SEQUENTIAL BAYESIAN QUADRATURE WITH HILBERT SPACE GAUSSIAN PROCESSES

Anonymous authors
Paper under double-blind review

ABSTRACT

Bayesian quadrature provides a probabilistic approach to numerical integration, but exact sequential Bayesian quadrature becomes increasingly expensive as function evaluations accumulate. We develop a scalable sequential Bayesian quadrature (BQ) method based on Hilbert space Gaussian process (HSGP) approximations of Matérn kernels. Both the quadrature estimator and an integral-targeted variance-reduction acquisition are represented in a finite Laplacian eigenbasis, while a geometry-stabilization constraint guarantees quasi-uniform sequential designs. For integrands in Sobolev spaces, we establish nonasymptotic mean-squared integration-error bounds that allow the target smoothness to differ from that of the Matérn reproducing kernel Hilbert space (RKHS), covering both noiseless and noisy observations. The error bounds explicitly track the feature budget, kernel lengthscale, regularization, sampling resolution, and observation noise, and determine how many features are sufficient and necessary to preserve the corresponding full-kernel quadrature rates. They further characterize response-independent parameter sequences that avoid repeated marginal-likelihood evaluation and optimization. Computation is carried out entirely in the finite feature space, substantially reducing the complexity for each integral evaluation relative to full-kernel BQ. Numerical experiments confirm the predicted trade-offs among accuracy, feature dimension, and runtime against full-kernel and approximate Bayesian quadrature baselines.

1 INTRODUCTION

Numerical integration is fundamental in probabilistic machine learning, Bayesian inference, uncertainty quantification, and scientific computing. Given a probability measure P on $\Omega \subset \mathbb{R}^d$, the goal is to estimate $I(f) := \int_\Omega f(\boldsymbol{x})\mathrm{d}P(\boldsymbol{x})$ from a limited number of evaluations of an expensive integrand f. Bayesian quadrature (BQ) places a Gaussian process (GP) prior on f and propagates the posterior through the integration functional (O'Hagan, 1991). However, BQ requires evaluating the kernel mean $\int_\Omega k(\boldsymbol{x}, \cdot)\mathrm{d}P(\boldsymbol{x})$, which is available in closed form only for certain measure-kernel pairs (Briol et al., 2019). Sequential BQ further selects evaluations according to their predicted contribution to the integral, for example through one-step posterior variance reduction (Osborne et al., 2012; Gunter et al., 2014; Kanagawa & Hennig, 2019). Its computational burden grows cubically with the design size n, since the kernel system and the acquisition must be repeatedly updated.

Finite-dimensional kernel approximations offer a natural route to scalability. Hilbert space Gaussian process (HSGP) methods use Laplacian eigenfunctions to replace the full kernel by a spectral expansion (Solin & Särkkä, 2020; Riutort-Mayol et al., 2023), while inducing-variable, Nyström, random-feature, and projection methods provide related low-dimensional representations. A central question in this literature is how accurately a finite representation must approximate the full GP or kernel estimator before the approximation becomes statistically negligible. Existing GP theory quantifies feature or inducing-point budgets for posterior approximation and contraction (Burt et al., 2020; Nieman et al., 2022), and for regression risk under random-feature, projected, and Nyström approximations, including misspecified settings (Rudi & Rosasco, 2017; Li et al., 2021; Lin & Cevher, 2020; Li et al., 2023; Huang et al., 2024). Kernel-level feature approximation has also been studied explicitly (Li et al., 2024). These results provide useful computational–statistical trade-offs for prediction and posterior approximation, but their error criteria are not integration er-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

rors. In particular, a small kernel, posterior, or regression approximation error does not directly specify how many features are sufficient to preserve the accuracy of a downstream quadrature rule.

Scalable Bayesian quadrature has developed in parallel, where the approximation is applied directly to the integration problem rather than to GP prediction. Kernel quadrature is closely connected to finite feature expansions (Bach, 2017); Nyström approximations have been analyzed for kernel means and quadrature (Chatalic et al., 2022; Hayakawa et al., 2023), while randomly pivoted Cholesky gives efficient kernel quadrature with error guarantees (Epperly & Moreno, 2023). These approaches demonstrate that substantial computational savings are possible without forming the full dense kernel system. Complementary spectral, Fourier, and recombination constructions further exploit structure in the kernel or quadrature rule (Warren et al., 2022; Warren & Ramos, 2024; Adachi et al., 2022). Sparse and Hilbert-space GP representations have also been used directly in quadrature (Li et al., 2025; Galvis-Florez et al., 2025). Exact-kernel quadrature theory, meanwhile, establishes convergence rates under adaptive sampling and smoothness misspecification (Kanagawa & Hennig, 2019; Kanagawa et al., 2020).

Despite these advances, the finite-approximation and quadrature theories mainly answer different questions. What remains unresolved is a single end-to-end characterization of how finite spectral resolution, kernel parameters, regularization, sampling geometry, target smoothness, and observation noise jointly determine the final integration error. This gap is particularly important when the function f lies in a Sobolev space of smoothness order s, i.e., $f \in H^s(\Omega)$, but the RKHS associated with the kernel has a Sobolev order τ which does not match with the order s. The finite approximation must simultaneously control spectral truncation, quadrature error in the retained space, regularization, and smoothness mismatch. A kernel approximation bound alone does not reveal how many features m are sufficient to keep truncation from becoming the statistical bottleneck.

To address the computational bottleneck and derive explicit error analysis, we develop a scalable sequential BQ method based on the standard Laplacian HSGP approximation to the Matérn kernels. We derive a theory characterizing how many features are sufficient to preserve the full kernel quadrature error rate under both oversmoothed and undersmoothed regimes, with and without observation noise. Our main contributions are as follows.

- **End-to-end parameter-explicit theory.** For quasi-uniform designs and $f \in H^s(\Omega)$ with $s > d/2$, we derive a non-asymptotic mean-squared integration error bound that explicitly tracks the number of features m, kernel lengthscale ℓ, normalized regularization λ, sampling resolution h_n, function smoothness s, RKHS smoothness τ, and observation noise σ_ϵ^2. The result covers oversmoothing ($s < \tau$), undersmoothing ($s > \tau$), and both noiseless and noisy observations. Under suitable parameter sequences, the bound recovers the rate $n^{-2\min\{s,\tau\}/d} + \sigma_\epsilon^2 n^{-1}$ for BQ based on Matérn kernels, where the deterministic term matches the benchmark full-kernel Matérn quadrature rate (Kanagawa et al., 2020), while the stochastic term attains the canonical n^{-1} noise scale (Cai et al., 2023).
- **Sufficient and necessary feature budgets.** The explicit bound determines how much spectral resolution is required to preserve this benchmark rate. In the noiseless case, feature budget $m \asymp n$ is rate-preserving for $s \leq \tau$, whereas $m \gtrsim n^{\tau/s} = o(n)$ suffices for $s > \tau$. With fixed nonzero noise, $m \gtrsim n^{d/(2s)} = o(n)$ suffices to reduce deterministic errors to the n^{-1} noise scale. We show that these lower bounds on feature budgets are tight in sequential designs.
- **Sequential BQ without likelihood optimization.** The estimator and integral-targeted acquisition are computed entirely in the HSGP feature space and combined with geometry stabilization, which guarantees quasi-uniformity in sequential designs. The theory permits (ℓ_n, λ_n, m_n) to depend only on design geometry, thereby avoiding repeated response-dependent marginal-likelihood evaluation and optimization. The feature-space linear-algebra cost is $O(nm^2 + m^3)$ rather than $O(n^3)$ for dense BQ, with further computational savings from eliminating repeated marginal-likelihood estimation (MLE). We use numerical experiments to compare our HSGP method with current sequential BQ benchmarks, and verify the predicted accuracy–feature–runtime tradeoffs.

2 Sequential Bayesian Quadrature with HSGP

We focus on Matérn kernels over bounded domains for which tensor-product Laplacian features are tractable. In addition, the Laplacian feature integrals admit closed-form expressions whenever the

characteristic function of the measure is available in closed form. The Matérn smoothness parameter is treated as fixed, while the remaining computational and regularization parameters may vary along the sequential design subject to the conditions derived from our error analysis.

2.1 FINITE-FEATURE BAYESIAN QUADRATURE

Suppose that $\Omega \subset \mathbb{R}^d$ is a bounded domain. A function f is observed at $\boldsymbol{X}_n = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subset \Omega$ through $Y_i = f(\boldsymbol{x}_i) + \epsilon_i$, $\mathbb{E}(\epsilon_i) = 0$, for $i = 1, \ldots, n$, and let $\boldsymbol{y}_n = (Y_1, \ldots, Y_n)^\top$. Our target is to estimate $I = \int_\Omega f(\boldsymbol{x}) \mathrm{d}P(\boldsymbol{x})$. For a kernel $k(\cdot, \cdot) : \Omega \times \Omega \to \mathbb{R}$, define $\boldsymbol{K}_n = [k(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j=1}^n$, the kernel-mean vector $\boldsymbol{z}_n := \int_\Omega [k(\boldsymbol{x}_1, \boldsymbol{x}), \ldots, k(\boldsymbol{x}_n, \boldsymbol{x})]^\top \mathrm{d}P(\boldsymbol{x})$, and $\kappa_P := \int_\Omega \int_\Omega k(\boldsymbol{x}, \boldsymbol{x}') \mathrm{d}P(\boldsymbol{x}) \mathrm{d}P(\boldsymbol{x}')$. With a nugget parameter $\eta > 0$, standard Bayesian quadrature (O'Hagan, 1991; Briol et al., 2019) gives the posterior mean and variance of the integral as

$$\widehat{I}_n = \boldsymbol{z}_n^\top (\boldsymbol{K}_n + \eta \boldsymbol{I}_n)^{-1} \boldsymbol{y}_n, \qquad \varrho_n^2 = \kappa_P - \boldsymbol{z}_n^\top (\boldsymbol{K}_n + \eta \boldsymbol{I}_n)^{-1} \boldsymbol{z}_n \tag{1}$$

where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix. The growing kernel system makes repeated evaluation of these quantities increasingly expensive in sequential BQ.

The Matérn kernel and its spectral density are defined as

- The Matérn kernel with length scale ℓ, scale σ^2, and smoothness $\nu > 0$,

$$k(\boldsymbol{x}, \boldsymbol{x}') = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu} \|\boldsymbol{x} - \boldsymbol{x}'\|_2}{\ell} \right)^\nu K_\nu \left(\frac{\sqrt{2\nu} \|\boldsymbol{x} - \boldsymbol{x}'\|_2}{\ell} \right), \tag{2}$$

  where K_ν is the modified Bessel function of the second kind.

- Matérn kernel spectral density:

$$S(\boldsymbol{\omega}) = \sigma^2 \frac{2^d \pi^{d/2} \Gamma(\tau)}{\Gamma(\nu)} \left(\frac{2\nu}{\ell^2} \right)^\nu \left(\frac{2\nu}{\ell^2} + \|\boldsymbol{\omega}\|_2^2 \right)^{-\tau}, \qquad \tau = \nu + d/2.$$

We replace the Matérn kernel by its HSGP representation on a fixed bounding box $D_L = (-L, L)^d$ containing Ω (Solin & Särkkä, 2020; Riutort-Mayol et al., 2023). Let $\boldsymbol{j} = (j_1, \ldots, j_d)^\top \in \mathbb{N}^d$. For a per-coordinate cutoff b, let $\mathcal{J}_b = [b]^d = \{1, \ldots, b\}^d$ and $m = b^d$ denote the total number of retained features. The rank-m HSGP approximation of kernel is

$$\hat{k}_m(\boldsymbol{x}, \boldsymbol{x}') := \sum_{\boldsymbol{j} \in [b]^d} S\left(\sqrt{\psi_{\boldsymbol{j}}}\right) \phi_{\boldsymbol{j}}(\boldsymbol{x}) \phi_{\boldsymbol{j}}(\boldsymbol{x}').$$

We consider the Dirichlet Laplacian eigenpairs for $\boldsymbol{x} = (x_1, \ldots, x_d)^\top$:

$$\phi_{\boldsymbol{j}}(\boldsymbol{x}) = L^{-d/2} \prod_{r=1}^d \sin\left(\frac{\pi j_r (x_r + L)}{2L} \right), \qquad \psi_{\boldsymbol{j}} = \left\| \frac{\pi \boldsymbol{j}}{2L} \right\|_2^2.$$

We collect the corresponding eigenfunctions into the vector $\boldsymbol{\phi}_m(\boldsymbol{x}) \in \mathbb{R}^m$ and define $\boldsymbol{\Phi}_n := [\phi_{\boldsymbol{j}}(\boldsymbol{x}_i)]_{1 \le i \le n, \boldsymbol{j} \in \mathcal{J}_b} \in \mathbb{R}^{n \times m}$, $\boldsymbol{\mu}_m := \int_\Omega \boldsymbol{\phi}_m(\boldsymbol{x}) \mathrm{d}P(\boldsymbol{x})$. The feature means $\boldsymbol{\mu}_m$ depend only on the integration measure and the Laplacian basis and can therefore be precomputed. Define $\boldsymbol{W}_m := \frac{1}{\sigma^2} \mathrm{diag}\left[S\left(\sqrt{\psi_1}\right), \ldots, S\left(\sqrt{\psi_m}\right) \right]$, which is a diagonal matrix independent of σ^2. At iteration n, let ℓ_n be the Matérn lengthscale and σ_n^2 be its scaling. We introduce the normalized regularization $\lambda_n := \eta_n / \sigma_n^2$.

The HSGP-BQ estimator can then be written entirely in terms of ℓ_n and λ_n as

$$\widehat{I}_m(f) := \boldsymbol{\mu}_m^\top \boldsymbol{C}_{m,n} \boldsymbol{\Phi}_n^\top \boldsymbol{y}_n, \qquad \boldsymbol{C}_{m,n} := \left(\lambda_n \boldsymbol{W}_m^{-1} + \boldsymbol{\Phi}_n^\top \boldsymbol{\Phi}_n \right)^{-1}. \tag{3}$$

Thus the overall kernel scale σ_n^2 does not separately affect the quadrature estimator, as it enters only through the ratio λ_n. The batch-equivalent linear-algebra cost is $O(nm^2 + m^3)$, and the sufficient statistics $\boldsymbol{\Phi}_n^\top \boldsymbol{\Phi}_n$ and $\boldsymbol{\Phi}_n^\top \boldsymbol{y}_n$ can be updated sequentially.

2.2 INTEGRAL-TARGETED SEQUENTIAL ACQUISITION

Proposition 1 (One-step reduction in the HSGP integral variance) *Fix the current parameter state $(\ell_n, \sigma_n^2, \eta_n, m_n)$ and design $\boldsymbol{X}_n$. For brevity, write $m = m_n = b_n^d$. The posterior variance of the integral under the HSGP approximation is*

$$\hat{\varrho}_n^2 = \eta_n \, \boldsymbol{\mu}_m^\top \boldsymbol{C}_{m,n} \boldsymbol{\mu}_m. \tag{4}$$

If a new observation is collected at $\boldsymbol{x} \in \Omega$ while the current parameter state is held fixed, the updated posterior variance based on $\boldsymbol{X}_n \cup \{\boldsymbol{x}\}$ satisfies

$$\hat{\varrho}_{n+1}^2(\boldsymbol{x}) - \hat{\varrho}_n^2 = -\eta_n \frac{\left[\boldsymbol{\phi}_m(\boldsymbol{x})^\top \boldsymbol{C}_{m,n} \boldsymbol{\mu}_m\right]^2}{1 + \boldsymbol{\phi}_m(\boldsymbol{x})^\top \boldsymbol{C}_{m,n} \boldsymbol{\phi}_m(\boldsymbol{x})}. \tag{5}$$

Proof of this proposition is in Appendix A.

Proposition 1 shows that the sequential acquisition can be evaluated entirely through the m-dimensional feature representation. Since η_n is independent of $\boldsymbol{x}$, from equation 5, we select the locations that maximize the one-step reduction in the current approximation of the integral variance:

$$a_n(\boldsymbol{x}) := \frac{\left[\boldsymbol{\phi}_m(\boldsymbol{x})^\top \boldsymbol{C}_{m,n} \boldsymbol{\mu}_m\right]^2}{1 + \boldsymbol{\phi}_m(\boldsymbol{x})^\top \boldsymbol{C}_{m,n} \boldsymbol{\phi}_m(\boldsymbol{x})}. \tag{6}$$

Related integral-targeted sequential criteria have been studied for exact BQ (Osborne et al., 2012; Gunter et al., 2014; Kanagawa & Hennig, 2019). We will show that different acquisition functions do not affect the convergence rate in Section 3.

Pure maximization of equation 6 does not by itself impose the geometric regularity required by deterministic sampling inequalities. Let $\boldsymbol{X}_n := \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$. Let $h_n := \sup_{\boldsymbol{x} \in \Omega} \min_{\boldsymbol{x}_k \in \boldsymbol{X}_n} \|\boldsymbol{x} - \boldsymbol{x}_k\|_2$ denote the fill distance. For a fixed $\gamma \in (0, 1)$, we use the stabilized feasible region

$$\Omega_{\gamma,n} := \{\boldsymbol{x} \in \Omega : \operatorname{dist}(\boldsymbol{x}, \boldsymbol{X}_n) \geq \gamma h_n\}, \quad \text{where } \operatorname{dist}(\boldsymbol{x}, \boldsymbol{X}_n) := \min_{\boldsymbol{x}_k \in \boldsymbol{X}_n} \|\boldsymbol{x} - \boldsymbol{x}_k\|_2, \tag{7}$$

and select

$$\boldsymbol{x}_{n+1} \in \arg\max_{\boldsymbol{x} \in \overline{\Omega}_{\gamma,n}} a_n(\boldsymbol{x}). \tag{8}$$

This separation principle is related to the γ-stabilization strategy previously studied for GP-based sequential designs (Wenzel et al., 2021). Here we combine it with the integration-specific acquisition equation 6 for BQ.

3 ERROR ANALYSIS

We now derive an end-to-end integration error bound that explicitly retains the HSGP lengthscale, feature budget, and regularization. The analysis treats the Matérn smoothness parameter ν as fixed and allows the target regularity to differ from the kernel regularity. Throughout, let $\tau = \nu + \frac{d}{2}$. For $u_n, v_n > 0$, $u_n \lesssim v_n$ and $v_n \gtrsim u_n$ mean $u_n \leq C v_n$ for some constant $C > 0$ depending only on Ω, s, d, ν, P. These parameters do not change with n. $u_n \asymp v_n$ if and only if $u_n \lesssim v_n$ and $u_n \gtrsim v_n$. The implicit constants may vary from line to line but are always independent of f and n.

We work under the following conditions.

