Generalized Bayesian Quadrature with Spectral Kernels

Abstract

Bayesian probabilistic integration, or Bayesian quadrature (BQ), has arisen as a popular means of numerical integral estimation with quantified uncertainty for problems where computational cost limits data availability. BQ leverages flexible Gaussian processes (GPs) to model an integrand which can be subsequently analytically integrated through properties of Gaussian distributions. However, BQ is inherently limited by the fact that the method relies on the use of a strict set of kernels for use in the GP model of the integrand, reducing the flexibility of the method in modeling varied integrand types. In this paper, we present spectral Bayesian quadrature, a form of Bayesian quadrature that allows for the use of *any* shift-invariant kernel in the integrand GP model while still maintaining the analytical tractability of the integral posterior, increasing the flexibility of BQ methods to address varied problem settings. Additionally our method enables integration with respect to a uniform expectation, effectively computing definite integrals of challenging integrands. We derive the theory and error bounds for this model, as well as demonstrate GBQ's improved accuracy, flexibility, and data efficiency, compared to traditional BQ and other numerical integration methods, on a variety of quadrature problems.

1 INTRODUCTION

Methods for estimation of non-analytical integrals through numerical methods play a key role across a broad spectrum of scientific fields, but these methods are often computationally expensive in nature. Methods such as finite-element or volumes, which are widely used in physical simulation to integrate partial differential equations, or Monte Carlo estimation, which is widely used in Bayesian statistics for estimation of posteriors, require a large number of function evaluations to reach a desired level of accuracy. In addition, many numerical integration methods fail to provide uncertainty quantification on their estimates, which is crucial in the applied settings in which physical simulation is often used.

Bayesian quadrature (BQ) (Diaconis, 1988; O'Hagan, 1991) is a probabilistic method which can remedy these concerns by offering performance on computationally-limited small data while admitting robust uncertainty bounds. BQ takes the form of a traditional quadrature rule:

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

for n evaluations of the function f, where weights w_i are instead learned through manipulation of a Bayesian non-parametric Gaussian process (GP) (C. E. Rasmussen and Williams, 2006) model on observations of the integrand $f(\boldsymbol{x})$.

The use of such a Bayesian non-parametric model for learning weights leverages the ability for GPs to perform well under data-scarcity as well as quantify uncertainty in a principled manner. In addition, the Gaussian nature of this model allows for the integral estimate of f to be a simple analytical integration of the GP prior on f using well-known characteristics of multi-variate Gaussian distributions. Previous work (Ghahramani and C. Rasmussen, 2003; Kandasamy, Schneider, and Póczos, 2015) has clearly demonstrated computational efficiency of BQ versus traditional methods such as Monte Carlo integration when the data dimensionality $d < 10$.

A chief advantage of using GPs in any probabilistic learning setting is the flexibility of choice of the GP kernel function k, which allows for a practitioner to inject domain knowledge of the problem space into the GP model. Characteristics such as as data smoothness or periodicity can easily be applied through choice or composition of specific kernel

Submitted to the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022). **To be used for reviewing only**.

functions tailored to these settings.

However, the traditional BQ formulation hamstrings this flexibility by limiting the choice of kernel in the integrand GP to only a small subset of kernels with known analytical kernel means, such as Gaussian or polynomial kernels. For well-known kernels that may not be analytically tractable in the BQ setting, but nonetheless might better model an integrand, traditional numerical quadrature methods must be used, reducing the computational efficiency that BQ offers. The question naturally arises of how practitioners might enable the full suite of kernel choices for use in the GP integrand prior while still maintaining the analytical tractability in the BQ setting, to most efficiently produce an accurate estimate to the integral of f.

In this paper, we expand on the literature of BQ and propose a solution to the problem of kernel choice with generalized Bayesian quadrature (GBQ), a method derived from random Fourier features (RFFs) by which any shift-invariant kernel can be used in the GP integrand prior while still allowing for analytical tractability in the BQ setting. By allowing for both kernel flexibility and analytical integration, we expand upon the ability of traditional BQ to model a variety of integrand types while still maintaining the computational efficiency BQ offers. We summarize our contributions here:

Contributions

- We propose generalized Bayesian quadrature (GBQ), a method of Bayesian quadrature that allows for the use of *any* shift-invariant kernel in the GP model of the integrand while still admitting an analytical estimate of the integral posterior mean and variance.
- We show that GBQ can directly be used to compute integrals over Gaussian and uniform measures within the same framework.
- We derive the upper-bounded error to this approximation as a function of data-availability.
- We demonstrate the accuracy and flexibility of this quadrature method versus traditional BQ, as well as data-efficiency versus typical Monte Carlo integration, on a selection of relevant domain problems.

2 RELATED WORKS

Quadrature methods of the type in equation 1 are well-studied due to their importance to a variety of fields, and there is a deep literature dating back centuries on methods for numerically approximating integrals. We will briefly review here relevant methods in relation to Bayesian quadrature.

Rather than deterministic quadrature weighting, various probabilistic quadrature approaches have been proposed (C. J. Oates and Sullivan, 2019) for integration when model observations are expensive, with one of the most popular methods being Bayesian quadrature. Many extensions to vanilla BQ have been developed over the years to increase performance and provide theoretical guarantees (Acerbi, 2018; Belhadji, Bardenet, and Chainais, 2019; Briol, Chris J. Oates, Girolami, and M. A. Osborne, 2015; Kennedy, 1998). Other applications include use in multi-fidelity modeling (Gessner, Gonzalez, and Mahsereci, 2020), Bayesian posterior estimation (Gunter et al., 2014; M. Osborne, Garnett, Roberts, et al., 2012), Bayesian optimization (Nguyen et al., 2020), and model selection (Chai et al., 2019; M. Osborne, Garnett, Ghahramani, et al., 2012).

The derivation of analytical forms, or empirical approximation, of kernel means, which is a significant component of the BQ formulation, is a problem that appears in numerous other fields. Namely, kernel mean embedding (Muandet et al., 2017), deep Gaussian processes (Damianou and Lawrence, 2013), and neural operators (Kovachki et al., 2021; Li et al., 2021) all attempt to do so through various means. There also exist empirical methods for the estimation of kernel means using random Fourier features (Muandet et al., 2017), as well as strong theoretical connections between the very concept of kernel-based quadrature and random Fourier features (Bach, 2017). In a related manner, methods have been proposed that seek to implement Fourier feature kernels through quadrature based methods (Mutny and Krause, 2018). However, to our knowledge, no methods directly solve kernel integrals analytically in the BQ setting using RFFs, as we propose to do in this paper.

The method that shares the most overlap with this work is Fourier neural operators (FNOs) (Li et al., 2021), which, as a part of a larger deep neural network architecture, estimate the convolution of shift-invariant kernels with a probability measure using parameters in Fourier space. While we take a similar approach to deriving kernel means using Fourier frequencies, the overall frameworks differ, with GBQ existing in the Gaussian process framework, thus offering uncertainty estimates for integral posteriors, while FNO's exist within a deterministic neural network architecture.

3 PRELIMINARIES

3.1 BAYESIAN QUADRATURE

We will now review various preliminary methods upon which GBQ is built, starting with Bayesian quadrature.

BQ assumes we have a function f that we are trying to integrate and a dataset $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^n$ with n noisy observations of f, where $\boldsymbol{x} \in \mathcal{R}^d$, $y_i = f(\boldsymbol{x}_i) + \epsilon$, and ϵ is i.i.d normal distributed noise. Typically, f is computationally expensive to evaluate, implying a small n and highlighting the need for uncertainty estimation in the final integral approximation. BQ does this by first placing a Gaussian process

(C. E. Rasmussen and Williams, 2006) prior on f, which we will briefly review here.

Gaussian Processes Gaussian processes are a Bayesian non-parametric method which model the target data generation function f we are attempting to learn as a joint multivariate Gaussian of the form:

$$f \sim \mathcal{GP}(\boldsymbol{\mu}(\boldsymbol{x}), k_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{x}')), \tag{2}$$

$$\boldsymbol{y} = f(\boldsymbol{x}) + \epsilon, \tag{3}$$

where $k_{\boldsymbol{\theta}}$ is a positive semi-definite *kernel function* with hyper-parameters $\boldsymbol{\theta}$, and $\boldsymbol{\mu}$ is a mean function. In the above, we assume an additive and independent Gaussian noise observation model with, $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, where $\boldsymbol{y}$ are noisy observations with standard deviation σ. k is typically chosen *a-priori* to encode known characteristics of the data $\mathcal{D}$ such as periodicity and smoothness.

For inference, the posterior-predictive distribution of f_* for a new data point $\{\boldsymbol{x}_*\}$, given the training data $\mathcal{D} = \{\boldsymbol{x}^i, y^i\}_{i=1}^n$, and Gram matrix $\boldsymbol{K}_{\boldsymbol{xx}} = k_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{x}')$, $\forall\, \boldsymbol{x}, \boldsymbol{x}'$, is given by $\mathcal{N}(\boldsymbol{\mu}(f*), \mathrm{Cov}(f*))$ where,

$$\mu(f*) = \boldsymbol{K}_{*x}(\boldsymbol{K}_{xx} + \sigma^2 \boldsymbol{I})^{-1}\boldsymbol{y}, \tag{4}$$

$$\mathrm{Cov}(f*) = \boldsymbol{K}_{**} - \boldsymbol{K}_{*x}(\boldsymbol{K}_{xx} + \sigma^2 \boldsymbol{I})^{-1}\boldsymbol{K}_{x*}. \tag{5}$$

In BQ, by setting a GP prior on the integrand f we can leverage the ability of GPs flexibly and accurately model functions with uncertainty on small data, but it is also advantageous in that we can directly and analytically integrate the integrand GP prior. This is performed using well-known characteristics of Gaussian distributions in order to form a posterior estimate $\langle \bar{f} \rangle$ of the integral of f.

Formally, the mean of the BQ estimate of $\langle \bar{f} \rangle$ is the expected value over measure $p(\boldsymbol{x})$ of the posterior mean of the GP prior (4) on f:

$$\begin{aligned} \langle \bar{f} \rangle &= \int_{\boldsymbol{x} \in \mathcal{R}} k(\boldsymbol{x}, \boldsymbol{X})^T \boldsymbol{K}^{-1} \boldsymbol{y}\, p(\boldsymbol{x})\, d\boldsymbol{x} \\ &= \boldsymbol{y}^T \boldsymbol{K}^{-1} \int_{\boldsymbol{x} \in \mathcal{R}} k(\boldsymbol{x}, \boldsymbol{X})\, p(\boldsymbol{x})\, d\boldsymbol{x} \\ &= \mu_{\boldsymbol{x}}(\boldsymbol{X})^T \boldsymbol{K}^{-1} \boldsymbol{y}, \end{aligned} \tag{6}$$

where $\mu_{\boldsymbol{x}}(\boldsymbol{X}) = [\mu_{\boldsymbol{x}}(\boldsymbol{x}_1) \dots \mu_{\boldsymbol{x}}(\boldsymbol{x}_n)]$ can be seen as the *kernel mean* over measure $p(\boldsymbol{x})$. The variance of this estimate is:

$$\mathbb{V}(\langle \bar{f} \rangle) = \int_{\boldsymbol{X} \in \mathcal{R}^d} \mu_{\boldsymbol{x}}(\boldsymbol{X}) p(\boldsymbol{X})\, d\boldsymbol{X} \tag{7}$$

which is notably independent of prior observations $\boldsymbol{X}$.

The mean formulation mirrors that of standard quadrature methods shown in equation (1), differing in that weights $\mu_{\boldsymbol{x}}(\boldsymbol{X})^T \boldsymbol{K}^{-1}$ are the result of probabilistic learning on observed data $\mathcal{D}$ and associated kernel choice, rather than decided *a priori* or by a heuristic.

Under a very limited selection of kernel and sampling measure choices, the mean (6) and variance (7) can be calculated analytically (Briol, Chris J. Oates, Girolami, M. A. Osborne, and Sejdinovic, 2019). Most commonly, a Gaussian kernel and Gaussian distribution for the measure $p(\boldsymbol{x})$, as proposed by (O'Hagan, 1991), is one such case. It is also prudent to note that the measure distribution can be fluid while retaining analytical tractability through use of importance sampling (Briol, Chris J. Oates, Cockayne, et al., 2017; Ghahramani and C. Rasmussen, 2003), while the choice of kernel is more restricted.

