Bayesian Quadrature for Ratios

Michael A. Osborne
Department of Engineering Science
University of Oxford
Oxford OX1 3PJ, UK
mosb@robots.ox.ac.uk

Roman Garnett
Robotics Institute
Carnegie Mellon University
Pittsburgh PA 15213, USA
rgarnett@andrew.cmu.edu

Stephen J. Roberts
Department of Engineering Science
University of Oxford
Oxford OX1 3PJ, UK
sjrob@robots.ox.ac.uk

Christopher Hart
Department of Physics
University of Oxford
Oxford OX1 3RH, UK
{Christopher.Hart,

Suzanne Aigrain
Department of Physics
University of Oxford
Oxford OX1 3RH, UK
Suzanne.Aigrain,

Neale P. Gibson
Department of Physics
University of Oxford
Oxford OX1 3RH, UK
Neale.Gibson}@astro.ox.ac.uk

Abstract

We describe a novel approach to quadrature for ratios of probabilistic integrals, such as are used to compute posterior probabilities. This approach offers performance superior to Monte Carlo methods by exploiting a Bayesian quadrature framework. We improve upon previous Bayesian quadrature techniques by explicitly modelling the non-negativity of our integrands, and the correlations that exist between them. It offers most where the integrand is multi-modal and expensive to evaluate. We demonstrate the efficacy of our method on data from the Kepler space telescope.

1 Introduction

Bayesian inference often requires the evaluation of nonanalytic definite integrals. In the main, techniques for numerical integration estimate the integral given the value of the integrand on a set of sample points, a set that is limited in size by the computational expense of evaluating the integrand. As discussed in (O'HAGAN, 1987), traditional Monte Carlo integration techniques do not make the best possible use of this valuable information. An alternative is found in Bayesian quadrature (O'HAGAN, 1991), which uses these samples within a Gaussian process model to perform inference about the integrand. The analytic niceties of the Gaussian then permit inference to be performed about the integral itself, the ultimate object of our interest. However, this use of a Gaussian process comes at a cost: as the Gaussian has unbounded support, it cannot reflect the knowledge that the integrand is a non-negative probability. This means that this model will assign non-zero probability mass to negative probabilities, giving rise to misleading results. A second problem is encountered when we wish to estimate the ratio of two integrals with common terms, as is the case when we marginalise hyperparameters by evaluating the ratio of two integrals over the likelihood, as in

$$p(y|z) = \frac{\int p(y|z,\phi)p(z|\phi)p(\phi)\ \mathrm{d}\phi}{\int p(z|\phi)p(\phi)\ \mathrm{d}\phi}.$$

Here we are required to model the correlation that exists between the common terms in order to not overestimate the importance of samples in those terms.

We address the first of these problems by modeling the non-negative terms in our integrand with a Gaussian process on their logarithm. This, and the second of our problems, destroy the analytic results relied upon by previous formulations of Bayesian quadrature. We propose to linearise our ratio of integrals as a function of the terms in the integrand, around suitable 'best-fit' values. This gives us an algorithm, Bayesian Quadrature for Ratios, that on synthetic examples outperforms traditional Monte Carlo approaches. Our algorithm is also applied to real data drawn from the Kepler mission, where sophisticated inference is needed to model light curves given very noisy observations.

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume 22 of JMLR: W&CP 22. Copyright 2012 by the authors.

2 Gaussian Processes

Gaussian processes (GPs) offer a powerful method to perform Bayesian inference about functions (RASMUSSEN and WILLIAMS, 2006). A GP is defined as a distribution over the functions $f : \Phi \to \mathbb{R}$ such that the distribution over the possible function values on any finite subset of Φ is multivariate Gaussian. For a function $f(\phi)$, the prior distribution over its values $\boldsymbol{f}$ on a subset $\boldsymbol{\phi} \subset \Phi$ are completely specified by a mean vector $\boldsymbol{\mu}$ and covariance matrix K

$$
\begin{aligned}
p(\boldsymbol{f}|I) &:= \mathcal{N}\big(\boldsymbol{f}; \boldsymbol{\mu}_f, K_f\big) \\
&:= \frac{1}{\sqrt{\det 2\pi K_f}} \exp\big(-\frac{1}{2}\,(\boldsymbol{f}-\boldsymbol{\mu}_f)^{\mathsf{T}}\,K_f^{-1}\,(\boldsymbol{f}-\boldsymbol{\mu}_f)\big),
\end{aligned}
$$

where I, the *context*, forms the background knowledge upon which all our probabilities are conditioned. Its ubqiuity leads us to henceforth drop it from explicit representation for notational convenience. The context, I, includes prior knowledge of both the mean and covariance functions, which generate $\boldsymbol{\mu}_f$ and K_f respectively. The prior mean function is chosen as appropriate for the problem at hand (often a constant), and the covariance function is chosen to reflect any prior knowledge about the structure of the function of interest, for example periodicity or differentiability. In this paper, we'll use Gaussian covariance functions,

$$K_f(\phi_1, \phi_2) := h_f^2\, \mathcal{N}(\phi_1; \phi_2, w_f)\,. \tag{1}$$

Here h_f specifies the output scale ('height') over f, while w_f defines a (squared) input scale ('width') over ϕ. Note that ϕ itself may be multi-dimensional, in which case w_f must actually be a covariance matrix. Where this is true for the remainder of the paper, we'll take w_f as diagonal.

Let us assume we have observations $(\boldsymbol{\phi}_s, \boldsymbol{f}_s)$ and are interested in making predictions about the function values $f_\star$ at input $\phi_\star$. We will assume that knowledge of function inputs such as $\boldsymbol{\phi}_s$ and $\phi_\star$ is incorporated into I (and will hence usually be hidden). With this information, we have the predictive equations

$$p\big(f_\star|\boldsymbol{f}_s\big) = \mathcal{N}\big(f_\star; m_{f|s}(\phi_\star), V_{f|s}(\phi_\star)\big)\,,$$

where we have, for the mean $m(a|b) := \int a\, p\big(a|b\big) \mathrm{d}a$ and variance $V(a|b) := \int \big(a - m(a|b)\big)^2\, p\big(a|b\big) \mathrm{d}a$,

$$
\begin{aligned}
&m_{f|s}(\phi_\star) := m(f_\star|\boldsymbol{f}_s) \\
&= \mu_f(\phi_\star) + K_f(\phi_\star, \boldsymbol{\phi}_s) K_f(\boldsymbol{\phi}_s, \boldsymbol{\phi}_s)^{-1}\big(\boldsymbol{f}_s - \mu_f(\boldsymbol{\phi}_s)\big)
\end{aligned} \tag{2}
$$

$$
\begin{aligned}
&V_{f|s}(\phi_\star) := V(f_\star|\boldsymbol{f}_s) \\
&= K_f(\phi_\star, \phi_\star) - K_f(\phi_\star, \boldsymbol{\phi}_s) K_f(\boldsymbol{\phi}_s, \boldsymbol{\phi}_s)^{-1} K_f(\boldsymbol{\phi}_s, \phi_\star)\,.
\end{aligned}
$$

