An Improved Bayesian Framework for Quadrature

Henry Chai **Roman Garnett**
Department of Computer Science & Engineering
Washington University in St. Louis
St. Louis, MO 63130
hchai@wustl.edu garnett@wustl.edu

Abstract

We present an improved Bayesian quadrature framework for estimating intractable integrals of constrained integrands. We derive the necessary approximation scheme for the use of the log transformation in this framework when the integrand is constrained to be non-negative. We also present a method for optimizing the hyperparameters associated with this framework in the original space of the integrand as opposed to in the transformed space. We demonstrate that this framework significantly improves upon the performance of previous Bayesian quadrature methods in terms of wall-clock time on both a synthetic and a real-world example.

1 Introduction

One problem that frequently arises when using Bayesian machine learning methods is estimating intractable integrals of the form $Z = \int f(x)\pi(x)\,\mathrm{d}x$ where $f(x)$ is a likelihood and $\pi(x)$ is a prior. Note that both quantities are probabilities and thus are always non-negative. Integrals of this type appear when performing model selection, the target application of this work. Model selection is a fundamental problem that naturally arises in the course of scientific inquiry: given a set of candidate models, how does one determine which model best explains an observed data set? Using Bayesian methods to address this question requires calculating integrals of the above form.

Many commonly used techniques to estimate such integrals rely on Monte Carlo estimators [1, 2, 3]. These techniques are agnostic to prior information about the integrand, such as non-negativity, and also converge slowly in terms of the number of required samples of the integrand, making them ill-suited for settings where the integrand is expensive. One alternative to these techniques is Bayesian quadrature (BQ) [4, 5, 6], which maintains a probabilistic belief on the integrand. This belief allows BQ to exploit information-theoretic principles to actively select informative sample locations, increasing sample efficiency. Recent work by Gunter et al. [7] improves upon the speed and accuracy of the standard BQ algorithm (SBQ) [6] by incorporating non-negativity information about the integrand in model selection problems. Their method models the square root of the integrand instead of the integrand itself; "undoing" this transformation softly incorporates the constraint.

While previous work [7, 8] has shown that specific algorithms can outperform Monte Carlo based methods and SBQ in a variety of settings where the integral is non-negative, a general framework for performing quadrature involving a broader class of constrained integrands has never been offered. Our contribution is to define such a framework. We also develop an instantiation of this framework that addresses some shortcomings of previous work. Specifically, our instantiation allows for integrands with wider dynamic ranges than the method in [7] and does not require conditioning on randomly-sampled candidate points as in [8]. Lastly, we develop a novel procedure for this framework whereby the associated hyperparameters are set by maximizing the marginal likelihood of true observations of the integrand. All previous work that used a transformation to exploit a priori information fit hyperparameters by maximizing the marginal likelihood of transformed observations. We demonstrate that doing so can lead to undesirable behavior and that our procedure gives a better-behaved model.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

2 Background

Given a finite set of models $\mathcal{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_m\}$, the Bayesian approach to model selection is to define a likelihood function for each model $P_{\mathcal{M}_i}\{D \mid \theta_i\}$, that depends on a vector of parameters θ_i. If the prior density of θ_i is given by $P\{\theta_i\}$, then the model evidence of $\mathcal{M}_i$ can be expressed as: $Z_i = P_{\mathcal{M}_i}\{D\} = \int P_{\mathcal{M}_i}\{D \mid \theta_i\} P\{\theta_i\} \, \mathrm{d}\theta_i$. Given an intractable integral of the form $Z = \int f(x)\pi(x) \, \mathrm{d}x$, SBQ operates by placing a Gaussian process (GP) prior on the function $f(x)$. GPs can be thought of as probability distributions over functions, where the joint distribution of any finite number of function values is multivariate normal. GPs are parametrized by a mean function $m(x)$ and a covariance function $K(x, x')$. Given a set of observations at locations $x_D = \{x_1, \ldots, x_n\}$ with corresponding values $f(x_D)$, a GP prior can be conditioned on these observations to arrive at a posterior GP with mean $m_D(x) = m(x) + K(x, x_D)K(x_D, x_D)^{-1}(f(x_D) - m(x_D))$ and covariance $K_D(x, x') = K(x, x') - K(x, x_D)K(x_D, x_D)^{-1}K(x_D, x')$. For more on GPs, see [9].

