A Locally Adaptive Bayesian Cubature Method

A Locally Adaptive Bayesian Cubature Method

Abstract

Bayesian cubature (BC) is a popular inferential perspective on the cubature of expensive integrands, wherein the integrand is emulated using a stochastic process model. Several approaches have been put forward to encode sequential adaptation (i.e. dependence on previous integrand evaluations) into this framework. However, these proposals have been limited to either estimating the parameters of a stationary covariance model or focusing computational resources on regions where large values are taken by the integrand. In contrast, many classical adaptive cubature methods focus computational resources on spatial regions in which local error estimates are largest. The contributions of this work are three-fold: First, we present a theoretical result that suggests there does not exist a direct Bayesian analogue of the classical adaptive trapezoidal method. Then we put forward a novel BC method that has empirically similar behaviour to the adaptive trapezoidal method. Finally we present evidence that the novel method provides improved cubature performance, relative to standard BC, in a detailed empirical assessment.

1 Introduction

In this paper we consider the numerical approximation of the integral

$$I(f^*) \coloneqq \int_D f^*(x)\,\mathrm{d}\pi(x), \tag{1}$$

of a continuous function $f^* : D \to \mathbb{R}$ with respect to a Borel reference measure π supported on a compact set $D \subset \mathbb{R}^d$. In particular, we consider the case where the evaluation of f^* is associated with a substantial computational cost. To control computational cost, a cubature method should attempt to control the number of evaluations of f^* required to obtain a desired level of accuracy for (1). In particular, a desirable attribute of a cubature method is to focus integrand evaluations on subregions of D in which the approximation of f^* is most difficult. If the user has no *a priori* knowledge about the locations of such regions then the cubature algorithm must be *locally adaptive* if it is to fulfill this requirement. Furthermore, any practical cubature method should provide an estimate of its precision, such as an *a posteriori* error estimate if the cubature method is classical, or a credible interval if a probabilistic cubature method is used.

The *Bayesian cubature* (BC) method for approximation of (1) can be traced back to Larkin (1972). Here, approximation of (1) is framed as an inferential task where the integrand f^* carries the status of a latent variable to be inferred. A distinguishing feature of BC, compared to classical approaches, is that the output of the method is a probability distribution on $\mathbb{R}$, simultaneously providing estimates and quantification of uncertainty regarding the value of the integral (1). The method finds application in machine learning (Osborne et al., 2012), statistics (Briol et al., 2019), signal processing (Prüher et al., 2018) and econometrics (Oettershagen, 2017), most typically in situations where evaluation of the integrand f^* is associated with a substantial computational cost. In the context of uncertainty quantification, for example, it becomes natural and parsemonious to combine the probabilistic output provided by BC with other probabilistic representations of uncertainty, such as measurement error and model error.

The general framework for BC can be expressed using two ingredients, the first of which is an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which a stochastic process $f : D \times \Omega \to \mathbb{R}$ is defined. This serves as a statistical model for the latent f^* and is endowed with the Bayesian semantics of *a priori* knowledge about the integrand. For instance, global properties, such as periodicity or monotonicity, and local properties, such as continuity and differentiability, may be known *a priori* and encoded. It is minimally required that sample paths of f are continuous and that f admits well-defined conditional processes, denoted $f|\mathcal{D}_n$, whenever $\mathcal{D}_n = \{(x_i, f^*(x_i))\}_{i=1}^n$ specifies n evaluations of the integrand on which the process is conditioned. Thus, in particular, the stochastic process f

Preliminary work. Under review by AISTATS 2020. Do not distribute.

can be integrated, giving rise to a random variable

$$\begin{aligned} I(f) : \Omega &\rightarrow \mathbb{R} \\ \omega &\mapsto \int_D f(x,\omega)\,\mathrm{d}\pi(x). \end{aligned}$$

The second ingredient is an *acquisition function* A, which – roughly speaking – maps a stochastic process (such as f) to a state $x \in D$. At iteration n of a BC method, the acquisition function is applied to the conditional process $f|\mathcal{D}_{n-1}$ and the output $x_n \in D$ represents the location where the integrand is next evaluated. The conditional process $f|\mathcal{D}_n$ can be integrated to produce a random variable $I(f)|\mathcal{D}_n$ on $\mathbb{R}$, whose distribution is the posterior marginal distribution for the integral (1); this is the output of the BC method. Note that we do not mandate a stopping rule based on an error estimate as part of a BC method; we are motivated by problems where f^* is associated with a substantial computational cost, so that one cannot practically expect to evaluate the integrand as many times as needed to achieve a pre-specified error threshold.

Through the choice of the stochastic process f and the acquisition function A, the behaviour of the BC method can be controlled. Here we overview existing work on BC, in terms of the framework just set out. Attention is limited to approaches that select the x_i according to some optimality criterion, as opposed to a set or sequence of x_i being *a priori* posited (for a discussion of the latter context, which has also been widely-studied, see Briol et al., 2019; Jagadeeswaran and Hickernell, 2019). The symbols $\mathbb{E}$, $\mathbb{V}$ and $\mathbb{C}$ are used to denote expectation, variance and covariance with respect to the underlying prior measure $\mathbb{P}$.

Non-Adaptive BC: The earliest contributions to this area, from Sul'din (1959, 1960); Larkin (1974); Diaconis (1988) and O'Hagan (1991), considered a Gaussian stochastic process model $f \sim \mathcal{GP}(m,k)$ for the integrand, with mean function $m(x) = \mathbb{E}[f(x)]$ and covariance function $k(x,y) = \mathbb{C}[f(x), f(y)]$ being *a priori* specified (Rasmussen and Williams, 2006). It can be shown that $\mathbb{V}[I(f)|\mathcal{D}_n]$, the posterior variance of the integral, depends on $\mathcal{D}_n$ only through the locations x_i and not the actual values $f^*(x_i)$ obtained. Thus the posterior variance can be arbitrarily small whilst the actual error can be arbitrarily large. These aforementioned authors proposed to select the x_i in a manner that minimises $\mathbb{V}[I(f)|\mathcal{D}_n]$, and as such no adaptation is achieved. Indeed, in those references the $\{x_i\}_{i=1}^n$ were pre-computed to globally minimise $\mathbb{V}[I(f)|\mathcal{D}_n]$ over the product space D^n, though we note that sequential (greedy) alternatives have been studied in Oettershagen (2017); Pronzato and Zhigljavsky (2018).

