat{g}(\xi)} \hat{\Phi}(\xi)^{-1} d\xi, \quad f, g \in H^r(\mathbb{R}^d),$$

where $\overline{\hat{g}(\xi)}$ denotes the complex conjugate of $\hat{g}(\xi)$.

For a measurable set Ω in $\mathbb{R}^d$, the (fractional order) Sobolev space $H^r(\Omega)$ is defined by the restriction of $H^r(\mathbb{R}^d)$; namely (see, e.g., [59, Eq. (1.8) and Definition 4.10])

$$H^r(\Omega) := \left\{ f : \Omega \to \mathbb{R} : f = g|_\Omega, \ \exists\, g \in H^r(\mathbb{R}^d) \right\}$$

with its norm defined by

$$\|f\|_{H^r(\Omega)} := \inf \left\{ \|g\|_{H^r(\mathbb{R}^d)} : g \in H^r(\mathbb{R}^d) \text{ s.t. } f = g|_\Omega \right\}.$$

If $r \in \mathbb{N}$ and Ω is an open set with Lipschitz boundary (see Definition 3), then $H^r(\Omega)$ is norm-equivalent to $W_2^r(\Omega)$ (see, e.g., [59, Eqs. (1.8), (4.20)]).

If $r > d/2$, the Sobolev space $H^r(\mathbb{R}^d)$ is an RKHS [61, Section 10]. In fact, the condition $r > d/2$ guarantees that the function $\hat{\Phi}(\xi) = (1 + \|\xi\|^2)^{-r}$ is integrable, so that $\hat{\Phi}(\xi)$ has a (inverse) Fourier transform

$$\Phi(x) = \frac{2^{1-r}}{\Gamma(r)} \|x\|^{r-d/2} K_{r-d/2}(\|x\|),$$

where Γ denotes the gamma function and $K_{r-d/2}$ is the modified Bessel function of the third kind of order $r - d/2$. The function Φ is positive definite, and the kernel $\Phi(x - y)$ gives $H^r(\mathbb{R}^d)$ as an RKHS. This kernel $\Phi(x - y)$ is essentially a Matérn kernel [33,34] with specific parameters. A *Wendland kernel* [60] also defines an RKHS that is norm-equivalent to $H^r(\mathbb{R}^d)$.

2.3 Kernel-Based Quadrature Rules

We briefly review basic facts regarding kernel-based quadrature rules necessary to describe our results. For details, we refer to [9,17].

Let $\Omega \subset \mathbb{R}^d$ be an open set, k be a measurable kernel on Ω, and $\mathcal{H}_k(\Omega)$ be the RKHS of k with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_k(\Omega)}$. Suppose P is a Borel probability measure on $\mathbb{R}^d$ with its support contained in Ω, and $\{(w_i, X_i)\}_{i=1}^n \subset (\mathbb{R} \times \Omega)^n$ is weighted points, which serve for quadrature. For an integrand f, define $Pf := \int f(x) \mathrm{d}P(x)$ and $P_n f := \sum_{i=1}^n w_i f(X_i)$, respectively, as the integral and a quadrature estimate as in (2). As mentioned in Sect. 1, a kernel quadrature rule aims at minimizing the worst-case error

$$e_n(P; \mathcal{H}_k(\Omega)) := \sup_{f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k(\Omega)} \le 1} |Pf - P_n f|. \tag{12}$$

Assume $\int \sqrt{k(x,x)} \, \mathrm{d}P(x) < \infty$, and define m_P, m_{P_n}[1] $\in \mathcal{H}_k(\Omega)$ by

$$m_P(y) := \int k(y,x) \mathrm{d}P(x), \quad m_{P_n}(y) := \sum_{i=1}^n w_i k(y, X_i), \quad y \in \Omega, \tag{13}$$

where the integral for m_P is understood as the Bochner integral. It is easy to see that, for all $f \in \mathcal{H}$,

$$Pf = \langle f, m_P \rangle_{\mathcal{H}_k(\Omega)}, \quad P_n f = \langle f, m_{P_n} \rangle_{\mathcal{H}_k(\Omega)}.$$

Worst-case error (12) can then be written as

$$e_n(P; \mathcal{H}_k(\Omega)) = \|m_P - m_{P_n}\|_{\mathcal{H}_k(\Omega)}, \tag{14}$$

[1] In the machine learning literature, the function m_P is known as *kernel mean embedding*, and the worst-case error is called the *maximum mean discrepancy*, which have been used in a variety of problems including two-sample testing [24,36,56].

and for any $f \in \mathcal{H}_k(\Omega)$

$$|P_n f - Pf| \leq \|f\|_{\mathcal{H}_k(\Omega)} e_n(P; \mathcal{H}_k(\Omega)). \tag{15}$$

It follows from (14) that

$$e_n^2(P; \mathcal{H}_k(\Omega)) = \int\int k(x, \tilde{x}) \mathrm{d}P(x) \mathrm{d}P(\tilde{x}) - 2 \sum_{i=1}^{n} w_i \int k(x, X_i) \mathrm{d}P(x) + \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j k(X_i, X_j). \tag{16}$$

The integrals in (16) are known in closed form for many pairs of k and P (see, e.g., Table 1 of [9]); for instance, it is known if k is a Wendland kernel and P is the uniform distribution on a ball in $\mathbb{R}^d$. One can then explicitly use formula (16) in order to obtain weighted points $\{(w_i, X_i)\}$ that minimizes worst-case error (12).

2.4 Examples of Kernel-Based Quadrature Rules

Bayesian quadrature This is a class of kernel-based quadrature rules that has been studied extensively in the literature on statistics and machine learning [4,7–9,13,22,25,27,35,46,48,49,51]. In Bayesian quadrature, design points $X_1, \ldots, X_n$ may be obtained jointly in a deterministic manner [9,13,35,48,51], sequentially (adaptively) [8,25,27,49] or randomly [4,7,9,22,46]. For instance, [9] proposed to generate design points randomly as a Markov chain Monte Carlo sample, or deterministically by a quasi-Monte Carlo rule, specifically as a higher-order digital net [15].

Given the design points being fixed, quadrature weights $w_1, \ldots, w_n$ are then obtained by the minimization of worst-case error (16), which can be done analytically by solving a linear system of size n. To describe this, let $X_1, \ldots, X_n$ be design points such that the kernel matrix $K := (k(X_i, X_j))_{i,j}^n \in \mathbb{R}^{n \times n}$ is invertible. The weights are then given by

$$\boldsymbol{w} := (w_1, \ldots, w_n)^T = K^{-1} \boldsymbol{z} \in \mathbb{R}^n, \tag{17}$$

where $\boldsymbol{z} := (m_P(X_i))_{i=1}^n \in \mathbb{R}^n$, with m_P defined in (13).

This way of constructing the estimate $P_n f$ is called *Bayesian quadrature*, since $P_n f$ can be seen as a posterior estimate in a certain Bayesian inference problem with f generated as sample of a Gaussian process (see, e.g., [27] and [9]).

Quasi-Monte Carlo Quasi-Monte Carlo (QMC) methods are equal-weight quadrature rules designed for the uniform distribution on a hypercube $[0, 1]^d$ [17]. Modern QMC methods make use of RKHSs and the associated kernels to define and calculate the worst-case error in order to obtain good design points (e.g., [14,18,26,54]). Therefore, such QMC methods are instances of kernel-based quadrature rules; see [42] and [17] for a review.

Kernel herding In the machine learning literature, an equal-weight quadrature rule called *kernel herding* [11] has been studied extensively [5,27,28,32]. It is an algorithm that greedily searches for design points so as to minimize the worst-case error in an RKHS. In contrast to QMC methods, kernel herding may be used with an arbitrarily distribution P on a generic measurable space, given that the integral $\int k(\cdot, x)\mathrm{d}P(x)$ admits a closed-form solution with a reproducing kernel k. It has been shown that a fast rate $O(n^{-1})$ is achievable for the worst-case error, when the RKHS is finite-dimensional [11]. While empirical studies indicate that the fast rate would also hold in the case of an infinite-dimensional RKHS, its theoretical proof remains an open problem [5].

3 Convergence Rates of Bayesian Quadrature

This section discusses the convergence rates of Bayesian quadrature in well-specified settings. It is shown that Bayesian quadrature can achieve the minimax optimal rates for deterministic quadrature rules in Sobolev spaces. The result also serves as a preliminary to Sect. 5, where misspecified cases are considered.

Let Ω be an open set in $\mathbb{R}^d$ and $X^n := \{X_1, \ldots, X_n\} \subset \Omega$. The main notion to express the convergence rate is fill distance $h_{X^n,\Omega}$ (8), which plays a central role in the literature on scattered data approximation [61] and has been used in the theoretical analysis of Bayesian quadrature in [9,44].

It is necessary to introduce some conditions on Ω. The first one is the *interior cone condition* [61, Definition 3.6], which is a regularity condition on the boundary of Ω. A *cone* $C(x, \xi(x), \theta, R)$ with vertex $x \in \mathbb{R}^d$, direction $\xi(x) \in \mathbb{R}^d$ ($\|\xi(x)\| = 1$), angle $\theta \in (0, 2\pi)$ and radius $R > 0$ is defined by

$$C(x, \xi(x), \theta, R) := \{x + \lambda y : \ y \in \mathbb{R}^d, \ \|y\| = 1, \ \langle y, \xi(x)\rangle \geq \cos\theta, \ \lambda \in [0, R]\}.$$

Definition 1 (*Interior cone condition*) A set $\Omega \subset \mathbb{R}^d$ is said to satisfy an interior cone condition if there exist an angle $\theta \in (0, 2\pi)$ and a radius $R > 0$ such that every $x \in \Omega$ is associated with a unit vector $\xi(x)$ so that the cone $C(x, \xi(x), \theta, R)$ is contained in Ω.

The interior cone condition requires that there is no 'pinch point' (i.e., a $\prec$-shape region) on the boundary of Ω; see also [44].

Next, the notions of special Lipschitz domain [57, p.181] and Lipschitz boundary[2] are defined as follows (see [57, p.189]; [6, Definition 1.4.4]).

Definition 2 (*Special Lipschitz domain*) For $d \geq 2$, an open set $\Omega \subset \mathbb{R}^d$ is called a special Lipschitz domain, if there exists a rotation of Ω, denoted by $\tilde{\Omega}$, and a function $\varphi : \mathbb{R}^{d-1} \to \mathbb{R}$ that satisfy the following:

1. $\tilde{\Omega} = \{(x, y) \in \mathbb{R}^d : y > \varphi(x)\}$;

[2] The definition of the Lipschitz boundary in [6] is identical to the definition of the *minimally smooth boundary* in [57, p.189]. This boundary condition was introduced by Elias M. Stein to prove the so-called *Stein's extension theorem* for Sobolev spaces [57, p.181].

2. φ is a Lipschitz function such that $|\varphi(x) - \varphi(x')| \leq M\|x - x'\|$ for all $x, x' \in \mathbb{R}^{d-1}$, where $M > 0$.

The smallest constant M for φ is called the Lipschitz bound of Ω.

Definition 3 (*Lipschitz boundary*) Let $\Omega \subset \mathbb{R}^d$ be an open set and $\partial\Omega$ be its boundary. Then, $\partial\Omega$ is called a Lipschitz boundary, if there exist constants $\varepsilon > 0$, $N \in \mathbb{N}$, $M > 0$, and open sets $U_1, U_2, \ldots, U_L \subset \mathbb{R}^d$, where $L \in \mathbb{N} \cup \{\infty\}$, such that the following conditions are satisfied:

1. For any $x \in \partial\Omega$, there exists an index i such that $B(x, \varepsilon) \subset U_i$, where $B(x, \varepsilon)$ is the ball centered at x and radius ε;
2. $U_{i_1} \cap \cdots \cap U_{i_{N+1}} = \emptyset$ for any distinct indices $\{i_1, \ldots, i_{N+1}\}$;
3. For each index i, there exists a special Lipschitz domain $\Omega_i \subset \mathbb{R}^d$ with Lipschitz bound b such that $U_i \cap \Omega = U_i \cap \Omega_i$ and $b \leq M$.

Examples of a set Ω having a Lipschitz boundary include: (i) Ω is an open bounded set whose boundary $\partial\Omega$ is C^1 embedded in $\mathbb{R}^d$; (ii) Ω is an open bounded convex set [57, p.189].

Proposition 1 *Let $\Omega \subset \mathbb{R}^d$ be a bounded open set such that an interior cone condition is satisfied and the boundary $\partial\Omega$ is Lipschitz, and P be a probability distribution on $\mathbb{R}^d$ with a bounded density function p such that* $\mathrm{supp}(P) \subset \Omega$. *For $r \in \mathbb{R}$ with $\lfloor r \rfloor > d/2$, k_r is a kernel on $\mathbb{R}^d$ that satisfies Assumption 1 and $\mathcal{H}_{k_r}(\Omega)$ is the RKHS of k_r restricted on Ω. Suppose that $X^n := \{X_1, \ldots, X_n\} \subset \Omega$ are finite points such that $G := (k_r(X_i, X_j))_{i,j=1}^n \in \mathbb{R}^{n\times n}$ is invertible, and $w_1, \ldots, w_n$ are the quadrature weights given by (17). Then, there exist constants $C > 0$ and $h_0 > 0$ independent of X^n, such that*

$$e_n(P; \mathcal{H}_{k_r}(\Omega)) \leq C h_{X^n,\Omega}^r,$$

provided that $h_{X^n,\Omega} \leq h_0$, where $e_n(P; \mathcal{H}_{k_r}(\Omega))$ is the worst-case error for the quadrature rule $\{(w_i, X_i)\}_{i=1}^n$.

Proof The proof idea is borrowed from [9, Theorem 1]. Let $f \in \mathcal{H}_{k_r}(\Omega)$ be arbitrary and fixed. Define a function $f_n \in \mathcal{H}_{k_r}(\Omega)$ by

$$f_n := \sum_{i=1}^n \alpha_i k_r(\cdot, X_i)$$

where $\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_n)^T = K^{-1}\boldsymbol{f} \in \mathbb{R}^n$ and $\boldsymbol{f} := (f(X_1), \ldots, f(X_n)) \in \mathbb{R}^n$. This function is an interpolant of f on X^n such that $f(X_i) = f_n(X_i)$ for all $X_i \in X^n$

It follows from the norm equivalence that $f \in H^r(\Omega)$ and

$$\|f\|_{H^r(\Omega)} \leq C_1 \|f\|_{\mathcal{H}_{k_r}(\Omega)}, \tag{18}$$

where $C_1 > 0$ is a constant.