In BQ, the limitation of the kernel to certain forms dependent on known closed-form analytical integration over the measure $p(\boldsymbol{x})$ gives up one of the greatest advantages of the GP prior: flexible selection of kernels for specific domains. To alleviate this issue, GBQ introduces random Fourier features into the BQ formulation for parametrization of the GP kernel.

3.2 RANDOM FOURIER FEATURES

As we shall see in Section 4 Random Fourier features enable the use of *any* shift-invariant kernel in the BQ-GP prior without sacrificing the analytical tractability of the integral posterior. This greatly increases the flexibility of the BQ to perform under a variety of problem conditions for which different kernels may be necessary.

Random Fourier features are obtained from the spectral representation of shift-invariant kernels given by Bochner's theorem:

Theorem 1 (Bochner's theorem (Rudin, 2011))**.** *A shift-invariant kernel $k(\boldsymbol{x}, \boldsymbol{x}') = k(\boldsymbol{x} - \boldsymbol{x}')$ is positive-definite if and only if it is the Fourier transform of a non-negative measure.*

Theorem 1 is the building block upon which (Rahimi and Recht, 2008) introduce random Fourier features (RFFs), which define a practical means by which Bochner's theorem can be applied in practice to estimate kernel functions in finite dimensions. Using the derivation from (Rahimi and Recht, 2008), if the probability density $p(\boldsymbol{\omega})$ is the Fourier transform of k:

$$\begin{aligned} k(\boldsymbol{x} - \boldsymbol{x}') &= \int_{\mathcal{R}^d} p(\boldsymbol{\omega}) e^{j\boldsymbol{\omega}(\boldsymbol{x} - \boldsymbol{x}')}\, d\boldsymbol{\omega}, \\ &= \int_{\mathcal{R}^d} p(\boldsymbol{\omega}) \cos(\boldsymbol{\omega}(\boldsymbol{x} - \boldsymbol{x}'))\, d\boldsymbol{\omega}. \end{aligned} \tag{8}$$

For brevity, equation (8) provides the formulation for the case that the kernel and data $\boldsymbol{x}$ are real-valued, but an alternative formulation exists for the case they are not.

It can be easily seen that the kernel function k is entirely defined by the choice of density $p(\boldsymbol{\omega})$, and several common kernels have known associated densities. For example, if

$p(\boldsymbol{\omega})$ is multivariate isotropic Gaussian, then (8) represents the radial basis function (RBF) kernel. By drawing from the associated $p(\boldsymbol{\omega})$ for our choice of kernel, RFFs approximate (8) with Monte Carlo by:

$$k(\boldsymbol{x}, \boldsymbol{x}') = k(\boldsymbol{x} - \boldsymbol{x}') \approx \frac{1}{R} \sum_{r=1}^{R} \cos(\boldsymbol{\omega}_r^T (\boldsymbol{x} - \boldsymbol{x}')) \quad (9)$$

where R is the number of Monte Carlo samples or *Fourier features*.

Alternatively, we can directly parametrize these features $\boldsymbol{\omega}$ as GP hyperparameters, which allows for optimal kernels to learned during training to best adapt to specific problem settings (Chang et al., 2017; Oliva et al., 2016; Tompkins et al., 2019; Zhen et al., 2020).

4 GENERALIZED BAYESIAN QUADRATURE

We build upon these concepts to devise our method, generalized Bayesian quadrature, which enables flexible Bayesian quadrature for use with any arbitrary shift-invariant kernel while maintaining analytical tractability of the kernel mean $\mu_{\boldsymbol{x}}(\boldsymbol{X})$. We begin by showing that a Gaussian density can be approximated with RFFs, which will lead to analytical tractability for general shift-invariant kernels.

4.1 PROBABILITY DENSITY FUNCTIONS AS RFF KERNELS

Analytical tractability of the BQ mean in (6) for any kernel represented by RFFs can be achieved by reformulating the kernel mean measure $p(\boldsymbol{x})$ as an RFF as well. In general, we can turn any positive-definite probability density function $p : \mathcal{X} \to [0, \infty)$ on $\mathcal{X} \subseteq \mathbb{R}^d$ into a kernel via the following construction:

$$\begin{aligned} &k_p : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}, \\ &k_p(\boldsymbol{x}, \boldsymbol{x}') = \longmapsto \begin{cases} p(\boldsymbol{x} - \boldsymbol{x}')\,, & \boldsymbol{x} - \boldsymbol{x}' \in \mathcal{X}, \\ 0\,, & \boldsymbol{x} - \boldsymbol{x}' \notin \mathcal{X}\,. \end{cases} \end{aligned} \quad (10)$$

It is easy to verify that a kernel defined as in Equation 10 is translation-invariant and positive-definite whenever p is. As examples of distributions with positive-definite densities we have the Gaussian and the Student-T (Rossberg, 1995).

RFF Representation of the Gaussian Given that an RBF kernel represents an un-normalized Gaussian, by sampling $\boldsymbol{\rho}$ from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and using a multivariate Gaussian normalizing constant $Z^{-1} = [(2\pi)^d |\Sigma|]^{-1/2}$, we can formulate a RFF kernel approximation of a Gaussian density function $q(\boldsymbol{x})$ as $\lim_{r\to\infty}$ as:

$$\begin{aligned} p(\boldsymbol{x}) \approx q(\boldsymbol{x}) &= Z^{-1} \exp\{-|\boldsymbol{x} - \boldsymbol{\mu}|^2\} \\ &\approx [(2\pi)^d |\boldsymbol{\Sigma}|]^{-1/2} \frac{1}{R} \sum_{r=1}^{R} \cos(\boldsymbol{\rho}_r^T (\boldsymbol{x} - \boldsymbol{\mu})). \end{aligned} \quad (11)$$

This form allows for the use of simple trigonometric identities to form an analytically integrable kernel mean formulation (6) over a Gaussian measure, which we will shortly demonstrate.

4.2 GENERALIZED BAYESIAN QUADRATURE POSTERIOR

We now reformulate the BQ mean and variance by substituting the RFF formulations of both the kernel and measure in equations (9) and (11) into the BQ mean in equation (6).

$$\begin{aligned} \langle \bar{f} \rangle = \boldsymbol{y}^T \boldsymbol{K}^{-1} \int_{\boldsymbol{x} \in \mathcal{R}} &\frac{1}{R} \sum_{r=1}^{R} \cos(\boldsymbol{\omega}_r^T (\boldsymbol{x} - \boldsymbol{X})) \\ &\times [(2\pi)^d |\boldsymbol{\Sigma}|]^{-1/2} \frac{1}{Z} \sum_{z=1}^{Z} \cos(\boldsymbol{\rho}_z^T (\boldsymbol{x} - \boldsymbol{\mu})) d\boldsymbol{x} \end{aligned} \quad (12)$$

The trigonometric form of both the kernel and measure distribution in this setting allow for the application of basic identities to rewrite the integrand as a linear function. Using the identity $\cos(\alpha)\cos(\beta) = \cos(\alpha+\beta)/2 + \cos(\alpha-\beta)/2$, we arrive at the following definition.

Definition 1 (Generalized Bayesian Quadrature)**.** *Given n noisy observations $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^n$ of a function f with $\boldsymbol{x} \in \mathcal{R}^d$, a kernel function k parametrized through random Fourier frequencies $\boldsymbol{\omega} \in \mathcal{R}^{R \times d}$ sampled from density p, kernel matrix $\boldsymbol{K} = [k(\boldsymbol{x}_i, \boldsymbol{x}'_j)]_{i,j=1}^n \in \mathbb{R}^{n \times n}$, and Fourier frequencies $\boldsymbol{\rho} \in \mathcal{R}^{Z \times d}$ sampled from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, we can formulate a mean estimate $\langle \bar{f} \rangle$ over a Gaussian measure of the indefinite integral of f, as:*

$$\begin{aligned} \langle \bar{f} \rangle = L \times \sum_{r=1}^{R} \sum_{z=1}^{Z} \Bigg[&\frac{h^d(\boldsymbol{x}^T(\boldsymbol{\omega}_r + \boldsymbol{\rho}_z) - (\boldsymbol{\omega}_r^T \boldsymbol{X} + \boldsymbol{\rho}_z^T \boldsymbol{\mu}))}{2 \prod_{j=1}^d (\omega_r^j + \rho_z^j)} \\ &+ \frac{h^d(\boldsymbol{x}^T(\boldsymbol{\omega}_r - \boldsymbol{\rho}_z) - (\boldsymbol{\omega}_r^T \boldsymbol{X} - \boldsymbol{\rho}_z^T \boldsymbol{\mu}))}{2 \prod_{j=1}^d (\omega_r^j - \rho_z^j)} \Bigg] + c \end{aligned} \quad (13)$$

$$L = \frac{\boldsymbol{y}^T \boldsymbol{K}^{-1}}{RZ[(2\pi)^d |\boldsymbol{\Sigma}|]^{1/2}} \quad (14)$$

where d is the dimensionality of $\boldsymbol{x}$, c is a constant of integration, and h^d is the function at the d-th index of the repeating series $h = [\sin, -\cos, -\sin, \cos, \sin, \dots]$.

Over closed bounds $\int_{\boldsymbol{a}}^{\boldsymbol{b}} f(\boldsymbol{x})\, d\boldsymbol{x}$, we multiply L by a truncation term $(\hat{\Phi}(\boldsymbol{b}) - \hat{\Phi}(\boldsymbol{a}))^{-1}$, where $\hat{\Phi}$ is an RFF estimate to the multivariate Gaussian CDF, which is calculable from Eq (11) [1].

In the case measure $p(\boldsymbol{x})$ is uniform, the indefinite integral estimate is:

$$\langle \bar{f} \rangle = \frac{1}{R} \sum_{r=1}^{R} \frac{h^d(\boldsymbol{\omega}^T(\boldsymbol{x}' - \boldsymbol{x}))}{\prod_{j=1}^{d} \omega_r^j}\,, \tag{15}$$

where h^d is defined as above.

See the appendix for full proof as well as accompanying variance derivation. Under definition 1 we obtain an analytical posterior for $\langle \bar{f} \rangle$ and $\mathbb{V}(\langle \bar{f} \rangle)$ that allows for flexible kernel choice through the use of RFFs.

4.3 APPROXIMATION ERROR

4.3.1 Gaussian Process and Random Fourier Features Error Bounds

The approximation error of GBQ extends from well-known error bounds derived from the literature of RFFs and BQ respectively. We present here an abbreviated form of this proof, the full version of which can be found in the supplement.

We begin with the following lemma outlining the error of the GP estimate $\langle \bar{f} \rangle$ to the integrand f under the assumption f is a member of the Hilbert space $\mathcal{H}_k$ defined by kernel k:

Lemma 1 (Durand, Maillard, and Pineau (2017, Theorem 1))**.** *Assume $f \in \mathcal{H}_k$ and that the observation noise ϵ is σ_ϵ-sub-Gaussian. Then the following holds with probability at least $1 - \delta$:*

$$\forall n \in \mathbb{N}, |f(\boldsymbol{x}) - \mu_n(\boldsymbol{x})| \leq \beta_k(\delta)\sigma_n(\boldsymbol{x}), \forall \boldsymbol{x} \in \mathcal{X}\,, \tag{16}$$

where μ_n and σ_n^2 denote the GP posterior mean and variance given n observations, according to (4) *and 5, respectively, and*

$$\beta_k(\delta) := \|f\|_k + \sigma_\epsilon \sqrt{\frac{2}{\lambda} \log\left(\frac{\det(\mathbf{I} + \lambda^{-1}\mathbf{K}_n)^{1/2}}{\delta}\right)}, \tag{17}$$

with

$$\mathbf{K}_n := [k(\boldsymbol{x}_i, \boldsymbol{x}_j')]_{i,j=1}^{n} \in \mathbb{R}^{n \times n}\,, \tag{18}$$

We follow with a lemma related to the error bounds on the RFF approximation to a shift-invariant kernel k.

[1] See supplementary for derivation.

