

Probabilistic Models for Integration Error in the Assessment of Functional Cardiac Models

Chris. J. Oates^{1,5}, Steven Niederer², Angela Lee²,
François-Xavier Briol^{3,4}, Mark Girolami^{4,5}

¹Newcastle University, ²King's College London, ³University of Warwick,

⁴Imperial College London, ⁵Alan Turing Institute

`chris.oates@ncl.ac.uk`

June 16, 2017

Abstract

This paper studies the numerical computation of integrals, representing estimates or predictions, over the output $f(x)$ of a computational model with respect to a distribution $p(dx)$ over uncertain inputs x to the model. For the functional cardiac models that motivate this work, neither f nor p possess a closed-form expression and evaluation of either requires ≈ 100 CPU hours, precluding standard numerical integration methods. Our proposal is to treat integration as an estimation problem, with a joint model for both the a priori unknown function f and the a priori unknown distribution p . The result is a posterior distribution over the integral that explicitly accounts for dual sources of numerical approximation error due to a severely limited computational budget. This construction is applied to account, in a statistically principled manner, for the impact of numerical errors that (at present) are confounding factors in functional cardiac model assessment.

1 Motivation: Predictive Assessment of Computer Models

This paper considers the problem of simulation-based assessment for computer models in general [9], motivated by an urgent need to assess the performance of sophisticated functional cardiac models [32]. In concrete terms, the problem that we consider can be expressed as the numerical approximation of integrals

$$p(f) = \int f(x)p(dx), \quad (1)$$

where $f(x)$ denotes a functional of the output from a computer model and x denotes unknown inputs (or ‘parameters’) of the model. The term $p(x)$ denotes a posterior distribution over model inputs. Although not our focus in this paper, we note that $p(x)$ is defined based on a prior $\pi_0(x)$ over these inputs and training data y assumed to follow the computer model $\pi(y|x)$ itself. The integral $p(f)$, in our context, represents a posterior prediction of actual cardiac behaviour. The computational model can be assessed through comparison of these predictions to test data generated from a real-world experiment.

The challenging nature of cardiac models – and indeed computer models in general – is such that a closed-form for both $f(x)$ and $p(dx)$ is precluded [30]. Instead, it is typical to be provided with a finite collection of samples $\{x_i\}_{i=1}^n$ obtained from $p(dx)$ through Monte Carlo (or related) methods [41]. The integrand $f(x)$ is then evaluated at these n input configurations, to obtain $\{f(x_i)\}_{i=1}^n$. Limited computational budgets necessitate that the number n is small and, in such situations, the error of an estimator for the integral $p(f)$ based on the data $\{(x_i, f(x_i))\}_{i=1}^n$ is subject to strict information-theoretic lower bounds [34]. The practical consequence is that an unknown (non-negligible) numerical error is introduced in the numerical approximation of $p(f)$, unrelated to the performance of the model. If this numerical error is ignored, it will constitute a confounding factor in the assessment of predictive performance for the computer model. It is therefore unclear how a fair model assessment can proceed. This motivates an attempt to understand the extent of numerical error in any estimate of $p(f)$. This is non-trivial; for example, the error distribution of the arithmetic mean $\frac{1}{n}\sum_{i=1}^n f(x_i)$ depends on the unknown f and p , and attempts to estimate this distribution solely from data, e.g. via a bootstrap or a central limit approximation, *cannot succeed* in general when the number of samples n is small, as argued in [36].

Our first contribution, in this paper, is to argue that approximation of $p(f)$ from samples $\{x_i\}_{i=1}^n$ and function evaluations $\{f(x_i)\}_{i=1}^n$ can be cast as an estimation task. Our second contribution is to derive a posterior distribution over the unknown value $p(f)$ of the integral. This distribution provides an interpretable quantification of the extent of numerical integration error that can be reasoned with and propagated through subsequent model assessment. Our third contribution is to establish theoretical properties of the proposed method. The method we present falls within the framework of *Probabilistic Numerics* and our work can be seen as a contribution to this emerging area [23, 6]. In particular, the method proposed is reminiscent of *Bayesian Quadrature* (BQ) [11, 37, 17, 38, 22]. In BQ, a Gaussian prior measure is placed on the unknown function f and is updated to a posterior when conditioned on the information $\{(x_i, f(x_i))\}_{i=1}^n$. This induces both a prior and a posterior over the value of $p(f)$ as push-forward measures under the projection operator $f \mapsto p(f)$. Since its introduction, several authors have related BQ to other methods such as the ‘herding’ approach from machine learning [24, 4], random feature approximations used in kernel methods [1], classical quadrature rules [42] and Quasi Monte Carlo (QMC) methods [5]. Most recently, [28] extended theoretical results for BQ to misspecified prior models, and [29] who provided efficient matrix algebraic methods for the implementation of BQ. However, as an important point of distinction, notice that BQ pre-supposes $p(dx)$ is known in closed-form - it does not apply in situations where $p(dx)$ is instead sampled. In this latter case $p(dx)$ will be called

an *intractable* distribution and, for model assessment, this scenario is typical.

To extend BQ to intractable distributions, this paper proposes to use a Dirichlet process mixture prior to estimate the unknown distribution $p(dx)$ from Monte Carlo samples $\{x_i\}_{i=1}^n$ [15]. It will be demonstrated that this leads to a simple expression for the closed-form terms which are required to implement the usual BQ. The overall method, called *Dirichlet process mixture Bayesian quadrature* (DPMBQ), constructs a (univariate) distribution over the unknown integral $p(f)$ that can be exploited to tease apart the intrinsic performance of a model from numerical integration error in model assessment. Note that BQ was used to estimate marginal likelihood in e.g. [39]. The present problem is distinct, in that we focus on predictive performance (of posterior expectations) rather than marginal likelihood, and its solution demands a correspondingly different methodological development.

On the computational front, DPMBQ demands a computational cost of $O(n^3)$. However, this cost is de-coupled from the often orders-of-magnitude larger costs involved in both evaluation of $f(x)$ and $p(dx)$, which form the main computational bottleneck. Indeed, in the modern computational cardiac models that motivate this research, the ≈ 100 CPU hour time required for a single simulation limits the number n of available samples to $\approx 10^3$ [32]. At this scale, numerical integration error cannot be neglected in model assessment. This raises challenges when making assessments or comparisons between models, since the intrinsic performance of models cannot be separated from numerical error that is introduced into the assessment. Moreover, there is an urgent ethical imperative that the clinical translation of such models is accompanied with a detailed quantification of the unknown numerical error component in model assessment. Our contribution explicitly demonstrates how this might be achieved.

The remainder of the paper proceeds as follows: In Section 2.1 we first recall the usual BQ method, then in Section 2.2 we present and analyse our novel DPMBQ method. Proofs of theoretical results are contained in the electronic supplement. Empirical results are presented in Section 3 and the paper concludes with a discussion in Section 4.

2 Probabilistic Models for Numerical Integration Error

Consider a domain $\Omega \subseteq \mathbb{R}^d$, together with a distribution $p(dx)$ on Ω . As in Eqn. 1, $p(f)$ will be used to denote the integral of the argument f with respect to the distribution $p(dx)$. All integrands are assumed to be (measurable) functions $f : \Omega \rightarrow \mathbb{R}$ such that the integral $p(f)$ is well-defined. To begin, we recall details for the BQ method when $p(dx)$ is known in closed-form [11, 37]:

2.1 Probabilistic Integration for Tractable Distributions (BQ)

In standard BQ [11, 37], a Gaussian Process (GP) prior $f \sim \text{GP}(m, k)$ is assigned to the integrand f , with mean function $m : \Omega \rightarrow \mathbb{R}$ and covariance function $k : \Omega \times \Omega \rightarrow \mathbb{R}$ [see 40,

for further details on GPs]. The implied prior over the integral $p(f)$ is then the push-forward of the GP prior through the projection $f \mapsto p(f)$:

$$p(f) \sim \text{N}(p(m), p \otimes p(k))$$

where $p \otimes p : \Omega \times \Omega \rightarrow \mathbb{R}$ is the measure formed by independent products of $p(dx)$ and $p(dx')$, so that under our notational convention the so-called *initial error* $p \otimes p(k)$ is equal to $\iint k(x, x')p(dx)p(dx')$. Next, the GP is conditioned on the information in $\{(x_i, f(x_i))\}_{i=1}^n$. The conditional GP takes a conjugate form $f|X, f(X) \sim \text{GP}(m_n, k_n)$, where we have written $X = (x_1, \dots, x_n)$, $f(X) = (f(x_1), \dots, f(x_n))^\top$. Formulae for the mean function $m_n : \Omega \rightarrow \mathbb{R}$ and covariance function $k_n : \Omega \times \Omega \rightarrow \mathbb{R}$ are standard can be found in [40, Eqns. 2.23, 2.24]. The BQ posterior over $p(f)$ is the push forward of the GP posterior:

$$p(f) | X, f(X) \sim \text{N}(p(m_n), p \otimes p(k_n)) \quad (2)$$

Formulae for $p(m_n)$ and $p \otimes p(k_n)$ were derived in [37]:

$$p(m_n) = f(X)^\top k(X, X)^{-1} \mu(X) \quad (3)$$

$$p \otimes p(k_n) = p \otimes p(k) - \mu(X)^\top k(X, X)^{-1} \mu(X) \quad (4)$$

where $k(X, X)$ is the $n \times n$ matrix with (i, j) th entry $k(x_i, x_j)$ and $\mu(X)$ is the $n \times 1$ vector with i th entry $\mu(x_i)$ where the function μ is called the *kernel mean* or *kernel embedding* [see e.g. 45]:

$$\mu(x) = \int k(x, x')p(dx') \quad (5)$$

Computation of the kernel mean and the initial error each requires that $p(dx)$ is known in general. The posterior in Eqn. 2 was studied in [5], where rates of posterior contraction were established under further assumptions on the smoothness of the covariance function k and the smoothness of the integrand. Note that the matrix inverse of $k(X, X)$ incurs a (naive) computational cost of $O(n^3)$; however this cost is *post-hoc* and decoupled from (more expensive) computation that involves the computer model.

2.2 Probabilistic Integration for Intractable Distributions

The dependence of Eqns. 3 and 4 on both the kernel mean and the initial error means that BQ cannot be used for intractable $p(dx)$ in general. To address this we construct a second non-parametric model for the unknown $p(dx)$, presented next.

Dirichlet Process Mixture Model Consider an infinite mixture model

$$p(dx) = \int \psi(dx; \phi)P(d\phi), \quad (6)$$

where $\psi : \Omega \times \Phi \rightarrow [0, \infty)$ is such that $\psi(\cdot; \phi)$ is a distribution on Ω with parameter $\phi \in \Phi$ and P is a mixing distribution defined on Φ . In this paper, each data point x_i is modelled as an independent draw from $p(dx)$ and is associated with a latent variable $\phi_i \in \Phi$ according to the generative process of Eqn. 6. i.e. $x_i \sim \psi(\cdot; \phi_i)$. To limit scope, the extension to correlated x_i is reserved for future work.