Assumption 1 *The domain $\Omega \subset (-B, B)^d$ is bounded with a Lipschitz boundary and satisfies the interior cone condition (Wendland, 2005). The probability measure P has a density $p \in L^2(\Omega)$. The bounding box $D_L = (-L, L)^d$ is fixed with $L \geq B + 1$.*

Assumption 2 *The integrand satisfies $f \in H^s(\Omega)$ for some $s > d/2$. Here $H^s(\Omega)$ is the Sobolev space with smoothness s on Ω (see Arcangéli & Torrens (2014); Wendland (2005)).*

Assumption 3 *The observation errors are independent, mean zero, and satisfy $\mathbb{E}[\epsilon^2] \leq \sigma_\epsilon^2 < \infty$. The noiseless setting corresponds to $\epsilon \equiv 0$.*

Assumption 4 *The fixed design $\boldsymbol{X}_n$ is quasi-uniform, with fill distance h_n and separation radius $q_n := \frac{1}{2} \min_{i \neq j} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2$, such that h_n / q_n is uniformly bounded over all $n \in \mathbb{N}$, and hence $h_n \asymp q_n \asymp n^{-1/d}$ as $n \to \infty$ (Wendland, 2005).*

3.1 Convergence Rate of HSGP-Bayesian Quadrature under Static Designs

Define the conditional worst-case mean-squared integration error

$$\mathcal{R}_n := \sup_{\|f\|_{H^s(\Omega)} \le 1} \mathbb{E}_\epsilon \left[\left(I(f) - \widehat{I}_m(f) \right)^2 \middle| \boldsymbol{X}_n \right].$$

Theorem 1 (Parameter-explicit HSGP-BQ error bound) *Suppose Assumptions 1–4 hold. Let* $A_\ell := \ell_n^\nu \max\left\{1, \frac{2\nu}{\ell_n^2}\right\}^{\tau/2}$, $B_\ell := \ell_n^{-\nu} \max\left\{1, \frac{\ell_n^2}{2\nu}\right\}^{\tau/2}$. *Suppose* $\ell_n > 0$, $m_n = b_n^d$ *and* $b_n > L$. *Define* $a_+ := \max\{a, 0\}$. *Then for all sufficiently large* n,

$$\mathcal{R}_n \lesssim \left\{ A_\ell \left(h_n^\tau B_\ell + h_n^{d/2} \lambda_n^{1/2} \right) (b_n/L)^{(\tau - s)_+} + \left(\lambda_n^{-1/2} B_\ell h_n^\nu + 1 \right) \left[h_n^s + (L/b_n)^s \right] \right\}^2 + \sigma_\epsilon^2 \left(\lambda_n^{-1/2} B_\ell h_n^\tau + h_n^{d/2} \right)^2.$$

The hidden constant is independent of n, ℓ_n, λ_n, *and* σ_ϵ^2.

Complete proofs and the parameter-explicit norm comparisons are deferred to Appendix B.

Theorem 1 does not prescribe a unique parameter schedule. Instead, it directly characterizes classes of sequences that preserve a desired integration rate.

Corollary 1 (Rate-preserving parameter sequences) *Suppose the assumptions of Theorem 1 hold. Suppose the response-independent sequences* (ℓ_n, λ_n, m_n) *with* $m_n = b_n^d$ *satisfy* $\ell_n = \ell > 0$, $L = B + 1$, $\lambda_n \asymp h_n^{2\nu}$. *If* $\epsilon \equiv 0$ *and*

$$n^{\min\{s,\tau\}/s} \lesssim m_n \lesssim \begin{cases} n, & s < \tau, \\ +\infty, & s \ge \tau, \end{cases},$$

then $\mathcal{R}_n \lesssim n^{-\frac{2}{d}\min\{s,\tau\}}$. *If* $\epsilon \neq 0$ *and*

$$n^{d/(2s)} \lesssim m_n \lesssim \begin{cases} n^{\frac{\nu}{\tau - s}}, & s < \tau, \\ +\infty, & s \ge \tau. \end{cases},$$

then $\mathcal{R}_n \lesssim (1 + \sigma_\epsilon^2) n^{-1}$.

The complete proof is in Appendix C. The deterministic term $n^{-\frac{2}{d}\min\{s,\tau\}}$ matches the corresponding full-kernel Matérn quadrature rate under quasi-uniform designs (Kanagawa et al., 2020; Novak, 1988), while the n^{-1} stochastic term reaches the canonical noise floor for noisy integration (Cai et al., 2023). Corollary 1 does not require a predetermined formula for ℓ_n, λ_n, or m_n, any response-independent sequence satisfying the conditions above inherits the same target rate.

Because the error bound includes the term $A_\ell B_\ell = \max\{(2\nu)/\ell_n^2, \ell_n^2/(2\nu)\}^{\tau/2} \ge 1$, we require that ℓ_n is bounded away from 0 and $+\infty$ in order to obtain the optimal convergence rate of $\mathcal{R}_n$ for noiseless observations. With noisy observations, ℓ_n is allowed to approach 0 or ∞, while a constant ℓ_n is still sufficient. Thus, we do not need to update ℓ_n in sequential sampling.

Let $L = B + 1$. In the noiseless setting, the rate $n^{-\frac{2}{d}\min\{s,\tau\}}$ is preserved by

$$m_n \begin{cases} \asymp n, & \text{if } s \le \tau, \\ \gtrsim n^{\tau/s}, & \text{if } s > \tau. \end{cases} \tag{9}$$

Hence a sublinear feature budget is sufficient in the undersmoothing regime $s > \tau$, the minimal cost is $O(n^{1+2\tau/s} + n^{3\tau/s}) = o(n^3)$. For the oversmoothing regime $s \le \tau$, the present bound requires full spectral resolution order $m_n \asymp n$, while the complexity can still be reduced by adjusting constant coefficients, for instance, by setting $m = n/2$.

If the noise variance $\sigma_\epsilon^2 > 0$ is fixed, the noise term has order n^{-1}. Since $s > d/2$, choosing

$$m_n \asymp n^{d/(2s)} \tag{10}$$

is sufficient to reduce the deterministic terms to the noise scale, yielding $\mathcal{R}_n \lesssim n^{-1}$. Because $d/(2s) < 1$, this feature budget is sublinear in n. Since $s > d/2$, if $\tau - s > 0$, we have $\nu/(\tau - s) > 1$. Thus, we can cap m at n instead of $n^{\frac{\nu}{\tau - s}}$.

Theorem 2 (Contraction of HSGP posterior integral variance) *Suppose the assumptions of Theorem 1 hold, and let the response-independent sequences (ℓ_n, λ_n, m_n) satisfy the conditions of Corollary 1. The posterior variance of the integral under the current HSGP approximation is defined in equation 4. Then, for all sufficiently large n, $\widehat{\varrho}_n^2/\sigma_n^2 \asymp B_\ell^2 h_n^{2\tau}$. Consequently, if ℓ_n is uniformly bounded away from zero and infinity, $\widehat{\varrho}_n^2/\sigma_n^2 \asymp h_n^{2\tau}$.*

Proof of this theorem is in Appendix D. Theorem 2 shows that the posterior variance targeted by the acquisition function $\widehat{\varrho}_n^2$ consists exactly of the finite-feature worst-case quadrature error and a regularization-weight penalty. Under quasi-uniform designs, $\widehat{\varrho}_n^2 \asymp \sigma_n^2 n^{-2\tau/d}$, where σ_n^2 is the scale parameter of Matérn kernel defined in equation 2.

3.2 From fixed designs to sequential HSGP-BQ

Next, we connect the fixed-design analysis above with the sequential rule in equation 8. This step is purely geometric and is independent of the particular acquisition function.

Lemma 1 (Geometry stabilization) *Let Ω be a bounded domain. Suppose the initial design $\boldsymbol{X}_{n_0}$ consists of distinct points and $n_0 \geq 2$. For some fixed $\gamma \in (0, 1)$, suppose every newly selected point $\boldsymbol{x}_{n+1}$ satisfies $\operatorname{dist}(\boldsymbol{x}_{n+1}, \boldsymbol{X}_n) \geq \gamma h_n$, for all $n \geq n_0$. Then $h_n \to 0$, and there exists $N \geq n_0$ such that, for every $n \geq N$, $h_n/q_n \leq 2/\gamma$. Consequently, under the domain assumptions above, the sequential designs are quasi-uniform and satisfy $h_n \asymp q_n \asymp n^{-1/d}$.*

Proof of Lemma 1 is in Appendix E. Indeed, the distance at which each point is inserted is at least γ times the preceding fill distance, while the fill distances are nonincreasing. This lemma guarantees quasi-uniform design points. In practice, the fill distance is analytically available when $d = 1$, and is approximated using a Sobol' point set when $d \geq 2$ (Sobol', 1967).

Let Ψ_n denote the updating rule from Corollary 1 that depends on n, h_n, or other geometric properties of $\boldsymbol{X}_n$, but not on the observed responses $\boldsymbol{y}_n$. Under Ψ_n, we set $L = B + 1$, $\lambda_n \asymp n^{-2\nu/d}$, $m_n \asymp n^\alpha$. According to equation 9 and equation 10, the updating rule Ψ_n is

$$\alpha = 1 \text{ if } s \leq \tau \text{ and } \epsilon \equiv 0; \;\; \alpha = \tau/s \text{ if } s > \tau \text{ and } \epsilon \equiv 0; \;\; \alpha = d/(2s) \text{ if } \epsilon \neq 0. \tag{11}$$

Based on Ψ_n and the γ-stabilizing constraint specified in equation 7 and equation 8, the resulting sequential procedure is summarized in Algorithm 1.

Algorithm 1 Sequential sampling

Input: Initial data $\{(\boldsymbol{x}_j, Y_j)\}_{j=1}^n$ and an iteration budget `step`
  Select hyper-parameters ν, s, c_λ, C_m, ℓ
  Compute $L = B + 1$, $\lambda_n = c_\lambda n^{-2\nu/d}$, $m_n = \lceil \max\{L^d + 1,\ C_m n^\alpha\} \rceil$, where α is selected by equation 11
  for $i = 1, \ldots,$ `step` **do**
    Compute $a_{n+i-1}(\boldsymbol{x})$ based on the current data $\{(\boldsymbol{x}_j, Y_j)\}_{j=1}^{n+i-1}$ using equation 6
    $\boldsymbol{x}_{n+i} \leftarrow \arg\max_{\boldsymbol{x} \in \overline{\Omega}_{\gamma, n+i-1}} a_{n+i-1}(\boldsymbol{x})$
    Observe $Y_{n+i} = f(\boldsymbol{x}_{n+i}) + \epsilon_{n+i}$
    Update $\lambda_{n+i} = c_\lambda (n+i)^{-2\nu/d}$, $m_{n+i} = \lceil \max\{L^d + 1,\ C_m (n+i)^\alpha\} \rceil$
  end for
  return HSGP-BQ estimator using equation 3

Unlike a fixed-hyperparameter sequential GP procedure, we allow the normalized regularization and feature budget to evolve with the design size. In addition, this response-independent update allows the same error analysis to cover both noiseless and noisy observations. Compared with other estimation methods such as frequentist MLE or Bayesian posterior sampling, our parameter updating avoids repeated marginal-likelihood evaluation and optimization.

Algorithm 1 still requires specification of the unknown function smoothness s. In practice, this issue can be addressed in at least two ways. One approach is to estimate s using higher-order quadratic variation methods, available for both noiseless (Loh et al., 2021) and noisy (Loh & Sun,

2023) observations. Such estimates can be used to construct a plausible range for s, based on which the kernel smoothness τ can be selected. Alternatively, one may use a sufficiently smooth kernel with a large τ and adapt to the unknown function regularity through a data-dependent choice of the lengthscale ℓ_n, for example via Bayesian updating procedures (van der Vaart & van Zanten, 2009; Fang & Bhadra, 2025). These approaches provide practical ways to relax the assumption that s is known a priori.

Proposition 2 (Sequential HSGP-BQ) *Consider the stabilized sequential rule equation 8, with fixed ν and response-independent parameter updates (λ_n, m_n) according to the rule specified in Corollary 1 and Algorithm 1, then under the same assumptions of Theorem 1,*

$$\mathcal{R}_n \lesssim n^{-\frac{2}{d}\min\{s,\tau\}} + \sigma_\epsilon^2 n^{-1}.$$

In particular, the same deterministic parameter conditions apply to both noiseless and noisy sequential observations.

Proposition 2 separates the sequential-design problem into two components. Geometry stabilization guarantees quasi-uniform sampling independently of the acquisition values, while Theorem 1 determines how the lengthscale, regularization, and feature budget may evolve without sacrificing integration accuracy.

The next proposition shows that the feature budget lower bounds in equation 9 and equation 10 are not only sufficient, but also necessary to preserve the optimal error rates. Therefore, our theoretical results are sharp in the worst case over admissible integration densities. The proof is in Appendix G.

Proposition 3 (Tight lower bounds on feature budgets) *Under the setting of Proposition 2, further suppose that $\Omega = (-B, B)^d$. Let $\mathcal{P}_B = \{p \in C(\overline{\Omega}) : \int_\Omega p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 1, 1/[2(2B)^d] \le p(\boldsymbol{x}) \le 3/[2(2B)^d], \forall \boldsymbol{x} \in \Omega\}$. If, for some $r > 0$, every fixed $p \in \mathcal{P}_B$ satisfies $\mathcal{R}_n \le C_p n^{-r}$ for some p-dependent constant $C_p > 0$ and for all sufficiently large n, then necessarily $m_n \gtrsim n^{rd/(2s)}$. Consequently, preserving the noiseless rate $n^{-2\min\{s,\tau\}/d}$ requires the feature-budget orders in equation 9, while preserving the n^{-1} rate under fixed nonzero noise requires the order in equation 10.*

4 Numerical experiments

We compare Full-kernel BQ, HSGP-BQ, Sparse Spectrum BQ (SSBQ) (Warren & Ramos, 2024), and randomly pivoted Cholesky kernel quadrature (RPCholesky-KQ) (Epperly & Moreno, 2023). In the figures, we denote the proposed integral-targeted acquisition with separation constraint γ by HSGP-γ-greedy. All methods within a replicate share the same random seed and initial design.

Synthetic setup. We integrate with respect to the uniform probability measure on $(-B, B)^d$, where $B = 1$, $d = 2$, and use the cusp function as the underlying true function $f_q(\boldsymbol{x}) = \sum_{j=1}^d \frac{1}{j}\left|x_j - \frac{1}{2}\sin(j)\right|^q$. Its integral is available analytically and f_q has Sobolev smoothness $s < q + 1/2$. The main experiment uses $d = 2$, $q = 2.25$ and a Matérn-$5/2$ kernel with $\ell = 0.5$ and $\sigma^2 = 1$. Thus $\tau = \nu + d/2 = 3.5 > 2.75 > s$. We start from a 100-point Latin hypercube design and add points sequentially up to $n = 500$, using $R = 20$ replicates. For the noisy case, the observations are $Y_i = f_q(\boldsymbol{x}_i) + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0, 0.1^2)$.

The parameters update every 10 observations. For our HSGP method, we use Algorithm 1. We set $\gamma = 0.8$, $s = q + 1/2 - 10^{-3}$ and take

$$m_n = 2\left\lceil \max\left\{L^d + 1,\ C_m n^\alpha\right\}/2\right\rceil,$$

such that the condition $m_n > L^d$ in Theorem 1 is satisfied. SSBQ uses the same m_n, which is always even as required by SSBQ. The difference is that SSBQ uses MLE to fit ℓ_n and λ_n. We use $(C_m, c_\lambda) = (0.8, 1.0)$ in the noiseless case and $(C_m, c_\lambda) = (10, 10^3)$ in the noisy case. Full-kernel BQ uses the same γ-greedy constraint on full-kernel acquisition to do sequential sampling. RPCholesky-KQ instead uses its native residual sampling on a fixed Sobol pool of 2^{12} points. Since RPCholesky-KQ does not require numerical optimization in sequential sampling, its sampling time cost is lower than the other methods. Once the sequential sampling finishes, SSBQ and HSGP methods use estimator defined in equation 3, while Full-kernel BQ and RPCholesky-KQ use equation 1 where the kernel mean $\boldsymbol{z}_n$ is approximated by 2^{12} quasi-Monte Carlo (QMC) points.

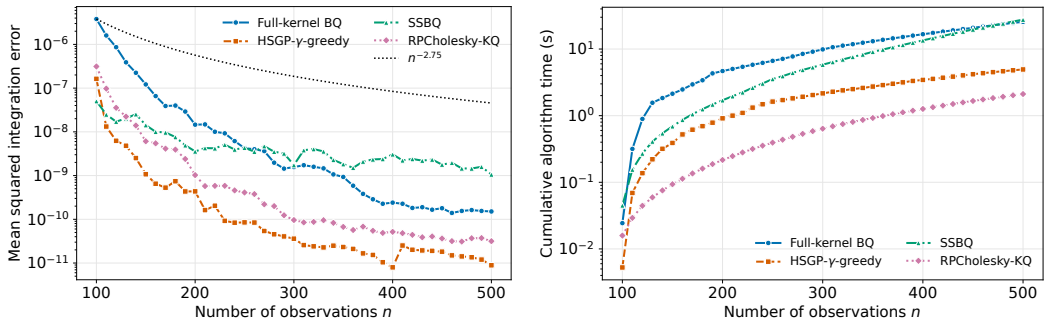


Figure 1: Noiseless synthetic experiment ($d = 2$, $q = 2.25$, $\nu = 5/2$). *(Left)* Mean squared integration error, the y-axis is log-scaled; the dotted guide is $n^{-2\min\{q+1/2,\tau\}/d} = n^{-2.75}$. *(Right)* Cumulative algorithm time. Curves are pointwise means over $R = 20$ replicates.

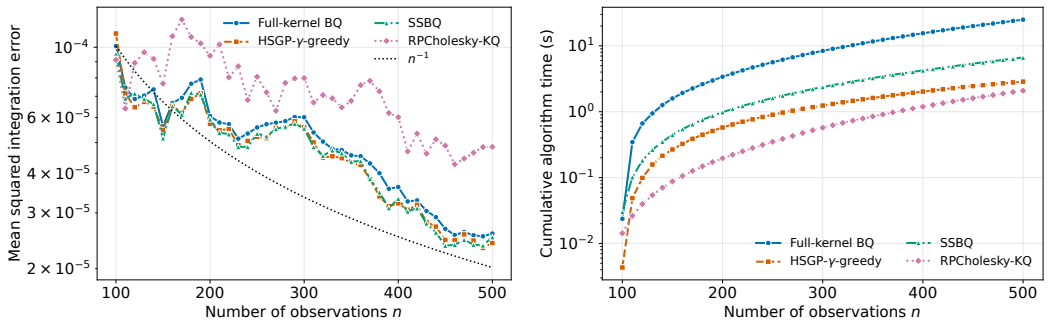


Figure 2: Noisy synthetic experiment with the same setup as Figure 1. *(Left)* Mean squared integration error; the dotted guide is n^{-1}. *(Right)* Cumulative algorithm time. Curves are pointwise means over $R = 20$ replicates.