Globally Adaptive BC: Subsequent authors considered parametric families of stationary Gaussian processes $f|\theta \sim \mathcal{GP}(m_\theta, k_\theta)$, where k_θ has the form $k_\theta(x,y) = \phi_\theta(\|x-y\|)$, $\phi_\theta : [0,\infty) \rightarrow \mathbb{R}$ (e.g. $\phi_\theta(s) = \theta_1^2 e^{-s^2/\theta_2^2}$), considering the parameter $\theta = (\theta_1, \theta_2)$ as a latent variable to also be inferred. This additional flexibility allows $\mathbb{V}[I(f)|\mathcal{D}_n]$ to depend on $\{f^*(x_i)\}_{i=1}^n$ and so some form of adaptivity may be achieved when, for example, the minimum expected variance acquisition function

$$A(f|\mathcal{D}_{n-1}) \in \operatorname*{arg\,min}_{x_n \in D} \mathbb{E}[\mathbb{V}[I(f)|\tilde{\mathcal{D}}_n]|\mathcal{D}_{n-1}] \qquad (2)$$

is used. Here $\mathbb{E}[\cdot|\mathcal{D}_{n-1}]$ denotes expectation with respect to $f|\mathcal{D}_{n-1}$ and $\tilde{\mathcal{D}}_n = \mathcal{D}_{n-1} \cup \{(x_n, f(x_n))\}$. In other words, x_n is selected to minimise the expectation of $\mathbb{V}[I(f)|\tilde{\mathcal{D}}_n]$ when the random variable $f(x_n)$ is distributed according to its marginal under $f|\mathcal{D}_{n-1}$. Adaptive selection of the x_i in this context was studied in Osborne (2010). The stationary (i.e. global) nature of the covariance model ϕ_θ has the limitation that the resulting set $\{x_i\}_{i=1}^n$ tends to focus equally on regions where the integrand is both well and not well approximated. Indeed, inferences for the parameter θ are principally driven by the "most difficult" part of the integrand, even if that region is spatially localised. Thus any stopping rule based on the posterior variance of the integral results in unnecessary computational effort devoted to regions in which the integrand can be easily approximated.

Locally Adaptive BC: The transformed stochastic process model $f(x,\omega) = T(g(x,\omega))$, where $T : \mathbb{R} \rightarrow \mathbb{R}$ is a pre-specified transformation and $g \sim \mathcal{GP}(m,k)$, has been proposed to encode global properties such as positivity (e.g. $T(z) = z^2$) into the stochastic process model. This was considered empirically in Gunter et al. (2014); Chai and Garnett (2019) and theoretically in Kanagawa and Hennig (2019). Coupled with the acquisition function (2), this construction behaves in such a way that regions in which $f^*(> 0)$ is large are allocated more of the computational budget.[1] Though appropriate in some situations (in particular, computation of marginal likelihood), such behaviour is not universally desirable (for instance, if f^* is easily approximated in the regions where f^* is large then such a strategy is likely to be inefficient).

Despite this extensive research development, the basic notion of allocating more computational resource to regions where approximation of the integrand is most difficult has not yet been realised in the context of a BC method. It is emphasised that adaptivity in

[1] The authors proposed also an indirect but more convenient alternative to (2), seeking instead the x for which the variance of $f(x)|\mathcal{D}_{n-1}$ is greatest.

this sense is ubiquitous throughout classical numerical analysis; for instance `QUADPACK` (Piessens, 1983) has been a standard integration library since its inception and all but one of its integration routines are adaptive. In addition, for sufficiently challenging integration problems it is known, both theoretically (Ritter, 2000, Chap. VII.3) and empirically (Rabe-Hesketh et al., 2002), that local adaptation is practically essential. It is therefore interesting and important to ask whether local adaptivity can also be exhibited by a suitably-designed BC method.

Outline: Our contributions in this paper are threefold: After recalling the classical adaptive trapezoidal method in Section 2 we then present a theoretical result, in Section 3, that suggests there does not exist a direct Bayesian analogue of this classical method. Then, in Section 4 we put forward a novel BC method that has empirically similar behaviour to the adaptive trapezoidal method. Its performance is empirically assessed in Section 5.

2 Background

In Section 2.1 the classical adaptive approach to cubature is briefly recalled, while standard background on the BC method is contained in Section 2.2.

2.1 Classical Adaptive Cubature

Classical approaches to (non-adaptive, for the moment) cubature can be categorised either as non-constructive (e.g. Monte Carlo, quasi Monte Carlo) or constructive (e.g. Newton-Cotes rules, Gaussian cubature). The latter are distinguished by the fact that they first construct an approximation to the integrand itself, typically an interpolant, and then exactly integrate this interpolant to obtain an approximation of (1). In either case, for a *linear* cubature method the output is an approximation

$$Q_n(f^*) \coloneqq \sum_{i=1}^{n} w_i f^*(x_i) \approx \int_D f^*(x)\,\mathrm{d}\pi(x) \qquad (3)$$

based on a set $\{x_i\}_{i=1}^n \subset D$ that must be specified. The point estimate $Q_n(f^*)$ is accompanied by an assessment of its error, $\epsilon = |I(f^*) - Q_n(f^*)|$, typically formulated as the difference $\tilde{\epsilon} = |Q_n(f^*) - Q_m(f^*)|$ of two cubature rules[2] (though we note that more general approaches based on extrapolation are also used; Richardson and Gaunt, 1927).

[2]This can be motivated as follows: If $Q_n(f^*)$ is provably better than $Q_m(f^*)$, say $|I(f^*) - Q_n(f^*)| \leq \frac{1}{2}|I(f^*) - Q_m(f^*)|$, then we have $\epsilon = |I(f^*) - Q_n(f^*)| \leq |Q_n(f^*) - Q_m(f^*)| =: \tilde{\epsilon}$, so $\tilde{\epsilon}$ is a genuine error bound.

The classical notion of local adaptivity is to recursively partition the integration domain $D = \cup_{r=1}^R D_r$ into sub-regions D_r over which local cubature rules of the form (3) are applied. An estimate $\tilde{\epsilon}_r$ of the error ϵ_r of these rules is produced for each region D_r and, if the estimated error is too big, those regions are sub-divided again until a global error tolerance $\sum_{r=1}^R \tilde{\epsilon}_r < \tau$ is satisfied.[3] Several such methods have been proposed, see Gonnet (2012). For example, recall the trapezoidal rule on $D = [a, b]$ with $\mathrm{d}\pi(x) = \mathrm{d}x$, which has the form,

$$\texttt{Trap}(f^*, a, b, n) \coloneqq \tfrac{b-a}{2n}\big(f^*(a) + f^*(b) + 2\textstyle\sum_{i=1}^{n-1} f^*\big(a + \tfrac{i(b-a)}{n}\big)\big). \qquad (4)$$

The trapezoidal rule forms the basis for the classical locally adaptive trapezoidal method:

Algorithm 1 Adaptive Trapezium Method

procedure $\textsc{AdapTrap}_{\rho,m,k}(f^*, a, b, \tau)$
  $Q_1 \leftarrow \texttt{Trap}(f^*, a, b, m)$
  $Q_2 \leftarrow \texttt{Trap}(f^*, a, b, 2m)$
  $\tilde{\epsilon} \leftarrow |Q_2 - Q_1|$
  if $\tilde{\epsilon} < \tau$ **then**
    $\hat{I} \leftarrow Q_2$
  else
    $\tau' \leftarrow \rho\tau$
    $\hat{I} \leftarrow \sum_{i=0}^{l-1} \texttt{AdapTrap}_{\rho,m,k}(f^*, a + \frac{(b-a)i}{k}, a + \frac{(b-a)(i+1)}{k}, \tau')$
  return $\hat{I}$

The `AdapTrap` method is an adaptive trapezoidal rule where the decision to subdivide into k uniform intervals is determined by the difference between the composite trapezoidal rule on $2m$ intervals and the composite trapezoidal rule on m intervals. The values $\tilde{\epsilon}$ thus form local error estimates and we accept our trapezoidal approximation to the integral on the subinterval only when $\tilde{\epsilon}$ is sufficiently small. The parameter ρ controls how the error tolerance τ scales at each recursive step of the algorithm and has natural choice $\rho = \frac{1}{k}$.

Generalisation of the `AdapTrap` algorithm is straightforward through the use of higher-order cubature rules (e.g. Simpson's rule or Gaussian quadrature) within each step of the procedure (Davis and Rabinowitz, 1984; Kahaner and Rechard, 1987; Berntsen et al., 1991). It is intuitively clear that any such method will attempt to allocate computational resources to those

[3]This setting differs slightly to the setting in which BC is used. Indeed, for the problems on which BC is used, f^* cannot in general be repeatedly evaluated until a global error tolerance is satisfied due to its prohibitive computational cost.

regions where approximation of f^* is most difficult. As argued in Section 1, this is not a feature of any existing BC method.

2.2 Standard Bayesian Cubature

In this section we briefly recall the pertinent aspects of the standard BC method.

Notation Let f_X^* with $[f_X^*]_i = f^*(x_i)$ contain evaluations of the integrand on the ordered n-tuple $X = (x_1, \ldots, x_n) \in D^n$. For $k : D \times D \to \mathbb{R}$ and $Y = (y_1, \ldots, y_m) \in D^m$, the matrix K_{XY} is defined as $[K_{XY}]_{ij} := k(x_i, y_j)$. Let also $K_X(y)$ be defined as $[K_X(y)]_i := k(x_i, y)$ whenever $y \in D$. The equivalent presentations of stochastic processes $f : D \times \Omega \to \mathbb{R}$ and $f(x) : \Omega \to \mathbb{R}$ are used, so that f_X where $[f_X]_i = f(x_i)$ is a random vector in $\mathbb{R}^n$.

Recall that a stochastic process f is Gaussian if and only if, for any $X \in D^n$, $n \in \mathbb{N}$, the random vector f_X is Gaussian-distributed. Thus a Gaussian process f is completely specified by its mean function $m(x) := \mathbb{E}[f(x)]$ and covariance function $k(x, y) := \mathbb{C}[f(x), f(y)]$ and we write $f \sim \mathcal{GP}(m, k)$. Under mild regularity conditions (which are beyond the scope of this work to discuss in detail; see Bogachev, 1998) it can be shown that the conditional stochastic processes $f|\mathcal{D}_n$ are well-defined, are also Gaussian, and have mean and covariance functions

$$m_{\mathcal{D}_n}(x) = f_X^{*\top} K_{XX}^{-1} K_X(x), \tag{5}$$

$$k_{\mathcal{D}_n}(x, y) = k(x, y) - K_X(x)^\top K_{XX}^{-1} K_X(y). \tag{6}$$

The output of the BC method is the random variable $I(f)|\mathcal{D}_n \sim \mathcal{N}(\mu_n, \sigma_n^2)$, which can be read off (5) and (6) as a univariate marginal:

$$\begin{aligned} \mu_n &= \textstyle\int_D m_{\mathcal{D}_n}(x)\,\mathrm{d}\pi(x) \\ &= f_X^{*\top} K_{XX}^{-1} \textstyle\int_D K_X(x)\,\mathrm{d}\pi(x), \qquad (7) \\ \sigma_n^2 &= \textstyle\int_D \int_D k_{\mathcal{D}_n}(x, y)\,\mathrm{d}\pi(x)\mathrm{d}\pi(y) \\ &= \textstyle\int_D \int_D k(x, y)\,\mathrm{d}\pi(x)\,\mathrm{d}\pi(y) \qquad (8) \\ &\quad - \left[\textstyle\int_D K_X(x)\,\mathrm{d}\pi(x)\right]^\top K_{XX}^{-1} \left[\textstyle\int_D K_X(y)\,\mathrm{d}\pi(y)\right]. \end{aligned}$$

The posterior mean (7) is seen to have the same form as (3). It is natural to select the design X in such a way that the posterior variance (8) is minimised. Since (8) does not depend on f^*, no adaptive estimation occurs in the standard BC method and the assessment of uncertainty provided by (8) is exclusively driven by the *a priori* specification of k and X. This behaviour is unsatisfactory, as posterior variance can be arbitrarily small whilst the actual error can be arbitrarily large. However, this property does allow optimal designs X to, in principle, be pre-computed (Sul'din, 1959, 1960; O'Hagan, 1991; Minka, 2000). Strategies to ensure analytic expressions for the integrals in (7) and (8) were proposed in Briol et al. (2019); Jagadeeswaran and Hickernell (2019). For large n, techniques have been put forward to facilitate the efficient inversion of the matrix K_{XX} (Karvonen and Särkkä, 2018; Karvonen et al., 2019; Jagadeeswaran and Hickernell, 2019).

Proposals that go beyond the standard BC method were outlined in Section 1. The simplest route to adaptivity is to consider a parametric family of covariance functions k_θ and to treat the parameter θ also as a latent variable to be inferred. For example, if $k_\theta(x, y) = \theta_1^2 e^{-\|x-y\|^2/\theta_2^2}$ with $\theta = (\theta_1, \theta_2)$, then estimation of θ_1 corresponds (roughly speaking) to estimating the amplitude of the integrand, while θ_2 corresponds to a characteristic lengthscale for the integrand. This form of adaptation (which may be realised either through full Bayesian inference for θ or as an empirical Bayes method) was first empirically demonstrated to produce reliable uncertainty assessment in Larkin (1974). However, the stationary form of the covariance model (i.e. the fact that two parameters θ_1 and θ_2 are required to describe the entire integrand) precludes the focussing of computational resources on those regions in which approximation of the integrand is most difficult.[4] As a result, for integrands involving spatially-localised variation, existing BC methods based on a stationary covariance model can be arbitrarily inefficient in terms of the number of evaluations of the integrand.