If one has a GP belief on the integrand of an intractable integral, then the posterior mean and variance of the integral can be derived using the fact that GPs are closed under the evaluation of linear functionals such as integration [6]. Specifically, if $f \sim \mathcal{GP}(m, K)$, then:

$$Z = \int f(x)\pi(x) \, \mathrm{d}x \sim N\bigg(\int m(x)\pi(x) \, \mathrm{d}x, \iint K(x, x')\pi(x)\pi(x') \, \mathrm{d}x \, \mathrm{d}x' \bigg). \tag{1}$$

Warped sequential active Bayesian integration (WSABI) [7] builds off of the BQ framework and incorporates non-negativity information by modelling the square root of the integrand. Let $g(x) = \sqrt{2(f(x) - \alpha)}$, so $f(x) = \alpha + {}^1\!/_2 g(x)^2$ for some small positive constant α. WSABI places a GP prior on g and conditions this prior on observations to arrive at a posterior, much like SBQ. However, when calculating a belief on Z, the key property of closure of GPs under linear functionals is not available using this method because the marginal distribution of any $f(x)$ is a non-central χ^2 distribution. Thus, WSABI has to approximate the posterior as a GP. Gunter et al. [7] propose two approximation schemes: linearization, which uses a first-order Taylor expansion around the posterior mean of the GP on $g(x)$, and moment-matching, which calculates the mean and covariance of the true posterior distribution and defines a GP with these values. For more details, see [7].

3 An improved framework for quadrature with constrained integrands

All of the methods described above can be viewed as instantiations of the following general framework for estimating intractable integrals of constrained integrands where the constraint on the integrand can be described as a warping of $\mathbb{R}$. Let $Z = \int f(x)\pi(x) \, \mathrm{d}x$ be the integral of interest.

1. Determine a warping function ξ such that ξ maps from $\mathbb{R}$ to the range of f. Let $g(x) = \xi^{-1}(f(x))$. Place a GP prior on $g \sim \mathcal{GP}(\mu, \Sigma)$.
2. Calculate the mean and variance of the resulting probabilistic belief on f (using the first and second moments): $m(\mu(x), \Sigma(x, x))$ and $K(\mu(x), \mu(x'), \Sigma(x, x'))$. Approximate the probabilistic belief on f by a GP: $f \sim \mathcal{GP}(m, K)$.
3. Iterate until the budget of evaluations is expended:
   (a) Select a location to sample using uncertainty sampling, which selects the point that maximizes the posterior variance of f: $x^* = \arg\max_x K_D(x, x)$, and observe $g(x^*)$.
   (b) Update the posterior belief $g \sim \mathcal{GP}(\mu_D, \Sigma_D)$, fitting the hyperparameters associated with the GP on g by maximizing the marginal likelihood of the observations of f, using the approximate GP belief on f.
4. Calculate the belief about the value of the integral given all observations D: $Z \sim N(\int m_D(x)\pi(x) \, \mathrm{d}x, \iint K_D(x, x')\pi(x)\pi(x') \, \mathrm{d}x \, \mathrm{d}x')$.

3.1 A moment-matched approximation for the log transformation

Although [7] considers the square root transformation, previous related work has also considered the log transformation [8]. However, in [8], the authors only derive a linearized approximation for the posterior GP. We present the moment-matched approximation: let $g(x) = \log f(x)$ so that

$f(x) = \exp g(x)$ After observing $D = \{x_D, g(x_D)\}$, let $g \sim \mathcal{GP}(\mu_D(x), \Sigma_D(x, x'))$. The mean and variance of the true posterior of f, a (confusingly named) log-normal distribution, are:

$$m_D(x) = \exp\big(\mu_D(x) + {}^1\!/_2\Sigma_D(x, x')\big),\ K_D(x, x') = m_D(x)\big(\exp\big(\Sigma_D(x, x')\big) - 1\big)m_D(x') \quad (2)$$

3.2 Hyperparameter optimization

Whenever an inference method uses GPs, an important consideration is how to set the associated hyperparameters. One commonly used method is to optimize the marginal likelihood of the observed data as a function of the hyperparameters. The motivation for fitting hyperparameters by maximizing the marginal likelihood is to best explain the observed data. However, when performing quadrature using the above framework, the goal is not to have the best possible explanation of the transformed data but rather to have an accurate belief about the original, untransformed data. This can be achieved by setting the hyperparameters so as to maximize the marginal likelihood of the (approximate) posterior belief on f (henceforth referred to as "fitting in f-space" as opposed to "fitting in g-space"). Gradient-based methods can be employed to solve this optimization problem as the gradient w.r.t. to the hyperparameters of the GP prior on g exists everywhere.