Synthetic experiments. In the noiseless experiment (Figure 1), our HSGP-γ-greedy method achieves lowest mean squared integration error at $n = 500$, while its cumulative time cost is slightly larger than RPCholesky-KQ. One possible explanation is that RPCholesky-KQ does not require costly numerical optimization for acquisition function. The SSBQ method performs poorly compared with other methods. The main reason is that SSBQ method uses MLE to fit ℓ_n and λ_n in sequential sampling. The MLE fitting is costly and may yield inaccurate estimations. In the noisy setting (Figure 2), the Full-kernel, SSBQ and HSGP methods reach a similar noise-limited scale at $n = 500$. However, there is a large gap between these three methods and RPCholesky-KQ. This result is reasonable since RPCholesky-KQ uses fixed lengthscale and nugget (Epperly & Moreno, 2023). When observations are noisy, a tunable nugget term is essential to integration accuracy. Additional simulation results including experiments using a Matérn-$3/2$ kernel are reported in Appendix H.1; posterior-variance contraction is examined separately in Appendix H.2.

QM9 setup and results. To test performance beyond a constructed two-dimensional integrand, we estimate the finite-population mean of isotropic polarizability in a 20,000-molecule QM9 dataset used in Epperly & Moreno (2023). We use the five covariates: r^2, atomic energy u_0^{atom}, dipole moment, HOMO energy, and LUMO energy. Thus, $d = 5$. Each covariate is rank-transformed to $(-1, 1)$, and responses are centered and scaled using the initial design. We start from a 100-point Latin hypercube design of distinct molecules, add points sequentially up to $n = 1000$, and use $R = 50$ replicates. All methods use an isotropic Matérn kernel with $\nu = 6$, $\ell = 0.5$, and $\sigma^2 = 1$. This relatively smooth kernel is analogous to the smooth Gaussian kernel benchmark used by RPCholesky-KQ while retaining a finite Matérn smoothness parameter. While the true smoothness s is unknown, we assume that the observations are noisy and $s/d > 1$. According to equation 11, we set $m_n = \min\{n, 2\lceil C_m\sqrt{n}/2\rceil\}$ with $(C_m, c_\lambda) = (10, 3 \cdot 10^3)$. Full-kernel

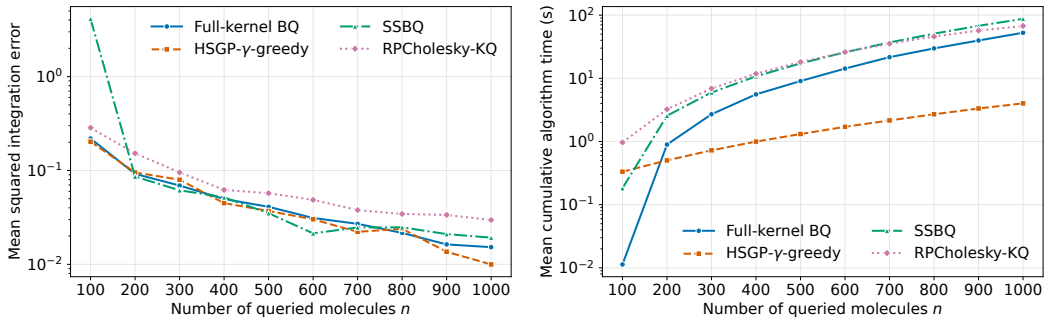


Figure 3: QM9 finite-population integration ($d = 5$, $\nu = 6$). *(Left)* Squared error in the estimated mean polarizability. *(Right)* Cumulative algorithm time.

BQ, HSGP-γ-greedy, and SSBQ use $\gamma = 0.8$ and evaluate their acquisition rules on 512 available candidates at each step. RPCholesky-KQ uses its native residual sampling without the γ-constraint.

At $n = 1000$ (Figure 3), HSGP-γ-greedy has the lowest terminal mean squared error, followed by Full-kernel BQ, SSBQ, and RPCholesky-KQ. HSGP-γ-greedy now is more efficient than RPCholesky-KQ, since no continuous optimization is required for the discrete real dataset. In addition, the kernel mean is now a finite sum of all kernel evaluations and does not require QMC approximation. Consequently, the Full-kernel BQ has a lower time cost than SSBQ and RPCholesky-KQ, but is still more costly than HSGP because it has no low-rank structure for computing acquisition. Therefore, we conclude that HSGP-γ-greedy method offers the most favorable accuracy-runtime trade-off in this experiment.

5 Discussion

We developed a sequential HSGP-based Bayesian quadrature method in which both quadrature and integral-targeted experimental design are carried out in a finite Laplacian feature space. The main theoretical message is that scalability need not be obtained by fixing the approximation or kernel parameters in advance. Instead, the feature budget, regularization, and lengthscale can evolve with the spatial resolution of the design while preserving the integration rates established for the underlying Sobolev class. A useful implication of the parameter-explicit analysis is that the finite approximation need not approximate the full kernel uniformly well. What matters for kernel integration is instead whether the truncation, regularization, and sampling errors remain below the quadrature benchmark at the current design resolution.

The experiments support a more specific computational conclusion. Our response-independent construction separates parameter selection from the realized function values. Standard GP workflows often refit hyperparameters by repeated marginal-likelihood optimization as the sequential design grows, coupling both the quadrature rule and subsequent design decisions to the observed responses. Here the parameter schedules can instead be determined from the design resolution alone. This does not remove the cost of acquisition optimization or feature-space linear algebra, but eliminates repeated likelihood fitting and makes the same deterministic analysis applicable to both noiseless and noisy observations. The geometry-stabilization condition plays a complementary role by controlling the spatial regularity required by the theory without prescribing the particular integration-targeted acquisition values.

Several extensions remain open. The current analysis focuses on Matérn kernels and uses tensor-product Laplacian features on a bounding box. It would be valuable to understand how alternative kernel families and spectral bases affect the approximation–accuracy trade-off and the resulting feature requirements. Moreover, the rate-preserving feature schedules depend on the target Sobolev smoothness s. When s is unknown, the present theory therefore provides an oracle benchmark rather than a fully adaptive procedure. Developing data-adaptive smoothness and feature-selection rules and extending the analysis to more general spectral bases and integration measures are important directions for future work.

AI USE STATEMENT

We used ChatGPT (GPT-5.6 Sol, OpenAI) to assist with three aspects of this work. First, it was used to edit and polish portions of the manuscript for clarity, grammar, and readability. Second, it was used to assist with the implementation and debugging of experimental code. Third, we used GPT-5.6 Sol to verify mathematical claims. All AI-assisted text, code, and mathematical suggestions were independently reviewed and checked by the authors. We take full responsibility for the final content of the manuscript, including all text, code, results, and claims produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All assumptions and theoretical results are stated explicitly in the main text, and complete proofs of all theorems and propositions, together with the auxiliary lemmas used in their derivations, are provided in the appendices.

We also provide an anonymized source-code archive as supplementary material, including the implementations of the proposed method and baselines, scripts for data processing and running the experiments, and instructions for reproducing the reported figures and numerical results.

REFERENCES

Masaki Adachi, Satoshi Hayakawa, Martin Jørgensen, Harald Oberhauser, and Michael A. Osborne. Fast Bayesian inference with batch Bayesian quadrature via kernel recombination. In *Advances in Neural Information Processing Systems*, volume 35, pp. 16533–16547. Curran Associates, Inc., 2022.

Rémi Arcangéli and Juan José Torrens. Sampling inequalities in Sobolev spaces. *Journal of Approximation Theory*, 182:18–28, 2014.

Francis Bach. On the equivalence between Kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research*, 18(21):1–38, 2017.

François-Xavier Briol, Chris J. Oates, Mark Girolami, Michael A. Osborne, and Dino Sejdinovic. Probabilistic integration: A role in statistical computation? *Statistical Science*, 34(1):1–22, 2019. doi: 10.1214/18-STS660.

David R. Burt, Carl Edward Rasmussen, and Mark van der Wilk. Convergence of sparse variational inference in Gaussian processes regression. *Journal of Machine Learning Research*, 21(131): 1–63, 2020.

Xu Cai, Chi Thanh Lam, and Jonathan Scarlett. On average-case error bounds for kernel-based Bayesian quadrature. *Transactions on Machine Learning Research*, 2023.

Antoine Chatalic, Nicolas Schreuder, Lorenzo Rosasco, and Alessandro Rudi. Nyström kernel mean embeddings. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 3006–3024. PMLR, 2022.

Ethan N. Epperly and Elvira Moreno. Kernel quadrature with randomly pivoted Cholesky. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Xiao Fang and Anindya Bhadra. Posterior concentration for Gaussian process priors under rescaled and hierarchical Matérn and confluent hypergeometric covariance functions. *Electronic Journal of Statistics*, 19:3333–3369, 2025.

Cristian A. Galvis-Florez, Ahmad Farooq, and Simo Särkkä. Provable quantum algorithm advantage for Gaussian process quadrature. *Transactions on Machine Learning Research*, 2025.

Tom Gunter, Michael A. Osborne, Roman Garnett, Philipp Hennig, and Stephen J. Roberts. Sampling for inference in probabilistic models with fast Bayesian quadrature. In *Advances in Neural Information Processing Systems*, volume 27, pp. 2789–2797. Curran Associates, Inc., 2014.

Satoshi Hayakawa, Harald Oberhauser, and Terry Lyons. Sampling-based Nyström approximation and kernel quadrature. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 12678–12699. PMLR, 2023.

Zhen Huang, Jiajin Sun, and Yian Huang. Quasi-Monte Carlo features for kernel approximation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20134–20165. PMLR, 2024.

Motonobu Kanagawa and Philipp Hennig. Convergence guarantees for adaptive Bayesian quadrature methods. In *Advances in Neural Information Processing Systems*, volume 32, pp. 6234–6245. Curran Associates, Inc., 2019.

Motonobu Kanagawa, Bharath K. Sriperumbudur, and Kenji Fukumizu. Convergence analysis of deterministic kernel-based quadrature rules in misspecified settings. *Foundations of Computational Mathematics*, 20(1):155–194, 2020.

Chengkun Li, Grégoire Clarté, Martin Jørgensen, and Luigi Acerbi. Fast post-process Bayesian inference with variational sparse Bayesian quadrature. *Statistics and Computing*, 35:167, 2025.

Jian Li, Yong Liu, and Weiping Wang. Optimal convergence rates for agnostic Nyström kernel learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19811–19836. PMLR, 2023.

Kevin Li, Max Balakirsky, and Simon Mak. Trigonometric quadrature Fourier features for scalable Gaussian process regression. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3484–3492. PMLR, 2024.

Zhu Li, Jean-Francois Ton, Dino Oglic, and Dino Sejdinovic. Towards a unified analysis of random Fourier features. *Journal of Machine Learning Research*, 22(108):1–51, 2021.

Junhong Lin and Volkan Cevher. Convergences of Regularized Algorithms and Stochastic Gradient Methods with Random Projections. *Journal of Machine Learning Research*, 21(20):1–44, 2020.

Wei-Liem Loh and Saifei Sun. Estimating the parameters of some common Gaussian random fields with nugget under fixed-domain asymptotics. *Bernoulli*, 29(3):2519–2543, 2023.

Wei-Liem Loh, Saifei Sun, and Jun Wen. On fixed-domain asymptotics, parameter estimation and isotropic Gaussian random fields with Matérn covariance functions. *Annals of Statistics*, 49(6): 3127–3152, 2021.

William McLean. *Strongly Elliptic Systems and Boundary Integral Equations*. Cambridge University Press, Cambridge, 2000. ISBN 978-0-521-66332-8.

Dennis Nieman, Botond Szabó, and Harry van Zanten. Contraction rates for sparse variational approximations in Gaussian process regression. *Journal of Machine Learning Research*, 23(205): 1–26, 2022.

Erich Novak. *Deterministic and Stochastic Error Bounds in Numerical Analysis*. Springer-Verlag, 1988.

Anthony O'Hagan. Bayes-Hermite quadrature. *Journal of Statistical Planning and Inference*, 29 (3):245–260, 1991.

Michael A. Osborne, Roman Garnett, Zoubin Ghahramani, David K. Duvenaud, Stephen J. Roberts, and Carl E. Rasmussen. Active learning of model evidence using Bayesian quadrature. In *Advances in Neural Information Processing Systems*, volume 25, pp. 46–54. Curran Associates, Inc., 2012.

Gabriel Riutort-Mayol, Paul-Christian Bürkner, Michael R. Andersen, Arno Solin, and Aki Vehtari. Practical Hilbert space approximate Bayesian Gaussian processes for probabilistic programming. *Statistics and Computing*, 33(1):17, 2023.

Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems*, volume 30, pp. 3215–3225, 2017.

I. M. Sobol'. On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Computational Mathematics and Mathematical Physics*, 7(4):86–112, 1967.

Arno Solin and Simo Särkkä. Hilbert space methods for reduced-rank Gaussian process regression. *Statistics and Computing*, 30(2):419–446, 2020.

A. W. van der Vaart and J. H. van Zanten. Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Annals of Statistics*, 37(5B):2655–2675, 2009.

Wenjia Wang and Bing-Yi Jing. Gaussian process regression: Optimality, robustness, and relationship with kernel ridge regression. *Journal of Machine Learning Research*.

Houston Warren and Fabio Ramos. Fast Fourier Bayesian quadrature. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4555–4563. PMLR, 2024.

Houston Warren, Rafael Oliveira, and Fabio Ramos. Generalized Bayesian quadrature with spectral kernels. In James Cussens and Kun Zhang (eds.), *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 2085–2095. PMLR, 2022.

Holger Wendland. *Scattered Data Approximation*, volume 17 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, 2005.

Tizian Wenzel, Gabriele Santin, and Bernard Haasdonk. A novel class of stabilized greedy kernel approximation algorithms: Convergence, stability and uniform point distribution. *Journal of Approximation Theory*, 262:105508, 2021.

A Proof of Proposition 1

We provide the derivation of the one-step reduction in the HSGP posterior variance of the integral. Throughout the proof, we fix the current parameter state (ℓ_n, λ_n, m_n) and write $m = m_n$ for simplicity. In particular, η_n, $\boldsymbol{W}_m$, and the feature dimension m are held fixed when comparing the designs $\boldsymbol{X}_n$ and $\boldsymbol{X}_n \cup \{\boldsymbol{x}\}$.

Under the HSGP approximation, the latent function is represented as

$$g_m(\boldsymbol{x}) = \boldsymbol{\phi}_m(\boldsymbol{x})^\top \boldsymbol{\theta}_m,$$

where the feature coefficient vector has the Gaussian prior

$$\boldsymbol{\theta}_m \sim \mathcal{N}\left(\mathbf{0}, \sigma_n^2 \boldsymbol{W}_m\right),$$
$$\text{where } \boldsymbol{W}_m = \frac{1}{\sigma_n^2}\text{diag}\left[S\left(\sqrt{\psi_1}\right), \ldots, S\left(\sqrt{\psi_m}\right)\right].$$

Given the "working" model under HSGP,

$$\boldsymbol{y}_n = \boldsymbol{\Phi}_n \boldsymbol{\theta}_m + \boldsymbol{\epsilon}_n, \qquad \boldsymbol{\epsilon}_n \sim \mathcal{N}\left(\mathbf{0}, \eta_n \boldsymbol{I}_n\right),$$

the posterior covariance matrix of $\boldsymbol{\theta}_m$ is

$$\text{Cov}\left(\boldsymbol{\theta}_m \mid \boldsymbol{y}_n\right) = \left(\frac{1}{\sigma_n^2}\boldsymbol{W}_m^{-1} + \frac{1}{\eta_n}\boldsymbol{\Phi}_n^\top \boldsymbol{\Phi}_n\right)^{-1}.$$

Since $\lambda_n = \eta_n / \sigma_n^2$, we have

$$\text{Cov}\left(\boldsymbol{\theta}_m \mid \boldsymbol{y}_n\right) = \eta_n \left(\lambda_n \boldsymbol{W}_m^{-1} + \boldsymbol{\Phi}_n^\top \boldsymbol{\Phi}_n\right)^{-1} = \eta_n \boldsymbol{C}_{m,n}.$$

The HSGP approximation to the integral is the linear functional

$$Pg_m = \int_\Omega g_m(\boldsymbol{x})\mathrm{d}P(\boldsymbol{x}) = \int_\Omega \boldsymbol{\phi}_m(\boldsymbol{x})^\top \boldsymbol{\theta}_m \mathrm{d}P(\boldsymbol{x}) = \boldsymbol{\mu}_m^\top \boldsymbol{\theta}_m,$$

where

$$\boldsymbol{\mu}_m = \int_\Omega \boldsymbol{\phi}_m(\boldsymbol{x})\mathrm{d}P(\boldsymbol{x}).$$

Therefore, conditional on the observations, its posterior variance is

$$\hat{\varrho}_n^2 := \mathrm{Var}\left(Pg_m \mid \boldsymbol{y}_n\right) = \boldsymbol{\mu}_m^\top \mathrm{Cov}\left(\boldsymbol{\theta}_m \mid \boldsymbol{y}_n\right) \boldsymbol{\mu}_m = \eta_n \boldsymbol{\mu}_m^\top \boldsymbol{C}_{m,n} \boldsymbol{\mu}_m,$$

which proves equation 4.

