Probabilistic Integration

François-Xavier Briol[1,*], Chris. J. Oates[2,3,*], Mark Girolami[1,4],
Michael A. Osborne[5] and Dino Sejdinovic[6]

[1]Department of Statistics, University of Warwick
[2]School of Mathematical and Physical Sciences, University of Technology Sydney
[3]The ARC Centre of Excellence for Mathematical and Statistical Frontiers
[4]The Alan Turing Institute for Data Science
[5]Department of Engineering Science, University of Oxford
[6]Department of Statistics, University of Oxford
[*]authors contributed equally

December 4, 2015

Abstract

Probabilistic numerical methods aim to model numerical error as a source of epistemic uncertainty that is subject to probabilistic analysis and reasoning, enabling the principled propagation of numerical uncertainty through a computational pipeline. In this paper we focus on numerical methods for integration. We present probabilistic (Bayesian) versions of both Markov chain and Quasi Monte Carlo methods for integration and provide rigorous theoretical guarantees for convergence rates, in both posterior mean and posterior contraction. The performance of probabilistic integrators is guaranteed to be no worse than non-probabilistic integrators and is, in many cases, asymptotically superior. These probabilistic integrators therefore enjoy the "best of both worlds", leveraging the sampling efficiency of advanced Monte Carlo methods whilst being equipped with valid probabilistic models for uncertainty quantification. Several applications and illustrations are provided, including examples from computer vision and system modelling using non-linear differential equations. A survey of open challenges in probabilistic integration is provided.

1 Introduction

The aim of this paper is to provide rigorous theoretical foundations for the probabilistic approach to integration introduced by O'Hagan (1991). A key feature of our analysis is the emphasis on connections with existing (non-probabilistic) Markov chain and Quasi Monte Carlo methods.

Context Numerical procedures, such as linear solvers, quadrature methods for integration and routines to approximately solve differential equations, are usually one of many building blocks in modern statistical inference procedures. These are typically considered as black-boxes that return a point estimate whose numerical error is considered to be negligible. Numerical methods are thus the only part of the statistical analysis for which uncertainty is not routinely accounted for in a fully probabilistic way (although analysis of errors and bounds on these are often available and highly developed). Failure to properly account for numerical error could potentially have drastic consequences on subsequent statistical inferences if the numerical error propagated through the computational pipeline is allowed to accumulate (Mosbach and Turner, 2009; Conrad et al., 2015).

Probabilistic numerics aims to explicitly model the epistemic uncertainty over the solution that remains after application of a particular numerical method (Hull and Swenson, 1966; Diaconis,

1988; Hennig et al., 2015). This confers several important benefits. Firstly, it provides a principled approach to quantify and propagate numerical uncertainty through computation, allowing for the possibility of errors with complex statistical structure. Secondly, it enables the user to control the uncertainty over the solution of the numerical procedure and identify key components of numerical uncertainty using statistical techniques such as analysis of variance. Thirdly, this probabilistic perspective can lead to new and effective numerical algorithms, as evidenced in recent work in the case of differential equations (Schober et al., 2014; Conrad et al., 2015; Dashti and Stuart, 2016)), linear algebra (Hennig, 2015) and optimization (Snoek et al., 2012; Hennig and Kiefel, 2013; Mahsereci and Hennig, 2015). The philosophical foundations for probabilistic numerics were first clearly exposed in the work of Diaconis (1988) and O'Hagan (1992) but elements can be traced back to Poincaré (1912) and Hull and Swenson (1966). We refer the interested reader to the recent exposition by Hennig et al. (2015).

Novel Contributions This paper develops probabilistic methods for numerical integration. Given a probability measure Π on a state space $\mathcal{X}$ with density function $\pi : \mathcal{X} \to [0, \infty)$, defined with respect to a reference measure σ, we aim to estimate integrals of the form

$$\Pi[f] := \int_{\mathcal{X}} f(\boldsymbol{x})\pi(\boldsymbol{x})\mathrm{d}\sigma(\boldsymbol{x}) \tag{1}$$

where $f : \mathcal{X} \to \mathbb{R}$ or $\mathbb{C}$ is a test function of interest. Two important scenarios motivate our work. Firstly, when evaluation of f is computationally intensive, so that only crude estimates for integrals can be obtained. Secondly, when many numerical integrals must be performed sequentially, so that small errors are able to accumulate.

In the case of integration, Bayesian Quadrature (BQ; O'Hagan, 1991) is a probabilistic numerics method that performs integration from a statistical perspective. Specifically, BQ assigns a Gaussian process prior measure over the integrand f and then, based on data $\mathcal{D} = \{\boldsymbol{x}_i, f_i\}_{i=1}^n$ with $\boldsymbol{x}_i \in \mathcal{X}$ and $f_i = f(\boldsymbol{x}_i)$, outputs a Gaussian process posterior measure $f|\mathcal{D}$ according to Bayes' rule. In turn this implies a Gaussian posterior distribution over $\Pi[f]$, since Π is a linear functional in f, representing a probabilistic model for uncertainty over the true value of $\Pi[f]$. The approach applies equally to a pre-determined set of states $\{\boldsymbol{x}_i\}_{i=1}^n$ or to states that are realisations from a random process, such as samples from a probability distribution which is often, but not necessarily, Π. In the latter, randomised case the method is known as Bayesian Monte Carlo (BMC; Rasmussen and Ghahramani, 2002).

Compared to non-probabilistic integrators, BQ has lacked rigorous theoretical foundations; this is addressed by the present paper. We propose and analyse Bayesian approaches to Quasi-Monte Carlo (QMC) and Markov Chain Monte Carlo (MCMC). The resulting Bayesian QMC (BQMC) and Bayesian MCMC (BMCMC) methods confer the benefits of efficient sampling schemes to a Bayesian approach to integration. In each case, we provide theoretical analysis of convergence rates for the posterior mean, which will always improve on the non-probabilistic counterparts, as well as rates for contraction of the Bayesian posterior. In doing so we lay to rest one of the principle critiques of Bayesian approaches to integration by establishing rigorous theoretical guarantees for these procedures.

The present paper significantly extends recent work by Briol et al. (2015) and provides a much more comprehensive exposition. A more abstract treatment of BMC has recently been provided by Bach (2015). Our work differs by focussing on explicit, constructive approaches to integration,

while Bach (2015) requires a specific importance sampling distribution which is not always available in closed-form.

Outline The paper is structured as follows. Sec. 2 provides an introduction to existing probabilistic numerics methodology for integration. Sec. 3 describes our proposed probabilistic integrators and analyses their theoretical properties. The two main theoretical results concern convergence and contraction rates, which will depend on both prior information and the method that is used to select states. Several technical issues are discussed in Sec. 4. Sec. 5 demonstrates the power of probabilistic integration in a variety of challenging applications and Sec. 6 surveys the remaining challenges in this area.

2 Background

Existing mathematical numerical analysis underpins our investigation. Challenging integration problems arise in almost every area of the sciences, engineering and applied mathematics. The development of high-performance approximations remains a central research problem in the numerical analysis and statistics communities. For this paper, we follow the majority of this literature and frame numerical integration as a problem of quadrature. To begin we review the reproducing kernel approach to numerical quadrature and describe the associated theoretical analysis.

2.1 Quadrature Rules and Numerical Error Analysis

2.1.1 Set-up and Notation

This paper considers the problem of computing integrals over a d-dimensional measure space $\mathcal{X}$ whose measure is denoted by σ. Our integrand is a measurable function $f : \mathcal{X} \to \mathbb{R}$ or $\mathbb{C}$ whose expectation $\Pi[f]$ we seek with respect to a probability measure Π. The measure Π is assumed to admit a density with respect to the reference measure σ, denoted by $\pi : \mathcal{X} \to [0, \infty)$. Write $\|f\|_2 := \left(\int_{\mathcal{X}} f(\boldsymbol{x})^2 \pi(\boldsymbol{x}) \mathrm{d}\sigma(\boldsymbol{x}) \right)^{1/2}$ and write $L^2(\Pi)$ for the set of functions which are square-integrable with respect to Π (i.e. $\|f\|_2 < \infty$). For vector arguments we also define $\|\boldsymbol{u}\|_2 = (u_1^2 + \cdots + u_d^2)^{1/2}$. We will make use of the notation $[u]_+ = \max\{0, u\}$.

A *quadrature rule* is any method for approximating integrals $\Pi[f]$ that can be written in the form

$$\hat{\Pi}[f] = \sum_{i=1}^{n} w_i f(\boldsymbol{x}_i), \tag{2}$$

for $n \in \mathbb{N}$ states $\{\boldsymbol{x}_i\}_{i=1}^n \subset \mathcal{X}$ and weights $\{w_i\}_{i=1}^n \subset \mathbb{R}$. The term *cubature rule* is sometimes used when the domain of integration is multi-dimensional (i.e. $d > 1$), although the two terms are often used interchangeably. The notation $\hat{\Pi}[f]$ is motivated by the fact that this expression can be re-written as the integral of f with respect to an empirical measure $\hat{\Pi}$ with density $\hat{\pi}(\boldsymbol{x}) = \sum_{i=1}^n w_i \delta(\boldsymbol{x} - \boldsymbol{x}_i)$, where $\delta(\cdot)$ is the Dirac delta measure and the weights w_i can be negative and need not sum to one. Well-known quadrature rules include (in $d = 1$ dimension) the Newton-Coates rules (trapezoid rule, midpoint rule, Simpson's rule), and Gaussian quadrature (Gauss-Legendre, Gauss-Hermite, Chebyshev-Gauss). In settings where f has additional regularity structure, π is unavailable or non-standard, or $\mathcal{X}$ is irregular or high-dimensional, these numerical rules can be

insufficient for practical purposes, motivating more efficient methods (including MCMC and QMC; see below).

2.1.2 Quadrature in Reproducing Kernel Hilbert Spaces

Analysis of the approximation properties of quadrature rules is naturally performed in terms of function spaces, and in particular in terms of reproducing kernel Hilbert space (RKHS) (e.g. Bach, 2015). Consider a Hilbert space $\mathcal{H}$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and associated norm $\| \cdot \|_{\mathcal{H}}$. $\mathcal{H}$ is said to be an RKHS if there exists a symmetric, positive definite function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ or $\mathbb{C}$, called a *kernel*, that satisfies two properties: **(1)** $k(\cdot, \boldsymbol{x}) \in \mathcal{H}$ for all $\boldsymbol{x} \in \mathcal{X}$ and; **(2)** $f(\boldsymbol{x}) = \langle f, k(\cdot, \boldsymbol{x}) \rangle_{\mathcal{H}}$ for all $\boldsymbol{x} \in \mathcal{X}$ and $f \in \mathcal{H}$ (the *reproducing* property). It can be shown that every kernel defines an RKHS and every RKHS admits a unique reproducing kernel (Berlinet and Thomas-Agnan, 2004, Sec. 1.3). For simplicity of presentation we generally assume that functions are real-valued below. In this paper all kernels k are assumed to satisfy

(A1) $\int_{\mathcal{X}} k(\boldsymbol{x}, \boldsymbol{x}) \pi(\boldsymbol{x}) \mathrm{d}\sigma(\boldsymbol{x}) < \infty$,

which guarantees $f \in L^2(\Pi)$ for all $f \in \mathcal{H}$. Indeed for $R = \int k(\boldsymbol{x}, \boldsymbol{x}) \pi(\boldsymbol{x}) \mathrm{d}\sigma(\boldsymbol{x}) < \infty$, we can upper bound $\|f\|_2$ using the reproducing property and Cauchy-Schwarz:

$$\|f\|_2^2 = \int_{\mathcal{X}} f(\boldsymbol{x})^2 \pi(\boldsymbol{x}) \mathrm{d}\sigma(\boldsymbol{x}) \leq \int_{\mathcal{X}} \|f\|_{\mathcal{H}}^2 k(\boldsymbol{x}, \boldsymbol{x}) \pi(\boldsymbol{x}) \mathrm{d}\sigma(\boldsymbol{x}) = R \|f\|_{\mathcal{H}}^2.$$

We refer the reader to Chapter 1 of Berlinet and Thomas-Agnan (2004) for properties and detailed examples of RKHSs. For an RKHS $\mathcal{H}$ with kernel k we define the *kernel mean* map $\mu_\pi : \mathcal{X} \to \mathbb{R}$ as

$$\mu_\pi(\boldsymbol{x}) := \int_{\mathcal{X}} k(\boldsymbol{x}, \boldsymbol{y}) \pi(\boldsymbol{y}) \mathrm{d}\sigma(\boldsymbol{y}), \tag{3}$$

which exists as an implication of (A1) (Smola et al., 2007). The name is justified by the fact that for all $f \in \mathcal{H}$ we have:

$$\begin{aligned} \Pi[f] = \int_{\mathcal{X}} f(\boldsymbol{x}) \pi(\boldsymbol{x}) \mathrm{d}\sigma(\boldsymbol{x}) &= \int_{\mathcal{X}} \left\langle f, k(\cdot, \boldsymbol{x}) \right\rangle_{\mathcal{H}} \pi(\boldsymbol{x}) \mathrm{d}\sigma(\boldsymbol{x}) \\ &= \left\langle f, \int_{\mathcal{X}} k(\cdot, \boldsymbol{x}) \pi(\boldsymbol{x}) \mathrm{d}\sigma(\boldsymbol{x}) \right\rangle_{\mathcal{H}} = \left\langle f, \mu_\pi \right\rangle_{\mathcal{H}}. \end{aligned}$$

The reproducing property permits an elegant theoretical analysis of quadrature rules, with many quantities of interest tractable analytically in $\mathcal{H}$. In the language of kernel means, quadrature rules of the form in Eqn. 2 can be written in the form $\hat{\Pi}[f] = \langle f, \mu_{\hat{\pi}} \rangle_{\mathcal{H}}$ where $\mu_{\hat{\pi}}$ is the approximation to the kernel mean given by

$$\mu_{\hat{\pi}}(\boldsymbol{x}) = \int_{\mathcal{X}} k(\boldsymbol{x}, \boldsymbol{y}) \hat{\pi}(\boldsymbol{y}) \mathrm{d}\sigma(\boldsymbol{y}) = \sum_{i=1}^{n} w_i k(\boldsymbol{x}, \boldsymbol{x}_i). \tag{4}$$

For fixed $f \in \mathcal{H}$, the integration error associated with $\hat{\Pi}$ can be then expressed as

$$\hat{\Pi}[f] - \Pi[f] = \left\langle f, \mu_{\hat{\pi}} \right\rangle_{\mathcal{H}} - \left\langle f, \mu_\pi \right\rangle_{\mathcal{H}} = \left\langle f, \mu_{\hat{\pi}} - \mu_\pi \right\rangle_{\mathcal{H}}.$$

An upper bound for the error is obtained by applying the Cauchy-Schwarz inequality:

$$|\hat{\Pi}[f] - \Pi[f]| \leq \|f\|_{\mathcal{H}} \|\mu_{\hat{\pi}} - \mu_{\pi}\|_{\mathcal{H}}. \tag{5}$$

The expression above decouples the smoothness (in $\mathcal{H}$) of the integrand f from the approximation accuracy of the kernel mean. Note that the smoothness of f does not depend on the quadrature rule and one can tailor quadrature rules to the approximation of μ_{π}, which in turn does not depend on the particular function f being integrated.

The performance of quadrature rules is usually quantified by the *worst-case error* in the RKHS (Dick et al., 2013), also called *maximum mean discrepancy* (MMD; Smola et al., 2007), given by $\|\hat{\Pi} - \Pi\|_{\text{op}}$ where

$$\|B\|_{\text{op}} := \sup_{\|f\|_{\mathcal{H}} \leq 1} B[f] \tag{6}$$

is the operator norm for bounded linear functionals $B : \mathcal{H} \to \mathbb{R}$. As a measure of quadrature accuracy the MMD is well-studied. Indeed, the above analysis shows that the MMD is characterised as the error in estimating the kernel mean:

Proposition 1. $\|\hat{\Pi} - \Pi\|_{op} = \|\mu_{\hat{\pi}} - \mu_{\pi}\|_{\mathcal{H}}$.

Numerical methods to solve integrals in RKHS thus attempt to minimise the MMD, and we will call *convergence rate* the rate at which this quantity tends to 0 as $n \to \infty$. The formulation of quadrature rules as minimising the MMD is natural and elegant since solving a least-squares problem in the feature space induced by the kernel gives minimax properties in the original space (Schölkopf and Smola, 2002). Indeed, the least-squares formulation is tractable in terms of kernel and kernel mean expressions: Letting $\boldsymbol{w} \in \mathbb{R}^n$ denote the vector of weights $\{w_i\}_{i=1}^n$, $\boldsymbol{z} \in \mathbb{R}^n$ be a vector such that $z_i = \mu_{\pi}(\boldsymbol{x}_i)$, and $\boldsymbol{K} \in \mathbb{R}^{n \times n}$ be the matrix with entries $K_{i,j} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, we have:

Proposition 2. $\|\hat{\Pi} - \Pi\|_{op}^2 = \boldsymbol{w}^T \boldsymbol{K} \boldsymbol{w} - 2\boldsymbol{w}^T \boldsymbol{z} + \Pi[\mu_{\pi}]$.

Proof. Direct calculation gives that

$$\begin{aligned} \|\mu_{\hat{\pi}} - \mu_{\pi}\|_{\mathcal{H}}^2 &= \sum_{i,j=1}^{n} w_i w_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) - 2 \sum_{i=1}^{n} w_i \int_{\mathcal{X}} k(\boldsymbol{x}, \boldsymbol{x}_i) \pi(\boldsymbol{x}) \mathrm{d}\sigma(\boldsymbol{x}) \\ &\quad + \int_{\mathcal{X}} \int_{\mathcal{X}} k(\boldsymbol{x}, \boldsymbol{y}) \pi(\boldsymbol{x}) \pi(\boldsymbol{y}) \mathrm{d}\sigma(\boldsymbol{x}) \mathrm{d}\sigma(\boldsymbol{y}) \;=\; \boldsymbol{w}^T \boldsymbol{K} \boldsymbol{w} - 2\boldsymbol{w}^T \boldsymbol{z} + \Pi[\mu_{\pi}] \end{aligned}$$

and the result follows immediately by applying Prop. 1. □

Several optimality properties for integration in RKHS were proven by Bakhvalov (1971) and collated in Sec. 4.2 of Novak and Woźniakowski (2008). Relevant to this work is the following:

Proposition 3. *An optimal (i.e. minimax) estimate* $\hat{\Pi}$ *can, without loss of generality, be taken in the form a quadrature rule (i.e. of the form* $\hat{\Pi}$ *in Eqn. 2).*

Remark 1. *Prop. 3 motivates restriction to the class of quadrature rules. Indeed, any non-linear estimator or so-called* adaptive *estimator, that learn about* f *"on-the-fly", can be matched in terms of accuracy by a quadrature rule as defined above.*

To obtain an optimal quadrature rule, the expression in Prop. 2 must be minimised in terms of both weights and states. Given states $\{\boldsymbol{x}_i\}_{i=1}^n$, this defines a convex minimisation problem over $\boldsymbol{w} \in \mathbb{R}^n$ whose solution is $\boldsymbol{w} = \boldsymbol{K}^{-1}\boldsymbol{z}$. Optimisation over $\{\boldsymbol{x}_i\}_{i=1}^n$ is, however, challenging (Minka, 2000; Chen et al., 2010). This paper proposes to exploit the well-known sampling efficiency of advanced MCMC and QMC methodologies to address this challenge, but Firstly, we review the Bayesian approach to numerical integration.

2.2 Probabilistic Integration

The probabilistic approach to integration was first clearly stated by Diaconis (1988) and later by O'Hagan (1991), who introduced the BQ nomenclature. Subsequent contributions include Minka (2000); Rasmussen and Ghahramani (2002); Osborne (2010); Huszar and Duvenaud (2012); Gunter et al. (2014) and Briol et al. (2015).

2.2.1 Bayesian Quadrature

Probabilistic integration begins by defining both a space $\mathcal{F}$ of functions $f : \mathcal{X} \to \mathbb{R}$ along with a prior probability measure over $\mathcal{F}$. This could in principle take any form, such as a student's t-process (Shah et al., 2014), a Mondrian forest (Lakshminarayanan et al., 2015) or a Bayesian neural network (Snoek et al., 2015). BQ models prior uncertainty over the integrand f with a Gaussian process (GP). Note that this choice of prior for the integrand affords closed-form inference for the integral; more on this below. A GP $\mathcal{GP}(m_0, k_0)$ is characterised by a mean function $m_0 : \mathcal{X} \to \mathbb{R}$ and a covariance function $k_0 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. From Loève's theorem, the set of valid covariance functions is exactly the set of valid kernel functions. In this paper, prior information takes the form "$f \in \mathcal{H}(k)$" for some RKHS $\mathcal{H}$ whose kernel is k. GPs are therefore a natural choice of probability model since this information can reasonably[1] be encoded by a GP with mean m_0 and covariance function $k_0 = k$.

Conditioning on data $\mathcal{D} = \{\boldsymbol{x}_i, f_i\}_{i=1}^n$ where $f_i = f(\boldsymbol{x}_i)$, we obtain a posterior, denoted $\mathbb{P}$, of the form $f \sim \mathcal{GP}(m_1, k_1)$. Write $\boldsymbol{f} \in \mathbb{R}^n$ for the vector of f_i values, $\boldsymbol{m}_0 \in \mathbb{R}^n$ for the vector of $m_0(\boldsymbol{x}_i)$ values, let $X = \{\boldsymbol{x}_i\}_{i=1}^n$ and write $\boldsymbol{k}(\boldsymbol{x}, X) = \boldsymbol{k}(X, \boldsymbol{x})^T$ for the $1 \times n$ vector whose ith entry is $k(\boldsymbol{x}, \boldsymbol{x}_i)$. Then, following Rasmussen and Williams (2006),

$$m_1(\boldsymbol{x}) = m_0(\boldsymbol{x}) + \boldsymbol{k}(\boldsymbol{x}, X)\boldsymbol{K}^{-1}(\boldsymbol{f} - \boldsymbol{m}_0) \tag{7}$$

$$k_1(\boldsymbol{x}, \boldsymbol{x}') = k(\boldsymbol{x}, \boldsymbol{x}') - \boldsymbol{k}(\boldsymbol{x}, X)\boldsymbol{K}^{-1}\boldsymbol{k}(X, \boldsymbol{x}'). \tag{8}$$

This GP provides a full distribution over functions $f \in \mathcal{F}$ which are consistent with both prior knowledge and data. Inversion of the kernel matrix is an expensive $O(n^3)$ operation; however we are motivated by cases where f is expensive to evaluate or states are expensive to select, where the costs of kernel computation are outweighed by the full probabilistic description offered by the GP, along with better estimates for the integral. This probabilistic description, further, may make more efficient state selection possible through the use of decision theory, resulting in the possibility of net

[1] Technically the subset $\mathcal{H}(k_0) \subset \mathcal{F}$ has measure zero under $\mathcal{GP}(m_0, k_0)$; this is unsatisfactory from a Bayesian perspective because the information "$f \in \mathcal{H}(k)$" is not faithfully encoded by the GP when $k_0 = k$. To properly encode an RKHS $\mathcal{H}(k)$ one can construct a covariance function $k_0(\boldsymbol{x}, \boldsymbol{y}) = \int_{\mathcal{X}} k(\boldsymbol{x}, \boldsymbol{z})k(\boldsymbol{z}, \boldsymbol{y})\pi(\boldsymbol{z})\mathrm{d}\sigma(\boldsymbol{z})$ that dominates k, in the sense of Lukic and Beder (2001). For simplicity of presentation we do not draw a distinction between k_0 and k in the main text.

computational savings for a desired level of accuracy (see Sec. 2.2.2). A sketch of the procedure is provided in Figure 1.