3 Bayesian Quadrature

Bayesian quadrature (O'HAGAN, 1991; RASMUSSEN and GHAHRAMANI, 2003) is a means of performing Bayesian inference about the value of a potentially nonanalytic integral

$$\langle f \rangle := \int f(\phi)\, p(\phi)\; \mathrm{d}\phi\,. \tag{3}$$

Note that we use a condensed notation; this and all integrals to follow are definite integrals over the entire domain of interest. We'll assume we are integrating with respect to a Gaussian prior

$$p(\phi) := \mathcal{N}(\phi; \nu_\phi, \lambda_\phi)\;, \tag{4}$$

although other convenient forms, or, if necessary, the use of an importance re-weighting trick, allow any other integral to be approximated (OSBORNE, 2010). If ϕ is a vector, ν_ϕ is a vector of identical size, and λ_ϕ an appropriate covariance matrix.

Quadrature involves evaluating $f(\phi)$ at a vector of sample points $\boldsymbol{\phi}_s$, giving $\boldsymbol{f}_s := f(\boldsymbol{\phi}_s)$. Of course, this evaluation is usually a computationally expensive operation. The resultant sparsity of our samples introduces uncertainty about the function f between them, and hence uncertainty about the integral $\langle f \rangle$.

We address the estimation of the value of our integral as a problem of Bayesian inference (O'HAGAN, 1992). In our case, both the values $f(\boldsymbol{\phi}_s)$ and their locations $\boldsymbol{\phi}_s$ represent valuable pieces of knowledge. As discussed by O'HAGAN (1987), traditional Monte Carlo, which approximates as

$$\langle f \rangle \simeq \frac{1}{|s|} \sum_{i=1}^{|s|} f(\phi_i)\,, \tag{5}$$

effectively ignores the information content of $\boldsymbol{\phi}_s$, leading to unsatisfactory behaviour[1].

We choose for f a GP prior with mean μ_f and the Gaussian covariance function (1). Here the scales h_f and w_f are *quadrature hyperparameters*, hyperparameters that specify the GP used for Bayesian quadrature. These quadrature hyperparameters, and the others that follow, will be fitted using maximum likelihood, and incorporated into the (hidden) context I.

Note that variables over which we have a multivariate Gaussian distribution are jointly Gaussian distributed with any projections of those variables. Because integration is a projection, we can hence use our computed

[1] For example, imagine that we had $|s| = 3$, and $\phi_1 = \phi_2$. In this case, the identical value $q(\phi_1) = q(\phi_2)$ will receive $^2/_3$ of the weight, whereas the equally useful $q(\phi_3)$ will receive only $^1/_3$.

samples $\boldsymbol{f}_s$ to perform analytic Gaussian process inference about the value of integrals over $f(\phi)$, such as $\langle f\rangle$. Our mean estimate for $\langle f\rangle$ given $\boldsymbol{f}_s$ is

$$\begin{aligned} m(\langle f\rangle|\boldsymbol{f}_s) &= \iint \langle f\rangle\, p(\langle f\rangle|f)\, p(f|\boldsymbol{f}_s)\, \mathrm{d}\langle f\rangle\, \mathrm{d}f \\ &= \iint \langle f\rangle\, \delta\Big(\langle f\rangle - \int f(\phi)\, p(\phi)\, \mathrm{d}\phi\Big) \\ &\qquad\qquad \mathcal{N}\big(f; m_{f|s}, V_{f|s}\big)\, \mathrm{d}\langle f\rangle\, \mathrm{d}f \\ &= \int m_{f|s}(\phi)\, p(\phi)\, \mathrm{d}\phi \\ &= \mu_f + \Upsilon_f(\boldsymbol{\phi}_s)^{\mathsf{T}}\, \tau_f(\boldsymbol{\phi}_s)\,, \end{aligned} \tag{6}$$

where for $\phi_i \in \boldsymbol{\phi}_s$,

$$\begin{aligned} \Upsilon_f(\phi_i) &:= \int K_f(\phi_i, \phi) p(\phi)\, \mathrm{d}\phi = h_f^2\, \mathcal{N}(\phi_i; \nu_\phi, \lambda_\phi + w_f) \\ \tau_f(\boldsymbol{\phi}_s) &:= K_f\big(\boldsymbol{\phi}_s, \boldsymbol{\phi}_s\big)^{-1}(\boldsymbol{f}_s - \mu_f)\,. \end{aligned}$$

Note that the form of our 'best estimate' for $\langle f\rangle$, (6), is an affine combination of the samples $\boldsymbol{f}_s$, just as for traditional quadrature or Monte Carlo techniques. Indeed, if μ_f is taken as the mean of $\boldsymbol{f}_s$ (as is usual for GP inference), the second term in (6) can be viewed as a correction factor to the Monte Carlo estimate (5). Note also that h_f represents a simple multiplicative factor to both $\Upsilon_f(\boldsymbol{\phi}_s)^{\mathsf{T}}$ and $K_f\big(\boldsymbol{\phi}_s, \boldsymbol{\phi}_s\big)$, and as such cancels out of (6). As per the appendix (OSBORNE, *et al.*, 2012), it is also possible to evaluate the Bayesian quadrature estimate for an integral over the product of functions.

4 Bayesian Quadrature for Ratios

We now present a novel approach to performing inference about probabilistic integrals. In inference, we are very commonly interested in making predictions about some variable $y(x)$, of which we receive potentially noise corrupted observations $z(x)$. This form encompasses problems of regression, tracking, classification and others. The prototypical example we consider is that of GP regression, e.g. in which x might be time and y an air temperature, although the algorithms described below are applicable to any choice of prediction model.