We next consider adding a new observation at a candidate location $\boldsymbol{x} \in \Omega$. Let

$$\boldsymbol{\phi} := \boldsymbol{\phi}_m(\boldsymbol{x}).$$

The augmented feature matrix is

$$\boldsymbol{\Phi}_{n+1} = \begin{bmatrix} \boldsymbol{\Phi}_n \\ \boldsymbol{\phi}^\top \end{bmatrix},$$

and hence

$$\boldsymbol{\Phi}_{n+1}^\top \boldsymbol{\Phi}_{n+1} = \boldsymbol{\Phi}_n^\top \boldsymbol{\Phi}_n + \boldsymbol{\phi}\boldsymbol{\phi}^\top.$$

Because the current parameter state is held fixed, the corresponding feature-space matrix after adding $\boldsymbol{x}$ is

$$\boldsymbol{C}_{m,n+1}(\boldsymbol{x}) := \left(\lambda_n \boldsymbol{W}_m^{-1} + \boldsymbol{\Phi}_{n+1}^\top \boldsymbol{\Phi}_{n+1}\right)^{-1} = \left(\lambda_n \boldsymbol{W}_m^{-1} + \boldsymbol{\Phi}_n^\top \boldsymbol{\Phi}_n + \boldsymbol{\phi}\boldsymbol{\phi}^\top\right)^{-1} = \left(\boldsymbol{C}_{m,n}^{-1} + \boldsymbol{\phi}\boldsymbol{\phi}^\top\right)^{-1}.$$

Applying the Sherman–Morrison identity gives

$$\boldsymbol{C}_{m,n+1}(\boldsymbol{x}) = \boldsymbol{C}_{m,n} - \frac{\boldsymbol{C}_{m,n}\boldsymbol{\phi}\boldsymbol{\phi}^\top \boldsymbol{C}_{m,n}}{1 + \boldsymbol{\phi}^\top \boldsymbol{C}_{m,n}\boldsymbol{\phi}}.$$

The posterior integral variance after adding the observation at $\boldsymbol{x}$ is therefore

$$\hat{\varrho}_{n+1}^2(\boldsymbol{x}) = \eta_n \boldsymbol{\mu}_m^\top \boldsymbol{C}_{m,n+1}(\boldsymbol{x}) \boldsymbol{\mu}_m = \eta_n \boldsymbol{\mu}_m^\top \boldsymbol{C}_{m,n} \boldsymbol{\mu}_m - \eta_n \frac{\boldsymbol{\mu}_m^\top \boldsymbol{C}_{m,n}\boldsymbol{\phi}\boldsymbol{\phi}^\top \boldsymbol{C}_{m,n} \boldsymbol{\mu}_m}{1 + \boldsymbol{\phi}^\top \boldsymbol{C}_{m,n}\boldsymbol{\phi}}.$$

Since $\boldsymbol{C}_{m,n}$ is symmetric,

$$\boldsymbol{\mu}_m^\top \boldsymbol{C}_{m,n}\boldsymbol{\phi}\boldsymbol{\phi}^\top \boldsymbol{C}_{m,n} \boldsymbol{\mu}_m = \left[\boldsymbol{\phi}^\top \boldsymbol{C}_{m,n} \boldsymbol{\mu}_m\right]^2.$$

Using the expression for $\hat{\varrho}_n^2$ above, we obtain

$$\hat{\varrho}_{n+1}^2(\boldsymbol{x}) - \hat{\varrho}_n^2 = -\eta_n \frac{\left[\boldsymbol{\phi}_m(\boldsymbol{x})^\top \boldsymbol{C}_{m,n} \boldsymbol{\mu}_m\right]^2}{1 + \boldsymbol{\phi}_m(\boldsymbol{x})^\top \boldsymbol{C}_{m,n}\boldsymbol{\phi}_m(\boldsymbol{x})},$$

which proves equation 5.

B Proof of Theorem 1

We prove the parameter-explicit error bound for a fixed quasi-uniform design. Throughout this section, we consider a fixed n and parameters $\ell_n, \lambda_n = \eta_n/\sigma_n^2$.

B.1 Error decomposition

The error is defined as the worst case mean square loss conditioned on design points $\boldsymbol{X}_n = \{\boldsymbol{x}_j\}_{j=1}^n$:

$$\mathcal{R}_n := \sup_{\|f\|_{H^s(\Omega)} \le 1} \mathbb{E}_\epsilon \left[\left(I - \widehat{I}_m \right)^2 \mid \boldsymbol{X}_n \right].$$

Let $\boldsymbol{\epsilon}_n = [\epsilon_1, \dots, \epsilon_n]^\top$ denote the noise vector, $\boldsymbol{F}_n = [f(\boldsymbol{x}_1), \dots, f(\boldsymbol{x}_n)]^\top$ denote the response vector. Then $\boldsymbol{y}_n = \boldsymbol{F}_n + \boldsymbol{\epsilon}_n$. Recall that the HSGP-BQ estimator is

$$\widehat{I}_m = \boldsymbol{\mu}_m^\top \boldsymbol{C}_{m,n} \boldsymbol{\Phi}_n^\top \boldsymbol{y}_n, \qquad \boldsymbol{C}_{m,n} := \left(\lambda_n \boldsymbol{W}_m^{-1} + \boldsymbol{\Phi}_n^\top \boldsymbol{\Phi}_n \right)^{-1}.$$

Let

$$\boldsymbol{\beta}_n := \boldsymbol{\Phi}_n \boldsymbol{C}_{m,n} \boldsymbol{\mu}_m,$$

then

$$\widehat{I}_m = \boldsymbol{\beta}_n^\top (\boldsymbol{F}_n + \boldsymbol{\epsilon}_n).$$

Since the noise is independent of f and has zero mean, it follows that

$$\begin{aligned} &\sup_{\|f\|_{H^s(\Omega)} \le 1} \mathbb{E}_\epsilon \left[\left(I(f) - \widehat{I}_m \right)^2 \mid \boldsymbol{X}_n \right] \\ &\le \sup_{\|f\|_{H^s(\Omega)} \le 1} \left(I(f) - \boldsymbol{\beta}_n^\top \boldsymbol{F}_n \right)^2 + \mathbb{E}_\epsilon \left[\left(\boldsymbol{\beta}_n^\top \boldsymbol{\epsilon}_n \right)^2 \mid \boldsymbol{X}_n \right]. \end{aligned} \tag{12}$$

The advantage of this decomposition is that the error rate of noiseless setting can be obtained directly by setting $\boldsymbol{\epsilon}_n = \boldsymbol{0}$.

For any $g \in L^2(\Omega)$, the deterministic part is further decomposed as

$$\left| I(f) - \boldsymbol{\beta}_n^\top \boldsymbol{F}_n \right| \le |I(f) - I(g)| + \left| I(g) - \boldsymbol{\beta}_n^\top \boldsymbol{G}_n \right| + \left| \boldsymbol{\beta}_n^\top (\boldsymbol{G}_n - \boldsymbol{F}_n) \right|, \tag{13}$$

where $\boldsymbol{G}_n = [g(\boldsymbol{x}_1), \dots, g(\boldsymbol{x}_n)]^\top$. We next illustrate how to select a proper g.

Let $\mathcal{G}_m(D_L)$ be the functional space:

$$\mathcal{G}_m(D_L) := \{\boldsymbol{x} \to \boldsymbol{\phi}_m(\boldsymbol{x})^\top \boldsymbol{\theta}_m : \boldsymbol{\theta}_m \in \mathbb{R}^m, \boldsymbol{x} \in D_L\}.$$

For $U \subset D_L$, $\mathcal{G}_m(U) := \{g|_U : g \in \mathcal{G}_m(D_L)\}$. Let

$$\|\cdot\|_{\mathcal{G}_m}^2 := \boldsymbol{\theta}_m^\top \boldsymbol{W}_m^{-1} \boldsymbol{\theta}_m.$$

B.2 Preliminary bounds

We first collect the approximation and quadrature bounds used in the proof. Throughout, $m = b^d$. Let

$$A_\ell := \ell_n^\nu \max\left\{1, \frac{2\nu}{\ell_n^2}\right\}^{\tau/2}, \qquad B_\ell := \ell_n^{-\nu} \max\left\{1, \frac{\ell_n^2}{2\nu}\right\}^{\tau/2}.$$

Lemma 2 *Suppose $b_n = m_n^{1/d} > L$. For every $f \in H^s(\Omega)$, there exists a finite-feature approximation $g_m \in \mathcal{G}_m(\Omega)$ such that*

$$\|f - g_m\|_{L^2(\Omega)} \lesssim (L/b_n)^s \|f\|_{H^s(\Omega)}, \qquad \|g_m\|_{\mathcal{G}_m} \lesssim A_\ell (b_n/L)^{(\tau-s)_+} \|f\|_{H^s(\Omega)},$$

where $(\tau - s)_+ = \max\{0, \tau - s\}$. Moreover, for a quasi-uniform fixed design,

$$\|f - g_m\|_n \lesssim \left[h_n^s + (L/b_n)^s\right] \|f\|_{H^s(\Omega)},$$

where

$$\|u\|_n := \left(\frac{1}{n} \sum_{i=1}^n u^2(\boldsymbol{x}_i) \right)^{1/2}.$$

Lemma 3 *Suppose P is a probability measure with density $p(\boldsymbol{x}) \in L^2(\Omega)$. Then for all sufficiently large n,*

$$\|\boldsymbol{\beta}_n\|_2 = \|\boldsymbol{\Phi}_n \boldsymbol{C}_{m,n} \boldsymbol{\mu}_m\|_2 \lesssim \lambda_n^{-1/2} B_\ell h_n^\tau + h_n^{d/2}.$$

In addition, for any $g_m \in \mathcal{G}_m(\Omega)$,

$$\left|I(g_m) - \boldsymbol{\beta}_n^\top \boldsymbol{G}_n\right| \lesssim \left(h_n^\tau B_\ell + h_n^{d/2} \lambda_n^{1/2}\right) \|g_m\|_{\mathcal{G}_m}.$$

B.3 Proof roadmap of Theorem 1

We first explain how to prove Theorem 1 by these lemmas and defer the detailed proof of lemmas to Section B.4 and B.5.

Suppose $g_m \in \mathcal{G}_m(\Omega)$ is defined in Lemma 2. Since the density $p(\boldsymbol{x}) \in L^2(\Omega)$,

$$|I(f) - I(g)| \le \|p\|_{L^2(\Omega)} \|f - g_m\|_{L^2(\Omega)} \lesssim (L/b_n)^s \|f\|_{H^s(\Omega)}. \tag{14}$$

By the second bound in Lemma 2 and the second bound in Lemma 3,

$$\begin{aligned}\left|I(g_m) - \boldsymbol{\beta}_n^\top \boldsymbol{G}_n\right| &\lesssim \left(h_n^\tau B_\ell + h_n^{d/2} \lambda_n^{1/2}\right) \|g_m\|_{\mathcal{G}_m} \\ &\lesssim \left(h_n^\tau B_\ell + h_n^{d/2} \lambda_n^{1/2}\right) A_\ell (b_n/L)^{(\tau-s)_+} \|f\|_{H^s(\Omega)}.\end{aligned} \tag{15}$$

Next, by the third bound in Lemma 2 and the first bound in Lemma 3,

$$\begin{aligned}&\left|\boldsymbol{\beta}_n^\top \left(\boldsymbol{G}_n - \boldsymbol{F}_n\right)\right| \\ &\le \|\boldsymbol{\beta}_n\|_2 \, n^{1/2} \|f - g_m\|_n \\ &\lesssim h_n^{d/2} \left(\lambda_n^{-1/2} B_\ell h_n^\nu + 1\right) \, n^{1/2} \left[h_n^s + (L/b_n)^s\right] \|f\|_{H^s(\Omega)} \\ &\asymp \left(\lambda_n^{-1/2} B_\ell h_n^\nu + 1\right) \left[h_n^s + (L/b_n)^s\right] \|f\|_{H^s(\Omega)}.\end{aligned} \tag{16}$$

The first inequality is from Cauchy–Schwarz inequality. The last equivalence is from Assumption 4.

Combining equation 14–equation 16, we have

$$\begin{aligned}\sup_{\|f\|_{H^s(\Omega)} \le 1} \left(I - \boldsymbol{\beta}_n^\top \boldsymbol{F}_n\right)^2 \lesssim \Bigg\{ &(L/b_n)^s + \left(h_n^\tau B_\ell + h_n^{d/2} \lambda_n^{1/2}\right) A_\ell (b_n/L)^{(\tau-s)_+} \\ &+ \left(\lambda_n^{-1/2} B_\ell h_n^\nu + 1\right) \left[h_n^s + (L/b_n)^s\right] \Bigg\}^2.\end{aligned} \tag{17}$$

For the noise term, by Lemma 3 and Assumption 3,

$$\begin{aligned}\mathbb{E}_\epsilon \left[\left(\boldsymbol{\beta}_n^\top \boldsymbol{\epsilon}_n\right)^2 \mid \boldsymbol{X}_n\right] &\le \sigma_\epsilon^2 \|\boldsymbol{\beta}_n\|_2^2 \\ &\lesssim \sigma_\epsilon^2 \left(\lambda_n^{-1/2} B_\ell h_n^\tau + h_n^{d/2}\right)^2.\end{aligned} \tag{18}$$

Finally, Theorem 1 follows from equation 17 and equation 18.

B.4 Proof of Lemma 2

First inequality. By Sobolev extension theorem, there exists an extension operator $E_s : H^s(\Omega) \to H^s(\mathbb{R}^d)$ such that

$$\tilde{f} := E_s f \in H^s(\mathbb{R}^d), \qquad \tilde{f}|_\Omega = f,$$

and

$$\|\tilde{f}\|_{H^s(\mathbb{R}^d)} \lesssim \|f\|_{H^s(\Omega)}.$$

Next, suppose χ_L is the function constructed by Lemma 4 with respect to Ω and D_L, then

$$\chi_L|_\Omega \equiv 1, \qquad \mathrm{supp}(\chi_L) \subset D_L, \qquad \|\chi_L\|_{W_\infty^{\lfloor s \rfloor+1}(\mathbb{R}^d)} \lesssim 1.$$

Let $\mathcal{P}_m$ be the projection operator on $\mathcal{G}_m(D_L)$. Define

$$g_m := \mathcal{P}_m(\chi_L E_s f) = \mathcal{P}_m(\chi_L \tilde{f}),$$

then

$$\begin{aligned}
\|f - g_m\|_{L^2(\Omega)}^2 &\le \|\chi_L\tilde{f} - \mathcal{P}_m(\chi_L\tilde{f})\|_{L^2(D_L)}^2 \\
&\overset{(1)}{=} \sum_{\boldsymbol{j}\notin[b_n]^d} \left\langle \chi_L\tilde{f}, \phi_{\boldsymbol{j}} \right\rangle_{L^2(D_L)}^2 \\
&\le \sum_{\boldsymbol{j}\notin[b_n]^d} \left[1 + \left(\frac{\pi b_n}{2L}\right)^2\right]^{-s} \left(1 + \left\|\frac{\pi \boldsymbol{j}}{2L}\right\|_2^2\right)^s \left\langle \chi_L\tilde{f}, \phi_{\boldsymbol{j}} \right\rangle_{L^2(D_L)}^2 \\
&\lesssim (L/b_n)^{2s} \sum_{\boldsymbol{j}\notin[b_n]^d} \left(1 + \left\|\frac{\pi \boldsymbol{j}}{2L}\right\|_2^2\right)^s \left\langle \chi_L\tilde{f}, \phi_{\boldsymbol{j}} \right\rangle_{L^2(D_L)}^2 \\
&\overset{(2)}{\lesssim} (L/b_n)^{2s} \|\chi_L\tilde{f}\|_{H^s(\mathbb{R}^d)}^2 \\
&\overset{(3)}{\lesssim} (L/b_n)^{2s} \|\chi_L\|_{W_\infty^{\lfloor s \rfloor+1}(\mathbb{R}^d)}^2 \|\tilde{f}\|_{H^s(\mathbb{R}^d)}^2 \\
&\overset{(4)}{\lesssim} (L/b_n)^{2s} \|f\|_{H^s(\Omega)}^2.
\end{aligned}$$

Identity (1) holds since $\{\phi_{\boldsymbol{j}}\}_{\boldsymbol{j}\in\mathbb{N}^d}$ is orthonormal basis on $L^2(D_L)$. Inequality (2) follows from Lemma 5. For inequality (3), we use Theorem 3.20 in McLean (2000). This Theorem 3.20 says if $v \in C^r(\mathbb{R}^d)$ for some integer $r \ge s$ and v has compact support. Then for $u \in H^s(\mathbb{R}^d)$,

$$\|uv\|_{H^s(\mathbb{R}^d)} \lesssim \|v\|_{W_\infty^r(\mathbb{R}^d)} \|u\|_{H^s(\mathbb{R}^d)}.$$

We apply this theorem with $u = \tilde{f}$, $v = \chi_L$, and $r = \lfloor s \rfloor + 1$. Inequality (4) holds since $\|\chi_L\|_{W_\infty^{\lfloor s \rfloor+1}(\mathbb{R}^d)} \lesssim 1$.

Second inequality. Following the construction:

$$g_m = \mathcal{P}_m(\chi_L E_s f),$$

write $g_m = \sum_{\boldsymbol{j}\in[b_n]^d} \theta_{\boldsymbol{j}}\phi_{\boldsymbol{j}}$, then

$$\begin{aligned}
\|g_m\|_{\mathcal{G}_m}^2 &= \boldsymbol{\theta}_m^\top \boldsymbol{W}_m^{-1} \boldsymbol{\theta}_m \\
&\asymp \sum_{\boldsymbol{j}\in[b_n]^d} \ell_n^{2\nu} \left(\frac{2\nu}{\ell_n^2} + \left\|\frac{\pi \boldsymbol{j}}{2L}\right\|_2^2\right)^\tau \theta_{\boldsymbol{j}}^2 \\
&\lesssim \ell_n^{2\nu} \max\left\{1, \frac{2\nu}{\ell_n^2}\right\}^\tau \sum_{\boldsymbol{j}\in[b_n]^d} \left(1 + \left\|\frac{\pi \boldsymbol{j}}{2L}\right\|_2^2\right)^\tau \theta_{\boldsymbol{j}}^2 \\
&\lesssim \ell_n^{2\nu} \max\left\{1, \frac{2\nu}{\ell_n^2}\right\}^\tau \left(1 + \left\|\frac{\pi b_n}{2L}\right\|_2^2\right)^{(\tau-s)_+} \sum_{\boldsymbol{j}\in[b_n]^d} \left(1 + \left\|\frac{\pi \boldsymbol{j}}{2L}\right\|_2^2\right)^s \theta_{\boldsymbol{j}}^2 \\
&\lesssim A_\ell^2 (b_n/L)^{2(\tau-s)_+} \|f\|_{H^s(\Omega)}^2.
\end{aligned}$$

Third inequality. Since the Laplacian basis functions are smooth,

$$g_m \in \mathcal{G}_m(\Omega) \subset H^s(\Omega).$$

Then

$$\begin{aligned}\|f-g_m\|_n &\overset{(1)}{\lesssim} h_n^s\|f-g_m\|_{H^s(\Omega)} + \|f-g_m\|_{L^2(\Omega)} \\ &\overset{(2)}{\lesssim} h_n^s\left(\|f\|_{H^s(\Omega)} + \|g_m\|_{H^s(\Omega)}\right) + (L/b_n)^s\|f\|_{H^s(\Omega)} \\ &\overset{(3)}{\lesssim} \left[h_n^s + (L/b_n)^s\right]\|f\|_{H^s(\Omega)}. \end{aligned} \tag{19}$$

Inequality (1) is from Lemma 28 in Wang & Jing; inequality (2) is by triangle inequality and the first bound. Inequality (3) is from the definition of g_m, Lemma 5, and Lemma 6 with smoothness s:

$$\begin{aligned}\|g_m\|^2_{H^s(\Omega)} &\lesssim \sum_{\boldsymbol{j}\in[b_n]^d}\left(1+\left\|\frac{\pi\boldsymbol{j}}{2L}\right\|_2^2\right)^s\theta_{\boldsymbol{j}}^2 \\ &\le \sum_{\boldsymbol{j}\in\mathbb{N}^d}\left(1+\left\|\frac{\pi\boldsymbol{j}}{2L}\right\|_2^2\right)^s\left|\langle\chi_L E_s f,\phi_{\boldsymbol{j}}\rangle_{L^2(D_L)}\right|^2 \\ &\lesssim \|\chi_L E_s f\|^2_{H^s(\mathbb{R}^d)} \\ &\lesssim \|f\|^2_{H^s(\Omega)}.\end{aligned}$$

Lemma 4 *Let $U=(-a,a)^d, V=(-b,b)^d, 0<a<b$ and $b-a\ge c>0$. For any $r\in\mathbb{N}$, there exists a constant $C(r,d,c)>0$ and $\zeta\in C^\infty(\mathbb{R}^d)$ such that*

$$\zeta\in[0,1],\ \zeta|_U\equiv 1,\ \operatorname{supp}(\zeta)=\left[-\frac{a+b}{2},\frac{a+b}{2}\right]^d,\ \|\zeta\|_{W^r_\infty(\mathbb{R}^d)}\le C(r,d,c).$$

The constant $C(r,d,c)$ depends only on r,d,c.