Lemma 2 (Sutherland and Schneider (2015, Proposition 1))**.** *Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a continuous shift-invariant positive-definite kernel with $k(\boldsymbol{x}, \boldsymbol{x}) = 1$ and such that $\nabla^2 k(\boldsymbol{x}, \boldsymbol{x})$ exists, for all $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d$. Suppose $\mathcal{X}$ is compact with diameter $\ell_{\mathcal{X}} < \infty$. Denote k's Fourier transform as P_k, which is a probability measure, and let $\sigma_k^2 := \mathbb{E}[\|\boldsymbol{\omega}\|_2^2]$ for $\omega \sim P_k$. Let $\tilde{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ denote k's RFF approximation with R frequencies according to* (9)*. Then the following holds for any $0 < \xi < \sigma_k \ell_{\mathcal{X}}$:*

$$\mathbb{P}\left[\sup_{\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}} |\tilde{k}(\boldsymbol{x}, \boldsymbol{x}') - k(\boldsymbol{x}, \boldsymbol{x}')| \geq \xi\right] \leq 66 \left(\frac{\sigma_k \ell_{\mathcal{X}}}{\xi}\right)^2 \exp\left(-\frac{R\xi^2}{4(d+2)}\right). \tag{19}$$

Therefore, for any $\delta \in (0, 1)$, we can achieve pointwise approximation error less than ξ with probability at least $1 - \delta$ if:

$$R \geq R(\xi, \delta, \sigma_k) := \frac{4(d+2)}{\xi^2} \left(\frac{2}{1 + \frac{2}{d}} \log \frac{\sigma_k \ell_{\mathcal{X}}}{\xi} + \log \frac{66}{\delta}\right) \tag{20}$$

4.3.2 Generalized Bayesian Quadrature Error

Next, we formulate an error bound on the RFF parametrization of the Gaussian (or any arbitrary) density shown in equation (11), as we build towards a final bound on GBQ.

Theorem 2 (Error of the RFF Density Approximation)**.** *Let $p : \mathcal{X} \to \mathbb{R}$ be a positive-definite probability density function defined on $\mathcal{X} \subset \mathbb{R}^d$ which is such that $\nabla^2 p(\mathbf{0})$ exists. Assume $\mathcal{X}$ is compact, and let $b_p > 0$ be any constant such that $b_p \geq \max_{\boldsymbol{x} \in \mathcal{X}} p(\boldsymbol{x})$. Let $\tilde{k}_p$ denote an RFF approximation with $Z \in \mathbb{N}$ frequencies to k_p as defined in* (10)*, and let $\tilde{p} : \boldsymbol{x} \mapsto \tilde{k}_p(\boldsymbol{x}, \mathbf{0})$, $\boldsymbol{x} \in \mathcal{X}$. Then, for any $\xi > 0$, the following holds:*

$$\mathbb{P}\left[\sup_{\boldsymbol{x} \in \mathcal{X}} |\tilde{p}(\boldsymbol{x}) - p(\boldsymbol{x})| \geq b_p \xi\right] \leq 66 \left(\frac{\sigma_{k_p} \ell_{\mathcal{X}}}{\xi}\right)^2 \exp\left(-\frac{Z_p \xi^2}{4(d+2)}\right) \tag{21}$$

where for the second statement we assume $\xi \leq \sigma_{k_p} \ell_{\mathcal{X}}$, and σ_{k_p}, $\ell_{\mathcal{X}}$, α_ξ and β_ξ are the same as defined in Lemma 2 for $k := \frac{1}{b_p} k_p$.

Finally, we combine these results to arrive at the upper bounded error for GBQ as a composition of the errors of GP approximation, RFF approximation, and RFF measure density estimation.

Theorem 3 (Upper-Bounded Generalized Bayesian Quadrature Error)**.** *Let $f \in \mathcal{H}_k$, where $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive-definite, translation-invariant kernel on $\mathcal{X} \subset \mathbb{R}^d$. Assume that:*

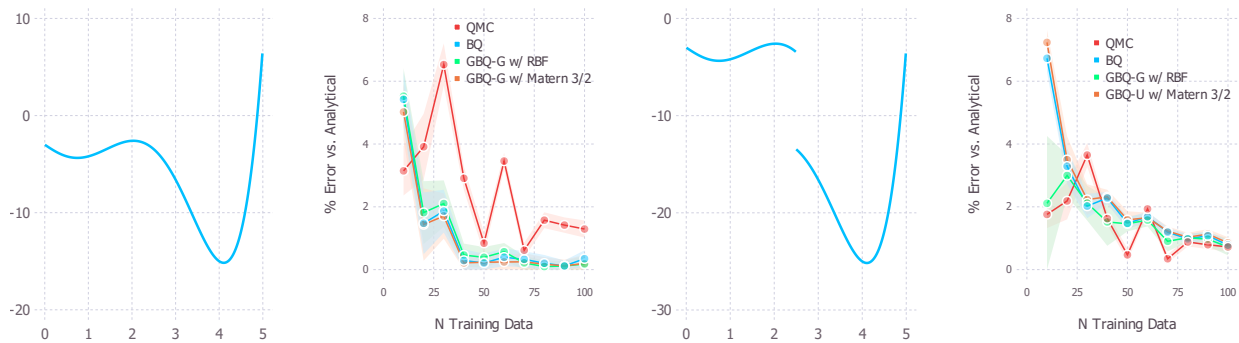

Figure 1: Function Plots and Bounded Error Graphs for 1D Continuous and Disjoint Polynomial Quadrature Experiments.

1. *$\mathcal{X}$ is compact with diameter $\ell_{\mathcal{X}} < \infty$ and volume $v_{\mathcal{X}} := \int_{\mathcal{X}} \mathrm{d}\boldsymbol{x} < \infty$;*
2. *$k(\mathbf{0}, \mathbf{0}) = 1$ and $\nabla^2 k(\mathbf{0}, \mathbf{0})$ exists;*
3. *and $p : \mathcal{X} \to [0, \infty)$ is a positive-definite probability density function.*

Then the following holds with probability at least $1 - \delta$:

$$\left| \int_{\mathcal{X}} f(\boldsymbol{x}) p(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} - \int_{\mathcal{X}} \tilde{\mu}_n(\boldsymbol{x}) \tilde{p}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \right| \leq \left(\frac{n}{\lambda} \beta_\epsilon \left(\frac{\delta}{4} \right) \xi_k + \beta_k \left(\frac{\delta}{4} \right) \max_{\boldsymbol{x} \in \mathcal{X}} \sigma_n(\boldsymbol{x}) \right) \times (1 + b_p \xi_p v_{\mathcal{X}}) + \|f\|_\infty b_p \xi_p v_{\mathcal{X}} \,, \quad (22)$$

where $\beta_\epsilon(\delta) := \|f\|_\infty + \sigma_\epsilon \sqrt{2 \log \left(\frac{n}{\delta} \right)}$, for an RFF approximation to k with $R \geq R\left(\xi_k, \frac{\delta}{4}, \sigma_k\right)$ frequencies and an RFF approximation to p with $Z \geq R\left(\xi_p, \frac{\delta}{4}, \sigma_{k_p}\right)$ frequencies, given $\xi_k > 0$ and $\xi_p > 0$.

We refer the reader to the supplementary for the full proof of theorems 2 and 3.

5 EXPERIMENTS

We demonstrate here the empirical results of GBQ compared to traditional Monte Carlo quadrature methods and BQ. Specifically, we measure percent error versus the analytical integral solution, with baselines of Monte Carlo (MC) integration, quasi Monte Carlo (QMC) using Halton sequence sampling (Halton, 1960), and BQ with the RBF kernel and a Gaussian measure.

For GBQ, we present results in the form of GBQ-Measure-Kernel, where the kernel is chosen from the RFF estimates to the RBF, Matérn 1/2 (M1/2), Matérn 3/2 (M3/2), and Matérn 5/2 (M5/2), and the measure is either uniform (U) or Gaussian (G). We hold static the number of integrand observations $f(\boldsymbol{x})$ available across all baselines and GBQ models. Additionally, we use the same GP kernel hyperparameters $\boldsymbol{\theta}$ in both BQ and GBQ, which are initialized at reasonable values according to the problem setting. For Fourier features $\boldsymbol{\omega}$ and $\boldsymbol{\rho}$ in equations (9) and (11), we sample using Halton sequences as well to produce a smoother coverage of the sample space, and we use 100 features in the 1D experiments and 300 in the 2D experiments. Finally, we implement these methods in Julia (Bezanson et al., 2015), and code has been made available [2].

We note that while our experiments consider the employment of the Matérn family of kernels, any shift-invariant kernel can be used in the GBQ integrand prior to adapt to a wide array of problem settings. While there are various analytical solutions to the Matérn family in traditional BQ, they require a kernel-specific integral to be calculated and implemented, and don't exist over all measures $p(\boldsymbol{x})$. We provide evidence to the flexibility of our method by showing that Matérn kernels can be implemented without change of problem formulation by simply sampling features $\boldsymbol{\omega}$ according to the appropriate frequency distribution.

For all experiments, at each training size n we report results as the average model-wise results over multiple runs under different random seeds, and include information on the error variance over runs. While experiments were run for all kernel-measure combinations for GBQ, for brevity we include here only those models that performed best on a given experiment.

5.1 1D EXPERIMENTS

Our first experiment is a simple 1D polynomial to empirically verify our theoretical results of section 4 regarding the efficacy of the GBQ method in both recreating results of traditional BQ using the RBF kernel as well as demonstrate the flexibility of kernel choice that GBQ offers.

[2] https://anonymous.4open.science/r/GBQ/

We model the integral of a polynomial of the form:

$$f(x) = 0.2x^3(x-4)^2 - 3x - 3\,, \tag{23}$$

in the first case, and disjointed version of the polynomial

$$f(x) = \begin{cases} 0.2x^3(x-4)^2 - 3x - 3, & x < 2.5\,, \\ 0.2x^3(x-4)^2 - 3x - 13, & x \geq 2.5\,, \end{cases} \tag{24}$$

in the second. The choice of the disjoint polynomial is in order to assess the value of the flexibility of GBQ in enabling varied kernel choice in BQ, and in this case we leverage Matérn kernels, which typically perform better than the RBF on non-smooth data. We run each experiment 10 times under different seeds at each n and report the aggregated mean and 95% confidence bounds in figure 1.

In the first experiment, which represents a smoother polynomial, BQ and GBQ both outperform QMC in accuracy as a function of data scarcity. We can see that GBQ-G-RBF is an excellent approximation to BQ, which similarly leverages an RBF kernel over a Gaussian measure, which helps to validate our theoretical results on both the accuracy of the RFF-based integration of the RFF-RBF kernel over a Gaussian measure, as well as the ability for RFFs to parametrize Gaussian distributions.

In the disjoint case, we see that at low n, GBQ has a slight advantage over BQ when using the Matérn kernel, but that results converge for all methods as training size increases. While QMC achieves better error at some points, it generally displays more variance over n in this experiment than the BQ and GBQ-based models.

5.2 2D EXPERIMENTS

We now move to a selection of 2D experiments, first of which is estimating the integral of a polynomial of the form:

$$f(x,y) = -0.005x^4 * 0.1x^3 + y^5(0.02x - 0.08) - 0.001y^2 + 0.2y + 0.5 \tag{25}$$

over the interval $x \in [-4, 4], y \in [-2.5, 2.5]$, as well as a disjoint 2D function:

$$f(x,y) = \begin{cases} e^{5x+5y}, & x < 0.5, y < 0.5\,, \\ 0, & x \geq 2.5, y \geq 2.5\,, \end{cases} \tag{26}$$

over the unit cube.

We perform both experiments over a range of training data sizes from 10 to 1000 n, with 5 runs per n at different random seeds. Plots of these functions can be seen in figure 2, and the means and standard deviations of the results are reported in tables 1 and 2.

In both experiments, we see that GBQ methods have universally lower mean error than QMC and BQ. The most performant kernel varies across n, but in several cases we see that the Matérn has the lowest error, supporting the case that flexibility of kernel choice is a valuable addition to the BQ method when considering both different integrand types as well as available training data.

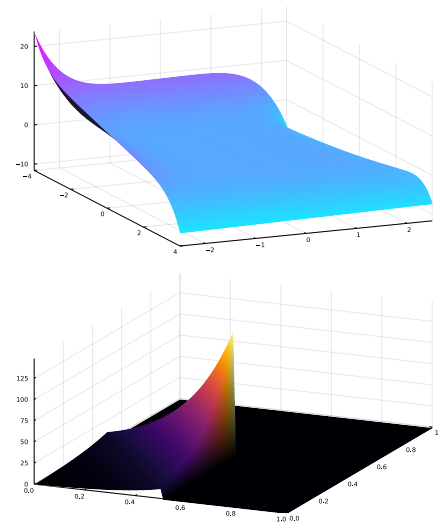

Figure 2: Plots of 2D experiment equations (25) (top) and (26) (bottom).