Associated with a matrix of inputs X_d (whose rows represent individual points in input space), we have a vector of observations $\boldsymbol{z}_d$; given these data, we are interested in inferring the value of y, $y_\star$, at a vector of inputs $\boldsymbol{x}_\star$. As with our convention above, we'll assume X_d and $\boldsymbol{x}_\star$ are always known and drop our probabilities' explicit dependence upon them. We have a prediction model $p\big(y_\star|\boldsymbol{z}_d, \phi\big)$, defined by hyperparameters $\phi \in \Phi$. These are commonly unknown *a priori* and under-determined by data, and hence must be marginalised. Therefore, we must evaluate

$$p\big(y_\star|\boldsymbol{z}_d\big) = \frac{\int p\big(y_\star|\boldsymbol{z}_d, \phi\big) p\big(\boldsymbol{z}_d|\phi\big) p(\phi)\ \mathrm{d}\phi}{\int p\big(\boldsymbol{z}_d|\phi\big) p(\phi)\ \mathrm{d}\phi}\,, \tag{7}$$

in which we have marginalised ϕ. We'll again assume the Gaussian prior (4) for $p(\phi)$, although, as before, other forms are possible. Typically, both our likelihood $p(\boldsymbol{z}_d|\phi)$ and predictions $p(y_\star|\boldsymbol{z}_d,\, \phi)$, treated as functions of ϕ, exhibit a functional form that renders these integrals non-analytic.

As such, we employ quadrature, evaluating

$$\begin{aligned} q(\phi) &:= p(y_\star|\boldsymbol{z}_d,\, \phi) \\ r(\phi) &:= p(\boldsymbol{z}_d|\phi) \end{aligned}$$

at samples $\boldsymbol{\phi}_s$, giving $\boldsymbol{q}_s := q(\boldsymbol{\phi}_s)$ and $\boldsymbol{r}_s := r(\boldsymbol{\phi}_s)$. Our evaluation of both q and r at the same vector of hyperparameter samples $\boldsymbol{\phi}_s$ is not absolutely necessary, but results in notational convenience and some computational savings, as we'll see later. Note that the more complex our model, and hence the greater the number of hyperparameters, the higher the dimension of the hyperparameter space we must sample in. The complexity of models we can practically consider is therefore limited by the curse of dimensionality. We put the problem of selecting the optimal sample locations $\boldsymbol{\phi}_s$ aside; we assume we have relatively useful samples, however obtained, and look to make optimal use of them.

Using such samples, traditional Monte Carlo would approximate as

$$\frac{\int q(\phi)\, r(\phi)\, p(\phi)\ \mathrm{d}\phi}{\int r(\phi)\, p(\phi)\ \mathrm{d}\phi} \simeq \frac{1}{|s|}\sum_{i=1}^{|s|} q(\phi_i)\,, \tag{8}$$

where, referring to (3) and (5), we substitute $q(\phi)$ for $f(\phi)$ and

$$p(\phi|\boldsymbol{z}_d) := \frac{r(\phi)\, p(\phi)}{\int r(\phi)\, p(\phi)\ \mathrm{d}\phi} \tag{9}$$

for $p(\phi|I)$. Note that the denominator integral renders (9) non-analytic, so that drawing samples from it requires some care (NEAL, 1993). We will now investigate the use of Bayesian Quadrature techniques for this problem.

We assign GP priors to the functions

$$\begin{aligned} \tilde{q} &:= \log(q) \\ \tilde{r} &:= \log(r)\,, \end{aligned}$$

with Gaussian covariances of the form (1). These choices of prior distribution are motivated by the fact that both q and r are strictly positive and possess a large dynamic range[2]. We now use these priors to perform inference about the functional $\varrho[\tilde{q},\tilde{r}]$, defined as

$$\varrho[\tilde{q},\tilde{r}] := p\big(y_\star|\tilde{q},\tilde{r},\boldsymbol{z}_d\big) = \frac{\int q(\phi)\,r(\phi)\,p(\phi)\;\mathrm{d}\phi}{\int r(\phi)\,p(\phi)\;\mathrm{d}\phi} \qquad (10)$$
$$= \frac{\int \exp\big(\tilde{q}(\phi)\big)\;\exp\big(\tilde{r}(\phi)\big)\,p(\phi)\;\mathrm{d}\phi}{\int \exp\big(\tilde{r}(\phi)\big)\,p(\phi)\;\mathrm{d}\phi}$$

with functional derivatives

$$\frac{\partial\varrho}{\partial\tilde{q}(\phi)}[\tilde{q},\tilde{r}] = \frac{\exp\big(\tilde{q}(\phi)\big)\;\exp\big(\tilde{r}(\phi)\big)\,p(\phi)}{\int \exp\big(\tilde{r}(\phi)\big)\,p(\phi)\;\mathrm{d}\phi}$$

and

$$\frac{\partial\varrho}{\partial\tilde{r}(\phi)}[\tilde{q},\tilde{r}] = \frac{\exp\big(\tilde{q}(\phi)\big)\;\exp\big(\tilde{r}(\phi)\big)\,p(\phi)}{\int \exp\big(\tilde{r}(\phi)\big)\,p(\phi)\;\mathrm{d}\phi} - \frac{\exp\big(\tilde{r}(\phi)\big)p(\phi)\int \exp\big(\tilde{q}(\phi)\big)\exp\big(\tilde{r}(\phi)\big)p(\phi)\,\mathrm{d}\phi}{\big(\int \exp\big(\tilde{r}(\phi)\big)p(\phi)\,\mathrm{d}\phi\big)^2}.$$

Note that $r(\phi)$ appears in both the numerator and denominator integrals of ϱ, introducing correlations between the values we estimate for them. For this reason, we must consider the ratio as a whole rather than performing inference for the numerator and denominator separately. This means that ϱ is not linear in our variables $\tilde{q}$ and $\tilde{r}$, unlike the $\langle f\rangle$ of (3), preventing the analytic stage of the inference process described in Section 3. As such, our introduction of the similarly non-linear transform exp does no real additional harm. This non-linearity means that we must perform inference about the functional $\varrho[\tilde{q},\tilde{r}]$ itself.

We make a *linearisation* approximation[3] for ρ, forcing ρ to be, as desired, affine in $\tilde{q}$ and $\tilde{r}$. Before proceeding, we introduce separate GP models over $q(\phi)$ and $r(\phi)$, the non-log functions. Then $m_{q|s}$ is the GP conditional mean (as per (2)) for q given observations $q(q_s)$. For these GPs (over the non-log quantities), we take zero prior means and Gaussian covariances of the form (1).