Proof of Lemma 4

Define

$$\begin{aligned}e(t) &= \begin{cases}\exp(-t^{-1}), & t>0,\\ 0, & t\le 0.\end{cases}\\ \gamma(t) &= \frac{e(t)}{e(t)+e(1-t)},\\ \psi(t) &= \gamma\left(\frac{2t+b+a}{b-a}\right)\gamma\left(\frac{b+a-2t}{b-a}\right),\\ \zeta(\boldsymbol{x}) &= \prod_{p=1}^d\psi(x_p).\end{aligned}$$

By construction, $\zeta\in[0,1]$, $\zeta\in C^\infty(\mathbb{R}^d)$, $\zeta|_U\equiv 1$, and ζ is supported on

$$\overline{\left\{t:\frac{2t+b+a}{b-a}>0\right\}\cap\left\{t:\frac{b+a-2t}{b-a}>0\right\}}^d=\left[-\frac{a+b}{2},\frac{a+b}{2}\right]^d.$$

In addition, for $r\in\mathbb{N}$,

$$\psi^{(r)}(t)=\begin{cases}\left(\frac{b-a}{2}\right)^{-r}\gamma^{(r)}\left(\frac{2t+b+a}{b-a}\right), & t\in\left(-\frac{a+b}{2},-a\right),\\ \left(\frac{a-b}{2}\right)^{-r}\gamma^{(r)}\left(\frac{b+a-2t}{b-a}\right), & t\in\left(a,\frac{a+b}{2}\right),\\ 0, & \text{else.}\end{cases}$$

Therefore,

$$\|\zeta\|_{W^r_\infty(\mathbb{R}^d)}=\sum_{|\alpha|\le r}\|D^\alpha\zeta\|_{L^\infty(\mathbb{R}^d)}$$

$$\leq \sum_{|\alpha|\leq r} \left(\frac{b-a}{2}\right)^{-|\alpha|} \left[\prod_{p=1}^{d} \sup_{t\in[0,1]} \left|\gamma^{(\alpha_p)}(t)\right|\right]$$
$$\leq \sum_{|\alpha|\leq r} \left(\frac{1}{2}c\right)^{-|\alpha|} \left[\prod_{p=1}^{d} \sup_{t\in[0,1]} \left|\gamma^{(\alpha_p)}(t)\right|\right].$$

Lemma 5 *If $u \in H^s(\mathbb{R}^d)$, $\mathrm{supp}(u) \subset D_L$, then*

$$\sum_{\boldsymbol{j}\in\mathbb{N}^d} \left(1 + \left\|\frac{\pi \boldsymbol{j}}{2L}\right\|_2^2\right)^s \left|\langle u, \phi_{\boldsymbol{j}}\rangle_{L^2(D_L)}\right|^2 \lesssim \|u\|^2_{H^s(\mathbb{R}^d)}.$$

Proof of Lemma 5

Let $s = \lfloor s \rfloor + \rho,\ \rho \in [0,1)$. We use the Sobolev norm defined as:

$$\|f\|^2_{H^s(\Omega)} := \sum_{|\alpha|\leq\lfloor s\rfloor} \|D^\alpha f\|^2_{L^2(\Omega)} + \mathbf{1}(\rho>0) \sum_{|\alpha|=\lfloor s\rfloor} \iint_{\Omega\times\Omega} \frac{|D^\alpha f(\boldsymbol{x}) - D^\alpha f(\boldsymbol{x}')|^2}{\|\boldsymbol{x}-\boldsymbol{x}'\|_2^{2\rho+d}} \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{x}', \tag{20}$$

where D^α is weak derivative with respect to a multi-index $\alpha \in \mathbb{N}_0^d$.

Notice that

$$\begin{aligned}
\phi_{\boldsymbol{j}}(\boldsymbol{x}) &= L^{-d/2} \prod_{p=1}^{d} \sin\left(\frac{\pi j_p}{2L}(x_p + L)\right) \\
&= L^{-d/2} \prod_{p=1}^{d} \frac{\exp\left(i\frac{\pi j_p}{2L}(x_p+L)\right) - \exp\left(-i\frac{\pi j_p}{2L}(x_p+L)\right)}{2i} \\
&= L^{-d/2} \sum_{\boldsymbol{\delta}\in\{\pm1\}^d} C_{\boldsymbol{\delta},\boldsymbol{j}} \exp\left(\frac{\pi}{2L} i \,\langle \boldsymbol{x},\ \boldsymbol{\delta}\odot\boldsymbol{j}\rangle_2\right),
\end{aligned} \tag{21}$$

where

$$\boldsymbol{\delta}\odot\boldsymbol{j} := [\delta_1 j_1, \ldots, \delta_d j_d]^\top,$$

and

$$|C_{\boldsymbol{\delta},\boldsymbol{j}}| \leq \left|\frac{1}{(2i)^d} \exp\left(i\frac{\pi}{2}\langle\boldsymbol{\delta},\ \boldsymbol{j}\rangle_2\right)\right| \leq 2^{-d}.$$

Let $\mathcal{F}$ denote the Fourier transform with the convention:

$$\mathcal{F}(u)(\boldsymbol{\omega}) = \int_{\mathbb{R}^d} u(\boldsymbol{x}) \exp(-i\boldsymbol{\omega}^\top \boldsymbol{x}) \mathrm{d}\boldsymbol{x},$$

then

$$\begin{aligned}
&\left|\langle u, \phi_{\boldsymbol{j}}\rangle_{L^2(D_L)}\right|^2 \\
&\overset{(1)}{=} L^{-d} \left|\int_{\mathbb{R}^d} u(\boldsymbol{x}) \sum_{\boldsymbol{\delta}\in\{\pm1\}^d} \overline{C}_{\boldsymbol{\delta},\boldsymbol{j}} \exp\left(-\frac{\pi}{2L} i\langle\boldsymbol{x},\ \boldsymbol{\delta}\odot\boldsymbol{j}\rangle_2\right) \mathrm{d}\boldsymbol{x}\right|^2 \\
&= L^{-d} \left|\sum_{\boldsymbol{\delta}\in\{\pm1\}^d} \overline{C}_{\boldsymbol{\delta},\boldsymbol{j}} \mathcal{F}(u)\left(\frac{\pi}{2L}\boldsymbol{\delta}\odot\boldsymbol{j}\right)\right|^2 \\
&\overset{(2)}{\lesssim} L^{-d} \sum_{\boldsymbol{\delta}\in\{\pm1\}^d} \left|\mathcal{F}(u)\left(\frac{\pi}{2L}\boldsymbol{\delta}\odot\boldsymbol{j}\right)\right|^2,
\end{aligned} \tag{22}$$

where Equation (1) holds since $\mathrm{supp}(u) \subset D_L$. For inequality (2), we use Cauchy-Schwarz inequality on the finite sum.

Case 1: If $\rho = 0$,

$$\begin{aligned}
&\sum_{\boldsymbol{j}\in\mathbb{N}^d}\left(1+\left\|\frac{\pi\boldsymbol{j}}{2L}\right\|_2^2\right)^s \left|\langle u,\phi_{\boldsymbol{j}}\rangle_{L^2(D_L)}\right|^2 \\
&\overset{(1)}{\lesssim} L^{-d}\sum_{\boldsymbol{j}\in\mathbb{N}^d}\sum_{\boldsymbol{\delta}\in\{\pm1\}^d}\left(1+\left\|\frac{\pi\boldsymbol{j}}{2L}\right\|_2^2\right)^s \left|\mathcal{F}(u)\left(\frac{\pi}{2L}\boldsymbol{\delta}\odot\boldsymbol{j}\right)\right|^2 \\
&\overset{(2)}{\lesssim} L^{-d}\sum_{\boldsymbol{j}\in\mathbb{N}^d}\sum_{\boldsymbol{\delta}\in\{\pm1\}^d}\sum_{|\alpha|\le s}\left|\mathcal{F}(D^\alpha u)\left(\frac{\pi}{2L}\boldsymbol{\delta}\odot\boldsymbol{j}\right)\right|^2 \\
&\le L^{-d}\sum_{\boldsymbol{j}\in\mathbb{Z}^d}\sum_{|\alpha|\le s}\left|\mathcal{F}(D^\alpha u)\left(\frac{\pi}{2L}\boldsymbol{j}\right)\right|^2 \\
&\overset{(3)}{\lesssim} \sum_{|\alpha|\le s}\|D^\alpha u\|^2_{L^2(\mathbb{R}^d)}.
\end{aligned}$$

Inequality (1) is from equation 22. For inequality (2), we use

$$\left(1+\|\boldsymbol{\omega}\|_2^2\right)^s \lesssim \sum_{|\alpha|\le s}\prod_{p=1}^d \omega_p^{2\alpha_p},$$

with $\omega = \frac{\pi}{2L}\boldsymbol{\delta}\odot\boldsymbol{j}$, and

$$\mathcal{F}(D^\alpha u)(\boldsymbol{\omega}) = \mathcal{F}(u)(\boldsymbol{\omega})\prod_{p=1}^d (i\omega_p)^{\alpha_p}, \qquad \operatorname{supp}(u)\subset D_L.$$

Inequality (3) is by Parseval identity: since

$$(4L)^{-d/2}\exp\left(i\frac{\pi}{2L}\langle\boldsymbol{x},\ \boldsymbol{j}\rangle_2\right)$$

is orthonormal Fourier basis on $(-2L,2L)^d$,

$$\begin{aligned}
&(4L)^{-d}\sum_{\boldsymbol{j}\in\mathbb{Z}^d}\left|\mathcal{F}(D^\alpha u)\left(\frac{\pi}{2L}\boldsymbol{j}\right)\right|^2 \\
&= \sum_{\boldsymbol{j}\in\mathbb{Z}^d}\left|\left\langle D^\alpha u(\boldsymbol{x}),(4L)^{-d/2}\exp\left(i\frac{\pi}{2L}\langle\boldsymbol{x},\ \boldsymbol{j}\rangle_2\right)\right\rangle_{L^2((-2L,2L)^d)}\right|^2 \\
&= \|D^\alpha u\|^2_{L^2((-2L,2L)^d)} \\
&\le \|D^\alpha u\|^2_{L^2(\mathbb{R}^d)}.
\end{aligned}$$

Case 2: If $\rho > 0$. Let $r = \lfloor s\rfloor$. Notice that

$$\begin{aligned}
\left(1+\|\boldsymbol{\omega}\|_2^2\right)^{r+\rho} &\lesssim \begin{cases}1, & \|\boldsymbol{\omega}\|_2<1,\\ \|\boldsymbol{\omega}\|_2^{2(r+\rho)}, & \|\boldsymbol{\omega}\|_2\ge 1\end{cases} \\
&\lesssim \sum_{|\alpha|\le r}\prod_{p=1}^d \omega_p^{2\alpha_p} + \|\boldsymbol{\omega}\|_2^{2\rho}\sum_{|\alpha|=r}\prod_{p=1}^d\omega_p^{2\alpha_p},
\end{aligned}$$

it follows that

$$\begin{aligned}
&\sum_{\boldsymbol{j}\in\mathbb{N}^d}\left(1+\left\|\frac{\pi\boldsymbol{j}}{2L}\right\|_2^2\right)^{r+\rho}\left|\langle u,\phi_{\boldsymbol{j}}\rangle_{L^2(D_L)}\right|^2 \\
&\lesssim L^{-d}\sum_{\boldsymbol{j}\in\mathbb{N}^d}\sum_{\boldsymbol{\delta}\in\{\pm1\}^d}\left\|\frac{\pi\boldsymbol{j}}{2L}\right\|_2^{2\rho}\sum_{|\alpha|=r}\left|\mathcal{F}(D^\alpha u)\left(\frac{\pi}{2L}\boldsymbol{\delta}\odot\boldsymbol{j}\right)\right|^2
\end{aligned}$$

$$+ L^{-d} \sum_{\boldsymbol{j} \in \mathbb{N}^d} \sum_{\boldsymbol{\delta} \in \{\pm 1\}^d} \sum_{|\alpha| \le r} \left| \mathcal{F}(D^\alpha u) \left(\frac{\pi}{2L} \boldsymbol{\delta} \odot \boldsymbol{j} \right) \right|^2. \tag{23}$$

It suffices to bound the term with $\left\| \frac{\pi \boldsymbol{j}}{2L} \right\|_2^{2\rho}$. Notice that for $\boldsymbol{j} \in \mathbb{N}^d$,

$$\begin{aligned}
&\int_{(-L,L)^d} \frac{\left| \exp\left(-i \frac{\pi}{2L} \boldsymbol{a}^\top \boldsymbol{j} \right) - 1 \right|^2}{\|\boldsymbol{a}\|_2^{2\rho+d}} \mathrm{d}\boldsymbol{a} \\
&= \left\| \frac{\pi \boldsymbol{j}}{2L} \right\|_2^{2\rho} \int_{\|\boldsymbol{a}\|_\infty < \pi \|\boldsymbol{j}\|_2 / 2} \frac{\left| \exp\left(-i \boldsymbol{a}^\top \boldsymbol{j} \,/\, \|\boldsymbol{j}\|_2 \right) - 1 \right|^2}{\|\boldsymbol{a}\|_2^{2\rho+d}} \mathrm{d}\boldsymbol{a} \\
&\ge \left\| \frac{\pi \boldsymbol{j}}{2L} \right\|_2^{2\rho} \int_{\|\boldsymbol{a}\|_\infty \le \pi/2} \frac{\left| \exp\left(-i \boldsymbol{a}^\top \boldsymbol{j} \,/\, \|\boldsymbol{j}\|_2 \right) - 1 \right|^2}{\|\boldsymbol{a}\|_2^{2\rho+d}} \mathrm{d}\boldsymbol{a} \\
&\overset{(1)}{\gtrsim} \left\| \frac{\pi \boldsymbol{j}}{2L} \right\|_2^{2\rho}.
\end{aligned} \tag{24}$$

For the last inequality, we use

$$\begin{aligned}
&\int_{\|\boldsymbol{a}\|_\infty \le \pi/2} \frac{\left| \exp\left(-i \boldsymbol{a}^\top \boldsymbol{j} \,/\, \|\boldsymbol{j}\|_2 \right) - 1 \right|^2}{\|\boldsymbol{a}\|_2^{2\rho+d}} \mathrm{d}\boldsymbol{a} \\
&\le \int_{\mathbb{R}^d} \frac{\left| \exp\left(-i \boldsymbol{a}^\top \boldsymbol{j} \,/\, \|\boldsymbol{j}\|_2 \right) - 1 \right|^2}{\|\boldsymbol{a}\|_2^{2\rho+d}} \mathrm{d}\boldsymbol{a} \\
&= \int_{\|\boldsymbol{a}\|_2 \le 1} \frac{\left| \exp\left(i \boldsymbol{a}^\top \boldsymbol{j} \,/\, \|\boldsymbol{j}\|_2 \right) - 1 \right|^2}{\|\boldsymbol{a}\|_2^{2\rho+d}} \mathrm{d}\boldsymbol{a} + \int_{\|\boldsymbol{a}\|_2 > 1} \frac{\left| \exp\left(i \boldsymbol{a}^\top \boldsymbol{j} \,/\, \|\boldsymbol{j}\|_2 \right) - 1 \right|^2}{\|\boldsymbol{a}\|_2^{2\rho+d}} \mathrm{d}\boldsymbol{a} \\
&\overset{(1)}{\le} \int_{\|\boldsymbol{a}\|_2 \le 1} \frac{\|\boldsymbol{a}\|_2^2}{\|\boldsymbol{a}\|_2^{2\rho+d}} \mathrm{d}\boldsymbol{a} + \int_{\|\boldsymbol{a}\|_2 > 1} \frac{4}{\|\boldsymbol{a}\|_2^{2\rho+d}} \mathrm{d}\boldsymbol{a} \\
&= \int_{S^{d-1}} \left(\int_0^1 R^{2-2\rho-d} R^{d-1} \mathrm{d}R + \int_1^{+\infty} 4 R^{-2\rho-d} R^{d-1} \mathrm{d}R \right) \mathrm{d}\boldsymbol{\theta} \\
&= \frac{\pi^{d/2}}{\Gamma(d/2)} \frac{4 - 3\rho}{\rho(1-\rho)} < \infty,
\end{aligned} \tag{25}$$

where inequality (1) follows from $|\exp(iz) - 1| \le |z|$ and $|\exp(iz) - 1| \le 2$.