The Dirichlet process (DP) is the natural conjugate prior for non-parametric discrete distributions [15]. Here we endow $P(d\phi)$ with a DP prior $P \sim \text{DP}(\alpha, P_b)$, where $\alpha > 0$ is a concentration parameter and $P_b(d\phi)$ is a base distribution over Φ . The base distribution P_b coincides with the prior expectation $\mathbb{E}[P(d\phi)] = P_b(d\phi)$, while α determines the spread of the prior about P_b . The DP is characterised by the property that, for any finite partition $\Phi = \Phi_1 \cup \dots \cup \Phi_m$, it holds that $(P(\Phi_1), \dots, P(\Phi_m)) \sim \text{Dir}(\alpha P_b(\Phi_1), \dots, \alpha P_b(\Phi_m))$ where $P(S)$ denotes the measure of the set $S \subseteq \Phi$. For $\alpha \rightarrow 0$, the DP is supported on the set of atomic distributions, while for $\alpha \rightarrow \infty$, the DP converges to an atom on the base distribution. This overall approach is called a DP *mixture* (DPM) model [16].

For a random variable Z , the notation $[Z]$ will be used as shorthand to denote the density function of Z . It will be helpful to note that for $\phi_i \sim P$ independent, writing $\phi_{1:n} = (\phi_1, \dots, \phi_n)$, standard conjugate results for DPs lead to the conditional

$$P \mid \phi_{1:n} \sim \text{DP}\left(\alpha + n, \frac{\alpha}{\alpha + n} P_b + \frac{1}{\alpha + n} \sum_{i=1}^n \delta_{\phi_i}\right)$$

where $\delta_{\phi_i}(d\phi)$ is an atomic distribution centred at the location ϕ_i of the i th sample in $\phi_{1:n}$. In turn, this induces a conditional $[dp \mid \phi_{1:n}]$ for the unknown distribution $p(dx)$ through Eqn. 6.

Kernel Means via Stick Breaking The *stick breaking* characterisation can be used to draw from the conditional DP [43]. A generic draw from $[P \mid \phi_{1:n}]$ can be characterised as

$$P(d\phi) = \sum_{j=1}^{\infty} w_j \delta_{\varphi_j}(d\phi), \quad w_j = \beta_j \prod_{j'=1}^{j-1} (1 - \beta_{j'}) \quad (7)$$

where randomness enters through the φ_j and β_j as follows:

$$\varphi_j \stackrel{\text{iid}}{\sim} \frac{\alpha}{\alpha + n} P_b + \frac{1}{\alpha + n} \sum_{i=1}^n \delta_{\phi_i}, \quad \beta_j \stackrel{\text{iid}}{\sim} \text{Beta}(1, \alpha + n)$$

In practice the sum in Eqn. 7 may be truncated at a large finite number of terms, N , with negligible truncation error, since weights w_j vanish at a geometric rate [25]. The truncated DP has been shown to provide accurate approximation of integrals with respect to the original DP [26]. For a realisation $P(d\phi)$ from Eqn. 7, observe that the induced distribution $p(dx)$ over Ω is

$$p(dx) = \sum_{j=1}^{\infty} w_j \psi(dx; \varphi_j). \quad (8)$$

Thus we have an alternative characterisation of $[p|\phi_{1:n}]$.

Our key insight is that one can take ψ and k to be a conjugate pair, such that both the kernel mean $\mu(x)$ and the initial error $p \otimes p(k)$ will be available in an explicit form for the distribution in Eqn. 8 [see Table 1 in 5, for a list of conjugate pairs]. For instance, in the one-dimensional case, consider $\varphi = (\varphi_1, \varphi_2)$ and $\psi(dx; \varphi) = N(dx; \varphi_1, \varphi_2)$ for some location and scale parameters φ_1 and φ_2 . Then for the Gaussian kernel $k(x, x') = \zeta \exp(-(x - x')^2/2\lambda^2)$, the kernel mean becomes

$$\mu(x) = \sum_{j=1}^{\infty} \frac{\zeta \lambda w_j}{(\lambda^2 + \varphi_{j,2})^{1/2}} \exp\left(-\frac{(x - \varphi_{j,1})^2}{2(\lambda^2 + \varphi_{j,2})}\right) \quad (9)$$

and the initial variance can be expressed as

$$p \otimes p(k) = \sum_{j=1}^{\infty} \sum_{j'=1}^{\infty} \frac{\zeta \lambda w_j w_{j'}}{(\lambda^2 + \varphi_{j,2} + \varphi_{j',2})^{1/2}} \exp\left(-\frac{(\varphi_{j,1} - \varphi_{j',1})^2}{2(\lambda^2 + \varphi_{j,2} + \varphi_{j',2})}\right). \quad (10)$$

Similar calculations for the multi-dimensional case are straight-forward and provided in the Supplemental Information.

The Proposed Model To put this all together, let θ denote all hyper-parameters that (a) define the GP prior mean and covariance function, denoted m_θ and k_θ below, and (b) define the DP prior, such as α and the base distribution P_b . It is assumed that $\theta \in \Theta$ for some specified set Θ . The marginal posterior distribution for $p(f)$ in the DPMBQ model is defined as

$$[p(f) | X, f(X)] = \iint [p(f) | X, f(X), p, \theta] [dp | X, \theta] [d\theta]. \quad (11)$$

The first term in the integral is BQ for a fixed distribution $p(dx)$. The second term represents the DPM model for the unknown $p(dx)$, while the third term $[d\theta]$ represents a hyper-prior distribution over $\theta \in \Theta$. The DPMBQ distribution in Eqn. 11 does not admit a closed-form expression. However, it is straight-forward to sample from this distribution without recourse to $f(x)$ or $p(dx)$. In particular, the second term can be accessed through the law of total probabilities:

$$[dp | X, \theta] = \int [dp | \phi_{1:n}] [\phi_{1:n} | X, \theta] d\phi_{1:n}$$

where the first term $[dp | \phi_{1:n}]$ is the stick-breaking construction and the term $[\phi_{1:n} | X, \theta]$ can be targeted with a Gibbs sampler. Full details of the procedure we used to sample from Eqn. 11, which is de-coupled from the much larger costs associated with the computer model, are provided in the Supplemental Information.

Theoretical Analysis The analysis reported below restricts attention to a fixed hyperparameter θ and a one-dimensional state-space $\Omega = \mathbb{R}$. The extension of theoretical results to multiple dimensions was beyond the scope of this paper.

Our aim in this section is to establish when DPMBQ is “consistent”. To be precise, a random distribution \mathbb{P}_n over an unknown parameter $\zeta \in \mathbb{R}$, whose true value is ζ_0 , is called *consistent* for ζ_0 at a *rate* r_n if, for all $\delta > 0$, we have $\mathbb{P}_n[(-\infty, \zeta_0 - \delta) \cup (\zeta_0 + \delta, \infty)] = O_P(r_n)$. Below we denote with f_0 and p_0 the respective true values of f and p ; our aim is to estimate $\zeta_0 = p_0(f_0)$. Denote with \mathcal{H} the reproducing kernel Hilbert space whose reproducing kernel is k and assume that the GP prior mean m is an element of \mathcal{H} . Our main theoretical result below establishes that the DPMBQ posterior distribution in Eqn. 11, which is a random object due to the n independent draws $x_i \sim p(dx)$, is consistent:

Theorem. *Let P_0 denote the true mixing distribution. Suppose that:*

1. f belongs to \mathcal{H} and k is bounded on $\Omega \times \Omega$.
2. $\psi(dx; \varphi) = N(dx; \varphi_1, \varphi_2)$.
3. P_0 has compact support $\text{supp}(P_0) \subset \mathbb{R} \times (\underline{\sigma}, \bar{\sigma})$ for some fixed $\underline{\sigma}, \bar{\sigma} \in (0, \infty)$.
4. P_b has positive, continuous density on a rectangle R , s.t. $\text{supp}(P_b) \subseteq R \subseteq \mathbb{R} \times [\underline{\sigma}, \bar{\sigma}]$.
5. $P_b(\{(\varphi_1, \varphi_2) : |\varphi_1| > t\}) \leq c \exp(-\gamma|t|^\delta)$ for some $\gamma, \delta > 0$ and $\forall t > 0$.

Then the posterior $\mathbb{P}_n = [p(f) \mid X, f_0(X)]$ is consistent for the true value $p_0(f_0)$ of the integral at the rate $n^{-1/4+\epsilon}$ where the constant $\epsilon > 0$ can be arbitrarily small.

The proof is provided in the Supplemental Information. Assumption (1) derives from results on consistent BQ [5] and can be relaxed further with the results in [28] (not discussed here), while assumptions (2-5) derive from previous work on consistent estimation with DPM priors [19]. For the case of BQ when $p(dx)$ is known and \mathcal{H} a Sobolev space of order $s > 1/2$ on $\Omega = [0, 1]$, the corresponding posterior contraction rate is $\exp(-Cn^{2s-\epsilon})$ [5, Thm. 1]. Our work, while providing only an upper bound on the convergence rate, suggests that there is an increase in the fundamental complexity of estimation for $p(dx)$ unknown compared to $p(dx)$ known. Interestingly, the $n^{-1/4+\epsilon}$ rate is slower than the classical Bernstein-von Mises rate $n^{-1/2}$ [47]. However, an out-of-hand comparison between these two quantities is not straight forward, as the former involves the interaction of two distinct non-parametric statistical models. It is known Bernstein-von Mises results can be delicate for non-parametric problems [see, for example, the counter-examples in 12]. Rather, this theoretical analysis guarantees consistent estimation in a regime that is non-standard.

3 Results

The remainder of the paper reports empirical results from application of DPMBQ to simulated data and to computational cardiac models.

3.1 Simulation Experiments

To explore the empirical performance of DPMBQ, a series of detailed simulation experiments were performed. For this purpose, a flexible test bed was constructed wherein the true distribution p_0 was a normal mixture model (able to approximate any continuous density) and the true integrand f_0 was a polynomial (able to approximate any continuous function). In this set-up it is possible to obtain closed-form expressions for all integrals $p_0(f_0)$ and these served as a gold-standard benchmark. To mimic the scenario of interest, a small number n of samples x_i were drawn from $p_0(dx)$ and the integrand values $f_0(x_i)$ were obtained. This information X , $f_0(X)$ was provided to DPMBQ and the output of DPMBQ, a distribution over $p(f)$, was compared against the actual value $p_0(f_0)$ of the integral.

For all experiments in this paper the Gaussian kernel k defined in Sec. 2.2 was used; the integrand f was normalised and the associated amplitude hyper-parameter $\zeta = 1$ fixed, whereas the length-scale hyper-parameter λ was assigned a $\text{Gam}(2, 1)$ hyper-prior. For the DPM, the concentration parameter α was assigned a $\text{Exp}(1)$ hyper-prior. These choices allowed for adaptation of DPMBQ to the smoothness of both f and p in accordance with the data presented to the method. The base distribution P_b for DPMBQ was taken to be normal inverse-gamma with hyper-parameters $\mu_0 = 0, \lambda_0 = \alpha_0 = \beta_0 = 1$, selected to facilitate a simplified Gibbs sampler. Full details of the simulation set-up and Gibbs sampler are reported in the Supplemental Information.