4 Experiments

We demonstrate some properties of our moment-matched log transform (MMLT) instantiation on a synthetic and a real-world example and compare its performance against WSABI, SBQ, traditional Monte Carlo (MC) and quasi-Monte Carlo (QMC) quadrature methods. For MMLT, WSABI and SBQ, we chose the GP prior to have a constant mean and a Matern covariance with $\nu = {}^3\!/_2$ and automatic relevance determination. Hyperparameters were fit in f-space when applicable with multiple restarts of the optimization procedure and all sample locations were selected iteratively using uncertainty sampling. The standard deviation of the noise in f-space was fixed to 10^{-6}. One shortcoming of MMLT (and indeed any use of the log transformation) is that the final belief on the mean of Z is an intractable integral itself for most choices of the covariance function. We estimated this integral using QMC, a defendable choice when the original integrand is expensive and thus cannot be estimated directly using QMC. We plan on investigating other approximation schemes in future work.

For our synthetic example, we consider estimating $Z(a, C) = \int_{-a}^{a} \exp(-Cx^2/2)(1/2a)\,\mathrm{d}x$ (an unnormalized, Gaussian likelihood $\exp(-Cx^2/2)$ against a uniform prior $(1/2a)$). Figure 1 shows the posterior belief for all three BQ methods for $Z(5, 10)$ along with sampled locations.

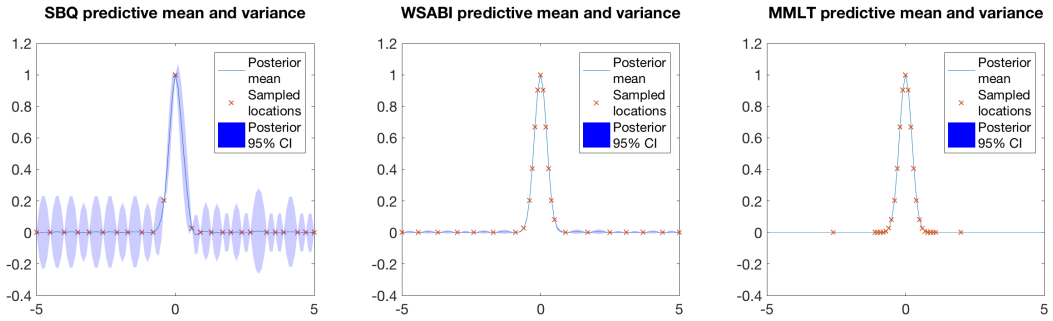


Figure 1: Posterior belief of f for SBQ, WSABI and MMLT

The belief maintained by SBQ includes significant mass below zero, WSABI's includes some mass below zero and MMLT outperforms both at incorporating the non-negativity information and clamps down on its uncertainty everywhere. Also, while SBQ and WSABI choose roughly uniformly distributed points, MMLT quickly finds the region of interest i.e. where the integrand transitions from having some mass to having functionally zero mass. Figure 2 shows the absolute errors of the log of the estimates of $Z(5, C)$ generated by MMLT, WSABI and SBQ for $C \in \{1, \ldots, 10\}$.

Using C as a proxy for the dynamic range of the integrand, as the dynamic range increases, the accuracy of WSABI and SBQ falls off faster than the accuracy of MMLT. Although this is a synthetic example, likelihoods frequently have dynamic ranges as large as the largest example, $Z(5, 10)$.