All existing work on the BC method, with the exception of the transformed stochastic process models of Gunter et al. (2014); Chai and Garnett (2019); Kanagawa and Hennig (2019), have been based upon a stationary covariance model.[5] Thus, in particular, no Bayesian analogues of classical locally adaptive methods have been proposed. In the next section we establish a cautionary result on the difficulties in developing a Bayesian analogue of the adaptive trapezoidal method. This serves as motivation for our novel proposal in Section 4.

3 A Bayesian `AdapTrap`?

The aim of this section is to discuss how one might naively attempt to create a direct Bayesian analogue of `AdapTrap`. To this end we recall the approach of

[4] The use of greedy sequential strategies for function approximation under a stationary covariance model leads to designs that are essentially space-filling (Cor. 11 of Santin et al., 2017).

[5] The latter exceptions propose to focus computational resources on regions in which $f^*(> 0)$ is large, which in general is not the same as focussing on regions where approximation of f^* is most difficult.

Diaconis (1988), who took a classical cubature rule of the form (3) and asked "for what prior does (3) arise as the mean of the posterior marginal distribution of the integral?".[6] Thus, in the context of creating an analogue of `AdapTrap`, we can follow Diaconis and seek a prior such that the mean of the posterior marginal for the integral is `Trap` in (4). Thus we must consider stochastic processes for which the conditional mean is the piecewise linear interpolant (over the range of $x_1, \ldots, x_n$) of the data $\mathcal{D}_n$ on which it is conditioned.

Let $C([a,b])$ denote the set of continuous real-valued functions on $[a,b]$ and consider the subset $F_{\rho,m,k,\tau} \subset C([a,b])$ of integrands f^* for which $\texttt{AdapTrap}_{\rho,m,k}$ fails to achieve its stated error tolerance τ upon termination, or for which $\texttt{AdapTrap}_{\rho,m,k}$ fails to terminate at all (this set is non-empty; e.g. Clancy et al., 2014). From an inferential perspective, the decision to employ $\texttt{AdapTrap}_{\rho,m,k}$ can be regarded as a belief that $f^* \notin F_{\rho,m,k,\tau}$. Proposition 3.1, presented next, suggests that stochastic process models giving rise to piecewise linear interpolants are incompatible with the use of $\texttt{AdapTrap}_{\rho,m,k}$, due to assigning non-zero probability mass to $F_{\rho,m,k,\tau}$ whenever $\tau > 0$. This result, whose proof is straight-forward and contained in the supplement, can be interpreted as an average-case analysis of `AdapTrap` (Ritter, 2000). Denote the error function $\mathrm{erf}(x) := \frac{1}{\sqrt{\pi}} \int_{-x}^{x} e^{-t^2} \, \mathrm{d}t$.

Proposition 3.1. Fix $a < b$, $\rho > 0$, $m \in \mathbb{N}$ and k a positive even integer. Let f^* be sampled from a centred Gaussian process on $C([a,b])$, whose law is denoted $\mathbb{P}^*$, such that the conditional mean $f^*|\mathcal{D}_n$ is the piecewise linear interpolant (over the range of $x_1, \ldots, x_n$) of the data $\mathcal{D}_n$ on which it is conditioned. If `AdapTrap` terminates, denote its error $\epsilon_{\rho,m,k,\tau}(f^*) := I(f^*) - \texttt{AdapTrap}_{\rho,m,k}(f^*, a, b, \tau)$, otherwise set $\epsilon_{\rho,m,k,\tau}(f^*) := \infty$. Then for every $\tau > 0$,

$$\mathbb{P}^*(|\epsilon_{\rho,m,k,\tau}| > \tau) \; > \; \mathrm{erf}(c\tau) \left[1 - \mathrm{erf}(\sqrt{3}c\tau)\right],$$

where $c > 0$ is a $\mathbb{P}^*$-dependent constant.

Though the probability mass assigned to $F_{\rho,m,k,\tau}$ can be made small, the fact that it is non-zero for all $\tau > 0$ calls into doubt whether direct Bayesian analogues of classical adaptive methods can exist, in contrast to the situation for non-adaptive methods (Karvonen et al., 2018). In Appendix A.2, further average-case analysis is provided, showing that for mis-specified ρ the expected number of steps of `AdapTrap` can be unbounded. Taken together, our analyses suggest that classical adaptive methods cannot be directly replicated in BC and a different strategy is needed. In Section 4 we therefore put forward a *de novo* BC method, which achieves adaptivity through a flexible non-stationary stochastic process model.

[6] Paraphrased. Conversely, Cor. 2.10 of Karvonen et al. (2018) showed that *all* non-adaptive cubature rules of the form (3) arise as the posterior mean of some stochastic process model.

4 Adaptive Bayesian Cubature

The aim of this section is to develop a novel BC method that is locally adaptive, in the sense of focussing integrand evaluations on spatial regions where approximation of f^* is most difficult. The forgoing discussion in Sections 1-3 suggests that this should be based on a *non-stationary* stochastic process model.

4.1 A Non-Stationary Process Model

Several non-stationary stochastic process models have been developed and in principle any of these could form the basis for a BC method. Three broad classes of non-stationary model are those based on deformation of the domain, partitioning of the domain, and convolution over the domain.[7] The *spatial deformation* approach considers a stochastic process of the form $f(x,\omega) = g(v(x),\omega)$, where g is a stationary stochastic process on D and v is a map from D to itself. Such models are flexible but conditioning on data in this context can be computationally difficult. The joint estimation of g and v was considered in a frequentist context in Sampson and Guttorp (1992) using thin-plate splines; analogous Bayesian approaches were developed in Damian et al. (2001); Schmidt and O'Hagan (2003); Damianou and Lawrence (2013). A *Bayesian partition model* represents a non-stationary process using piecewise stationary processes, each fitted on one element of a partition of D (Kim et al., 2005; Gramacy and Lee, 2008). The advantage of such a model is its simplicity and ease to fit, but an unfortunate consequence is that continuity of the process across elements of the partition is not easily enforced. The *process convolution* approach takes a collection of local covariance models and then – roughly speaking – convolves them to obtain a new, non-stationary global covariance model (Higdon et al., 1999; Paciorek, 2003). Theoretical results on the flexibility of these models have been established (Dunlop et al., 2018).