We perform the linearisation of $\rho[\tilde{q},\tilde{r}]$ around the point defined by $\tilde{q}_0 := \log(m_{q|s})$ and $\tilde{r}_0 := \log(m_{r|s})$. We make the definitions $\rho_0 := \rho[\tilde{q}_0,\tilde{r}_0]$, $\frac{\partial\rho_0}{\partial\tilde{r}(\phi)} := \frac{\partial\rho}{\partial\tilde{r}(\phi)}[\tilde{q}_0,\tilde{r}_0]$, and for future notational convenience, assume that if we condition on ϱ_0, we are also conditioning on its functional derivatives. This linearisation gives us a convenient mean for ϱ,

$$m(\varrho[\tilde{q},\tilde{r}]|\varrho_0) = \varrho_0 + \int \frac{\partial\varrho_0}{\partial\tilde{q}(\phi)}\big(\tilde{q}(\phi)-\tilde{q}_0(\phi)\big)\mathrm{d}\phi + \int \frac{\partial\varrho_0}{\partial\tilde{r}(\phi)}\big(\tilde{r}(\phi)-\tilde{r}_0(\phi)\big)\mathrm{d}\phi\,. \quad (11)$$

That is, we can now write

$$\varrho_0 := \varrho[\tilde{q}_0,\tilde{r}_0] = \frac{m(\langle qr\rangle|\boldsymbol{q}_s,\,\boldsymbol{r}_s)}{m(\langle r\rangle|\boldsymbol{r}_s)} \qquad (12)$$

along with its functional derivatives

$$\frac{\partial\varrho_0}{\partial\tilde{q}(\phi)} := \frac{\partial\varrho}{\partial\tilde{q}(\phi)}[\tilde{q}_0,\tilde{r}_0] = \frac{m(q(\phi)|\boldsymbol{q}_s)\;m(r(\phi)|\boldsymbol{r}_s)\;p(\phi)}{m(\langle r\rangle|\boldsymbol{r}_s)}$$
$$\frac{\partial\varrho_0}{\partial\tilde{r}(\phi)} := \frac{\partial\varrho}{\partial\tilde{r}(\phi)}[\tilde{q}_0,\tilde{r}_0] = \frac{m(r(\phi)|\boldsymbol{r}_s)\,p(\phi)}{m(\langle r\rangle|\boldsymbol{r}_s)}\left(m(q(\phi)|\boldsymbol{q}_s) - \frac{m(\langle qr\rangle|\boldsymbol{q}_s,\,\boldsymbol{r}_s)}{m(\langle r\rangle|\boldsymbol{r}_s)}\right).$$

Note that ϱ_0 and its functional derivatives are analytic; our choice of $\tilde{q}_0$ and $\tilde{r}_0$ allowed us to resolve the integrals required.

We now motivate our linearisation. Of course, we don't actually have access to entire $\tilde{q}$ and $\tilde{r}$ functions, over which we will have to marginalise. Imagine, for the purposes of building intuition, these functions being parameterised by the arbitrarily large but finite vectors of their values, $\tilde{\boldsymbol{q}}_f$ and $\tilde{\boldsymbol{r}}_f$, at the vector $\boldsymbol{\phi}_f = \{\phi_1,\ldots,\phi_{|f|}\}$. Our integrals over $\tilde{q}$ and $\tilde{r}$ can be arbitrarily well-represented by equivalent sums over these values $\tilde{\boldsymbol{q}}_f$ and $\tilde{\boldsymbol{r}}_f$, which we use to represent ϱ for any values of $\tilde{\boldsymbol{q}}_f$ and $\tilde{\boldsymbol{r}}_f$—inference for ϱ is now unnecessary. Hence for some $\boldsymbol{\alpha}_f$ and some $\boldsymbol{\beta}_f$

$$m(\varrho|\tilde{\boldsymbol{q}}_s,\,\tilde{\boldsymbol{r}}_s) \simeq \iint \frac{\sum_{i=1}^{|f|}\alpha_i\,e^{\tilde{q}_i}\,e^{\tilde{r}_i}}{\sum_{i=1}^{|f|}\beta_i\,e^{\tilde{r}_i}} \times\mathcal{N}\big(\tilde{\boldsymbol{q}}_f;m(\tilde{\boldsymbol{q}}_f|\tilde{\boldsymbol{q}}_s),C(\tilde{\boldsymbol{q}}_f|\tilde{\boldsymbol{q}}_s)\big) \times\mathcal{N}\big(\tilde{\boldsymbol{r}}_f;m(\tilde{\boldsymbol{r}}_f|\tilde{\boldsymbol{r}}_s)\,,C(\tilde{\boldsymbol{r}}_f|\tilde{\boldsymbol{r}}_s)\big)\mathrm{d}\tilde{\boldsymbol{q}}_f\,\mathrm{d}\tilde{\boldsymbol{r}}_f \qquad (13)$$

The multivariate Gaussians over $\tilde{\boldsymbol{q}}_f$ and $\tilde{\boldsymbol{r}}_f$ significantly restrict the volume of function space that needs to be integrated over. The other term in our integrand is a ratio of weighted sums of $\exp(\tilde{\boldsymbol{q}})$ and $\exp(\tilde{\boldsymbol{r}})$. For $1 \le i \le |f|$, the Gaussian over $\tilde{q}_i$ will vary as $\exp(-\tilde{q}_i^2)$, where the ratio of sums will vary only as $\exp(\tilde{q}_i)$, and similarly for $\tilde{r}_j$. This forms the justification for our linearisation approach—around the narrow peaks of our GPs, the slow variation of our ratio of integrals can be well approximated as affine. We can visualise this fact in two dimensions, as per Figure 1.

[2] In practice, we use the transform $\log\left({}^{r(\phi)}/_{\gamma_r}+1\right)$ where $\gamma_r := 100\max(\boldsymbol{r}_s)$. This give better resolution of some numerical issues, and allows us to assume the transformed quantity has zero mean. For the sake of simplicity, we leave further technical details to the appendix (OSBORNE, *et al.*, 2012).

[3] Note that this linearisation is equivalent to taking another GP for ρ, with the affine covariance $K_\varrho\big[(\tilde{q},\tilde{r}),(\tilde{q}',\tilde{r}')\big] := \int\tilde{q}(\phi)\tilde{q}'(\phi)\mathrm{d}\phi + \int\tilde{r}(\phi)\tilde{r}'(\phi)\mathrm{d}\phi + \omega^2$

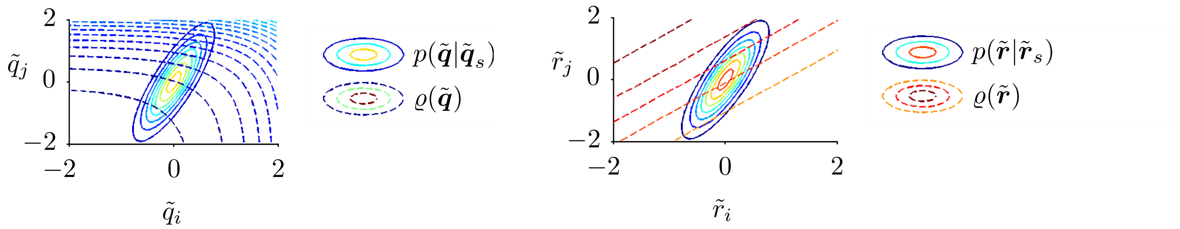


Figure 1: A comparison of bivariate Gaussians ($p(\tilde{\boldsymbol{q}}|\tilde{\boldsymbol{q}}_s)$ and $p(\tilde{\boldsymbol{r}}|\tilde{\boldsymbol{r}}_s)$) and a ratio of weighted sums of exponentials of those two variables (ϱ); both terms as per the integrand of (13). It can be seen that around the narrow peaks of either Gaussian, the ratio can be reasonably approximated as linear.