Denote by $\mathbb{E}[\cdot|\mathcal{D}]$, $\mathbb{V}[\cdot|\mathcal{D}]$ the expectation and variance taken with respect to the posterior distribution $\mathbb{P}[\cdot|\mathcal{D}]$ over f given data $\mathcal{D}$. As integration is a linear functional, we obtain a Gaussian posterior distribution over the value of the integral:

Proposition 4. *In the posterior, the integral $\Pi[f]$ is Gaussian with*

$$\mathbb{E}[\Pi[f]|\mathcal{D}] = \Pi[m_0] + \boldsymbol{z}^T \boldsymbol{K}^{-1}(\boldsymbol{f} - \boldsymbol{m}_0) \tag{9}$$

$$\mathbb{V}[\Pi[f]|\mathcal{D}] = \Pi[\mu_\pi] - \boldsymbol{z}^T \boldsymbol{K}^{-1}\boldsymbol{z}. \tag{10}$$

where $\boldsymbol{z}$ is a $n \times 1$ vector containing evaluations of the kernel mean $z_i = \mu_\pi(\boldsymbol{x}_i)$.

Proof. An application of Fubini's theorem produces

$$\begin{aligned}
\mathbb{E}[\Pi[f]|\mathcal{D}] &= \mathbb{E}\left[\int_{\mathcal{X}} f(\boldsymbol{x})\pi(\boldsymbol{x})\mathrm{d}\sigma(\boldsymbol{x})\middle|\mathcal{D}\right] = \int_{\mathcal{X}} \mathbb{E}[f(\boldsymbol{x})|\mathcal{D}]\pi(\boldsymbol{x})\mathrm{d}\sigma(\boldsymbol{x}) = \int_{\mathcal{X}} m_1(\boldsymbol{x})\pi(\boldsymbol{x})\mathrm{d}\sigma(\boldsymbol{x}) \\
\mathbb{V}[\Pi[f]|\mathcal{D}] &= \int_{\mathcal{F}} \left[\int_{\mathcal{X}} f(\boldsymbol{x})\pi(\boldsymbol{x})\mathrm{d}\sigma(\boldsymbol{x}) - \int_{\mathcal{X}} m_1(\boldsymbol{x})\pi(\boldsymbol{x})\mathrm{d}\sigma(\boldsymbol{x})\right]^2 d\mathbb{P}[f|\mathcal{D}] \\
&= \int_{\mathcal{X}}\int_{\mathcal{X}}\int_{\mathcal{F}} [f(\boldsymbol{x}) - m_1(\boldsymbol{x})][f(\boldsymbol{x}') - m_1(\boldsymbol{x}')]d\mathbb{P}[f|\mathcal{D}]\pi(\boldsymbol{x})\pi(\boldsymbol{x}')\mathrm{d}\sigma(\boldsymbol{x})\mathrm{d}\sigma(\boldsymbol{x}') \\
&= \int_{\mathcal{X}}\int_{\mathcal{X}} k_1(\boldsymbol{x}, \boldsymbol{x}')\pi(\boldsymbol{x})\pi(\boldsymbol{x}')\mathrm{d}\sigma(\boldsymbol{x})\mathrm{d}\sigma(\boldsymbol{x}').
\end{aligned}$$

The proof is completed by substituting m_1 and k_1 from Eqns. 7 and 8. $\Pi[\mu_\pi]$ exists by (A1). □

We now have a probabilistic model for the epistemic uncertainty over the value of the integral that is due to employing a quadrature rule with a finite number n of function evaluations. By re-parametrising $f \mapsto f - m_0$ we can, without loss of generality, suppose that $m_0 \equiv 0$ for the remainder of the paper. Then the posterior mean takes the form of a quadrature rule

$$\hat{\Pi}_{\mathrm{BQ}}[f] := \sum_{i=1}^{n} w_i^{\mathrm{BQ}} f(\boldsymbol{x}_i) \tag{11}$$

where $\boldsymbol{w}^{\mathrm{BQ}} := \boldsymbol{K}^{-1}\boldsymbol{z}$. This BQ rule happens to have strong approximation properties: Huszar and Duvenaud (2012) point out that Eqn. 10 is identical to the expression in Prop. 2 with optimally chosen weights $\boldsymbol{w} = \boldsymbol{w}^{\mathrm{BQ}}$, so that the posterior variance is exactly equal to the worst case error (MMD) squared. $\hat{\Pi}_{\mathrm{BQ}}$ is therefore minimax over all quadrature rules based on the (fixed) states $\{\boldsymbol{x}_i\}_{i=1}^n$. The posterior variance $\mathbb{V}[\Pi[f]|\mathcal{D}]$ does not depend on function values $\{f_i\}_{i=1}^n$, but only on the location of the states $\{\boldsymbol{x}_i\}_{i=1}^n$ and the kernel of $\mathcal{H}$. This is useful as it allows us to pre-compute state locations that can be used to integrate multiple integrals within the same RKHS $\mathcal{H}$.

For BQ, weights are automatically constrained to be $\boldsymbol{w}^{\mathrm{BQ}} = \boldsymbol{K}^{-1}\boldsymbol{z}$ but there is flexibility in the selection of states $\{\boldsymbol{x}_i\}_{i=1}^n$ and several proposals appear in the literature. For example O'Hagan (1991) used classical Gauss-Hermite states, Rasmussen and Ghahramani (2002) generated states using MC and Osborne (2010); Briol et al. (2015) selected states by targeting posterior variance, both directly and indirectly. As noted in Huszar and Duvenaud (2012), the selection of states involves an exploration-exploitation trade-off. First, as states concentrate around regions of high

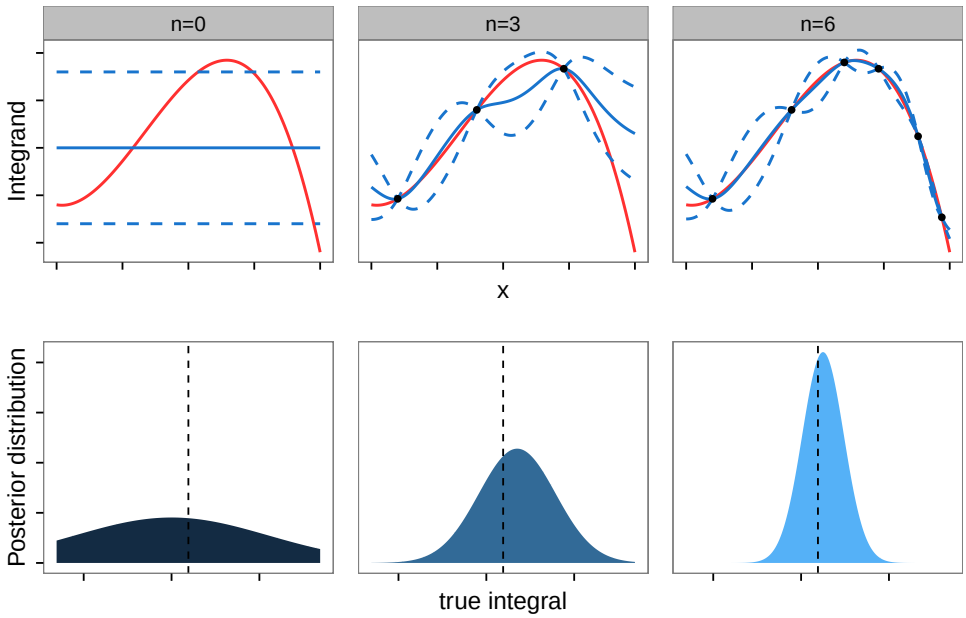


Figure 1: Sketch of Bayesian Quadrature. The top row shows the approximation of the integrand f (in red) by the GP posterior mean m_1 (in blue) as the number n of function evaluations is increased. The dashed lines represent 95% credible intervals. The bottom row shows the Gaussian distribution with mean $\mathbb{E}[\Pi[f]|\mathcal{D}]$ and variance $\mathbb{V}[\Pi[f]|\mathcal{D}]$ that models our uncertainty over the solution of the integral as n increases (the dashed black line gives the true value of the integral). When $n = 0$, the approximation of the integral is fully specified by the GP prior. As the number of states n increases, the approximation of f becomes more precise and the Gaussian posterior distribution contracts onto the true value of the integral.

probability mass under Π, the values of the kernel mean vector $\boldsymbol{z}$ will increase and the posterior variance (Eqn. 10) will decrease accordingly. This therefore encourages exploitation of the density. However, as design points get closer to each other, the eigenvalues of $\boldsymbol{K}$ will increase and therefore the eigenvalues of $\boldsymbol{K}^{-1}$ will decrease, leading to an increase of the posterior variance. This therefore encourages exploration of the density. The following sections discuss different schemes for selection of states that aim to address this trade-off.

2.2.2 Optimisation-Based Quadrature Rules

An *Optimal* BQ (OBQ) rule selects states $\{\boldsymbol{x}_i\}_{i=1}^n$ to globally minimise the posterior variance (equivalent to globally minimising the MMD). It is known that OBQ corresponds to classical quadrature rules (e.g. Gauss-Hermite) for specific choices of RKHS $\mathcal{H}$ (Diaconis, 1988; O'Hagan, 1991; Särkka et al., 2015). Nevertheless, optimal set points can rarely be found analytically. In most cases, OBQ is unfortunately intractable in practice as the corresponding optimization problem is usually NP-hard (Schölkopf and Smola, 2002, Sec. 10.2.3).

A pragmatic approach to select states is the greedy algorithm, sequentially minimising the posterior variance at each iteration. This rule, called *Sequential* BQ (SBQ), is straightforward to implement, e.g. using general-purpose numerical optimisation, and is a probabilistic integration

method that is often used in practice (Osborne et al., 2012; Gunter et al., 2014). Recently, more sophisticated optimization algorithms have been used to select states. For example, Briol et al. (2015) used conditional gradient algorithms (also called Frank-Wolfe algorithm (Lacoste-Julien et al., 2015) or kernel herding (Chen et al., 2010)) that, in effect, produce a linear approximation to the posterior variance based on its derivative. This method, called *Frank-Wolfe* BQ (FWBQ) notably provided the first results for convergence of a general-purpose BQ method (which was shown to be up to exponential) and posterior contraction (up to super-exponential).

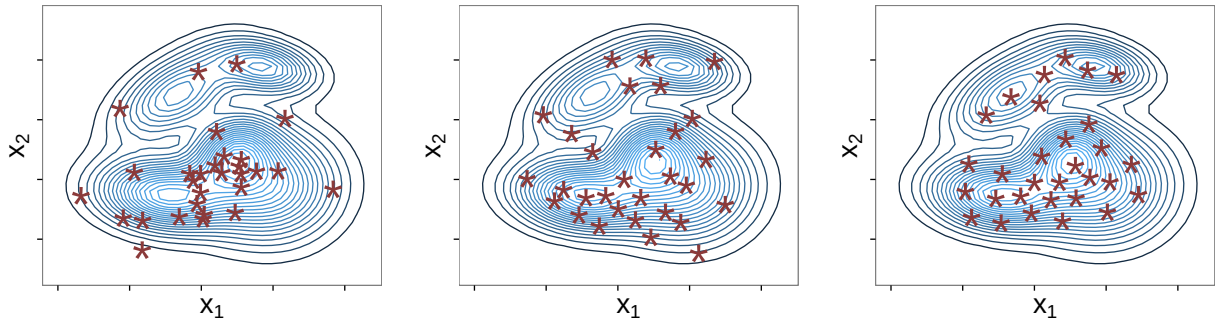
However there are a number of weaknesses with SBQ and FWBQ that motivate the present work. Firstly, they do not scale well to high-dimensional settings due to the need to repeatedly solve high-dimensional optimisation problems. For selection of states, (MC)MC and QMC methods offer considerable potential and this is our focus below. Secondly, theoretical guarantees for FWBQ only hold in finite dimensional RKHS, while nothing at all is known for SBQ. The results in this paper apply to infinite-dimensional RKHS.

2.2.3 Monte Carlo and Quasi Monte Carlo Methods

A *Monte Carlo* (MC) method is defined as a quadrature rule based on uniform weights $w_i^{\mathrm{MC}} := 1/n$ and states $\{\boldsymbol{x}_i\}_{i=1}^n$ that are formally considered as random variables. The simplest of those methods consists of independently sampling states independently from Π (Fig. 2, left). For unnormalised densities π, MCMC methods proceed similarly but induce a dependence structure among the $\{\boldsymbol{x}_i\}_{i=1}^n$. In either case we denote the (random) estimators by $\hat{\Pi}_{\mathrm{MC}}$ (when $\boldsymbol{x}_i = \boldsymbol{x}_i^{\mathrm{MC}}$) and $\hat{\Pi}_{\mathrm{MCMC}}$ (when $\boldsymbol{x}_i = \boldsymbol{x}_i^{\mathrm{MCMC}}$) respectively. Uniformly weighted estimators are well-suited to many challenging integration problems since they provide a dimension-independent convergence rate for the MMD of $O_P(n^{-1/2})$ (Thm. 5 below). They are also widely applicable and straight-forward to analyse; for instance the central limit theorem (CLT) gives that $\sqrt{n}(\hat{\Pi}_{\mathrm{MC}}[f] - \Pi[f]) \to \mathcal{N}(0, \tau_f^{-1})$ where $\tau_f^{-1} = \Pi[f^2] - \Pi[f]^2$ and the convergence is in distribution. However, the CLT is not well-suited as a measure of *epistemic* uncertainty (i.e. as an explicit model for numerical error) since (i) it is only valid asymptotically, and (ii) τ_f is unknown, depending on the integral $\Pi[f]$ that we are trying to compute. This motivates instead probabilistic integration for the class of MC estimators (i.e. BMC; Rasmussen and Ghahramani, 2002).

A related class of methods is QMC. These methods exploit knowledge of the RKHS $\mathcal{H}$ to spread the states in an efficient, deterministic way over the domain $\mathcal{X}$ (Figure 2, middle). QMC also approximates integrals using a quadrature rule $\hat{\Pi}_{\mathrm{QMC}}[f]$ that has uniform weights $w_i^{\mathrm{QMC}} := 1/n$. These methods benefit from an extensive theoretical literature (Dick and Pillichshammer, 2010). The (in some cases) optimal convergence rates as well as sound statistical properties of QMC have led to interest in the machine learning and statistics communities (e.g. Rahimi and Recht, 2007; Yang et al., 2014; Gerber and Chopin, 2015; Oates and Girolami, 2015).

Remark 2. *The restriction of MC methods to uniform weights can be motivated by the fact that the class of uniform-weighted estimators is rich enough to find estimators that achieve optimal convergence rates (Novak and Woźniakowski, 2010). However this is a "fixed d" result and, even for QMC methods, optimal convergence rates as $n \to \infty$ usually come at the cost of a rate constant C_d that diverges as $d \to \infty$. Non-uniformity of weights has been suggested as one possible solution to the dimensionality problem (Novak and Woźniakowski, 2010, p109). There have been several attempts at constructing rigorous methods based on non-uniform weights. Examples include the universal algorithm of Krieg and Novak (2015), which gives non-uniform but positive weights, and Smolyak*



(a) MC (b) QMC (c) FW

Figure 2: Illustration of states used for quadrature, based on a Gaussian mixture Π. (a) Monte Carlo (MC) sampling from Π. (b) A *Sobol sequence* - a specific type of Quasi MC (QMC) point sequence - mapped to Π. (c) States obtained using the Frank-Wolfe (FW) algorithm. QMC and FW usually far outperfom MC due to their better coverage of Π.

algorithms (Novak and Woźniakowski, 2010, Chapter 15) that also allow for negative weights. Both algorithms provide better rates than uniform-weight rules for high-dimensional integration. The methods that we propose below are based on non-uniform weights and in Sec. 5.2.3 we present results in a high-dimensional setting.

Our goal is to establish probabilistic integrators based on both (MC)MC and QMC.

3 Methods

In this section we outline our theoretical framework for establishing consistency and contraction of BQ estimators based on (MC)MC and QMC states.

3.1 Bayesian QMC and Bayesian (MC)MC

(MC)MC and QMC methods have been extensively studied in the literature. Relative to existing optimisation-based approaches, like SBQ and FWBQ, they are computationally inexpensive, more widely applicable and well-suited to many challenging integration problems, such as in high-dimensions where optimisation algorithms are known to struggle. This makes them well placed to generate states for use in BQ. We pursue these ideas in detail below.

This paper studies the two-step procedure that first uses (MC)MC or QMC in order to select states and then assigns BQ (minimax) weights to those states. Thus we define

$$\hat{\Pi}_{\mathrm{BMC}}[f] := \sum_{i=1}^{n} w_i^{\mathrm{BQ}} f(\boldsymbol{x}_i^{\mathrm{MC}}) \tag{12}$$

$$\hat{\Pi}_{\mathrm{BQMC}}[f] := \sum_{i=1}^{n} w_i^{\mathrm{BQ}} f(\boldsymbol{x}_i^{\mathrm{QMC}}) \tag{13}$$

$$\hat{\Pi}_{\mathrm{BMCMC}}[f] := \sum_{i=1}^{n} w_i^{\mathrm{BQ}} f(\boldsymbol{x}_i^{\mathrm{MCMC}}) \tag{14}$$

This two-step procedure means that no modification to existing (MC)MC or QMC sampling schemes is necessary. Moreover each estimator is associated with a full posterior probability distribution described in Sec. 2.2.

BMC was first described by Rasmussen and Ghahramani (2002) while BQMC has been described by Hickernell et al. (2005); Marques et al. (2013); Särkka et al. (2015). To date we are not aware of any theoretical analysis of the BQ posterior distributions that are associated with these methods. The goal of the next section is to establish theoretical guarantees for consistency of these point estimators and contraction of their associated posterior distributions.

3.2 Proof Techniques for Consistency and Contraction

We begin by establishing consistency of probabilistic integrators and discuss how the rate of convergence depends on (i) the RKHS $\mathcal{H}$, (ii) the selection of states $\{\boldsymbol{x}_i\}_{i=1}^n$ and (iii) the domain of integration $\mathcal{X}$. Following this, we then turn to posterior contraction.

We describe three proof techniques to obtain convergence and contraction rates for all probabilistic integration methods that feature in this paper. The first, below, is a generalization of Briol et al. (2015, Thm. 1):

Lemma 1 (Bayesian re-weighting)**.** *Consider the quadrature rule* $\hat{\Pi}[f] = \sum_{i=1}^n w_i f(\boldsymbol{x}_i)$ *and the corresponding re-weighted rule* $\hat{\Pi}_{BQ}[f] = \sum_{i=1}^n w_i^{BQ} f(\boldsymbol{x}_i)$*. Suppose we have a convergence rate* δ_n *for* $\hat{\Pi}$ *(i.e.* $\|\hat{\Pi} - \Pi\|_{op} \leq \delta_n$*). Then* $\|\hat{\Pi}_{BQ} - \Pi\|_{op} \leq \delta_n$*.*

Proof. From Prop. 2 we have

$$\|\hat{\Pi} - \Pi\|_{\text{op}}^2 = \boldsymbol{w}^T \boldsymbol{K} \boldsymbol{w} - 2\boldsymbol{w}^T \boldsymbol{z} + \Pi[\mu_\pi].$$

The right hand side is minimised by $\boldsymbol{w} = \boldsymbol{w}^{\text{BQ}} = \boldsymbol{K}^{-1}\boldsymbol{z}$ and the value at the minimum is $\|\hat{\Pi}_{\text{BQ}} - \Pi\|_{\text{op}}^2$. □

Thus probabilistic integrators obtained by re-weighting existing quadrature rules will be *at least as good* as their non-probabilistic versions. Probabilistic integrators can in practice be several orders of magnitude more accurate than their non-probabilistic counterparts (Huszar and Duvenaud, 2012).

A second approach to obtain convergence rates is to look at probabilistic integration as a functional approximation problem. The convergence rate of quadrature rules can be shown to be at least as good as the corresponding functional approximation rates in $L^2(\Pi)$. (The converse also holds; see Bach (2015, Sec. 3.4).) This is summarised as follows:

Lemma 2 (Regression bound)**.** *Fix states* $X = \{\boldsymbol{x}_i\}_{i=1}^n$*. Then we have* $|\hat{\Pi}_{BQ}[f] - \Pi[f]| \leq \|f - \mathbb{E}[f|\mathcal{D}]\|_2$*, where* $\hat{\Pi}_{BQ}$ *is the BQ rule based on* X*.*

Proof. From linearity and Gaussianity we have

$$\hat{\Pi}_{\text{BQ}}[f] = \int_{\mathcal{X}} \mathbb{E}[f(\boldsymbol{x})|\mathcal{D}]\pi(\boldsymbol{x})\mathrm{d}\sigma(\boldsymbol{x})$$

For the BQ estimate Jensen's inequality leads us to see that

$$\begin{aligned} |\hat{\Pi}_{\text{BQ}}[f] - \Pi[f]|^2 &= \left(\int_{\mathcal{X}} f(\boldsymbol{x})\pi(\boldsymbol{x})\mathrm{d}\sigma(\boldsymbol{x}) - \int_{\mathcal{X}} \mathbb{E}[f(\boldsymbol{x})|\mathcal{D}]\pi(\boldsymbol{x})\mathrm{d}\sigma(\boldsymbol{x})\right)^2 \\ &\leq \int_{\mathcal{X}} (f - \mathbb{E}[f|\mathcal{D}])^2(\boldsymbol{x})\pi(\boldsymbol{x})\mathrm{d}\sigma(\boldsymbol{x}) = \|f - \mathbb{E}[f|\mathcal{D}]\|_2^2, \end{aligned}$$

as required. □

Lemmas 1 and 2 refer to the point estimators provided by BQ rules. However, our primary focus is to quantify the change in probability mass as the number of samples increases. In the case of probabilistic integrators, the posterior probability mass concentrates around the true value of the integral as n increases; this is called *posterior contraction*. Theorem 3 below formalises this result and shows that, for BQ, point estimator consistency implies posterior contraction. For measurable A we write $\mathbb{P}[A|\mathcal{D}] = \mathbb{E}[1_A|\mathcal{D}]$ where 1_A is the indicator function of the event A.

Lemma 3 (BQ contraction)**.** *Assume $f \in \mathcal{H}$. Suppose that $\|\hat{\Pi}_{BQ} - \Pi\|_{op} \leq \delta_n$ where $\delta_n \downarrow 0$. Define $I_D = (\Pi[f] - D, \Pi[f] + D)$ to be an open interval of diameter $2D$ centred on the true integral. Then $\mathbb{P}[I_D^c|\mathcal{D}]$, the posterior mass on $I_D^c = \mathbb{R} \setminus I_D$, vanishes at the rate*

$$\mathbb{P}[I_D^c|\mathcal{D}] = o(\exp(-C\delta_n^{-2})) \tag{15}$$

where $C = D^2/2$.

The proofs of Lemma 3 and subsequent results are reserved for Appendix A.