The process convolution approach was used for the experiments in this paper. This choice allows for substantial flexibility to incorporate *a priori* knowledge and to adapt, in principle, to non-stationary features of the integrand.[8] Following Paciorek (2003), we adopted

[7] This discussion is not intended to be comprehensive and work that does not naturally fall into any of the three categories identified, such as Ba et al. (2012), is not discussed.

[8] Although partition models are closer in spirit to classical adaptive methods, the fact that they only provide an

a hierarchical Gaussian process model with spatially-dependent lengthscale field. The first part of the model specifies that $f|\theta \sim \mathcal{GP}(m_\theta, k_\theta)$. The mean function $m_\theta = c$ is here taken as a constant $c \in \mathbb{R}$ and, letting $\phi : [0, \infty) \to \mathbb{R}$ be a positive definite radial basis function, the covariance function has the form

$$k_\theta(x, y) = \frac{\sigma^2 \sqrt{\ell(x)\ell(y)}}{\sqrt{\ell(x)^2 + \ell(y)^2}} \phi\left(\frac{\|x - y\|}{\sqrt{\ell(x)^2 + \ell(y)^2}}\right).$$

The parameters to be jointly inferred are $\theta = \{c, \sigma, \ell(\cdot)\}$, where $\sigma > 0$ is an amplitude parameter and $\ell : D \to [0, \infty)$ is a lengthscale field. The second part of the hierarchical model specifies a prior distribution for θ. The lengthscale $\ell(\cdot)$ was itself parametrised as a piecewise linear and non-negative function throughout. Specific choices for ϕ, the prior for θ and the parametrisation of $\ell(\cdot)$ are deferred to Section 5.

4.2 Adaptive Selection of the Point Set

A sequential approach to selecting the x_i was adopted, based on the minimum expected variance acquisition function (2) of Osborne (2010). This can be viewed as a specific instance of sequential Bayesian optimal experimental design (BOED; Chaloner and Verdinelli, 1995).[9] As is typical in BOED, (2) is an intractable global optimisation problem over D that must in practice be approximated (e.g. Overstall et al., 2018). Two practical algorithms are now presented. In what follows we let $\mathcal{D}_0$ be pre-specified and let $D_n \subset D$ denote a finite set of reference points in D over which the optimisation (e.g. grid search) required at stage n of the algorithm is performed; full details are reserved for Appendix D. Recall that we do not mandate a stopping rule as part of a BC method. However, if required then the standard deviation of $I(f)|\mathcal{D}_n$ can be used to decide when the algorithm should be terminated. For completeness we present our algorithms with an explicit stopping rule included.

Algorithm 3, which is reserved for the supplement, uses Markov chain Monte Carlo (MCMC) to approximate the intractable acquisition function (2), in an idealised approach that we call `AdapBC`. The computational requirement of MCMC is assumed to be negligible compared to the cost of evaluating the integrand. However, the need to ensure convergence of the Markov chain introduces practical difficulties for the user and therefore we focus on an empirical Bayes (EB) alternative

Algorithm 2 (E) Adaptive Bayesian Cubature

1: **procedure** E-AdapBC(f^*, τ)
2: $n \leftarrow 1$, $\tilde{\epsilon} \leftarrow \infty$
3: **while** $\tilde{\epsilon} \geq \tau$ **do**
4: $\theta_n \leftarrow \arg\max_\theta p(\mathcal{D}_{n-1} \,|\, \theta) - r(\theta)$
5: Sample $(f_m)_{m=1}^M \sim f \,|\, \theta_n, \mathcal{D}_{n-1}$ ▷ $M \gg 1$
6: **for** each x in D_n **do**
7: $\tilde{\mathcal{D}}_n \leftarrow \mathcal{D}_{n-1} \cup \{(x, f_m(x))\}$
8: $E(x) \leftarrow \mathbb{E}[\mathbb{V}[I(f)|\theta_n, \tilde{\mathcal{D}}_n]|\theta_n, \mathcal{D}_{n-1}]$
9: Pick $x_n \in \arg\min_{x \in D_n} E(x)$
10: $\mathcal{D}_n \leftarrow \mathcal{D}_{n-1} \cup \{(x_n, f^*(x_n)\}$
11: $n \leftarrow n + 1$, $\tilde{\epsilon} \leftarrow \mathbb{V}[I(f)|\theta_n, \mathcal{D}_n]^{\frac{1}{2}}$
12: **return** $I(f)|\theta_n, \mathcal{D}_n$

in Algorithm 2, called `E-AdapBC`, wherein the parameter θ is estimated rather than being marginalised. To avoid over-confident estimation[10], we regularised the EB estimator using an additional penalty term $r(\theta)$ specified in Appendix D. An advantage of `E-AdapBC` over `AdapBC` is that the computation of the expected variance in line 8 of Algorithm 2 has a closed form, *vis a vis* (8). This completes the methodological development; in the next section the proposed methods are empirically assessed.

5 Experimental Assessment

The purpose of this section is to investigate whether (`AdapBC` and) `E-AdapBC` provide the local adaptation that is missing from standard BC. For the remainder, we use `StdBC` to signify the simplified version of `E-AdapBC` in which the lengthscale field $\ell(\cdot)$ is simply a constant, to be estimated. All other settings (e.g. the choice of ϕ), were taken to be identical between `StdBC` and `E-AdapBC`. All methods that we consider incur an auxiliary computational cost that is orders of magnitude larger than that which would be associated with a classical cubature method. BC methods are motivated by situations where evaluation of f^* is associated with a substantial computational cost (an explicit example is provided in Section 5.3), so that such auxiliary computation can be justified. For this reason, computational cost is quantified in the results that follow only through the number of evaluations of the integrand.

A BC method is considered to perform well if, loosely speaking, the posterior mean $\mu_n(f^*) := \mathbb{E}[I(f)|\theta_n, \mathcal{D}_n]$ provides an accurate point estimate of (1) and the posterior spread $\sigma_n(f^*) := \mathbb{V}[I(f)|\theta_n, \mathcal{D}_n]^{\frac{1}{2}}$ is well-

approximate notion of conditioning precludes their use for rigorous uncertainty quantification in a BC method.

[9] Recall that all the standard notions of optimality, such as A- and D optimality, coincide in the univariate Gaussian context and correspond to minimising the *a priori* expected variance of the quantity of interest.

[10] The use of EB in the context of the BC method was shown to result in over-confident estimation at small n in Briol et al. (2019).