We now define

$$\Delta_{\tilde{q}} := m_{\tilde{q}|s} - \tilde{q}_0 = m_{\tilde{q}|s} - \log(m_{q|s})$$
$$\Delta_{\tilde{r}} := m_{\tilde{q}|s} - \tilde{r}_0 = m_{\tilde{r}|s} - \log(m_{r|s})\,,$$

the differences between the GP means over our transformed quantities and the transformed GP means over original quantities. We expect $\Delta_{\tilde{r}}$ to be small, as per figure 2. This implies that $\tilde{r}_0$ is close to the peaks of our Gaussians over $\tilde{r}$, rendering our linearisation appropriate.

We now return to our predictive posterior;

$$\begin{aligned} & p(y_\star|\varrho_0, \Delta_{\tilde{q}}, \Delta_{\tilde{r}}, \tilde{\boldsymbol{q}}_s, \tilde{\boldsymbol{r}}_s, \boldsymbol{z}_d) \\ &= \iiint p(y_\star|\tilde{q}, \tilde{r}, \boldsymbol{z}_d)\, p(\varrho|\varrho_0, \tilde{q}, \tilde{r}) \\ &\qquad \times p(\tilde{q}|\tilde{\boldsymbol{q}}_s)\, p(\tilde{r}|\tilde{\boldsymbol{r}}_s)\, \mathrm{d}\varrho\, \mathrm{d}\tilde{q}\, \mathrm{d}\tilde{r} \\ &= \iiint \varrho[\tilde{q}, \tilde{r}]\, p(\varrho|\varrho_0, \tilde{q}, \tilde{r}) \\ &\qquad \times \mathcal{N}\big(\tilde{q}; m_{\tilde{q}|s}, C_{\tilde{q}|s}\big)\, \mathcal{N}\big(\tilde{r}; m_{\tilde{r}|s}, C_{\tilde{r}|s}\big)\, \mathrm{d}\varrho\, \mathrm{d}\tilde{q}\, \mathrm{d}\tilde{r} \\ &= \iint m(\varrho[\tilde{q}, \tilde{r}]|\varrho_0) \\ &\qquad \times \mathcal{N}\big(\tilde{q}; m_{\tilde{q}|s}, C_{\tilde{q}|s}\big)\, \mathcal{N}\big(\tilde{r}; m_{\tilde{r}|s}, C_{\tilde{r}|s}\big)\, \mathrm{d}\tilde{q}\, \mathrm{d}\tilde{r} \\ &= m\big(\varrho[m_{\tilde{q}|s}, m_{\tilde{r}|s}]\big|\varrho_0\big) \\ &= \varrho_0 + \int \frac{\partial \varrho_0}{\partial \tilde{q}(\phi)} \Delta_{\tilde{q}}(\phi) \mathrm{d}\phi + \int \frac{\partial \varrho_0}{\partial \tilde{r}(\phi)} \Delta_{\tilde{r}}(\phi) \mathrm{d}\phi. \quad (14) \end{aligned}$$

As with any linearisation approximation, this final estimate is the value at the selected point $(\tilde{q}_0, \tilde{r}_0)$, plus two correction factors modelling the influence of the first derivatives. These correction factors contain a further two non-analytic integrals, over ϕ. Fortunately, they're not too far away from being analytic; almost all terms with dependence on ϕ within those integrals are Gaussian. The exceptions are the $\log(m_{q|s})$ and $\log(m_{r|s})$ terms within $\Delta_{\tilde{q}}$ and $\Delta_{\tilde{r}}$. As such, we perform another stage of Bayesian quadrature by treating $\epsilon_{rq} := m_{r|s}\,\Delta_{\tilde{q}}$, $\epsilon_{rr} := m_{r|s}\,\Delta_{\tilde{r}}$ and $\epsilon_{qr} := m_{q|s}\,\Delta_{\tilde{r}}$ as unknown functions of ϕ. For these functions we take Gaussian process priors with zero prior mean and Gaussian covariance (1). The choice of this prior is informed by Figure 2b, which is smooth and possesses a zero mean. Note that the choice to place a GP on ϵ_{rq} rather than directly on $m_{r|s}$, $\log(m_{q|s})$ or similar quantities simplifies our analysis and improves numerical performance.

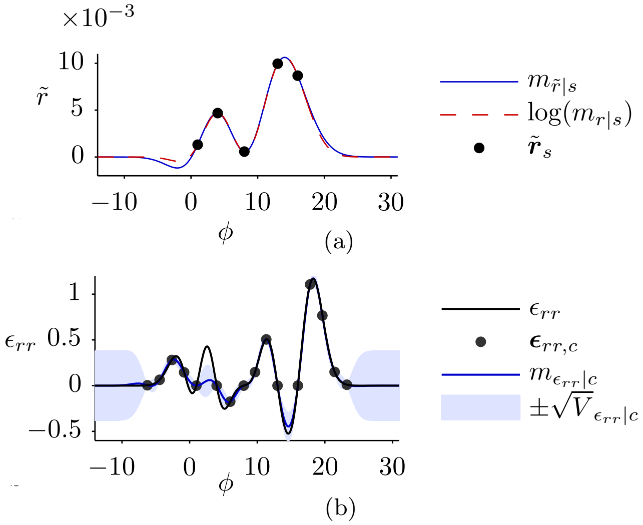


Figure 2: (a) Note that $\log(m_{r|s})$ resembles $m_{\tilde{r}|s}$; (b) ϵ_{rr} represents the difference between the two functions, scaled by $m_{r|s}$.