We see that $\sum_{i=1}^n w_i f(X_i) = \int f_n(x)\mathrm{d}P(x)$. In fact, recalling that the weights $\boldsymbol{w} := (w_1, \ldots, w_n)^T$ are defined as $\boldsymbol{w} = K^{-1}\boldsymbol{z}$, where $\boldsymbol{z} := (z_1, \ldots, z_n)^T$ with $z_i := \int k_r(x, X_i)\mathrm{d}P(x)$, it follows that

$$\sum_{i=1}^n w_i f(X_i) = \boldsymbol{w}^T \boldsymbol{f} = \boldsymbol{z}^T K^{-1} \boldsymbol{f} = \boldsymbol{z}^T \boldsymbol{\alpha}$$
$$= \sum_{i=1}^n \alpha_i \int k_r(x, X_i)\mathrm{d}P(x) = \int f_n(x)\mathrm{d}P(x).$$

Using this identity, we have

$$\left| \int f(x)\mathrm{d}P(x) - \sum_{i=1}^n w_i f(X_i) \right| = \left| \int f(x)\mathrm{d}P(x) - \int f_n(x)\mathrm{d}P(x) \right|$$
$$\leq \|f - f_n\|_{L_1(\Omega)} \|p\|_{L_\infty(\Omega)}$$
$$\leq C_0 \|f\|_{H^r(\Omega)} h^r_{X^n,\Omega} \|p\|_{L_\infty(\Omega)} \quad (19)$$
$$\leq C_0 C_1 \|f\|_{\mathcal{H}_{k_r}(\Omega)} h^r_{X^n,\Omega} \|p\|_{L_\infty(\Omega)}, \quad (20)$$

where (19) follows from Theorem 11.32 and Corollary 11.33 in [61] (where we set $m := 0$, $p := 2$, $q := 1$, $k := \lfloor r \rfloor$ and $s := r - \lfloor r \rfloor$), and (20) from (18). Note that constant C_0 depends only on r, d and the constants in the interior cone condition (which follows from the fact that Theorem 11.32 in [61] is derived from Proposition 11.30 in [61]). Setting $C := C_0 C_1 \|p\|_\infty$ completes the proof. □

Remark 1 – Typically, the fill distance $h_{X^n,\Omega}$ decreases to 0 as the number n of design points increases. Therefore, the upper bound $C h^r_{X^n\Omega}$ provides a faster rate of convergence for $e_n(P; W_2^r(\Omega))$ by a larger value of the degree r of smoothness.

– The condition $h_{X^n,\Omega} \leq h_0$ requires that the design points $X^n = \{X_1, \ldots, X_n\}$ must cover the set Ω to a certain extent in order to guarantee the error bound to hold. This requirement arises since we have used a result from the scattered data approximation literature [61, Corollary 11.33] to derive inequality (19) in our proof. In the literature, such a condition is necessary and we refer an interested reader to Section 11 of [61] and references therein.

– The constant $h_0 > 0$ depends only on the constants θ and R in the interior cone condition (Definition 1). The explicit form is $h_0 := Q(\lfloor r \rfloor, \theta)R$, where $Q(\lfloor r \rfloor, \theta) := \frac{\sin\theta \sin\psi}{8\lfloor r \rfloor^2 (1+\sin\theta)(1+\sin\psi)}$ with $\psi := 2 \arcsin \frac{\sin\theta}{4(1+\sin\theta)}$ [61, p.199].

The following is an immediate corollary to Proposition 1.

Corollary 1 *Assume that Ω, P and r satisfy the conditions in Proposition 1. Suppose that $X^n := \{X_1, \ldots, X_n\} \subset \Omega$ are finite points such that $G := (k_r(X_i, X_j))_{i,j=1}^n \in \mathbb{R}^{n\times n}$ is invertible and $h_{X^n,\Omega} = O(n^{-\alpha})$ for some $0 < \alpha \leq 1/d$ as $n \to \infty$, and $w_1, \ldots, w_n$ are the quadrature weights given by (17) based on X^n. Then, we have*

$$e_n(P; \mathcal{H}_{k_r}(\Omega)) = O(n^{-\alpha r}) \quad (n \to \infty), \quad (21)$$

where $e_n(P; \mathcal{H}_{k_r}(\Omega))$ is the worst-case error of the quadrature rule $\{(w_i, X_i)\}_{i=1}^n$.

Remark 2 – Result (21) implies that the same rate is attainable for the Sobolev space $H^r(\Omega)$ (instead of $H_{k_r}(\Omega)$):

$$e_n(P; H^r(\Omega)) = O(n^{-\alpha r}) \quad (n \to \infty) \tag{22}$$

with (the sequence of) the same weighted points $\{(w_i, X_i)\}_{i=1}^\infty$. This follows from the norm equivalence between $\mathcal{H}_{k_r}(\Omega)$ and $H^r(\Omega)$.

– If the fill distance satisfies $h_{X^n,\Omega} = O(n^{-1/d})$ as $n \to \infty$, then $e_n(P; H^r(\Omega)) = O(n^{-r/d})$. This rate is minimax optimal for the deterministic quadrature rules for the Sobolev space $H^r(\Omega)$ on a hypercube [40, Proposition 1 in Section 1.3.12]. Corollary 1 thus shows that Bayesian quadrature achieves the minimax optimal rate in this setting.

– The decay rate for the fill distance $h_{X^n,\Omega} = O(n^{-1/d})$ holds when, for example, the design points $X^n = \{X_1, \ldots, X_n\}$ are equally spaced grid points in Ω. Note that this rate cannot be improved: If the fill distance decreased at a rate faster than $O(n^{-1/d})$, then $e_n(P; H^r(\Omega))$ would decrease more quickly than the minimax optimal rate, which is a contradiction.

4 Main Results

This section presents the main results on misspecified settings. Two results based on different assumptions are discussed: one on the quadrature weights in Sect. 4.1 and the other on the design points in Sect. 4.2. The approximation theory for Sobolev spaces developed by [37] is employed in the results.

4.1 Convergence Rates Under an Assumption on Quadrature Weights

Theorem 1 *Let $\Omega \subset \mathbb{R}^d$ be an open set whose boundary is Lipschitz, P be a probability distribution on $\mathbb{R}^d$ with $\mathrm{supp}(P) \subset \Omega$, r be a real number with $r > d/2$, and s be a natural number with $s \leq r$. Let k_r denote a kernel on $\mathbb{R}^d$ satisfying Assumption 1, and $\mathcal{H}_{k_r}(\Omega)$ the RKHS of k_r restricted on Ω. Then, for any $\{(w_i, X_i)\}_{i=1}^n \in (\mathbb{R} \times \Omega)^n$, $f \in C_B^s(\Omega) \cap H^s(\Omega) \cap L_1(\Omega)$, and $\sigma > 0$, we have*

$$\begin{aligned} |P_n f - Pf| \leq c_1 \left(\sum_{i=1}^n |w_i| + 1 \right) \sigma^{-s} \|f\|_{C_B^s(\Omega)} \\ + c_2 (1 + \sigma^2)^{\frac{r-s}{2}} e_n(P; \mathcal{H}_{k_r}(\Omega)) \|f\|_{H^s(\Omega)}, \end{aligned} \tag{23}$$

where $c_1, c_2 > 0$ are constants independent of $\{(w_i, X_i)\}_{i=1}^n$, f and σ.

Proof We first derive some inequalities used for proving the assertion. It follows from norm equivalence that $f \in W_2^s(\Omega)$, where $W_2^s(\Omega)$ is the Sobolev space defined via weak derivatives. Since Ω has a Lipschitz boundary, Stein's extension theorem [57,

p.181] guarantees that there exists a bounded linear extension operator $\mathfrak{E} : W_2^s(\Omega) \to W_2^s(\mathbb{R}^d)$ such that

$$\mathfrak{E}(f)(x) = f(x), \quad \forall x \in \Omega, \tag{24}$$

$$\|\mathfrak{E}(f)\|_{W_2^s(\mathbb{R}^d)} \le C_1 \|f\|_{W_2^s(\Omega)}, \tag{25}$$

where C_1 is a constant independent of the choice of f. From the norm equivalence and (25), there is a constant C_2 such that

$$\|\mathfrak{E} f\|_{H^s(\mathbb{R}^d)} \le C_2 \|f\|_{H^s(\Omega)}. \tag{26}$$

Since $f \in L_1(\Omega)$, the extension also satisfies $\mathfrak{E}(f) \in L_1(\mathbb{R}^d)$ [57, p.181]. In addition, by the construction of $\mathfrak{E}$ [57, Eqs.(24)(31) on p.191], one can show [38, Section 3.2.2] that $\mathfrak{E}$ is also a linear bounded operator from $C_B^s(\Omega)$ to $C_0^s(\mathbb{R}^d)$, that is,

$$\|\mathfrak{E} f\|_{C_0^s(\mathbb{R}^d)} \le C_3 \|f\|_{C_B^s(\Omega)}, \tag{27}$$

for some constant $C_3 > 0$. Below we write $\tilde{f} := \mathfrak{E}(f)$ for notational simplicity.

Let $g_\sigma \in H^r(\mathbb{R}^d)$ be the approximate function of $\tilde{f}$ defined as (B.10) by Calderón's formula ("Appendix B.2"; we set $f := \tilde{f}$). The property $\tilde{f} \in C_0^s(\mathbb{R}^d) \cap H^s(\mathbb{R}^d) \cap L_1(\mathbb{R}^d)$ enables the use of Proposition 3.7 of [37] (where we set $k := s$ and $\alpha := 0$; see Proposition A.1 in "Appendix A" for a review), which gives in combination with (27) that

$$\|\tilde{f} - g_\sigma\|_{L_\infty(\mathbb{R}^d)} \le C\sigma^{-s} \|\tilde{f}\|_{C_0^s(\mathbb{R}^d)} \le C_4 \sigma^{-s} \|f\|_{C_B^s(\Omega)}, \tag{28}$$

for some constant $C_4 > 0$ which is independent of f.

From $\tilde{f} \in C_0^s(\mathbb{R}^d) \cap H^s(\mathbb{R}^d) \cap L_1(\mathbb{R}^d)$, Lemma B.6 in "Appendix B.2" can be applied, by which together with (26) we have

$$\|g_\sigma\|_{H^r(\mathbb{R}^d)} \le C_5'(1+\sigma^2)^{\frac{r-s}{2}} \|\tilde{f}\|_{H^s(\mathbb{R}^d)} \le C_5 (1+\sigma^2)^{\frac{r-s}{2}} \|f\|_{H^s(\Omega)} \tag{29}$$

for some constants C_5 and C_5', which are independent of σ and $\tilde{f}$.

With the decomposition

$$|P_n f - Pf| \le \underbrace{|P_n f - P_n g_\sigma|}_{(A)} + \underbrace{|P_n g_\sigma - P g_\sigma|}_{(B)} + \underbrace{|P g_\sigma - Pf|}_{(C)},$$

each of the terms (A), (B) and (C) will be bounded in the following.

First, the term (A) is bounded as

$$\begin{aligned}(A) &\le \sum_{i=1}^n |w_i|\,|f(X_i) - g_\sigma(X_i)| \\ &= \sum_{i=1}^n |w_i| \left|\tilde{f}(X_i) - g_\sigma(X_i)\right| \quad (\because \{X_i\}_{i=1}^n \subset \Omega \text{ and (24)})\end{aligned}$$

$$\leq \left(\sum_{i=1}^{n} |w_i|\right) \|\tilde{f} - g_\sigma\|_{L_\infty(\mathbb{R}^d)} \overset{(28)}{\leq} C_4 \left(\sum_{i=1}^{n} |w_i|\right) \sigma^{-s} \|f\|_{C_B^s(\Omega)}.$$

For the term (B), it follows from the norm equivalence and restriction that for some constant D

$$\|g_\sigma|_\Omega\|_{\mathcal{H}_{k_r}(\Omega)} \leq D\|g_\sigma\|_{H^r(\mathbb{R}^d)}. \tag{30}$$

This inequality and (29) give

$$\begin{aligned}(B) &\leq \|g_\sigma|_\Omega\|_{\mathcal{H}_{k_r}(\Omega)} \left\|m_{P_n} - m_P\right\|_{\mathcal{H}_{k_r}(\Omega)} \\ &\leq D\|g_\sigma\|_{H^r(\mathbb{R}^d)} e_n(P; \mathcal{H}_{k_r}(\Omega)) \\ &\leq DC_5(1+\sigma^2)^{\frac{r-s}{2}} e_n(P; \mathcal{H}_{k_r}(\Omega))\|f\|_{H^s(\Omega)}.\end{aligned}$$

Finally, the term (C) is bounded as

$$(C) \leq \int \left|g_\sigma(x) - \tilde{f}(x)\right| \mathrm{d}P(x) \leq \|g_\sigma - \tilde{f}\|_{L_\infty(\mathbb{R}^d)} \overset{(28)}{\leq} C_4\sigma^{-s}\|f\|_{C_B^s(\Omega)}.$$

Combining these three bounds, the assertion is obtained. □

Remark 3 – The integrand f is assumed to satisfy $f \in H^s(\Omega) \cap C_B^s(\Omega) \cap L_1(\Omega)$, which is slightly stronger than just assuming $f \in H^s(\Omega)$.
– In upper bound (23), the constant $\sigma > 0$ controls the trade-off between the two terms: $c_2(1+\sigma^2)^{\frac{r-s}{2}} e_n(P; \mathcal{H}_{k_r}(\Omega))\|f\|_{H^s(\Omega)}$ and $c_1\left(\sum_{i=1}^n |w_i| + 1\right) \cdot \sigma^{-s}\|f\|_{C_B^s(\Omega)}$. In the proof, the integrand f is approximated by a band-limited function $g_\sigma \in H^r(\Omega)$, where σ is the highest spectrum that g_σ possesses. Thus, the trade-off in the upper bound corresponds to the trade-off between the accuracy of approximation of f by g_σ and the penalty incurred on the regularity of g_σ.