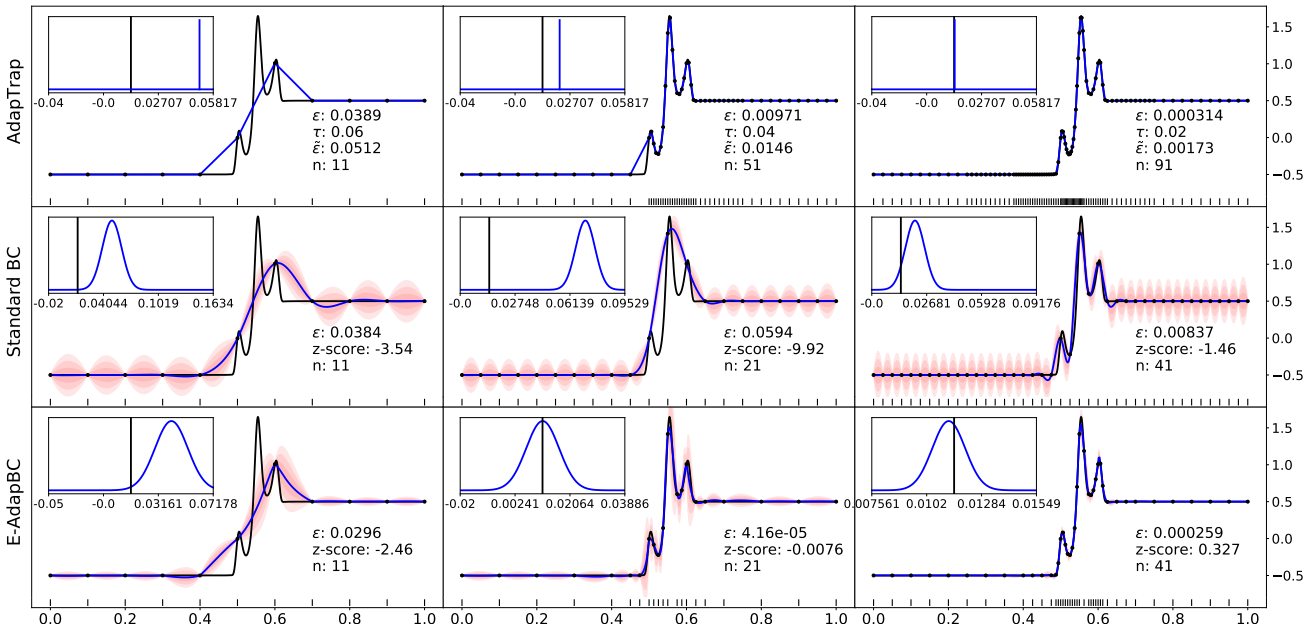


Figure 1: Comparison of `AdapTrap`, `StdBC` and `E-AdapBC`. [Here —— represents the true integrand f^*, —— represents the mean of the conditional process $f|\mathcal{D}_n$ and ■ represents pointwise credible intervals. The tick marks ❙❙❙❙❙ indicate where the integrand was evaluated. For `StdBC` and `E-AdapBC` the error $\epsilon := |\mu_n(f^*) - I(f^*)|$, the z-score $[\mu_n(f^*) - I(f^*)]/\sigma_n(f^*)$ and the number of integrand evaluations n are reported. For `AdapTrap` the error ϵ, the global error tolerance τ, the estimated error $\tilde{\epsilon} := \sum_r \tilde{\epsilon}_r$ and the number of integrand evaluations n are reported. Inset panels compare the true value $I(f^*) \approx 0.011$ to the distribution $I(f)|\theta_n, \mathcal{D}_n$. Settings for all methods are detailed in Appendix E.]

calibrated as an indicator of the true error $|\mu_n(f^*) - I(f^*)|$; in this paper calibratedness is quantified by $Z_n(f^*) := \frac{\mu_n(f^*) - I(f^*)}{\sigma_n(f^*)}$ whose values should be plausible as samples from $\mathcal{N}(0,1)$ when the BC method is well-calibrated (Briol et al., 2019). The ideas are illustrated next in Section 5.1. In Section 5.2 the results of detailed synthetic assessment are presented and in Section 5.3 we report results based on a realistic integration task involving trajectories of an autonomous robot. All results in this paper can be reproduced in `Python` using code available at `github.com/[anonymised]`.

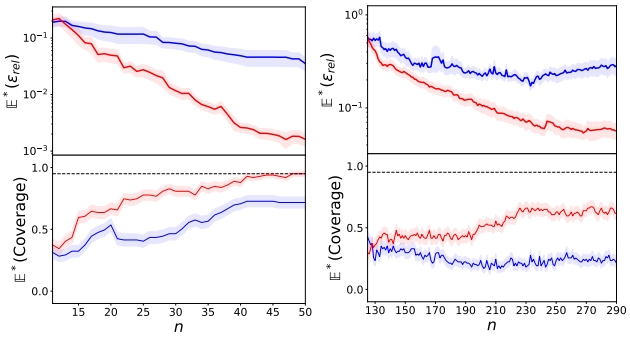
5.1 Illustration of Adaptation

Figure 1 compares the performance of `AdapTrap` (top), `StdBC` (middle) and `E-AdapBC` (bottom) on a toy integrand f^* in dimension $d = 1$. Full details of the specific settings used for all methods are reserved for Appendix E.1. Theoretical analysis of `StdBC` indicates that the points X at which the integrand is evaluated are essentially space-filling (Cor. 11 of Santin et al., 2017). In contrast, both `AdapTrap` and `E-AdapBC` deploy their computational resources in the region where f^* is varying the most. `AdapTrap` provides an accurate point estimate for (1) and a deterministic error estimate $\tilde{\epsilon}$. In each case $\epsilon < \tau$, i.e. the true error has been controlled succesfully by `AdapTrap`. In contrast, both BC methods provide distributional output whose uncertainty is well-calibrated once n is large enough that the regions of highest variation have been found. Of course, Figure 1 studies a single integrand and a more systematic assessment is performed next.

5.2 Synthetic Assessment

To assess the proposed methods on a wider range of test problems, we automatically generated integrands f_i^*, $i = 1, \ldots, 100$, in a manner described in Appendix E.2. The negligible cost of evaluating the synthetic f_i^* ensures that their integrals $I(f_i^*)$ can be accurately approximated using a classical method, providing a gold-standard for assessment. The methods `AdapBC` and `E-AdapBC` were compared to `StdBC`.[11] Figure 2 (top row) displays the mean of the relative errors $\left|\frac{\mu_n(f_i^*) - I(f_i^*)}{I(f_i^*)}\right|$ for `StdBC` and `E-AdapBC`. Results are reported for the case $\mathrm{d}\pi(x) = \mathrm{d}x$ and in dimension $d = 1$ (left) and $d = 3$ (right). It can be seen that the conclusions of Figure 1 hold in broad terms over

[11] The f_i^* can take both positive and negative values, so the methods of Gunter et al. (2014); Chai and Garnett (2019) cannot be directly applied.