We must now choose sample points $\boldsymbol{\phi}_c$ at which to evaluate our ϵ functions. For simplicity, and due to their similar structure, we choose a common vector $\boldsymbol{\phi}_c$ for all such functions, giving us observations $\boldsymbol{\epsilon}_{rq,c}$, $\boldsymbol{\epsilon}_{rr,c}$ and $\boldsymbol{\epsilon}_{qr,c}$. $\boldsymbol{\phi}_c$ should firstly include $\boldsymbol{\phi}_s$, at which points we know that ϵ is equal to zero. Note firstly that a simple heuristic for determining the 'best' samples (in the sense of samples with which to fit a GP) is to select those samples at extrema. With reference to Figure 2b, note that the peaks and troughs of ϵ occur at points far-removed from $\boldsymbol{\phi}_s$, but no further away than a few input scales. This is where the transformed mean for q or r is liable to differ most from the mean of the transformed variable (and the appropriate Δ extremised). That is, we would like to select $\boldsymbol{\phi}_c$ as points as far away from all points in $\boldsymbol{\phi}_s$ as possible, while still remaining within a few input scales. Finding $\boldsymbol{\phi}_c$ hence requires solving the largest empty sphere problem, which we can solve with the use of a Voronoi diagram (VORONOI, 1907; SHAMOS and HOEY, 1975; OKABE and SUZUKI, 1997). Where this requires computation in excess of that afforded by our allowance, we instead construct a KD-tree (BENTLEY, 1975) for our $\boldsymbol{\phi}_s$, and select $\boldsymbol{\phi}_c$ as the centres of the hyper-rectangles defined by the splitting hyperplanes of the tree.

By marginalising over the unknown ϵ functions, then, we arrive at our final posterior;

$$p(y_\star|\boldsymbol{\epsilon}_{rq,c}, \boldsymbol{\epsilon}_{rr,c}, \boldsymbol{\epsilon}_{qr,c}, \varrho_0, \tilde{\boldsymbol{q}}_s, \tilde{\boldsymbol{r}}_s, \boldsymbol{z}_d) = \varrho_0 + C_{\tilde{q}} + C_{\tilde{r}} \,,$$

where the two correction factors are

$$\begin{aligned} C_{\tilde{q}} &:= \frac{m(\langle q\epsilon_{rq}\rangle|\boldsymbol{q}_s, \boldsymbol{\epsilon}_{rq,c})}{m(\langle r\rangle|\boldsymbol{r}_s)} \\ C_{\tilde{r}} &:= \frac{1}{m(\langle r\rangle|\boldsymbol{r}_s)} \times \Bigg(m(\langle q\epsilon_{rr}\rangle|\boldsymbol{q}_s, \boldsymbol{\epsilon}_{rr,c}) \\ &\qquad - m(\langle \epsilon_{rr}\rangle|\boldsymbol{\epsilon}_{rr,c}) \, \frac{m(\langle qr\rangle|\boldsymbol{q}_s, \boldsymbol{r}_s)}{m(\langle r\rangle|\boldsymbol{r}_s)} \Bigg) \,. \end{aligned}$$

Note that if we wish to compute the posterior mean for $y_\star$, we need merely redefine $q(\phi)$ as $m(y_\star|\boldsymbol{z}_d, \phi)$. Of course, as the mean is not usually strictly non-negative, we can place a GP on q directly, hence requiring no correction factor for $\tilde{q}$. Similarly, we can compute the posterior variance for $y_\star$ by redefining $q(\phi)$ as the second moment $\int y_\star^2 \, p(y_\star|\boldsymbol{z}_d, \phi) \, \mathrm{d}y_\star$. The posterior variance is then given by (14) minus the squared posterior mean.

5 Empirical Evaluation

We now turn to an empirical evaluation of our novel algorithm, BQR, built around the approach outlined in section 4. Explicitly, we used (14) to compute the posterior density for a predictant $y_\star$, and appropriate modifications to compute its mean and variance.

We compared our method against a number of alternatives. Firstly, we use maximum likelihood (ML), which approximates the likelihood function $r(\phi)$ as the Dirac delta function $\delta(\phi - \phi_m)$, where ϕ_m is the hyperparameter sample with maximal likelihood. Clearly, this approach is inappropriate for multimodal likelihoods. Next, we use traditional Monte Carlo, MC, which uses (8) to estimate our ratio of integrals. We also compare against Naïve Bayesian Quadrature, NBQ, in which both denominator and numerator integrals in (7) are treated independently and the product $q(\phi)r(\phi)$ in the numerator modelled using a single GP. This allows us to quantify the influence of modelling correlations between numerator and denominator integrals. We also test against what we call BQZ, the algorithm identical to BQR except that the first derivatives of ϱ_0 are assumed to be zero. This allows us to evaluate the significance of the correction factors due to our linearisation, which are assumed zero for BQZ. As such, BQZ does not make use of the logarithmic transform.

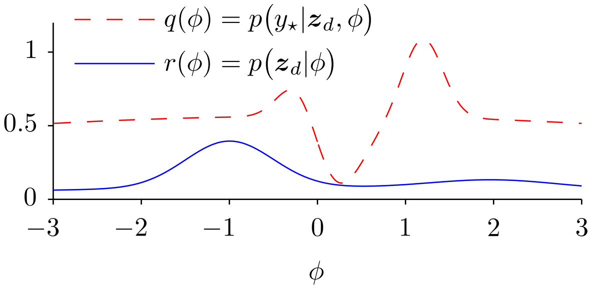


Figure 3: Test functions used for a one-dimensional integration example.

Table 1: RMSE for synthetic one-dimensional integral.

	ML	MC	NBQ	BQZ	BQR
slice	0.0120	0.0184	0.0167	0.0044	**0.0043**
HMC	0.0120	0.0217	0.0203	0.0046	**0.0042**

The samples $\boldsymbol{\phi}_s$ required to perform quadrature were obtained using Markov Chain Monte Carlo methods, to simulate samples from the posterior for ϕ, (9). In particular, we used Hybrid Monte Carlo (HMC) (DUANE, *et al.*, 1987) and Slice Sampling (NEAL, 2003). While techniques exist (MINKA, 2000) to obtain samples more suitable for Bayesian quadrature, we wished to give MC every possible advantage relative to Bayesian alternatives.

The diagonal input scale matrices w_q and w_r (giving a number of quadrature hyperparameters equal to twice the dimension of ϕ) were found by maximising their respective marginal likelihoods using a multi-start gradient-ascent method. The input scales $w_{\tilde{r}}$ were taken as identical to w_r (and similarly $w_{\tilde{q}} = w_q$), and the input scales for all ϵ quantities were taken as half of w_r: these choices worked well in practice (see figure 2). We have stronger prior information for our prior means μ_q and μ_r, which were also found to have a great influence on ultimate performance. Firstly, likelihood functions are typically highly peaked and localised around ϕ_m, and so we take a zero mean for μ_r. Given this, our integrals are usually dominated by $r(\phi_m)$; we took μ_q as $q(\phi_m)$ so that regression for q was most accurate near this critical value.