The following result, which is a corollary of Theorem 1, provides a rate of convergence for the quadrature error in a misspecified setting. It is derived by assuming certain rates for the quantity $\sum_{i=1}^n |w_i|$ and the worst-case error $e_n(P; \mathcal{H}_{k_r})$.

Corollary 2 *Let Ω, P, r, s, k_r, and $\mathcal{H}_{k_r}(\Omega)$ be the same as Theorem 1. Suppose that $\{(w_i, X_i)\}_{i=1}^n \in (\mathbb{R} \times \Omega)^n$ satisfies $e_n(P; \mathcal{H}_{k_r}(\Omega)) = O(n^{-b})$ and $\sum_{i=1}^n |w_i| = O(n^c)$ for some $b > 0$ and $c \geq 0$, respectively, as $n \to \infty$. Then, for any $f \in C_B^s(\Omega) \cap H^s(\Omega) \cap L_1(\Omega)$, we have*

$$|P_n f - Pf| = O(n^{-bs/r + c(r-s)/r}) \quad (n \to \infty). \tag{31}$$

Proof Let $\sigma_n := n^\theta > 0$, where $\theta > 0$ will be determined later. Plugging $e_n(P; \mathcal{H}_{k_r}(\Omega)) = O(n^{-b})$ and $\sum_{i=1}^n |w_i| = O(n^c)$ to (23) with $\sigma := \sigma_n$ leads

$$|P_n f - Pf| = O(n^{c-\theta s}) + O(n^{\theta(r-s)-b}).$$

Setting $\theta = (b+c)/r$, which balances the two terms in the right-hand side, completes the proof. □

Remark 4 – The exponent of the rate in (31) consists of two terms: $-bs/r$ and $c(r-s)/r$. The first term $-bs/r$ corresponds to a degraded rate from the original $O(n^{-b})$ by the factor of smoothness ratio s/r, while the second term $c(r-s)/r$ makes the rate slower. The effect of the second term increases as the constant c or the gap $(r-s)$ of misspecification becomes larger.

- The obtained rate recovers $O(n^{-b})$ for $r = s$ (well-specified case) regardless of the value of c.
- Consider the misspecified case $r > s$. If $c > 0$, the term $c(r-s)/r$ always makes the rate slower. It is thus better to have $c = 0$, as in this case we have the rate $O(n^{-bs/r})$ in the misspecified setting. The weights with $\max_{i=1,\dots,n} |w_i| = O(n^{-1})$, such as equal weights $w_i = 1/n$, realize $c = 0$.
- As mentioned earlier, the minimax optimal rate for the worst-case error in the Sobolev space $H^r(\Omega)$ with Ω being a cube in $\mathbb{R}^d$ and P being the Lebesgue measure on Ω is $O(n^{-r/d})$ [40, Proposition 1 in Section 1.3.12]. If design points satisfy $b = r/d$ and $c = 0$ in this setting, Corollary 2 provides the rate $O(n^{-s/d})$ for $f \in H^s(\Omega) \cap C_B^s(\Omega) \cap L_1(\Omega)$. This rate is the same as the minimax optimal rate for $H^s(\Omega)$ and hence implies some adaptivity to the order of differentiability.
- The assumption $\sum_{i=1}^n |w_i| = O(n^c)$ can be also interpreted from a probabilistic viewpoint. Assume that the observation involves noise, $Y_i := f(X_i) + \varepsilon_i$ $(i = 1, \dots, n)$, where ε_i is independent noise with $\mathbb{E}[\varepsilon_i^2] = \sigma_{\mathrm{noise}}^2$ ($\sigma_{\mathrm{noise}} > 0$ is a constant) for $i = 1, \dots, n$, and that Y_i are used for numerical integration. The expected squared error is decomposed as

$$\mathbb{E}_{\varepsilon_1,\dots,\varepsilon_n}\left[\left(\sum_{i=1}^n w_i Y_i - Pf\right)^2\right] = \mathbb{E}_{\varepsilon_1,\dots,\varepsilon_n}\left[\left(P_n f - Pf + \sum_{i=1}^n w_i \varepsilon_i\right)^2\right]$$
$$= |P_n f - Pf|^2 + \sigma_{\mathrm{noise}}^2 \sum_{i=1}^n w_i^2.$$

  In the last expression, the first term $|P_n f - Pf|^2$ is the squared error in the noiseless case, and the second term $\sigma_{\mathrm{noise}}^2 \sum_{i=1}^n w_i^2$ is the error due to noise. Since $\sum_{i=1}^n w_i^2 \le (\sum_{i=1}^n |w_i|)^2 = O(n^{2c})$, the error in the second term may be larger as c increases. Hence, quadrature weights having smaller c are preferable in terms of robustness to the existence of noise; this in turn makes the quadrature rule more robust to the misspecification of the degree of smoothness.

Theorem 1 and Corollary 2 require a control on the absolute sum of the quadrature weights $\sum_{i=1}^n |w_i|$. This is possible with, for instance, equal-weight quadrature rules that seek for good design points. However, the control of $\sum_{i=1}^n |w_i|$ could be difficult for quadrature rules that obtain the weights by optimization based on prefixed design points. This includes the case of Bayesian quadrature that optimizes the weights without any constraint. To deal with such methods, in the next section we will develop

theoretical guarantees that do not rely on the assumption on the quadrature weights, but on a certain assumption on the design points.

4.2 Convergence Rates Under an Assumption on Design Points

This subsection provides convergence guarantees in a misspecified settings under an assumption on the design points. The assumption is described in terms of separation radius (6), which is (the half of) the minimum distance between distinct design points. The separation radius of points $X^n := \{X_1, \ldots, X_n\} \subset \mathbb{R}^d$ is denoted by q_{X^n}. Note that if $X^n \subset \Omega$ for some Ω, then the separation radius lower bounds the fill distance, i.e., $q_{X^n} \leq h_{X^n,\Omega}$.

Henceforth, we will consider a bounded domain Ω, and without loss of generality, we assume that it satisfies $\mathrm{diam}(\Omega) \leq 1$.

Theorem 2 *Let $\Omega \subset \mathbb{R}^d$ be an open bounded set with $\mathrm{diam}(\Omega) \leq 1$ such that the boundary is Lipschitz, P be a probability distribution on $\mathbb{R}^d$ such that $\mathrm{supp}(P) \subset \Omega$, r be a real number with $r > d/2$, and s be a natural number with $s \leq r$. Let k_r denote a kernel on $\mathbb{R}^d$ satisfying Assumption 1, and $\mathcal{H}_{k_r}(\Omega)$ the RKHS of k_r restricted on Ω. For any $\{(w_i, X_i)\}_{i=1}^n \in (\mathbb{R} \times \Omega)^n$ and $f \in C_B^s(\Omega) \cap H^s(\Omega)$, we have*

$$|P_n f - Pf| \leq C \max\left(\|f\|_{C_B^s(\Omega)}, \|f\|_{H^s(\Omega)}\right)\left(q_{X^n}^{-(r-s)} e_n(P; \mathcal{H}_{k_r}(\Omega)) + q_{X^n}^s\right), \tag{32}$$

where $C > 0$ is a constant depending neither on $\{(w_i, X_i)\}_{i=1}^n$ nor on the choice of f and $e_n(P; \mathcal{H}_{k_r}(\Omega))$ is the worst-case error in $\mathcal{H}_{k_r}(\Omega)$ for $\{(w_i, X_i)\}_{i=1}^n$.

Proof By the same argument as the first part of the proof for Theorem 1, there exists an extension of f to $\tilde{f} \in W_2^s(\mathbb{R}^d) \cap C_0^s(\mathbb{R}^d)$ such that

$$\tilde{f}(x) = f(x), \quad \forall x \in \Omega, \tag{33}$$

$$\|\tilde{f}\|_{H^s(\mathbb{R}^d)} \leq C_1 \|f\|_{H^s(\Omega)}, \tag{34}$$

$$\|\tilde{f}\|_{C_0^s(\mathbb{R}^d)} \leq C_2 \|f\|_{C_B^s(\Omega)}, \tag{35}$$

for some positive constants C_i $(i = 1, 2)$. Note also that $f \in L^1(\Omega)$, since $f \in C_B^s(\Omega)$ and Ω is bounded. This implies $\tilde{f} \in L_1(\mathbb{R}^d)$ [57, p.181].

From the above inequalities, there is a constant $C_3 > 0$ independent of the choice of f such that

$$\max\left(\|\tilde{f}\|_{C_0^s(\mathbb{R}^d)}, \|\tilde{f}\|_{H^s(\mathbb{R}^d)}\right) \leq C_3 \max\left(\|f\|_{C_B^s(\Omega)}, \|f\|_{H^s(\Omega)}\right). \tag{36}$$

For notational simplicity, write

$$\sigma_n := \frac{C_d}{q_{X^n}} \tag{37}$$

where $C_d := 24(\frac{\sqrt{\pi}}{3}\Gamma(\frac{d+2}{2}))^{\frac{2}{d+1}}$ with Γ being the Gamma function. From Theorems A.1 and A.2 in "Appendix A" (which are restatements of Theorems 3.5 and 3.10 of [37]), there exists a function $\tilde{f}_{\sigma_n} \in H^r(\mathbb{R}^d)$ such that

$$\tilde{f}_{\sigma_n}(X_i) = \tilde{f}(X_i), \quad (i = 1, \dots, n), \tag{38}$$

$$\|\tilde{f} - \tilde{f}_{\sigma_n}\|_{L_\infty(\mathbb{R}^d)} \le C_{s,d}\sigma_n^{-s} \max(\|\tilde{f}\|_{C_0^s(\mathbb{R}^d)}, \|\tilde{f}\|_{H^s(\mathbb{R}^d)}), \tag{39}$$

where $C_{s,d}$ is a constant depending only on s and d. Combining (39) and (36) obtains

$$\|\tilde{f} - \tilde{f}_{\sigma_n}\|_{L_\infty(\mathbb{R}^d)} \le C_4\sigma_n^{-s} \max\left(\|f\|_{C_B^s(\Omega)}, \|f\|_{H^s(\Omega)}\right), \tag{40}$$

where $C_4 := C_{s,d}C_3$.

From Assumption 1 and $\tilde{f} \in C_B^s(\mathbb{R}^d) \cap H^s(\mathbb{R}^d) \cap L_1(\mathbb{R}^d)$, Lemma A.1 (see "Appendix A") gives

$$\|\tilde{f}_{\sigma_n}\|_{\mathcal{H}_{k_r}(\mathbb{R}^d)} \le C_{s,d,k_r}\sigma_n^{r-s} \max(\|\tilde{f}\|_{C_0^s(\mathbb{R}^d)}, \|\tilde{f}\|_{H^s(\mathbb{R}^d)}),$$

where C_{s,d,k_r} is a constant only depending on r, s, d and k_r. It follows from this inequality and (36) that

$$\|\tilde{f}_{\sigma_n}\|_{\mathcal{H}_{k_r}(\mathbb{R}^d)} \le C_5\sigma_n^{r-s} \max\left(\|f\|_{C_B^s(\Omega)}, \|f\|_{H^s(\Omega)}\right), \tag{41}$$

where $C_5 := C_{s,d,k_r}C_3$.

We are now ready to prove the assertion. In the decomposition

$$|P_n f - Pf| = |P_n\tilde{f} - P\tilde{f}| \le \underbrace{|P_n\tilde{f} - P_n\tilde{f}_{\sigma_n}|}_{(A)} + \underbrace{|P_n\tilde{f}_{\sigma_n} - P\tilde{f}_{\sigma_n}|}_{(B)} + \underbrace{|P\tilde{f}_{\sigma_n} - P\tilde{f}|}_{(C)},$$

the term (A) is zero from (38).

With $\|\tilde{f}_{\sigma_n}|_\Omega\|_{\mathcal{H}_{k_r}(\Omega)} \le \|\tilde{f}_{\sigma_n}\|_{\mathcal{H}_{k_r}(\mathbb{R}^d)}$ ([2], Section 5), the term (B) can be bounded as

$$\begin{aligned}
(B) &= \left|\sum_{i=1}^n w_i \tilde{f}_{\sigma_n}|_\Omega(X_i) - \int \tilde{f}_{\sigma_n}|_\Omega(x)\mathrm{d}P(x)\right| \\
&\le \left|\left\langle \tilde{f}_{\sigma_n}|_\Omega, m_{P_n} - m_P\right\rangle_{\mathcal{H}_{k_r}(\Omega)}\right| \quad (\because \tilde{f}_{\sigma_n}|_\Omega \in \mathcal{H}_{k_r}(\Omega)) \\
&\le \left\|\tilde{f}_{\sigma_n}|_\Omega\right\|_{\mathcal{H}_{k_r}(\Omega)} e_n(P; \mathcal{H}_{k_r}(\Omega)) \\
&\le \left\|\tilde{f}_{\sigma_n}\right\|_{\mathcal{H}_{k_r}(\mathbb{R}^d)} e_n(P; \mathcal{H}_{k_r}(\Omega)) \\
&\overset{(41)}{\le} C_5\sigma_n^{r-s} \max\left(\|f\|_{C_B^s(\Omega)}, \|f\|_{H^s(\Omega)}\right) e_n(P; \mathcal{H}_{k_r}(\Omega)).
\end{aligned}$$

The term (C) is upper bounded as

$$(C) \leq \|\tilde{f}_{\sigma_n} - \tilde{f}\|_{L_\infty(\mathbb{R}^d)} \overset{(39)}{\leq} C_4 \sigma_n^{-s} \max\left(\|f\|_{C_B^s(\Omega)}, \|f\|_{H^s(\Omega)}\right).$$

These bounds complete the proof. □

Remark 5 – From $q_{X^n} \leq h_{X^n}$, the separation radius q_{X^n} typically converges to zero as $n \to \infty$. For the upper bound in (32), the factor $q_{X^n}^{-(r-s)}$ in the first term diverges to infinity as $n \to \infty$, while the second term goes to zero. Thus, q_{X^n} should decay to zero in an appropriate speed depending on the rate of $e_n(P; \mathcal{H}_{k_r}(\Omega))$, in order to make the quadrature error small in the misspecified setting.

– Note that as the gap between r and s becomes large, the effect of the separation radius becomes serious; this follows from the expression $q_{X^n}^{-(r-s)}$.