For comparison, we considered the default 50% confidence interval description of numerical error

$$\left(\bar{f} - t^* \frac{s}{\sqrt{n}}, \bar{f} + t^* \frac{s}{\sqrt{n}} \right) \quad (12)$$

where $\bar{f} = n^{-1} \sum_{i=1}^n f(x_i)$, $s^2 = (n-1)^{-1} \sum_{i=1}^n (f(x_i) - \bar{f})^2$ and t^* is the 50% level for a Student's t -distribution with $n-1$ degrees of freedom. It is well-known that Eqn. 12 is a poor description of numerical error when n is small [c.f. "Monte Carlo is fundamentally unsound" 36]. For example, with $n = 2$, in the extreme case where, due to chance, $f(x_1) \approx f(x_2)$, it follows that $s \approx 0$ and no numerical error is acknowledged. This fundamental problem is *resolved through the use of prior information* on the form of both f and p in the DPMBQ method. The proposed method is further distinguished from Eqn. 12 in that the distribution over numerical error is fully non-parametric, not e.g. constrained to be Student- t .

Empirical Results Coverage frequencies are shown in Fig. 1a for a specific integration task (f_0, p_0) , that was deliberately selected to be difficult for Eqn. 12 due to the rare event represented by the mass at $x = 2$. These were compared against central 50% posterior credible intervals produced under DPMBQ. These are the frequency with which the confidence/credible interval contain the true value of the integral, here estimated with 100 independent realisations for DPMBQ and 1000 for the (less computational) standard method (standard errors are shown for both). Whilst it offers correct coverage in the asymptotic limit, Eqn. 12 can be seen to be over-confident when n is small, with coverage often less than 50%. In contrast, DPMBQ accounts for the fact p is being estimated and provides conservative estimation about the extent of numerical error when n is small.

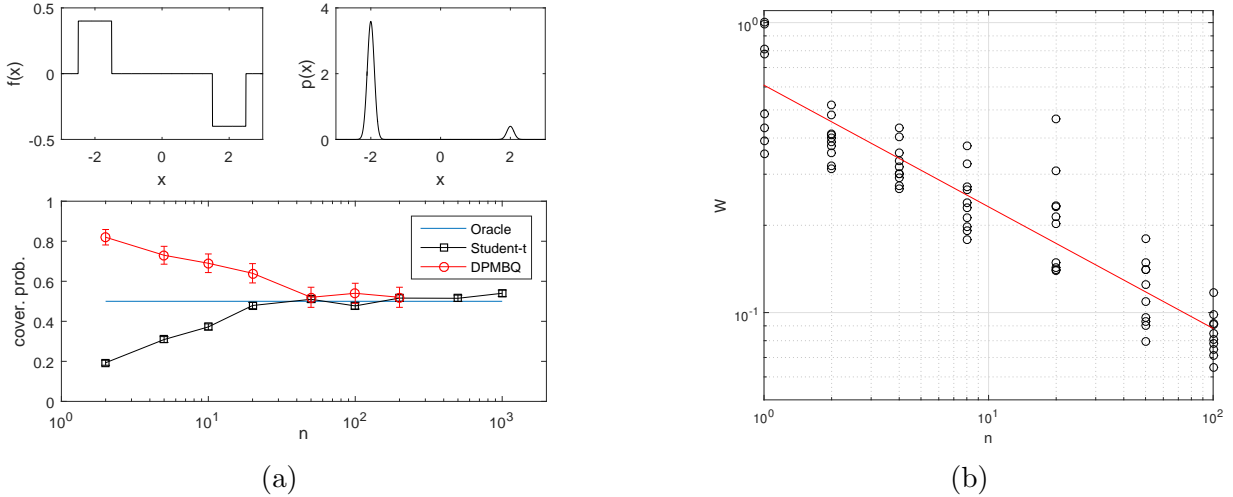


Figure 1: Simulated data results. (a) Comparison of coverage frequencies for the simulation experiments. (b) Convergence assessment: Wasserstein distance (W) between the posterior in Eqn. 11 and the true value of the integral, is presented as a function of the number n of data points. [Circles represent independent realisations and the linear trend is shown in red.]

To present results that do not depend on a fixed coverage level (e.g. 50%), we next measured convergence in the Wasserstein distance:

$$W = \int |p(f) - p_0(f_0)| d[p(f) | X, f(X)]$$

In particular we explored whether the theoretical rate of $n^{-1/4+\epsilon}$ was realised. (Note that the theoretical result applied just to fixed hyper-parameters, whereas the experimental results reported involved hyper-parameters that were marginalised, so that this is a non-trivial experiment.) Results in Fig. 1b demonstrated that W scaled with n at a rate which was consistent with the theoretical rate claimed.

Full experimental results on our polynomial test bed, reported in detail in the Supplemental Information, revealed that W was larger for higher-degree polynomials (i.e. more complex integrands f), while W was insensitive to the number of mixture components (i.e. to more complex distributions p). The latter observation may be explained by the fact that the kernel mean μ is a smoothed version of the distribution p and so is not expected to be acutely sensitive to variation in p itself.

3.2 Application to a Computational Cardiac Model

The Model The computation model considered in this paper is due to [31] and describes the mechanics of the left and right ventricles through a heart beat. In brief, the model geometry (Fig. 2a, top right) is described by fitting a C^1 continuous cubic Hermite finite

element mesh to segmented magnetic resonance images (MRI; Fig. 2a, top left). Cardiac electrophysiology is modelled separately by the solution of the mono-domain equations and provides a field of activation times across the heart. The passive material properties and afterload of the heart are described, respectively, by a transversely isotropic material law and a three element Windkessel model. Active contraction is simulated using a phenomenological cellular model, with spatial variation arising from the local electrical activation times. The active contraction model is defined by five input parameters: t_r and t_d are the respective constants for the rise and decay times, T_0 is the reference tension, a_4 and a_6 respectively govern the length dependence of tension rise time and peak tension. These five parameters were concatenated into a vector $x \in \mathbb{R}^5$ and constitute the model inputs.

The model is fitted based on training data y that consist of functionals $g_j : \mathbb{R}^5 \rightarrow \mathbb{R}$, $j = 1, \dots, 10$, of the pressure and volume transient morphology during baseline activation and when the heart is paced from two leads implanted in the right ventricle apex and the left ventricle lateral wall. These 10 functionals are defined in the Supplemental Information; a schematic of the model and fitted measurements are shown in Fig. 2a (bottom panel).

Test Functions The distribution $p(dx)$ was taken to be the posterior distribution over model inputs x that results from an improper flat prior on x and a squared-error likelihood function: $\log p(x) = \text{const.} + \frac{1}{0.1^2} \sum_{j=1}^{10} (y_j - g_j(x))^2$. The training data $y = (y_1, \dots, y_{10})$ were obtained from clinical experiment. The task we considered is to compute posterior expectations for functionals $f(x)$ of the model output produced when the model input x is distributed according to $p(dx)$. This represents the situation where a fitted model is used to predict response to a causal intervention, representing a clinical treatment.

For assessment of the DPMBQ method, which is our principle aim in this experiment, we simply took the test functions f to be each of the physically relevant model outputs g_j in turn (corresponding to no causal intervention). This defined 10 separate numerical integration problems as a test bed. Benchmark values for $p_0(g_j)$ were obtained, as described in the Supplemental Information, at a total cost of $\approx 10^5$ CPU hours, which would not be routinely practical.

Empirical Results For each of the 10 numerical integration problems in the test bed, we computed coverage probabilities, estimated with 100 independent realisations (standard errors are shown), in line with those discussed for simulation experiments. These are shown in Fig. 2b, where we compared Eqn. 12 with central 50% posterior credible intervals produced under DPMBQ. It is seen that Eqn. 12 is usually reliable but *can* sometimes be over-confident, with coverage probabilities less than 50%. This over-confidence can lead to spurious conclusions on the predictive performance of the computational model. In contrast, DPMBQ provides a uniformly conservative quantification of numerical error (cover. prob. $\geq 50\%$).

The DPMBQ method is further distinguished from Eqn. 12 in that it entails a *joint* distribution for the 10 integrals (the unknown p is shared across integrals - an instance of transfer learning across the 10 integration tasks). Fig. 2b also appears to show a correlation

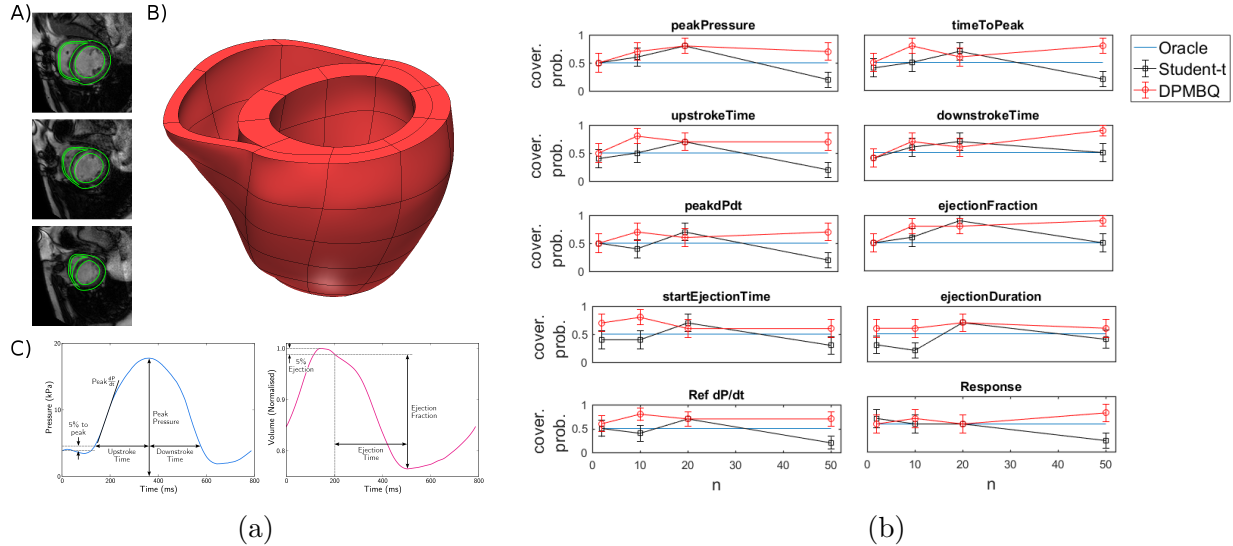


Figure 2: Cardiac model results: (a) Computational cardiac model. A) Segmentation of the cardiac MRI. B) Computational model of the left and right ventricles. C) Schematic image showing the features of pressure (left) and volume transient (right). (b) Comparison of coverage frequencies, for each of 10 numerical integration tasks defined by functionals g_j of the cardiac model output.

structure in the standard approach (black lines), but this is an artefact of the common sample set $\{x_i\}_{i=1}^n$ that was used to simultaneously estimate all 10 integrals; Eqn. 12 is still applied *independently* to each integral.