Our first experiment was a synthetic, one-dimensional example for the q and r functions depicted in figure 3, and for a zero mean, unit variance Gaussian prior $p(\phi)$. These functions were created using a mixture of Gaussians, such that we could determine the exact result (0.5709) for our ratio of integrals (10). We then compared the root mean squared error (RMSE) between the estimates produced by our various methods and this exact result. Table 1 tabulates the scores over the last 200 samples (thus permitting a 50 point 'burn-in') and 100 trial sample chains, and figure 4 the results as a function of the number of samples. Our methods comfortably outperform MC and NBQ, and for the majority of the sample history our correction factors give a small improvement. For clarity, ML results were not plotted; its RMSE plateaus once ϕ_m is found, typically about 10 samples in. On the basis of these results, we choose to perform solely slice sampling (of up to 500 samples) henceforth, to again favour MC relative to our methods.

We now consider two examples of GP regression, in which we must marginalise over hyperparameters ϕ. For GPs, computing a single likelihood $r(\phi)$ requires the computationally onerous inversion (or finding the Cholesky factor) of a covariance matrix of size equal to the number of data, D. Hence evaluating N hyperparameter samples takes a considerable quantity of time: $O(D^3N)$. In comparison, the cost of evaluating BQR (dominated by the cost of finding the Cholesky factor of a covariance matrix of size equal to N) is typically modest, $O(N^3)$. Having at great expense evaluated all our hyperparameter samples, it seems prudent to use them in the most intelligent way possible.

For both examples, we took independent Gaussian priors (such that λ_ϕ is diagonal) for the various hyperparameters of the model, each with zero mean and a variance of four. We first tested on a synthetic regression example drawn from FRIEDMAN (1991) and RASMUSSEN and GHAHRAMANI (2003), with eight hyperparameters to marginalise. Specifically, we used the function

$$f(x_1, x_2, x_3, x_4, x_5) := 10\sin(\pi x_1 x_2) + 20(x_3 - 1/2) + 10x_4 + 5x_5$$

with zero mean, unit variance Gaussian noise. We performed GP regression for 100 test points given 100 observations, all independently drawn from the uniform $[0,1]^5$ distribution; we marginalised five input scales, an output scale, a noise variance and a prior mean. We evaluated the RMSE of the predictive means produced by our various methods both as a function of the number of samples, displayed in figure 5a, and over all but the first 50 samples, listed in table 2. It can be seen that all non-MC methods perform roughly equally, due to the presence of a strong dominant peak in the likelihood function.

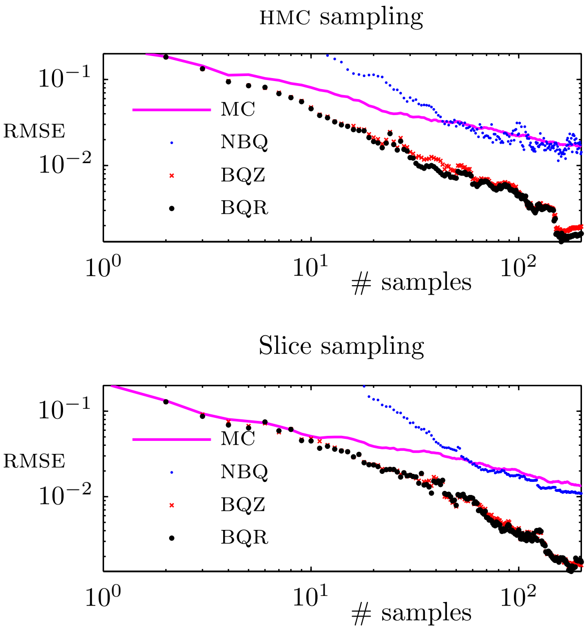
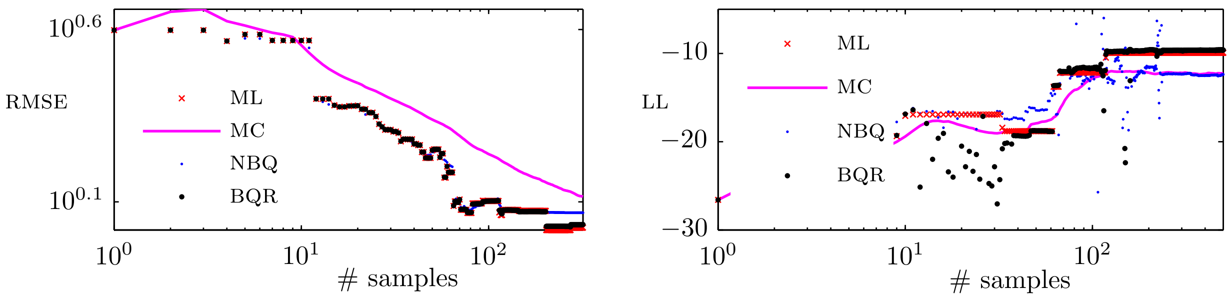


Figure 4: Error in estimates for one-dimensional integral as a function of the number of samples.

We also tested on light curves from the first month of operation of the Kepler mission (BORUCKI, *et al.*, 2011). Here we are required to regress flux from a star, as a function of time, in order to infer the rotation rate and other properties of dark spots on the star's surface. For this, we used a GP model with a constant mean and decaying-periodic covariance, giving six hyperparameters in total to marginalise. The data is corrupted by non-trivial noise mechanisms, and the final results are sensitive to our regression, so it is important that correct inference about our model hyperparameters is performed. A large number of datasets exist; we choose one in particular for which the likelihood function is highly multi-modal. Give the lack of ground truth for this dataset, we evaluated predictive performance by splitting the data into 151 point training ($\boldsymbol{z}_d$) and testing ($\boldsymbol{z}_\star := \{z_1, \ldots, z_{|\star|}\}$) vectors, and computed LL $= \sum_{i=1}^{|\star|} \log \mathcal{N}(z_i; m_i, C_i)$ for the predictive means $\mathbf{m}$ and variances $\mathbf{C}$ produced by each method. This allows us to evaluate the the quality of our predictive uncertainties. LL is plotted in figure 5a as a function of the number of samples, and over all but the first 50 samples, listed in table 3. It can be seen that BQR is best able to cope with the complicated, multimodal likelihood surface.

Our algorithm, BQR was the most broadly successful of those tested in these experiments. Note that ML is completely unqualified to represent the full posterior: for GP prediction, for example, it will always provide a unimodal Gaussian posterior. In contrast,



(a)

(b)

Figure 5: Performance for GP regression on (a) synthetic and (b) Kepler data. BQZ not plotted for clarity.

Table 2: RMSE for GP prediction on synthetic data.

ML	MC	NBQ	BQZ	BQR
1.1722	1.5551	1.2246	1.1805	1.1805

Table 3: LL for GP prediction on Kepler data.