Based on Theorem 2, we establish below a rate of convergence in a misspecified setting by assuming a certain rate of decay for the separation radius as the number of design points increases.

Corollary 3 *Let $\Omega, P, r, s, k_r, \mathcal{H}_{k_r}(\Omega)$ be the same as in Theorem 2. Suppose $\{(w_i, X_i)\}_{i=1}^n \in (\mathbb{R} \times \Omega)^n$ is design points such that $e_n(P; \mathcal{H}_{k_r}(\Omega)) = O(n^{-b})$ and $q_{X^n} = \Theta(n^{-a})$ for some $b > 0$ and $a > 0$, respectively, as $n \to \infty$. Then, for any $f \in C_B^s(\Omega) \cap H^s(\Omega)$, we have*

$$|P_n f - Pf| = O(n^{-\min(b-a(r-s),as)}) \quad (n \to \infty). \tag{42}$$

In particular, the rate in the right-hand side is optimized when $a = b/r$, which gives

$$|P_n f - Pf| = O(n^{-\frac{bs}{r}}) \quad (n \to \infty). \tag{43}$$

Proof Plugging $e_n(P; \mathcal{H}_{k_r}(\Omega)) = O(n^{-b})$ and $q_{X^n} = \Theta(n^{-a})$ into (32) yields

$$|P_n f - Pf| = O(n^{a(r-s)-b}) + O(n^{-as}) = O(n^{-\min(b-a(r-s),as)}),$$

which proves (42). The second assertion is obvious. □

Remark 6 As stated in the assertion, the best rate for the bound is achieved when $a = b/r$. The resulting rate in (43) coincides with that of Corollary 2 (see (31)) with $c = 0$. Therefore, observations similar to those for Theorem 1 can be made with the rate in (43).

5 Bayesian Quadrature in Misspecified Settings

To demonstrate the results of Sect. 4, a rate of convergence for Bayesian quadrature in misspecified settings is derived. To this end, an upper bound on the integration error of Bayesian quadrature is first provided, when the smoothness of an integrand is

overestimated. It is obtained by combining Theorem 2 in Sect. 4 and Proposition 1 in Sect. 3.

Theorem 3 *Let $\Omega \subset \mathbb{R}^d$ be a bounded open set with $\mathrm{diam}(\Omega) \leq 1$ such that an interior cone condition is satisfied and the boundary is Lipschitz, P be a probability distribution on $\mathbb{R}^d$ with a bounded density function p such that $\mathrm{supp}(P) \subset \Omega$, r be a real number with $\lfloor r \rfloor > d/2$, and s be a natural number with $s \leq r$. Suppose that k_r is a kernel on $\mathbb{R}^d$ satisfying Assumption 1, $X^n := \{X_1, \ldots, X_n\} \subset \Omega$ is design points such that $G := (k_r(X_i, X_j))_{i,j=1}^n \in \mathbb{R}^{n\times n}$ is invertible, and $w_1, \ldots, w_n$ are the Bayesian quadrature weights in (17) based on k_r. Assume that there exist constants $c_q > 0$ and $\delta > 0$ independent of X^n, such that $1 - s/r < \delta \leq 1$ and*

$$h_{X^n,\Omega} \leq c_q q_{X^n}^{\delta}. \tag{44}$$

Then, there exist positive constants C and h_0 independent of X^n, such that for any $f \in C_B^s(\Omega) \cap H^s(\Omega)$, we have

$$|P_n f - Pf| \leq C \max\left(\|f\|_{C_B^s(\Omega)}, \|f\|_{H^s(\Omega)}\right) h_{X^n,\Omega}^{r-(r-s)/\delta}, \tag{45}$$

provided that $h_{X^n,\Omega} \leq h_0$.

Proof Under the assumptions, Theorem 2 gives that

$$|P_n f - Pf| \leq C_1 \max\left(\|f\|_{C_B^s(\Omega)}, \|f\|_{H^s(\Omega)}\right) \left(q_{X^n}^{-(r-s)} e_n(P; \mathcal{H}_{k_r}(\Omega)) + q_{X^n}^s\right), \tag{46}$$

where $C_1 > 0$ is a constant and $e_n(P; \mathcal{H}_{k_r}(\Omega))$ is the worst-case error of $\{(w_i, X_i)\}_{i=1}^n$ in $\mathcal{H}_{k_r}(\Omega)$. On the other hand, Proposition 1 implies that there exist constants $C_2 > 0$ and $h_0 > 0$ independent of the choice of X^n, such that

$$e_n(P; \mathcal{H}_{k_r}(\Omega)) \leq C_2 h_{X^n,\Omega}^r, \tag{47}$$

provided that $h_{X^n,\Omega} \leq h_0$. Note also that (44) implies that

$$q_{X^n}^{-1} \leq c_q^{1/\delta} h_{X^n,\Omega}^{-1/\delta}. \tag{48}$$

From $q_{X^n} \leq h_{X^n,\Omega}$ and the above inequalities, it follows that

$$\begin{aligned}
&|P_n f - Pf| \\
&\overset{(46)}{\leq} C_1 \max\left(\|f\|_{C_B^s(\Omega)}, \|f\|_{H^s(\Omega)}\right) \left(q_{X^n}^{-(r-s)} e_n(P; \mathcal{H}_{k_r}(\Omega)) + q_{X^n}^s\right) \\
&\overset{(47)}{\leq} C_1 \max\left(\|f\|_{C_B^s(\Omega)}, \|f\|_{H^s(\Omega)}\right) \left(C_2 q_{X^n}^{-(r-s)} h_{X^n,\Omega}^r + q_{X^n}^s\right) \\
&\overset{(48)}{\leq} C_1 \max\left(\|f\|_{C_B^s(\Omega)}, \|f\|_{H^s(\Omega)}\right) \left(C_2 c_q^{(r-s)/\delta} h_{X^n,\Omega}^{r-(r-s)/\delta} + q_{X^n}^s\right)
\end{aligned}$$

$$\overset{(\star)}{\leq} C_1 \max\left(\|f\|_{C_B^s(\Omega)}, \|f\|_{H^s(\Omega)}\right)\left(C_2 c_q^{(r-s)/\delta} h_{X^n,\Omega}^{r-(r-s)/\delta} + h_{X^n}^s\right)$$
$$\overset{(\dagger)}{\leq} C_3 \max\left(\|f\|_{C_B^s(\Omega)}, \|f\|_{H^s(\Omega)}\right) h_{X^n,\Omega}^{r-(r-s)/\delta},$$

where C_1, C_2 and C_3 are positive constants independent of the choice of design points X^n, and we used $q_{X^n} \leq h_{X^n,\Omega}$ in $(\star)$, $0 < h_{X^n} \leq 1$ and $0 < r - (r-s)/\delta \leq s$ in $(\dagger)$. □

Remark 7 – Condition (44) implies that

$$c' h_{X^n,\Omega}^{1/\delta} \leq q_{X^n} \leq h_{X^n,\Omega}, \tag{49}$$

where $c' := c_q^{-1/\delta}$ is independent of X^n. This condition is stronger for a larger value of δ, requiring that distinct design points should not be very close to each other. Note that the lower bound $1 - s/r < \delta$ is necessary for the upper bound of error (45) to have a positive exponent, while the upper bound $\delta \leq 1$ follows from $q_{X^n} \leq h_{X^n,\Omega}$, which holds by definition. The constraint $1 - s/r < \delta$ and (49) thus imply that a stronger condition is required for X^n as the degree of misspecification becomes more serious (i.e., as the ratio s/r becomes smaller).

– If condition (44) is satisfied for $\delta = 1$, then the design points X^n are called *quasi-uniform* [53, Section 7.3]. In this case, the bound in (45) is

$$|P_n f - Pf| \leq C \max\left(\|f\|_{C_B^s(\Omega)}, \|f\|_{H^s(\Omega)}\right) h_{X,\Omega}^s. \tag{50}$$

This is the same order of approximation as that of Proposition 1 when $r = s$. Proposition 1 provides an error bound for Bayesian quadrature in a well-specified case, where one knows the degree of smoothness s of the integrand. Therefore, (50) suggests that if the design points are quasi-uniform, then Bayesian quadrature can be adaptive to the (unknown) degree of the smoothness s of the integrand f, even in a situation where one only knows its upper bound $r \geq s$.

We obtain the following as a corollary of Theorem 3. The proof is obvious and omitted.

Corollary 4 *Let Ω, P, r, s, k_r, X^n, G and w_i $(i = 1, \ldots, n)$ be the same as Theorem 3. Assume that there exist constants $c_q > 0$ and $\delta > 0$ independent of X^n, such that $1 - s/r < \delta \leq 1$ and*

$$h_{X^n,\Omega} \leq c_q q_{X^n}^{\delta},$$

and further $h_{X^n,\Omega} = O(n^{-\alpha})$ as $n \to \infty$ for some $0 < \alpha \leq 1/d$. Then, for all $f \in C_B^s(\Omega) \cap H^s(\Omega)$, we have

$$|P_n f - Pf| = O(n^{-\alpha[r-(r-s)/\delta]}) \quad (n \to \infty). \tag{51}$$

In particular, the best possible rate in the right-hand side is achieved when $\delta = 1$ *and* $\alpha = 1/d$*, giving that*

$$|P_n f - Pf| = O(n^{-s/d}) \quad (n \to \infty). \tag{52}$$

Remark 8 – The rate $O(n^{-s/d})$ in (52) matches the minimax optimal rate of deterministic quadrature rules for the worst-case error in the Sobolev space $H^s(\Omega)$ with Ω being a cube [40, Proposition 1 in Section 1.3.12]. Therefore, it is shown that the optimal rate may be achieved by Bayesian quadrature, even in the misspecified setting (under a slightly stronger assumption that $f \in H^s(\Omega) \cap C_B^s(\Omega)$). In other words, Bayesian quadrature may achieve the optimal rate *adaptively*, without knowing the degree s of smoothness of a test function: One just needs to know its upper bound $r \geq s$.

– The main assumptions required for the optimal rate (52) are that (i) $h_{X^n,\Omega} = O(n^{-1/d})$ and that (ii) $h_{X^n,\Omega} \leq c_q q_{X^n}^{\delta}$ for $\delta = 1$. Recall that (i) is the same assumption that is required for the optimal rate $O(n^{-r/d})$ in the well-specified setting $f \in H^r(\Omega)$ (Corollary 1). On the other hand, (ii) is the one required for the finite sample bound in Theorem 3. Both these assumptions are satisfied, for instance, if $X_1, \ldots, X_n$ are grid points in Ω.

6 Simulation Experiments

We conducted simulation experiments to empirically assess the obtained theoretical results. MATLAB code for reproducing the results is available at https://github.com/motonobuk/kernel-quadrature. We focus on Bayesian quadrature in these experiments.

6.1 Problem Setting

Domain, distribution and design points The domain is $\Omega := [0, 1] \subset \mathbb{R}$, and the measure of quadrature P is the uniform distribution over $[0, 1]$. For design points, we consider the following two configurations:

– *Uniform* $X^n = \{X_1, \ldots, X_n\}$ are equally spaced grid points in $[0, 1]$ with $X_1 = 0$ and $X_n = 1$, that is, $X_i = (i - 1)/(n - 1)$ for $i = 1, \ldots, n$.
– *Non-uniform* $X^n = \{X_1, \ldots, X_n\}$ are non-equally spaced points in $[0, 1]$, such that $X_i = (i - 1)/(n - 1)$ if i is odd, and $X_i = X_{i-1} + (n - 1)^{-2}$ if i is even.

For the *uniform* design points, both the fill distance $h_{X^n,\Omega}$ and the separation radius $q_{X^n,\Omega}$ decay at the rate $O(n^{-1})$. On the other hand, for the *non-uniform* points the separation radius decays at the rate $O(n^{-2})$, while the rate of the fill distance remains the same $O(n^{-1})$ as for the uniform points. Using these two different sets of design points, we can observe the effect of the separation radius to the performance of kernel quadrature.

Kernels As before, r denotes the assumed degree of smoothness used for computing quadrature weights, and s denotes the true smoothness of test integrands, both

expressed in terms of Sobolev spaces. As kernels of the corresponding Sobolev spaces, we used Wendland kernels [61, Definition 9.11], which are given as follows [61, Corollary 9.14]. Define the following univariate functions:

$$\begin{aligned}
&\phi_{1,0}(t) := (1-t)_+, \quad \phi_{1,1}(t) := (1-t)_+^3(3t+1),\\
&\phi_{1,2}(t) := (1-t)_+^5(24t+15t+3),\\
&\phi_{1,3}(t) := (1-t)_+^7(315t^3+285t^2+105t+15), \quad t \geq 0,
\end{aligned}$$

where $(x)_+ := \min(0, x)$. The Wendland kernel k_r whose RKHS is norm-equivalent to the Sobolev space $H^r([0,1])$ of order r $(= 1, 2, 3, 4)$ is then defined by $k_r(x, y) := \phi_{d,r-1}(|x-y|/\delta)$ for $x, y \in [0,1]$ [61, Theorem 10.35], where δ is a scale parameter and we set it to be 0.1.

Evaluation measure For each pair of r $(= 1, 2, 3, 4)$ and s $(= 1, 2, 3, 4)$, we first computed quadrature weights $w_1, \ldots, w_n$ by minimizing the worst-case error in $H^r([0,1])$ and then evaluated the quadrature rule $(w_i, X_i)_{i=1}^n$ by computing the worst-case error in $H^s([0,1])$, that is, $\sup_{\|f\|_{H^s([0,1])} \leq 1} |P_n f - Pf|$. More concretely, we computed the weights $w_1, \ldots, w_n$ by formula (17) for Bayesian quadrature using the kernel k_r and then evaluated worst-case error (12) by computing the square root of (16) using the kernel k_s. In this way, one can evaluate the performance of kernel quadrature under various settings. For instance, the case $s < r$ is a situation where the true smoothness s is smaller than the assumed one r, the misspecified setting we have dealt in the paper.

6.2 Results

The simulation results are shown in Fig. 1 (*Uniform* design points) and Fig. 2 (*Non-uniform* design points). In the figures, we also report the exponents in the empirical rates of the fill distance $h_{X^n,\Omega}$, the separation radius q_{X^n}, and the absolute sum of weights $\sum_{i=1}^n |w_i|$ in the top of each subfigure; see the captions of Figs. 1 and 2 for details. Based on these, we can draw the following observations.