4 Discussion

Numerical analysis often focuses the convergence order of numerical methods, but in non-asymptotic regimes the language of probabilities can provide a richer, more intuitive and more useful description of numerical error. This paper cast the computation of integrals $p(f)$ as an estimation problem amenable to Bayesian methods [27, 11, 6]. The difficulty of this problem depends on our level of prior knowledge (rendering the problem trivial if a closed-form solution is *a priori* known) and, in the general case, on how much information we are prepared to obtain on the objects f and p through numerical computation [23]. In particular, we distinguish between three states of prior knowledge: (1) f known, p unknown, (2) f unknown, p known, (3) both f and p unknown. Case (1) is the subject of Monte Carlo methods [41] and concerns classical problems in applied probability such as estimating confidence intervals for expectations based on Markov chains. Notable recent work in this direction is [10], who obtained a point estimate \hat{p} for p using a kernel smoother and then, in effect, used $\hat{p}(f)$ as an estimate for the integral. The decision-theoretic risk associated with error in \hat{p} was explored in [7]. Independent of integral estimation, there is a large literature

on density estimation [48]. Our probabilistic approach provides a Bayesian solution to this problem, as a special case of our more general framework. Case (2) concerns functional analysis, where [34] provide an extensive overview of theoretical results on approximation of unknown functions in an information complexity framework. As a rule of thumb, estimation improves when additional smoothness can be *a priori* assumed on the value of the unknown object [see 5]. The main focus of this paper was Case (3), until now unstudied, and a transparent, general statistical method called DPMBQ was proposed.

The path-finding nature of this work raises several important questions for future theoretical and applied research. First, these methods should be extended to account for the low-rank phenomenon that is often encountered in multi-dimensional integrals [13]. Second, there is no reason, in general, to restrict attention to function values obtained at the locations in X . Indeed, one could first estimate $p(dx)$, then select suitable locations X' from at which to evaluate $f(X')$. This touches on aspects of statistical experimental design; the practitioner seeks a set X' that minimises an appropriate loss functional at the level of $p(f)$; see again [7]. Third, whilst restricted to Gaussians in our experiments, further methodological work will be required to establish guidance for the choice of kernel k in the GP and choice of base distribution P_b in the DPM [c.f. chapter 4 of 40].

There is an urgent ethical imperative to account for confounding due to numerical error in cardiac model assessment [32]. To address this problem, we have proposed the DPMBQ method. However, the method should be of independent interest in machine learning for computer models in general [e.g. 21].

Acknowledgments

CJO was supported by the ARC Centre of Excellence for Mathematical and Statistical Frontiers. SN was supported by an EPSRC Intermediate Career Fellowship. 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. CJO and MG were supported by the Programme on Data-Centric Engineering at the Alan Turing Institute. This material was based upon work partially supported by the National Science Foundation under Grant DMS-1127914 to the Statistical and Applied Mathematical Sciences Institute. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

A Supplemental Text

This supplement contains proofs, additional derivations and experimental results that complement the material in the Main Text.

A.1 Proof of Theorem

Denote by p_0 the true distribution that gives rise to the observations in X . Consider inference for p_0 under the DPM model for X . Let $\mu_0(x) = p_0(k(\cdot, x)) \in \mathcal{H}$ denote the exact kernel mean. Let $\|\cdot\|_{\mathcal{H}}$ and $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denote the norm and inner product associated with \mathcal{H} . An important bound is derived from Cauchy-Schwarz:

$$\left| p_0(f_0) - \sum_{i=1}^n w_i f_0(x_i) \right| \leq \|f_0\|_{\mathcal{H}} \left\| \mu_0 - \sum_{i=1}^n w_i k(\cdot, x_i) \right\|_{\mathcal{H}}$$

This motivates us to study approximation of the kernel mean μ_0 in a Hilbert space context. Let $\mu(x) = p(k(\cdot, x)) \in \mathcal{H}$ be the generic unknown kernel mean in the case where p is an uncertain distribution. The reproducing property in \mathcal{H} can be used to bound kernel mean approximation error:

$$\begin{aligned} \|\mu_0 - \mu\|_{\mathcal{H}}^2 &= \langle \mu_0 - \mu, \mu_0 - \mu \rangle_{\mathcal{H}} \\ &= \left\langle \int k(\cdot, x)(p_0(x) - p(x))dx, \int k(\cdot, x')(p_0(x') - p(x'))dx' \right\rangle_{\mathcal{H}} \\ &= \iint \langle k(\cdot, x), k(\cdot, x') \rangle_{\mathcal{H}} (p_0(x) - p(x))(p_0(x') - p(x')) \, dx dx' \\ &\leq \sup_{x, x' \in \Omega} |k(x, x')| \times \|p_0 - p\|_1^2 \\ &\leq 4 \sup_{x, x' \in \Omega} |k(x, x')| \times d_{\text{Hell}}(p_0, p)^2. \end{aligned}$$

The DPM model provides a posterior distribution over $p(dx)$; in turn this implies a posterior distribution over the kernel mean $\mu(x)$. Denote the Hellinger distance $d_{\text{Hell}}(p_0, p)$ and recall that, for two densities p_0, p , we have $\|p_0 - p\|_1 \leq 2d_{\text{Hell}}(p_0, p)$. Under assumptions (A2-5) of the theorem, [19, Thm. 6.2] established that the DP location-scale mixture model satisfies $d_{\text{Hell}}(p_0, p) = O_P(n^{-1/2+\epsilon})$, where $\epsilon > 0$ denotes a generic positive constant that can be arbitrarily small. Thus, in the posterior, $\|\mu_0 - \mu\|_{\mathcal{H}}^2 = O_P(n^{-1+\epsilon})$.

Let $\mu_{0,n}(\cdot) = \mu_0(X)k(X, X)^{-1}k(X, \cdot) \in \mathcal{H}$. The idealised BQ posterior, where $p(dx)$ is known, takes the form

$$[p(f) \mid p_0, X, f_0(X)] = N(\langle f_0, \mu_{0,n} \rangle_{\mathcal{H}}, \|\mu_0 - \mu_{0,n}\|_{\mathcal{H}}^2),$$

as shown in [5]. Let $\mu_n(\cdot) = \mu(X)k(X, X)^{-1}k(X, \cdot) \in \mathcal{H}$. For the DPMBQ posterior, where $p(dx)$ is unknown, we have the conditional distribution

$$[p(f) \mid p, X, f_0(X)] = N(\langle f_0, \mu_n \rangle_{\mathcal{H}}, \|\mu - \mu_n\|_{\mathcal{H}}^2).$$

Our aim is to relate the DPMBQ posterior to the idealised BQ posterior. To this end, it is claimed that:

$$\|\mu_0 - \mu_n\|_{\mathcal{H}}^2 \leq \|\mu_0 - \mu_{0,n}\|_{\mathcal{H}}^2 + n^{1/2}\|\mu_0 - \mu\|_{\mathcal{H}}^2. \quad (13)$$

Here we have decomposed the estimation error $\mu_0 - \mu_n$ into a term $\mu_0 - \mu_{0,n}$, that represents the error of the idealised BQ method, and a term $\mu_0 - \mu$ that captures the fact that the true mean element μ_0 is unknown.

To prove the claim, we follow Lemma 2 in [5]: Write $\epsilon(X) = \mu(X) - \mu_0(X)$ and deduce that

$$\begin{aligned}
& \|\mu_0 - \mu_n\|_{\mathcal{H}}^2 \\
&= \left\| \int k(x, \cdot) p_0(dx) - \mu(X)^\top k(X, X)^{-1} k(X, \cdot) \right\|_{\mathcal{H}}^2 \\
&= p_0 \otimes p_0(k) - 2\mu(X)^\top k(X, X)^{-1} \mu_0(X) + \mu(X)^\top k(X, X)^{-1} \mu(X) \\
&= p_0 \otimes p_0(k) - 2(\epsilon(X) + \mu_0(X))^\top k(X, X)^{-1} \mu_0(X) \\
&\quad + (\epsilon(X) + \mu_0(X))^\top k(X, X)^{-1} (\epsilon(X) + \mu_0(X)) \\
&= p_0 \otimes p_0(k) - 2\mu_0(X)^\top k(X, X)^{-1} \mu_0(X) + \mu_0(X)^\top k(X, X)^{-1} \mu_0(X) \\
&\quad + \epsilon(X)^\top k(X, X)^{-1} \epsilon(X) \\
&= \|\mu_0 - \mu_{0,n}\|_{\mathcal{H}}^2 + \epsilon(X)^\top k(X, X)^{-1} \epsilon(X). \tag{14}
\end{aligned}$$

Let $\mathcal{H} \otimes \mathcal{H}$ denote the tensor product of Hilbert spaces [2, Sec. 1.4.6]. Then the second term in Eqn. 14 is non-negative and can be bounded using the reproducing properties of both \mathcal{H} and $\mathcal{H} \otimes \mathcal{H}$:

$$\begin{aligned}
\epsilon(X)^\top k(X, X)^{-1} \epsilon(X) &= \sum_{i, i'=1}^n [k(X, X)^{-1}]_{i, i'} \langle \mu - \mu_0, k(\cdot, x_i) \rangle_{\mathcal{H}} \langle \mu - \mu_0, k(\cdot, x_{i'}) \rangle_{\mathcal{H}} \\
&= \left\langle (\mu - \mu_0) \otimes (\mu - \mu_0), \sum_{i, i'=1}^n [k(X, X)^{-1}]_{i, i'} \times k(\cdot, x_i) \otimes k(\cdot, x_{i'}) \right\rangle_{\mathcal{H} \otimes \mathcal{H}} \\
&\leq \|\mu_0 - \mu\|_{\mathcal{H}}^2 \left\| \sum_{i, i'=1}^n [k(X, X)^{-1}]_{i, i'} k(\cdot, x_i) \otimes k(\cdot, x_{i'}) \right\|_{\mathcal{H} \otimes \mathcal{H}},
\end{aligned}$$

where the final inequality is Cauchy-Schwarz. The latter factor evaluates to $n^{1/2}$, again using the reproducing property for $\mathcal{H} \otimes \mathcal{H}$:

$$\begin{aligned}
& \left\| \sum_{i=1}^n \sum_{i'=1}^n [k(X, X)^{-1}]_{i, i'} k(\cdot, x_i) \otimes k(\cdot, x_{i'}) \right\|_{\mathcal{H} \otimes \mathcal{H}}^2 \\
&= \sum_{i, i', j, j'} [k(X, X)^{-1}]_{i, i'} [k(X, X)^{-1}]_{j, j'} \\
&\quad \times \langle k(\cdot, x_i) \otimes k(\cdot, x_{i'}), k(\cdot, x_j) \otimes k(\cdot, x_{j'}) \rangle_{\mathcal{H} \otimes \mathcal{H}} \\
&= \sum_{i, i', j, j'} [k(X, X)^{-1}]_{i, i'} [k(X, X)^{-1}]_{j, j'} [k(X, X)]_{i, j} [k(X, X)]_{i', j'} \\
&= \text{tr}[k(X, X) k(X, X)^{-1} k(X, X) k(X, X)^{-1}] \\
&= n.
\end{aligned}$$

This establishes that the claim holds.