ML	MC	NBQ	BQZ	BQR
-10.448	-12.563	-12.948	-10.310	**-10.262**

the other methods, by mixing the Gaussian predictions from samples with different hyperparameters, can capture more complicated distributions. This means that while ML is able to produce effective mean predictions, it is prone to underestimates of predictive variances, such as lead it to a worse overall fit on the real Kepler data. MC's wasteful use of samples means that its estimates converge more slowly than Bayesian alternatives, despite our use of sampling methods explicitly designed to meet its needs. NBQ estimates exhibit occasional wild fluctations, largely due to the poor conditioning of a covariance matrix over hyperparameter samples that are excessively similar. This leads its GP to assign excessive probability mass to negative likelihoods. Due to our approximations, BQR is not immune to this problem, but our approach does render it significantly more robust: its correction factors ameliorate this effect in trying to force the integrand to be positive. Our modelling of the correlations between integrals over $r(\phi)$ also grant BQR superior performance for multi-modal, heavy-tailed likelihoods.

We conclude that constraining functions to be positive was overall probably less significant than dealing with the correlations in the numerator and denominator, given the relative performances of NBQ (which does not model correlations), BQZ (which does not make use of the log transform) and BQR.

6 Conclusions

Our algorithm, BQR, outperformed competitors in real and synthetic tasks requiring the numerical computation of posterior probabilities. We have successfully demonstrated that it is possible to use Bayesian methods to resolve the questions of approximation required to perform Bayesian inference. In particular, we have demonstrated the worth of acknowledging relevant prior information: here, that our integrands are non-negative and correlated.

In testing, we have focused on small numbers of samples, suitable for applications where evaluating the integrand is computationally demanding. This is due to the not-insignificant computational burden imposed by Bayesian quadrature's requirement to find the Cholesky factor of covariance matrices of size equal to the number of samples. This difficulty can be somewhat eased by integrating sparse GP methods (QUIÑONERO-CANDELA and RASMUSSEN, 2005; SNELSON and GHAHRAMANI, 2006; WALDER, *et al.*, 2008; LÁZARO-GREDILLA, *et al.*, 2010), which approximate the covariance matrix as sparse, an avenue we would like to investigate. We could also investigate the minor modification of our approach for general integration tasks where the integrand is non-negative, to compute, for example, marginal likelihoods.

7 Acknowledgments

The Kepler data used in this work were obtained from the Multimission Archive at the Space Telescope Science Institute (MAST). We thank the entire Kepler team for their many years of hard work. M.A.O. was funded by the ORCHID project (`http://www.orchid.ac.uk/`). N.P.G. and S.A. acknowledge support from STFC grant ST/G002266/2. We would like to thank David Duvenaud, Jan Calliess and anonymous reviewers for many helpful comments.

References

Bentley, J. (1975). Multidimensional binary search trees used for associative searching. *Communications of the ACM*, vol. 18, no. 9, pp. 509–517.

Borucki, W., Koch, D., Basri, G., Batalha, N., Boss, A., Brown, T., Caldwell, D., Christensen-Dalsgaard, J., Cochran, W., DeVore, E., *et al.* (2011). Characteristics of Kepler planetary candidates based on the first data set. *The Astrophysical Journal*, vol. 728, p. 117.

Duane, S., Kennedy, A., Pendleton, B. and Roweth, D. (1987). Hybrid Monte Carlo. *Physics letters B*, vol. 195, no. 2, pp. 216–222.

Friedman, J.H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, vol. 19, no. 1, pp. pp. 1–67. ISSN 00905364. Available at: `http://www.jstor.org/stable/2241837`

Lázaro-Gredilla, M., Quiñonero-Candela, J., Rasmussen, C. and Figueiras-Vidal, A. (2010). Sparse spectrum Gaussian process regression. *The Journal of Machine Learning Research*, vol. 11, pp. 1865–1881.

Minka, T.P. (2000). Deriving quadrature rules from Gaussian processes. Tech. Rep., Statistics Department, Carnegie Mellon University.

Neal, R. (2003). Slice sampling. *Annals of Statistics*, pp. 705–741.

Neal, R.M. (1993). Probabilistic inference using Markov chain Monte Carlo methods. Tech. Rep. CRG-TR-93-1, University of Toronto. Available at: `http://www.cs.toronto.edu/~R./ftp/review.pdf`

O'Hagan, A. (1987). Monte Carlo is fundamentally unsound. *The Statistician*, vol. 36, pp. 247–249.

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

O'Hagan, A. (1992). Some Bayesian numerical analysis. In: Bernardo, J.M., Berger, J.O., Dawid, A.P. and Smith, A.F.M. (eds.), *Bayesian Statistics 4*, pp. 345–363. Oxford University Press.

Okabe, A. and Suzuki, A. (1997). Locational optimization problems solved through Voronoi diagrams. *European Journal of Operational Research*, vol. 98, no. 3, pp. 445–456.

Osborne, M.A. (2010). *Bayesian Gaussian Processes for Sequential Prediction, Optimisation and Quadrature.* Ph.D. thesis, University of Oxford. Available at `www.robots.ox.ac.uk/~mosb/full_thesis.pdf`.

Osborne, M.A., Garnett, R., Roberts, S.J., Hart, C., Aigrain, S. and Gibson, N.P. (2012). Bayesian quadrature for ratios: Now with even more Bayesian quadrature. Available at: `http://www.robots.ox.ac.uk/~mosb/papers/BQ_aistats_appendix.pdf`

Quiñonero-Candela, J. and Rasmussen, C. (2005). A unifying view of sparse approximate Gaussian process regression. *The Journal of Machine Learning Research*, vol. 6, pp. 1939–1959.

Rasmussen, C.E. and Ghahramani, Z. (2003). Bayesian Monte Carlo. In: Becker, S. and Obermayer, K. (eds.), *Advances in Neural Information Processing Systems*, vol. 15. MIT Press, Cambridge, MA.

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

Shamos, M. and Hoey, D. (1975). Closest-point problems. In: *Foundations of Computer Science, 1975., 16th Annual Symposium on*, pp. 151–162. IEEE.

Snelson, E. and Ghahramani, Z. (2006). Sparse Gaussian processes using pseudo-inputs. *Advances in neural information processing systems*, vol. 18, p. 1257.

Voronoi, G. (1907). Nouvelles applications des paramètres continus à la théorie des formes quadratiques. *Journal für die Reine und Angewandte Mathematik*, vol. 133, pp. 97–178.

Walder, C., Kim, K. and Schölkopf, B. (2008). Sparse multiscale Gaussian process regression. In: *Proceedings of the 25th international conference on Machine learning*, pp. 1112–1119. ACM.