For the lower bound, define

$$F(\boldsymbol{v}) := \int_{\|\boldsymbol{a}\|_\infty \le \pi/2} \frac{\left| \exp\left(-i \boldsymbol{a}^\top \boldsymbol{v} \right) - 1 \right|^2}{\|\boldsymbol{a}\|_2^{2\rho+d}} \mathrm{d}\boldsymbol{a}.$$

Since $\rho \in (0,1)$, $\|F\|_{L^\infty(S^{d-1})} < \infty$. By dominated convergence theorem, $F(\boldsymbol{v})$ is continuous on the compact set S^{d-1}. In addition, $\exp\left(-i \boldsymbol{a}^\top \boldsymbol{v} \right) \ne 1$ almost surely. Consequently,

$$\begin{aligned}
&\int_{\|\boldsymbol{a}\|_\infty \le \pi/2} \frac{\left| \exp\left(-i \boldsymbol{a}^\top \boldsymbol{j} \,/\, \|\boldsymbol{j}\|_2 \right) - 1 \right|^2}{\|\boldsymbol{a}\|_2^{2\rho+d}} \mathrm{d}\boldsymbol{a} \\
&\ge \inf_{\boldsymbol{v} \in S^{d-1}} F(\boldsymbol{v}) \\
&> 0,
\end{aligned}$$

Finally, combining equation 23 and equation 24 yields

$$\begin{aligned}
&L^{-d} \sum_{\boldsymbol{j} \in \mathbb{N}^d} \sum_{\boldsymbol{\delta} \in \{\pm 1\}^d} \left\| \frac{\pi \boldsymbol{j}}{2L} \right\|_2^{2\rho} \sum_{|\alpha| = r} \left| \mathcal{F}(D^\alpha u) \left(\frac{\pi}{2L} \boldsymbol{\delta} \odot \boldsymbol{j} \right) \right|^2 \\
&\lesssim L^{-d} \sum_{\boldsymbol{j} \in \mathbb{N}^d} \sum_{\boldsymbol{\delta} \in \{\pm 1\}^d} \sum_{|\alpha| = r} \int_{(-L,L)^d} \frac{\left| \exp\left(-i \frac{\pi}{2L} \boldsymbol{a}^\top \boldsymbol{\delta} \odot \boldsymbol{j} \right) - 1 \right|^2 \left| \mathcal{F}(D^\alpha u) \left(\frac{\pi}{2L} \boldsymbol{\delta} \odot \boldsymbol{j} \right) \right|^2}{\|\boldsymbol{a}\|_2^{2\rho+d}} \mathrm{d}\boldsymbol{a}
\end{aligned}$$

$$
\begin{aligned}
&\leq L^{-d} \sum_{\boldsymbol{j}\in\mathbb{Z}^d} \sum_{|\alpha|=r} \int_{(-L,L)^d} \frac{\left|\exp\left(-i\frac{\pi}{2L}\boldsymbol{a}^\top \boldsymbol{j}\right)-1\right|^2 \left|\mathcal{F}(D^\alpha u)\left(\frac{\pi}{2L}\boldsymbol{j}\right)\right|^2}{\|\boldsymbol{a}\|_2^{2\rho+d}} \mathrm{d}\boldsymbol{a} \\
&= L^{-d} \sum_{|\alpha|=r} \sum_{\boldsymbol{j}\in\mathbb{Z}^d} \int_{(-L,L)^d} \frac{\left|\mathcal{F}\left[D^\alpha u(\cdot+\boldsymbol{a}) - D^\alpha u(\cdot)\right]\left(\frac{\pi \boldsymbol{j}}{2L}\right)\right|^2}{\|\boldsymbol{a}\|_2^{2\rho+d}} \mathrm{d}\boldsymbol{a} \\
&\lesssim \sum_{|\alpha|=r} \int_{(-L,L)^d} \frac{\int_{(-2L,2L)^d} \left|D^\alpha u(\boldsymbol{z}+\boldsymbol{a}) - D^\alpha u(\boldsymbol{z})\right|^2 \mathrm{d}\boldsymbol{z}}{\|\boldsymbol{a}\|_2^{2\rho+d}} \mathrm{d}\boldsymbol{a} \\
&\leq \sum_{|\alpha|=r} \iint_{\mathbb{R}^d\times\mathbb{R}^d} \frac{|D^\alpha u(\boldsymbol{x}) - D^\alpha u(\boldsymbol{x}')|^2}{\|\boldsymbol{x}-\boldsymbol{x}'\|_2^{2\rho+d}} \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{x}' \\
&\leq \|u\|^2_{H^s(\mathbb{R}^d)}.
\end{aligned}
$$

B.5 Proof of Lemma 3

Let

$$
\begin{aligned}
\boldsymbol{\gamma}_n &:= \boldsymbol{C}_{m,n}\boldsymbol{\mu}_m, \\
h(\boldsymbol{t}) &:= \lambda_n\, \boldsymbol{\gamma}_n^\top \boldsymbol{\phi}_m(\boldsymbol{t}), \\
u_n &:= \int_\Omega h(\boldsymbol{t}) \mathrm{d}P(\boldsymbol{t}).
\end{aligned}
$$

First, as $p(\boldsymbol{x}) \in L^2(\Omega)$,

$$
u_n \lesssim \|h\|_{L^2(\Omega)}. \tag{26}
$$

The term u_n also admits the lower bound

$$
u_n = \lambda_n \|\boldsymbol{\beta}_n\|_2^2 + \lambda_n^2 \boldsymbol{\gamma}_n^\top \boldsymbol{W}_m^{-1} \boldsymbol{\gamma}_n \geq \lambda_n \|\boldsymbol{\beta}_n\|_2^2. \tag{27}
$$

Since $\|h\|^2_{\mathcal{G}_m} = \lambda_n^2 \boldsymbol{\gamma}_n^\top \boldsymbol{W}_m^{-1} \boldsymbol{\gamma}_n$, we also have

$$
\|h\|^2_{\mathcal{G}_m} \leq u_n.
$$

By Lemma 6,

$$
\|h\|^2_{H^\tau(\Omega)} \lesssim B_\ell^2 \|h\|^2_{\mathcal{G}_m} \leq B_\ell^2 u_n. \tag{28}
$$

In addition,

$$
\|h\|_n^2 = \frac{\lambda_n^2}{n} \|\beta_n\|_2^2 \leq \frac{\lambda_n}{n} u_n. \tag{29}
$$

Combining equation 26, equation 28, and equation 29 with Sobolev sampling inequality, which is formula (1.2) in Arcangéli & Torrens (2014), yields

$$
\begin{aligned}
u_n &\lesssim \|h\|_{L^2(\Omega)} \\
&\lesssim h_n^\tau \|h\|_{H^\tau(\Omega)} + h_n^{d/2} n^{1/2} \|h\|_n \\
&\lesssim \left(h_n^\tau B_\ell + h_n^{d/2} \lambda_n^{1/2}\right) u_n^{1/2}.
\end{aligned}
$$

Therefore, by equation 27,

$$
\|\boldsymbol{\beta}_n\|_2 \leq \lambda_n^{-1/2} u_n^{1/2} \lesssim \lambda_n^{-1/2} h_n^\tau B_\ell + h_n^{d/2}.
$$

For the second bound,

$$
\begin{aligned}
\left|I(g_m) - \boldsymbol{\beta}_n^\top \boldsymbol{G}_n\right| &= \left|(\boldsymbol{\mu}_m - \boldsymbol{\Phi}_n^\top \boldsymbol{\Phi}_n \boldsymbol{\gamma}_n)^\top \boldsymbol{\theta}_m\right| \\
&= \lambda_n \left|\boldsymbol{\gamma}_n^\top \boldsymbol{W}_m^{-1} \boldsymbol{\theta}_m\right| \\
&\overset{(1)}{\leq} \lambda_n \left(\boldsymbol{\gamma}_n^\top \boldsymbol{W}_m^{-1} \boldsymbol{\gamma}_n\right)^{1/2} \left(\boldsymbol{\theta}_m^\top \boldsymbol{W}_m^{-1} \boldsymbol{\theta}_m\right)^{1/2} \\
&= \|h\|_{\mathcal{G}_m} \|g_m\|_{\mathcal{G}_m} \\
&\lesssim \left(h_n^\tau B_\ell + h_n^{d/2} \lambda_n^{1/2}\right) \|g_m\|_{\mathcal{G}_m}.
\end{aligned}
$$

Inequality (1) is by Cauchy–Schwarz inequality and $\boldsymbol{W}_m$ is a diagonal matrix.

Lemma 6 *For any $g_m \in \mathcal{G}_m(\Omega)$,*

$$\|g_m\|_{H^\tau(\Omega)} \lesssim \ell_n^{-\nu} \max\left\{1, \frac{\ell_n^2}{2\nu}\right\}^{\tau/2} \|g_m\|_{\mathcal{G}_m}.$$

Proof of Lemma 6

let g_m be denoted by $\sum_{\boldsymbol{j}\in[b_n]^d} \theta_{\boldsymbol{j}}\phi_{\boldsymbol{j}}$, then

$$\begin{aligned}\|g_m\|_{\mathcal{G}_m}^2 &\asymp \sum_{\boldsymbol{j}\in[b_n]^d} \ell_n^{2\nu}\left(\frac{2\nu}{\ell_n^2} + \left\|\frac{\pi\boldsymbol{j}}{2L}\right\|_2^2\right)^\tau \theta_{\boldsymbol{j}}^2 \\ &\gtrsim \ell_n^{2\nu}\min\left\{1, \frac{2\nu}{\ell_n^2}\right\}^\tau \sum_{\boldsymbol{j}\in[b_n]^d}\left(1 + \left\|\frac{\pi\boldsymbol{j}}{2L}\right\|_2^2\right)^\tau \theta_{\boldsymbol{j}}^2.\end{aligned}$$

Equivalently,

$$\sum_{\boldsymbol{j}\in[b_n]^d}\left(1 + \left\|\frac{\pi\boldsymbol{j}}{2L}\right\|_2^2\right)^\tau \theta_{\boldsymbol{j}}^2 \lesssim \ell_n^{-2\nu}\max\left\{1, \frac{\ell_n^2}{2\nu}\right\}^\tau \|g_m\|_{\mathcal{G}_m}^2.$$

On the other hand, from equation 21,

$$\begin{aligned}D^\alpha\phi_{\boldsymbol{j}}(\boldsymbol{x}) &= D^\alpha L^{-d/2}\sum_{\boldsymbol{\delta}\in\{\pm1\}^d} C_{\boldsymbol{\delta},\boldsymbol{j}}\exp\left(\frac{\pi}{2L} i\left\langle\boldsymbol{x},\ \boldsymbol{\delta}\odot\boldsymbol{j}\right\rangle_2\right) \\ &= L^{-d/2}\sum_{\boldsymbol{\delta}\in\{\pm1\}^d} C_{\boldsymbol{\delta},\boldsymbol{j}}\prod_{p=1}^d\left(i\frac{\pi}{2L}\delta_p j_p\right)^{\alpha_p}\exp\left(\frac{\pi}{2L} i\left\langle\boldsymbol{x},\ \boldsymbol{\delta}\odot\boldsymbol{j}\right\rangle_2\right). \qquad (30)\end{aligned}$$

Since $D^\alpha\phi_{\boldsymbol{j}}(\boldsymbol{x})$ remains orthogonal over D_L, and

$$\|D^\alpha\phi_{\boldsymbol{j}}(\boldsymbol{x})\|_{L^2(D_L)}^2 \le \prod_{p=1}^d\left(\frac{\pi j_p}{2L}\right)^{2\alpha_p},$$

it follows that

$$\begin{aligned}\|D^\alpha g_m\|_{L^2(\Omega)}^2 &\le \|D^\alpha g_m\|_{L^2(D_L)}^2 \\ &\lesssim \sum_{\boldsymbol{j}\in[b_n]^d}\prod_{p=1}^d\left(\frac{\pi j_p}{2L}\right)^{2\alpha_p}\theta_{\boldsymbol{j}}^2.\end{aligned}$$

We use the definition of Sobolev norm in equation 20. For the integer part,

$$\begin{aligned}&\sum_{|\alpha|\le\lfloor\tau\rfloor}\|D^\alpha g_m\|_{L^2(\Omega)}^2 \\ &\lesssim \sum_{|\alpha|\le\lfloor\tau\rfloor}\sum_{\boldsymbol{j}\in[b_n]^d}\prod_{p=1}^d\left(\frac{\pi j_p}{2L}\right)^{2\alpha_p}\theta_{\boldsymbol{j}}^2 \\ &\lesssim \sum_{|\alpha|\le\lfloor\tau\rfloor}\sum_{\boldsymbol{j}\in[b_n]^d}\left(1 + \left\|\frac{\pi\boldsymbol{j}}{2L}\right\|_2^2\right)^{|\alpha|}\theta_{\boldsymbol{j}}^2 \\ &\lesssim \sum_{\boldsymbol{j}\in[b_n]^d}\left(1 + \left\|\frac{\pi\boldsymbol{j}}{2L}\right\|_2^2\right)^\tau\theta_{\boldsymbol{j}}^2 \\ &\lesssim \ell_n^{-2\nu}\max\left\{1, \frac{\ell_n^2}{2\nu}\right\}^\tau\|g_m\|_{\mathcal{G}_m}^2.\end{aligned}$$

If τ is not an integer, let $\rho = \tau - \lfloor \tau \rfloor$. For the fractional part,

$$
\begin{aligned}
&\sum_{|\alpha|=\lfloor\tau\rfloor} \iint_{\Omega\times\Omega} \frac{|D^\alpha g_m(\boldsymbol{x}) - D^\alpha g_m(\boldsymbol{x}')|^2}{\|\boldsymbol{x}-\boldsymbol{x}'\|_2^{2\rho+d}} \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{x}' \\
\leq &\sum_{|\alpha|=\lfloor\tau\rfloor} \iint_{D_L\times D_L} \frac{|D^\alpha g_m(\boldsymbol{x}) - D^\alpha g_m(\boldsymbol{x}')|^2}{\|\boldsymbol{x}-\boldsymbol{x}'\|_2^{2\rho+d}} \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{x}' \\
\leq &\sum_{|\alpha|=\lfloor\tau\rfloor} \int_{\|\boldsymbol{a}\|_\infty\leq 2L} \frac{\int_{D_{2L}} |D^\alpha g_m(\boldsymbol{x}'+\boldsymbol{a}) - D^\alpha g_m(\boldsymbol{x}')|^2 \mathrm{d}\boldsymbol{x}'}{\|\boldsymbol{a}\|_2^{2\rho+d}} \mathrm{d}\boldsymbol{a} \\
\overset{(1)}{\lesssim} &\sum_{|\alpha|=\lfloor\tau\rfloor} \int_{\|\boldsymbol{a}\|_\infty\leq 2L} \frac{\sum_{\boldsymbol{j}\in[b_n]^d} \theta_{\boldsymbol{j}}^2 \left\|\frac{\pi\boldsymbol{j}}{2L}\right\|_2^{2|\alpha|} \sum_{\boldsymbol{\delta}\in\{\pm1\}^d} |\exp\left(\frac{\pi}{2L} i\langle \boldsymbol{a},\ \boldsymbol{\delta}\odot\boldsymbol{j}\rangle_2\right) - 1|^2}{\|\boldsymbol{a}\|_2^{2\rho+d}} \mathrm{d}\boldsymbol{a} \\
\lesssim &\sum_{|\alpha|=\lfloor\tau\rfloor} \sum_{\boldsymbol{j}\in[b_n]^d} \theta_{\boldsymbol{j}}^2 \left(1 + \left\|\frac{\pi\boldsymbol{j}}{2L}\right\|_2^2\right)^{|\alpha|} \sum_{\boldsymbol{\delta}\in\{\pm1\}^d} \int_{\|\boldsymbol{a}\|_\infty\leq 2L} \frac{|\exp\left(\frac{\pi}{2L} i\langle \boldsymbol{a},\ \boldsymbol{\delta}\odot\boldsymbol{j}\rangle_2\right) - 1|^2}{\|\boldsymbol{a}\|_2^{2\rho+d}} \mathrm{d}\boldsymbol{a} \\
\overset{(2)}{\lesssim} &\sum_{|\alpha|=\lfloor\tau\rfloor} \sum_{\boldsymbol{j}\in[b_n]^d} \theta_{\boldsymbol{j}}^2 \left(1 + \left\|\frac{\pi\boldsymbol{j}}{2L}\right\|_2^2\right)^{|\alpha|+\rho} \\
\lesssim &\ \ell_n^{-2\nu} \max\left\{1, \frac{\ell_n^2}{2\nu}\right\}^\tau \|g_m\|_{\mathcal{G}_m}^2.
\end{aligned}
$$

Inequality (1) is from formula equation 30, which yields

$$
\begin{aligned}
&D^\alpha \phi_{\boldsymbol{j}}(\boldsymbol{x}+\boldsymbol{a}) - D^\alpha \phi_{\boldsymbol{j}}(\boldsymbol{x}) \\
= &\sum_{\boldsymbol{\delta}\in\{\pm1\}^d} C_{\boldsymbol{\delta},\boldsymbol{j}} \prod_{p=1}^d \left(i\frac{\pi}{2L}\delta_p j_p\right)^{\alpha_p} \exp\left(\frac{\pi}{2L} i\,\langle \boldsymbol{x},\ \boldsymbol{\delta}\odot\boldsymbol{j}\rangle_2\right) \left[\exp\left(\frac{\pi}{2L} i\langle \boldsymbol{a},\ \boldsymbol{\delta}\odot\boldsymbol{j}\rangle_2\right) - 1\right].
\end{aligned}
$$

For inequality (2), we use analysis from equation 25 and

$$
\begin{aligned}
&\int_{\|\boldsymbol{a}\|_\infty\leq 2L} \frac{|\exp\left(\frac{\pi}{2L} i\langle \boldsymbol{a},\ \boldsymbol{\delta}\odot\boldsymbol{j}\rangle_2\right) - 1|^2}{\|\boldsymbol{a}\|_2^{2\rho+d}} \mathrm{d}\boldsymbol{a} \\
\leq &\left\|\frac{\pi\boldsymbol{j}}{2L}\right\|_2^{2\rho} \int_{\mathbb{R}^d} \frac{|\exp\left(i\langle \boldsymbol{a}, \boldsymbol{\delta}\odot\boldsymbol{j}\ /\ \|\boldsymbol{j}\|_2\rangle_2\right) - 1|^2}{\|\boldsymbol{a}\|_2^{2\rho+d}} \mathrm{d}\boldsymbol{a} \\
\leq &\frac{\pi^{d/2}}{\Gamma(d/2)} \frac{4-3\rho}{\rho(1-\rho)} \left\|\frac{\pi\boldsymbol{j}}{2L}\right\|_2^{2\rho}.
\end{aligned}
$$

Combining the results of integer and fractional parts, we have

$$
\|g_m\|_{H^\tau(\Omega)}^2 \lesssim \ell_n^{-2\nu} \max\left\{1, \frac{\ell_n^2}{2\nu}\right\}^\tau \|g_m\|_{\mathcal{G}_m}^2.
$$

C Proof of Corollary 1

From Theorem 1, under the constraint that

$$
b_n > L, \qquad \ell_n > 0,
$$

we have

$$
\begin{aligned}
\mathcal{R}_n \lesssim &\left\{A_\ell \left(h_n^\tau B_\ell + h_n^{d/2}\lambda_n^{1/2}\right)(b_n/L)^{(\tau-s)_+} \right. \\
&\left. + \left(\lambda_n^{-1/2} B_\ell h_n^\nu + 1\right)\left[h_n^s + (L/b_n)^s\right]\right\}^2 + \sigma_\epsilon^2 \left(\lambda_n^{-1/2} B_\ell h_n^\tau + h_n^{d/2}\right)^2.
\end{aligned}
$$

Now if $\ell_n = \ell > 0$ is a fixed constant independent of n, it follows that

$$\mathcal{R}_n \lesssim \left\{ \left(h_n^\tau + h_n^{d/2} \lambda_n^{1/2} \right) (b_n/L)^{(\tau - s)_+} + \left(\lambda_n^{-1/2} h_n^\nu + 1 \right) \left[h_n^s + (L/b_n)^s \right] \right\}^2 + \sigma_\epsilon^2 \left(\lambda_n^{-1/2} h_n^\tau + h_n^{d/2} \right)^2.$$