From Lemmas 1 and 3 in [5], we have that the idealised BQ estimate based on the bounded kernel k satisfies $\|\mu_0 - \mu_{0,n}\|_{\mathcal{H}} = O_P(n^{-1/2})$. Indeed, $\|\mu_0 - \mu_{0,n}\|_{\mathcal{H}} \leq \|\mu_0 - \hat{\mu}_{0,n}\|_{\mathcal{H}}$, where

$$\hat{\mu}_{0,n} = \frac{1}{n} \sum_{i=1}^n k(\cdot, x_i)$$

is the Monte Carlo estimate for the kernel mean [Lemma 3 of 5]. As k is bounded, the norm $\|\mu_0 - \hat{\mu}_{0,n}\|_{\mathcal{H}}$ vanishes as $O_P(n^{-1/2})$ [Lemma 1 of 5]. Combining the above results in Eqn. 13, we obtain

$$\begin{aligned} \|\mu_0 - \mu_n\|_{\mathcal{H}}^2 &= O_P(n^{-1}) + n^{1/2} \times O_P(n^{-1+\epsilon}) \\ &= O_P(n^{-1/2+\epsilon}). \end{aligned}$$

To finish, recall that for DPMBQ we have the random variable representation

$$p(f) = \langle f_0, \mu_n \rangle_{\mathcal{H}} + \|\mu - \mu_n\|_{\mathcal{H}} \xi,$$

where $\xi \sim N(0, 1)$ is independent of X . Thus, from the triangle inequality followed by Cauchy-Schwarz:

$$\begin{aligned} |p_0(f_0) - p(f)| &= |\langle f_0, \mu_0 \rangle_{\mathcal{H}} - \langle f_0, \mu_n \rangle_{\mathcal{H}} - \|\mu - \mu_n\|_{\mathcal{H}} \xi| \\ &\leq |\langle f_0, \mu_0 - \mu_n \rangle_{\mathcal{H}}| + \|\mu - \mu_n\|_{\mathcal{H}} |\xi| \\ &\leq |\langle f_0, \mu_0 - \mu_n \rangle_{\mathcal{H}}| + [\|\mu - \mu_0\|_{\mathcal{H}} + \|\mu_0 - \mu_n\|_{\mathcal{H}}] |\xi| \\ &\leq \|f_0\|_{\mathcal{H}} \|\mu_0 - \mu_n\|_{\mathcal{H}} + O_P(n^{-1/2+\epsilon}) + O_P(n^{-1/4+\epsilon}) \\ &= O_P(n^{-1/4+\epsilon}). \end{aligned}$$

Denote the DPMBQ posterior distribution with $\mathbb{P}_n = [p(f) | X, f(X)]$. Then for $\delta > 0$ fixed, the posterior mass $\mathbb{P}_n[(\infty, p_0(f_0) - \delta) \cup (p_0(f_0) + \delta, \infty)] = O_P(n^{-1/4+\epsilon})$. This completes the proof.

A.2 Computational Details

This section describes the computation for DPMBQ. The model admits the following straightforward sampler:

1. draw θ from the hyper-prior $[\theta]$
2. draw $\phi_{1:n}$ from $[\phi_{1:n} | X, \theta]$ (via a Gibbs sampler)
3. draw p from $[dp | \phi_{1:n}]$ (via stick-breaking)
4. draw $p(f)$ from $[p(f) | X, f(X), p, \theta]$ (via BQ)

For step (2), it is convenient (but not essential) to use a conjugate base distribution P_b . In the case of a Gaussian model ψ , the normal inverse-gamma distribution, parametrised with $\mu_0 \in \mathbb{R}$, $\lambda_0, \alpha_0, \beta_0 \in (0, \infty)$, permits closed-form conditionals and facilitates an efficient Gibbs sampler. Full details are provided in supplemental Sec. A.2.1. (Note that the conjugate base distribution does not fall within the scope of the theorem; however the use of a more general Metropolis-within-Gibbs scheme enables computation from such models with trivial modification.) In all experiments below we fixed hyper-parameters to default values $\lambda_0 = \alpha_0 = \beta_0 = 1$, $\mu_0 = 0$; there was no noticeable dependence of inferences on these choices, which are several levels removed from $p(f)$, the unknown of interest.

This direct scheme admits several improvements: e.g. (a) stratified or QMC sampling of θ in step (1); (b) Rao-Blackwellisation of the additional randomisation in $p(f)$, to collapse steps (3) and (4) [3]; (c) the Gibbs sampler of [14] can be replaced by more sophisticated alternatives, such as [33]. Indeed, one need not sample from the prior $[\theta]$ and instead target the hyper-parameter posterior with MCMC. In experiments, the straight-forward scheme outlined here was more than adequate to obtain samples from the DPMBQ model. Thus we implemented this basic sampler and leave the above extensions as possible future work.

A.2.1 Gibbs Sampler

This section derives the conditional distributions that are needed for an efficient Gibbs sampler that targets $[\phi | X, \theta]$. The main result is presented in the proposition below:

Proposition. *Consider the multivariate Gaussian model $\psi(dx; \phi) = \text{N}(dx | \phi_1, \text{diag}(\phi_2))$, with mean vector $\phi_1 \in \mathbb{R}^d$ and marginal variance vector $\phi_2 \in \mathbb{R}^d$. Consider the base distribution $P_b(d\phi)$ composed of independent normal inverse-gamma $\text{NIG}(\phi_{1,k}, \phi_{2,k} | \mu_0, \lambda_0, \alpha_0, \beta_0)$ components with $\mu_0 \in \mathbb{R}$, $\lambda_0, \alpha_0, \beta_0 \in (0, \infty)$ for $k = 1, \dots, d$. Denote $\phi_i = (\phi_{i,1}, \phi_{i,2})$ and $\phi_{(-i)} = (\phi_1, \dots, \phi_{i-1}, \phi_{i+1}, \dots, \phi_n)$. For this conjugate choice, we have the closed-form posterior conditional*

$$[\phi_i | \phi_{(-i)}, X, \theta] = \omega_0 Q_i + \sum_{j \neq i} \omega_j \delta_{\phi_j}$$

where Q_i is composed of independent $\text{NIG}(\phi_{i,1,k}, \phi_{i,2,k} | \mu_{i,k}, \lambda_{i,k}, \alpha_{i,k}, \beta_{i,k})$ components and

$$\begin{aligned} \begin{bmatrix} \omega_0 \\ \omega_j \end{bmatrix} &\propto \begin{bmatrix} \alpha \prod_{k=1}^d \frac{1}{2\pi^{1/2}} \frac{\lambda_0^{1/2}}{\lambda_{i,k}^{1/2}} \frac{\beta_0^{\alpha_0}}{\beta_{i,k}^{\alpha_{i,k}}} \frac{\Gamma(\alpha_{i,k})}{\Gamma(\alpha_0)} \\ \text{N}(x_i | \phi_{j,1}, \text{diag}(\phi_{j,2})) \end{bmatrix} \\ \mu_{i,k} &= \frac{\lambda_0 \mu_0 + x_{i,k}}{\lambda_0 + 1} \\ \lambda_{i,k} &= \lambda_0 + 1 \\ \alpha_{i,k} &= \alpha_0 + \frac{1}{2} \\ \beta_{i,k} &= \beta_0 + \frac{1}{2} (\lambda_0 \mu_0^2 + x_{i,k}^2 - \lambda_{i,k} \mu_{i,k}^2). \end{aligned}$$

Proof. From Theorem 1 of [15], also known as “Bayes’ theorem for DPs”, we have that the prior $P \sim \text{DP}(\alpha, P_b)$ and the likelihood $\phi_i \sim P$ (independent) lead to a posterior

$$P \mid \phi_{(-i)} \sim \text{DP} \left(\alpha + n - 1, \frac{1}{\alpha + n - 1} \left(\alpha P_b + \sum_{j \neq i} \delta_{\phi_j} \right) \right).$$

It follows that, for a measurable set A ,

$$\begin{aligned} \text{Prob}[\phi_i \in A \mid \phi_{(-i)}] &= \mathbb{E}[P(A) \mid \phi_{(-i)}] \\ &= \frac{1}{\alpha + n - 1} \left(\alpha P_b(A) + \sum_{j \neq i} \delta_{\phi_j}(A) \right). \end{aligned}$$

From (standard) Bayes’ theorem,

$$\begin{aligned} [\phi_i \mid \phi_{(-i)}] &= \frac{[X \mid \phi_{1:n}] [\phi_i \mid \phi_{(-i)}]}{[X \mid \phi_{(-i)}]} \\ &\propto [X \mid \phi_{1:n}] [\phi_i \mid \phi_{(-i)}] \propto [x_i \mid \phi_i] [\phi_i \mid \phi_{(-i)}] \end{aligned}$$

and combining the two above results, in the case of a Gaussian model $\psi(\text{d}x_i; \phi_i)$ with mean vector $\phi_{i,1}$ and marginal variance vector $\phi_{i,2}$, leads to

$$\begin{aligned} [\phi_i \mid \phi_{(-i)}] &\propto \text{N}(x_i \mid \phi_{i,1}, \text{diag}(\phi_{i,2})) \times \left(\alpha P_b(\phi_i) + \sum_{j \neq i} \delta_{\phi_j}(\phi_i) \right) \\ &= \alpha \text{N}(x_i \mid \phi_{i,1}, \text{diag}(\phi_{i,2})) P_b(\phi_i) + \sum_{j \neq i} \text{N}(x_i \mid \phi_{j,1}, \text{diag}(\phi_{j,2})) \delta_{\phi_j}(\phi_i), \end{aligned}$$

where $\phi_i = (\phi_{i,1}, \phi_{i,2})$ with $\phi_{i,1} \in \mathbb{R}^d$ and $\phi_{i,2} \in (0, \infty)^d$.

For closed-form expressions, P_b must be taken conjugate to the Gaussian model:

$$\begin{aligned} P_b(\phi_i) &= \prod_{k=1}^d \text{NIG}(\phi_{i,1,k}, \phi_{i,2,k} \mid \mu_0, \lambda_0, \alpha_0, \beta_0) \\ &= \prod_{k=1}^d \text{N}(\phi_{i,1,k} \mid \mu_0, \lambda_0^{-1} \phi_{i,2,k}) \text{IG}(\phi_{i,2,k} \mid \alpha_0, \beta_0), \end{aligned}$$

in the obvious notation $\phi_{i,j} = (\phi_{i,j,1}, \dots, \phi_{i,j,d})$. Thus

$$\begin{aligned} \text{N}(x_i \mid \phi_{i,1}, \text{diag}(\phi_{i,2})) P_b(\phi_i) &= \text{N}(x_i \mid \phi_{i,1}, \text{diag}(\phi_{i,2})) \\ &\quad \times \prod_{k=1}^d \text{NIG}(\phi_{i,1,k}, \phi_{i,2,k} \mid \mu_0, \lambda_0, \alpha_0, \beta_0) \\ &= \omega_0 \times \prod_{k=1}^d \text{NIG}(\phi_{i,1,k}, \phi_{i,2,k} \mid \mu_{i,k}, \lambda_{i,k}, \alpha_{i,k}, \beta_{i,k}), \end{aligned}$$

where

$$\begin{aligned}
\omega_0 &= \prod_{k=1}^d \frac{1}{2\pi^{1/2}} \frac{\lambda_0^{1/2}}{\lambda_{i,k}^{1/2}} \frac{\beta_0^{\alpha_0}}{\beta_{i,k}^{\alpha_{i,k}}} \frac{\Gamma(\alpha_{i,j})}{\Gamma(\alpha_0)} \\
\mu_{i,k} &= \frac{\lambda_0 \mu_0 + x_{i,k}}{\lambda_0 + 1} \\
\lambda_{i,k} &= \lambda_0 + 1 \\
\alpha_{i,k} &= \alpha_0 + \frac{1}{2} \\
\beta_{i,k} &= \beta_0 + \frac{1}{2}(\lambda_0 \mu_0^2 + x_{i,k}^2 - \lambda_{i,k} \mu_{i,k}^2).
\end{aligned}$$

This completes the proof. \square

In all experiments the Gibbs sampler was initialised at $\phi_{i,1,k} = x_{i,k}$ and $\phi_{i,2,k} = 1$ and run until a convergence criteria was satisfied. In this way we produced samples from $[\phi_{1:n} | X, \theta]$ for the direct sampling scheme outlined in the main text.