Optimal rates in the well-specified case In both Figs. 1 and 2, the black solid lines are the worst-case errors in the well-specified case $s = r$. The empirical convergence rates of these worst-case errors are very close to the optimal rates derived in Sect. 3 (see Corollary 1 and its remarks), confirming the theoretical results. Proposition 1 and Corollary 1 also show that the worst-case error in the well-specified case is determined by the fill distance and is independent of the separation radius. The simulation results are consistent with this, since for both Figs. 1 and 2 the fill distance decays essentially at the rate $O(n^{-1})$, while the separation radius decays quicker for Fig. 2 than for Fig. 1.

Adaptability to lesser smoothness Let us look at Fig. 1 for the misspecified case $s < r$, i.e., where the true smoothness s is smaller than the assumed one r. For every pair of $s < r$, the rates are very close to the optimal ones, showing that adaptation to the unknown lesser smoothness in fact occurs. This is consistent with Corollaries 3 and 4, which imply that adaptation occurs if the design points are quasi-uniform. Figure 2

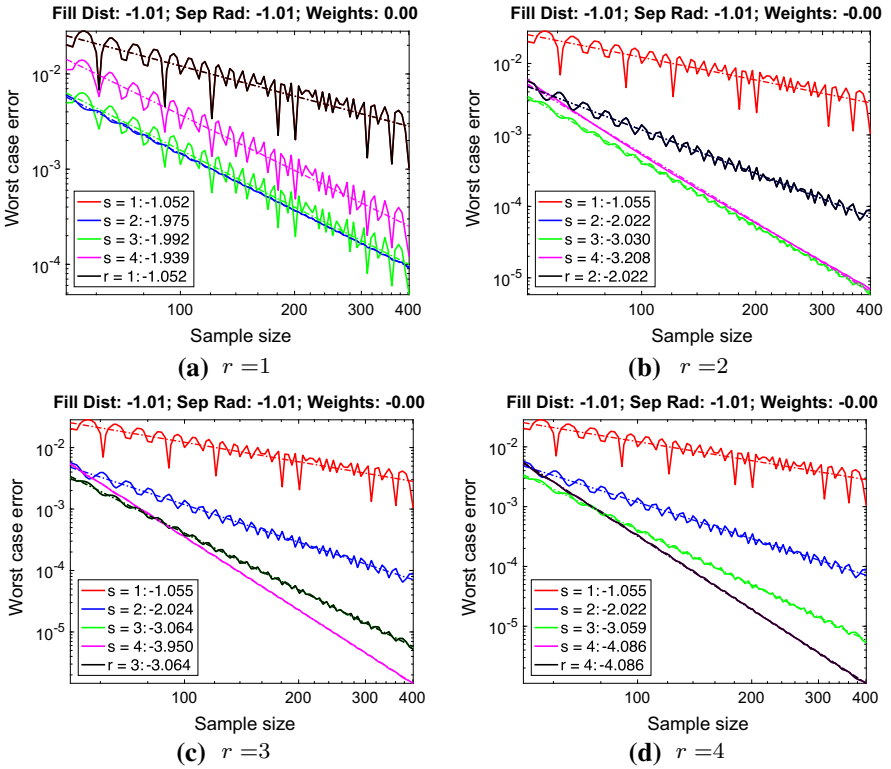

Fig. 1 Design points are *Uniform*, i.e., equally spaced grid points in [0, 1]; see Sect. 6.1 for details. The solid lines are the worst-case errors, and the dotted lines are the corresponding linear fits. The subfigures **a**–**d** are, respectively, the results for the weights computed using the kernel k_r with $r = 1, 2, 3, 4$. Black lines are the worst-case errors for the well-specified case $s = r$ (i.e., the worst-case error is evaluated in the same Sobolev space where the weights are obtained). Note that black lines overlap the corresponding lines for $s = r$ (e.g., in the subfigure **a** the red line for $s = 1$ does not appear since the black line completely overlaps it). In each legend, we report the exponents of the empirical rates of the worst-case errors. For instance, in the subfigure **d**, the worst-case error for $s = 1$ decays at the rate $O(n^{-1.055})$. On the top of each figure, the exponents in the empirical rates of the fill distance $h_{X^n,\Omega}$, the separation radius q_{X^n} and the absolute sum of weights $\sum_{i=1}^n |w_i|$ are shown. For instance, for the subfigure (**d**), we have $h_{X^n,\Omega} = O(n^{-1.01})$, $q_{X^n} = O(n^{-1.01})$ and $\sum_{i=1}^n |w_i| = O(n^{0.00})$ (Color figure online)

shows also some adaptability, but the rates for $s = 1$ with $r > s$ are slower than the optimal one. This will be discussed below, in a discussion on the effect of the separation radius.

Adaptability to greater smoothness While the case $s > r$ is not covered by our theoretical analysis, Figs. 1 and 2 show some adaptation to the greater smoothness. This phenomenon is also observed by Bach [4, Section 5], who showed (for quadrature weights obtained with *regularized* matrix inversion) that if $2r \geq s > r$, then the optimal rate is still attainable in an adaptive way. Bach [4, Section 6] verified this finding in experiments with quadrature weights *without* regularization. In our experiments, this phenomenon is observed for all cases of $2r \geq s > r$ expect for the case $r = 2$

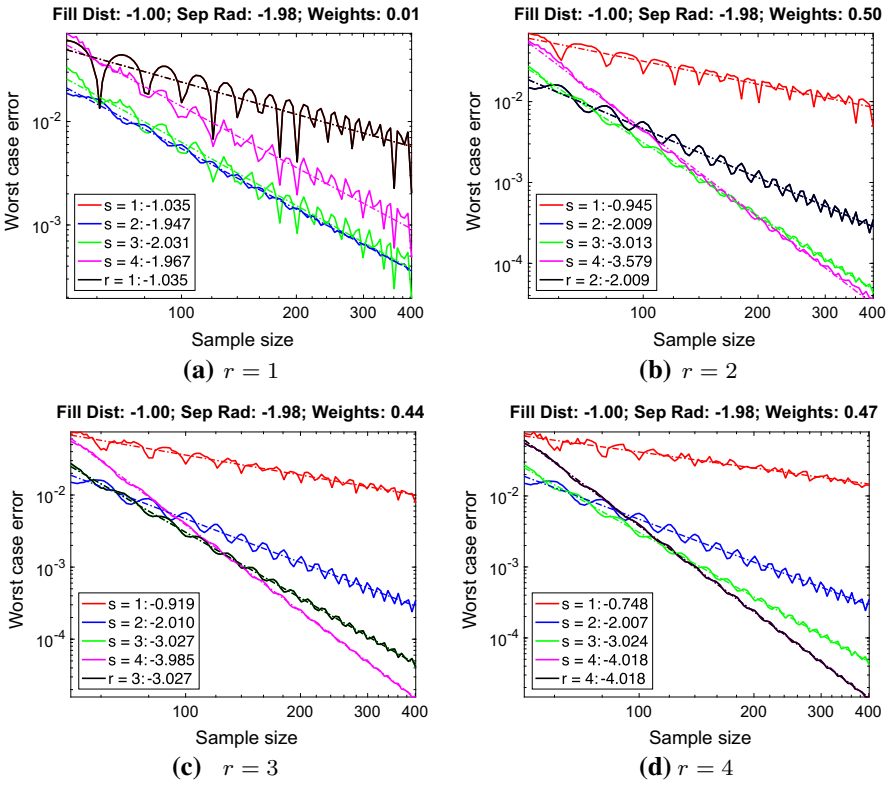

Fig. 2 Design points are *Non-uniform*, i.e., non-equally spaced points in [0, 1]; see Sect. 6.1 for details. The solid lines are the worst-case errors, and the dotted lines are the corresponding linear fits. The subfigures **a**–**d** are, respectively, the results for the weights computed using the kernel k_r with $r = 1, 2, 3, 4$. Black lines are the worst-case errors for the well-specified case $s = r$ (i.e., the worst-case error is evaluated in the same Sobolev space where the weights are obtained). Note that black lines overlap the corresponding lines for $s = r$ (e.g., in the subfigure **a** the red line for $s = 1$ does not appear since the black line completely overlaps it). In each legend, we report the exponents of the empirical rates of the worst-case errors. For instance, in the subfigure **d**, the worst-case error for $s = 1$ decays at the rate $O(n^{-0.748})$. On the top of each figure, the exponents in the empirical rates of the fill distance $h_{X^n,\Omega}$, the separation radius q_{X^n} and the absolute sum of weights $\sum_{i=1}^n |w_i|$ are shown. For instance, for the subfigure (**d**), we have $h_{X^n,\Omega} = O(n^{-1.00})$, $q_{X^n} = O(n^{-1.98})$ and $\sum_{i=1}^n |w_i| = O(n^{0.47})$ (Color figure online)

and $s = 4$ in both Figs. 1 and 2. Note, however, that in [4], design points are assumed to be *randomly* generated from a specific proposal distribution, so the results there are not directly applicable to deterministic quadrature rules.

The effect of the separation radius In Fig. 1, the rate for $s = 1$, that is, $O(n^{-1.052})$, remains essentially the same for different values of $r = 1, 2, 3, 4$. This rate is essentially the optimal rate for $s = 1$, thus showing the adaptability of Bayesian quadrature to the unknown lesser smoothness (for $r = 2, 3, 4$). On the other hand, in Fig. 2 on non-uniform design points, the rate for $s = 1$ becomes slower as r increases. That is, the rates are $O(n^{-1.035})$ for $r = 1$ (the well-specified case), $O(n^{-0.945})$ for $r = 2$, $O(n^{-0.919})$ for $r = 3$ and $O(n^{-0.748})$ for $r = 4$. This phenomenon may be attributed

to the fact that the separation radius of the design points for Fig. 2 decays faster than those for Fig. 1. Corollary 4 shows that the rates in the misspecified case $s < r$ become slower as the separation radius decays more quickly and/or as the gap $r - s$ (or the degree of misspecification) increases, and this is consistent with the simulation results.

The effect of the weights While the sum of absolute weights $\sum_{i=1}^{n} |w_i|$ remains constant in Fig. 1, this quantity increases in Fig. 2. In the notation of Corollary 2, $\sum_{i=1}^{n} |w_i| = O(n^c)$ with $c = 0$ for Fig. 1 while $c \approx 0.5$ for Fig. 2 with $r = 2, 3, 4$. Therefore, the observation given in the preceding paragraph is also consistent with Corollary 2, since it states that larger c makes the rates slower in the misspecified case. Note that the separation radius and the quantity $\sum_{i=1}^{n} |w_i|$ are intimately related in the case of Bayesian quadrature, since the weights are computed from the inverse of the kernel matrix as (17) and thus affected by the smallest eigenvalue of the kernel matrix, while this smallest eigenvalue strongly depends on the separation radius and the smoothness of the kernel; see, e.g., [52] [61, Section 12] and references therein.

7 Discussion

In this paper, we have discussed the convergence properties of kernel quadrature rules with deterministic design points in misspecified settings. In particular, we have focused on settings where quadrature weighted points are generated based on misspecified assumptions on the degree of smoothness, that is, the situation where the integrand is less smooth than assumed.

We have revealed conditions for quadrature rules under which adaptation to the unknown lesser degree of smoothness occurs. In particular, we have shown that a kernel quadrature rule is adaptive if the sum of absolute weights remains constant, or if the spacing between design points is not too small (as measured by the separation radius). Moreover, by focusing on Bayesian quadratures as working examples, we have shown that they can achieve minimax optimal rates of the unknown degree of smoothness, if the design points are quasi-uniform. We expect that this result provides a practical guide for developing kernel quadratures that are robust to the misspecification of the degree of smoothness; such robustness is important in modern applications of quadrature methods, such as numerical integration in sophisticated Bayesian models, since they typically involve complicated or black box integrands, and thus, misspecification is likely to happen.

There are several important topics to be investigated as part of future work.

Other RKHSs This paper has dealt with Sobolev spaces as RKHSs of kernel quadrature. However, there are many other important RKHSs of interest where similar investigation can be carried out. For instance, Gaussian RKHSs (i.e., the RKHSs of Gaussian kernels) have been widely used in the literature on Bayesian quadrature. Such an RKHS consists of functions with infinite degree of smoothness. This makes theoretical analysis challenging: Our analysis relies on the approximation theory developed by Narcowich and Ward [37], which only applies to the standard Sobolev spaces. Similarly, the theory of [37] is also not applicable to Sobolev spaces with dominating

mixed smoothness, which have been popular in the QMC literature. In order to analyze quadrature rules in these RKHSs, we therefore need to extend the approximation theory of [37] to such spaces. Overall, this is an important but challenging theoretical problem. (We also mention that relevant results are available in follow-up papers [38,39]. While these results do not directly provide the desired generalizations due to the same reasons mentioned above, these could still be potentially useful for our purpose.)

Sequential (adaptive) quadrature Another important direction is the analysis for kernel quadratures that sequentially select design points. Such methods are also called *adaptive*, since the selection of the next point X_{n+1} depends on the function values $f(X_1), \dots, f(X_n)$ of the already selected points $X_1, \dots, X_n$. Note that the adaptability here is different from that of the current paper where we used it in the context of adaptability of quadrature to unknown degree of smoothness. For instance, the WSABI algorithm by [25] is an example of adaptive Bayesian quadrature which is considered as state of the art for the application of Bayesian model evidence calculation. Such adaptive methods have been known to be able to outperform non-adaptive methods in the following case: The hypothesis space is imbalanced or non-convex (see, e.g., Section 1 of [41]). In the worst-case error, the hypothesis space is the unit ball in the RKHS $\mathcal{H}$, which is balanced and convex and so adaptation does *not* help. In fact, it is known that the optimal rate can be achieved without adaptation. However, if the hypothesis space is imbalanced (i.e., f being in the hypothesis space does *not* imply that $-f$ is in the hypothesis space), then adaptive methods may perform better. For instance, the WSABI algorithm focuses on *nonnegative* integrands, which means that the hypothesis is imbalanced, and thus, adaptive selection helps. Our analysis in this paper has focused on the worst-case error defined by the unit ball in an RKHS, which is balanced and convex. A future direction is thus to consider the setting of imbalanced or non-convex hypothesis spaces, such as the one consisting of nonnegative functions, which will enable us to analyze the convergence behavior of sequential or adaptive Bayesian quadrature in misspecified settings.