This result demonstrates that the posterior distribution is well-behaved; probability mass tends to zero outside of any open neighbourhood of the true solution as n increases. Hence, if our prior is well calibrated (see Sec. 4.1), the posterior distribution provides an appropriate description of epistemic uncertainty over the solution of the integral as a result of performing a finite number n of computations.

As a self-contained introduction of the proof techniques established above, in Appendix B we obtain a convergence rate for OBQ as originally formulated in the seminal paper of O'Hagan (1991). These techniques will be used below to establish theoretical properties for BQMC and B(MC)MC.

3.3 Theoretical Results

We have now established that to prove posterior mean convergence and posterior contraction, it is sufficient to prove convergence of the MMD (Lemma 3). In this section we explore some immediate consequences of this result for Sobolev-like spaces by leveraging established theory on MMD convergence. This will allow us in Sec. 5 to provide theoretical guarantees on several problems of practical importance in machine learning and computer vision.

3.3.1 Bayesian (MC)MC

Analysis of MCMC methods deals with the rate constant, while rates themselves scale as the MC rates. In this section we therefore focus on BMC (Rasmussen and Ghahramani, 2002), leaving the analysis of rate constants for BMCMC as future work. We begin by providing a general result for MC estimation. This requires a slight strengthening of (A1):

(A2) $k_{\max} := \sup_{\boldsymbol{x} \in \mathcal{X}} k(\boldsymbol{x}, \boldsymbol{x})^{1/2} < \infty$.

This implies that f is bounded on $\mathcal{X}$.

Recall that in MC, states $\{\boldsymbol{x}_i^{\mathrm{MC}}\}_{i=1}^n$ are sampled independently from Π and weighted uniformly. For MC estimators the MMD converges at the classical (dimension-independent) MC rate (e.g. Altun and Smola, 2006, Thm. 15):

Proposition 5 (MC Methods)**.** *Under (A2) we have* $\|\hat{\Pi}_{MC} - \Pi\|_{op} = O_P(n^{-1/2})$.

Prop. 5 exemplifies a powerful framework in which to study the convergence properties of (MC)MC methods in an RKHS that has become popular in machine learning (Smola et al., 2007). In the case of B(MC)MC, the regression bound (Lemma 2) enables us to obtain rates for the MMD that improve on the MC rate in certain cases.

When the states are random variables, it is possible to discuss the average-case scenario. Let $\mathcal{X} = [0,1]^d$, Π be uniform and σ be the Lebesgue measure. Write $\mathfrak{F}$ for the Fourier transform operator. Define the *Sobolev space*

$$\mathcal{H}_\alpha := \{f \in L^2(\Pi) \text{ such that } \|f\|_{\mathcal{S},\alpha} < \infty\}, \tag{16}$$

equipped with the norm

$$\|f\|_{\mathcal{S},\alpha} := \|\mathfrak{F}^{-1}[(1 + \|\boldsymbol{\xi}\|_2^2)^{\alpha/2}\mathfrak{F}[f]]\|_2. \tag{17}$$

Here α is the *order* of the space. It can be shown that $\mathcal{H}_\alpha$ is the set of functions f whose weak derivatives $(\partial x_1)^{u_1} \dots (\partial x_d)^{u_d} f$ exist in $L^2(\Pi)$ for $u_1 + \dots + u_d \leq \alpha$. Any radial kernel whose Fourier transform decays at a rate α (e.g. Matérn kernel) induces an RKHS that is norm-equivalent to $\mathcal{H}_\alpha$.

Theorem 1 (BMC in $\mathcal{H}_\alpha$)**.** *Let $\mathcal{H}$ be an RKHS that is norm-equivalent to $\mathcal{H}_\alpha$, where $\alpha \in \mathbb{N}$ and $\alpha > d/2$. Then*

$$\|\hat{\Pi}_{BMC} - \Pi\|_{op} = O_P(n^{-\alpha/d+\epsilon}) \tag{18}$$

$$\mathbb{P}[I_D^c|\mathcal{D}] = o_P(\exp(-Cn^{2\alpha/d-\epsilon})) \tag{19}$$

where $\epsilon > 0$ can be arbitrarily small.

Remark 3. *$O_P(n^{-\alpha/d-1/2})$ is an information-theoretic lower bound on the performance of any random quadrature rule in $\mathcal{H}_\alpha$ (Novak and Woźniakowski, 2010). Thus BMC converges at a near-optimal rate. During the completion of this work, Bach (2015) obtained a similar result but for fixed n. The focus of that work is different and the analysis does not imply the asymptotic results that we have described.*

3.3.2 Bayesian QMC

In the previous section we showed that BMC is nearly rate-optimal in $\mathcal{H}_\alpha$, so that there is little need to develop BQMC methods in this space (those will in fact also attain this optimal rate). We therefore consider spaces of functions whose *mixed* partial derivatives exist, for which much faster convergence rates can be obtained using QMC methods. To formulate BQMC we consider collections of states $\{\boldsymbol{x}_i\}_{i=1}^n$ that constitute QMC point sequences. Specifically, we consider higher-order digital nets. For the benefit of readers who may not be familiar with QMC, we briefly recall essential definitions in Appendix C, but the reader is referred to Dick and Pillichshammer (2010) for further details.

Let $\mathcal{X} = [0,1]^d$ and σ be the Lebesgue measure. Write $\mathfrak{F}$ for the Fourier transform operator. Define the *Sobolev space of dominating mixed smoothness* as

$$\mathcal{S}_\alpha := \{f \in L^2(\Pi) \text{ such that } \|f\|_{\mathcal{S},\alpha} < \infty\}, \tag{20}$$

equipped with the norm

$$\|f\|_{\mathcal{S},\alpha} := \left\| \mathfrak{F}^{-1}\left[\prod_{i=1}^{d}(1+\xi_i^2)^{\alpha/2}\mathfrak{F}[f]\right]\right\|_2. \tag{21}$$

Here α is the *order* of the space. It can be shown that $\mathcal{S}_\alpha$ is the set of functions f whose weak derivatives $(\partial x_1)^{u_1}\dots(\partial x_d)^{u_d} f$ exist in $L^2(\Pi)$ for $u_i \leq \alpha$, $i = 1,\dots,d$. Moreover $\mathcal{S}_\alpha$ is an RKHS that is norm-equivalent to the RKHS generated by a tensor product of Matérn kernels (Sickel and Ullrich, 2009), or indeed a tensor product of any other univariate Sobolev space -generating kernel.

Theorem 2 (BQMC in $\mathcal{S}_\alpha$)**.** *Let $\mathcal{X} = [0,1]^d$, σ be the Lebesgue measure and take Π to be uniform on $\mathcal{X}$. Let $\mathcal{H}$ be an RKHS that is norm-equivalent to $\mathcal{S}_\alpha$. Consider the BQMC estimator $\hat{\Pi}_{BQMC}$ whose states $\{\boldsymbol{x}_i^{QMC}\}_{i=1}^n$ are a higher-order digital $(t,\alpha,1,\alpha m \times m, d)$ net over $\mathbb{Z}_b$ for some prime b where $n = b^m$ (defined in Appendix C). Then we have*

$$\|\hat{\Pi}_{BQMC} - \Pi\|_{op} = O(n^{-\alpha+\epsilon}), \tag{22}$$

$$\mathbb{P}[I_D^c|\mathcal{D}] = o(\exp(-Cn^{2\alpha-\epsilon})), \tag{23}$$

where $\epsilon > 0$ can be arbitrarily small.

Remark 4. *This result is optimal for any deterministic quadrature rule in $\mathcal{S}_\alpha$ (Dick, 2011). These results should be understood to hold on the sub-sequence $n = b^m$; indeed QMC methods cannot give guarantees for all $n \in \mathbb{N}$ (Owen, 2014).*

Remark 5. *In Sec. 5.2.3 we discuss the possibility of constructing BQMC rules in high-dimensional spaces by considering weighted versions of Sobolev spaces of dominating mixed smoothness.*

In practice many of the integration problems that we face actually involve integrands f that are infinitely differentiable, but are expensive to evaluate. We therefore provide additional results, in Appendix D, that cover spaces of infinitely differentiable functions. The strong prior assumption of infinite differentiability leads to exponential convergence of the MMD as the number of states n goes to infinity.

This concludes our theoretical analysis. We have established optimal and near-optimal rates of convergence (and hence contraction) for both BMC and BQMC in a general function space setting. This directly addresses the criticism that BQ lacks theoretical foundations. In the following section we turn to methodological considerations that are relevant to implementation of these methods.

4 Implementation

Below we discuss a number of practical considerations that are important in applications of BQMC and B(MC)MC, as well as some methodological extensions. Additionally we have described an extension that can produce unbiased estimates (Appendix E) and provided a discussion of scalability for BQ (Appendix H).

4.1 Calibration

The theoretical results above deal with asymptotic scaling, but a question remains on whether the posterior uncertainty is well-calibrated for finite values of n. i.e. whether the scale of the posterior

uncertainty matches the scale of the actual numerical error. A particular distinction of B(MC)MC from BQMC and optimisation-based schemes (see Sec. 2.2.2) is that the choice of states $\{\boldsymbol{x}_i\}_{i=1}^n$ does not depend on the kernel. This is on one hand a weakness, since we do not leverage the kernel to cleverly select states, but on the other hand a strength, since this permits fully off-line learning of the kernel (known in statistics as *calibration*) after evaluation of the integrand.

Calibration of B(MC)MC amounts to eliciting appropriate values for kernel hyper-parameters conditional upon the sampled states. In this paper we take an *empirical Bayes* approach, choosing hyper-parameters that maximise the log-marginal likelihood:

$$\log \mathbb{P}(\boldsymbol{f}|\{\boldsymbol{x}_i\}_{i=1}^n) = -\frac{1}{2}\boldsymbol{f}^T \boldsymbol{K}^{-1}\boldsymbol{f} - \frac{1}{2}\log|\boldsymbol{K}| - \frac{n}{2}\log(2\pi). \tag{24}$$

This is guided by the recent analysis of Szabó et al. (2015) who show that empirical Bayes credible sets in the function space $\mathcal{H}$ give correct uncertainty quantification (i.e. correct coverage rates) for sufficiently regular elements f in $\mathcal{H}$ (specifically, for $f \in \mathcal{H}$ that satisfy an additional technical condition known as a *polished tail* condition). The regression bound (Lemma 2) implies that the posterior credible sets for $\Pi[f]$ also provide correct coverage rates when calibrated using empirical Bayes, under the polished tail condition. Although no analytical solution is available for the empirical Bayes hyper-parameters, an approximate solution can easily be obtained numerically. Alternative approaches such as marginalisation of hyper-parameters (e.g. Osborne, 2010; Nickl and Söhl, 2015) or "learning the kernel" (Ong et al., 2005; Duvenaud et al., 2013) could be used but were not considered here.

4.2 Tractable and Intractable Kernel Means

BQ requires that the kernel mean $\mu_\pi(\boldsymbol{x}) = \Pi[k(\cdot, \boldsymbol{x})]$ is available in closed-form. This is the case for several kernel-density pairs (k, π) and a subset of these pairs are recorded in Table 1. These pairs are fairly widely applicable; for example the control functional kernel (Oates et al., 2015) provides a closed-form expression for the kernel mean whenever the gradient $\partial \log \pi(\boldsymbol{x})$ is available; this includes the important setting of Bayesian posterior inference, where the p.d.f. π is only available up to proportionality (see Sec. 4.3).

In the event that the kernel-density pair (k, π) of interest does not lead to a closed-form kernel mean, it is sometimes possible to determine another kernel-density pair (k', π') for which $\Pi'[k'(\cdot, \boldsymbol{x})]$ is available and such that (i) $f\pi/\pi' \in \mathcal{H}(k')$, (ii) $\mathrm{supp}(\pi) \subseteq \mathrm{supp}(\pi')$. Then one can construct an importance sampling estimator

$$\Pi[f] = \int_{\mathcal{X}} f(\boldsymbol{x})\pi(\boldsymbol{x})\mathrm{d}\sigma(\boldsymbol{x}) = \int_{\mathcal{X}} \frac{f(\boldsymbol{x})\pi(\boldsymbol{x})}{\pi'(\boldsymbol{x})}\pi'(\boldsymbol{x})\mathrm{d}\sigma(\boldsymbol{x}) = \Pi'[f\pi/\pi']. \tag{25}$$

and proceed as above (O'Hagan, 1991). Bach (2015) derives an optimal importance sampling distribution for BMC and provides an approximation algorithm when the distribution is not tractable, which greatly widens the applicability of BQ.

However, since $\mathcal{H}$ should represent prior information, such strategies may be seen as lacking statistical justification. We therefore provide a discussion of methods to approximate intractable kernel means in Appendix I. In summary, a MC estimate of the kernel mean based on m samples can be used in place of the exact kernel mean with no loss in efficiency, provided that $m = O(n^{1/2}\delta_n^{-2})$ where δ_n is the rate of the exact BQ estimator. The use of approximate kernel means is not considered further in the present paper because, from a probabilistic numerics point of view, the

$\mathcal{X}$	Π	k	Reference
$[0,1]^d$	Unif($\mathcal{X}$)	Wendland TP	Oates and Girolami (2015)
$[0,1]^d$	Unif($\mathcal{X}$)	Matérn Weighted TP	Sec. 5.2.3
$[0,1]^d$	Unif($\mathcal{X}$)	Korobov TP	Appendix D
$[0,1]^d$	Unif($\mathcal{X}$)	Exponentiated quadratic	Appendix J
$\mathbb{R}^d$	Mixt. of Gaussians	Exponentiated quadratic	O'Hagan (1991)
$\mathbb{S}^d$	Unif($\mathcal{X}$)	Gegenbauer	Sec. 5.2.1
Arbitrary	Unif($\mathcal{X}$) / Mixt. of Gauss.	trigonometric	Integration by parts
Arbitrary	Unif($\mathcal{X}$)	Splines	Minka (2000)
Arbitrary	Known moments	Polynomial TP	Briol et al. (2015)
Arbitrary	Known $\partial \log \pi(\boldsymbol{x})$	Control functional	Sec. 4.3

Table 1: A non-exhaustive list of distribution (Π) and kernel (k) pairs that provide a closed-form expression for the kernel mean ($\mu_\pi(\boldsymbol{x}) = \Pi[k(\cdot, \boldsymbol{x})]$) and the initial error $\Pi[\mu_\pi]$. Here TP refers to the tensor product of one-dimensional kernels.

additional source of uncertainty that is due to numerical error in the kernel mean must also be reflected in the posterior variance (to avoid a philosophical "infinite regress").

4.3 Intractable Densities

Often integration problems involve distributions Π whose densities π are only known up to proportionality (e.g. posterior probability distributions). This would appear to preclude the possibility of obtaining a closed-form kernel mean, but Oates et al. (2015) showed this is not the case. Suppose that we have access to $\eta(\boldsymbol{x}) \propto \pi(\boldsymbol{x})$ such that η is differentiable on $\mathcal{X}$. Then we can proceed as follows: Firstly, we define $\boldsymbol{u} : \mathcal{X} \to \mathbb{R}^n$ componentwise as $u_i(\boldsymbol{x}) := (\partial/\partial x_i) \log \eta(\boldsymbol{x})$. Secondly we specify an RKHS $\mathcal{H}_0$ whose elements are differentiable functions $\phi : \mathcal{X} \to \mathbb{R}$. Thirdly, we construct the set $\mathcal{H}$ whose elements are of the form

$$f(\boldsymbol{x}) = c + \underbrace{\sum_{i=1}^{d} (\partial/\partial x_i)\phi_i(\boldsymbol{x}) + \sum_{i=1}^{d} \phi_i(\boldsymbol{x}) u_i(\boldsymbol{x})}_{\psi(\boldsymbol{x})} \tag{26}$$

where $c \in \mathbb{R}$, $\phi_i \in \mathcal{H}_0$ for $i = 1, \ldots, d$. The function ψ is called a *control functional* from the fact that (under suitable the boundary conditions) we have $\Pi[\psi] = 0$. It can be shown that $\mathcal{H}$ can be endowed with the structure of an RKHS such that the reproducing kernel k associated with $\mathcal{H}$ gives rise to a closed-form (in fact constant) kernel mean. Exact BQ can therefore be performed in $\mathcal{H}$. The RKHS $\mathcal{H}_0$ can typically be selected, for a given integration problem, so that the integrand can be written in the form of Eqn. 26. Full details can be found in Oates et al. (2015).

Excellent performance has been reported for the posterior mean point estimate obtained using control functionals. We note that the interpretation of the posterior variance requires care since the assumption $f \in \mathcal{H}$ is somewhat delicate. In applications below involving control functionals, we only focus on the point estimate.

4.4 Noisy Function Evaluations

To counteract spectral decay in the kernel matrix and improve numerical stability, the kernel matrix $\boldsymbol{K}$ is often replaced in practice by the matrix $\boldsymbol{K} + \lambda \boldsymbol{I}$ for some small $\lambda > 0$. Such regularisation can be interpreted in several ways. If added solely to improve numerical stability, $\lambda \boldsymbol{I}$ is sometimes referred to as *jitter* or a *nugget* term. Of particular interest here is the interpretation that the observed function values f_i are corrupted by noise. Such situations could arise when f is computationally intensive to evaluate and an inexact or noisy surrogate function is used instead for this purpose (Bastos and O'Hagan, 2009). In either case the posterior variance is naturally and appropriately inflated. Below we explore the impact of noisy data in more detail.

We consider a homoscedastic Gaussian noise model in which $\boldsymbol{y} = \boldsymbol{f} + \boldsymbol{e}$ is observed, where $\boldsymbol{e} \sim \mathcal{N}(\boldsymbol{0}, \tau^{-1}\boldsymbol{I})$. In this case, using the conjugacy of Gaussian variables, it is possible to get a closed-form expression for the induced quadrature rule $\hat{\Pi}^{\boldsymbol{e}}_{\mathrm{BQ}}$ and other quantities of interest by replacing $\boldsymbol{f}$ by $\boldsymbol{y}$ and adding a constant term to the diagonal of the kernel matrix of size $\lambda = \tau^{-1}$ (Rasmussen and Williams, 2006). This leads to a probabilistic integrator with

$$\|\hat{\Pi}^{\boldsymbol{e}}_{\mathrm{BQ}} - \Pi\|^2_{\mathrm{op}} \;=\; \|\hat{\Pi}_{\mathrm{BQ}} - \Pi\|^2_{\mathrm{op}} + \tau^{-1}\|\boldsymbol{w}^{\mathrm{BQ}}\|^2_2. \tag{27}$$

Since the term $\|\boldsymbol{w}^{\mathrm{BQ}}\|_2$ can in general decay more slowly (as $n \to \infty$) compared to the MMD term $\|\hat{\Pi}_{\mathrm{BQ}} - \Pi\|_{\mathrm{op}}$, it comes as no surprise that asymptotic convergence rates are much slower in the noisy data regime, as demonstrated by the following:

Proposition 6 (BMC with noisy data)**.** *In the setting of Sec. 3.3.1 and under the homosecdastic Gaussian noise model, we achieve*

$$\|\hat{\Pi}^{\boldsymbol{e}}_{BMC} - \Pi\|_{op} \;=\; O_P(n^{-\alpha/(2\alpha+d)}), \tag{28}$$

while for a Gaussian kernel $k(\boldsymbol{x}, \boldsymbol{y}) = \exp(-\|\boldsymbol{x} - \boldsymbol{y}\|^2_2)$, we have

$$\|\hat{\Pi}^{\boldsymbol{e}}_{BMC} - \Pi\|_{op} \;=\; O_P(n^{-1/2+\epsilon}) \tag{29}$$

where $\epsilon > 0$ can be arbitrarily small.

Clearly the effect of measurement noise is to destroy the asymptotic efficiency of BMC over a simple MC estimator; in fact the BMC estimator becomes *worse* than the MC estimator in these instances. A similar observation is made in Bach (2015).

5 Results

The aims of this section are two-fold; (i) to validate the preceding theoretical analysis and (ii) to demonstrate the applicability and effectiveness of probabilistic integrators in a range of challenging integration problems arising in contemporary machine learning and statistical applications.

5.1 Empirical Validation

Initially we examine whether the theoretical convergence rates obtained above are actually observed in examples with finite n. Then, in a controlled setting, we probe the impact of calibration on the quality of the estimator performance.

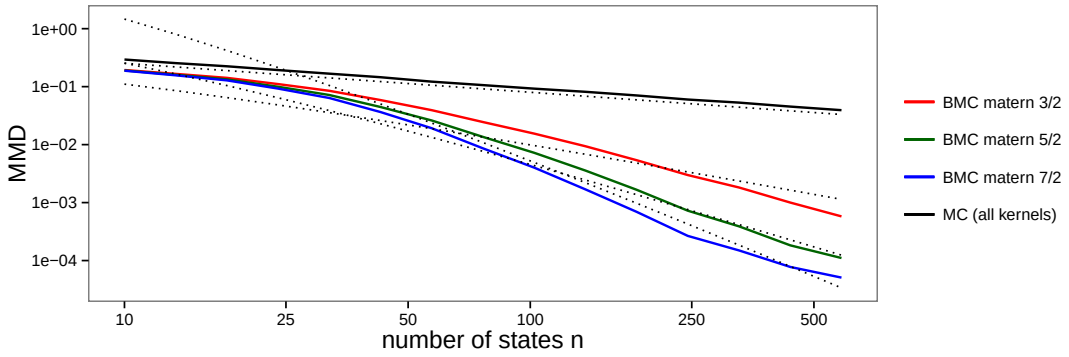


Figure 3: BMC on $\mathcal{X} = [0,1]$ with Matérn kernels of smoothness $\beta \in \{3/2, 5/2, 7/2\}$ and length-scale $\sigma = 0.05$. Each of the lines represents an average of 100 runs of BMC. The BMC methods appear to attain their theoretical convergence rates (dotted black lines).

BMC Beginning with BMC, we examined whether theoretical convergence rates are realised in practice. Our initial investigation focuses on $\mathcal{X} = [0,1]^d$ and RKHS that are characterised by tensor products of Matérn kernels. States $\boldsymbol{x}_i$ were generated independently and uniformly over $\mathcal{X}$. Results in Fig. 3 demonstrate that theoretical convergence rates are indeed observed in practice. Clearly the BMC estimators far outperform MC estimators, with the extent of the performance gain depending on how much smoothness on the integrand can be assumed *a priori*.

In foreseen applications of BMC, kernel parameters may not be available *a priori* and calibration of these parameters will be required. Within the setting considered above, we investigated the performance of the empirical Bayes approach to elicit kernel parameters, as described in Sec. 4.1. Results, reserved for Appendix F, were consistent with the recent analysis of Szabó et al. (2015) that guarantees conservative posterior coverage when the integrand f is sufficiently smooth.

BQMC Our initial investigation of BQMC focuses on $\mathcal{X} = [0,1]^d$ and Π uniform over $\mathcal{X}$. For integration in the space $\mathcal{S}_\alpha(\mathcal{X})$ we employed higher-order digital nets[2] of order β based on Sobol point sequences ($b = 2$) for increasing values of $m \in \mathbb{N}$, so that the total number of states was $n = 2^m$ (see Appendix C).