A.2.2 Tensor Structure for Multi-Dimensional Integrals

This section describes how multi-dimensional integration problems on a tensor-structured domain $\Omega = \Omega_1 \otimes \dots \otimes \Omega_d$ can be decomposed into a tensor product of univariate integration problems. This construction was used to produce the results in the Main Text, as well as in Sec. A.3.2 of the Supplement.

Assume a tensor product kernel

$$k(x, x') = k_1(x_1, x'_1) \times \dots \times k_d(x_d, x'_d)$$

on $\Omega \times \Omega$, together with a product model

$$\psi(\mathrm{d}x; \phi) = \psi_1(\mathrm{d}x_1; \phi_1) \times \dots \times \psi_d(\mathrm{d}x_d; \phi_d).$$

Then a generic draw from $[p | \phi_{1:n}]$ has the form

$$p(\mathrm{d}x) = \sum_{j=1}^{\infty} w_j \psi_1(\mathrm{d}x_1; \varphi_{j,1}) \times \dots \times \psi_d(\mathrm{d}x_d; \varphi_{j,d}),$$

where $\varphi_j \sim P$ are independent with $\varphi_j = (\varphi_{j,1}, \dots, \varphi_{j,d})$, and the corresponding kernel mean is

$$\mu(x) = \sum_{j=1}^{\infty} w_j \prod_{i=1}^d \left(\int_{\Omega_i} k_i(x_i, x'_i) \psi_i(x'_i; \varphi_{j,i}) \mathrm{d}x'_i \right).$$

The initial error $p \otimes p(k)$ is derived as

$$\begin{aligned}
p \otimes p(k) &= \sum_{j,j'=1}^{\infty} w_j w_{j'} \prod_{i=1}^d \\
&\int_{\Omega_i} k_i(x_i, x'_i) \psi_i(x_i; \varphi_{j,i}) \psi_i(x'_i; \varphi_{j',i}) \mathrm{d}x_i \mathrm{d}x'_i.
\end{aligned}$$

For an efficient Gibbs sampler, as in Sec. A.2.1, the prior model on the mixing distribution $P(d\phi)$ was taken as a tensor product of $\text{DP}(\alpha, P_{b,i})$ priors where $P_{b,i}(dx_i)$ is a base distribution on Ω_i . The experiments of Sec. A.3.2 were performed as explained above, where the individual components k_i , ψ_i and $P_{b,i}$ were taken to be the same as used for the simulation examples in Sec. A.3.

A.3 Experimental Set-Up and Results

Two simulation studies were undertaken, based on polynomial test functions where the true integral is known in closed-form (Sec. A.3.1) and based on differential equations where the true integral must be estimated with brute-force computation (Sec. A.3.2).

A.3.1 Flexible Polynomial Test Bed

To assess the performance of the DPMBQ method, we considered independent data x_1, \dots, x_n generated from a known distribution $p(dx)$. In addition, the function $f(x)$ was fixed and known, so that overall the exact value of the integral $p(f)$ provided a known benchmark.

For illustration, we focused on the generic class of one-dimensional test problems obtained when $p(dx)$ is a Gaussian mixture distribution

$$p(dx) = \sum_{i=1}^m r_i \text{N}(dx; c_i, s_i^2)$$

defined on $\Omega = \mathbb{R}$, where $c_i \in \mathbb{R}$, $r_i, s_i \in [0, \infty)$, $\sum_{i=1}^m r_i = 1$, and the function $f(x)$ is a polynomial

$$f(x) = \sum_{i=1}^q a_i x^{b_i}$$

where $a_i \in \mathbb{R}$ and $b_i \in \mathbb{N}_0$. For this problem class, the integral $p(f)$ is computable in closed-form and the generic approximation properties of Gaussian mixtures and polynomials provide an expressive test-bed. In addition, the GP prior with mean function $m_\theta(x) = 0$ and Gaussian covariance function

$$k_\theta(x, x') = \zeta \exp(-(x - x')^2 / 2\lambda^2)$$

was employed with $\zeta = 1$ fixed. This choice provides a closed-form kernel mean for assessment purposes, with standard Gaussian calculations analogous to those performed in the Main Text.

Illustration Consider the toy problem where $f(x) = 1 + x - 0.1x^3$, $p(dx) = \text{N}(dx; 0, 1)$, such that the true integral $p(f) = 1$ is known in closed-form. For the kernel k_θ we initially fixed the hyper-parameter λ at a default value $\lambda = 1$. The concentration hyper-parameter α was initially fixed to $\alpha = 1$ (the *unit information* DP prior). For all experiments, the stick breaking construction described in the Main Text was truncated after the first $N = 500$

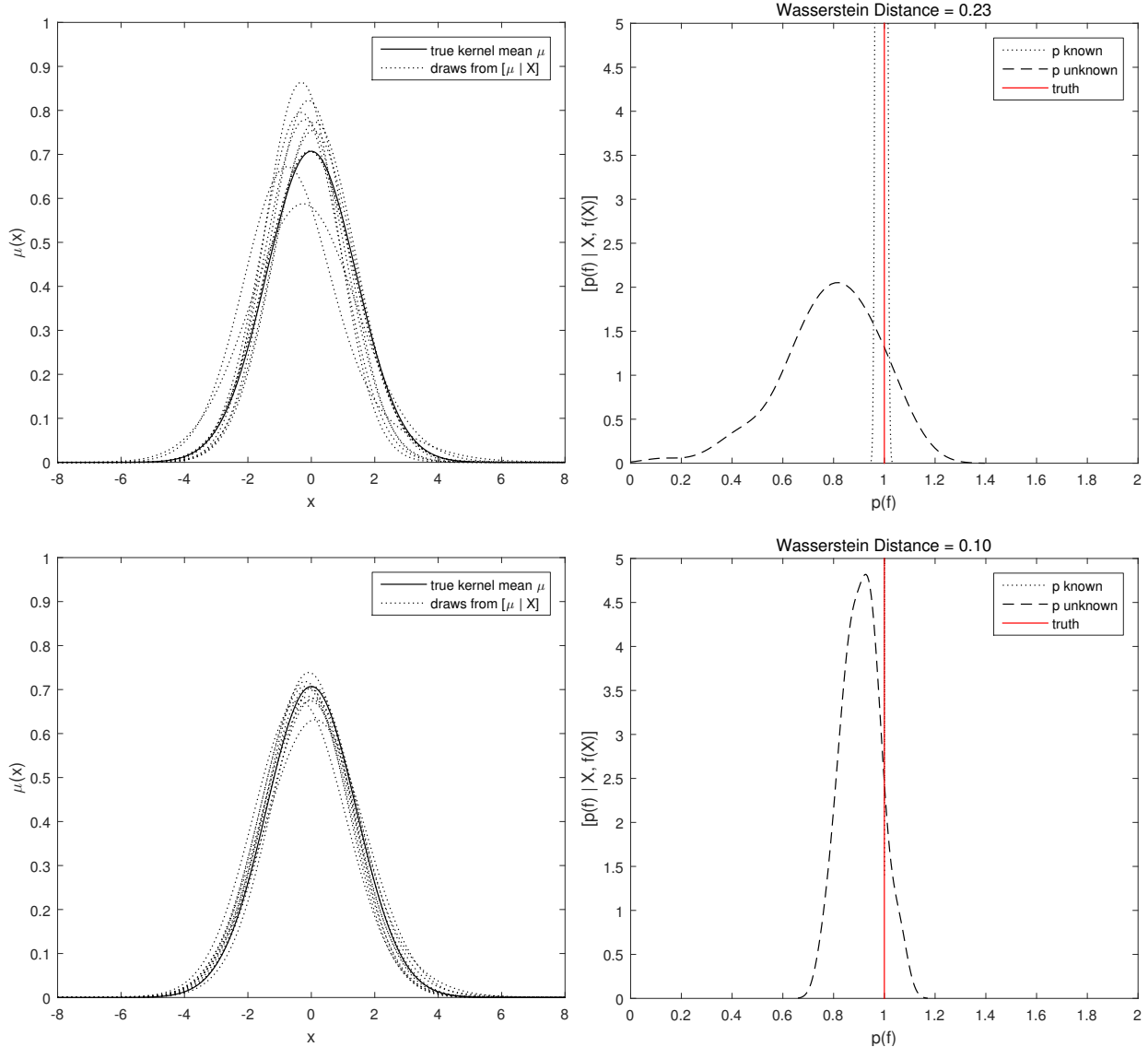


Figure 3: Illustration; computation of $p(f)$ where both $f(x)$ and $p(dx)$ are *a priori* unknown. Partial information on $p(dx)$ is provided as n draws $x_i \sim p(dx)$. Partial information on $f(x)$ is provided by the values $f(x_i)$ at each of the n locations. Left: Bayesian estimation of the kernel mean μ , that characterises the unknown distribution $p(dx)$. Right: Posterior distribution over the value of the integral $p(f)$ (dashed line); for reference, the truth (red line) and the posterior that would be obtained *if* $p(dx)$ was known (dotted line) are also shown. Two sample sizes, (top) $n = 10$, (bottom) $n = 100$, are presented.

terms; at this level results were invariant to further increases in N . In Fig. 3 we present realisations of the posterior distributions $[\mu | X]$ and $[p(f) | X, f(X)]$ at two sample sizes, (a) $n = 10$ and (b) $n = 100$. In this case each posterior contains the true value $p_0(f_0)$ of the integral in its effective support region. The posterior variance is greatly inflated with