In the disjoint polynomial experiment, we intentionally include GBQ-G with the RBF (the BQ equivalent) in table 2, even though it was not high performing among the GBQ methods, to demonstrate the potential performance enhancement GBQ offers through kernel choice. We see GBQ-G-RBF track closely with BQ, while GBQ-U with the Matérn 1/2 and GBQ-U-RBF in combination perform better at all n, and frequently with implied worst-case error bounds well below the BQ mean error.

An interesting experimental result was the importance of consistent methodology used for solving the kernel mean $\boldsymbol{\mu_x}(\boldsymbol{X})$ and producing the kernel matrix $\boldsymbol{K}$, when applied in the BQ posterior mean formulation (6). Anecdotally, we found that using the combination of a kernel mean derived from traditional BQ and a kernel that was estimated through RFFs, and vice-versa, produced significantly unstable posterior integral mean estimates. These results suggest the benefit of using the full-stack GBQ method with RFF parametrization of both the kernel and measure distribution in order to achieve the best experimental results.

6 DISCUSSION

In this paper, we have introduced generalized Bayesian quadrature, a method for performing Bayesian quadrature using any shift-invariant kernel while maintaining posterior tractability. We derive the upper bound on the error of this approximation, while also demonstrating the practical bene-

Table 1: 2D Polynomial of Equation (25) Integration Results (% Error)

n	QMC	BQ	GBQ-U RBF	GBQ-G RBF	GBQ-G M5/2
10	98.78 ± 7.23	8.57 ± 6.77	17.03 ± 9.06	10.27 ± 5.32	$\mathbf{4.88 \pm 3.73}$
25	76.57 ± 16.34	9.69 ± 7.45	$\mathbf{8.32 \pm 7.16}$	8.53 ± 7.39	11.08 ± 10.63
50	44.92 ± 5.7	7.81 ± 2.64	14.77 ± 2.6	7.33 ± 3.07	$\mathbf{5.72 \pm 5.22}$
100	31.02 ± 3.46	4.02 ± 3.5	$\mathbf{1.97 \pm 0.88}$	4.04 ± 2.93	2.41 ± 1.71
250	7.97 ± 1.6	1.22 ± 1.13	$\mathbf{1.03 \pm 0.93}$	2.14 ± 0.77	1.86 ± 1.1
500	6.07 ± 0.85	0.68 ± 0.63	$\mathbf{0.49 \pm 0.53}$	1.34 ± 1.6	1.56 ± 1.65
750	5.51 ± 0.65	0.73 ± 0.26	$\mathbf{0.48 \pm 0.38}$	1.22 ± 1.24	1.35 ± 1.21
1000	3.94 ± 0.46	0.41 ± 0.26	$\mathbf{0.36 \pm 0.26}$	1.41 ± 1.36	1.52 ± 1.34

Table 2: 2D Disjoint Polynomial of Equation (26) Integration Results (% Error)

n	QMC	BQ	GBQ-U RBF	GBQ-U M1/2	GBQ-G RBF
10	164.04 ± 0.34	38.42 ± 0.72	$\mathbf{8.26 \pm 3.82}$	95.64 ± 12.34	30.79 ± 3.68
25	20.28 ± 0.75	10.59 ± 0.75	$\mathbf{2.64 \pm 1.0}$	5.06 ± 4.49	10.17 ± 0.95
50	23.38 ± 0.28	26.08 ± 0.3	17.42 ± 0.7	$\mathbf{12.96 \pm 10.34}$	27.14 ± 0.69
100	26.8 ± 0.24	38.93 ± 0.23	25.06 ± 0.3	$\mathbf{5.92 \pm 7.4}$	38.26 ± 0.28
250	4.41 ± 0.16	11.99 ± 0.16	$\mathbf{2.74 \pm 0.33}$	2.99 ± 2.03	12.01 ± 0.28
500	3.48 ± 0.09	12.63 ± 0.09	3.46 ± 0.12	$\mathbf{2.08 \pm 0.85}$	12.68 ± 0.1
750	3.24 ± 0.07	12.38 ± 0.07	3.01 ± 0.06	$\mathbf{2.02 \pm 0.58}$	12.24 ± 0.1
1000	0.86 ± 0.05	9.62 ± 0.05	$\mathbf{0.61 \pm 0.05}$	0.73 ± 0.18	9.48 ± 0.08

fits on a selection of quadrature problems when compared to traditional numerical integration methods and baseline BQ.

More broadly, we note the wider applicability of the methods proposed in this paper. Our chief theoretical contribution comes within the framework of Bayesian quadrature, but in essence it is providing the analytical solution to a kernel mean when the kernel and measure distribution are approximated by RFFs. However, kernel means have a wide array of use cases as discussed in 2, and represent fertile ground for future applications of our theoretical results.

Additionally, as part of the process of applying GBQ over closed-bounds in multiple dimensions, raising necessity for a truncation term composed of multivariate cumulative distribution functions, we devised a method to parametrize distributions using RFFs and analytically integrate this estimate in order to produce a CDF. For many distributions which offer no closed-form multivariate CDF, this method might be of use.

Future research may look into these applications as well as extending the flexibility and computational aspects of the method. Potential extensions include learning the RFF kernel through its spectral density, leveraging low-rank GP posteriors for computational efficiency improvements in kernel matrix inversion in the BQ mean, and composing multiple levels of GBQ together into deeper architectures for applications to highly nonlinear problems. The introduction proposed in this paper has demonstrated both theoretical and empirical promise that will provide a solid launching point for these pursuits.

REFERENCES

Acerbi, Luigi (2018). "Variational Bayesian Monte Carlo." In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc.

Bach, Francis (Jan. 1, 2017). "On the equivalence between kernel quadrature rules and random feature expansions." In: *The Journal of Machine Learning Research* 18.1, pp. 714–751.

Belhadji, Ayoub, Rémi Bardenet, and Pierre Chainais (Dec. 31, 2019). "Kernel quadrature with DPPs." In: *arXiv:1906.07832 [cs, stat]*. arXiv: 1906.07832.

Bezanson, Jeff et al. (July 19, 2015). "Julia: A Fresh Approach to Numerical Computing." In: *arXiv:1411.1607 [cs]*. arXiv: 1411.1607.

Briol, François-Xavier, Chris J. Oates, Jon Cockayne, et al. (Aug. 6, 2017). "On the sampling problem for Kernel Quadrature." In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. Sydney, NSW, Australia: JMLR.org, pp. 586–595.

Briol, François-Xavier, Chris J. Oates, Mark Girolami, and Michael A. Osborne (Dec. 6, 2015). "Frank-Wolfe Bayesian Quadrature: Probabilistic Integration with Theoretical Guarantees." In: *arXiv:1506.02681 [stat]*. arXiv: 1506.02681.

Briol, François-Xavier, Chris J. Oates, Mark Girolami, Michael A. Osborne, and Dino Sejdinovic (Feb. 2019).

"Probabilistic Integration: A Role in Statistical Computation?" In: *Statistical Science* 34.1, pp. 1–22.

Chai, Henry et al. (May 24, 2019). "Automated Model Selection with Bayesian Quadrature." In: *Proceedings of the 36th International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, pp. 931–940.

Chang, Wei-Cheng et al. (May 23, 2017). "Data-driven Random Fourier Features using Stein Effect." In: *arXiv:1705.08525 [cs, stat]*. arXiv: 1705.08525.

Damianou, Andreas and Neil D. Lawrence (Apr. 29, 2013). "Deep Gaussian Processes." In: *Artificial Intelligence and Statistics*. Artificial Intelligence and Statistics. PMLR, pp. 207–215.

Diaconis, P. (1988). "Bayesian Numerical Analysis." In.

Durand, Audrey, Odalric-Ambrym Maillard, and Joelle Pineau (Aug. 2, 2017). "Streaming kernel regression with provably adaptive mean, variance, and regularization." In: *arXiv:1708.00768 [cs, stat]*. arXiv: 1708.00768.

Gessner, Alexandra, Javier Gonzalez, and Maren Mahsereci (Aug. 6, 2020). "Active Multi-Information Source Bayesian Quadrature." In: *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*. Uncertainty in Artificial Intelligence. PMLR, pp. 712–721.

Ghahramani, Zoubin and Carl Rasmussen (2003). "Bayesian Monte Carlo." In: *Advances in Neural Information Processing Systems*. Vol. 15. MIT Press.

Gunter, Tom et al. (Nov. 3, 2014). "Sampling for Inference in Probabilistic Models with Fast Bayesian Quadrature." In: *arXiv:1411.0439 [stat]*. arXiv: 1411.0439.

Halton, J. H. (Dec. 1, 1960). "On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals." In: *Numerische Mathematik* 2.1, pp. 84–90.

Kandasamy, Kirthevasan, Jeff Schneider, and Barnabás Póczos (July 25, 2015). "Bayesian active learning for posterior estimation." In: *Proceedings of the 24th International Conference on Artificial Intelligence*. Buenos Aires, Argentina: AAAI Press, pp. 3605–3611.

Kennedy, Marc (Dec. 1, 1998). "Bayesian quadrature with non-normal approximating functions." In: *Statistics and Computing* 8.4, pp. 365–375.

Kovachki, Nikola et al. (Dec. 16, 2021). "Neural Operator: Learning Maps Between Function Spaces." In: *arXiv:2108.08481 [cs, math]*. arXiv: 2108.08481.

Li, Zongyi et al. (May 16, 2021). "Fourier Neural Operator for Parametric Partial Differential Equations." In: *arXiv:2010.08895 [cs, math]*. arXiv: 2010.08895.

Muandet, Krikamol et al. (2017). "Kernel Mean Embedding of Distributions: A Review and Beyond." In: *Foundations and Trends® in Machine Learning* 10.1, pp. 1–141. arXiv: 1605.09522.

Mutny, Mojmir and Andreas Krause (2018). "Efficient High Dimensional Bayesian Optimization with Additivity and Quadrature Fourier Features." In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc.

Nguyen, Thanh et al. (June 3, 2020). "Distributionally Robust Bayesian Quadrature Optimization." In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. International Conference on Artificial Intelligence and Statistics. PMLR, pp. 1921–1931.

O'Hagan, A. (Nov. 1, 1991). "Bayes–Hermite quadrature." In: *Journal of Statistical Planning and Inference* 29.3, pp. 245–260.

Oates, C. J. and T. J. Sullivan (Nov. 1, 2019). "A modern retrospective on probabilistic numerics." In: *Statistics and Computing* 29.6, pp. 1335–1351.

Oliva, Junier B. et al. (May 2, 2016). "Bayesian Nonparametric Kernel-Learning." In: *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*. Artificial Intelligence and Statistics. PMLR, pp. 1078–1086.

Osborne, Michael, Roman Garnett, Zoubin Ghahramani, et al. (2012). "Active Learning of Model Evidence Using Bayesian Quadrature." In: *Advances in Neural Information Processing Systems*. Vol. 25. Curran Associates, Inc.

Osborne, Michael, Roman Garnett, Stephen Roberts, et al. (Mar. 21, 2012). "Bayesian Quadrature for Ratios." In: *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*. Artificial Intelligence and Statistics. PMLR, pp. 832–840.

Rahimi, Ali and Benjamin Recht (2008). "Random Features for Large-Scale Kernel Machines." In: *Advances in Neural Information Processing Systems*. Vol. 20. Curran Associates, Inc.

Rasmussen, Carl Edward and Christopher K. I. Williams (2006). *Gaussian processes for machine learning*. Cambridge, Mass: MIT Press. 248 pp.

Rossberg, H. -J. (Aug. 1, 1995). "Positive definite probability densities and probability distributions." In: *Journal of Mathematical Sciences* 76.1, pp. 2181–2197.

Rudin, Walter (2011). *Fourier Analysis on Groups*. Hoboken: John Wiley & Sons.

Sutherland, Danica J. and Jeff Schneider (July 12, 2015). "On the error of random fourier features." In: *Proceedings*

of the Thirty-First Conference on Uncertainty in Artificial Intelligence. Arlington, Virginia, USA: AUAI Press, pp. 862–871.