Balancing $\lambda_n^{-1/2} h_n^\nu + 1$ and $h_n^\tau + h_n^{d/2} \lambda_n^{1/2}$ yields

$$\lambda_n \asymp h_n^{2\nu}.$$

Since $s, \tau > d/2$, the error bound becomes

$$\mathcal{R}_n \lesssim \left[h_n^\tau \, (b_n/L)^{(\tau - s)_+} + (L/b_n)^s \right]^2 + \sigma_\epsilon^2 h_n^d.$$

C.1 Noiseless case

In noiseless case, $\epsilon \equiv 0$. Solving the inequality

$$(L/b_n)^s \lesssim h_n^{\min\{s,\tau\}}, \qquad h_n^\tau (b_n/L)^{(\tau - s)_+} \lesssim h_n^{\min\{s,\tau\}},$$

it follows that

$$L h_n^{-\min\{s,\tau\}/s} \lesssim b_n \lesssim \begin{cases} L/h_n, & s < \tau, \\ +\infty, & s \geq \tau. \end{cases}$$

Equivalently,

$$L^d h_n^{-d \min\{s,\tau\}/s} \lesssim m_n \lesssim \begin{cases} (L/h_n)^d, & s < \tau, \\ +\infty, & s \geq \tau. \end{cases}$$

Since $L = B + 1$, which is a constant independent of n, the tuning

$$\lambda_n \asymp h_n^{2\nu}, \qquad h_n^{-d \min\{s,\tau\}/s} \lesssim m_n \lesssim \begin{cases} h_n^{-d}, & s < \tau, \\ +\infty, & s \geq \tau, \end{cases}$$

yields

$$\mathcal{R}_n \lesssim h_n^{2 \min\{s,\tau\}} + \sigma_\epsilon^2 h_n^d.$$

Under quasi-uniform condition in Assumption 4, $h_n \asymp n^{-1/d}$, then

$$\mathcal{R}_n \lesssim n^{-\frac{2}{d} \min\{s,\tau\}} + \sigma_\epsilon^2 n^{-1}.$$

And the tuning rule is $\ell_n = \ell$, $L = B + 1$,

$$\lambda_n \asymp h_n^{2\nu}, \qquad n^{\min\{s,\tau\}/s} \lesssim m_n \lesssim \begin{cases} n, & s < \tau, \\ +\infty, & s \geq \tau, \end{cases}$$

C.2 Noisy case

Solving the inequality

$$(L/b_n)^s \lesssim h_n^{d/2}, \qquad h_n^\tau (b_n/L)^{(\tau - s)_+} \lesssim h_n^{d/2},$$

it follows that

$$L h_n^{-d/(2s)} \lesssim b_n \lesssim L h_n^{-\frac{\nu}{(\tau - s)_+}},$$

where $h_n^{-\frac{\nu}{(\tau - s)_+}}$ is understood as $+\infty$ when $\tau \leq s$. When $h_n \asymp n^{-1/d}$ and $L = B + 1$, we require that

$$n^{d/(2s)} \lesssim m_n \lesssim \begin{cases} n^{\frac{\nu}{\tau - s}}, & s < \tau, \\ +\infty, & s \geq \tau. \end{cases}$$

This tuning yields

$$\mathcal{R}_n \lesssim (1 + \sigma_\epsilon^2) n^{-1}.$$

D PROOF OF THEOREM 2

According to Proposition 1, the HSGP approximated posterior variance is

$$\hat{\varrho}_n^2 = \eta_n \, \boldsymbol{\mu}_m^\top \boldsymbol{C}_{m,n} \boldsymbol{\mu}_m,$$

where

$$\boldsymbol{\mu}_m = \int_\Omega \boldsymbol{\phi}_m(\boldsymbol{x}) \mathrm{d}P(\boldsymbol{x}), \qquad \boldsymbol{C}_{m,n} = \left(\lambda_n \boldsymbol{W}_m^{-1} + \boldsymbol{\Phi}_n^\top \boldsymbol{\Phi}_n\right)^{-1}.$$

Let

$$\boldsymbol{\beta}_n = \boldsymbol{\Phi}_n \boldsymbol{C}_{m,n} \boldsymbol{\mu}_m.$$

Notice that

$$\boldsymbol{\mu}_m - \boldsymbol{\Phi}_n^\top \boldsymbol{\beta}_n = \lambda_n \boldsymbol{W}_m^{-1} \boldsymbol{C}_{m,n} \boldsymbol{\mu}_m,$$

by direct computation,

$$\lambda_n \boldsymbol{\mu}_m^\top \boldsymbol{C}_{m,n} \boldsymbol{\mu}_m = \left(\boldsymbol{\mu}_m - \boldsymbol{\Phi}_n^\top \boldsymbol{\beta}_n\right)^\top \boldsymbol{W}_m \left(\boldsymbol{\mu}_m - \boldsymbol{\Phi}_n^\top \boldsymbol{\beta}_n\right) + \lambda_n \|\boldsymbol{\beta}_n\|_2^2.$$

By Cauchy–Schwarz inequality, for any $g_m \in \mathcal{G}_m(\Omega)$,

$$\sup_{\|g_m\|_{\mathcal{G}_m} \le 1} \left| I(g_m) - \boldsymbol{\beta}_n^\top \boldsymbol{G}_n \right|^2 = \left(\boldsymbol{\mu}_m - \boldsymbol{\Phi}_n^\top \boldsymbol{\beta}_n\right)^\top \boldsymbol{W}_m \left(\boldsymbol{\mu}_m - \boldsymbol{\Phi}_n^\top \boldsymbol{\beta}_n\right).$$

Thus,

$$\hat{\varrho}_n^2 = \sigma_n^2 \left[\sup_{\|g_m\|_{\mathcal{G}_m} \le 1} \left| I(g_m) - \boldsymbol{\beta}_n^\top \boldsymbol{G}_n \right|^2 + \lambda_n \|\boldsymbol{\beta}_n\|_2^2 \right].$$

D.1 UPPER BOUND

From Lemma 3,

$$\hat{\varrho}_n^2 / \sigma_n^2 \lesssim \left(h_n^\tau B_\ell + h_n^{d/2} \lambda_n^{1/2} \right)^2.$$

Under the tuning rule in Corollary 1,

$$\lambda_n \asymp B_\ell^2 h_n^{2\nu},$$

then

$$\hat{\varrho}_n^2 / \sigma_n^2 \lesssim B_\ell^2 h_n^{2\tau}.$$

D.2 LOWER BOUND

We apply Corollary 1 to the constant function $f \equiv \frac{1}{\|1\|_{H^s(\Omega)}} \in H^s(\Omega)$, under the tuning rule in the corollary,

$$|1 - \mathbf{1}^\top \boldsymbol{\beta}_n| \to 0.$$

Thus, for sufficiently large n,

$$|\mathbf{1}^\top \boldsymbol{\beta}_n| \ge 1/2.$$

By Cauchy–Schwarz inequality,

$$\|\boldsymbol{\beta}_n\|_2 \ge \frac{1}{2\sqrt{n}}.$$

Thus,

$$\hat{\varrho}_n^2 / \sigma_n^2 \ge \lambda_n \|\boldsymbol{\beta}_n\|_2^2 \gtrsim \lambda_n n^{-1} \asymp B_\ell^2 h_n^{2\tau}.$$

E Proof of Lemma 1

By construction of $\Omega_{\gamma,n}$, every newly selected point satisfies

$$\operatorname{dist}(\boldsymbol{x}_{n+1}, \boldsymbol{X}_n) \geq \gamma h_n, \qquad n \geq n_0.$$

Moreover, since $\boldsymbol{X}_n \subseteq \boldsymbol{X}_{n+1}$, the fill distances are non-increasing:

$$h_{n+1} \leq h_n.$$

Hence h_n converges to some limit $h_\infty \geq 0$.

We first show that $h_\infty = 0$. Suppose, to the contrary, that $h_\infty > 0$. Then, for every $j > i \geq n_0$,

$$\|\boldsymbol{x}_{j+1} - \boldsymbol{x}_{i+1}\|_2 \geq \operatorname{dist}(\boldsymbol{x}_{j+1}, \boldsymbol{X}_j) \geq \gamma h_j \geq \gamma h_\infty.$$

Thus the sequence of newly selected points contains infinitely many points whose pairwise distances are bounded below by the positive constant γh_∞. This is impossible because Ω is bounded and its closure is compact, hence every infinite sequence in Ω has a convergent subsequence. Therefore,

$$h_n \to 0.$$

Since $n_0 \geq 2$, define

$$\delta_0 := \min_{1 \leq i < j \leq n_0} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2.$$

Since the initial design consists of distinct points, $\delta_0 > 0$. There exists $N \geq n_0 + 1$ such that

$$\gamma h_{N-1} \leq \delta_0.$$

Fix any $n \geq N$ and any two distinct points $\boldsymbol{x}_i, \boldsymbol{x}_j \in \boldsymbol{X}_n$, with $i < j$. If $j \leq n_0$, then both points belong to the initial design, and by the monotonicity of h_n,

$$\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 \geq \delta_0 \geq \gamma h_{N-1} \geq \gamma h_n.$$

If $j > n_0$, then $\boldsymbol{x}_j$ was selected after $\boldsymbol{X}_{j-1}$ had been constructed. Hence,

$$\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 \geq \operatorname{dist}(\boldsymbol{x}_j, \boldsymbol{X}_{j-1}) \geq \gamma h_{j-1} \geq \gamma h_n.$$

Therefore every pair of distinct points in $\boldsymbol{X}_n$ is separated by at least γh_n. Recalling that

$$q_n = \frac{1}{2} \min_{1 \leq i < j \leq n} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2,$$

we obtain

$$\frac{h_n}{q_n} \leq \frac{2}{\gamma}$$

for every $n \geq N$.

Finally, the standard packing and covering arguments on a bounded domain give

$$q_n \lesssim n^{-1/d}, \qquad h_n \gtrsim n^{-1/d}.$$

Combining these inequalities yields

$$h_n \asymp q_n \asymp n^{-1/d}.$$

F Proof of Proposition 2

By Lemma 1, the geometry-stabilized sequential rule generates a quasi-uniform design sequence. Hence, for all sufficiently large n,

$$h_n \asymp n^{-1/d}.$$

Since the response-independent parameter sequences (ℓ_n, λ_n, m_n) satisfy the conditions of Corollary 1, applying that corollary gives

$$\mathcal{R}_n \lesssim h_n^{2\min\{s,\tau\}} + \sigma_\epsilon^2 h_n^d.$$

Substituting $h_n \asymp n^{-1/d}$ yields

$$\mathcal{R}_n \lesssim n^{-2\min\{s,\tau\}/d} + \sigma_\epsilon^2 n^{-1},$$

which proves the proposition.

G PROOF OF PROPOSITION 3

We prove the result for $\Omega = (-B, B)^d$. Let $p_0 = (2B)^{-d}$ denote the uniform density on Ω. For a density $p \in \mathcal{P}_B$, write $\boldsymbol{X}_n(p)$ for the design generated by the stabilized sequential rule, $I_p(f) = \int_\Omega f(\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$, and $\mathcal{R}_n(p)$ for the corresponding worst-case mean-squared integration error. We use the same density-independent initial design and a fixed deterministic tie-breaking rule for all p, and assume that the feature means are evaluated exactly. We first introduce an auxiliary result in the following lemma.

Lemma 7 *Let $K \in \mathbb{N}$, and let $\mathcal{L}_1, \ldots, \mathcal{L}_K$ be real linear functionals on $C_c^\infty(\Omega)$. There exist constants $c_0 > 0, C_0 > 0$, depending only on B, d, s, such that one can find $g \in C_c^\infty(\Omega)$ satisfying*

$$\mathcal{L}_j(g) = 0, \quad 1 \le j \le K, \qquad \|g\|_\infty \le 1, \qquad \|g\|^2_{L^2(\Omega)} \ge c_0, \qquad \|g\|_{H^s(\Omega)} \le C_0(K+1)^{s/d}.$$

Proof of Lemma 7 Choose a nonzero bump function $\psi \in C_c^\infty((0,1)^d)$ with $0 \le \psi \le 1$ and use its zero extension to $\mathbb{R}^d$. Let $t = \lceil (2(K+1))^{1/d} \rceil$ and $M = t^d$. For each multi-index $\boldsymbol{u} \in \{0, \ldots, t-1\}^d$, define $\psi_{\boldsymbol{u}}(\boldsymbol{x}) = \psi\Big(\frac{t(\boldsymbol{x}+B\mathbf{1})}{2B} - \boldsymbol{u}\Big)$. Each $\psi_{\boldsymbol{u}}$ is compactly supported in a distinct grid cell of $\Omega = (-B, B)^d$. Hence the M functions $\psi_{\boldsymbol{u}}$ have mutually disjoint supports contained in Ω.

For $\boldsymbol{c} = (c_{\boldsymbol{u}}) \in \mathbb{R}^M$, let $g = g_{\boldsymbol{c}} = \sum_{\boldsymbol{u} \in \{0,\ldots,t-1\}^d} c_{\boldsymbol{u}} \psi_{\boldsymbol{u}}$, and define the $K \times M$ matrix A by $A_{j,\boldsymbol{u}} = \mathcal{L}_j(\psi_{\boldsymbol{u}})$. Consider the nonempty compact polytope $Q = \{\boldsymbol{c} \in \mathbb{R}^M : A\boldsymbol{c} = 0,\ |c_{\boldsymbol{u}}| \le 1 \text{ for every } \boldsymbol{u} \in \{0, \ldots, t-1\}^d\}$. By linearity of $\mathcal{L}_j$, $\mathcal{L}_j(g) = \sum_{\boldsymbol{u} \in \{0,\ldots,t-1\}^d} c_{\boldsymbol{u}} \mathcal{L}_j(\psi_{\boldsymbol{u}}) = (A\boldsymbol{c})_j = 0$ for every $j = 1, \ldots, K$.

At any extreme point $\boldsymbol{c}$ of Q, at least $M - \mathrm{rank}(A) \ge M - K \ge M/2$ coordinates satisfy $|c_{\boldsymbol{u}}| = 1$, where the last inequality follows from $M = t^d \ge 2(K+1)$. Otherwise, there would exist a nonzero vector supported on the unsaturated coordinates and lying in the null space of A, contradicting extremality. Hence $\sum_{\boldsymbol{u} \in \{0,\ldots,t-1\}^d} c_{\boldsymbol{u}}^2 \ge M/2$.

Since the M functions $\psi_{\boldsymbol{u}}$ have disjoint supports, we have $\|g\|_\infty \le 1$, and

$$\|g\|^2_{L^2(\Omega)} = \left(\frac{2B}{t}\right)^d \|\psi\|^2_{L^2((0,1)^d)} \sum_{\boldsymbol{u} \in \{0,\ldots,t-1\}^d} c_{\boldsymbol{u}}^2 \ge \frac{(2B)^d}{2} \|\psi\|^2_{L^2((0,1)^d)}.$$

Hence, the L^2 norm of g is bounded below uniformly in K.

For each multi-index α, the chain rule gives $D^\alpha \psi_{\boldsymbol{u}}(\boldsymbol{x}) = [t/(2B)]^{|\alpha|} (D^\alpha \psi)\big(t(\boldsymbol{x} + B\mathbf{1})/(2B) - \boldsymbol{u}\big)$. Since the supports of the M functions $\psi_{\boldsymbol{u}}$ are mutually disjoint, $\|D^\alpha g\|^2_{L^2(\Omega)} = \sum_{\boldsymbol{u}} c_{\boldsymbol{u}}^2 \|D^\alpha \psi_{\boldsymbol{u}}\|^2_{L^2(\Omega)}$. By scaling, $\|D^\alpha \psi_{\boldsymbol{u}}\|^2_{L^2(\Omega)} = t^{2|\alpha|-d}(2B)^{d-2|\alpha|} \|D^\alpha \psi\|^2_{L^2((0,1)^d)}$. Since $|c_{\boldsymbol{u}}| \le 1$ and there are $M = t^d$ coefficients, $\sum_{\boldsymbol{u}} c_{\boldsymbol{u}}^2 \le t^d$, and hence $\|D^\alpha g\|_{L^2(\Omega)} \le C_\alpha t^{|\alpha|}$, where C_α depends only on B, α, and the fixed function ψ.

In particular, $\|g\|_{L^2(\Omega)} \le C$ for some $C > 0$. Let q be an integer such that $q > s$. Using the standard integer-order Sobolev norm and the bound above for all $|\alpha| \le q$, we obtain $\|g\|_{H^q(\Omega)} \le C t^q$. The Sobolev interpolation inequality therefore gives $\|g\|_{H^s(\Omega)} \le C \|g\|^{1-s/q}_{L^2(\Omega)} \|g\|^{s/q}_{H^q(\Omega)} \le C t^s$. Since $t = \lceil [2(K+1)]^{1/d} \rceil \lesssim (K+1)^{1/d}$, we conclude that $\|g\|_{H^s(\Omega)} \le C_0 (K+1)^{s/d}$. □

Proof of Proposition 3

We prove the result by contradiction by constructing a single density $p_\star \in \mathcal{P}_B$ for which an insufficient feature budget violates the assumed rate. Suppose, to the contrary, that for a given $r > 0$, $m_n \not\gtrsim n^{rd/(2s)}$. Then $\limsup_{n\to\infty} n^r / m_n^{2s/d} = \infty$. Since $m_n \ge 1$, for every fixed k, $m_n + k \le (k+1)m_n$, so we can recursively choose an increasing sequence n_k such that

$$n_k^r / (m_{n_k} + k)^{2s/d} \ge 2^{3k}, \qquad \forall k \in \mathbb{N}. \tag{31}$$

We now recursively construct a sequence of functions $\{g_k\}_{k \ge 1}$ and probability densities $\{p^{(k)}\}_{k \ge 0}$. Set $p^{(0)} = p_0$. Suppose that $g_1, \ldots, g_{k-1}$ and $p^{(k-1)}$ have been constructed. We run the sequential

procedure in Algorithm 1 under the density $p^{(k-1)}$ through iteration n_k, and denote the resulting points by $\boldsymbol{x}_1^{(k-1)}, \ldots, \boldsymbol{x}_{n_k}^{(k-1)}$. We choose g_k to satisfy the conclusion of Lemma 7 with the following $K_k = 2m_{n_k} + k$ linear constraints:

$$\begin{aligned}
\int_\Omega g_k(\boldsymbol{x})\mathrm{d}\boldsymbol{x} &= 0, \\
\int_\Omega \phi_{\boldsymbol{j}}(\boldsymbol{x}) g_k(\boldsymbol{x})\mathrm{d}\boldsymbol{x} &= 0, \qquad \boldsymbol{j} \in \mathcal{J}_{b_{n_k}}, \\
\sum_{i=1}^{n_k} \phi_{\boldsymbol{j}}(\boldsymbol{x}_i^{(k-1)}) g_k(\boldsymbol{x}_i^{(k-1)}) &= 0, \qquad \boldsymbol{j} \in \mathcal{J}_{b_{n_k}}, \\
\int_\Omega g_k(\boldsymbol{x}) g_j(\boldsymbol{x})\mathrm{d}\boldsymbol{x} &= 0, \qquad 1 \le j < k,
\end{aligned} \tag{32}$$

and that for constants $c_0 > 0, C_1 > 0$ independent of k,

$$\|g_k\|_\infty \le 1, \qquad \|g_k\|_{L^2(\Omega)}^2 \ge c_0, \qquad \|g_k\|_{H^s(\Omega)} \le C_1 (m_{n_k} + k)^{s/d}. \tag{33}$$

Here the last bound follows from Lemma 7 since $K_k + 1 = 2m_{n_k} + k + 1 \lesssim m_{n_k} + k$.