(a) (b)

Figure 2: Synthetic assessment in (a) $d = 1$ and (b) $d = 3$ for `StdBC` (——) and `E-AdapBC` (——), where 100 integrands were randomly generated. Top row: the mean relative error against the number of evaluations n. Bottom row: the coverage frequencies for 95% credible intervals for each method. The notional coverage (- - -) is indicated. [Standard errors displayed.]

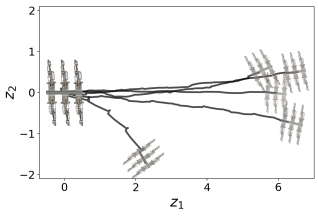
an ensemble of integrands, though of course there exist particular integrands for which `StdBC` happens due to chance to outperform `E-AdapBC`. The bottom row of Figure 2 reports coverage frequencies for the 95% highest-posterior density interval. Over-confidence is apparent at small values of n, especially for `StdBC` and for $d = 3$, but for larger n (when the most variable regions of the integrand are discovered) the methods are better calibrated. The impact of the choice of radial basis function $\phi(\cdot)$ and the parametrisation of the lengthscale field $\ell(\cdot)$ was investigated in Appendix E.3. Results for `AdapBC` were broadly similar to `E-AdapBC` after manual tuning of the MCMC and these are deferred to Appendix E.4.

5.3 Autonomous Robot Assessment

The final experiment concerns an application of `E-AdapBC` to autonomous robotics (Chrono, 2019a). Here $x \in \mathbb{R}^3$ represents parameters that describe the performance of a set of actuators in an autonomous walking robot. The notional value and actual value of x will not be equal in general and there is interest in understanding the effect of parameter variability on the actual trajectory of the robot; see Figure 3a. Let $(z_1(x), z_2(x))$ denote the spatial coordinates of the robot after a fixed sequence of commands have been completed. Conceptually, the variability in the parameters can be represented (after re-parametrisation) as $x \sim \mathcal{N}(0, I_{3\times 3})$ and there is interest in evaluating moments $I(f^*)$ where $f^* \in \{z_1, z_2, z_1^2, z_2^2\}$. The situation typifies instances where f^* is associated with a substantial computational cost, since simulation of the robot moving requires the numerical solution of a system of ordinary differential equations. The methods `StdBC` and `E-AdapBC` were each applied to this task, with full details contained in Appendix E.5. The intractability of the true integrals $I(f^*)$ precludes a direct assessment as in Section 5.2. Instead, we focus on estimation accuracy (only) and report an approximate bound based on Jensen's inequality and Monte Carlo:

$$\begin{aligned}\mathbb{E}[(I(f) - I(f^*))^2|\theta_n, \mathcal{D}_n] &\leq \mathbb{E}[I((f-f^*)^2)|\theta_n, \mathcal{D}_n] \\ &= \int_D \mathbb{E}[(f(x) - f^*(x))^2|\theta_n, \mathcal{D}_n]\mathrm{d}\pi(x) \\ &\approx \tfrac{1}{m}\textstyle\sum_{i=1}^m \mathbb{E}[(f(y_i) - f^*(y_i))^2|\theta_n, \mathcal{D}_n] \qquad (9)\end{aligned}$$

where $y_i \overset{\text{iid}}{\sim} \mathcal{N}(0, I_{d\times d})$ and $m = 264$. For each integrand, `E-AdapBC` outperformed `StdBC` as quantified by (9); see Figure 3b.



(a)

f^*	Mean Sq. Error (9)	
	`StdBC`	`E-AdapBC`
z_1	0.895 ±0.07	**0.293** ±0.03
z_2	14.3 ±2.05	**2.28** ±0.26
z_1^2	0.884 ±0.11	**0.336** ±0.07
z_2^2	1527 ±268.13	**132.13** ±11.13

(b)

Figure 3: Autonomous robot assessment. (a) Trajectories produced by the robot. (b) Error as quantified in (9), for each of integrand f^* relating to the final position of the robot. [Standard errors displayed.]

6 Conclusion

This paper highlighted the important issue of local adaptivity in the context of BC methods and discussed why naive constructions based on lifting classical adaptive methods to the Bayesian framework can fail. To address these issues, a novel locally adaptive BC method was proposed and demonstrated to perform well in both a synthetic and realistic empirical assessment. The construction was quite general, in the sense that essentially any sufficiently flexible Bayesian regression model can be used, and investigation of alternative regression models can form the basis of further work. Also of interest, non-myopic alternatives to (2) have been proposed for BC (Jiang et al., 2019) and these could also be investigated.

Our focus was on cubature, but local adaptation can be considered in the context of other probabilistic numerical methods (Hennig et al., 2015). For example, adaptive time-stepping has recently received attention in the probabilistic numerical solution of ordinary differential equations (Chkrebtii and Campbell, 2019) and analogous methods for partial differential equations have yet to be developed.

References

Ba, S., Joseph, V. R., et al. (2012). Composite Gaussian process models for emulating expensive functions. *The Annals of Applied Statistics*, 6(4):1838–1860.

Berntsen, J., Espelid, T. O., and Sørevik, T. (1991). On the subdivision strategy in adaptive quadrature algorithms. *Journal of Computational and Applied Mathematics*, 35(1-3):119–132.

Bogachev, V. I. (1998). *Gaussian Measures*. Number 62 in Mathematical Surveys and Monographs. American Mathematical Society.

Briol, F.-X., Oates, C. J., Girolami, M., Osborne, M. A., and Sejdinovic, D. (2019). Probabilistic Integration: A Role in Statistical Computation? *Statistical Science*, 34(1):1–22. Appeared with discussion and rejoinder.

Chai, H. and Garnett, R. (2019). Improving Quadrature for Constrained Integrands. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*.

Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304.

Chkrebtii, O. and Campbell, D. (2019). Adaptive step-size selection for state-space probabilistic differential equation solvers. *Statistics and Computing*. To appear.

Chrono, P. (2019a). Chrono: An Open Source Framework for the Physics-Based Simulation of Dynamic Systems. Accessed: 2019-09-15.

Chrono, P. (2019b). Make a spider robot in solidworks and simulate it. Accessed: 2019-10-6.

Clancy, N., Ding, Y., Hamilton, C., Hickernell, F. J., and Zhang, Y. (2014). The cost of deterministic, adaptive, automatic algorithms: Cones, not balls. *Journal of Complexity*, 30(1):21–45.

Damian, D., Sampson, P. D., and Guttorp, P. (2001). Bayesian estimation of semi-parametric non-stationary spatial covariance structures. *Environmetrics*, 12(2):161–178.