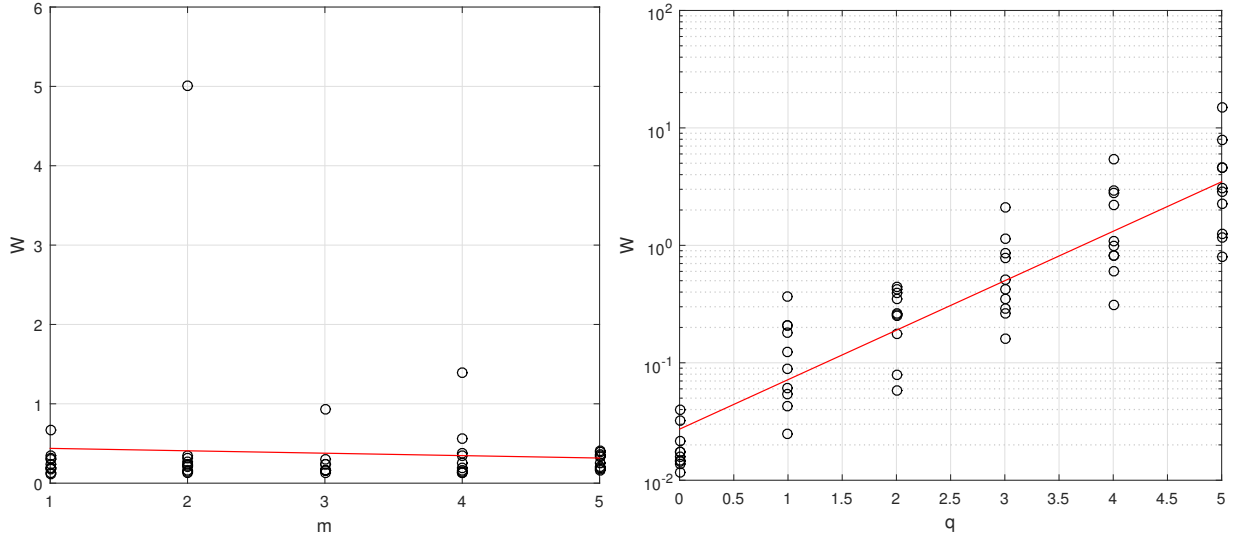


Figure 4: Empirical investigation. The Wasserstein distance, W , between the posterior $[p(f) | X, f(X)]$ and the true value of the integral is presented as a function of (left) the number m of mixture components that constitute $p(dx)$, and (right) the degree q of the polynomial function $f(x)$ whose integral is to be determined. [Circles represent independent realisations of W , while in (right) a linear trend line (red) is shown.]

respect to the idealised case in which $p(dx)$, and hence the kernel mean μ , is known. This is intuitively correct and reflects the increased difficulty of the problem in which both $f(x)$ and $p(dx)$ are *a priori* unknown.

Detailed Results To explore estimator convergence in detail, we considered the general simulation set-up above and measured estimator performance with the Wasserstein (or *earth movers'*) distance:

$$W = \int |p(f) - p_0(f_0)| d[p(f) | X, f(X)].$$

Consistent estimation, as defined in the Main Text, is implied by convergence in Wasserstein distance. It should be noted that consistent estimation does not imply correct coverage of posterior credible intervals [18]; this aspect is left for future work.

There are three main questions that we address below; these concern dependence of the approximation properties of the posterior $[p(f) | X, f(X)]$ on (i) the number n of data, (ii) the complexity of the distribution $p(dx)$, and (iii) the complexity of the function $f(x)$. Our results can be summarised as follows:

- **Effect of the number n of data:** As n increases, we expect contraction of the posterior measure over $[\mu | X]$ onto the true kernel mean. Hence, in the limit of infinite data, the resultant integral estimates will coincide with those of BQ. However, the rate of convergence of the proposed method could be much slower compared to the idealised case in which $p(dx)$, and hence $\mu(x)$, is *a priori* known.

The problem of Fig. 3 was considered in a more general setting where the hyper-parameters θ are assigned prior distributions and are subsequently marginalised out. For these results, the kernel parameter λ was assigned a $\text{Gam}(2, 1)$ hyper-prior and the concentration parameter α was assigned a $\text{Exp}(1)$ hyper-prior; these were employed for the remainder.

Results in Fig. 5 showed that the posterior $[p(f) \mid X, f(X)]$ appears to converge to the true value of the integrand (in the Wasserstein sense) as the number n of data are increased. The slope of the trend line was $\approx -1/4$, in close agreement with the theoretical analysis. This does not resemble the rapid posterior contraction results established in BQ when $p(dx)$ is *a priori* known, which can be exponential for the Gaussian kernel [5]. This reflects the more challenging nature of the estimation problem when $p(dx)$ is unavailable in closed-form.

- **Effect of the complexity of $p(dx)$:** It is anticipated that a more challenging inference problem for $p(dx)$ entails poorer estimation performance for $p(f)$. To investigate, the complexity of $p(dx)$ was measured as the number m of mixture components. For this experiment, the number m of mixture components was fixed, with weights (r_1, \dots, r_m) drawn from $\text{Dir}(2)$. The location parameters c_i were independent draws from $\text{N}(0, 1)$ and the scale parameters s_i were independent draws from $\text{Exp}(1)$.

Results in Fig. 4 (left), which were based on $n = 20$, did not demonstrate a clear effect. This was interesting and can perhaps be explained by the fact that $\mu(x)$ is a kernel-smoothed version of $p(dx)$ and thus is somewhat robust to fluctuations in $p(dx)$.

- **Effect of the complexity of $f(x)$:** A more challenging inference problem for $f(x)$ ought to also entail poorer estimation. To investigate, the complexity of $f(x)$ was measured as the degree q of this polynomial. For each experiment, q was fixed and the coefficients a_i were independent draws from $\text{N}(0, 1)$.

Results in Fig. 4 (right), based on $n = 20$, showed that the posterior was more accurate for larger q , this time in agreement with intuition.

A.3.2 Goodwin Oscillator

Our second simulation experiment considered the computation of Bayesian forecasts based on a 5-dimensional computer model.

For a manageable benchmark we took a computer model that is well-understood; the *Goodwin oscillator*, which is prototypical for larger models of complex chemical systems [20]. The oscillator considers a competitive molecular dynamic, expressed as a system of ordinary differential equations (ODEs), that induces oscillation between the concentration $z_i(t; x)$ of two species S_i ($i = 1, 2$). Parameters, denoted x and *a priori* unknown, included two synthesis rate constants, two degradation rate constants and one exponent parameter. Full details, that include the prior distributions over parameters used in the experiment below, can be found in [35]. From an experimental perspective, we suppose that concentrations of both species are

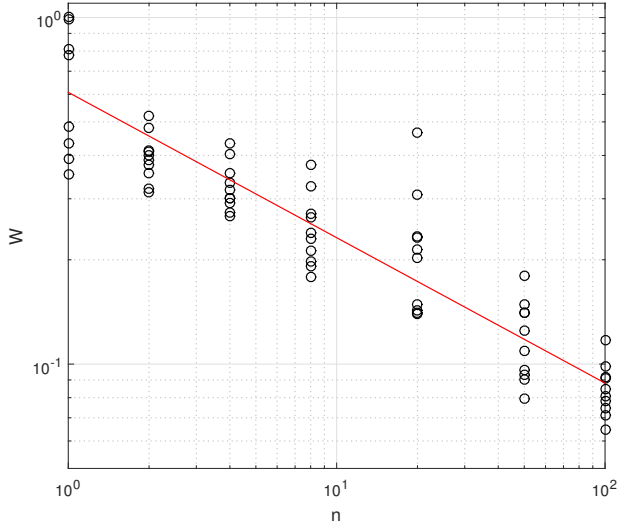


Figure 5: Empirical investigation. The example of Fig. 3 was again considered, this time marginalising over hyper-parameters λ and α . The Wasserstein distance, W , between the posterior $[p(f) | X, f(X)]$ and the true value of the integral, is presented as a function of the number n of data points. [Circles represent independent realisations, while a linear trend line (red) is shown.]

observed at 41 discrete time points t_j with uniform spacing in $[0, 40]$. Observation occurred through an independent Gaussian noise process $y_{i,j} = z_i(t_j; x) + \epsilon_{i,j}$ where $\epsilon_{i,j} \sim \mathcal{N}(0, 0.1^2)$. Data-generating parameters were identical to [35] with model dimension $g = 3$. Fig. 6 (left) shows the full data $y = (y_{i,j})$.

The forecast that we consider here is for the concentration of S_1 at the later time $t = 50$. In particular we defined $f(x)$ to be equal to $z_1(50)$ and obtained n samples x_i from the posterior $[x | y]$ using tempered population Markov chain Monte Carlo (MCMC), in all aspects identical to [35]. Then, $f(x_i)$ was evaluated and stored for each x_i ; the locations $X = (x_i)$ and function evaluations $f(X)$ are the starting point for the DPMBQ method.

This prototypical model is small enough for numerical error to be driven to zero via repeated numerical simulation of the ODEs, providing us with a benchmark. Nevertheless, the key features that motivate our work are present here: (i) The forecast function $f(x)$ is expensive and black-box, being a long-range solution of a system of ODEs and requiring that the global solution error is carefully controlled. (ii) The task of obtaining samples x_i is costly, as each evaluation of the likelihood $[y | x]$, and hence the posterior $[x | y]$, requires the solution of a system of ODEs.

Performance was examined through the Wasserstein distance to the true forecast $p_0(f_0)$, the latter obtained through brute-force simulation. The multi-dimensional integral was modelled as a tensor product of one-dimensional integrals, as described in Sec. A.2.2 in the supplement. This allowed the uni-variate model from Sec. A.3 to be re-used at minimal effort. Results, in Fig. 6 (right), indicated that the posterior was consistent. Note that the Wasserstein distances are large for this problem, reflecting the greater uncertainties that are

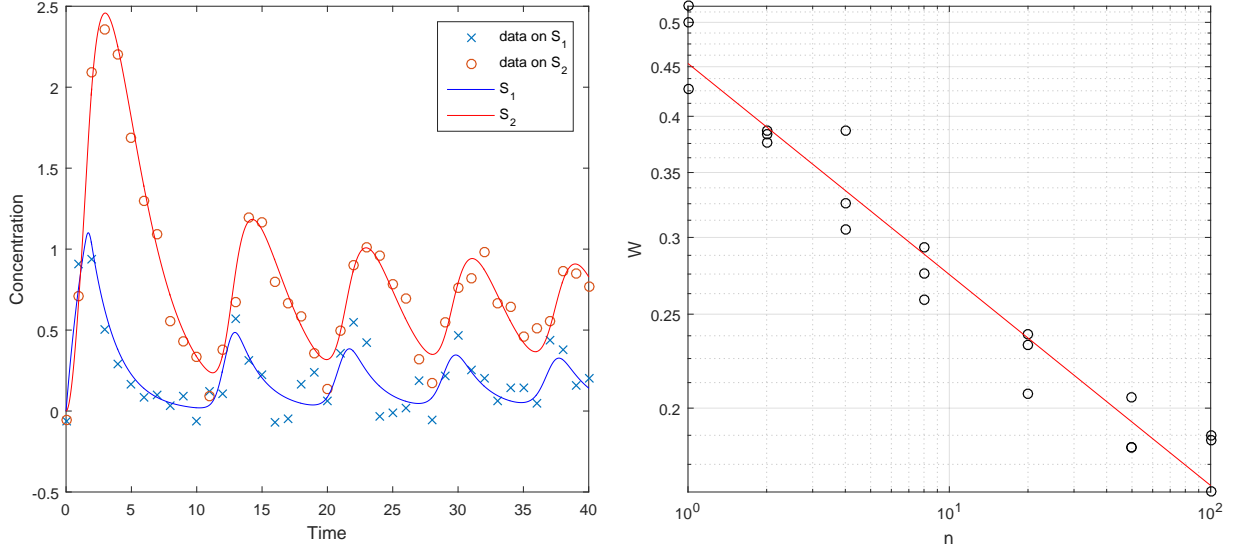


Figure 6: Application to Bayesian forecasting. Left: Data on two species, S_1 and S_2 , generated from the Goodwin oscillator, a system of differential equations that contain five unknown parameters. The forecast $p(f)$ under consideration is the posterior expected concentration of species S_1 at the later time point $t = 50$. Right: The Wasserstein distance, W , from the proposed posterior $[p(f) | X, f(X)]$ to the true integral is shown. Here n represents the number of samples x_i that were obtained from the posterior $[x | y]$ over the unknown parameters.

associated with a 5-dimensional integration problem with only $n < 10^2$ draws from $p(dx)$.