Random design points We have focused on *deterministic* quadrature rules in this paper. In the literature, however, the use of *random* design points has also been popular. For instance, the design points of Bayesian quadrature might be i.i.d. with a certain proposal distribution or generated as an MCMC sequence. Likewise, QMC methods usually apply randomization to deterministic design points. Our forthcoming paper will deal with such situations and provide more general results than the current paper.

Acknowledgements The open access funding is provided by the Max Planck Society. We would like to express our gratitude to the editor and anonymous referees for their constructive feedback that greatly improved the paper. Most of this work has been done when MK was working at the Institute of Statistical Mathematics, Tokyo.

Appendix A: Key Results of Narcowich and Ward [37]

Here we review some key results from [37], which are needed in the proofs for our results. One reason for including this is that a certain assumption about a function of interest, that is, its integrability, is lacking in the results of [37]; see Remark A.1 for details. Therefore, for the sake of completeness (as well as for the ease of the reader), we provide restatements of those results.

For $\sigma > 0$, below we denote by $\mathcal{B}_\sigma$ a subset of $L_2(\mathbb{R}^d)$ such that each $f \in \mathcal{B}_\sigma$ has a spectral density whose support is contained in the (closed) ball $B(0,\sigma)$ with radius σ, i.e.,

$$\mathcal{B}_\sigma := \left\{ f \in L_2(\mathbb{R}^d) : \ \mathrm{supp}(\hat{f}) \subset B(0,\sigma) \right\}.$$

This is a Paley–Weiner class of band-limited functions. Thus, the functions in $\mathcal{B}_\sigma$ are analytic (and thus, they are continuous) and vanish at infinity. Therefore, $\mathcal{B}_\sigma \subset L_2(\mathbb{R}^d) \cap C_0(\mathbb{R}^d)$.

The following theorem is a restatement of Theorem 3.5 of [37].

Theorem A.1 *Let $X^n := \{X_1, \ldots, X_n\} \subset \mathbb{R}^d$ be n distinct points with separation radius $q_{X^n} := \frac{1}{2} \min_{i \neq j} \|X_i - X_j\|$, such that $\mathrm{diam}(X^n) := \max_{i,j} \|X_i - X_j\| \leq 1$. Let $\sigma > 0$ be a constant such that*

$$\sigma \geq \sigma_0 := \frac{24}{q_{X^n}} \left\{ \frac{\sqrt{\pi}}{3} \Gamma\left(\frac{d+2}{2} \right) \right\}^{\frac{2}{d+2}}.$$

Then, for any $f \in C_0(\mathbb{R}^d) \cap L_2(\mathbb{R}^d)$, there exists $f_\sigma \in \mathcal{B}_\sigma$ that satisfies

$$f(X_i) = f_\sigma(X_i), \quad i = 1, \ldots, n,$$

and

$$\begin{aligned} &\max\left(\|f - f_\sigma\|_{C_0(\mathbb{R}^d)}, \|f - f_\sigma\|_{L_2(\mathbb{R}^d)} \right) \\ &\quad \leq C_d \inf_{g \in \mathcal{B}_\sigma} \max\left(\|f - g\|_{C_0(\mathbb{R}^d)}, \|f - g\|_{L_2(\mathbb{R}^d)} \right) \end{aligned}$$

with $C_d := 5 + 2^{d+3}$.

In the above theorem, f_σ is an interpolant of f on X^n. Thus, the theorem guarantees that such a f_σ can be taken as a band-limited function with a sufficiently large band-length σ. More precisely, the lower bound σ_0 for σ is proportional to the reciprocal of the separation radius q_{X^n}. This means that the band-length σ should increase as the minimum distance between distinct design points decreases.

The following proposition is a restatement of Proposition 3.7 of [37], which establishes an upper bound on the L_1-error for the approximate function defined in (B.10)—see "Appendix B.2."

Proposition A.1 *Let $s \in \mathbb{N}$ and $\alpha \in \mathbb{N}_0^d$ be a multi-index such that $|\alpha| < s$. Suppose $f \in C_0^s(\mathbb{R}^d) \cap H^s(\mathbb{R}^d) \cap L_1(\mathbb{R}^d)$ and g_σ is the approximate function defined in (B.10). Then, for any $\sigma > 0$,*

$$\|\partial^\alpha f - \partial^\alpha g_\sigma\|_{L_\infty(\mathbb{R}^d)} \leq C_{s-|\alpha|}\sigma^{|\alpha|-s}\|f\|_{C_0^s(\mathbb{R}^d)},$$

where $C_{k-|\alpha|} > 0$ is a constant depending only on the value of $k-|\alpha|$ and the function ψ of Lemma B.2.

The following theorem, which is Theorem 3.10 in [37], provides an upper bound on the approximation error of the interpolant f_σ.

Theorem A.2 *Let $s \in \mathbb{N}$ and $\alpha \in \mathbb{N}_0^d$ be a multi-index such that $|\alpha| < s$. Suppose $f \in C_0^s(\mathbb{R}^d) \cap H^s(\mathbb{R}^d) \cap L_1(\mathbb{R}^d)$, f_σ is the interpolant from Theorem A.1 with $\sigma > 0$ and $X^n := \{X_1, \ldots, X_n\}$ satisfies the conditions in Theorem A.1. Then, there is a constant $C_{|\alpha|,s,d}$ that depends only on $|\alpha|$, s and d such that*

$$\left\|\partial^\alpha f - \partial^\alpha f_\sigma\right\|_{L_\infty(\mathbb{R}^d)} \leq C_{|\alpha|,s,d}\sigma^{|\alpha|-s}\max\left(\|f\|_{C_0^s(\mathbb{R}^d)}, \|f\|_{H^s(\mathbb{R}^d)}\right).$$

The following proposition, which is Proposition 3.11 in [37], provides an upper bound on a Sobolev norm of the interpolant f_σ.

Proposition A.2 *Let $s \in \mathbb{N}$. Suppose $f \in C_0^s(\mathbb{R}^d) \cap H^s(\mathbb{R}^d) \cap L_1(\mathbb{R}^d)$, f_σ is the interpolant from Theorem A.1 with $\sigma > 0$ and $X^n := \{X_1, \ldots, X_n\}$ satisfies the conditions in Theorem A.1. Then, there is a constant $C_{s,d}$ that depends only on s and d such that*

$$\|f_\sigma\|_{H^s(\mathbb{R}^d)} \leq C_{s,d}\max\left(\|f\|_{C_0^s(\mathbb{R}^d)}, \|f\|_{H^s(\mathbb{R}^d)}\right).$$

Remark A.1 We have the following comments on Propositions A.1, A.2 and Theorem A.2.

- In the original statement of Proposition 3.7 in [37], the assumption $f \in L_1(\mathbb{R}^d)$ is missing. However, since this assumption is required for the function g_σ to be well defined (see Lemma B.5), we have included it in Proposition A.1. Since Theorem 3.10 and Proposition 3.11 of [37] depend on Proposition 3.7, we have included the assumption $f \in L_1(\mathbb{R}^d)$ in Theorem A.2 and Proposition A.2.
- In the original statement of Proposition 3.11 in [37], the condition $\sigma \geq 1$ is required. This condition is implicitly satisfied by σ in Proposition A.2 as the condition on σ in Theorem A.1 implies $\sigma \geq 1$, which can be seen from the fact that $q_{X^n} \leq 1/2$ (follows from the assumption $\mathrm{diam}(X^n) \leq 1$) and the definition of the lower bound σ_0 of σ.

Appendix A.1: The Sobolev Norm of the Interpolant f_σ

Here we provide an upper bound on the Sobolev (RKHS) norm of the interpolant f_σ in Theorem A.1. The result essentially follows from an argument in p.298 of [37], but we prove it for completeness.

Lemma A.1 *Let $r \in \mathbb{R}$, $r > d/2$ and $s \in \mathbb{N}$, $r \geq s$. Let k_r be a kernel on $\mathbb{R}^d$ such that $k_r(x, y) := \Phi(x - y)$, where $\Phi : \mathbb{R}^d \to \mathbb{R}$ satisfies*

$$C_1(1 + \|\xi\|^2)^{-r} \leq \hat{\Phi}(\xi), \quad \xi \in \mathbb{R}^d$$

for some constant $C_1 > 0$ independent of ξ. Suppose $f \in C_0^s(\mathbb{R}^d) \cap H^s(\mathbb{R}^d) \cap L_1(\mathbb{R}^d)$, f_σ is the interpolant from Theorem A.1 with $\sigma > 0$ and $X^n := \{X_1, \ldots, X_n\}$ satisfies the conditions in Theorem A.1. Then, we have

$$\|f_\sigma\|_{\mathcal{H}_{k_r}} \leq C_{s,d,k_r} \sigma^{r-s} \max\left(\|f\|_{C_0^s(\mathbb{R}^d)}, \|f\|_{H^s(\mathbb{R}^d)}\right), \tag{A.1}$$

where C_{s,d,k_r} is a constant only depending on r, s, d, and k_r (note that the dependency on the kernel k_r is via the constant C_1).

Proof As in Remark A.1, we have $\sigma \geq 1$. We then have

$$\begin{aligned}
\|f_\sigma\|_{\mathcal{H}_{k_r}}^2 &= \int_{\|\xi\| \leq \sigma} |\hat{f}_\sigma(\xi)|^2 \hat{\Phi}(\xi)^{-1} d\xi \quad (\because f \in \mathcal{B}_\sigma) \\
&\leq C_1^{-1} \int_{\|\xi\| \leq \sigma} |\hat{f}_\sigma(\xi)|^2 (1 + \|\xi\|^2)^r d\xi \\
&\leq C_1^{-1} (1 + \sigma^2)^{r-s} \int_{\|\xi\| \leq \sigma} |\hat{f}_\sigma(\xi)|^2 (1 + \|\xi\|^2)^s d\xi \quad (\because r - s \geq 0) \\
&\leq C_1^{-1} (1 + \sigma^2)^{r-s} \|f_\sigma\|_{H^s(\mathbb{R}^d)}^2 \leq C_1^{-1} 2^{r-s} \sigma^{2(r-s)} \|f_\sigma\|_{H^s(\mathbb{R}^d)}^2 \quad (\because \sigma \geq 1).
\end{aligned}$$

Therefore, by using Proposition A.2, it follows that

$$\begin{aligned}
\|f_\sigma\|_{\mathcal{H}_{k_r}} &\leq C_1^{-1/2} 2^{(r-s)/2} \sigma^{r-s} \|f_\sigma\|_{H^s(\mathbb{R}^d)} \\
&\leq C_1^{-1/2} 2^{(r-s)/2} \sigma^{r-s} C_{s,d} \max\left(\|f\|_{C_0^s(\mathbb{R}^d)}, \|f\|_{H^s(\mathbb{R}^d)}\right),
\end{aligned}$$

where $C_{s,d}$ is a constant only depending on s and d. The proof completes by setting $C_{s,d,k_r} := C_1^{-1/2} 2^{(r-s)/2} C_{s,d}$. □

Appendix B: Approximation in Sobolev Spaces

Appendix B.1: Fundamental Lemma

In the proof of Theorem 1, we used Proposition 3.7 of [37], which assumes the existence of a function $\psi : \mathbb{R}^d \to \mathbb{R}$ satisfying the properties in Lemma B.2. Since the existence

of this function is not proved in [37], we will first prove it for completeness. Lemma B.2 is a variant of Lemma 1.1 of [19], from which we borrowed the proof idea.

Lemma B.2 *Let $s \in \mathbb{N}$. Then, there exists a function $\psi : \mathbb{R}^d \to \mathbb{R}$ satisfying the following properties:*

(a) *ψ is radial;*
(b) *ψ is a Schwartz function;*
(c) *$\mathrm{supp}(\hat{\psi}) \subset B(0,1)$;*
(d) *$\int_{\mathbb{R}^d} x^\beta \psi(x) \mathrm{d}x = 0$ for every multi-index β satisfying $|\beta| := \sum_{i=1}^d \beta_i \leq s$, where $x^\beta := \prod_{i=1}^d x_i^{\beta_i}$;*
(e) *ψ satisfies*

$$\int_0^\infty |\hat{\psi}(t\xi)|^2 \frac{\mathrm{d}t}{t} = 1, \quad \forall \xi \in \mathbb{R}^d \backslash \{0\}. \tag{B.2}$$

Proof Define a function $u \in L_1(\mathbb{R}^d)$ as the inverse Fourier transform of a function $\hat{u} \in L_1(\mathbb{R}^d)$ defined by $\hat{u}(\xi) := \exp\left(-1/(1-\|\xi\|^2)\right)$ if $\|\xi\| < 1$ and $\hat{u}(\xi) = 0$ otherwise. Then, $\hat{u}$ is radial, Schwartz, and satisfies $\mathrm{supp}(\hat{u}) \subset B(0,1)$ [1, Sec. 2.28]. Also note that u is real valued, since $\hat{u}$ is symmetric.

Let $m \in \mathbb{N}$ satisfy $m > s/2$. Define a function $h : \mathbb{R}^d \to \mathbb{R}$ by $h := \Delta^m u$, where Δ denotes the Laplacian defined by $\Delta f := \sum_{i=1}^d \frac{\partial^2 f}{\partial x_i^2}$. Note that we have (see, e.g., p.117 of [57])

$$\hat{h}(\xi) = C_m \|\xi\|^{2m} \hat{u}(\xi), \tag{B.3}$$

where C_m is a constant depending only on m. From this expression, it follows that $\hat{h}$ is radial and Schwartz (and so is h) and that $\mathrm{supp}(\hat{h}) \subset B(0,1)$. Thus, the function h satisfies the required properties (a) (b) and (c). Later we will define the function ψ in the assertion based on h.