Based on the sequence of functions $\{g_k\}_{k\ge1}$, we recursively construct a sequence of probability densities on Ω. Let $p_0(\boldsymbol{x}) = (2B)^{-d}$ be the uniform density on $\Omega = (-B, B)^d$, and set $p^{(0)} = p_0$. Define $p^{(k)}(\boldsymbol{x}) = p_0(\boldsymbol{x})\big(1 + \frac{1}{2}\sum_{j\le k} 2^{-j} g_j(\boldsymbol{x})\big)$. The first constraint in equation 32 and $\|g_j\|_\infty \le 1$ imply that $p^{(k)} \in \mathcal{P}_B$, so the recursion is well defined. Define the limiting density $p_\star(\boldsymbol{x}) = p_0(\boldsymbol{x})\big(1 + \frac{1}{2}\sum_{j=1}^{\infty} 2^{-j} g_j(\boldsymbol{x})\big)$. Then $\{p^{(k)}\}_{k\ge1}$ converges uniformly to $p_\star$, and hence $p_\star \in C(\overline{\Omega})$. Since $\int_\Omega g_k(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 0$ and $\|g_k\|_\infty \le 1$, we have $\int_\Omega p_\star(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 1$. In addition, $\sum_{j=1}^{\infty} 2^{-j} = 1$ implies that $p_0(\boldsymbol{x})/2 \le p_\star(\boldsymbol{x}) \le 3p_0(\boldsymbol{x})/2$ for all $\boldsymbol{x} \in \Omega$, i.e., $p_\star \in \mathcal{P}_B$.

For a fixed k, recall that $\boldsymbol{x}_1^{(k-1)}, \ldots, \boldsymbol{x}_{n_k}^{(k-1)}$ denote the design points generated under $p^{(k-1)}$. Since $p_\star(\boldsymbol{x}) - p^{(k-1)}(\boldsymbol{x}) = (p_0(\boldsymbol{x})/2)\sum_{j\ge k} 2^{-j} g_j(\boldsymbol{x})$, the nested structure $\mathcal{J}_{b_{n_k}} \subseteq \mathcal{J}_{b_{n_j}}$ for every $j \ge k$ and the second constraint in equation 32 imply

$$\int_\Omega \phi_{\boldsymbol{j}}(\boldsymbol{x}) \left\{ p_\star(\boldsymbol{x}) - p^{(k-1)}(\boldsymbol{x}) \right\} \mathrm{d}\boldsymbol{x} = 0, \qquad \boldsymbol{j} \in \mathcal{J}_{b_{n_k}}. \tag{34}$$

We now show by induction that the designs generated under $p_\star$ and $p^{(k-1)}$ agree through iteration n_k. Their common density-independent initial designs agree. Suppose the two designs agree through some iteration $n < n_k$. Then their fill distances, stabilized feasible sets, geometry-dependent parameters, feature matrices, and spectral weight matrices are identical. Since $\mathcal{J}_{b_n} \subseteq \mathcal{J}_{b_{n_k}}$, equation 34 also implies equality of their feature means at iteration n. Hence the acquisition functions in equation 6 agree at every candidate point, and the common tie-breaking rule selects the same next point. Induction therefore gives $\boldsymbol{X}_{n_k}(p_\star) = \boldsymbol{X}_{n_k}(p^{(k-1)})$.

Define the functions $f_k = \frac{g_k}{C_1(m_{n_k}+k)^{s/d}}$, so that $\|f_k\|_{H^s(\Omega)} \le 1$ by equation 33. By the third constraint in equation 32, $\boldsymbol{\Phi}_{n_k}^\top \big(f_k(\boldsymbol{x}_1), \ldots, f_k(\boldsymbol{x}_{n_k})\big)^\top = 0$ along the trajectory generated under $p_\star$. Using equation 3 and the mean-zero observation noise therefore gives $\mathbb{E}_\epsilon\{\widehat{I}_{m_{n_k}}(f_k)\} = 0$. On the other hand, by the first and last constraints in equation 32, $\{g_k\}_{k\ge1}$ are pairwise orthogonal in $L^2(\Omega)$ and each has zero integral. Hence, for each $k \ge 1$,

$$\begin{aligned}
I_{p_\star}(f_k) &= \int_\Omega f_k(\boldsymbol{x}) p_\star(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= \frac{1}{C_1(2B)^d(m_{n_k}+k)^{s/d}} \left\{ \int_\Omega g_k(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \frac{1}{2}\sum_{j=1}^{\infty} 2^{-j} \int_\Omega g_k(\boldsymbol{x}) g_j(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \right\} \\
&= \frac{\|g_k\|_{L^2(\Omega)}^2}{C_1 2^{k+1}(2B)^d(m_{n_k}+k)^{s/d}}.
\end{aligned}$$

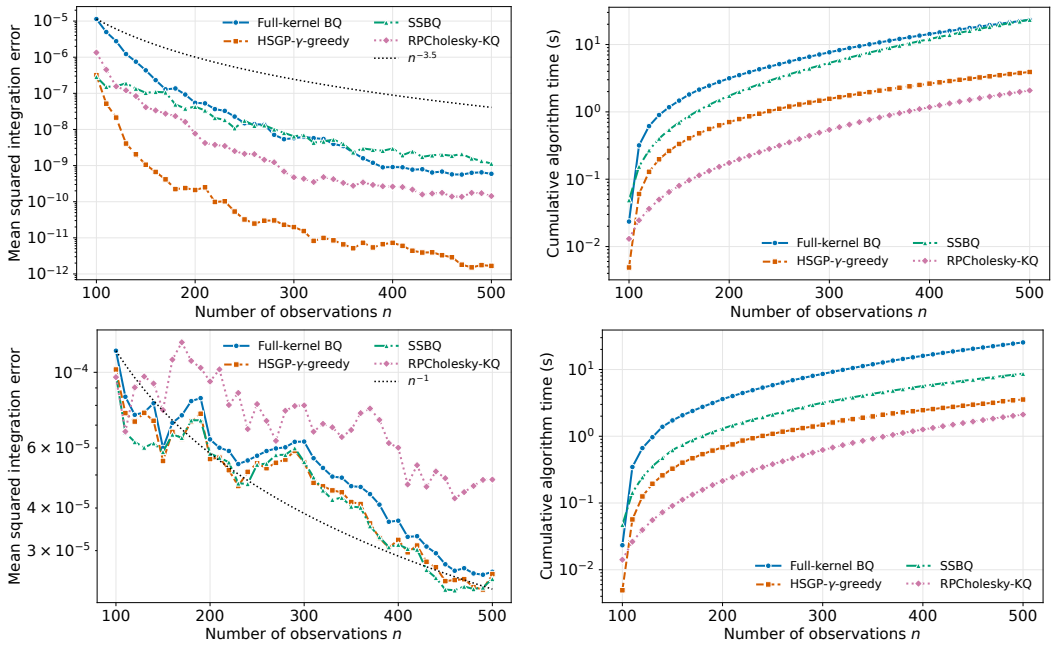


Figure 4: Matérn-5/2 simulation with $q = 3.5$. *(Top)* Noiseless observations; *(Bottom)* noise standard deviation 0.1. *(Left)* Mean squared integration error; *(Right)* cumulative algorithm time.

Since the mean-squared error is bounded below by the squared bias and $\mathbb{E}_\epsilon\{\widehat{I}_{m_{n_k}}(f_k)\} = 0$, using equation 31 and equation 33, we have that as $k \to \infty$,

$$n_k^r \mathcal{R}_{n_k}(p_\star) \geq n_k^r \{I_{p_\star}(f_k)\}^2 \geq \frac{c_0^2 n_k^r}{2^{2k+2}(2B)^{2d} C_1^2 (m_{n_k}+k)^{2s/d}} \geq \frac{c_0^2}{4(2B)^{2d} C_1^2} \cdot 2^k \to \infty.$$

This contradicts the assumed bound $\mathcal{R}_n(p_\star) \leq C_{p_\star} n^{-r}$ for all sufficiently large n. Therefore, $m_n \gtrsim n^{rd/(2s)}$ must hold.

For the noiseless benchmark, taking $r = 2\min\{s,\tau\}/d$ gives $m_n \gtrsim n$ when $s \leq \tau$ and $m_n \gtrsim n^{\tau/s}$ when $s > \tau$, matching equation 9. For fixed nonzero noise, taking $r = 1$ gives $m_n \gtrsim n^{d/(2s)}$, matching equation 10.

H Additional Numerical Results

H.1 Extra sequential experiments

We conduct the Matérn-5/2 experiment with test function of different smoothness. Let $q = 3.5$, $d = 2$, then $s = q + 1/2 - 10^3 > q = \tau$. So the kernel is undersmoothed relative to the integrand. We again sequentially add design points from the initial $n = 100$ points up to $n = 500$, use $R = 20$ replicates, $\gamma = 0.8$, $L = B + 1 = 2$, and update parameters every ten observations. Under Algorithm 1, the hyper-parameters (C_m, c_λ) are $(1.5, 0.1)$ in the noiseless setting and $(25, 300)$ in the noisy setting. We test both noiseless and noisy cases. The other settings are the same as Section 4.

Figure 4 reports the Matérn-5/2 experiments in the undersmoothing regime $s > \tau$. The qualitative performance observed in Section 4 for the oversmoothing regime $s < \tau$ remains unchanged after crossing the smoothness boundary. In the noiseless undersmoothing experiment, HSGP-γ-greedy achieves the lowest terminal mean squared integration error. RPCholesky-KQ is faster at 2.08 seconds, but its error, 1.43×10^{-10}, is approximately two orders of magnitude larger than that of HSGP-γ-greedy. With noisy observations, Full-kernel BQ, HSGP-γ-greedy, and SSBQ reach a similar noise-limited accuracy, while HSGP-γ-greedy is 2.4 times faster than SSBQ and 7.2 times

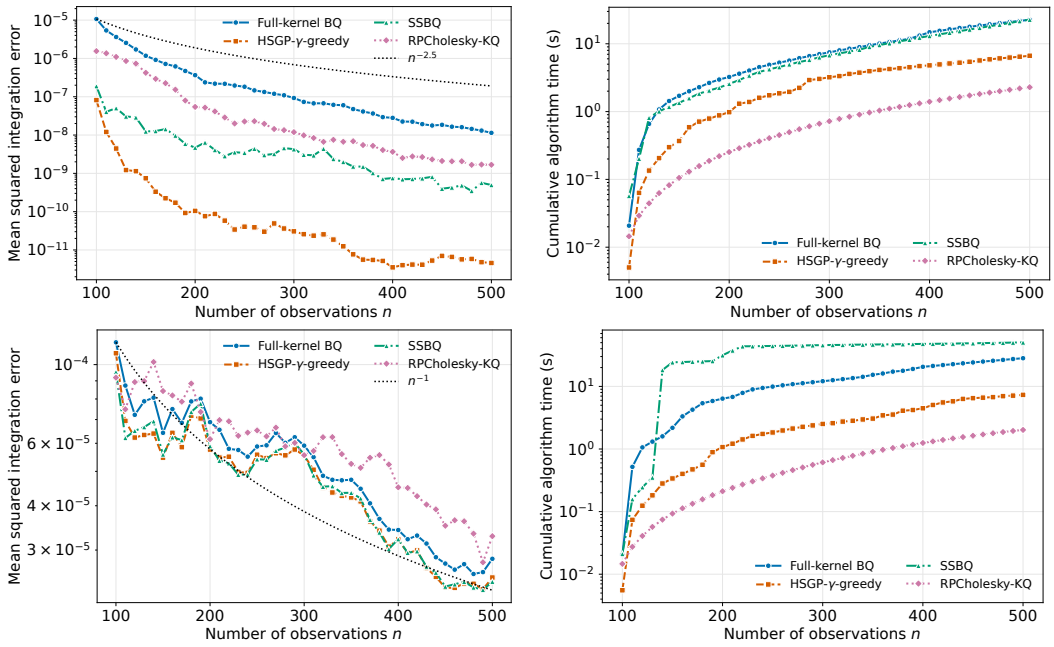


Figure 5: Matérn-$3/2$ simulation with $q = 2.5$. *(Top)* Noiseless observations; *(Bottom)* noise standard deviation 0.1. *(Left)* Mean squared integration error; *(Right)* cumulative algorithm time.

faster than Full-kernel BQ. Although RPCholesky-KQ has the lowest runtime, its terminal error is nearly twice that of HSGP-γ-greedy. Together with the oversmoothing results in Section 4, these experiments show that the performance of HSGP-γ-greedy is robust to whether the Matérn kernel over- or undersmooths the integrand.

To assess the robustness of the empirical comparison to kernel smoothness, we additionally conduct simulation using a Matérn-$3/2$ kernel, which is less smooth than the integrand. Figure 5 shows the results of Matérn-$3/2$ simulation experiments. Here $\tau = 1.5 + d/2 = 2.5 = q < s$. Replacing the Matérn-$5/2$ kernel with the less smooth Matérn-$3/2$ kernel does not change the qualitative comparison among the methods. Under noiseless setting, at $n = 500$, HSGP-γ-greedy achieves a mean squared integration error of 4.57×10^{-12} with a mean cumulative computational time of 6.65 seconds. Its error is approximately 2.5×10^3, 1.1×10^2, and 3.7×10^2 times smaller than those of Full-kernel BQ, SSBQ, and RPCholesky-KQ, respectively. Meanwhile, HSGP-γ-greedy is about 3.4 times faster than both Full-kernel BQ and SSBQ. Although RPCholesky-KQ has a lower computational cost, its integration error is more than two orders of magnitude larger. We also consider a noisy setting in which independent Gaussian noise with standard deviation 0.1 is added to each observation. In this case, the relevant error reference rate is the noise-limited rate n^{-1}. At $n = 500$, HSGP-γ-greedy achieves a mean squared integration error of 2.51×10^{-5} with a mean cumulative computational time of 7.32 seconds. It is approximately 3.9 and 6.9 times faster than Full-kernel BQ and SSBQ, respectively. RPCholesky-KQ is faster than HSGP-γ-greedy but has an approximately 1.31 times larger error.

These results indicate that the strong performance of HSGP-γ-greedy is not specific to a particular Matérn kernel: it retains the most favorable accuracy–computational cost trade-off for both the Matérn-$3/2$ and Matérn-$5/2$ kernels.

H.2 Contraction of the HSGP posterior integral variance

We empirically examine the contraction of the posterior integral variance characterized by Theorem 2. For each saved HSGP-BQ replicate and each evaluation size n, we reconstruct the normalized

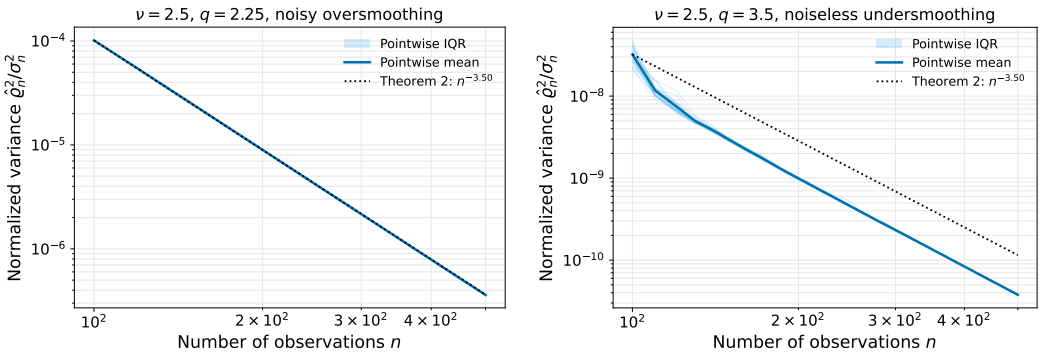


Figure 6: Contraction of the normalized HSGP posterior integral variance. *(Left)* Noisy oversmoothing with $q = 2.25$ and $\sigma_\epsilon = 0.1$. *(Right)* Noiseless undersmoothing with $q = 3.5$. The solid blue curve is the pointwise mean. The dotted line shows the $n^{-3.5}$ reference slope predicted by Theorem 2 and is vertically aligned with the mean at $n = 100$. The shaded region is the pointwise 0.25-0.75 interquartile range.

posterior integral variance as

$$\frac{\widehat{\varrho}_n^2}{\sigma_n^2} = \lambda_n \boldsymbol{\mu}_{m_n}^\top \left(\lambda_n \boldsymbol{W}_{m_n}^{-1} + \boldsymbol{\Phi}_n^\top \boldsymbol{\Phi}_n\right)^{-1} \boldsymbol{\mu}_{m_n}.$$

This calculation uses the design points and parameter sequences saved during the original experiments and therefore does not require rerunning the sequential acquisition procedure.

We consider the two main settings: noisy oversmoothing with $q = 2.25$, $s < q + 1/2 < \tau$, and $\sigma_\epsilon = 0.1$, and noiseless undersmoothing with $q = 3.5$, $s > q = \tau$, and $\sigma_\epsilon = 0$. In both settings, $d = 2$, $\nu = 5/2$, $\tau = \nu + d/2 = 3.5$, $B = 1$, $L = 2$, $\ell_n = 0.5$, and $\sigma_n^2 = 1$. The sequential sampling data is from the Matérn-$5/2$ simulation in Section 4 and Appendix H.1. Each experiment contains 20 replicates with $n_{\text{init}} = 100$ and $n_{\text{total}} = 500$.

Under quasi-uniform sampling, Theorem 2 gives

$$\frac{\widehat{\varrho}_n^2}{\sigma_n^2} \lesssim h_n^{2\tau} \asymp n^{-2\tau/d}.$$

For the present Matérn-$5/2$ kernel in dimension $d = 2$, the predicted reference rate is therefore

$$n^{-2\tau/d} = n^{-7/2}.$$

As shown in Figure 6, the two mean curves closely follow the $n^{-7/2}$ reference slope, despite the different integrand regularities, noise levels, and feature-budget schedules. This behavior is consistent with Theorem 2. Once the feature and regularization conditions are satisfied, the contraction of the normalized posterior integral variance is controlled primarily by the kernel regularity τ and the geometric resolution of the design, rather than by the smoothness of the particular integrand. The narrow interquartile bands further indicate that the observed contraction is stable across the 20 independently initialized sequential designs.