Damianou, A. and Lawrence, N. (2013). Deep Gaussian processes. In *Artificial Intelligence and Statistics*, pages 207–215.

Davis, P. J. and Rabinowitz, P. (1984). *Methods of Numerical Integration*. Academic Press.

Diaconis, P. (1988). Bayesian numerical analysis. *Statistical Decision Theory and Related Topics IV*, 1:163–175.

Dick, J. and Pillichshammer, F. (2010). *Digital nets and sequences: discrepancy theory and quasi–Monte Carlo integration*. Cambridge University Press.

Dunlop, M., Girolami, M., Stuart, A., and Teckentrup, A. (2018). How Deep Are Deep Gaussian Processes? *Journal of Machine Learning Research*, 19(1):2100–2145.

Gonnet, P. (2012). A Review of Error Estimation in Adaptive Quadrature. *ACM Computing Surveys (CSUR)*, 44(4):22.

Graham, R. L., Knuth, D. E., and Patashnik, O. (1994). *Concrete Mathematics: A Foundation for Computer Science*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2nd edition.

Gramacy, R. B. and Lee, H. K. H. (2008). Bayesian Treed Gaussian Process Models with an Application to Computer Modeling. *Journal of the American Statistical Association*, 103(483):1119–1130.

Gunter, T., Osborne, M. A., Garnett, R., Hennig, P., and Roberts, S. J. (2014). Sampling for Inference in Probabilistic Models with Fast Bayesian Quadrature. In *Advances in Neural Information Processing Systems*, pages 2789–2797.

Hennig, P., Osborne, M. A., and Girolami, M. (2015). Probabilistic numerics and uncertainty in computations. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2179):20150142.

Higdon, D. M., Swall, J. L., and Kern, J. C. (1999). *Bayesian Statistics*, volume 6, chapter Non-Stationary Spatial Modeling.

Jagadeeswaran, R. and Hickernell, F. J. (2019). Fast Automatic Bayesian Cubature Using Lattice Sampling. *Statistics and Computing*. To appear.

Jiang, S., Chai, H., Gonzalez, J., and Garnett, R. (2019). Efficient nonmyopic Bayesian optimization and quadrature. *arXiv:1909.04568*.

Kahaner, D. K. and Rechard, O. W. (1987). TWODQD an adaptive routine for two-dimensional integration. *Journal of Computational and Applied Mathematics*, 17(1-2):215–234.

Kanagawa, M. and Hennig, P. (2019). Convergence Guarantees for Adaptive Bayesian Quadrature Methods. *arXiv:1905.10271*.

Karvonen, T., Oates, C. J., and Särkkä, S. (2018). A Bayes–Sard cubature method. In *32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*.

Karvonen, T. and Särkkä, S. (2018). Fully Symmetric Kernel Quadrature. *SIAM Journal on Scientific Computing*, 40(2):A697–A720.

Karvonen, T., Särkkä, S., and Oates, C. J. (2019). Symmetry Exploits for Bayesian Cubature Methods. *Statistics and Computing*. To appear.

Kim, H.-M., Mallick, B. K., and Holmes, C. C. (2005). Analyzing Nonstationary Spatial Data Using Piecewise Gaussian Processes. *Journal of the American Statistical Association*, 100(470):653–668.

Larkin, F. (1972). Gaussian measure in Hilbert space and applications in numerical analysis. *Rocky Mountain Journal of Mathematics*, 2(3):379–422.

Larkin, F. (1974). Probabilistic error estimates in spline interpolation and quadrature. In *IFIP Congress; Information Processing*, volume 74, pages 605–609.

Minka, T. (2000). Deriving Quadrature Rules from Gaussian Processes. Technical report, Statistics Department, Carnegie Mellon University.

Oettershagen, J. (2017). *Construction of Optimal Cubature Algorithms with Applications to Econometrics and Uncertainty Quantification*. PhD thesis, Institut für Numerische Simulation, Universität Bonn.

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

Osborne, M. (2010). *Bayesian Gaussian Processes for Sequential Prediction, Optimisation and Quadrature*. PhD thesis, University of Oxford.

Osborne, M. A., Duvenaud, D., Garnett, R., Rasmussen, C. E., Roberts, S. J., and Ghahramani, Z. (2012). Active Learning of Model Evidence Using Bayesian Quadrature. In *Advances in Neural Information Processing Systems*.

Overstall, A. M., McGree, J. M., and Drovandi, C. C. (2018). An approach for finding fully Bayesian optimal designs using normal-based approximations to loss functions. *Statistics and Computing*, 28(2):343–358.

Paciorek, C. J. (2003). *Nonstationary Gaussian processes for regression and spatial modelling*. PhD thesis, Carnegie Mellon University.

Piessens, R. (1983). *Quadpack : a subroutine package for automatic integration*. Springer-Verlag.

Pronzato, L. and Zhigljavsky, A. (2018). Bayesian quadrature and energy minimization for space-filling design. *arXiv:1808.10722*.

Prüher, J., Karvonen, T., Oates, C. J., Straka, O., and Särkkä, S. (2018). Improved calibration of numerical integration error in sigma-point filters. *arXiv:1811.11474*.

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal*, 2(1):1–21.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Richardson, L. F. and Gaunt, J. A. (1927). The deferred approach to the limit. *Philosophical Transactions of the Royal Society of London. Series A*, 226(636-646):299–361.

Ritter, K. (2000). *Average-Case Analysis of Numerical Problems*, volume 1733 of *Lecture Notes in Mathematics*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Roininen, L., Girolami, M., Lasanen, S., and Markkanen, M. (2019). Hyperpriors for Matérn fields with applications in Bayesian inversion. *Inverse Problems & Imaging*, 13(1):1–29.

Sampson, P. and Guttorp, P. (1992). Nonparametric Estimation of Nonstationary Spatial Covariance Structure. *Journal of the American Statistical Association*, 87(417):108–119.

Santin, G., Haasdonk Communicated by De Rossi, B. A., and Francomano, E. (2017). Convergence rate of the data-independent P-greedy algorithm in kernel-based approximation. *Dolomites Research Notes on Approximation*, 10.

Schmidt, A. M. and O'Hagan, A. (2003). Bayesian Inference for Non-Stationary Spatial Covariance Structure via Spatial Deformations. *Journal of the Royal Statistical Society, Series B*, 65(3):743–758.

Sul'din, A. V. (1959). Wiener measure and its applications to approximation methods. I. *Izv. Vyssh. Uchebn. Zaved. Mat.*, 6(13):145–158.

Sul'din, A. V. (1960). Wiener measure and its applications to approximation methods. II. *Izv. Vyssh. Uchebn. Zaved. Mat.*, 5(18):165–179.