We next show that h satisfies the property (d). Let $\beta \in \mathbb{N}_0^d$ be any multi-index satisfying $|\beta| \leq s$, and let $p_\beta(x) := x^\beta$. It follows that $p_\beta h$ is Schwartz, and thus, $p_\beta h \in L_1(\mathbb{R}^d)$. Then, we have

$$\int x^\beta h(x) \mathrm{d}x = \widehat{(p_\beta h)}(0), \tag{B.4}$$

which follows from $p_\beta h \in L_1(\mathbb{R}^d)$ and from the definition of Fourier transform. Note that we have $\widehat{p_\beta h}(\xi) = i^{|\beta|} \partial^\beta \hat{h}(\xi)$, which can be expanded as

$$\partial^\beta \hat{h}(\xi) \overset{(B.3)}{=} \partial^\beta \left[C_m \|\xi\|^{2m} \hat{u}(\xi) \right] = C_m \sum_{\gamma \in \mathbb{N}_0^d : \gamma \leq \beta} \binom{\beta}{\gamma} \partial^\gamma \left[\|\xi\|^{2m} \right] \partial^\beta \left[\hat{u}(\xi) \right], \tag{B.5}$$

where $\gamma \leq \beta$ is defined by that $\gamma_i \leq \beta_i$ for all $i = 1, \ldots, d$, and $\binom{\beta}{\gamma} := \frac{\prod_{i=1}^d \beta_i!}{\prod_{i=1}^d \gamma_i!}$. Using the multinomial theorem, the mixed partial derivative $\partial^\gamma \left[\|\xi\|^{2m} \right]$ in the above equation can be further expanded as

$$\partial^\gamma \left[\|\xi\|^{2m} \right] = \sum_{\alpha \in \mathbb{N}_0^d : |\alpha| = m} \frac{m!}{\prod_{i=1}^d \alpha_i!} \prod_{i=1}^d \frac{d^{\gamma_i}}{d\xi_i^{\gamma_i}} \left[\xi_i^{2\alpha_i} \right]. \tag{B.6}$$

From this, it follows that $\partial^\gamma \left[\|\xi\|^{2m} \right]\big|_{\xi=0} = 0$, and thus, (B.5) gives that $\partial^\beta \hat{h}(0) = 0$. Therefore, from (B.4) and $\widehat{p_\beta h}(\xi) = i^{|\beta|} \partial^\beta \hat{h}(\xi)$, it holds that $\int_{\mathbb{R}} x^\beta h(x) \mathrm{d}x = 0$, which is the property (d).

We next show that $\int_0^\infty |\hat{h}(t\xi)|^2 \frac{\mathrm{d}t}{t} < \infty$ for all $\xi \in \mathbb{R}^d \backslash \{0\}$. Since $\hat{h}$ is bounded and $\mathrm{supp}(\hat{h}) \subset B(0,1)$, we have $\int_1^\infty |\hat{h}(t\xi)|^2 \frac{\mathrm{d}t}{t} < \infty$. Also, since $|\hat{h}(t\xi)| = O(t^{2m})$ as $t \to +0$ (which follows from $\hat{h}(t\xi) = (-1)^m \|t\xi\|^{2m} \hat{u}(t\xi)$ with $\hat{u}$ being bounded), we have $\int_0^1 |\hat{h}(t\xi)|^2 \frac{\mathrm{d}t}{t} < \infty$. Therefore, $\int_0^\infty |\hat{h}(t\xi)|^2 \frac{\mathrm{d}t}{t} < \infty$.

Note that since $\hat{h}$ is radial, $\int_0^\infty |\hat{h}(t\xi)|^2 \frac{\mathrm{d}t}{t}$ only depends on the norm $\|\xi\|$. Furthermore, $\int_0^\infty |\hat{h}(t\xi)|^2 \frac{\mathrm{d}t}{t}$ remains the same for different values of the norm $\|\xi\| > 0$ due to the property of the Haar measure $\mathrm{d}t/t$. In other words, there is a constant $0 < C < \infty$ satisfying $\int_0^\infty |\hat{h}(t\xi)|^2 \frac{\mathrm{d}t}{t} = C$ for all $\xi \in \mathbb{R}^d \backslash \{0\}$. The proof is completed by defining ψ in the assertion as $\psi(x) := C^{-1/2} h(x)$. □

Notation 1 Note that ψ being radial implies that $\hat{\psi}$ is radial, so $\hat{\psi}(t\xi)$ in (B.2) depends on ξ only through its norm $\|\xi\|$. Therefore, we may henceforth use the notation

$$\hat{\psi}(t\|\xi\|) \tag{B.7}$$

to denote $\hat{\psi}(t\xi)$, to emphasize its dependence on the norm. Similarly, we use the notation $\hat{\psi}(t)$ to imply $\hat{\psi}(t\xi)$ for some (and any) $\xi \in \mathbb{R}^d$ with $\|\xi\| = 1$.

Appendix B.2: Approximation Via Calderón's Formula

The following result is known as Calderón's formula [19, Theorem 1.2] and will be used in defining an approximate function (B.10). We use below the notation $f * g$ for any functions $f : \mathbb{R}^d \to \mathbb{R}$ and $g : \mathbb{R}^d \to \mathbb{R}$ to denote their convolution: $(f * g)(x) := \int f(x-y) g(y) \mathrm{d}y$.

Theorem B.3 (Calderón's formula) *Let $\psi \in L_1$ be a radial function satisfying (B.2), and for $t > 0$ define*

$$\psi_t(x) := \frac{1}{t^d} \psi(x/t), \quad x \in \mathbb{R}^d. \tag{B.8}$$

Then, for any $f \in L_2$, we have

$$f(x) = \int_0^\infty (\psi_t * \psi_t * f)(x) \frac{\mathrm{d}t}{t}, \quad x \in \mathbb{R}^d, \tag{B.9}$$

where the improper integral in (B.9) is to be interpreted in the following L_2 sense: if $0 < \varepsilon < \delta < \infty$ and $f_{\varepsilon,\delta}(x) := \int_\varepsilon^\delta (\psi_t * \psi_t * f)(x) \frac{\mathrm{d}t}{t}$, then $\|f - f_{\varepsilon,\delta}\|_{L_2} \to 0$ as $\varepsilon \to +0$ and $\delta \to \infty$ independently.

Note that it is easy to verify from (B.8) that $\|\psi\|_{L_1} = \|\psi_t\|_{L_1}$ holds for all $t > 0$. Let ψ be the function in Lemma B.2. Following Section 3.2 of [37], we consider the following approximation of f based on Calderón's formula (B.9):

$$g_\sigma(x) := \int_{1/\sigma}^{\infty} (\psi_t * \psi_t * f)(x) \, \frac{\mathrm{d}t}{t}. \tag{B.10}$$

The integral in (B.10) is also improper and should be interpreted as follows. Let $\delta > 1/\sigma$ and define

$$g_{\sigma,\delta} := \int_{1/\sigma}^{\delta} (\psi_t * \psi_t * f)(x) \, \frac{\mathrm{d}t}{t}. \tag{B.11}$$

Then, g_σ in (B.10) is defined to be a function in L_2 such that $\lim_{\delta\to\infty} \|g_\sigma - g_{\sigma,\delta}\|_{L_2} = 0$. Such g_σ exists (as a limit of $g_{\sigma,\delta}$), as shown in Lemma B.5. Since there is no proof of this result in [37], we provide a proof for the sake of completeness. To this end, we first need the following lemma.

Lemma B.3 *Let $g_{\sigma,\delta}$ be defined as in (B.11) with $\delta > 1/\sigma$. For all $1 \le p \le \infty$, if $f \in L_p$, then $g_{\sigma,\delta} \in L_p$.*

Proof For $1 \le p \le \infty$, we have

$$\begin{aligned}
\|g_{\sigma,\delta}\|_{L_p} &= \left\| \int_{1/\sigma}^{\delta} \psi_t * \psi_t * f \, \frac{\mathrm{d}t}{t} \right\|_{L_p} \\
&\le \int_{1/\sigma}^{\delta} \|\psi_t * \psi_t * f\|_{L_p} \, \frac{dt}{t} \quad (\because \text{Minkowski's inequality}) \\
&\le \int_{1/\sigma}^{\delta} \|\psi_t\|_{L_1}^2 \|f\|_{L_p} \frac{\mathrm{d}t}{t} \quad (\because \text{Young's inequality}) \\
&= \int_{1/\sigma}^{\delta} \|\psi\|_{L_1}^2 \|f\|_{L_p} \frac{\mathrm{d}t}{t} = \|\psi\|_{L_1}^2 \|f\|_{L_p} (\log(\delta) - \log(1/\sigma)) < +\infty,
\end{aligned}$$

where in the last line we used the assumption $f \in L_p$ and the fact $\psi \in L_1$, which is a consequence of ψ being a Schwartz function (see Lemma B.2). □

Lemma B.4 *Assume $f \in L_1$, and let $g_{\sigma,\delta}$ be defined as in (B.11) with $\delta > 1/\sigma$. Then, the Fourier transform of $g_{\sigma,\delta}$ is given by*

$$\hat{g}_{\sigma,\delta}(\xi) = \begin{cases} \hat{f}(\xi) \int_{\|\xi\|/\sigma}^{\min(1,\|\xi\|\delta)} (\hat{\psi}(t))^2 \frac{\mathrm{d}t}{t}, & \text{if } \|\xi\| < \sigma \\ 0, & \text{otherwise} \end{cases}.$$

Proof We have

$$\hat{g}_{\sigma,\delta}(\xi) = \int \int_{1/\sigma}^{\delta} (\psi_t * \psi_t * f)(x) \frac{\mathrm{d}t}{t} e^{-i\xi^T x} \mathrm{d}x$$

$$= \int_{1/\sigma}^{\delta} \int (\psi_t * \psi_t * f)(x) e^{-i\xi^T x} \mathrm{d}x \frac{\mathrm{d}t}{t} \quad (\because \text{Fubini's theorem})$$
$$= \hat{f}(\xi) \int_{1/\sigma}^{\delta} (\hat{\psi}_t(\xi))^2 \frac{\mathrm{d}t}{t} = \hat{f}(\xi) \int_{1/\sigma}^{\delta} (\hat{\psi}(t\xi))^2 \frac{\mathrm{d}t}{t}.$$

In the above derivation, Fubini's theorem is applicable since $\psi_t * \psi_t * f \in L_1$ (which follows from $\psi \in L_1$, $f \in L_1$ and Young's inequality; see the proof of Lemma B.3).

Recall that $\hat{\psi}$ is radial, so that the value of $\hat{\psi}(t\xi)$ only depends on the norm of its argument $\|t\xi\| = t\|\xi\|$. By a change of variables $\tau := t\|\xi\|$, and recalling the notation $\hat{\psi}(t\|\xi\|) := \hat{\psi}(t\xi)$, it holds that

$$\int_{1/\sigma}^{\delta} (\hat{\psi}(t\|\xi\|))^2 \frac{\mathrm{d}t}{t} = \int_{\|\xi\|/\sigma}^{\|\xi\|\delta} (\hat{\psi}(\tau))^2 \frac{d\tau}{\tau}$$
$$= \begin{cases} \int_{\|\xi\|/\sigma}^{\min(1, \|\xi\|\delta)} (\hat{\psi}(\tau))^2 \frac{d\tau}{\tau}, & \text{if } \|\xi\| < \sigma \\ 0, \quad \text{otherwise} \end{cases}, \tag{B.12}$$

where the last line follows from the property $\mathrm{supp}(\psi) \subset B(0, 1)$. The proof is completed by combining this and the above expression of $\hat{g}_{\sigma,\delta}(\xi)$. □

We are now ready to show that the improper integral in (B.10) is well defined as a limit of $g_{\sigma,\delta}$ in L_2: The following lemma characterizes this limiting function in L_2 in terms of its Fourier transform.

Lemma B.5 *Assume $f \in L_1 \cap L_2$. Let $g_{\sigma,\delta}$ be defined as in (B.11) with $\delta > 1/\sigma$, and $g_\sigma \in L_2$ be the inverse Fourier transform of $\hat{g}_\sigma \in L_2$ defined by*

$$\hat{g}_\sigma(\xi) = \begin{cases} \hat{f}(\xi) \int_{\|\xi\|/\sigma}^{1} (\hat{\psi}(t))^2 \frac{\mathrm{d}t}{t}, & \text{if } \|\xi\| < \sigma \\ 0, \quad \text{otherwise} \end{cases}.$$

Then, we have $\lim_{\delta\to\infty} \|g_\sigma - g_{\sigma,\delta}\|_{L_2} = 0$.

Proof First note that by Lemma B.3, the assumption $f \in L_1 \cap L_2$ implies $g_{\sigma,\delta} \in L_1 \cap L_2$, so we have $\hat{g}_{\sigma,\delta} \in L_1 \cap L_2$. Below we will show $\lim_{\delta\to\infty} \|\hat{g}_\sigma - \hat{g}_{\sigma,\delta}\|_{L_2} = 0$, from which the assertion follows because of the Fourier transform being an isometry from L_2 to L_2. By Lemma B.4 (which is applicable as $f \in L_1$) we have

$$\|\hat{g}_\sigma - \hat{g}_{\sigma,\delta}\|_{L_2}^2 = \int_{\|\xi\|<\sigma} |\hat{f}(\xi)|^2 \left| \int_{\min(1, \|\xi\|\delta)}^{1} (\hat{\psi}(t))^2 \frac{\mathrm{d}t}{t} \right|^2 d\xi.$$

Therefore,

$$\lim_{\delta\to\infty} \|\hat{g}_\sigma - \hat{g}_{\sigma,\delta}\|_{L_2}^2 = \int_{\|\xi\|<\sigma} |\hat{f}(\xi)|^2 \lim_{\delta\to\infty} \left| \int_{\min(1, \|\xi\|\delta)}^{1} (\hat{\psi}(t))^2 \frac{\mathrm{d}t}{t} \right|^2 d\xi$$

$$= \int_{\|\xi\|<\sigma} |\hat{f}(\xi)|^2 \left| \int_1^1 (\hat{\psi}(t))^2 \frac{\mathrm{d}t}{t} \right|^2 d\xi = 0, \tag{B.13}$$

where (B.13) follows from the dominated convergence theorem (which follows from $f \in L_2$). □

Appendix B.3: The Sobolev Norm of the Approximate Function

In the main body of the paper, we use the following lemma, which is not provided in [37].