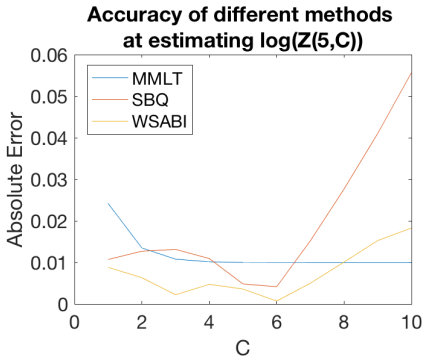
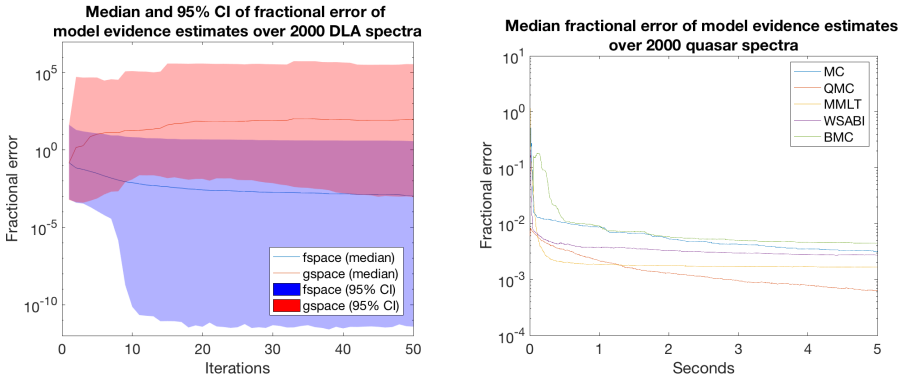


Figure 2: Accuracy of SBQ, WSABI and MMLT at estimating the synthetic example

Our real-world application is a model selection problem from astrophysics: predicting whether a damped Lyman-α absorber (DLA) exists on the sightline between a quasar and earth. For a complete description of the problem, see [10]. Figure 3 contains the results of two experiments on this dataset.



(a) (b)

Figure 3: Experiments on DLA application

Figure 3a shows the fractional error of estimates of the model evidence using MMLT with hyperparameters fit in g-space and f-space. The version of MMLT that fits in f-space significantly outperforms the model that fits in g-space. Since the underlying data has a massive dynamic range, the model that fits in g-space is forced to have a large output scale. Conversely, when fitting in f-space, the dynamic range of observed values shrinks significantly, allowing the model to maintain a reasonable degree of certainty about the value of the integrand everywhere. Figure 3b shows the fractional error of different quadrature methods: MMLT outperforms SBQ, WSABI and MC in terms of wall-clock time. Note that although QMC outperforms MMLT on this application, QMC is not well-suited for most model selection problems due to the inability to build an appropriate low-discrepancy sequence.

5 Conclusion and Future Work

We have presented a framework for performing Bayesian quadrature of constrained integrands where the constraint on the integrand can be defined as a warping of $\mathbb{R}$. We showed how previous methods can be described using this framework, developed an instantiation of this framework with some desirable properties and proposed a new methodology for setting associated hyperparameters. Using both synthetic and real-world data, we have shown that our new instantiation can outperform previous BQ algorithms developed to estimate intractable integrals of non-negative integrands as well as traditional Monte Carlo based methods. Future work includes determining how to deal with noisy observations being pushed through the transformation ξ and incorporating observations of the integrand's directional derivative in different directions.

6 References

References

[1] R.M. Neal. Annealed importance sampling. Statistics and Computing, 11(2):125—-139, 2001.

[2] X. Meng and W.H. Wong. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. Statistica Sinica, 6(4):831—-860, 1996.

[3] J. Skilling. Nested sampling. Bayesian inference and maximum entropy methods in science and engineering, 735:395—-405, 2004.

[4] P. Diaconis. Bayesian numerical analysis. Statistical Decision Theory and Related Topics, 4 (1):163—175, 1988.

[5] A. O'Hagan. Bayes-hermite quadrature. Journal of Statistical Planning and Inference, 29: 245—260, 1991.

[6] C.E. Rasmussen and Z. Ghahramani. Bayesian monte carlo. Advances in Neural Information Processing Systems, 2003.

[7] T. Gunter, M.A. Osborne, R. Garnett, P. Hennig, and S.J. Roberts. Sampling for inference in probabilistic models with fast bayesian quadrature. Advances in Neural Information Processing Systems, 2014.

[8] M. Osborne, R. Garnett, Z. Ghahramani, D.K. Duvenaud, S.J. Roberts, and C.E. Rasmussen. Active learning of model evidence using bayesian quadrature. Advances in Neural Information Processing Systems, 2012.

[9] C.E. Rasmussen and C.K.I. Williams. Gaussian Processes for Machine Learning. MIT Press, 2006.

[10] S. Bird R. Garnett, S. Ho and J. Schneider. Detecting damped lyman-α absorbers with gaussian processes. Journal of Statistical Planning and Inference, 472(2):1850–1865, 2017.