An extension of this framework, not considered here, would use a probabilistic ODE solver in tandem with DPMBQ to model the approximate nature of numerical solution to the ODEs in the reported forecasts [44, 23].

A.3.3 Cardiac Model Experiment

Test Functionals g_j Used in the Cardiac Model Experiment The 10 functionals g_j , that are the basis for clinical data on the cardiac model in the main text, are defined in the next paragraph:

The left ventricle pressure curve during baseline activation is characterised by the peak value (Peak Pressure), the time of the peak value (Time to Peak) and the time for pressure to rise (Upstroke Time) from 5% of the pressure change to the peak value and then fall back down (Down Stroke Time). The volume transient is described by the ratio of the left ventricle volume of blood ejected over the maximal left ventricle volume (Ejection Fraction), the time that the ventricle volume has decreased by 5% of the maximal volume (Start Ejection Time) and the time taken between the start of ejection and the point where the heart reaches its smallest left ventricle volume (Ejection Duration). The effect of pacing the heart is measured by the percentage change in the maximum rate of pressure development at baseline (Ref dPdt) and during pacing (Peak dPdt), defined as the acute haemodynamic

response (Response).

Brute-Force Computation for a Benchmark The samples $\{x_i\}_{i=1}^n$ from $p(dx)$ can in principle be obtained via any sophisticated Markov chain Monte Carlo (MCMC) methods, such as [46, 8]. Recall that each evaluation of $p(dx)$ requires $\approx 10^3$ hours, so that the MCMC method must be efficient. To reduce the computational overhead required for this project, we circumvented MCMC and instead exploited an existing, detailed empirical approximation to $p(dx)$ that had been pre-computed by a subset of the authors. This consisted of a collection of $m \approx 10^3$ weighted states (x_i, p_i) , where the x_i were selected via an *ad-hoc* adaptive Latin hypercube method, and such that the weights $p_i \propto p(x_i)$. Then, in this work, an (approximate) sample of size $n \ll m$ was obtained by sampling with replacement from the empirical distribution defined by this weighted point set. For our assessment of DPMBQ, benchmark values for each integral were computed as $\sum_{i=1}^m p_i f(x_i)$ for $m \approx 10^3$; note that this required a total of $\approx 10^5$ CPU hours and would not be routinely practical.

References

- [1] F Bach. On the Equivalence Between Quadrature Rules and Random Features. *arXiv:1502.06800*, 2015.
- [2] A Berline and C Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer Science & Business Media, 2011.
- [3] D Blackwell. Conditional Expectation and Unbiased Sequential Estimation. *Annals of Mathematical Statistics*, 18(1):105–110, 1947.
- [4] F-X Briol, CJ Oates, M Girolami, and MA Osborne. Frank-Wolfe Bayesian quadrature: Probabilistic integration with theoretical guarantees. In *Advances in Neural Information Processing Systems*, pages 1162–1170, 2015.
- [5] F-X Briol, CJ Oates, M Girolami, MA Osborne, and D Sejdinovic. Probabilistic Integration: A Role for Statisticians in Numerical Analysis? *arXiv:1512.00933*, 2015.
- [6] J Cockayne, CJ Oates, T Sullivan, and M Girolami. Bayesian probabilistic numerical methods. *arXiv:1702.03673*, 2017.
- [7] SN Cohen. Data-driven nonlinear expectations for statistical uncertainty in decisions. *arXiv:1609.06545*, 2016.
- [8] Patrick R Conrad, Youssef M Marzouk, Natesh S Pillai, and Aaron Smith. Accelerating asymptotically exact mcmc for computationally intensive models via local approximations. *Journal of the American Statistical Association*, 111(516):1591–1607, 2016.
- [9] PS Craig, M Goldstein, JC Rougier, and AH Seheult. Bayesian Forecasting for Complex Systems Using Computer Simulators. *Journal of the American Statistical Association*, 96(454):717–729, 2001.
- [10] B Delyon and F Portier. Integral Approximation by Kernel Smoothing. *Bernoulli*, 22(4):2177–2208, 2016.
- [11] P Diaconis. Bayesian Numerical Analysis. *Statistical Decision Theory and Related Topics IV*, 1:163–175, 1988.
- [12] P Diaconis and D Freedman. On the Consistency of Bayes Estimates. *Annals of Statistics*, 14(1):1–26, 1986.
- [13] J Dick, FY Kuo, and IH Sloan. High-Dimensional Integration: The Quasi-Monte Carlo Way. *Acta Numerica*, 22:133–288, 2013.
- [14] MD Escobar and M West. Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995.
- [15] TS Ferguson. A Bayesian Analysis of Some Nonparametric Problems. *Annals of Statistics*, 1(2):209–230, 1973.

- [16] TS Ferguson. Bayesian Density Estimation by Mixtures of Normal Distributions. *Recent Advances in Statistics*, 24(1983):287–302, 1983.
- [17] Z Ghahramani and CE Rasmussen. Bayesian Monte Carlo. In *Advances in Neural Information Processing Systems*, volume 15, pages 489–496, 2002.
- [18] S Ghosal and A Van Der Vaart. Convergence Rates of Posterior Distributions for Non-IID Observations. *Annals of Statistics*, 35(1):192–223, 2007.
- [19] S Ghosal and AW Van Der Vaart. Entropies and Rates of Convergence for Maximum Likelihood and Bayes Estimation for Mixtures of Normal Densities. *Annals of Statistics*, 29(5):1233–1263, 2001.
- [20] BC Goodwin. Oscillatory Behavior in Enzymatic Control Processes. *Advances in Enzyme Regulation*, 3:425–437, 1965.
- [21] RB Gramacy and HKH Lee. Adaptive Design and Analysis of Supercomputer Experiments. *Technometrics*, 51(2):130–145, 2009.
- [22] T Gunter, MA Osborne, R Garnett, P Hennig, and SJ Roberts. Sampling for Inference in Probabilistic Models With Fast Bayesian Quadrature. In *Advances in Neural Information Processing Systems*, pages 2789–2797, 2014.
- [23] P Hennig, MA Osborne, and M Girolami. Probabilistic Numerics and Uncertainty in Computations. *Proceedings of the Royal Society A*, 471(2179):20150142, 2015.
- [24] F Huszár and D Duvenaud. Optimally-Weighted Herding is Bayesian Quadrature. In *Uncertainty in Artificial Intelligence*, volume 28, pages 377–386, 2012.
- [25] H Ishwaran and LF James. Gibbs Sampling Methods for Stick-Breaking Priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.
- [26] H Ishwaran and M Zarepour. Exact and Approximate Sum Representations for the Dirichlet Process. *Canadian Journal of Statistics*, 30(2):269–283, 2002.
- [27] JB Kadane and GW Wasilkowski. Average case epsilon-complexity in computer science: A Bayesian view. *Bayesian Statistics 2, Proceedings of the Second Valencia International Meeting*, pages 361–374, 1985.
- [28] M Kanagawa, BK Sriperumbudur, and K Fukumizu. Convergence Guarantees for Kernel-Based Quadrature Rules in Misspecified Settings. In *Advances in Neural Information Processing Systems*, volume 30, 2016.
- [29] T Karvonen and S Särkkä. Fully symmetric kernel quadrature. *arXiv:1703.06359*, 2017.
- [30] MC Kennedy and A O’Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B*, 63(3):425–464, 2001.
- [31] AWC Lee, A Crozier, ER Hyde, P Lamata, M Truong, M Sohal, T Jackson, JM Behar, S Claridge, A Shetty, E Sammut, G Plank, CA Rinaldi, and S Niederer. Biophysical Modeling to Determine the Optimization of Left Ventricular Pacing Site and AV/VV Delays in the Acute and Chronic Phase of Cardiac Resynchronization Therapy. *Journal of Cardiovascular Electrophysiology*, 28(2):208–215, 2016.
- [32] GR Mirams, P Pathmanathan, RA Gray, P Challenor, and RH Clayton. White paper: Uncertainty and Variability in Computational and Mathematical Models of Cardiac Physiology. *The Journal of Physiology*, 594(23):6833–6847, 2016.
- [33] RM Neal. Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.
- [34] E Novak and H Woźniakowski. Tractability of Multivariate Problems, Volume II : Standard Information for Functionals. *EMS Tracts in Mathematics 12*, 2010.
- [35] CJ Oates, T Papamarkou, and M Girolami. The Controlled Thermodynamic Integral for Bayesian Model Evidence Evaluation. *Journal of the American Statistical Association*, 2016. To appear.
- [36] A O’Hagan. Monte Carlo is fundamentally unsound. *Journal of the Royal Statistical Society, Series D*, 36(2/3):247–249, 1987.
- [37] A O’Hagan. Bayes–Hermite Quadrature. *Journal of Statistical Planning and Inference*, 29(3):245–260, 1991.
- [38] M Osborne, R Garnett, S Roberts, C Hart, S Aigrain, and N Gibson. Bayesian quadrature for ratios. In *Artificial Intelligence and Statistics*, pages 832–840, 2012.

- [39] MA Osborne, DK Duvenaud, R Garnett, CE Rasmussen, SJ Roberts, and Z Ghahramani. Active learning of model evidence using Bayesian quadrature. In *Advances in Neural Information Processing Systems*, 2012.
- [40] C Rasmussen and C Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.
- [41] C Robert and G Casella. *Monte Carlo Statistical Methods*. Springer Science & Business Media, 2013.
- [42] S Särkkä, J Hartikainen, L Svensson, and F Sandblom. On the relation between Gaussian process quadratures and sigma-point methods. *Journal of Advances in Information Fusion*, 11(1):31–46, 2016.
- [43] J Sethuraman. A Constructive Definition of Dirichlet Priors. *Statistica Sinica*, 4(2):639–650, 1994.
- [44] J Skilling. Bayesian Solution of Ordinary Differential Equations. In *Maximum Entropy and Bayesian Methods*, pages 23–37. Springer, 1992.
- [45] A Smola, A Gretton, L Song, and B Schölkopf. A Hilbert Space Embedding for Distributions. *Algorithmic Learning Theory, Lecture Notes in Computer Science*, 4754:13–31, 2007.
- [46] H Strathmann, D Sejdinovic, S Livingstone, Z Szabo, and A Gretton. Gradient-free hamiltonian monte carlo with efficient kernel exponential families. In *Advances in Neural Information Processing Systems*, pages 955–963, 2015.
- [47] R Von Mises. *Mathematical Theory of Probability and Statistics*. Academic, London, 1974.
- [48] MP Wand and MC Jones. *Kernel Smoothing*. CRC Press, 1994.