Tompkins, Anthony et al. (Apr. 11, 2019). "Black Box Quantiles for Kernel Learning." In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. The 22nd International Conference on Artificial Intelligence and Statistics. PMLR, pp. 1427–1437.

Zhen, Xiantong et al. (Nov. 21, 2020). "Learning to Learn Kernels with Variational Random Features." In: *Proceedings of the 37th International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, pp. 11409–11419.

Generalized Bayesian Quadrature with Spectral Kernels: Supplementary Materials

A Generalized Bayesian Quadrature Derivation

A.1 GBQ Integral Mean Over Uniform Measures

We start by deriving the general solution to integrals of the form:

$$\begin{aligned}\langle f \rangle &= \int_{\boldsymbol{x}\in\mathcal{R}^d} \frac{1}{R}\sum_{r=1}^{R} \cos(\boldsymbol{\omega}_r^T(\boldsymbol{x}-\boldsymbol{X}))\, d\boldsymbol{x} \\ &\overset{\dagger}{=} \frac{1}{R}\sum_{r=1}^{R} \int_{\boldsymbol{x}\in\mathcal{R}^d} \cos(\boldsymbol{\omega}_r^T(\boldsymbol{x}-\boldsymbol{X}))\, d\boldsymbol{x}\end{aligned} \tag{1}$$

where $\boldsymbol{x} \in \mathcal{R}^d$ and $\boldsymbol{\omega} \in \mathcal{R}^{R\times d}$, and $\dagger$ makes use of the fact that $\int_{x\in\mathcal{R}} x + x\, dx = \int_{x\in\mathcal{R}} x + \int_{x\in\mathcal{R}} x$. Equation (1) represents the integral of a RFF estimated kernel over a uniform measure, or the uniform kernel mean. Using u substitution to integrate out a single x^j variable from the vector-valued $\boldsymbol{x}$ results in:

$$\langle f \rangle_{x^j} = \frac{1}{R}\sum_{r=1}^{R} \int_{\boldsymbol{x}^i\in\mathcal{R}^{d-1}} \frac{\sin(\boldsymbol{\omega}_r^T(\boldsymbol{x}-\boldsymbol{X}))}{\omega_r^j}\, d\boldsymbol{x}^i. \tag{2}$$

where

$$\boldsymbol{x}^i := \begin{bmatrix} x^1 \\ \vdots \\ x^{j-1} \\ x^{j+1} \\ \vdots \\ x^d \end{bmatrix} \tag{3}$$

If we integrate (2) again over a new variable x^t in $\boldsymbol{x}^i$, we start to see a general pattern emerge:

$$\langle f \rangle_{x^j x^t} = \frac{1}{R}\sum_{r=1}^{R} \int_{\boldsymbol{x}^{i'}\in\mathcal{R}^{d-2}} \frac{-\cos(\boldsymbol{\omega}_r^T(\boldsymbol{x}-\boldsymbol{X}))}{\omega_r^j\omega_r^t}\, d\boldsymbol{x}^{i'}. \tag{4}$$

The repeated integral of the cos follows a repeating pattern through $h = [\sin, -\cos, -\sin, \cos, \sin, \dots]$, while the integral of u-substituted $\boldsymbol{\omega}^T(\boldsymbol{x}-\boldsymbol{X})$ simply results in the multiplication of the integrand denominator by w_r^j for each variable x^j in $\boldsymbol{x}$ we integrate over. Thus, integrating over the entirety of $\boldsymbol{x} \in \mathcal{R}^d$ will result in an indefinite integral of the form :

$$\langle f \rangle = \frac{1}{R}\sum_{r=1}^{R} \frac{h^d(\boldsymbol{\omega}_r^T(\boldsymbol{x}-\boldsymbol{X}))}{\prod_{j=1}^{d}\omega_r^j} \tag{5}$$

where h is defined as the repeating series above and h^d represents the d-th index of h. For a RFF kernel parametrized by $\boldsymbol{\omega}$, equation (5) represents the indefinite uniform expectation expectation; in other words, the uniform kernel mean, which can now be substituted back into the BQ mean formulation to produce the GBQ integral mean over a uniform measure:

$$\langle \bar{f} \rangle = \frac{\boldsymbol{y}^T \boldsymbol{K}^{-1}}{R} \sum_{r=1}^{R} \frac{h^d(\boldsymbol{\omega}_r^T(\boldsymbol{x} - \boldsymbol{X}))}{\prod_{j=1}^{d} \omega_r^j} \tag{6}$$

A.2 GBQ Integral Mean Over Uniform Measures

For the variance, we leverage the BQ variance formulation, which simply involves solving for the expectation of (5) over $p(\boldsymbol{x}')$:

$$\mathbb{V}(\hat{f}) = \int_{\boldsymbol{x} \in \mathcal{R}^d} \int_{\boldsymbol{X} \in \mathcal{R}^d} k(\boldsymbol{x}, X)\, p(\boldsymbol{x}) p(\boldsymbol{X})\, d\boldsymbol{x} d\boldsymbol{X}\,. \tag{7}$$

We then substitute (5) for $\int_{\boldsymbol{x} \in \mathcal{R}^d} k(\boldsymbol{x}, \boldsymbol{X})\, p(\boldsymbol{x})$ in (7)

$$\begin{aligned} \mathbb{V}(\langle \bar{f} \rangle) &= \int_{\boldsymbol{X} \in \mathcal{R}^d} \frac{1}{R} \sum_{r=1}^{R} \frac{h^d(\boldsymbol{\omega}_r^T(\boldsymbol{x} - \boldsymbol{X}))}{\prod_{j=1}^{d} \omega_r^j} p(\boldsymbol{X}) d\boldsymbol{X} \\ &= \frac{1}{R} \sum_{r=1}^{R} \int_{\boldsymbol{X} \in \mathcal{R}^d} \frac{h^d(\boldsymbol{\omega}_r^T(\boldsymbol{x} - \boldsymbol{X}))}{\prod_{j=1}^{d} \omega_r^j} d\boldsymbol{X} \end{aligned} \tag{8}$$

Using the same techniques as the mean derivation in the previous section, we can easily arrive at the indefinite form of the variance estimate:

$$\mathbb{V}(\langle \bar{f} \rangle) = \frac{1}{R} \sum_{r=1}^{R} \frac{-1^d h^{2d}(\boldsymbol{\omega}_r^T(\boldsymbol{x} - \boldsymbol{X}))}{\prod_{j=1}^{d} \omega_r^j \omega_r^j} \tag{9}$$

where h^{2d} is the $2d$-th index of h as previously defined, and the term -1^d is introduced due to the fact that $\boldsymbol{x}'$ is negative in the integrand.

A.3 RFF Kernel Means Over Gaussian Measures

We now turn our attention to integrals of the form:

$$\begin{aligned} \langle f \rangle &= \int_{\boldsymbol{x} \in \mathcal{R}^d} \frac{1}{R} \sum_{r=1}^{R} \cos(\boldsymbol{\omega}_r^T(\boldsymbol{x} - \boldsymbol{X})) p(\boldsymbol{x})\, d\boldsymbol{x} \\ &= \frac{1}{R} \sum_{r=1}^{R} \int_{\boldsymbol{x} \in \mathcal{R}^d} \cos(\boldsymbol{\omega}_r^T(\boldsymbol{x} - \boldsymbol{X})) p(\boldsymbol{x})\, d\boldsymbol{x}\,, \end{aligned} \tag{10}$$

where $p(\boldsymbol{x})$ is a Gaussian. Equation 10 represents the RFF estimate to the kernel expectation over a Gaussian distribution, or the Gaussian kernel mean.

As in the main paper, we parametrize $p(\boldsymbol{x}')$ as an RFF approximation to the multivariate Gaussian, and can rewrite 10 as:

$$\begin{aligned} \langle \bar{f} \rangle &= \boldsymbol{y}^T \boldsymbol{K}^{-1} \int_{\boldsymbol{x} \in \mathcal{R}^d} \frac{1}{R} \sum_{r=1}^{R} \cos(\boldsymbol{\omega}_r^T(\boldsymbol{x} - \boldsymbol{X})) \times \frac{1}{Z|(2\pi)^d \boldsymbol{\Sigma}|^{1/2}} \sum_{z=1}^{Z} \cos(\boldsymbol{\rho}_z^T(\boldsymbol{x} - \boldsymbol{\mu})) d\boldsymbol{x} \\ &= \frac{\boldsymbol{y}^T \boldsymbol{K}^{-1}}{RZ|(2\pi)^d \boldsymbol{\Sigma}|^{1/2}} \sum_{r=1}^{R} \sum_{z=1}^{Z} \int_{\boldsymbol{x} \in \mathcal{R}^d} \cos(\boldsymbol{\omega}_r^T(\boldsymbol{x} - \boldsymbol{X})) \cos(\boldsymbol{\rho}_z^T(\boldsymbol{x} - \boldsymbol{\mu})) d\boldsymbol{x} \end{aligned} \tag{11}$$

Looking at the inner-term integrand term in (11), we can apply the trigonometric identity $\cos(\alpha)\cos(\beta) = \cos(\alpha+\beta)/2 + \cos(\alpha-\beta)/2$ and rewrite the integral as:

$$\int_{\boldsymbol{x}\in\mathcal{R}^d} \cos(\boldsymbol{\omega}_r^T(\boldsymbol{x}-\boldsymbol{X}))\cos(\boldsymbol{\rho}_z^T(\boldsymbol{x}-\boldsymbol{\mu}))d\boldsymbol{x} = \int_{\boldsymbol{x}\in\mathcal{R}^d} \frac{\cos(\boldsymbol{\omega}_r^T(\boldsymbol{x}-\boldsymbol{X}) + \boldsymbol{\rho}_z^T(\boldsymbol{x}-\boldsymbol{\mu}))}{2} + \frac{\cos(\boldsymbol{\omega}_r^T(\boldsymbol{x}-\boldsymbol{X}) - \boldsymbol{\rho}_z^T(\boldsymbol{x}-\boldsymbol{\mu}))}{2}\, d\boldsymbol{x}, \tag{12}$$

which we can reorganize as:

$$\int_{\boldsymbol{x}\in\mathcal{R}^d} \frac{cos(\boldsymbol{x}^T(\boldsymbol{\omega}_r+\boldsymbol{\rho}_z) - (\boldsymbol{\omega}_r^T\boldsymbol{X} + \boldsymbol{\rho}_z^T\boldsymbol{\mu}))}{2} + \frac{cos(\boldsymbol{x}^T(\boldsymbol{\omega}_r-\boldsymbol{\rho}_z) - (\boldsymbol{\omega}_r^T\boldsymbol{X} - \boldsymbol{\rho}_z^T\boldsymbol{\mu}))}{2}\, d\boldsymbol{x}. \tag{13}$$

Using the same method of u-substitution and properties of the integrals of cos and sin as in (2) and (5), we can solve the indefinite integral, and return to place the terms not involving $\boldsymbol{x}$ from (11), as:

$$\langle\bar{f}\rangle = L \times \sum_{r=1}^{R}\sum_{z=1}^{Z}\Big[\frac{h^d(\boldsymbol{x}^T(\boldsymbol{\omega}_r+\boldsymbol{\rho}_z) - (\boldsymbol{\omega}_r^T\boldsymbol{X} + \boldsymbol{\rho}_z^T\boldsymbol{\mu}))}{2\prod_{j=1}^d(\omega_r^j+\rho_z^j)} + \frac{h^d(\boldsymbol{x}^T(\boldsymbol{\omega}_r-\boldsymbol{\rho}_z) - (\boldsymbol{\omega}_r^T\boldsymbol{X} - \boldsymbol{\rho}_z^T\boldsymbol{\mu}))}{2\prod_{j=1}^d(\omega_r^j-\rho_z^j)}\Big] + c \tag{14}$$

$$L = \frac{\boldsymbol{y}^T\boldsymbol{K}^{-1}}{RZ[(2\pi)^d|\boldsymbol{\Sigma}|]^{1/2}} \tag{15}$$

where h^d is defined as in (5). This is the same formulation we use in the main paper in definition 1, and represents the GBQ integral mean over the Gaussian measure.

A.4 RFF Kernel Variance Over Gaussian Measures

To find the variance of $\langle\bar{f}\rangle$ over a Gaussian measure, we find the expectation of the Gaussian kernel mean, which we calculate as a part of (14), over $p(\boldsymbol{X})$

To do so, we first rewrite the kernel mean

$$\boldsymbol{\mu}_{\boldsymbol{x}}(\boldsymbol{X}) = \sum_{r=1}^{R}\sum_{z=1}^{Z}\Big[\frac{h^d(\boldsymbol{x}^T(\boldsymbol{\omega}_r+\boldsymbol{\rho}_z) - (\boldsymbol{\omega}_r^T\boldsymbol{X} + \boldsymbol{\rho}_z^T\boldsymbol{\mu}))}{2\prod_{j=1}^d(\omega_r^j+\rho_z^j)} + \frac{h^d(\boldsymbol{x}^T(\boldsymbol{\omega}_r-\boldsymbol{\rho}_z) - (\boldsymbol{\omega}_r^T\boldsymbol{X} - \boldsymbol{\rho}_z^T\boldsymbol{\mu}))}{2\prod_{j=1}^d(\omega_r^j-\rho_z^j)}\Big] \times \frac{1}{RZ[(2\pi)^d|\boldsymbol{\Sigma}|]^{1/2}} \tag{16}$$

as

$$\boldsymbol{\mu}_{\boldsymbol{x}}(\boldsymbol{X}) = \frac{1}{RZ[(2\pi)^d|\boldsymbol{\Sigma}|]^{1/2}}\sum_{r=1}^{R}\sum_{z=1}^{Z}\Big[\frac{h^d(\gamma - \boldsymbol{\omega}_r^T\boldsymbol{X})}{2\prod_{j=1}^d(\omega_r^j+\rho_z^j)} + \frac{h^d(\delta - \boldsymbol{\omega}_r^T\boldsymbol{X})}{2\prod_{j=1}^d(\omega_r^j-\rho_z^j)}\Big] \tag{17}$$

Where we have combined the terms inside the h^ds not involving $\boldsymbol{X}$ with γ and δ.

We next substitute (17) into (7), while also introducing the RFF estimate to $p(\boldsymbol{X})$, to obtain the integral variance.

$$\begin{aligned}\mathbb{V}(\langle\bar{f}\rangle) &= \int_{\boldsymbol{X}\in\mathcal{R}^d} \frac{\cos(\boldsymbol{\rho}_u^T(\boldsymbol{X}-\boldsymbol{\mu}))}{RZU[(2\pi)^d|\boldsymbol{\Sigma}|]}\sum_{r=1}^{R}\sum_{z=1}^{Z}\sum_{u=1}^{U}\Big[\frac{h^d(\gamma - \boldsymbol{\omega}_r^T\boldsymbol{X})}{2\prod_{j=1}^d(\omega_r^j+\rho_z^j)} + \frac{h^d(\delta - \boldsymbol{\omega}_r^T\boldsymbol{X})}{2\prod_{j=1}^d(\omega_r^j-\rho_z^j)}\Big]\, d\boldsymbol{X} \\ &= \tau\sum_{r=1}^{R}\sum_{z=1}^{Z}\sum_{u=1}^{U}\int_{\boldsymbol{X}\in\mathcal{R}^d}\cos(\boldsymbol{\rho}_u^T(\boldsymbol{X}-\boldsymbol{\mu}))\Big[\frac{h^d(\gamma - \boldsymbol{\omega}_r^T\boldsymbol{X})}{2\prod_{j=1}^d(\omega_r^j+\rho_z^j)} + \frac{h^d(\delta - \boldsymbol{\omega}_r^T\boldsymbol{X})}{2\prod_{j=1}^d(\omega_r^j-\rho_z^j)}\Big]\, d\boldsymbol{X}\end{aligned} \tag{18}$$

where have introduced another index of $\boldsymbol{\rho}$ in $u = 1, \ldots, U$, and substituted τ for $[RZU(2\pi)^d|\boldsymbol{\Sigma}|]^{-1}$

We note here that h^d could be any of $[\cos, \sin, -\cos, -\sin]$, so we cannot necessarily leverage the same identity we used previously on the products of cosines. However, a very similar identity exists for the product of cosines and sines in $\sin(\alpha)\cos(\beta) = \sin(\alpha+\beta)/2 + \sin(\alpha-\beta)$, and the negativity of either function is trivial. We will continue with the variance proof under the assumption that $h^d = \cos$, but it is straightforward to derive the variance under other cases.

Using the identity $\cos(\alpha)\cos(\beta) = \cos(\alpha+\beta)/2 + \cos(\alpha-\beta)$, we simplify the integrand in (18) to:

$$\int_{\boldsymbol{X}\in\mathcal{R}^d} \frac{\cos(\boldsymbol{\rho}_u^T(\boldsymbol{X}-\boldsymbol{\mu}) + \gamma - \boldsymbol{\omega}_r^T\boldsymbol{X})}{4\prod_{j=1}^d(\omega_r^j + \rho_z^j)} + \frac{\cos(\boldsymbol{\rho}_u^T(\boldsymbol{X}-\boldsymbol{\mu}) - \gamma + \boldsymbol{\omega}_r^T\boldsymbol{X})}{4\prod_{j=1}^d(\omega_r^j + \rho_z^j)} + \frac{\cos(\boldsymbol{\rho}_u^T(\boldsymbol{X}-\boldsymbol{\mu}) + \delta - \boldsymbol{\omega}_r^T\boldsymbol{X})}{4\prod_{j=1}^d(\omega_r^j - \rho_z^j)} + \frac{\cos(\boldsymbol{\rho}_u^T(\boldsymbol{X}-\boldsymbol{\mu}) - \delta + \boldsymbol{\omega}_r^T\boldsymbol{X})}{4\prod_{j=1}^d(\omega_r^j - \rho_z^j)} \tag{19}$$

and simplify further to

$$\int_{\boldsymbol{X}\in\mathcal{R}^d} \frac{\cos(\boldsymbol{X}^T(\boldsymbol{\rho}_u - \boldsymbol{\omega}_r) - \boldsymbol{\rho}_u^T\boldsymbol{\mu} + \gamma)}{4\prod_{j=1}^d(\omega_r^j + \rho_z^j)} + \frac{\cos(\boldsymbol{X}^T(\boldsymbol{\rho}_u + \boldsymbol{\omega}_r) - \boldsymbol{\rho}_u^T\boldsymbol{\mu} - \gamma)}{4\prod_{j=1}^d(\omega_r^j + \rho_z^j)} + \frac{\cos(\boldsymbol{X}^T(\boldsymbol{\rho}_u - \boldsymbol{\omega}_r) - \boldsymbol{\rho}_u^T\boldsymbol{\mu} + \delta)}{4\prod_{j=1}^d(\omega_r^j - \rho_z^j)} + \frac{\cos(\boldsymbol{X}^T(\boldsymbol{\rho}_u + \boldsymbol{\omega}_r) - \boldsymbol{\rho}_u^T\boldsymbol{\mu} - \delta))}{4\prod_{j=1}^d(\omega_r^j - \rho_z^j)} \tag{20}$$

Lastly, we use the same tactics u-substitution and knowledge of the repeating nature of trigonometric integrals to calculate the indefinite integral which can be evaluated to produce the GBQ variance over a Gaussian measure:

$$\begin{aligned}\mathbb{V}(\langle\bar{f}\rangle) = \tau\sum_{r=1}^{R}\sum_{z=1}^{Z}\sum_{u=1}^{U}\Bigg[&\frac{h^d(\boldsymbol{X}^T(\boldsymbol{\rho}_u - \boldsymbol{\omega}_r) - \boldsymbol{\rho}_u^T\boldsymbol{\mu} + \gamma)}{4\prod_{j=1}^d(\omega_r^j + \rho_z^j)(\rho_u^j - \omega_r^j)} + \frac{h^d(\boldsymbol{X}^T(\boldsymbol{\rho}_u + \boldsymbol{\omega}_r) - \boldsymbol{\rho}_u^T\boldsymbol{\mu} - \gamma)}{4\prod_{j=1}^d(\omega_r^j + \rho_z^j)(\rho_u^j + \omega_r^j)} \\ &+ \frac{h^d(\boldsymbol{X}^T(\boldsymbol{\rho}_u - \boldsymbol{\omega}_r) - \boldsymbol{\rho}_u^T\boldsymbol{\mu} + \delta)}{4\prod_{j=1}^d(\omega_r^j - \rho_z^j)(\rho_u^j - \omega_r^j)} + \frac{h^d(\boldsymbol{X}^T(\boldsymbol{\rho}_u + \boldsymbol{\omega}_r) - \boldsymbol{\rho}_u^T\boldsymbol{\mu} - \delta))}{4\prod_{j=1}^d(\omega_r^j - \rho_z^j)(\rho_u^j + \omega_r^j)}\Bigg]\end{aligned} \tag{21}$$

where h^d is defined as before, and we can substitute back γ, τ, and δ. We remark that in the case h^d is not cos in the kernel mean (17) that the associated proof follows a very similar form.

A.5 Multivariate CDF of the RFF Formulated Gaussian

Using the already established methods from the previous two sections, particularly section A.1, on the integrals of RFF-parametrized kernels and distributions, it is trivial to show through u-substitution and the straightforward integrals of trigonometric functions that the indefinite integral an RFF Gaussian over the bounds is:

$$\begin{aligned}\hat{\Phi}(\boldsymbol{x}) &= \int_{\boldsymbol{x}\in\mathcal{R}^d} \frac{1}{R[(2\pi)^d|\boldsymbol{\Sigma}|]^{1/2}}\sum_{r=1}^{R}\cos(\boldsymbol{\rho}^T(\boldsymbol{x}-\boldsymbol{\mu}))\,d\boldsymbol{x} \\ &= \frac{1}{R[(2\pi)^d|\boldsymbol{\Sigma}|]1/2}\sum_{r=1}^{R}\int_{\boldsymbol{x}\in\mathcal{R}^d}\cos(\boldsymbol{\rho}^T(\boldsymbol{x}-\boldsymbol{\mu}))\,d\boldsymbol{x} \\ &= \frac{h^d(\boldsymbol{\rho}^T(\boldsymbol{x}-\boldsymbol{\mu}))}{R[(2\pi)^d|\boldsymbol{\Sigma}|]^{1/2}\prod_{j=1}^d\rho_r^j}\end{aligned} \tag{22}$$

where h^d is defined as in previous sections. Subsequent application of this indefinite integral over definite bounds $[\boldsymbol{a}, \boldsymbol{b}]$ can then be used to estimate the CDF of a multivariate Gaussian over a domain.

B Proofs for the theoretical results

B.1 Background

We consider a standard GP posterior mean and variance, respectively, as:

$$\mu_n(\boldsymbol{x}) := \mathbf{k}_n(\boldsymbol{x})^\mathsf{T}(\mathbf{K}_n + \lambda\mathbf{I})^{-1}\mathbf{y}_n \tag{23}$$

$$\sigma_n^2(\boldsymbol{x}) := k(\boldsymbol{x}, \boldsymbol{x}') - \mathbf{k}_n(\boldsymbol{x})^\mathsf{T}(\mathbf{K}_n + \lambda\mathbf{I})^{-1}\mathbf{k}_n(\boldsymbol{x}) \tag{24}$$

where we use notation shortcuts for the vector $\mathbf{k}(\boldsymbol{x}) := [k(\boldsymbol{x}, \boldsymbol{x}_i)]_{i=1}^n \in \mathbb{R}^n$ and the kernel matrix $\mathbf{K} := [k(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j=1}^n \in \mathbb{R}^{n\times n}$. Correspondingly, the our method employs Fourier features to approximate a GP posterior mean as:

$$\hat{\mu}_n(\boldsymbol{x}) := \tilde{\mathbf{k}}(\boldsymbol{x})^\mathsf{T}\left(\mathbf{K}_n + \lambda\mathbf{I}\right)^{-1}\mathbf{y}\,, \tag{25}$$

where $\tilde{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is formally defined according to the next statement.

Definition 2. *Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ denote a translation-invariant positive-definite kernel on $\mathcal{X} \subset \mathbb{R}^d$, $d \in \mathbb{N}$. The random Fourier feature approximation is defined as:*

$$\tilde{k}(\boldsymbol{x}, \boldsymbol{x}') := \boldsymbol{\phi}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\phi}(\boldsymbol{x}'), \quad \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}\,, \tag{26}$$

where:

$$\boldsymbol{\phi}(\boldsymbol{x}) := \sqrt{\frac{1}{R}}\begin{bmatrix}\sin(\boldsymbol{\omega}_1^\mathsf{T}\boldsymbol{x})\\ \cos(\boldsymbol{\omega}_1^\mathsf{T}\boldsymbol{x})\\ \vdots\\ \sin(\boldsymbol{\omega}_R^\mathsf{T}\boldsymbol{x})\\ \cos(\boldsymbol{\omega}_R^\mathsf{T}\boldsymbol{x})\end{bmatrix}, \quad \boldsymbol{\omega}_i \overset{i.i.d.}{\sim} P_k\,, \quad \boldsymbol{x} \in \mathcal{X}\,, \tag{27}$$

with P_k denoting the probability distribution that corresponds to the Fourier transform of the kernel k. Equivalently, we can also write:

$$\tilde{k}(\boldsymbol{x}, \boldsymbol{x}') = \frac{1}{R}\sum_{i=1}^{R}\cos(\boldsymbol{\omega}_i^\mathsf{T}(\boldsymbol{x} - \boldsymbol{x}')\,, \quad \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}\,. \tag{28}$$

B.2 Auxiliary results

We will make use of guarantees for RFFs to bound the kernel approximation error. In particular, we consider the following result from Sutherland and Schneider (2015).

Lemma 3 (Sutherland and Schneider (2015, Proposition 1), full version)**.** *Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a continuous shift-invariant positive-definite kernel with $k(\boldsymbol{x}, \boldsymbol{x}) = 1$ and such that $\nabla^2 k(\boldsymbol{x}, \boldsymbol{x})$ exists, for all $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d$. Suppose $\mathcal{X}$ is compact with diameter $\ell_\mathcal{X} < \infty$. Denote k's Fourier transform as P_k, which is a probability measure, and let $\sigma_k^2 := \mathbb{E}[\|\boldsymbol{\omega}\|_2^2]$ for $\omega \sim P_k$. Let $\tilde{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ denote k's RFF approximation with R frequencies according to Definition 2. For any $\xi > 0$, let:*

$$\alpha_\xi := \min\left(1, \sup_{\boldsymbol{x},\boldsymbol{x}'\in\mathcal{X}} \frac{1}{2} + \frac{1}{2}k(2\boldsymbol{x}, 2\boldsymbol{x}') - k(\boldsymbol{x}, \boldsymbol{x}')^2 + \frac{1}{3}\xi\right)\,, \tag{29}$$

$$\beta_d := \left(\left(\frac{d}{2}\right)^{-\frac{d}{d+2}} + \left(\frac{d}{2}\right)^{\frac{2}{d+2}}\right)2^{\frac{6d+2}{d+2}}\,. \tag{30}$$

Then the following holds for any $\xi > 0$:

$$\begin{aligned}\mathbb{P}\left[\sup_{\boldsymbol{x},\boldsymbol{x}'\in\mathcal{X}} |\tilde{k}(\boldsymbol{x},\boldsymbol{x}') - k(\boldsymbol{x},\boldsymbol{x}')| \geq \xi\right] &\leq \beta_d \left(\frac{\sigma_k \ell_{\mathcal{X}}}{\xi}\right)^{\frac{2}{1+\frac{2}{d}}} \exp\left(-\frac{R\xi^2}{4(d+2)\alpha_\xi}\right) \\ &\leq 66 \left(\frac{\sigma_k \ell_{\mathcal{X}}}{\xi}\right)^{2} \exp\left(-\frac{R\xi^2}{4(d+2)}\right),\end{aligned} \tag{31}$$

where for the second statement we assume $\xi \leq \sigma_k \ell_{\mathcal{X}}$. Therefore, for any $\delta \in (0,1)$, we can achieve pointwise approximation error less than ξ with probability at least $1-\delta$ if:

$$R \geq \frac{4(d+2)\alpha_\xi}{\xi^2}\left(\frac{2}{1+\frac{2}{d}} \log \frac{\sigma_k \ell_{\mathcal{X}}}{\xi} + \log \frac{\beta_d}{\delta}\right). \tag{32}$$

Compared to the original statement of the result in Sutherland and Schneider (2015), note that we use the number of Fourier frequencies R, instead of the dimensionality of the feature vector, i.e., $D := 2R$, so that some constants are changed. Considering the result above, as $\max_{d\in\mathbb{N}} \beta_d = 66$ (see Sutherland and Schneider, 2015) and $\alpha_\xi \leq 1$, we can also set the minimum number of features for a given error bound $\xi > 0$ and $\delta \in (0,1)$ as:

$$R(\xi,\delta,\sigma_k) := \frac{4(d+2)}{\xi^2}\left(\frac{2}{1+\frac{2}{d}} \log \frac{\sigma_k \ell_{\mathcal{X}}}{\xi} + \log \frac{66}{\delta}\right), \tag{33}$$

though a tighter bound is available via Equation 32. Therefore, the restatement of the result in the main paper as Lemma 2 is still valid.

The norm of the observations vector $\mathbf{y}$ in a Gaussian process can be bounded in terms of the integrand f's extremes and the number of data points, as in the following result.

Lemma 4. *Given $\delta \in (0,1)$, assuming i.i.d. Gaussian observation noise $\epsilon \sim \mathcal{N}(0,\sigma_\epsilon^2)$, we have that:*

$$\mathbb{P}\left[\|\mathbf{y}\|_2 \leq \sqrt{n}\left(\|f\|_\infty + \sigma_\epsilon \sqrt{2\log\left(\frac{n}{\delta}\right)}\right)\right] \geq 1-\delta. \tag{34}$$

Proof. Starting from the definition of the 2-norm, we have:

$$\|\mathbf{y}\|_2^2 = \sum_{i=1}^n y_i^2 = \sum_{i=1}^n (f(\boldsymbol{x}_i)+\epsilon_i)^2 \leq n \max_{i\in\{1,\dots,n\}} (f(\boldsymbol{x}_i)+\epsilon_i)^2. \tag{35}$$

Assuming i.i.d. Gaussian observation noise $\epsilon \sim \mathcal{N}(0,\sigma_\epsilon^2)$, the following holds:

$$\forall \beta > 0, \quad \mathbb{P}\left[|\epsilon| \geq \beta\sigma_\epsilon\right] \leq \exp\left(-\beta^2/2\right), \tag{36}$$

By applying a union bound, we have:

$$\begin{aligned}\mathbb{P}\left[\exists i \in \{1,\dots,n\} : y_i \geq f(\boldsymbol{x}_i) + \beta\sigma_\epsilon\right] &\leq \sum_{i=1}^n \mathbb{P}\left[\epsilon_i \geq \beta\sigma_\epsilon\right] \\ &\leq n\mathbb{P}\left[|\epsilon| \geq \beta\sigma_\epsilon\right] \\ &\leq n\exp\left(-\beta^2/2\right)\end{aligned} \tag{37}$$

Solving for $n\exp\left(-\beta^2/2\right) = \delta$ and taking the complement, we then obtain:

$$\mathbb{P}\left[\forall i \in \{1,\dots,n\}, \quad y_i \leq \|f\|_\infty + \sigma_\epsilon\sqrt{2\log\left(\frac{n}{\delta}\right)}\right] \geq 1-\delta. \tag{38}$$

The result then follows by applying the latter to Equation 35. □

B.3 The probability distribution approximation via RFF

For the approximation of p by $\tilde{p}$, we use the following fact.

Theorem 4 (Bochner's theorem (Rudin, 1990))**.** *A function $u : \mathcal{X} \to \mathbb{R}$, $\mathcal{X} \subset \mathbb{R}^d$ is positive-definite if and only if it is the Fourier transform of a non-negative measure.*

By Bochner's theorem (Theorem 4), as previously applied to positive-definite kernels (Theorem 1, main paper), we can also trivially conclude that any *positive-definite* probability density function is by itself the Fourier transform of a probability measure, so that it admits a Fourier-feature representation of the form in Definition 2. A probability density function $p : \mathbb{R}^d \to [0, \infty)$ is positive-definite if, for all $n \in \mathbb{N}$, $\{\alpha_i\}_{i=1}^n \subset \mathbb{R}$ and all $\{\boldsymbol{x}_i\}_{i=1}^n \subset \mathbb{R}^d$ the following holds:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j p(\boldsymbol{x}_i - \boldsymbol{x}_j) \geq 0 \,. \tag{39}$$

Not every probability density function is positive-definite, but examples include Gaussian and Student-T distributions (Rossberg, 1995). In particular, we can make a kernel k_p from a probability density function p on $\mathcal{X}$ by:

$$\begin{aligned} k_p : \mathcal{X} \times \mathcal{X} &\longrightarrow \mathbb{R} \\ \boldsymbol{x}, \boldsymbol{x}' &\longmapsto \begin{cases} p(\boldsymbol{x} - \boldsymbol{x}') \,, & \boldsymbol{x} - \boldsymbol{x}' \in \mathcal{X}, \\ 0 \,, & \boldsymbol{x} - \boldsymbol{x}' \notin \mathcal{X} \,. \end{cases} \end{aligned} \tag{40}$$

It is easy to verify that a kernel defined as above is positive-definite if p is positive-definite. The kernel is also translation-invariant, since $k_p(\mathbf{v} + \boldsymbol{x}, \mathbf{v} + \boldsymbol{x}') = k_p(\boldsymbol{x}, \boldsymbol{x}')$, for any $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$ and any $\mathbf{v} \in \mathbb{R}^d$. Similarly, we have the equivalence $p(\boldsymbol{x}) = k_p(\boldsymbol{x}, \mathbf{0})$ and a corresponding $\tilde{p}(\boldsymbol{x}) = \tilde{k}_p(\boldsymbol{x}, \mathbf{0})$, for $\boldsymbol{x} \in \mathcal{X}$, by applying Definition 2 to k_p. As a result, we can use Lemma 3 to k_p to bound the approximation error in $|p(\boldsymbol{x}) - \tilde{p}(\boldsymbol{x})|$.

Theorem 5 (Restatement of Theorem 2)**.** *Let $p : \mathcal{X} \to \mathbb{R}$ be a positive-definite probability density function defined on $\mathcal{X} \subset \mathbb{R}^d$ which is such that $\nabla^2 p(\mathbf{0})$ exists. Assume $\mathcal{X}$ is compact, and let $b_p > 0$ be any constant such that $b_p \geq \max_{\boldsymbol{x} \in \mathcal{X}} p(\boldsymbol{x})$. Let $\tilde{k}_p$ denote an RFF approximation with $R_p \in \mathbb{N}$ frequencies to k_p as defined in Equation 40, and let $\tilde{p} : \boldsymbol{x} \mapsto \tilde{k}_p(\boldsymbol{x}, \mathbf{0})$, $\boldsymbol{x} \in \mathcal{X}$. Then, for any $\xi > 0$, the following holds:*

$$\begin{aligned} \mathbb{P}\left[\sup_{\boldsymbol{x} \in \mathcal{X}} |\tilde{p}(\boldsymbol{x}) - p(\boldsymbol{x})| \geq b_p \xi\right] &\leq \beta_d \left(\frac{\sigma_{k_p} \ell_{\mathcal{X}}}{\xi}\right)^{\frac{2}{1+\frac{2}{d}}} \exp\left(-\frac{R_p \xi^2}{4(d+2)\alpha_\xi}\right) \\ &\leq 66 \left(\frac{\sigma_{k_p} \ell_{\mathcal{X}}}{\xi}\right)^2 \exp\left(-\frac{R_p \xi^2}{4(d+2)}\right) \end{aligned} \tag{41}$$

where for the second statement we assume $\xi \leq \sigma_{k_p} \ell_{\mathcal{X}}$, and σ_{k_p}, $\ell_{\mathcal{X}}$, α_ξ and β_ξ are the same as defined in Lemma 3 for $k := \frac{1}{b_p} k_p$.

Proof. The result follows by applying Lemma 3 to a normalised version $\bar{k}_p := \frac{1}{b_p} k_p$ of k_p (Equation 40), which is such that $\bar{k}_p(\boldsymbol{x}, \boldsymbol{x}') = 1$. Noticing that:

$$\begin{aligned} \sup_{\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}} |\tilde{k}_p(\boldsymbol{x}, \boldsymbol{x}') - k_p(\boldsymbol{x}, \boldsymbol{x}')| &= \sup_{\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}} |\tilde{k}_p(\boldsymbol{x} - \boldsymbol{x}', \mathbf{0}) - k_p(\boldsymbol{x} - \boldsymbol{x}', \mathbf{0})| \\ &= \sup_{\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X} : \boldsymbol{x} - \boldsymbol{x}' \in \mathcal{X}} |\tilde{p}(\boldsymbol{x} - \boldsymbol{x}') - p(\boldsymbol{x} - \boldsymbol{x}')| \\ &\leq \sup_{\boldsymbol{x} \in \mathcal{X}} |\tilde{p}(\boldsymbol{x}) - p(\boldsymbol{x})| \,, \end{aligned} \tag{42}$$

so that $\sup_{\boldsymbol{x},\boldsymbol{x}'\in\mathcal{X}} |\tilde{k}_p(\boldsymbol{x},\boldsymbol{x}') - k_p(\boldsymbol{x},\boldsymbol{x}')| \geq b_p\xi$ implies $\sup_{\boldsymbol{x}\in\mathcal{X}} |\tilde{p}(\boldsymbol{x}) - p(\boldsymbol{x})| \geq b_p\xi$, concludes the proof. □

Given $\xi_p > 0$ such that $\sup_{\boldsymbol{x}\in\mathcal{X}} |p(\boldsymbol{x}) - \tilde{p}(\boldsymbol{x})| \,\mathrm{d}\boldsymbol{x} \leq \xi_p$, the integration error is bounded by:

$$\int_{\mathcal{X}} |p(\boldsymbol{x}) - \tilde{p}(\boldsymbol{x})| \,\mathrm{d}\boldsymbol{x} \leq b_p\xi_p \int_{\mathcal{X}} \mathrm{d}\boldsymbol{x} \leq b_p\xi_p v_{\mathcal{X}} \,, \tag{43}$$

where $v_{\mathcal{X}} := \int_{\mathcal{X}} \mathrm{d}\boldsymbol{x}$ denotes the volume of the domain $\mathcal{X}$. The latter can be bounded by the volume of a hyper-sphere of diameter $\ell_{\mathcal{X}}$ in $\mathbb{R}^d$, i.e.:

$$v_{\mathcal{X}} \leq \frac{\pi^d \ell_{\mathcal{X}}^d}{2^d \Gamma\left(\frac{d}{2}+1\right)} \,, \tag{44}$$

where Γ denotes Euler's gamma function.

B.4 Quadrature approximation error

We now combine our results to bound the quadrature approximation error.

Theorem 6 (Restatement of Theorem 3)**.** *Let $f \in \mathcal{H}_k$, where $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive-definite, translation-invariant kernel on $\mathcal{X} \subset \mathbb{R}^d$. Assume that:*

1. *$\mathcal{X}$ is compact with diameter $\ell_{\mathcal{X}} < \infty$ and volume $v_{\mathcal{X}} := \int_{\mathcal{X}} \mathrm{d}\boldsymbol{x} < \infty$;*
2. *$k(\mathbf{0},\mathbf{0}) = 1$ and $\nabla^2 k(\mathbf{0},\mathbf{0})$ exists;*
3. *and $p : \mathcal{X} \to [0,\infty)$ is a positive-definite probability density function.*

Then, given any $\delta \in (0,1)$, the following holds with probability at least $1-\delta$:

$$\begin{aligned} &\left| \int_{\mathcal{X}} f(\boldsymbol{x})p(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} - \int_{\mathcal{X}} \hat{\mu}_n(\boldsymbol{x})\tilde{p}(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} \right| \\ &\leq \left(\frac{n}{\lambda}\beta_\epsilon\left(\frac{\delta}{4}\right)\xi_k + \beta_k\left(\frac{\delta}{4}\right)\max_{\boldsymbol{x}\in\mathcal{X}} \sigma_n(\boldsymbol{x}) \right)(1 + b_p\xi_p v_{\mathcal{X}}) + \|f\|_\infty b_p\xi_p v_{\mathcal{X}} \,, \end{aligned} \tag{45}$$

for an RFF approximation to k with $R_k \geq R\left(\xi_k, \frac{\delta}{4}, \sigma_k\right)$ frequencies and an RFF approximation to p with $R_p \geq R\left(\xi_p, \frac{\delta}{4}, \sigma_{k_p}\right)$ frequencies, given $0 < \xi_k \leq \sigma_k \ell_{\mathcal{X}}$ and $0 < \xi_p \leq \sigma_{k_p}\ell_{\mathcal{X}}$, where:

$$\beta_\epsilon(\delta) := \|f\|_\infty + \sigma_\epsilon \sqrt{2\log\left(\frac{n}{\delta}\right)} \tag{46}$$

$$\beta_k(\delta) := \|f\|_k + \sigma_\epsilon \sqrt{\frac{2}{\lambda}\log\left(\frac{\det(\mathbf{I} + \lambda^{-1}\mathbf{K}_n)^{1/2}}{\delta}\right)} \tag{47}$$

$$R(\xi,\delta,\sigma_k) := \frac{4(d+2)}{\xi^2}\left(\frac{2}{1+\frac{2}{d}}\log\frac{\sigma_k\ell_{\mathcal{X}}}{\xi} + \log\frac{66}{\delta}\right) . \tag{48}$$

Proof. In the spectral Bayesian quadrature formulation, we have the following approximation:

$$\int_{\mathcal{X}} f(\boldsymbol{x})p(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} \approx \mathbf{y}^\mathsf{T}(\mathbf{K}_n + \lambda\mathbf{I})^{-1} \int_{\mathcal{X}} \tilde{\mathbf{k}}_n(\boldsymbol{x})\tilde{p}(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} = \int_{\mathcal{X}} \hat{\mu}(\boldsymbol{x})\tilde{p}(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} \,, \tag{49}$$

where $\hat{\mu}_n(\boldsymbol{x}) := \tilde{\mathbf{k}}_n(\boldsymbol{x})^\mathsf{T}(\mathbf{K}_n + \lambda\mathbf{I})^{-1}\mathbf{y}$ We will bound the approximation error by starting with the following decomposition:

$$\begin{aligned}
&\left|\int_{\mathcal{X}} f(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - \int_{\mathcal{X}} \hat{\mu}_n(\boldsymbol{x})\tilde{p}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\right| \\
&\leq \left|\int_{\mathcal{X}} f(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - \int_{\mathcal{X}} \hat{\mu}_n(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\right| + \left|\int_{\mathcal{X}} \hat{\mu}_n(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - \int_{\mathcal{X}} \hat{\mu}_n(\boldsymbol{x})\tilde{p}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\right| \\
&\leq \|f - \hat{\mu}_n\|_\infty + \|\hat{\mu}_n\|_\infty \int_{\mathcal{X}} |p(\boldsymbol{x}) - \tilde{p}(\boldsymbol{x})|\,\mathrm{d}\boldsymbol{x}\,.
\end{aligned} \tag{50}$$

We first observe that:

$$\forall \boldsymbol{x} \in \mathcal{X}, \quad |f(\boldsymbol{x}) - \hat{\mu}_n(\boldsymbol{x})| \leq |f(\boldsymbol{x}) - \mu(\boldsymbol{x})| + |\mu(\boldsymbol{x}) - \hat{\mu}(\boldsymbol{x})|\,. \tag{51}$$

Assuming $f \in \mathcal{H}_k$, given $\delta_\mu \in (0,1)$, we can apply Lemma 1 (main paper) to bound the first term on the right-hand side as:

$$\mathbb{P}\left[\sup_{\boldsymbol{x}\in\mathcal{X}} |f(\boldsymbol{x}) - \mu_n(\boldsymbol{x})| \leq \sup_{\boldsymbol{x}\in\mathcal{X}} \beta_k(\delta_\mu)\sigma_n(\boldsymbol{x})\right] \geq 1 - \delta_\mu\,. \tag{52}$$

For the second-term on the right-hand side of Equation 51, we have that:

$$\begin{aligned}
|\mu_n(\boldsymbol{x}) - \hat{\mu}_n(\boldsymbol{x})| &\leq \|\mathbf{k}_n(\boldsymbol{x}) - \tilde{\mathbf{k}}_n(\boldsymbol{x})\|_2 \|(\mathbf{K}_n + \lambda\mathbf{I})^{-1}\mathbf{y}\|_2 \\
&\leq \|\mathbf{k}_n(\boldsymbol{x}) - \tilde{\mathbf{k}}_n(\boldsymbol{x})\|_2 \|(\mathbf{K}_n + \lambda\mathbf{I})^{-1}\|_2 \|\mathbf{y}\|_2\,, \\
&\leq \frac{\|\mathbf{y}\|_2}{\lambda} \|\mathbf{k}_n(\boldsymbol{x}) - \tilde{\mathbf{k}}_n(\boldsymbol{x})\|_2\,.
\end{aligned} \tag{53}$$

since $\|(\mathbf{K}_n + \lambda\mathbf{I})^{-1}\|_2 \leq \lambda^{-1}$. Applying Lemma 4, given $\delta_\epsilon \in (0,1)$, yields:

$$\mathbb{P}\left[\|\mathbf{y}\|_2 \leq \sqrt{n}\beta_\epsilon(\delta_\epsilon)\right] \geq 1 - \delta_\epsilon\,. \tag{54}$$

where $\beta_\epsilon(\delta) := \|f\|_\infty + \sigma_\epsilon\sqrt{2\log\left(\frac{n}{\delta}\right)}$. In addition, considering the kernel approximation guarantee in Lemma 3, for a given number of Fourier frequencies $R_k \geq R(\delta_k, \xi_k)$, leads us to:

$$\mathbb{P}\left[\sup_{\boldsymbol{x}\in\mathcal{X}} \|\mathbf{k}_n(\boldsymbol{x}) - \tilde{\mathbf{k}}_n(\boldsymbol{x})\|_2 \leq \sqrt{n}\xi_k\right] \geq 1 - \delta_k\,. \tag{55}$$

Therefore, we have:

$$\mathbb{P}\left[\|f - \hat{\mu}_n\|_\infty \leq \beta_k(\delta_\mu)\max_{\boldsymbol{x}\in\mathcal{X}} \sigma_n(\boldsymbol{x}) + \frac{1}{\lambda} n\xi_k\beta_\epsilon(\delta_\epsilon)\right] \geq 1 - \delta_\mu - \delta_\epsilon - \delta_k\,, \tag{56}$$

which follows by applying a union bound on the complementary events in the equations above. Lastly, note that, under the assumption that the event in Equation 56 holds, the following is also true:

$$\|\hat{\mu}_n\|_\infty \leq \|f\|_\infty + \frac{1}{\lambda} n\xi_k\beta_\epsilon(\delta_\epsilon) + \max_{\boldsymbol{x}\in\mathcal{X}} \beta_k(\delta_k)\sigma_n(\boldsymbol{x})\,. \tag{57}$$

Regarding the probability density approximation, let $v_{\mathcal{X}} := \int_{\mathcal{X}} \mathrm{d}\boldsymbol{x}$ represent the volume of $\mathcal{X}$. Assume $R_p \geq R(\delta_p, \xi_p)$ Fourier frequencies for $\tilde{p}$, for $\delta_p \in (0,1)$. Then Theorem 5 tells us that:

$$\mathbb{P}\left[\int_{\mathcal{X}} |p(\boldsymbol{x}) - \tilde{p}(\boldsymbol{x})|\,\mathrm{d}\boldsymbol{x} \leq b_p\xi_p v_{\mathcal{X}}\right] \geq 1 - \delta_p\,. \tag{58}$$

The final result follows by applying a union bound to combine the events in equations 56, 57 and 58 into Equation 50. □

References

H.-J. Rossberg. Positive definite probability densities and probability distributions. *Journal of Mathematical Sciences*, 76(1):2181–2197, 1995.

Walter Rudin. The Basic Theorems of Fourier Analysis. In *Fourier Analysis on Groups*, number 2, chapter 1, pages 1–34. John Wiley & Sons, Ltd, 1990.

Dougal J. Sutherland and Jeff Schneider. On the error of random Fourier features. In *Uncertainty in Artificial Intelligence - Proceedings of the 31st Conference, UAI 2015*, pages 862–871, 2015.