The Sobolev embedding theorem implies that such higher-order digital nets provide optimal $O(n^{-\alpha+\epsilon})$ rates whenever $\alpha \leq \beta$, since $\mathcal{S}_\beta \subseteq \mathcal{S}_\alpha$. Here we consider tensor products of Matérn kernels and show in Appendix J that closed-form kernel means exist for $\beta = \alpha + 1/2$ whenever $\alpha \in \mathbb{N}$. (Alternatively we could consider Wendland kernels, which provide integer smoothness.) Results in Figure 4 present values of the MMD for increasing numbers of states n and for orders $\alpha \in \{1, 2, 3\}$. The BQMC methods are clearly seen to achieve the theoretical lower rate bounds provided in Theorem 2. Indeed, there is a suggestion that convergence may be faster, which may be explained by the "extra" smoothness of the kernel ($\beta - \alpha = 1/2$). Relative to QMC (not shown), the BQMC method always produces a smaller MMD, in line with Bayesian re-weighting (Lemma 1). At large values of n the convergence rate sometimes appears to change - this is due to numerical instability. An empirical investigation of numerical stability is provided in Appendix G.

[2]Higher-order digital $(t, \beta, 1, \beta m \times m, d)$ nets over $\mathbb{Z}_2$ were generated in `MATLAB R2015a` using code provided by J. Dick at `https://quasirandomideas.wordpress.com/2010/06/17/`[Accessed 24 Nov. 2015].

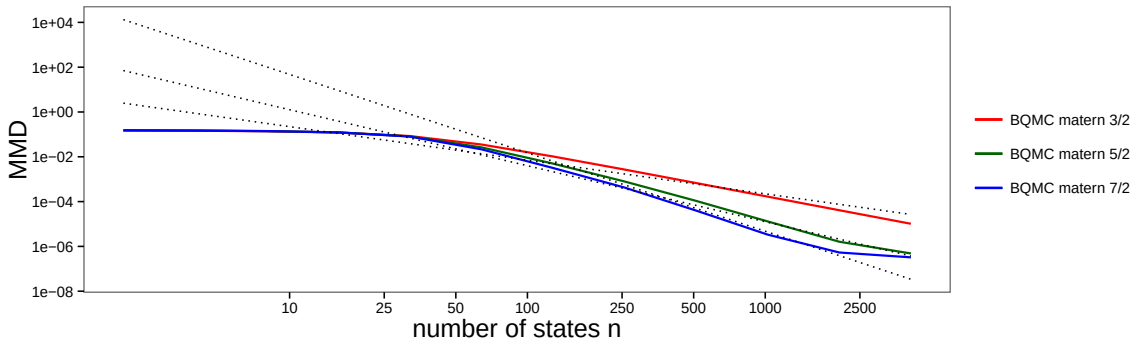


Figure 4: BQMC on $\mathcal{X} = [0, 1]$ with higher-order digital nets of order α when using a Matérn kernel of smoothness $\beta = \alpha + 1/2 \in \{3/2, 5/2, 7/2\}$ and length-scale $\sigma = 0.01$. The BQMC methods are seen to outperform their theoretical convergence rates proved in Thm. 2. In particular, they obtain the optimal convergence rate for $\mathcal{S}_\beta$ (dotted black lines) which is optimal for the space considered whereas the theorem only shows they can achieve the rate of $\mathcal{S}_\alpha$.

In practice it is common to employ sub-optimal QMC point sequences (e.g. Halton or Sobol) for problems that exhibit additional smoothness. In such cases BQMC can provide faster convergence rates than QMC because the latter does not exploit this additional smoothness. Empirical evidence, provided in Fig. 5, supports this claim.

5.2 Applications

The remainder of the paper applies BMCMC and BQMC to three different and challenging problems arising in contemporary machine learning and statistical applications.

5.2.1 Probabilistic Integration on the sphere

Probabilistic integration methods can be defined on arbitrary nonlinear manifolds. The possibility of probabilistic integration in non-Euclidean spaces was suggested as far back as Diaconis (1988) but has only recently been implemented, in the context of computer vision (Brouillat et al., 2009; Marques et al., 2015). Below we formulate and analyse BQMC on the sphere. The method is applied to compute illumination integrals used in the rendering of surfaces.

Spherical Integration In this section we provide the first theoretical study of spherical BQMC and describe a particular class of kernel for which the kernel mean is available in closed-form. We work on the d-sphere

$$\mathbb{S}^d = \{\boldsymbol{x} = (x_1, \ldots, x_{d+1}) \in \mathbb{R}^{d+1} \ : \ \|\boldsymbol{x}\|_2 = \sqrt{x_1^2 + \cdots + x_{d+1}^2} = 1\} \tag{30}$$

in order to estimate integrals of the form

$$\Pi[f] = \int_{\mathbb{S}^d} f(\boldsymbol{x})\pi(\boldsymbol{x})\mathrm{d}\sigma(\boldsymbol{x}), \tag{31}$$

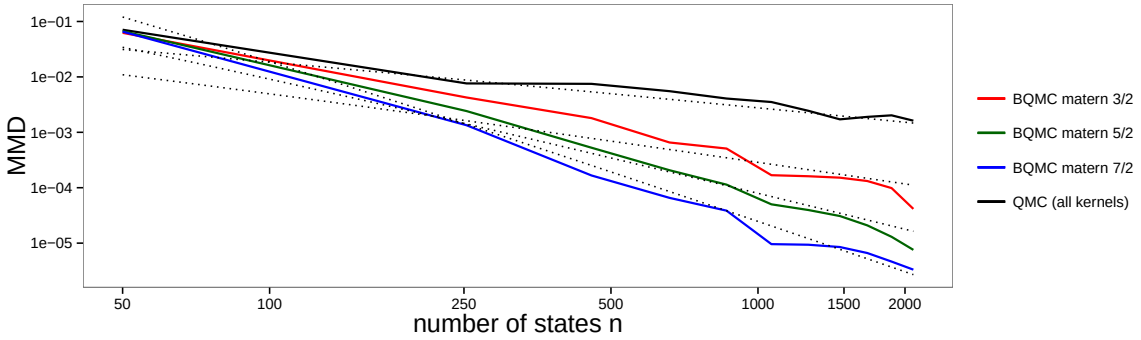


Figure 5: BQMC on $\mathcal{X} = [0, 1]$ with a *Sobol* sequence when using a Matérn kernel of smoothness β and length-scale $\sigma = 0.01$. The use of a Halton sequence was also explored but resulted in similar results (omitted here for clarity). The Sobol sequence is known to provide at least $O(n^{-1+\epsilon})$ rates for integrands with one derivative, but BQMC manages to obtain the optimal rate in Sobolev spaces for each of the Matérn kernel with $\beta \in \{3/2, 5/2, 7/2\}$ (dotted lines). Results show that BQMC converges more quickly than QMC.

where σ is the spherical measure (i.e. uniform over $\mathbb{S}^d$ with $\int_{\mathbb{S}^d} \mathrm{d}\sigma = 1$). For simplicity we focus on the case where the measure Π is uniform over $\mathbb{S}^d$. We specifically focus on the case $d = 2$ that will be used in the computer vision application below.

The function spaces that we consider are Sobolev-like spaces $\mathcal{H}_\alpha(\mathbb{S}^d)$ for $\alpha > d/2$, defined to be the RKHS with reproducing kernel

$$k(\boldsymbol{x}, \boldsymbol{x}') = \sum_{l=0}^{\infty} \lambda_l P_l^{(d)}(\boldsymbol{x} \cdot \boldsymbol{x}') \qquad \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{S}^d. \tag{32}$$

where $\lambda_l \asymp (1 + l)^{-2\alpha}$ (here $a_l \asymp b_l$ is taken to mean that there exists $c_1, c_2 \in \mathbb{R}$ such that $c_1 a_l \leq b_l \leq c_2 a_l$) and $P_l^{(d)}$ are normalised Gegenbauer polynomials (for $d = 2$ these are also known as Legendre polynomials) (Brauchart et al., 2014). A particularly simple expression for the kernel in $d = 2$ and Sobolev-like space $\alpha = 3/2$ can be obtained by taking $\lambda_0 = 4/3$ along with $\lambda_l = -\lambda_0 \times (-1/2)_l/(3/2)_l$ where $(a)_l = a(a+1)\ldots(x+l-1) = \Gamma(a+l)/\Gamma(a)$ is the Pochhammer symbol. Specifically, these choices produce

$$k(\boldsymbol{x}, \boldsymbol{x}') = \frac{8}{3} - \|\boldsymbol{x} - \boldsymbol{x}'\|_2, \qquad \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{S}^2. \tag{33}$$

This kernel is associated with a tractable kernel mean $\mu_\pi(\boldsymbol{x}) = \int_{\mathbb{S}^2} k(\boldsymbol{x}, \boldsymbol{x}')\mathrm{d}\sigma(\boldsymbol{x}') = \frac{4}{3}$ and hence the initial error is also available $\Pi[\mu_\pi] = \int_{\mathbb{S}^2} \mu_\pi(\boldsymbol{x})\mathrm{d}\sigma(\boldsymbol{x}') = 4/3$.

The states $\{\boldsymbol{x}_i\}_{i=1}^n$ could be generated as MC samples. In that case, analogous results to those obtained in Sec. 3.3.1 can be obtained using our proof techniques from Sec. 3.2. Specifically, from Thm. 7 of Brauchart et al. (2014) and Bayesian re-weighting (Lemma 1), classical MC leads to slow convergence $\|\hat{\Pi}_{\mathrm{MC}} - \Pi\|_{\mathrm{op}} = O_P(n^{-1/2})$. The regression bound argument (Lemma 2) together with a functional approximation result in Le Gia et al. (2012, Thm. 3.2), gives a faster rate for BMC of $\|\hat{\Pi}_{\mathrm{BMC}} - \Pi\|_{\mathrm{op}} = O_P(n^{-3/4})$. (For brevity the details are omitted.)

Rather than focus on MC methods, we present stronger results based on spherical QMC point sets. We briefly introduce the concept of a *spherical t-design* (Delsarte et al., 1977) which is define

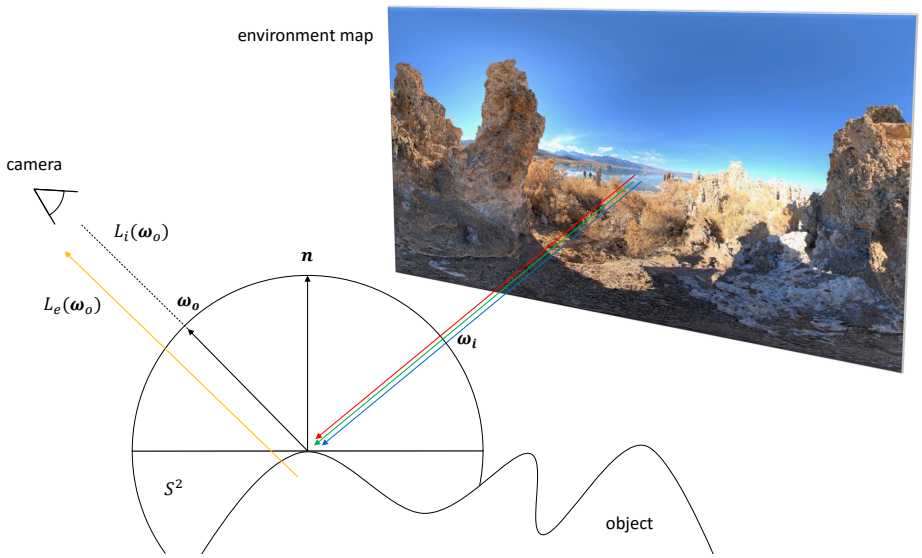


Figure 6: Application to illumination integrals in computer vision. The cartoon features the California lake environment map that was used in our experiments.

as a set $\{\boldsymbol{x}_i\}_{i=1}^n \subset \mathbb{S}^d$ satisfying:

$$\int_{\mathbb{S}^d} f(\boldsymbol{x}) \mathrm{d}\sigma(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{x}_i) \tag{34}$$

for all polynomials $f : \mathbb{S}^d \to \mathbb{R}$ of degree at most t. (i.e. f is the restriction to $\mathbb{S}^d$ of a polynomial in the usual Euclidean sense $\mathbb{R}^{d+1} \to \mathbb{R}$.).

The following properties of spherical t-designs follow from Hesse and Sloan (2005); Bondarenko et al. (2013) and Bayesian re-weighting (Lemma 1):

Theorem 3. *For all $d \geq 2$ there exists C_d such that for all $n \geq C_d t^d$ there exists a spherical t-design on $\mathbb{S}^d$ with n points. Moreover, for $\alpha = 3/2$ and $d = 2$, the use of a spherical t-designs leads to a rate $\|\hat{\Pi}_{BQMC} - \Pi\|_{op} = O(n^{-3/4})$ and $\mathbb{P}[I_D^c|\mathcal{D}] = o(\exp(-Cn^{3/2}))$.*

The rate in Thm. 3 is best-possible in the space $\mathcal{H}_{3/2}(\mathbb{S}^2)$ (Brauchart et al., 2014) and, unlike the result for BMC, is fully deterministic[3]. Although explicit spherical t-designs are not currently known in closed-form, approximately optimal point sets have been computed numerically to high accuracy[4].

Global Illumination integrals We applied spherical integration in the context of global illumination (Pharr and Humphreys, 2004). This problem occurs when one wants to render virtual objects based on a realistic model of a given environment (e.g. a view of a lake; see Fig. 6). In those cases, the models are based on four main factors: a geometric model for the object being

[3]Empirical evidence in Marques et al. (2015) suggests that BQMC attains faster rates than BMC in RKHS that are smoother than $\mathcal{H}_{3/2}(\mathbb{S}^2)$.

[4]Our experiments were based on such point sets provided by R. Womersley on his website `http://web.maths.unsw.edu.au/~rsw/Sphere/EffSphDes/sf.html`[Accessed 24 Nov. 2015].

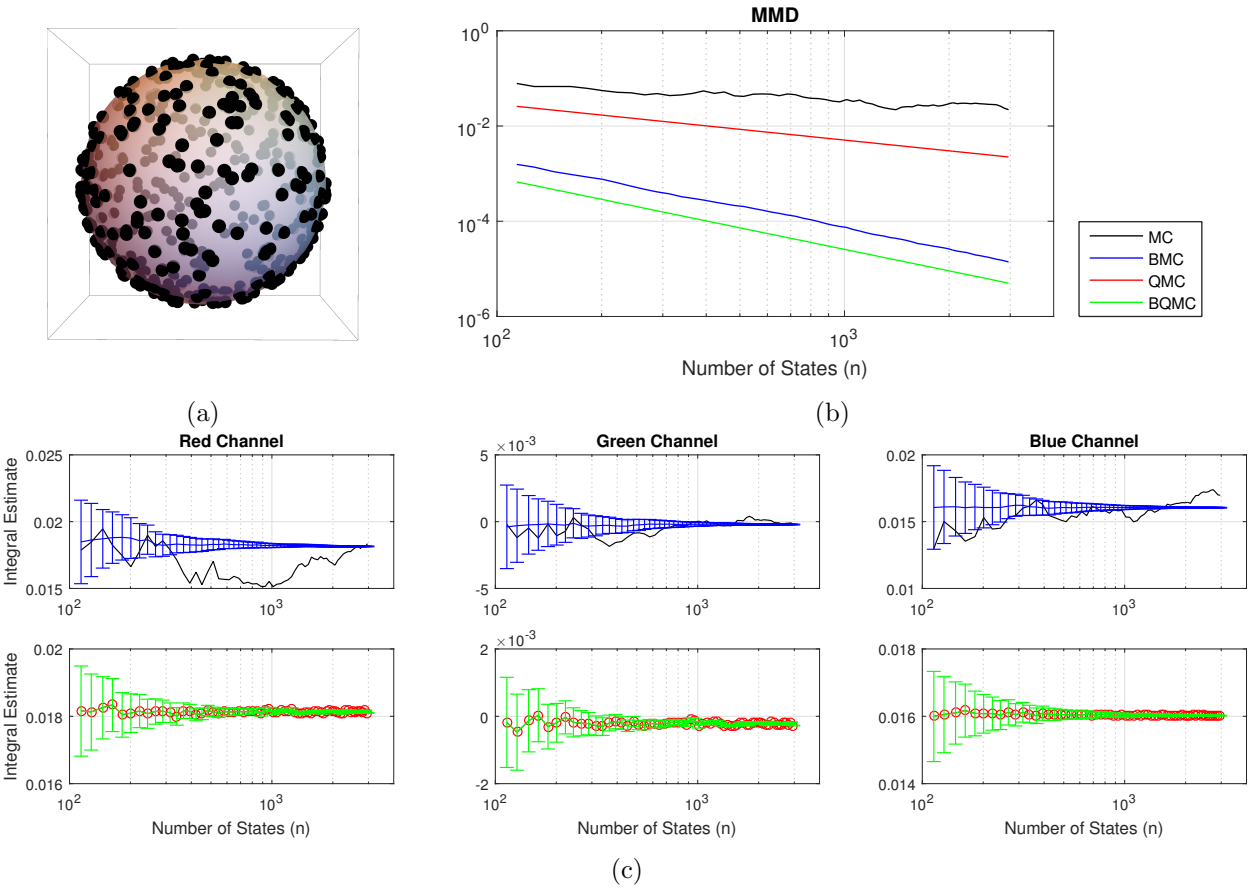


Figure 7: Application to illumination integrals in computer vision. (a) A spherical t-design over $\mathbb{S}^2$. (b) The MMD, or worst-case-error, for Monte Carlo (MC), Bayesian MC (BMC), Quasi MC (QMC) and Bayesian QMC (BQMC). (c) Probabilistic integration over the sphere was employed to estimate the RGB colour intensities for the California lake environment. [Error bars for BMC and BQMC represent two posterior standard deviations (i.e. 95% credible intervals). Red circles are used to highlight QMC estimates, which are closely aligned with the BQMC estimates.]

rendered, a model for the reflectivity of the surface of the object, the angle at which we observe the object and a description of the light sources (provided by an *environment map*). The light emitted from the environment will interact with the object in multiple ways and our goal is to estimate the total amount of light arriving at the camera. This can be formulated as an illumination integral[5]

$$L_o(\boldsymbol{\omega}_o) = L_e(\boldsymbol{\omega}_o) + \int_{\mathbb{S}^2} L_i(\boldsymbol{\omega}_i)\rho(\boldsymbol{\omega}_i, \boldsymbol{\omega}_o)[\boldsymbol{\omega}_i \cdot \boldsymbol{n}]_+ \mathrm{d}\sigma(\boldsymbol{\omega}_i), \tag{35}$$

expressed with respect to the spherical measure σ. Here $L_o(\boldsymbol{\omega}_o)$ is the *outgoing radiance*, i.e. the outgoing light in the direction $\boldsymbol{\omega}_o$. $L_e(\boldsymbol{\omega}_o)$ represents the amount of light emitted by the object itself (which we will assume to be known) and $L_i(\boldsymbol{\omega}_i)$ is the light hitting the object from direction

[5]It is noted by Marques et al. (2015) that slightly improved empirical performance can be obtained by replacing the $[\boldsymbol{\omega}_i \cdot \boldsymbol{n}]_+$ term with the smoother $\boldsymbol{\omega}_i \cdot \boldsymbol{n}$ term and restricting the domain of integration to the hemisphere $\boldsymbol{\omega}_i \cdot \boldsymbol{n} \geq 0$. For simplicity we present the problem as an integral over $\mathbb{S}^2$.

$\boldsymbol{\omega}_i$. The term $\rho(\boldsymbol{\omega}_i, \boldsymbol{\omega}_o)$ is the *bidirectional reflectance distribution* function (BRDF), which models the fraction of light entering the pixel through direction $\boldsymbol{\omega}_i$ and being reflected towards direction $\boldsymbol{\omega}_o$. Here $\boldsymbol{n}$ is a unit vector normal to the surface of the object. Our investigation is motivated by strong empirical results for BQMC in this context obtained by Marques et al. (2015)[6].

In order to assess the performance of BQMC we consider a typical illumination integration problem based on the California lake environment map shown in Fig. 6[7]. The goal here is to compute intensities for each of the three RGB colour channels corresponding to observing a virtual object from a fixed direction $\boldsymbol{\omega}_o$. We consider the case of an object directly facing the camera ($\boldsymbol{w}_o = \boldsymbol{n}$). For the BRDF we took $\rho(\boldsymbol{\omega}_i, \boldsymbol{\omega}_o) = (2\pi)^{-1} \exp(\boldsymbol{\omega}_i \cdot \boldsymbol{\omega}_o - 1)$. The integral in Eqn. 35 was viewed here as an integral with respect to a uniform measure Π and the integrand $f(\boldsymbol{\omega}_i) = L_i(\boldsymbol{\omega}_i)\rho(\boldsymbol{\omega}_i, \boldsymbol{\omega}_o)[\boldsymbol{\omega}_i \cdot \boldsymbol{\omega}_o]_+$ was modeled using the kernel in Eqn. 33. In contrast, Marques et al. (2015) viewed Eqn. 35 as an integral with respect to $\pi(\boldsymbol{\omega}_i) \propto \rho(\boldsymbol{\omega}_i, \boldsymbol{\omega}_o)$ and coupled this with a Gaussian kernel restricted to the hemisphere. The approach that we propose has two advantages; (i) it provides a closed-form expression for the kernel mean, (ii) a rougher kernel may be more appropriate in the context of illumination integrals, as pointed out by Brouillat et al. (2009).

Results in Fig. 7 demonstrate a reduction in MMD for the BMC and BQMC methodologies over their MC and QMC counterparts. Moreover we observe similar rates of convergence for BMC and BQMC, in line with the theoretical results presented above. Translating this performance into the RGB-space, we see that BMC and BQMC provide an appropriate quantification of uncertainty in the value of the integral at all values of n that were considered. For this particular test function the BQMC point estimate was almost identical to the QMC estimate at all values of n. Empirical results reported by Marques et al. (2015), based on Gaussian kernels, showed a RMSE rate of $O(n^{-0.72})$, which is similar to the theoretical $O(n^{-3/4})$ rate that we provide here. A more detailed comparison of the methods is reserved for future research.

5.2.2 Integration with Intractable Densities

Here we consider Bayesian parameter estimation for a non-linear differential equation model. This problem involves an intractable probability density, which means that the kernel mean is likely to be intractable for most common kernels. We will therefore follow the methodology proposed in Sec. 4.3 and make use of the control functional kernel for which a tractable kernel mean can be obtained. The particular probabilistic integration method that we will consider is BMCMC, where the underlying Markov chain used to obtain states is provided by a Riemann manifold Hamiltonian Monte Carlo (RMHMC) method (Girolami and Calderhead, 2011).

We consider nonlinear dynamical systems of the form

$$\frac{d\boldsymbol{u}}{ds} = \boldsymbol{f}(\boldsymbol{u}, s; \boldsymbol{\theta}), \quad \boldsymbol{u}(0) = \boldsymbol{u}_0. \tag{36}$$

where the state variables are assumed to be observed under noise at discrete times $s_1 < s_2 < \cdots < s_n$, denoting the observations by $\boldsymbol{y}(s_j)$. We consider a Gaussian observation process with likelihood

$$p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{u}_0, \sigma) = \prod_{j=1}^{n} \mathcal{N}(\boldsymbol{y}(s_j)|\boldsymbol{u}(s_j; \boldsymbol{\theta}, \boldsymbol{u}_0), \tau^{-1}\boldsymbol{I}) \tag{37}$$

[6]The authors call their method BMC, but states arose from a deterministic (spiral point) algorithm.

[7]This environment map is freely available at: `http://www.hdrlabs.com/sibl/archive.html`.

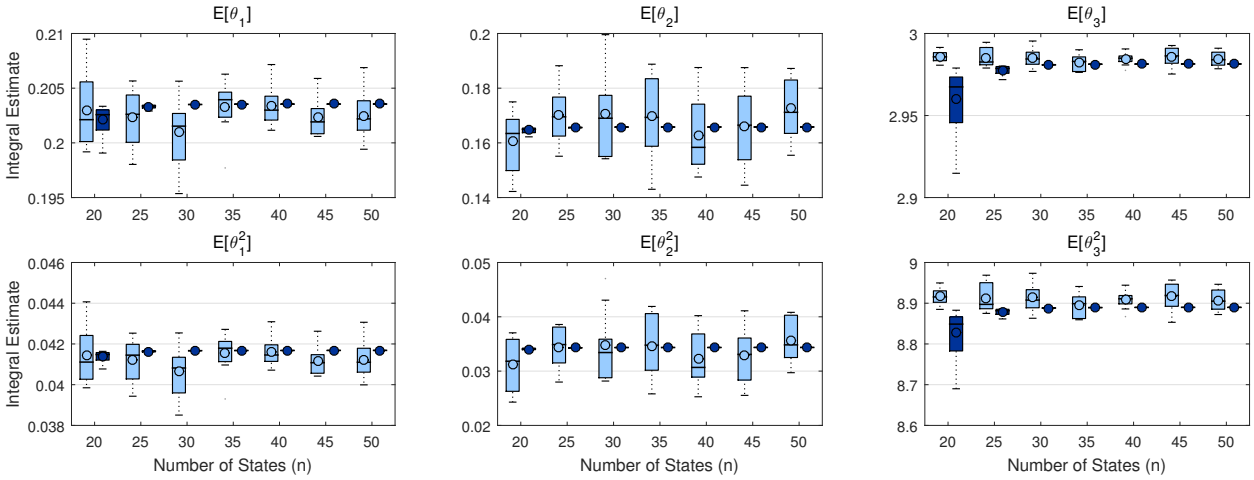


Figure 8: Integration with intractable densities; results. Light blue = standard Monte Carlo. Dark blue = Bayesian MCMC. The boxplots show the distribution of the posterior mean estimate under subsampling from the empirical distribution produced by MCMC.

where $\boldsymbol{u}(s_j; \boldsymbol{\theta}, \boldsymbol{u}_0)$ denotes the solution of the system in Eqn. 36. Each evaluation of the likelihood requires the numerical solution of a system of differential equations, so that the cost of the probabilistic integration scheme was far out-weighed by the computational cost of obtaining MCMC samples. An important challenge is therefore to obtain accurate estimates in the low n regime.

The particular non-linear ODE model that we consider in our numerical experiments is the Fitzhugh–Nagumo model

$$\dot{U}_1 \;=\; \theta_3 \left(U_1 - \frac{U_1^3}{3} + U_2 \right), \qquad \dot{U}_2 \;=\; -\left(\frac{U_1 - \theta_1 + \theta_2 U_2}{\theta_3} \right) \tag{38}$$

with data-generating parameters and data identical to those used in Girolami and Calderhead (2011). The parameter prior $p(\boldsymbol{\theta})$ and RMHMC sampling scheme were identical to the implementation provided by Girolami and Calderhead (2011). We consider here the problem of estimating the first and second posterior moments of the ODE parameters $\boldsymbol{\theta}$. The control functional kernel described in Sec. 4.3 was used with kernel parameters selected manually, informed by performance at cross-validation. Results in Fig. 8 show that the BMCMC estimates are more accurate compared to the standard MCMC estimates when $n \geq 20$. These results suggest that BQ is a useful addition to the computational Bayesian's toolbox.

5.2.3 High-Dimensional Probabilistic Integration

Our aim in this final section is to demonstrate how recent theoretical advances in high-dimensional QMC theory enable tractable high-dimensional probabilistic integration. We will concentrate on BQMC, but the methodology proposed below could be applied to other probabilistic integrators. The presentation culminates in an application to semi-parametric regression with random effects, where marginalisation of random effects requires solution of a challenging $d = 50$ dimensional integral. We emphasise that while $d = 50$ may not seem very large, it is much higher than available in previous applications of probabilistic integration; the largest values of d we found in the literature is $d = 19$ in Hennig et al. (2015) and $d = 20$ in Osborne et al. (2012).

Weighted Spaces The formulation of high (and infinite) -dimensional QMC results requires a construction known as a *weighted* Hilbert space. These spaces, defined below, are motivated by the observation that many integrands encountered in applications seem to vary more in lower dimensional projections compared to higher dimensional projections. Our presentation below follows Sec. 2.5.4 of Dick and Pillichshammer (2010).

As usual with QMC, we work in $\mathcal{X} = [0,1]^d$, σ is the Lebesgue measure and with Π uniform over $\mathcal{X}$. Let $\mathcal{I} = \{1, 2, \ldots, d\}$. For each subset $u \subseteq \mathcal{I}$, define a weight $\gamma_u \in (0, \infty)$ and denote the collection of all weights by $\boldsymbol{\gamma} = \{\gamma_u\}_{u \subseteq \mathcal{I}}$. Consider the space $\mathcal{H}_{\boldsymbol{\gamma}}$ of functions of the form

$$f(\boldsymbol{x}) = \sum_{u \subseteq \mathcal{I}} f_u(\boldsymbol{x}_u) \tag{39}$$

where f_u belongs to an RKHS $\mathcal{H}_u$ with reproducing kernel k_u and $\boldsymbol{x}_u$ denotes the components of $\boldsymbol{x}$ that are indexed by u. We point out that this construction is not restrictive, since any function can be written in this form by considering only $u = \mathcal{I}$. We turn $\mathcal{H}_{\boldsymbol{\gamma}}$ into a Hilbert space by defining an inner product

$$\langle f, g \rangle_{\boldsymbol{\gamma}} := \sum_{u \subseteq \mathcal{I}} \gamma_u^{-1} \langle f_u, g_u \rangle_u \tag{40}$$

where $\boldsymbol{\gamma} = \{\gamma_u : u \subseteq \mathcal{I}\}$. Constructed in this way, $\mathcal{H}_{\boldsymbol{\gamma}}$ is an RKHS with reproducing kernel

$$k_{\boldsymbol{\gamma}}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{u \subseteq \mathcal{I}} \gamma_u k_u(\boldsymbol{x}, \boldsymbol{x}'). \tag{41}$$

Intuitively, the weights γ_u can be taken to be small whenever the function f does not depend heavily on the $|u|$-way interaction of the states $\boldsymbol{x}_u$. Thus most of the γ_u will be small for a function f that is effectively low-dimensional. A measure of the dimensionality of the function is given by $\sum_{u \subseteq \mathcal{I}} \gamma_u$.

The (canonical) *weighted* Sobolev space $\mathcal{S}_{\alpha,\boldsymbol{\gamma}}$ is defined by taking each of the component spaces $\mathcal{H}_u$ to be Sobolev spaces of dominating mixed smoothness $\mathcal{S}_\alpha$. i.e. the space $\mathcal{H}_u$ is norm-equivalent to a tensor product of $|u|$ one-dimensional Sobolev spaces, each with smoothness parameter α. Constructed in this way, $\mathcal{S}_{\alpha,\boldsymbol{\gamma}}$ is an RKHS with kernel

$$k_{\alpha,\boldsymbol{\gamma}}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{u \subseteq \mathcal{I}} \gamma_u \prod_{i \in u} \left(\sum_{k=1}^{\alpha} \frac{B_k(x_i) B_k(x_i')}{(k!)^2} - (-1)^\alpha \frac{B_{2\alpha}(|x_i - x_i'|)}{(2\alpha)!} \right) \tag{42}$$

where the B_k are Bernoulli polynomials. For example, taking $\alpha = 1$ we have the kernel

$$k_{1,\boldsymbol{\gamma}}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{u \subseteq \mathcal{I}} \gamma_u \prod_{i \in u} \left(\frac{x_i^2}{2} + \frac{(x_i')^2}{2} - \frac{x_i}{2} - \frac{x_i'}{2} - \frac{|x_i - x_i'|}{2} + \frac{1}{3} \right) \tag{43}$$

and tractable kernel mean $\mu_\pi(\boldsymbol{x}) = \int_{\mathcal{X}} k_{1,\boldsymbol{\gamma}}(\boldsymbol{x}, \boldsymbol{x}') d\boldsymbol{x}' = \gamma_\emptyset$. In finite dimensions $d < \infty$, we can construct a higher-order digital net that attains optimal QMC rates for weighted Sobolev spaces:

Theorem 4. *Let $\mathcal{H}$ be an RKHS that is norm-equivalent to $\mathcal{S}_{\alpha,\boldsymbol{\gamma}}$. Then BQMC based on a digital $(t, \alpha, 1, \alpha m \times m, d)$-net over $\mathbb{Z}_b$ attains the optimal rate*

$$\|\hat{\Pi}_{BQMC} - \Pi\|_{op} = O(n^{-\alpha+\epsilon}) \tag{44}$$

for any $\epsilon > 0$, where $n = b^m$. Hence $\mathbb{P}[I_D^c | \mathcal{D}] = o(\exp(-Cn^{2\alpha-\epsilon}))$.

Remark 6. *The QMC rules in Theorem 4 do not explicitly take into account the values of the weights $\boldsymbol{\gamma}$. An algorithm that tailors QMC points to specific weights $\boldsymbol{\gamma}$ is known as the "component by component" (CBC) algorithm; further details can be found in (Kuo, 2003). In principle the CBC algorithm can lead to improved rate constants in high dimensions, because effort is not wasted in directions where f varies little, but the computational overheads are also greater. We did not consider CBC algorithms for BQMC in this paper.*

Remark 7. *The weighted Hilbert space framework allows us to bound the MMD* independently of dimension *providing that*

$$\sum_{u \in \mathcal{I}} \gamma_u < \infty \tag{45}$$

(Sloan and Woźniakowski, 1998). This justifies the "high-dimensional" terminology; the posterior variance can bounded independently of dimension for these RKHSs. Analogous results were provided by Fasshauer et al. (2012) for the Gaussian kernel. Further details are provided in Sec. 4.1 of (Dick et al., 2013).

Semi-Parametric Random Effects Regression For illustration we observe that weighted Sobolev spaces provide an RKHS that appropriately models features of integrals that appear in semi-parametric random effects regression. Below we consider a problem posed and studied by Kuo et al. (2008). The context is inference for the parameters $\boldsymbol{\beta}$ of a Poisson semi-parametric random effects regression model

$$\begin{aligned} Y_j|\lambda_j &\sim \text{Po}(\lambda_j) \\ \log(\lambda_j) &= \beta_0 + \beta_1 z_{1,j} + \beta_2 z_{2,j} + u_1\phi_1(z_{2,j}) + \cdots + u_d\phi_d(z_{2,j}) \\ u_j &\sim N(0, \tau^{-1}) \text{ independently.} \end{aligned} \tag{46}$$

Here $z_{1,j} \in \{0,1\}$, $z_{2,j} \in (0,1)$ and $\phi_j(z) = [z - \kappa_j]_+$ where $\kappa_j \in (0,1)$ are pre-determined knots (wlog $\kappa_j < \kappa_{j+1}$). We took $d = 50$ equally spaced knots in $[\min \boldsymbol{z}_2, \max \boldsymbol{z}_2]$. Inference for $\boldsymbol{\beta}$ requires multiple evaluations of the observed data likelihood

$$p(\boldsymbol{y}|\boldsymbol{\beta}) = \int_{\mathbb{R}^d} p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}) p(\boldsymbol{u}) \mathrm{d}\boldsymbol{u} \tag{47}$$

and therefore is a natural candidate for probabilistic integration methods, in order to propagate the cumulative uncertainty of estimating multiple numerical integrals into the posterior distribution $p(\boldsymbol{\beta}|\boldsymbol{y})$.

In order to transform this integration problem to the unit cube we perform the change of variables $x_j = \Phi^{-1}(u_j)$ so that we wish to evaluate

$$p(\boldsymbol{y}|\boldsymbol{\beta}) = \int_{[0,1]^d} p(\boldsymbol{y}|\boldsymbol{\beta}, \Phi^{-1}(\boldsymbol{x})) \mathrm{d}\boldsymbol{x}. \tag{48}$$

Here $\Phi^{-1}(\boldsymbol{x})$ denotes the standard Gaussian inverse CDF applied to each component of the vector $\boldsymbol{x}$.

Probabilistic integration proceeds under the hypothesis that the integrand of interest $f(\boldsymbol{x}) = p(\boldsymbol{y}|\boldsymbol{\beta}, \Phi^{-1}(\boldsymbol{x}))$ belongs to (or at least can be well approximated by functions in) $\mathcal{S}_{\alpha,\boldsymbol{\gamma}}$ for some smoothness parameter α and some weights $\boldsymbol{\gamma}$. Intuitively, the integrand $f(\boldsymbol{x})$ is such that an

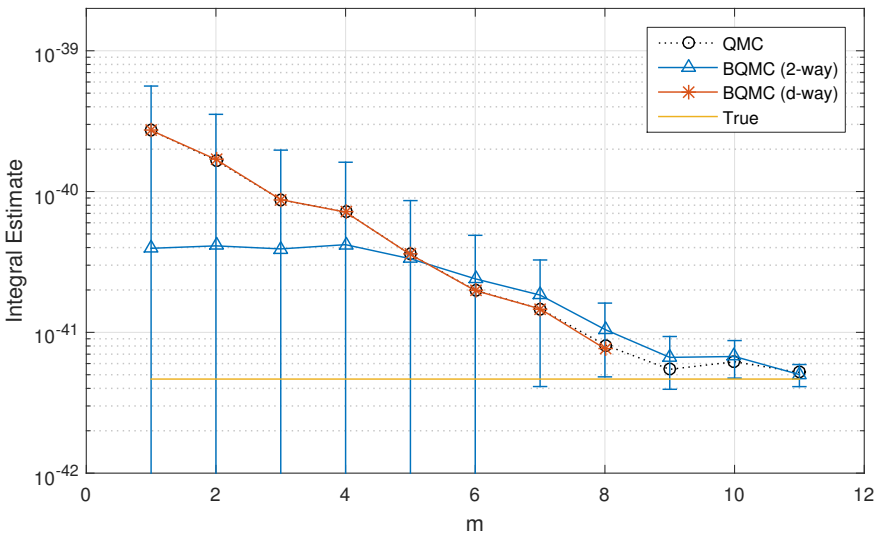


Figure 9: Application to semi-parametric random effects regression in $d = 50$ dimensions, based on $n = 2^m$ samples from a higher-order digital net. [Here error bars show two posterior standard deviations from the posterior mean. To improve visibility results are shown on the log-scale; error bars are symmetric on the linear scale. A brute-force QMC estimate was used to approximate the true value of the integral.]

increase in the value of x_j at the knot κ_j can be compensated for by a decrease in the value of x_{j+1} at a neighbouring knot κ_{j+1}, but not by changing values of $\boldsymbol{x}$ at more remote knots. Therefore we expect $f(\boldsymbol{x})$ to exhibit strong individual and pairwise dependence on the x_j, but expect higher-order dependency to be much weaker. This motivates the weighted space assumption. We chose weights $\boldsymbol{\gamma}$, in the terminology of Kuo et al. (2008), to be the "order two" weights $\gamma_u = 1$ for $|u| \leq d_{\max}$, $d_{\max} = 2$, $\gamma_u = 0$ otherwise, which corresponds to an assumption of low order interaction terms (though f can still depend on all of its arguments). This choice of weights was shown by Kuo et al. (2008) to achieve the same performance as the popular "product" weights $\gamma_u = 2^{-|u|}$, so we therefore used the order two weights to reduce the computational burden. We briefly mention recent work by Sinescu et al. (2012) that provides more detailed theoretical analysis for the choice of weights $\boldsymbol{\gamma}$.

In terms of frequentist point accuracy, results in Fig. 9 (with $\alpha = 1$) demonstrate that the BQMC posterior distribution provides accuracy comparable to the standard QMC estimate, with BQMC more accurate than QMC at smaller sample sizes ($n \leq 2^5$). To understand the effect of the weighted space construction here, we compared against BQMC with d-way interactions ($u \in \{\emptyset, \mathcal{I}\}$). We found that the d-way BQMC closely resembled standard QMC and thus integral estimates based on 2-way interactions were more accurate at smaller sample sizes, although in general the performance of all methods was comparable to standard QMC on this problem. In terms of uncertainty quantification, the 90% posterior credible regions more-or-less cover the truth for this problem, suggesting that the uncertainty estimates are sensible.

Although we did not consider it here, Kuo et al. (2008) demonstrated how centring and scaling transformations of the integrand $f(\boldsymbol{x})$ can further boost empirical performance in this example.

6 Conclusion

The increasing sophistication of complex computational models, of which numerical integration is one component, demands an improved understanding of how numerical error accumulates and propagates through sequential computation. In (now common) settings where integrands are computationally intensive, or very many numerical integrals are required, then an attractive and statistically principled solution is to model the numerical error explicitly. This paper lays firm theoretical foundations for the probabilistic approach to integration.

The general methodology that we describe above provides a unified framework in which existing MC and QMC methods can be adapted to produce high-performance probabilistic integrators. It was shown that these probabilistic integrators can achieve super-exponential rates for posterior contraction and several empirical experiments demonstrated correct posterior coverage. These rates, obtained in Sobolev-type spaces, are important, fundamental and novel contributions. However, there remain many important open questions for probabilistic integration that we did not address here:

Theory

- Our results concerned the asymptotic frequentist coverage of posterior credible intervals. An important area of future research, that is perhaps more important in foreseen applications, will be to obtain corresponding non-asymptotic results for frequentist coverage. In addition, the robustness of the posterior coverage to mis-specified RKHS deserves to be explored in greater detail.

- Our theoretical analysis focused on BMC rather than BMCMC, justified by the fact that our results were "only" asymptotic scaling relationships and therefore will be identical for both methods. For the future non-asymptotic analysis it will be important to extend these results to the case of correlated samples, such as arise in MCMC and other sampling schemes, such as sequential MC.

- From a perspective of convenience, we restricted attention to simple domains of integration, such as the (hyper)cube and the (hyper)sphere. The extension of these results to general integration domains, and to more general stochastic and path integrals, should be developed.

Methodology

- The requirement of a tractable kernel mean must be overcome to enable BQ for arbitrary RKHS and arbitrary probability distributions. While we have sketched details for how this can be achieved (Appendix I), it remains an open problem to reconcile such an approach with a formal probabilistic interpretation.

- On the other hand, it can be argued that the RKHS framework is rather restrictive. One important property that is not easily encoded in an RKHS is non-negativity of the integrand (e.g. as encountered in Sec. 5.2.3). Recent work addresses this issue using approximations (Osborne et al., 2012; Gunter et al., 2014), but there does not yet exist an exact solution.

- Advances in computational approaches for kernel methods should be investigated to mitigate the headline $O(n^3)$ computational complexity of BQ methods.

Application

- This paper did not present results in which BQ is employed within a larger computational pipeline, even though this is the primary application area for probabilistic integration. Our focus was instead the theoretical foundations of BQ. With foundations now established, it will be important to explore the practical challenges and of probabilistic integration within a larger computational framework.

- An important and untouched feature of the probabilistic formulation is the possibility to perform transfer learning when several related integrals require evaluation. This is a direction that we will explore in future work.

Acknowledgements

The authors acknowledge helpful conversations with Alessandro Barp, Jon Cockayne, Josef Dick, David Duvenaud, Philipp Hennig, Aretha Teckentrup and Houying Zhu. The authors are grateful to Ricardo Marques for providing code used in producing these results. FXB was supported by the EPSRC grant [EP/L016710/1]. MG was supported by the EPSRC grant [EP/J016934/1, EP/K034154/1], an EPSRC Established Career Fellowship, the EU grant [EU/259348] and a Royal Society Wolfson Research Merit Award.

References

Y. Altun and A. Smola. Unifying divergence minimization and statistical inference via convex duality. In *Learning Theory*, pages 139–153. Springer, 2006.

S. Ambikasaran and E. Darve. The inverse fast multipole method. *arXiv:1407.1572*, 2014.

F. Bach. Sharp analysis of low-rank kernel matrix approximations. In *International Conference on Learning Theory*, pages 185–209, 2013.

F. Bach. On the equivalence between quadrature rules and random features. *arXiv:1502.06800*, 2015. doi: HAL-01118276-v2.

N. S. Bakhvalov. On the optimality of linear methods for operator approximation in convex classes of functions. *USSR Computational Mathematics and Mathematical Physics*, 11(4):244–249, 1971.

L. S. Bastos and A. O'Hagan. Diagnostics for Gaussian process emulators. *Technometrics*, 51(4): 425–438, 2009.

A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer Science+Business Media, New York, 2004.

A. Bondarenko, D. Radchenko, and M. Viazovska. Optimal asymptotic bounds for spherical designs. *Annals of Mathematics*, 178(2):443–452, 2013.

J. Brauchart, E. Saff, I. H. Sloan, and R. Womersley. QMC designs: Optimal order quasi Monte Carlo integration schemes on the sphere. *Mathematics of Computation*, 83:2821–2851, 2014.

F.-X. Briol, C. J. Oates, M. Girolami, and M. A. Osborne. Frank-Wolfe Bayesian Quadrature: Probabilistic integration with theoretical guarantees. In *Advances In Neural Information Processing Systems*, 2015.

J. Brouillat, C. Bouville, B. Loos, C. Hansen, and K. Bouatouch. A Bayesian Monte Carlo approach to global illumination. *Computer Graphics Forum*, 28(8):2315–2329, 2009.

Y. Chen, M. Welling, and A. Smola. Super-samples from kernel herding. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2010.

P. Conrad, M. Girolami, S. Särkka, A. Stuart, and K. Zygalakis. Probability measures for numerical solutions of differential equations. *arXiv:1506.04592*, 2015.

B. Dai, B. Xie, N. He, Y. Liang, A. Raj, M. Balcan, and L. Song. Scalable kernel methods via doubly stochastic gradients. In *Advances in Neural Information Processing Systems*, pages 3041–3049, 2014.

M. Dashti and A. Stuart. The Bayesian approach to inverse problems. In *Handbook of Uncertainty Quantification*. 2016.

P. Delsarte, J. M. Goethals, and J. J. Seidel. Spherical codes and designs. *Geometriae Dedicata*, 6 (3):363–388, 1977.

P. Diaconis. Bayesian numerical analysis. *Statistical Decision Theory and Related Topics IV*, pages 163–175, 1988.

J. Dick. Higher order scrambled digital nets achieve the optimal rate of the root mean square error for smooth integrands. *Annals of Statistics*, 39(3):1372–1398, 2011.

J. Dick and F. Pillichshammer. *Digital Nets and Sequences - Discrepancy Theory and Quasi-Monte Carlo Integration*. Cambridge University Press, 2010.

J. Dick, G. Larcher, F. Pillichshammer, and H. Woźniakowski. Exponential convergence and tractability of multivariate integration for Korobov spaces. *Mathematics of Computation*, 80 (274):905–905, 2011.

J. Dick, F. Y. Kuo, and I. H. Sloan. High-dimensional integration: The quasi-Monte Carlo way. *Acta Numerica*, 22:133–288, 2013.

D. Duvenaud, J. Lloyd, R. Grosse, J Tenenbaum, and Z. Ghahramani. Structure discovery in nonparametric regression through compositional kernel search. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1166–1174, 2013.

A. El Alaoui and M. W. Mahoney. Fast randomized kernel methods with statistical guarantees. In *Advances in Neural Information Processing Systems*, 2015.

G. Fasshauer, F. Hickernell, and H. Woźniakowski. On dimension-independent rates of convergence for function approximation with Gaussian kernels. *SIAM Journal on Numerical Analysis*, 50(1): 247–271, 2012.

M. Gerber and N. Chopin. Sequential quasi-Monte Carlo. *Journal of the Royal Statistical Society B: Statistical Methodology*, 77(3):509–579, 2015.

M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(2):123–214, 2011.

R. B. Gramacy and D. W. Apley. Local Gaussian process approximation for large computer experiments. *Journal of Computational and Graphical Statistics*, 24(2):561–578, 2015.

T. Gunter, R. Garnett, M. Osborne, P. Hennig, and S. Roberts. Sampling for inference in probabilistic models with fast Bayesian quadrature. In *Advances in Neural Information Processing Systems*, pages 2789–2797, 2014.

P. Hennig. Probabilistic interpretation of linear solvers. *SIAM Journal on Optimization*, 25(1): 234–260, 2015.

P. Hennig and M. Kiefel. Quasi-Newton methods: A new direction. *Journal of Machine Learning Research*, 14:843–865, 2013.

P. Hennig, M. A. Osborne, and M. Girolami. Probabilistic numerics and uncertainty in computations. *Journal of the Royal Society A*, 471(2179), 2015.

K. Hesse and I. A. Sloan. Worst-case errors in a Sobolev space setting for cubature over the sphere S2. *Bulletin of the Australian Mathematical Society*, 71(1):81–105, 2005.

F. J. Hickernell, C. Lemieux, and A. B. Owen. Control variates for quasi-Monte Carlo. *Statistical Science*, 20(1):1–31, 2005.

T. E. Hull and J. R. Swenson. Tests of probabilistic models for propagation of roundoff errors. *Communications of the ACM*, 9(2):108–113, 1966.

F. Huszar and D. Duvenaud. Optimally-weighted herding is Bayesian quadrature. In *Uncertainty in Artificial Intelligence*, pages 377–385, 2012.

M. Katzfuss. A multi-resolution approximation for massive spatial datasets. *arXiv:1507.04789*, 2015.

D. Krieg and E. Novak. A universal algorithm for multivariate integration. *arXiv:1507.06853*, 2015.

F. Y. Kuo. Component-by-component constructions achieve the optimal rate of convergence for multivariate integration in weighted Korobov and Sobolev spaces. *Journal of Complexity*, 19(3): 301–320, 2003.

F. Y. Kuo and H. Woźniakowski. Gauss-Hermite quadratures for functions from Hilbert spaces with Gaussian reproducing kernels. *BIT Numerical Mathematics*, 52(2):425–436, 2012.

F. Y. Kuo, W. T. M. Dunsmuir, I. H. Sloan, M. P. Wand, and R. S. Womersley. Quasi-Monte Carlo for highly structured generalised response models. *Methodology and Computing in Applied Probability*, 10(2):239–275, 2008.

S. Lacoste-Julien, F. Lindsten, and F. Bach. Sequential Kernel Herding : Frank-Wolfe Optimization for Particle Filtering. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, pages 544–552, 2015.

B. Lakshminarayanan, D.M. Roy, and Y.W. Teh. Mondrian forests for large-scale regression when uncertainty matters. *arXiv:1506.03805*, 2015.

M. Lazaro-Gredilla, J. Quinonero-Candela, C. E. Rasmussen, and A. R. Figueiras-Vidal. Sparse spectrum Gaussian process regression. *Journal of Machine Learning Research*, 11:1865–1881, 2010.

Q. T. Le Gia, I. H. Sloan, and H. Wendland. Multiscale approximation for functions in arbitrary Sobolev spaces by scaled radial basis functions on the unit sphere. *Applied and Computational Harmonic Analysis*, 32:401–412, 2012.

M. N. Lukic and J. H. Beder. Stochastic processes with sample paths in reproducing kernel Hilbert spaces. *Transactions of the American Mathematical Society*, 353(10):3945–3969, 2001.

M. Mahsereci and P. Hennig. Probabilistic line searches for stochastic optimization. In *Advances In Neural Information Processing Systems*, 2015.

R. Marques, C. Bouville, M. Ribardiere, L. P. Santos, and K. Bouatouch. A spherical Gaussian framework for Bayesian Monte Carlo rendering of glossy surfaces. *IEEE Transactions on Visualization and Computer Graphics*, 19(10):1619–1632, 2013.

R. Marques, C. Bouville, L.P. Santos, and K. Bouatouch. Efficient quadrature rules for illumination integrals: from Quasi Monte Carlo to Bayesian Monte Carlo. *Synthesis Lectures on Computer Graphics and Animation*, 7(2):1–92, 2015.

T. Minka. Deriving quadrature rules from Gaussian processes. Technical report, Statistics Department, Carnegie Mellon University, 2000.

S. Mosbach and A. G. Turner. A quantitative probabilistic investigation into the accumulation of rounding errors in numerical ODE solution. *Computers & Mathematics with Applications*, 57(7): 1157–1167, 2009.

F. Narcowich, J. Ward, and H. Wendland. Sobolev bounds on functions with scattered zeros, with applications to radial basis function surface fitting. *Mathematics of Computation*, 74(250): 743–763, 2005.

R. Nickl and J. Söhl. Nonparametric Bayesian posterior contraction rates for discretely observed scalar diffusions. *arXiv:1510.05526*, 2015.

E. Novak and H. Woźniakowski. *Tractability of Multivariate Problems Volume I: Linear Information*. European Mathematical Society Publishing House, EMS Tracts in Mathematics 6, 2008.

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

C. J. Oates and M. Girolami. Control functionals for quasi Monte Carlo integration. *arXiv:1501.03379*, 2015.

C. J. Oates, M. Girolami, and N. Chopin. Control functionals for Monte Carlo integration. *arXiv:1410.2392*, 2015.

A. O'Hagan. Bayes–Hermite quadrature. *Journal of Statistical Planning and Inference*, 29:245–260, 1991.

A. O'Hagan. Some Bayesian numerical analysis. *Bayesian Statistics*, 4:345–363, 1992.

C. S. Ong, A. Smola, and B. Williamson. Learning the kernel with hyperkernels. *Journal of Machine Learning Research*, 6:1043–1071, 2005.

M. A. Osborne. *Bayesian Gaussian processes for sequential prediction, optimisation and quadrature*. PhD thesis, University of Oxford, 2010.

M. A. Osborne, D. Duvenaud, R. Garnett, C. E. Rasmussen, S. Roberts, and Z. Ghahramani. Active learning of model evidence using Bayesian quadrature. In *Advances In Neural Information Processing Systems*, pages 46–54, 2012.

Art B Owen. A constraint on extensible quadrature rules. *Numerische Mathematik*, pages 1–8, 2014.

M. Pharr and G. Humphreys. *Physically based rendering: From theory to implementation*. Morgan Kaufmann, 2004.

H. Poincaré. *Calcul des probabilités*. Gauthier-Villars, 1912.

J. Quinonero-Candela and C. E. Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6:1939–1959, 2005.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances In Neural Information Processing Systems*, pages 1177–1184, 2007.

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

C. E. Rasmussen and Z. Ghahramani. Bayesian Monte Carlo. In *Advances in Neural Information Processing Systems*, pages 489–496, 2002.

S. Särkka, J. Hartikainen, L. Svensson, and F. Sandblom. On the relation between Gaussian process quadratures and sigma-point methods. 2015.

M. Schober, D. Duvenaud, and P. Hennig. Probabilistic ODE solvers with Runge-Kutta means. In *Advances in Neural Information Processing Systems*, pages 739–747, 2014.

B. Schölkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, 2002.

Amar Shah, Andrew Gordon Wilson, and Zoubin Ghahramani. Student-t processes as alternatives to gaussian processes. *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, 2014.

Q. Shi, J. Petterson, G. Dror, J. Langford, A. Smola, and S. V. N. Vishwanathan. Hash kernels for structured data. *Journal of Machine Learning Research*, 10:2615–2637, 2009.

W. Sickel and T. Ullrich. Tensor products of Sobolev–Besov spaces and applications to approximation from the hyperbolic cross. *Journal of Approximation Theory*, 161(2):748–786, 2009.

V. Sinescu, F. Y. Kuo, and I. H. Sloan. On the choice of weights in a function space for quasi-Monte Carlo methods for a class of generalised response models in statistics. In *Monte Carlo and Quasi-Monte Carlo Methods*. 2012.

I. H. Sloan and H. Woźniakowski. When are Quasi-Monte Carlo algorithms efficient for high dimensional integrals? *Journal of Complexity*, 14(1):1–33, 1998.

A. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *Proceedings of the 18th International Conference on Algorithmic Learning Theory*, pages 13–31, 2007.

J. Snoek, H. Larochelle, and R. P. Adams. Practical Bayesian optimization of machine learning algorithms. In *Advances In Neural Information Processing Systems*, pages 2951–2959, 2012.

J. Snoek, O. Rippel, K. Swersky, R. Kiros, N. Satish, N. Sundaram, M. A. Patwary, Prabhat, and R. P. Adams. Scalable Bayesian optimization using deep neural networks. In *International Conference on Machine Learning*, pages 2171–2180, 2015.

B. Szabó, A. van der Vaart, and J. van Zanten. Frequentist coverage of adaptive nonparametric Bayesian credible sets. *The Annals of Statistics*, 43(4):1391–1428, 2015.

A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.

A. van Der Vaart and H. van Zanten. Information rates of nonparametric Gaussian process methods. *Journal of Machine Learning Research*, 12:2095–2119, 2011.

A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):480–492, 2012.

Z. Wang, M. Zoghi, F. Hutter, D. Matheson, and N. De Freitas. Bayesian optimization in high dimensions via random embeddings. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence.*, pages 1778–1784, 2013.

H. Wendland. Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Advances in Computational Mathematics*, 4:389–396, 1995.

J. Yang, V. Sindhwani, H. Avron, and M. Mahoney. Quasi-Monte Carlo feature maps for shift-invariant kernels. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, 2014.

A Proof of Theoretical Results

Proof of Lemma 3. Assume without loss of generality that $D < \infty$. The posterior distribution over $\Pi[f]$ is Gaussian with mean $m_n := \hat{\Pi}[f]$ and variance v_n. Since $v_n = \|\hat{\Pi}_{\mathrm{BQ}} - \Pi\|^2_{\mathrm{op}}$ we have $v_n \leq \delta_n^2$. Now the posterior probability mass on I_D^c is given by

$$\mathbb{P}[I_D^c|\mathcal{D}] = \int_{I_D^c} \phi(r|m_n, v_n)\mathrm{d}r, \tag{49}$$

where $\phi(r|m_n, v_n)$ is the p.d.f. of the $\mathcal{N}(m_n, v_n)$ distribution. From the definition of D we get the upper bound

$$\begin{aligned}
\mathbb{P}[I_D^c|\mathcal{D}] \quad &\leq \quad \int_{-\infty}^{\Pi[f]-D} \phi(r|m_n, v_n)\mathrm{d}r + \int_{\Pi[f]+D}^{\infty} \phi(r|m_n, v_n)\mathrm{d}r & (50)\\
&= \quad 1 + \Phi\Big(\underbrace{\frac{\Pi[f] - m_n}{\sqrt{v_n}}}_{(*)} - \frac{D}{\sqrt{v_n}}\Big) - \Phi\Big(\underbrace{\frac{\Pi[f] - m_n}{\sqrt{v_n}}}_{(*)} + \frac{D}{\sqrt{v_n}}\Big). & (51)
\end{aligned}$$

From the definition of the MMD we have that the terms $(*)$ are bounded by $\|f\|_{\mathcal{H}} < \infty$, so that asymptotically as $v_n \downarrow 0$ we have

$$\begin{aligned}
\mathbb{P}[I_D^c|\mathcal{D}] \quad &\lesssim \quad 1 + \Phi\big(-D/\sqrt{v_n}\big) - \Phi\big(D/\sqrt{v_n}\big)\\
&\lesssim \quad 1 + \Phi\big(-D/\delta_n\big) - \Phi\big(D/\delta_n\big)\\
&\lesssim \quad \mathrm{erfc}\big(D/\sqrt{2}\delta_n\big) \;=\; o(\exp(-C\delta_n^{-2}))
\end{aligned} \tag{52}$$

where $C = D^2/2$ and we have used the fact that $\mathrm{erfc}(x) \lesssim x^{-1}\exp(-x^2/2)$. □

The following technical lemma is elementary but we could not find a proof in the existing literature. We therefore provide one below:

Lemma 4. *An expectation $\mathbb{E}_X$ over MC samples $X = \{\boldsymbol{x}_i\}_{i=1}^n$ obeys a scaling relationship*

$$\mathbb{E}_X h_X^\gamma \quad = \quad O(n^{-\gamma/d+\epsilon}) \tag{53}$$

for $\epsilon > 0$ arbitrarily small, where $h_X = \sup_{\boldsymbol{x}\in\mathcal{X}} \min_{i=1,\ldots,n} \|\boldsymbol{x} - \boldsymbol{x}_i\|$ is the fill distance *of X in $\mathcal{X}$.*

Proof. Define a uniform grid of reference points $\{\boldsymbol{g}_i\}_{i=1}^G \subset \mathcal{X}$ consisting of all $G = m^d$ $(m > 1)$ states of the form $\boldsymbol{g} = (g_1, \ldots, g_d)$ where $g_i \in \{0, \frac{1}{m-1}, \ldots, \frac{m-2}{m-1}, 1\}$. Consider the event $E = [\forall i \exists j : \|\boldsymbol{g}_i - \boldsymbol{x}_j\| \leq \frac{1}{m-1}]$ where for each grid point $\boldsymbol{g}_i$ there is a state $\boldsymbol{x}_j$ within a distance $\frac{1}{m-1}$ of it. From the triangle inequality it follows that E implies the event $[h_X \leq \frac{3}{m-1}]$. i.e. there cannot be "large holes" in X. Now, writing $\mathbb{P}_X[A] = \mathbb{E}_X[A]$, we have

$$\mathbb{P}_X[E^c] \quad = \quad \mathbb{P}_X\left[\exists i : \forall j, \|\boldsymbol{g}_i - \boldsymbol{x}_j\| > \frac{1}{m-1}\right] \;\leq\; \sum_{i=1}^{G} \underbrace{\mathbb{P}_X\left[\forall j, \|\boldsymbol{g}_i - \boldsymbol{x}_j\| > \frac{1}{m-1}\right]}_{(*)}. \tag{54}$$

The probability of the event $(*)$ is largest when $\boldsymbol{g}_i$ lies on one of the corners of $\mathcal{X}$; e.g. $\boldsymbol{g}_i = (0, \ldots, 0)$. In that case $\mathbb{P}_X[(*)] = (1-V)^n$ where $V = 2^{-d}\pi^{d/2}/\Gamma(d/2+1)(m-1)^d$ is the volume of the intersection of $\mathcal{X}$ with the ball of radius $\frac{1}{m-1}$ centred on $\boldsymbol{g}_i$. Thus

$$\mathbb{P}\left[h_X \leq \frac{3}{m-1}\right] \quad \geq \quad \mathbb{P}_X[E] \quad \geq \quad 1 - G\left[1 - \frac{2^{-d}\pi^{d/2}}{\Gamma(d/2+1)(m-1)^d}\right]^n \tag{55}$$

Letting $\zeta = \frac{3}{m-1}$ implies that $m = 1 + \frac{3}{\zeta}$ and $G = (1+\frac{3}{\zeta})^d$. In this reparametrisation we have

$$\begin{aligned} \mathbb{P}_X[h_X \leq \zeta] &\geq 1 - \left(1+\frac{3}{\zeta}\right)^d \left[1 - \frac{6^{-d}\pi^{d/2}}{\Gamma(d/2+1)}\zeta^d\right]^n & (56) \\ &\geq 1 - \left(\frac{4}{\zeta}\right)^d (1 - C_d\zeta^d)^n, & (57) \end{aligned}$$

where we have written $C_d = 6^{-d}\pi^{d/2}/\Gamma(d/2+1)$. While Eqn. 57 holds only for ζ of the form $\frac{3}{m-1}$, it can be made to hold for all $0 < \zeta < 1$ by replacing C_d with $\tilde{C}_d = 2^{-d}C_d$. This is because for any $0 < \zeta < 1$ there exists $m > 1$ such that $\tilde{\zeta} = \frac{3}{m-1}$ satisfies $\frac{\zeta}{2} \leq \tilde{\zeta} < \zeta$, along with the fact that $\mathbb{P}_X[h_X \leq \zeta] \leq \mathbb{P}_X[h_X \leq \tilde{\zeta}]$.

From the reverse Markov inequality, since $h_X^\gamma \leq 1$ with probability one, we have that for all $\zeta < \mathbb{E}[h_X^\gamma]$,

$$\mathbb{P}_X[h_X^\gamma > \zeta] \quad \geq \quad \frac{\mathbb{E}_X[h_X^\gamma] - \zeta}{1-\zeta} \tag{58}$$

and upon rearranging

$$\mathbb{E}_X[h_X^\gamma] \quad \leq \quad 1 - (1-\zeta)\mathbb{P}_X[h_X \leq \zeta^{1/\gamma}]. \tag{59}$$

Combining Eqns. 57 and 59 leads to

$$\begin{aligned} \mathbb{E}_X[h_X^\gamma] &\leq \zeta + (1-\zeta)\left(\frac{4}{\zeta^{1/\gamma}}\right)^d (1 - C_d\zeta^{d/\gamma})^n \\ &\leq \zeta + \left(\frac{4}{\zeta^{1/\gamma}}\right)^d (1 - C_d\zeta^{d/\gamma})^n. \end{aligned} \tag{60}$$

Now, letting $\zeta = n^{-\delta}$ for some fixed $\delta > 0$ and varying n, we have that

$$\mathbb{E}_X[h_X^\gamma] \quad \leq \quad \frac{1}{n^\delta} + 4^d \underbrace{n^{d\delta/\gamma}(1 - \tilde{C}_d n^{-d\delta/\gamma})^n}_{(**)} \tag{61}$$

where $(**) \sim \exp(-\tilde{C}_d n^{1-d\delta/\gamma})$. The right hand side of Eqn. 61 is asymptotically minimised by taking $\delta = \frac{\gamma}{d} - \epsilon$ for $\epsilon > 0$ arbitrarily small. We therefore conclude that for $\delta < \gamma/d$ and $\epsilon > 0$ arbitrarily small, $\mathbb{E}_X[h_X^\gamma] = O(n^{-\gamma/d+\epsilon})$, as required. □

Proof of Thm. 1. Initially consider fixed states $X = \{\boldsymbol{x}_i\}_{i=1}^n$ (i.e. fixing the random seed) and $\mathcal{H} = \mathcal{H}_\alpha$. Define h_X as in Lemma 4. From standard results in functional approximation (Narcowich et al., 2005, Thm. 1.1) there exists $C > 0$ such that

$$\|f - \mathbb{E}[f|\mathcal{D}]\|_2 \leq C h_X^\alpha \|f\|_{\mathcal{H}}. \tag{62}$$

From the regression bound (Lemma 2),

$$|\hat{\Pi}_{\mathrm{BMC}}[f] - \Pi[f]| \leq \|f - \mathbb{E}[f|\mathcal{D}]\|_2. \tag{63}$$

Combining Eqns. 62 and 63 produces $\|\hat{\Pi}_{\mathrm{BMC}} - \Pi\|_{\mathrm{op},\mathcal{H}_\alpha} \leq C h_X^\alpha$, where we have made the $\mathcal{H}_\alpha$-dependence of the MMD explicit in the notation. Now, taking an expectation $\mathbb{E}_X$ over the states $X = \{\boldsymbol{x}_i\}_{i=1}^n$, viewed as independent draws from Π, we have

$$\mathbb{E}_X \|\hat{\Pi}_{\mathrm{BMC}} - \Pi\|_{\mathrm{op},\mathcal{H}_\alpha} \quad \leq \quad C \mathbb{E}_X h_X^\alpha. \tag{64}$$

From Lemma 4 we have a scaling relationship

$$\mathbb{E}_X h_X^\alpha \quad = \quad O(n^{-\alpha/d+\epsilon}) \tag{65}$$

for $\epsilon > 0$ arbitrarily small. From Markov's inequality, convergence in mean implies convergence in probability and thus, combining Eqns. 64 and 65, we have

$$\|\hat{\Pi}_{\mathrm{BMC}} - \Pi\|_{\mathrm{op},\mathcal{H}_\alpha} \quad = \quad O_P(n^{-\alpha/d+\epsilon}). \tag{66}$$

This completes the proof for $\mathcal{H} = \mathcal{H}_\alpha$. More generally, if $\mathcal{H}$ is norm-equivalent to $\mathcal{H}_\alpha$ then the result follows from the fact that $\|\hat{\Pi}_{\mathrm{BMC}} - \Pi\|_{\mathrm{op},\mathcal{H}} \leq \lambda \|\hat{\Pi}_{\mathrm{BMC}} - \Pi\|_{\mathrm{op},\mathcal{H}_\alpha}$ for some $\lambda > 0$. □

Proof of Thm. 2. From Theorem 15.21 of Dick and Pillichshammer (2010), the QMC rule $\hat{\Pi}_{\mathrm{QMC}}$ based on a higher-order digital $(t, \alpha, 1, \alpha m \times m, d)$ net over $\mathbb{Z}_b$ for some prime b satisfies

$$\|\hat{\Pi}_{\mathrm{QMC}} - \Pi\|_{\mathrm{op}} \leq C_{d,\alpha} \frac{(\log n)^{d\alpha}}{n^\alpha} = O_P(n^{-\alpha+\epsilon}) \tag{67}$$

for $\mathcal{S}_\alpha$ the Sobolev space of dominating mixed smoothness order α, where $C_{d,\alpha} > 0$ is a constant that depends only on d and α (but not on n). The result follows immediately from Bayesian reweighting (Lemma 1) and norm equivalence. The contraction rate is obtained by applying Lemma 3. □

Proof of Prop. 6. Initially consider fixed states $\{\boldsymbol{x}_i\}_{i=1}^n$ (i.e. fixing the random seed). Fix a particular integration problem whose true integrand is $f_0 \in \mathcal{H}$. Since the MMD (squared) coincides with the posterior variance, we have from Jensen's inequality

$$\|\hat{\Pi}^{\boldsymbol{e}}_{\mathrm{BMC}} - \Pi\|^2_{\mathrm{op}} \;=\; \mathbb{E}[\Pi[f] - \mathbb{E}[\Pi[f]]]^2 \;=\; \mathbb{E}[\Pi[f - \mathbb{E}[f]]]^2 \;\leq\; \mathbb{E}\|f - \mathbb{E}[f]\|_2^2. \tag{68}$$

Here $\mathbb{E} = \mathbb{E}[\cdot|\{\boldsymbol{x}_i^{\mathrm{MC}}, y_i\}_{i=1}^n]$ denotes an expectation with respect to the posterior GP that includes a model for the observation noise. Noting that $\mathbb{E}[f]$ is the variational minimiser of the posterior least squares loss, we have $\mathbb{E}\|f - \mathbb{E}[f]\|_2^2 \leq \mathbb{E}\|f - f_0\|_2^2$. Now, taking an expectation $\mathbb{E}_X$ over the states $\{\boldsymbol{x}_i\}_{i=1}^n$, viewed as independent draws from Π, we have

$$\mathbb{E}_X \|\hat{\Pi}^{\boldsymbol{e}}_{\mathrm{BMC}} - \Pi\|^2_{\mathrm{op}} \quad \leq \quad \mathbb{E}_X \mathbb{E}\|f - f_0\|_2^2. \tag{69}$$

Since the left hand side of Eqn. 64 is independent of f_0, it suffices to exhibit a particular regression problem f_0 for which the right hand side converges at a known rate. Following van Der Vaart and van Zanten (2011), suppose in addition that $f_0 \in \mathcal{C}_\alpha \cap \mathcal{H}_\alpha$ for $\alpha > d/2$. Here $\mathcal{C}_\alpha$ is the Hölder space on $[0,1]^d$ and $\mathcal{H}_\alpha$ is the Sobolev space on $[0,1]^d$, which each contain, for example, the function $f_0 \equiv 0$. Then from Theorem 5 of van Der Vaart and van Zanten (2011) we have a scaling relationship

$$\mathbb{E}_X \mathbb{E} \|f - f_0\|_2^2 \quad \sim \quad n^{-2\alpha/(2\alpha+d)}. \tag{70}$$

Tsybakov (2008) proves that this rate is minimax for the noisy regression problem. From Markov's inequality, convergence in mean implies convergence in probability and thus, combining Eqns. 69 and 70, we have

$$\|\hat{\Pi}^{\boldsymbol{e}}_{\mathrm{BMC}} - \Pi\|_{\mathrm{op}} \quad = \quad O_P(n^{-\alpha/(2\alpha+d)}). \tag{71}$$

On the other hand, if we have a Gaussian kernel then we suppose in addition that f_0 is a restriction to $[0,1]^d$ of an element of $\mathcal{A}^{\gamma,r}(\mathbb{R}^d)$, for $r \geq 1$ and $\gamma > 0$, defined to be the set of functions whose Fourier transform $\mathfrak{F} f_0$ satisfies

$$\int \exp(\gamma \|\boldsymbol{\xi}\|^r) |\mathfrak{F} f_0|^2(\boldsymbol{\xi}) \mathrm{d}\boldsymbol{\xi} \quad < \quad \infty. \tag{72}$$

Again, the function $f_0 \equiv 0$ belongs to $\mathcal{A}^{\gamma,r}(\mathbb{R}^d)$. This time, from Theorem 10 of van Der Vaart and van Zanten (2011) we have a scaling relationship

$$\mathbb{E}_X \mathbb{E} \|f - f_0\|_2^2 \quad \sim \quad (\log n)^{2/r}/n. \tag{73}$$

Since the function $f_0 \equiv 0$ belongs to $\mathcal{A}^{\gamma,r}(\mathbb{R}^d)$ for all $r \geq 1$ we conclude, via Markov's inequality as before, that

$$\|\hat{\Pi}^{\boldsymbol{e}}_{\mathrm{BMC}} - \Pi\|_{\mathrm{op}} \quad = \quad O_P(n^{-1/2+\epsilon}) \tag{74}$$

where $\epsilon > 0$ can be arbitrarily small. This completes the proof. □

Proof of Thm. 3. Bondarenko et al. (2013) showed that for all $d \geq 2$ there exists C_d such that for all $n \geq C_d t^d$ there exists a spherical t-design on $\mathbb{S}^d$ with n points. On the other hand, Hesse and Sloan (2005) showed that such a design would achieve $\|\hat{\Pi}_{\mathrm{QMC}} - \Pi\|_{\mathrm{op}} = O(n^{-3/4})$ in the case where $\alpha = 3/2$ and $d = 2$. (A recent survey of these results is provided by Brauchart et al. (2014).) The result follows from Bayesian re-weighting (Lemma 1). □

Proof of Thm. 4. The QMC rate

$$\|\hat{\Pi}_{\mathrm{QMC}} - \Pi\|_{\mathrm{op}} = O(b^{-\alpha m} m^{d\alpha}) \tag{75}$$

is proven in Theorem 15.21 of Dick and Pillichshammer (2010). The number of quadrature points in such a net is $n = b^m$, so that this rate is just the familiar $O(n^{-\alpha+\epsilon})$. The result follows immediately from Bayesian re-weighting (Lemma 1). □

Appendices B to J are provided as supplementary text.

Supplemental Text

B Illustration of Proof Techniques

In this Appendix we obtain a convergence rate for OBQ as originally formulated in the seminal paper of O'Hagan (1991). As a stepping stone, we initially consider an *ad-hoc* rule for one-dimensional integration that could reasonably be called *Bayesian Gauss-Hermite* quadrature (BGHQ). Indeed, suppose $\mathcal{X} = \mathbb{R}$, σ is Lebesgue measure and Π is the $N(0, \nu_\pi^2)$ distribution. Then the BGHQ estimator, denoted by $\hat{\Pi}_{\mathrm{BGHQ}}[f]$, corresponds to BQ with a Gaussian kernel $k(x, y) = \exp(-(x - y)^2/2\nu_k^2)$ and to states $\{x_i\}_{i=1}^n$ that are chosen at the zeros of the generalised Hermite polynomials $H_n^{[\nu_\pi^2]}$ of degree n, defined by the rescaling

$$H_n^{[\nu_\pi^2]}(x) := (2\nu_\pi^2)^{-n/2} H_n(x/\sqrt{2}\nu_\pi) \tag{76}$$

where H_n are the standard Hermite polynomials. A simple re-weighting argument, based on Lemma 1, produces the following:

Theorem 5 (BGHQ convergence rate)**.** *Let $\nu_\pi/\nu_k < 1$. The BGHQ rule satisfies*

$$\|\hat{\Pi}_{BGHQ} - \Pi\|_{op} = O((\nu_\pi/\nu_k)^{2n}) \tag{77}$$

and hence the posterior mean converges exponentially. Furthermore, the posterior contracts super-exponentially:

$$\mathbb{P}[I_D^c|\mathcal{D}] = o(\exp(-C(\nu_k/\nu_\pi)^{4n})), \tag{78}$$

where I_D and C were defined in Lemma 3.

Proof. BGHQ is a re-weighted version of Gauss-Hermite quadrature (GHQ), a quadrature rule with states $\{x_i^{\mathrm{GHQ}}\}_{i=1}^n$ chosen at the zeros of the generalized Hermite polynomials $H_n^{[\nu_\pi]}$ of degree n and weights w_i^{GHQ} chosen such that $\hat{\Pi}_{\mathrm{GHQ}}$ is exact for all polynomials of degree $2n - 1$ or less:

$$w_i^{\mathrm{GHQ}} := \frac{n!}{\nu_\pi^{n-1} n^2 H_{n-1}^{[\nu_\pi]}(x_i^{\mathrm{GHQ}})^2}. \tag{79}$$

Theorem 4.1 in Kuo and Woźniakowski (2012) establishes a rate for the worst case error of

$$\|\hat{\Pi}_{\mathrm{GHQ}} - \Pi\|_{\mathrm{op}} = 2^{-1/4}\Big(\frac{\nu_\pi}{\nu_k}\Big)^n (1 + o(1)) = O\Big(\Big(\frac{\nu_\pi}{\nu_k}\Big)^n\Big). \tag{80}$$

The result for BGHQ immediately follows from Bayesian re-weighting (Lemma 1). Furthermore the contraction rate can be obtained by applying Lemma 3. □

Remark 8. *Särkka et al. (2015) showed that, in fact, the classical GHQ weights are exactly the BGHQ weights when the latter is performed in an RKHS with kernel*

$$k(x, x') = \sum_{i=1}^{2n-1} \sum_{j=1}^{2n-1} \frac{1}{i!j!} \lambda_{i,j} H_i(x) H_j(x') \tag{81}$$

for a particular choice of the $\lambda_{i,j}$.

Now we turn to the OBQ method proposed in O'Hagan (1991) and documented further in O'Hagan (1992)[8]. (Recall that OBQ selects states $\{\boldsymbol{x}_i\}_{i=1}^n$ to globally minimise the worst-case integration error.) An immediate corollary of Theorem 5 provides convergence rates for OBQ in one dimension:

Corollary 1 (**OBQ convergence rate**). *Take $\mathcal{X} = \mathbb{R}$, σ be the Lebesgue measure and Π be the $N(0, \nu_\pi^2)$ distribution. Suppose that $\nu_\pi/\nu_k < 1$. Consider OBQ based on the Gaussian kernel $k(x, y) = \exp(-(x-y)^2/2\nu_k^2)$. Then $\|\hat{\Pi}_{OBQ} - \Pi\|_{op} = O((\nu_\pi/\nu_k)^{2n})$ and $\mathbb{P}[I_D^c|\mathcal{D}] = o(\exp(-C(\nu_k/\nu_\pi)^{4n}))$.*

Proof. From the definition of OBQ we have

$$\hat{\Pi}_{\text{OBQ}} := \arg\min_{\hat{\Pi}} \|\hat{\Pi} - \Pi\|_{\text{op}} \tag{82}$$

where the minimum is taken over the set of all valid quadrature rules $\hat{\Pi}$; i.e. over the location of states $\{x_i\}_{i=1}^n \subset \mathcal{X}$ and weights $\{w_i\}_{i=1}^n \in \mathbb{R}$. Consider a specific quadrature rule $\hat{\Pi} = \hat{\Pi}_{\text{BGHQ}}$: From Theorem 5, we have that

$$\|\hat{\Pi}_{\text{OBQ}} - \Pi\|_{\text{op}} \leq \|\hat{\Pi}_{\text{BGHQ}} - \Pi\|_{\text{op}} = 2^{-1/4}\Big(\frac{\nu_\pi}{\nu_k}\Big)^n(1 + o(1)) \tag{83}$$

as required. Furthermore the contraction rate can be obtained by applying Lemma 3. □

To the best of our knowledge this is the first formal proof that OBQ converges at an exponential rate in an infinite dimensional RKHS setting (Briol et al., 2015, proves this only for finite-dimensional RKHS).

C Digital Nets for BQMC

This appendix provides a concise definition of higher-order digital nets.

Definition 1 (Digital net). *Let b be a prime and let $m, m' \geq 1$ be integers, where $m' \geq m$. Let $\boldsymbol{C}_1, \ldots, \boldsymbol{C}_d$ be $m' \times m$ matrices over the finite field $\mathbb{F}_b = \{0, 1, \ldots, b-1\}$ of order b. Below we construct $n = b^m$ states on $\mathcal{X} = [0,1)^d$: For $0 \leq i \leq b^m - 1$, let $i = i_0 + i_1 b + \cdots + i_{m-1}b^{m-1}$ be the b-adic expansion of i (i.e. the i_j are the unique $i_j \in \mathbb{F}_b$ for which the equality is satisfied). Identify i with the vector $\boldsymbol{i} = (i_0, \ldots, i_{m-1})^T \in \mathbb{F}_b^m$. For $1 \leq j \leq d$ multiply the matrix $\boldsymbol{C}_j$ by $\boldsymbol{i}$, i.e.,*

$$\boldsymbol{C}_j\boldsymbol{i} := (y_{j,1}(\boldsymbol{i}), \ldots, y_{j,m'}(\boldsymbol{i}))^T \in \mathbb{F}_b^{m'} \tag{84}$$

and set

$$[\boldsymbol{x}_{i+1}]_j := \frac{y_{j,1}(\boldsymbol{i})}{b} + \cdots + \frac{y_{j,m'}(\boldsymbol{i})}{b^{m'}}. \tag{85}$$

The point set $\{\boldsymbol{x}_i\}_{i=1}^n$ is called a digital net *over $\mathbb{F}_b$, with generating matrices $\boldsymbol{C}_1, \ldots, \boldsymbol{C}_d$.*

[8]Somewhat confusingly, this approach was originally named *Bayes-Hermite quadrature*, but Hermite polynomials do not feature in its construction.

Definition 2 (Higher-order digital net)**.** *In the setting of Defn. 1, additionally let* $\alpha \geq 1$ *and* $0 < \beta \leq \min(1, \alpha m/m')$ *be real numbers and let* $0 \leq t \leq \beta m'$ *be a natural number. Write* $\boldsymbol{C}_j = (\boldsymbol{c}_{j,1}, \dots, \boldsymbol{c}_{j,m'})^T$*. If for all* $1 \leq i_{j,v_j} < \cdots < i_{j,1}$*, where* $0 \leq v_j$ *for all* $j = 1, \dots, d$*, with*

$$\sum_{j=1}^{d} \sum_{l=1}^{\min(v_j,\alpha)} i_{j,l} \leq \beta m' - t, \tag{86}$$

the vectors $\boldsymbol{c}_{1,i_{1,v_1}}, \dots, \boldsymbol{c}_{1,i_{1,1}}, \dots, \boldsymbol{c}_{d,i_{d,v_d}}, \dots, \boldsymbol{c}_{d,i_{d,1}} \in \mathbb{F}_b^m$ *are linearly independent over* $\mathbb{F}_b$*, then the digital net with generating matrices* $\boldsymbol{C}_1, \dots, \boldsymbol{C}_d$ *is called a* higher-order digital $(t, \alpha, \beta, m' \times m, d)$ net *over* $\mathbb{F}_b$*.*

The definition of a digital net is constructive, in the sense that it specifies a unique collection of states $\{\boldsymbol{x}_i\}_{i=1}^n$ where $n = b^m$. In contrast, the definition of a higher-order digital net is non-constructive and it is not immediately clear whether any digital nets are also higher-order digital nets. However, explicit constructions of higher-order digital $(t, \alpha, 1, \alpha m \times m, d)$ nets over $\mathbb{Z}_b$ for all prime numbers b and $\alpha, d, m \in \mathbb{N}$ are known and are given in Dick and Pillichshammer (2010, Sec. 15.2). (The natural number t is a deterministic function of the generating matrices and α and is not important for this paper.)

D Extending BQMC to Infinitely Smooth Functions

In this appendix we provide additional results for BQMC that cover spaces of infinitely differentiable functions; so-called *Korobov spaces* (Dick, 2011). For simplicity we present only the case where the integrand f is a periodic function on $\mathcal{X} = [0,1]^d$, with σ the Lebesgue measure and Π is uniform, but we allow for the possibility that $f : \mathcal{X} \to \mathbb{C}$ produces complex values. Periodicity allows us to leverage the Fourier series representation

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{\omega} \in \mathbb{Z}^d} \hat{f}(\boldsymbol{\omega}) \exp(2\pi i \boldsymbol{\omega} \cdot \boldsymbol{x}), \tag{87}$$

where the Fourier coefficients

$$\hat{f}(\boldsymbol{\omega}) = \int_{\mathcal{X}} f(\boldsymbol{x}) \exp(-2\pi i \boldsymbol{\omega} \cdot \boldsymbol{x}) \mathrm{d}\boldsymbol{x} \tag{88}$$

are assumed to decay exponentially fast; $\hat{f}(\boldsymbol{\omega}) = O(h^{|\boldsymbol{\omega}|})$ where $\boldsymbol{\omega} = (\omega_1, \dots, \omega_d)$, $|\boldsymbol{\omega}| = |\omega_1| + \cdots + |\omega_d|$ and $0 < h < 1$, implying that f is infinitely differentiable. Following recent work by Dick et al. (2011) we focus on the particular Korobov space that is the RKHS generated by the kernel

$$k(\boldsymbol{x}, \boldsymbol{y}) = \sum_{\boldsymbol{\omega} \in \mathbb{Z}^d} W_{\boldsymbol{\omega}} \exp(2\pi i \boldsymbol{\omega} \cdot (\boldsymbol{x} - \boldsymbol{y})), \tag{89}$$

where coefficients $W_{\boldsymbol{\omega}} \geq 0$ satisfy $W_{\mathbf{0}} = 1$ and

$$\sum_{\boldsymbol{\omega} \in \mathbb{Z}^d} W_{\boldsymbol{\omega}} < \infty. \tag{90}$$

Smoothness of the functions $f \in \mathcal{H}(k)$ depends on how rapidly this sum converges as $|\boldsymbol{\omega}| \to \infty$. Further details can be found in Dick et al. (2011). We note that the kernel mean $\Pi[k(\cdot, \boldsymbol{y})]$ is available in closed form.

The additional prior information that is provided by a Korobov space is enough to provide exponential convergence rates for BQMC:

Proposition 7 (BQMC in Korobov spaces)**.** *Let $\mathcal{X} = [0,1]^d$ and let Π be uniform on $\mathcal{X}$. Consider a Korobov space $\mathcal{H}(k)$ as defined above. Then, for n prime, there exists a choice of states $\{\boldsymbol{x}_i\}_{i=1}^n$ such that the corresponding BQMC rule satisfies*

$$\|\hat{\Pi}_{BQMC} - \Pi\|_{op} = O(h^{1/4(d!n)^{1/d}} n^{1/2}) \tag{91}$$

for some d-dependent constant $c_d > 0$ that does not depend on n.

Proof. The proof is analogous to the Sobolev case: We leverage an established QMC worst case error bound for Korobov spaces; in this case Theorem 2 in Dick et al. (2011). Bayesian re-weighting (Lemma 1) completes the proof. □

To limit scope we do not discuss the explicit construction of the states whose existence is guaranteed by Theorem 7. Sec. 6 of Dick et al. (2011) provides further details.

E De-biasing the Probabilistic Integrals

A consequence of incorporating prior information is that the point estimate provided by the posterior mean $\mathbb{E}[\Pi[f]|\mathcal{D}]$ is no longer unbiased, in the frequentist sense, as an estimator for $\Pi[f]$. While this is a non-issue from the probabilistic numerics perspective, the availability of unbiased estimators could help to broaden the applicability of probabilistic integrators. Here we present a simple modification that leads to unbiased estimation, as described in recent work in the MC (Oates et al., 2015) and QMC (Oates and Girolami, 2015) literature.

Consider splitting the states $\{\boldsymbol{x}_i\}_{i=1}^m \cup \{\boldsymbol{x}_i\}_{i=m+1}^n$. The first m states are used to train a GP $f \sim \mathcal{GP}(m_1, k_1)$. The remaining $n-m$ states are used to evaluate a uniformly weighted quadrature rule

$$\hat{\Pi}_{\mathrm{UB}}[f] := \frac{1}{n-m} \sum_{i=m+1}^{n} f(\boldsymbol{x}_i) - m_1(\boldsymbol{x}_i) + \Pi[m_1]. \tag{92}$$

When $m < n$ and the $\{\boldsymbol{x}_i\}_{i=m+1}^n$ are marginally distributed as Π, $\hat{\Pi}_{\mathrm{UB}}[f]$ is an unbiased estimator of $\Pi[f]$. (In a loose sense, the case $m = n$ corresponds to the BMC estimator.) Based on the data $\{\boldsymbol{x}_i, f_i\}_{i=1}^m$, this produces a probability model for $\Pi[f]$ that is Gaussian with mean $\hat{\Pi}_{\mathrm{UB}}[f]$ and variance $\frac{1}{n-m}\mathbb{V}[\Pi[f]|\{\boldsymbol{x}_i, f_i\}_{i=1}^m]$, where $\mathbb{V}[\Pi[f]|\{\boldsymbol{x}_i, f_i\}_{i=1}^m]$ is the BQ variance after observing only the first m samples.

F Calibration via Empirical Bayes

We consider calibration for BMC in $\mathcal{X} = [0,1]$ with Π uniform over $\mathcal{X}$. The Matérn kernel with $\beta = 5/2$ was employed and the length scale τ was considered to be unknown. For each value of n we estimates an appropriate value $\hat{\tau}_n$ for τ using empirical Bayes, as described in Sec. 4.1. BMC then proceeded on the basis of these estimated hyper-parameters. Results in Fig. 10, based on the

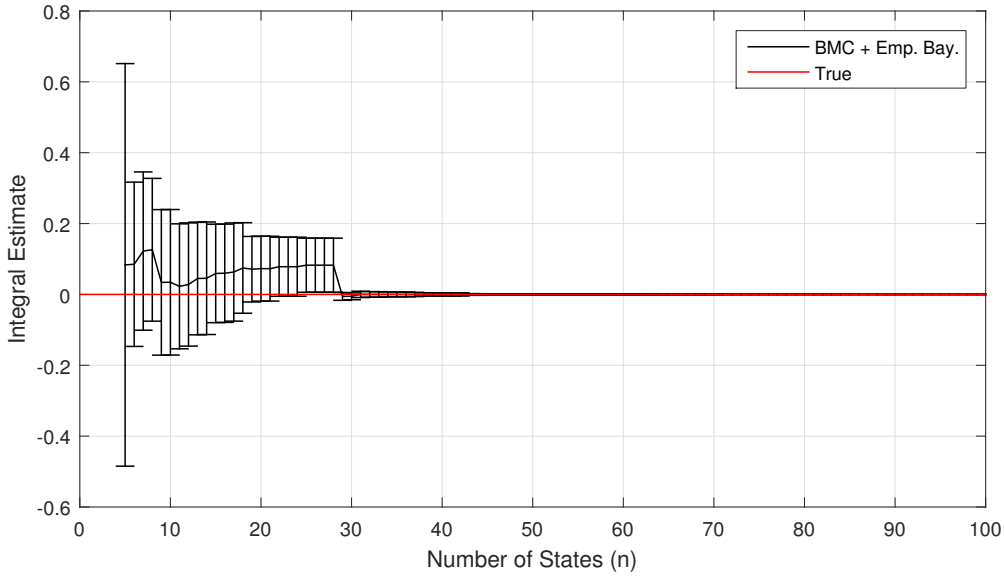


Figure 10: Calibration of kernel hyper-parameters using empirical Bayes. Results are shown for calibration of BMC on $\mathcal{X} = [0, 1]$ with Π uniform over $\mathcal{X}$. The Matérn kernel with $\beta = 5/2$ was employed and the length scale τ was considered to be unknown. Error bars show 90% posterior credible intervals.

integrand $f(x) = \sin(4\pi x)$, show that the posterior uncertainty is well-calibrated, with the truth typically covered by the posterior credible interval.

These results are in line with the recent work of Szabó et al. (2015), which guarantees appropriate posterior coverage when the integrand f is sufficiently smooth. A more extensive study of calibration for BQ was outside the scope of the present paper.

G Numerical Stability

As discussed in Sec. 4.4, computation of BQ weights can require numerical regularisation and this has the potential to negatively impact on the performance of the BQ estimator. Poorly conditioned kernel matrices $\boldsymbol{K}$ occur when two (or more) states $\boldsymbol{x}_i$, $\boldsymbol{x}_j$ are not well distinguished by the kernel k, as can occur when $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are close together. As n increases, so does the potential for poor conditioning.

Figure 11 describes the impact of numerical regularization in a challenging case where the value of the length-scale parameter $\sigma = 1.5$ in the Matérn kernel is high for a function on $\mathcal{X} = [0, 1]$, suggesting that functions can be well-approximated using a small number of points. This kernel therefore fails to clearly distinguish between nearby states. Results show that the BQMC methods with Matérn kernel with $\beta = 3/2$, $\beta = 5/2$ and $\beta = 7/2$ initially (small n) obey the theoretical convergence rates provided in Sec. 3.3.2, however after a few hundred observations the rate of convergence ceases to hold due to the numerical regularisation. This is consistent with the analysis of noisy data performed in Sec. 4.4.

Arguably, this issue is insignificant since BQMC dramatically outperforms QMC for $n \leq 100$ which is sufficient for precise estimation (MMD$< 10^{-5}$). Indeed, this example was chosen in

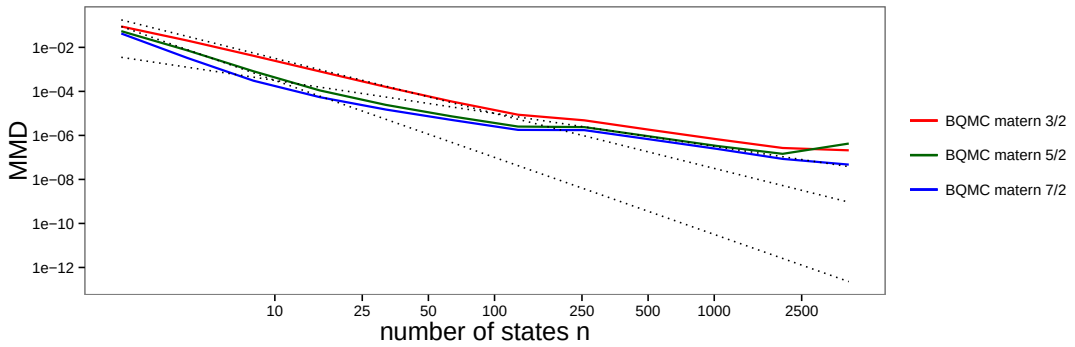


Figure 11: Numerical stability in Bayesian re-weighting. We consider the MMD for BQMC in $\mathcal{X} = [0, 1]$ using higher-order digital nets. We use the Matérn kernel with $\beta \in \{1/2, 5/2, 7/2\}$ with $\tau = 1.5$, so that only a few function evaluations are required for accurate integration. For this values of β, the kernel matrix becomes poorly conditioned when $n > 25$ and regularisation puts us into the "noisy function evaluation" regime, for which performance is known to be poor (see Sec. 4.4). Note however that a small error of 10^{-4} is already obtained for values of $n > 10$.

order to stretch the limits of the method. Furthermore, the issue will only occur in low-dimensional problems, since the curse of dimensionality will prevent states from being "too close". However these numerical issues are not present in the deterministic QMC method, so that QMC may be preferred to BQMC if the aim is to obtain extremely precise approximations (in which case modelling the numerical error is probably unnecessary).

H Scalability of B(MC)MC and BQMC

H.1 Scalability in the Number of States

In situations where f is cheap to evaluate, the naive $O(n^3)$ computational cost associated with kernel matrix inversion renders BQ unsuitable relative to the $O(n)$ cost of MC and QMC methods However, when f is expensive to evaluate, BQ methods can prove considerably more effective than their standard counterparts.

Exact inversion can be achieved at low cost through exploiting structure in the kernel matrix. Examples include: the use of kernels with compact support (e.g. Wendland, 1995) to induce sparsity in the kernel matrix; tensor product kernels (e.g. O'Hagan, 1991) in the context of inverting kernel matrices defined by tensor products of point sets in multivariate problems; using Toeplitz solvers for a stationary kernel evaluated on an evenly-spaced point set; and making use of low-rank kernels (e.g. polynomial kernels).

In addition there are many "approximate" techniques: (i) Reduced rank approximations reduce the computational cost to $O(nm^2)$ where $m << n$ is a parameter controlling the accuracy of the approximation, essentially an effective degree of freedom (Quinonero-Candela and Rasmussen, 2005; Bach, 2013; El Alaoui and Mahoney, 2015). (ii) Explicit feature maps designed for additive kernels (Vedaldi and Zisserman, 2012). (iii) Local approximations (Gramacy and Apley, 2015), training only on nearest neighbour data. (iv) Multi-scale approximations, whereby the high-level structure is modelled using a full GP and approximation schemes are applied to lower-level structure (Katzfuss,

2015). (v) Fast multipole methods (Ambikasaran and Darve, 2014). (vi) Random approximations of the kernel itself, rather than the kernel matrix, such as random Fourier features (RFF; Rahimi and Recht, 2007), spectral methods (Lazaro-Gredilla et al., 2010; Bach, 2015) and hash kernels (Shi et al., 2009). (RFF have previously been successfully applied in BQ by Briol et al. (2015).) (vii) Parallel programming provides an alternative perspective on complexity reduction, as discussed in (e.g.) Dai et al. (2014).

This does not represent an exhaustive list of the (growing) literature on GP computation. Note that the latter do not come with probability models for the additional source of numerical error introduced by the approximation.

H.2 Scalability in Dimension

High-dimensional integrals that arise in applications are, in many cases, effectively low-dimensional problems. This can occur either (i) when the distribution Π is effectively concentrated in a low-dimensional manifold in $\mathcal{X}$ (this is responsible for the excellent performance of (MC)MC in certain high-dimensional settings), or (ii) when the integrand f depends on only a subset of its inputs, possibly after a transformation (this is responsible for the excellent performance of QMC methods in certain high-dimensional settings; Dick et al., 2013). The B(MC)MC and BQMC methods that we study provably deliver performance that is at least equivalent to (MC)MC and QMC in settings (i) and (ii) respectively (see Sec. 5.2.3 for an empirical example with $d = 50$). Conversely, when neither Π nor f are effectively low-dimensional, all approaches to integration necessarily suffer from a curse of dimension. For example, for Π uniform on $\mathcal{X} = [0, 1]^d$ and f belonging to a general Sobolev space of order α, no deterministic integration algorithm can exceed the $O(n^{-\alpha/d})$ rate. Clearly this rate becomes arbitrarily slow as d tends to infinity. Nevertheless, we note that BQ estimators remain coherent, reverting to the prior in this degenerate limit. Having weights that tend to zero is natural from a Bayesian point of view since our approximation of the integrand f will become very poor as d grows with n fixed. Note also that de-biased probabilistic integrators (Sec. E) have weights $\boldsymbol{w}$ tending to $\boldsymbol{w}^{\mathrm{MC}}$ when d goes to infinity. Thus the de-biased estimator collapses onto the MC estimator, rather than the prior estimator.

We briefly note a number of alternative approaches exist for problems in which the effective dimensionality is low. In particular, low-dimensional random embeddings project the ambient space into a lower dimensional space using a randomized map, perform computation in that space and then map back the results to the original space (see e.g. Wang et al., 2013, in the context of Bayesian optimisation).

I Approximation of Kernel Means

If the kernel mean μ_π is not tractable, then it is not possible to compute the quantities $z_i = \mu_\pi(\boldsymbol{x}_i)$ analytically. To address this issue, we consider approximating this kernel mean and study the impact of the approximation scheme on the accuracy of the BQ estimator. We focus on

approximations of the form

$$\tilde{z}_i := \sum_{j=1}^{m} \gamma_j k(\boldsymbol{y}_j, \boldsymbol{x}_i), \tag{93}$$

$$\tilde{\pi}_{\boldsymbol{\gamma}}(\boldsymbol{x}) := \sum_{j=1}^{m} \gamma_j \delta(\boldsymbol{x} - \boldsymbol{y}_j), \tag{94}$$

that are based on auxiliary states $\{\boldsymbol{y}_j\}_{j=1}^m$ and auxiliary weights $\{\gamma_j\}_{j=1}^m$. Substituting the approximation $\tilde{\boldsymbol{z}} = (\tilde{z}_1, \ldots, \tilde{z}_n)^T$ in place of $\boldsymbol{z}$ produces approximate BQ weights

$$\tilde{\boldsymbol{w}} := \boldsymbol{K}^{-1}\tilde{\boldsymbol{z}}, \tag{95}$$

$$\tilde{\pi}(\boldsymbol{x}) := \sum_{j=1}^{m} \tilde{w}_j \delta(\boldsymbol{x} - \boldsymbol{y}_j). \tag{96}$$

Note that the the kernel matrix $\boldsymbol{K}$ has size n, independent of m. The cost of computing the approximate weights is now $O(nm) + O(n^3)$, instead of the usual $O(n^3)$, so m can be taken as large as $O(n^2)$ without increasing overall computational complexity.

Denote the approximate BQ estimator $\tilde{\Pi}_{\mathrm{BQ}}$ (i.e. the quadrature rule based on states $\{\boldsymbol{x}_i\}_{i=1}^n$ and approximate BQ weights $\tilde{\boldsymbol{w}}_{\mathrm{BQ}}$). The effect of this approximation can be understood as follows:

Proposition 8. $\|\tilde{\Pi}_{BQ} - \Pi\|_{op}^2 \leq \|\hat{\Pi}_{BQ} - \Pi\|_{op}^2 + \sqrt{n}\|\mu_{\pi_{\boldsymbol{\gamma}}} - \mu_\pi\|_{\mathcal{H}}^2$.

For the error term $\|\mu_{\pi_{\boldsymbol{\gamma}}} - \mu_\pi\|_{\mathcal{H}}$, Prop. 5 shows that an approximation based on MC with states $\{\boldsymbol{y}_j\}_{j=1}^m$ and uniform weights $\gamma_j = 1/m$ provides a convergence rate of $O_P(m^{-1/2})$. To preserve the overall BQ convergence rate of δ_n, in this case, we must therefore take $m = O(n^{1/2}\delta_n^{-2})$. On the other hand, to avoid increasing computational complexity relative to the exact BQ case, we must take $m = O(n^2)$. Combining these rates shows that (e.g.) *if* the exact BQ estimator converges at $\delta_n = O(n^{-3/4})$, *then* taking $m = O(n^2)$ produces an overall rate $\|\tilde{\Pi}_{\mathrm{BQ}} - \Pi\|_{\mathrm{op}} = O_P(n^{-3/4})$, which is faster than the rate achieved by MC estimation *despite* intractability of the kernel mean. This demonstrates that the efficient estimation of integrals provided by the BQ point estimator may be applicable beyond the class of kernel-density pairs that lead to closed-form kernel means.

Proof of Prop. 8. Let $\tilde{\boldsymbol{z}} = \boldsymbol{z} + \boldsymbol{\epsilon}$. Then

$$\begin{aligned} \|\tilde{\Pi}_{\mathrm{BQ}} - \Pi\|_{\mathrm{op}}^2 &= \tilde{\boldsymbol{w}}_{\mathrm{BQ}}^T \boldsymbol{K} \tilde{\boldsymbol{w}}_{\mathrm{BQ}} - 2\tilde{\boldsymbol{w}}_{\mathrm{BQ}}^T \boldsymbol{z} + \Pi[\mu_\pi] \\ &= \|\hat{\Pi}_{\mathrm{BQ}} - \Pi\|_{\mathrm{op}}^2 + \boldsymbol{\epsilon}^T \boldsymbol{K}^{-1} \boldsymbol{\epsilon}. \end{aligned}$$

We use $\otimes$ to denote the tensor product of RKHS. Now, since $\epsilon_i = \tilde{z}_i - z_i = \mu_{\pi_{\boldsymbol{\gamma}}}(\boldsymbol{x}_i) - \mu_\pi(\boldsymbol{x}_i) = \langle \mu_{\pi_{\boldsymbol{\gamma}}} - \mu_\pi, k(\cdot, \boldsymbol{x}_i)\rangle_{\mathcal{H}}$, we have

$$\begin{aligned} \boldsymbol{\epsilon}^T \boldsymbol{K}^{-1} \boldsymbol{\epsilon} &= \sum_{i,i'} [\boldsymbol{K}^{-1}]_{i,i'} \langle \mu_{\pi_{\boldsymbol{\gamma}}} - \mu_\pi, k(\cdot, \boldsymbol{x}_i)\rangle_{\mathcal{H}} \langle \mu_{\pi_{\boldsymbol{\gamma}}} - \mu_\pi, k(\cdot, \boldsymbol{x}_{i'})\rangle_{\mathcal{H}} \\ &= \left\langle (\mu_{\pi_{\boldsymbol{\gamma}}} - \mu_\pi) \otimes (\mu_{\pi_{\boldsymbol{\gamma}}} - \mu_\pi), \sum_{i,i'} [\boldsymbol{K}^{-1}]_{i,i'} k(\cdot, \boldsymbol{x}_i) \otimes k(\cdot, \boldsymbol{x}_{i'}) \right\rangle_{\mathcal{H}\otimes\mathcal{H}} \\ &\leq \|\mu_{\pi_{\boldsymbol{\gamma}}} - \mu_\pi\|_{\mathcal{H}}^2 \left\| \sum_{i,i'} [\boldsymbol{K}^{-1}]_{i,i'} k(\cdot, \boldsymbol{x}_i) \otimes k(\cdot, \boldsymbol{x}_{i'}) \right\|_{\mathcal{H}\otimes\mathcal{H}}. \end{aligned}$$

It remains to show that the second term is equal to $\sqrt{n}$. Indeed,

$$\left\|\sum_{i,i'}[\boldsymbol{K}^{-1}]_{i,i'}k(\cdot,\boldsymbol{x}_i)\otimes k(\cdot,\boldsymbol{x}_{i'})\right\|_{\mathcal{H}}^2 = \sum_{i,i',l,l'}[\boldsymbol{K}^{-1}]_{i,i'}[\boldsymbol{K}^{-1}]_{l,l'}\langle k(\cdot,\boldsymbol{x}_i)\otimes k(\cdot,\boldsymbol{x}_{i'}), k(\cdot,\boldsymbol{x}_l)\otimes k(\cdot,\boldsymbol{x}_{l'})\rangle_{\mathcal{H}}$$
$$= \sum_{i,i',l,l'}[\boldsymbol{K}^{-1}]_{i,i'}[\boldsymbol{K}^{-1}]_{l,l'}[\boldsymbol{K}]_{il}[\boldsymbol{K}]_{i',l'} = \mathrm{tr}[\boldsymbol{K}\boldsymbol{K}^{-1}\boldsymbol{K}\boldsymbol{K}^{-1}] = n.$$

This completes the proof. □

J Computing the Kernel Mean

Below we provide formulae for the mean element in the case of Gaussian and Matérn kernels and particular choices of distribution Π.

J.1 Gaussian Kernel

Here we provide computational details for tensor products of the Gaussian kernel $k(x,y) := \lambda^2\exp(-\tau(x-y)^2)$. From the tensor construction we can wlog consider a one-dimensional state space $\mathcal{X}=[a,b]$. Let Π be uniform over $\mathcal{X}$. We derive below the mean element μ_π as well as initial error $\Pi[\mu_\pi]$ for this particular case:

$$\mu_\pi(x) = \int_a^b k(x,y)\pi(y)\mathrm{d}y = \frac{\lambda^2}{b-a}\int_a^b \exp(-\tau(x-y)^2)\mathrm{d}y \tag{97}$$
$$= \frac{\sqrt{\pi}\lambda^2}{2\sqrt{\tau}(b-a)}\Big[\mathrm{erf}\,(\sqrt{\tau}(x-a)) - \mathrm{erf}\,(\sqrt{\tau}(x-b))\Big] \tag{98}$$

The initial error is given by:

$$\Pi[\mu_\pi] = \int_a^b \mu_\pi(x)\pi(x)\mathrm{d}x \tag{99}$$
$$= \frac{\sqrt{\pi}\lambda^2}{2\sqrt{\tau}(b-a)^2}\int_a^b \Big[\mathrm{erf}(\sqrt{\tau}(x-a)) - \mathrm{erf}(\sqrt{\tau}(x-b))\Big]\mathrm{d}x \tag{100}$$
$$= \frac{\lambda^2}{\sqrt{2\tau}(b-a)^2}\exp(-\tau(b-a))\times \tag{101}$$
$$\Big[\frac{\sqrt{2}}{\sqrt{\tau}}\sigma - \exp(\tau(b-a)^2)\Big(\frac{\sqrt{2}}{\sqrt{\tau}}\sigma + (b-a)\sqrt{2\pi}\,\mathrm{erf}(\sqrt{\tau}(a-b))\Big)\Big] \tag{102}$$

These expressions can easily be generalised to multiple dimensions by taking tensor products.

J.2 Matérn Kernel

Here we provide computational details for the Matérn kernel on bounded intervals $\mathcal{X}=[a,b]\subset\mathbb{R}$. The Matérn kernels are translation-invariant and so can be written in the form $k(x,y) := \phi(r)$, where $r=|x-y|$. In general, the Matérn kernel is defined as:

$$k_\beta(x,y) = \phi_\beta(r) := \frac{2^{1-\beta}}{\Gamma(\beta)}\Big(\frac{\sqrt{2\beta}r}{\tau}\Big)^\beta K_\beta\Big(\frac{\sqrt{2\beta}r}{\tau}\Big). \tag{103}$$

This expression may be quite complex to compute in general, but simplifies significantly when β is a half integer (i.e. $\beta = p + 1/2$ and $p \in \mathbb{N}$):

$$\phi_{p+1/2}(r) := \exp\Big(-\frac{\sqrt{2\beta}r}{\tau}\Big)\frac{\Gamma(p+1)}{\Gamma(2p+1)}\sum_{i=0}^{p}\frac{(p+i)!}{i!(p-i)!}\Big(\frac{\sqrt{8\beta}r}{\tau}\Big)^{p-i}. \tag{104}$$

We will consider four cases when $\beta \in \{\frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \frac{7}{2}\}$:

$$\begin{aligned}
\phi_{1/2}(r) &:= \lambda^2 \exp\Big(-\frac{r}{\tau}\Big), && (105)\\
\phi_{3/2}(r) &:= \lambda^2\Big(1+\frac{\sqrt{3}r}{\tau}\Big)\exp\Big(-\frac{\sqrt{3}r}{\tau}\Big) && (106)\\
\phi_{5/2}(r) &:= \lambda^2\Big(1+\frac{\sqrt{5}r}{\tau}+\frac{5r^2}{3\tau^2}\Big)\exp\Big(-\frac{\sqrt{5}r}{\tau}\Big), && (107)\\
\phi_{7/2}(r) &:= \lambda^2\Big(1+\frac{\sqrt{7}r}{\tau}+\frac{14r^2}{5\tau^2}+\frac{7^{3/2}r^3}{15\tau^3}\Big)\exp\Big(-\frac{\sqrt{7}r}{\tau}\Big). && (108)
\end{aligned}$$

For Π uniform over $\mathcal{X}$, the kernel mean is given by:

$$\begin{aligned}
\mu_{\pi,\beta}(x) &= \int_a^b \phi_\beta(|x-y|)\pi(y)\mathrm{d}y \\
&= \frac{1}{b-a}\Big[\int_a^x \phi_\beta(x-y)\mathrm{d}y + \int_x^b \phi_\beta(y-x)\mathrm{d}y\Big],
\end{aligned} \tag{109}$$

which correspond to the following expressions for the kernel means when $\beta \in \{\frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \frac{7}{2}\}$:

$$\begin{aligned}
\mu_{\pi,1/2}(x) &= \lambda^2\tau \times \frac{\Big(2-\exp\big(\frac{a-x}{\tau}\big)-\exp\big(\frac{x-b}{\tau}\big)\Big)}{b-a}, && (110)\\
\mu_{\pi,3/2}(x) &= \frac{\lambda^2}{b-a}\times\Big[\frac{4\tau}{\sqrt{3}}-\frac{1}{3}\exp\Big(\frac{\sqrt{3}(x-b)}{\tau}\Big)\times\Big(3b+2\sqrt{3}\tau-3x\Big) \\
&\qquad -\frac{1}{3}\exp\Big(\frac{\sqrt{3}(a-x)}{\tau}\Big)\times\Big(-3a+2\sqrt{3}\tau+3x\Big)\Big], && (111)\\
\mu_{\pi,5/2}(x) &= \frac{\lambda^2}{(b-a)}\times\Big[\frac{16\tau}{3\sqrt{5}}-\frac{1}{15\tau}\exp\Big(\frac{\sqrt{5}(x-b)}{\tau}\Big) \\
&\qquad \times\big(8\sqrt{5}\tau^2+25\tau(b-x)+5\sqrt{5}(b-x)^2\big) \\
&\qquad -\frac{1}{15\tau}\exp\Big(\frac{\sqrt{5}(a-x)}{\tau}\Big) \\
&\qquad \times\big(8\sqrt{5}\tau^2+25\tau(x-a)+5\sqrt{5}(a-x)^2\big)\Big], && (112)
\end{aligned}$$

$$
\begin{aligned}
\mu_{\pi,7/2}(x) \;=\;& \frac{\lambda^2}{105\tau^2(b-a)}\Big[96\sqrt{7}\tau^3 \\
&- \exp\Big(\frac{\sqrt{7}(x-b)}{\tau}\Big)\Big(48\sqrt{7}\tau^3 - 231\tau^2(x-b) \\
&\qquad +63\sqrt{7}\tau(x-b)^2 - 49(x-b)^3\Big) \\
&- \exp\Big(\frac{\sqrt{7}(a-x)}{\tau}\Big)\Big(48\sqrt{7}\tau^3 + 231\tau^2(x-a) \\
&\qquad +63\sqrt{7}\tau(x-a)^2 + 49(x-a)^3\Big)\Big]. \qquad (113)
\end{aligned}
$$

The initial errors $\Pi[\mu_{\pi,\beta}]$ are

$$
\begin{aligned}
\Pi[\mu_{\pi,1/2}] &= 2\lambda^2\tau \times \frac{\Big((b-a)+\tau\big(\exp\big(\frac{a-b}{\tau}\big)-1\big)\Big)}{(b-a)^2}, \qquad (114)\\
\Pi[\mu_{\pi,3/2}] &= \frac{2\lambda^2\tau}{3(b-a)^2}\Big[2\sqrt{3}(b-a) - 3\tau \\
&\qquad +3\exp\Big(\frac{\sqrt{3}(a-b)}{\tau}\Big) \times \Big(\sqrt{3}(b-a)+3\tau\Big)\Big], \qquad (115)
\end{aligned}
$$

$$
\begin{aligned}
\Pi[\mu_{\pi,5/2}] &= \frac{\lambda^2}{15(b-a)^2}\Big[2\tau\big(8\sqrt{5}(b-a)-15\tau\big) + 2\exp\Big(\frac{\sqrt{5}(a-b)}{\tau}\Big) \\
&\quad \times\big(5a^2+10ab+5b^2+7\sqrt{5}(b-a)\tau+15\tau^2\big)\Big], \qquad (116)\\
\Pi[\mu_{\pi,7/2}] &= \frac{\lambda^2}{105\tau(b-a)^2}\Big[-6\tau^2(16\sqrt{7}(a-b)+35\tau) + 2\exp\Big(\frac{\sqrt{7}(a-b)}{\tau}\Big) \\
&\quad \times\Big(7\sqrt{7}(b^3-a^3) + 84b^2\tau + 57\sqrt{7}b\tau^2 + 105\tau^3 \\
&\quad +21a^2(\sqrt{7}b+4\tau) - 3a(7\sqrt{7}b^2+56b\tau+19\sqrt{7}\tau^2)\Big)\Big]. \qquad (117)
\end{aligned}
$$

Again, these expressions can easily be generalised to multiple dimensions by taking tensor products.