Lemma B.6 *Let $r, s \in \mathbb{R}$, $r, s > 0$ such that $r \geq s$ and let $\sigma > 0$ be a constant. If $f \in H^s(\mathbb{R}^d) \cap L_1(\mathbb{R}^d)$, the function g_σ defined in (B.10) satisfies*

$$\|g_\sigma\|_{H^r} \leq (1+\sigma^2)^{\frac{r-s}{2}} \|f\|_{H^s},$$

where $C > 0$ is a constant independent of f and σ.

Proof Note that from (B.2), if $\|\xi\| < \sigma$, we have $\int_{\|\xi\|/\sigma}^1 |\hat{\psi}(t)|^2 \frac{\mathrm{d}t}{t} \leq \int_0^1 |\hat{\psi}(t)|^2 \frac{\mathrm{d}t}{t} \leq 1$. Therefore, by Lemma B.5 we have

$$\begin{aligned}
\|g_\sigma\|_{H^r}^2 &= \int_{B(0,\sigma)} (1+\|\xi\|^2)^r |\widehat{g_\sigma}(\xi)|^2 d\xi \\
&\leq \int_{B(0,\sigma)} (1+\|\xi\|^2)^r |\hat{f}(\xi)|^2 d\xi \\
&\leq (1+\sigma^2)^{r-s} \int_{B(0,\sigma)} (1+\|\xi\|^2)^s |\hat{f}(\xi)|^2 d\xi \\
&\leq (1+\sigma^2)^{r-s} \|f\|_{H^s}^2,
\end{aligned}$$

yielding the result. □

References

1. Adams, R.A., Fournier, J.J.F.: Sobolev Spaces, 2nd edn. Academic Press, New York (2003)
2. Aronszajn, N.: Theory of reproducing kernels. Transactions of the American Mathematical Society, 68(3) pp. 337–404 (1950)
3. Avron, H., Sindhwani, V., Yang, J., Mahoney, M.W.: Quasi-Monte Carlo feature maps for shift-invariant kernels. Journal of Machine Learning Research **17**(120), 1–38 (2016)
4. Bach, F.: On the equivalence between kernel quadrature rules and random feature expansions. Journal of Machine Learning Research **18**(19), 1–38 (2017)
5. Bach, F., Lacoste-Julien, S., Obozinski, G.: On the equivalence between herding and conditional gradient algorithms. In: J. Langford, J. Pineau (eds.) Proceedings of the 29th International Conference on Machine Learning (ICML2012), pp. 1359–1366. Omnipress (2012)
6. Brenner, S.C., Scott, L.R.: The Mathematical Theory of Finite Element Methods, 3rd edn. Springer (2008)

7. Briol, F.X., Oates, C.J., Cockayne, J., Chen, W.Y., Girolami, M.: On the sampling problem for kernel quadrature. In: D. Precup, Y.W. Teh (eds.) Proceedings of the 34th International Conference on Machine Learning, *Proceedings of Machine Learning Research*, vol. 70, pp. 586–595. PMLR (2017)
8. Briol, F.X., Oates, C.J., Girolami, M., Osborne, M.A.: Frank-Wolfe Bayesian quadrature: Probabilistic integration with theoretical guarantees. In: C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, R. Garnett (eds.) Advances in Neural Information Processing Systems 28, pp. 1162–1170. Curran Associates, Inc. (2015)
9. Briol, F.X., Oates, C.J., Girolami, M., Osborne, M.A., Sejdinovic, D.: Probabilistic integration: A role in statistical computation? Statistical Science (2018). To appear
10. Chen, W.Y., Mackey, L., Gorham, J., Briol, F.X., Oates, C.: Stein points. In: J. Dy, A. Krause (eds.) Proceedings of the 35th International Conference on Machine Learning, *Proceedings of Machine Learning Research*, vol. 80, pp. 844–853. PMLR (2018)
11. Chen, Y., Welling, M., Smola, A.: Supersamples from kernel-herding. In: P. Grünwald, P. Spirtes (eds.) Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI 2010), pp. 109–116. AUAI Press (2010)
12. Cucker, F., Zhou, D.X.: Learning Theory: An approximation theory view point. Cambridge University Press (2007)
13. Diaconis, P.: Bayesian numerical analysis. Statistical decision theory and related topics IV **1**, 163–175 (1988)
14. Dick, J.: Explicit constructions of quasi-Monte Carlo rules for the numerical integration of high-dimensional periodic functions. SIAM Journal on Numerical Analysis **45**, 2141–2176 (2007)
15. Dick, J.: Walsh spaces containing smooth functions and quasi–Monte Carlo rules of arbitrary high order. SIAM Journal on Numerical Analysis **46**(3), 1519–1553 (2008)
16. Dick, J.: Higher order scrambled digital nets achieve the optimal rate of the root mean square error for smooth integrands. The Annals of Statistics **39**(3), 1372–1398 (2011)
17. Dick, J., Kuo, F.Y., Sloan, I.H.: High dimensional numerical integration - the Quasi-Monte Carlo way. Acta Numerica **22** 133-288 (2018)
18. Dick, J., Nuyens, D., Pillichshammer, F.: Lattice rules for nonperiodic smooth integrands. Numerische Mathematik **126**(2), 259–291 (2014)
19. Frazier, M., Jawerth, B., Weiss, G.L.: Littlewood-Paley Theory and the Study of Function Spaces. American Mathematical Society (1991)
20. Fuselier, E., Hangelbroek, T., Narcowich, F.J., Ward, J.D., Wright, G.B.: Kernel based quadrature on spheres and other homogeneous spaces. Numerische Mathematik **127**(1), 57–92 (2014)
21. Gerber, M., Chopin, N.: Sequential quasi Monte Carlo. Journal of the Royal Statistical Society. Series B. Statistical Methodology **77**(3), 509-579 (2015)
22. Ghahramani, Z., Rasmussen, C.E.: Bayesian monte carlo. In: S. Becker, S. Thrun, K. Obermayer (eds.) Advances in Neural Information Processing Systems 15, pp. 505–512. MIT Press (2003)
23. Goda, T., Dick, J.: Construction of interlaced scrambled polynomial lattice rules of arbitrary high order. Foundations of Computational Mathematics **15**(5), 1245–1278 (2015)
24. Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., Smola, A.: A kernel two-sample test. Jounal of Machine Learning Research **13**, 723–773 (2012)
25. Gunter, T., Osborne, M.A., Garnett, R., Hennig, P., Roberts, S.J.: Sampling for inference in probabilistic models with fast Bayesian quadrature. In: Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, K.Q. Weinberger (eds.) Advances in Neural Information Processing Systems 27, pp. 2789–2797. Curran Associates, Inc. (2014)
26. Hickernell, F.J.: A generalized discrepancy and quadrature error bound. Mathematics of Computation **67**(221), 299–322 (1998)
27. Huszár, F., Duvenaud, D.: Optimally-weighted herding is Bayesian quadrature. In: N. de Freitas, K. Murphy (eds.) Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI2012), pp. 377–385. AUAI Press (2012)
28. Kanagawa, M., Nishiyama, Y., Gretton, A., Fukumizu, K.: Filtering with state-observation examples via kernel monte carlo filter. Neural Computation **28**(2), 382–444 (2016)
29. Kanagawa, M., Sriperumbudur, B.K., Fukumizu, K.: Convergence guarantees for kernel-based quadrature rules in misspecified settings. In: D.D. Lee, M. Sugiyama, U.V. Luxburg, I. Guyon, R. Garnett (eds.) Advances in Neural Information Processing Systems 29, pp. 3288–3296. Curran Associates, Inc. (2016)

30. Karvonen, T., Oates, C.J., Särkkä, S.: A Bayes-Sard cubature method. In: Advances in Neural Information Processing Systems 31. Curran Associates, Inc. (2018). To appear
31. Kersting, H., Hennig, P.: Active uncertainty calibration in Bayesian ODE solvers. In: Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence (UAI 2016), pp. 309–318. AUAI Press (2016)
32. Lacoste-Julien, S., Lindsten, F., Bach, F.: Sequential kernel herding: Frank-Wolfe optimization for particle filtering. In: G. Lebanon, S.V.N. Vishwanathan (eds.) Proceedings of the 18th International Conference on Artificial Intelligence and Statistics, *Proceedings of Machine Learning Research*, vol. 38, pp. 544–552. PMLR (2015)
33. Matèrn, B.: Spatial variation. Meddelanden fran Statens Skogsforskningsinstitut **49**(5) (1960)
34. Matèrn, B.: Spatial Variation, 2nd edn. Springer-Verlag (1986)
35. Minka, T.: Deriving quadrature rules from Gaussian processes. Tech. rep., Statistics Department, Carnegie Mellon University (2000)
36. Muandet, K., Fukumizu, K., Sriperumbudur, B.K., Schölkopf, B.: Kernel mean embedding of distributions : A review and beyond. Foundations and Trends in Machine Learning **10**(1–2), 1–141 (2017)
37. Narcowich, F.J., Ward, J.D.: Scattered-data interpolation on $\mathbb{R}^n$: Error estimates for radial basis and band-limited functions. SIAM Journal on Mathematical Analysis **36**, 284–300 (2004)
38. Narcowich, F.J., Ward, J.D., Wendland, H.: Sobolev bounds on functions with scattered zeros, with applications to radial basis function surface fitting. Mathematics of Computation **74**(250), 743–763 (2005)
39. Narcowich, F.J., Ward, J.D., Wendland, H.: Sobolev error estimates and a Bernstein inequality for scattered data interpolation via radial basis functions. Constructive Approximation **24**(2), 175–186 (2006)
40. Novak, E.: Deterministic and Stochastic Error Bounds in Numerical Analysis. Springer-Verlag (1988)
41. Novak, E.: Some results on the complexity of numerical integration. In: R. Cools, D. Nuyens (eds.) Monte Carlo and Quasi-Monte Carlo Methods. Springer Proceedings in Mathematics & Statistics, vol. 163, pp. 161–183. Springer, Cham (2016)
42. Novak, E., Wózniakowski, H.: Tractability of Multivariate Problems, Vol. II: Standard Information for Functionals. EMS (2010)
43. Oates, C., Niederer, S., Lee, A., Briol, F.X., Girolami, M.: Probabilistic models for integration error in the assessment of functional cardiac models. In: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (eds.) Advances in Neural Information Processing Systems 30, pp. 110–118. Curran Associates, Inc. (2017)
44. Oates, C.J., Cockayne, J., Briol, F.X., Girolami, M.: Convergence rates for a class of estimators based on Stein's method. Bernoulli (2018). To appear
45. Oates, C.J., Girolami, M.: Control functionals for quasi-Monte Carlo integration. In: A. Gretton, C.C. Robert (eds.) Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, *Proceedings of Machine Learning Research*, vol. 51, pp. 56–65. PMLR (2016)
46. Oates, C.J., Girolami, M., Chopin, N.: Control functionals for Monte Carlo integration. Journal of the Royal Statistical Society, Series B **79**(2), 323–380 (2017)
47. Oates, C.J., Papamarkou, T., Girolami, M.: The controlled thermodynamic integral for Bayesian model evidence evaluation. Journal of the American Statistical Association **111**(514), 634–645 (2016)
48. O'Hagan, A.: Bayes–Hermite quadrature. Journal of Statistical Planning and Inference **29**, 245–260 (1991)
49. Osborne, M.A., Duvenaud, D.K., Garnett, R., Rasmussen, C.E., Roberts, S.J., Ghahramani, Z.: Active learning of model evidence using Bayesian quadrature. In: F. Pereira, C.J.C. Burges, L. Bottou, K.Q. Weinberger (eds.) Advances in Neural Information Processing Systems 25, pp. 46–54. Curran Associates, Inc. (2012)
50. Paul, S., Chatzilygeroudis, K., Ciosek, K., Mouret, J.B., Osborne, M.A., Whiteson, S.: Alternating optimisation and quadrature for robust control. In: The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), pp. 3925–3933 (2018)
51. Särkkä, S., Hartikainen, J., Svensson, L., Sandblom, F.: On the relation between Gaussian process quadratures and sigma-point methods. Journal of Advances in Information Fusion **11**(1), 31–46 (2016)
52. Schaback, R.: Error estimates and condition numbers for radial basis function interpolation. Advances in Computational Mathematics **3**(3), 251–264 (1995)
53. Schaback, R., Wendland, H.: Kernel techniques: From machine learning to meshless methods. Acta Numerica **15**, 543–639 (2006)

54. Sloan, I.H., Wózniakowski, H.: When are quasi-Monte Carlo algorithms efficient for high dimensional integrals? Journal of Complexity **14**(1), 1–33 (1998)
55. Sommariva, A., Vianello, M.: Numerical cubature on scattered data by radial basis functions. Computing **76**, 295–310 (2006)
56. Sriperumbudur, B.K., Gretton, A., Fukumizu, K., Schölkopf, B., Lanckriet, G.R.: Hilbert space embeddings and metrics on probability measures. Jounal of Machine Learning Research **11**, 1517–1561 (2010)
57. Stein, E.M.: Singular Integrals and Differentiability Properties of Functions. Princeton University Press, Princeton, NJ (1970)
58. Steinwart, I., Christmann, A.: Support Vector Machines. Springer (2008)
59. Triebel, H.: Theory of Function Spaces III. Birkhäuser Verlag (2006)
60. Wendland, H.: Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. Advances in Computational Mathematics **4**(1), 389–396 (1995)
61. Wendland, H.: Scattered Data Approximation. Cambridge University Press, Cambridge, UK (2005)
62. Xi, X., Briol, F.X., Girolami, M.: Bayesian quadrature for multiple related integrals. In: J. Dy, A. Krause (eds.) Proceedings of the 35th International Conference on Machine Learning, *Proceedings of Machine Learning Research*, vol. 80, pp. 5373–5382. PMLR (2018)

Affiliations

Motonobu Kanagawa[1] · Bharath K. Sriperumbudur[2] · Kenji Fukumizu[3]

✉ Motonobu Kanagawa
motonobu.kanagawa@uni-tuebingen.de; motonobu.kanagawa@gmail.com

Bharath K. Sriperumbudur
bks18@psu.edu

Kenji Fukumizu
fukumizu@ism.ac.jp

1 University of Tübingen and Max Planck Institute for Intelligent Systems, Max-Planck-Ring 4, 72076 Tübingen, Germany

2 Department of Statistics, Pennsylvania State University, University Park, PA 16802, USA

3 The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan