Improving Quadrature for Constrained Integrands

Henry Chai
hchai@wustl.edu

Roman Garnett
garnett@wustl.edu

Department of Computer Science and Engineering
Washington University in St. Louis

Abstract

We present an improved Bayesian framework for performing inference of affine transformations of constrained functions. We focus on quadrature with nonnegative functions, a common task in Bayesian inference. We consider constraints on the range of the function of interest, such as nonnegativity or boundedness. Although our framework is general, we derive explicit approximation schemes for these constraints, and argue for the use of a log transformation for functions with high dynamic range such as likelihood surfaces. We propose a novel method for optimizing hyperparameters in this framework: we optimize the marginal likelihood in the original space, as opposed to in the transformed space. The result is a model that better explains the actual data. Experiments on synthetic and real-world data demonstrate our framework achieves superior estimates using less wall-clock time than existing Bayesian quadrature procedures.

1 Introduction

Integrals over model (hyper)parameters are frequently encountered in Bayesian inference. Model selection, for example, is a fundamental concern in the course of scientific inquiry: which of several candidate models best explains an observed dataset $\mathcal{D}$? The Bayesian approach requires the computation of *model evidence,* an integral of the form $Z = \int f(\mathcal{D} \mid \theta)\, \pi(\theta)\, \mathrm{d}\theta$ where θ is a vector of model parameters, $f(\mathcal{D} \mid \theta)$ is a likelihood, and $\pi(\theta)$ is a prior. Computing a marginal predictive distribution similarly requires integrating a predictive density $p(y \mid x, \mathcal{D}, \theta)$ against a posterior distribution $p(\theta \mid \mathcal{D})$. Note that the integrand in both these scenarios is known *a priori* to be nonnegative, as it is the product of probability densities. Unfortunately, these integrals are often computationally intractable and thus must be approximated.

Numerous common techniques to estimate such integrals rely on *Monte Carlo* estimators [15, 17, 23]. These methods are agnostic to prior information about the integrand, such as nonnegativity, and also converge slowly in terms of the number of required samples, rendering them ill-suited for settings where the integrand is expensive to evaluate. One alternative is *Bayesian quadrature* (BQ) [4, 14, 18, 20], which relies on a probabilistic belief on the integrand that can be conditioned on observations to derive a posterior belief about the value of the integral or any other affine transformation. The theoretical properties of kernel quadrature methods (including BQ) have been studied at length: these methods can achieve faster convergence rates than Monte Carlo estimators [1, 2, 13], even when the underlying model is misspecified [11, 12], a commonly-cited pitfall of kernel-based methods.

Recent work by Gunter et al. [9] and Osborne et al. [19] have improved the speed and accuracy of classical BQ methods such as *Bayesian Monte Carlo* (BMC) [20] for estimating integrals of *nonnegative* functions. These two methods reason about the square root and the log of the integrand, respectively, instead of the integrand itself. By "undoing" these transformations, we may softly incorporate the nonnegativity constraint. Although previous work [9, 19] has demonstrated that suitably modified BQ can outperform Monte Carlo methods and BMC for estimating integrals of nonnegative functions, a general framework for quadrature with the use of transformations has never been offered.

Our contribution is to define a Bayesian framework for a wide variety of inference tasks, including quadrature, involving a broader class of constrained functions. We provide complete details of this framework for two important classes of constrained functions: nonnega-

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

tive functions and functions bounded on an interval. Common examples arising in machine learning include likelihoods and classification (e.g., validation) error. We then apply our framework to quadrature, where we address some shortcomings of previous work. Specifically, our approach can make effective use of a log transform to efficiently estimate integrals involving extreme dynamic range. This is in contrast to the method in [9], which cannot handle such dynamic range, and to [19], which relied on a series of abstruse and inefficient approximations. Finally, we develop a novel training procedure whereby hyperparameters are fit by maximizing the marginal likelihood of true observations of the integrand. All previous related work instead fit hyperparameters by maximizing the marginal likelihood of transformed observations. We demonstrate this can lead to undesirable behavior and that our procedure yields a better-behaved model, *even if adopted into previous procedures such as* [9]. We conduct experiments with real-world data showing that our proposed framework and novel hyperparameter optimization method outperforms previous BQ algorithms.

2 Bayesian Quadrature

Let $Z = \int f(x)\,\pi(x)\,\mathrm{d}x$ be an intractable integral.[1] Bayesian quadrature operates by placing a Gaussian process (GP) prior on the function f, $p(f) = \mathcal{GP}(\mu, \Sigma)$ [21]. GPs are probability distributions over functions, where the joint distribution of any finite number of function values is multivariate normal; this belief is parametrized by a mean function $\mu(x)$ and a covariance function $\Sigma(x, x')$. Given a set of observations at locations $\mathbf{x} = \{x_1, \dots, x_n\}$ with corresponding function values $\mathbf{f} = f(\mathbf{x})$, the GP prior can be conditioned on these observations to arrive at a posterior GP with mean $\mu_{\mathcal{D}}(x) = \mu(x) + \Sigma(x, \mathbf{x})\Sigma(\mathbf{x}, \mathbf{x})^{-1}\bigl(\mathbf{f} - \mu(\mathbf{x})\bigr)$ and covariance $\Sigma_{\mathcal{D}}(x, x') = \Sigma(x, x') - \Sigma(x, \mathbf{x})\Sigma(\mathbf{x}, \mathbf{x})^{-1}\Sigma(\mathbf{x}, x')$.

Given a GP belief on a function, we may derive a belief over integrals of that function using the fact that GPs are closed under linear transformations such as integration [20]. Specifically, if $p(f) = \mathcal{GP}(\mu, \Sigma)$, then our integral of interest $Z = \int f(x)\,\pi(x)\,\mathrm{d}x$ is normal:

$$p(Z) = \mathcal{N}\bigl(\textstyle\int \mu(x)\pi(x)\,\mathrm{d}x, \iint \Sigma(x, x')\pi(x)\pi(x')\,\mathrm{d}x\,\mathrm{d}x'\bigr). \tag{1}$$

Warped sequential active Bayesian integration (WSABI) [9] builds off BQ to incorporate nonnegativity information about an integrand f with a warped GP [24]. Specifically, WSABI places a GP prior on $g(x) = \sqrt{2(f(x) - \alpha)}$, for some small positive constant α. This prior is then conditioned on observations to arrive at a posterior, like BQ. Warped GPs have been previously used for a variety of machine learning tasks [22, 26]. However, when applied to quadrature, warped GPs lack the key property of closure under linear transformations. In particular, the marginal predictive distribution of an arbitrary function value $f(x)$ is no longer Gaussian but instead depends on the choice of warping function; in the case of WSABI, these marginals are non-central χ^2 distributions, which are inconvenient for quadrature. WSABI approximates the posterior belief about f as a GP using one of two proposed 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 on f and adopts a GP matching these moments [9]. Either approximation gives a GP belief about f approximately incorporating the nonnegativity constraint, and we may use standard results such as (1) to reason about integrals, etc. Below we will describe a general procedure following these ideas, then describe how to improve upon the procedure in numerous ways in practice.

3 Inference on Constrained Functions

We propose a framework for inferring affine functionals of functions with contraints on their range. Let $f\colon \mathcal{X} \to \mathcal{Y} \subset \mathbb{R}$ be a function of interest with range constrained to a subset $\mathcal{Y}$ of the real line; for example, a nonnegative function would have $\mathcal{Y} = (0, \infty)$, and a function bounded on an interval would have $\mathcal{Y} = (a, b)$. Let $Z = L[f]$ be an affine functional of f we wish to infer.

1. Determine an invertible warping ξ mapping $\mathbb{R}$ onto $\mathcal{Y}$, the domain of f. Define an *unconstrained* function $g\colon \mathcal{X} \to \mathbb{R}$ by $g(x) = \xi^{-1}\bigl(f(x)\bigr)$ and place a GP prior on g, $p(g) = \mathcal{GP}(\mu, \Sigma)$.

2. Observe g at locations chosen by an appropriate sampling policy, yielding data $\mathcal{D} = \{\mathbf{x}, g(\mathbf{x})\}$.

3. Derive a posterior belief on the transformed function, $p(g \mid \mathcal{D}) = \mathcal{GP}(\mu_{\mathcal{D}}, \Sigma_{\mathcal{D}})$.

4. Calculate the posterior mean $m_{\mathcal{D}}$ and covariance $K_{\mathcal{D}}$ functions of the induced posterior belief on f. If needed, these can be approximated as polynomials in the posterior moments of g; see below for details. Approximate the belief on f by a moment-matched GP: $p(f \mid \mathcal{D}) \approx \mathcal{GP}(m_{\mathcal{D}}, K_{\mathcal{D}})$.

5. Derive a posterior belief about Z (e.g., (1)):

$$p(Z \mid \mathcal{D}) = \mathcal{N}\bigl(L[m_{\mathcal{D}}], L^2[K_{\mathcal{D}}]\bigr) \tag{2}$$

where $L^2[K] = L\bigl[L[K(x, \cdot)]\bigr] = L\bigl[L[K(\cdot, x)]\bigr]$ (see (1) for an example).

[1] For notational simplicity, the following will be written as if $x \in \mathbb{R}$, but all results extend to $x \in \mathbb{R}^d$.

Table 1: Induced moments of $f = \xi(g)$ for various transformations ξ, if $p(g) = \mathcal{GP}(\mu, \Sigma)$. We provide the *raw* second moment $C(x, x')$ in this table; the covariance function can be computed by $K(x, x') = C(x, x') - m(x)\,m(x')$. Some entries for the second raw moment refer to values of the first moment for that transform.

transform	first moment $m(x) = \mathbb{E}\bigl[f(x)\bigr]$	second raw moment $C(x, x') = \mathbb{E}\bigl[f(x)f(x')\bigr]$
$\xi(f) = \alpha + f^2$ [9]	$\alpha + \mu(x)^2 + \Sigma(x, x)$	$2\Sigma(x, x')^2 + 4\mu(x)\,\Sigma(x, x')\,\mu(x') + m(x)\,m(x')$
$\xi(f) =$ any polynomial in f	polynomial in μ and Σ	polynomial in μ and Σ
$\xi(f) = \exp(f)$	$\exp\bigl(\mu(x) + {}^{1}\!/_{2}\Sigma(x, x)\bigr)$	$m(x)\,\exp\bigl(\Sigma(x, x')\bigr)\,m(x')$
$\xi(f) = \Phi(f)$	$\Phi\left(\frac{\mu(x)}{\sqrt{\Sigma(x, x) + 1}}\right)$	$\Phi\left(\begin{bmatrix}\mu(x)\\ \mu(x')\end{bmatrix}, \begin{bmatrix}\Sigma(x, x) + 1 & \Sigma(x, x')\\ \Sigma(x', x) & \Sigma(x', x') + 1\end{bmatrix}\right)$

In short, we maintain a GP belief on a warped version of f that removes the constraint. We then approximate a GP belief on f given data via moment matching, after which we can easily reason about affine functionals. Particular instances of this framework have appeared in the literature; for example, WSABI (specifically the –M variant [9]) implements this framework using the square root transform to infer integrals of nonnegative functions. However, we will discuss the framework in greater generality and provide practical advice.

The above framework is agnostic to several design choices. First, we do not specify the warping function ξ in step (1). WSABI, for example, relies intimately on the square root map. This induces nonnegativity, but we will demonstrate that it does not yield useful models for functions with high dynamic range. We will provide details to work with a wide range of warping functions, including polynomials, log transformations, and sigmoidal transformations such as the probit.

Further, we do not specify how exactly the posterior belief in the transformed space $p(g \mid \mathcal{D})$ is derived in step (3), in particular how any associated hyperparameters are fit. We will discuss this issue in detail later and provide a novel approach.

Finally, we make no assumptions about the mechanism for choosing observation locations $\mathbf{x}$ in step (2). These could be sampled proportional to some distribution, à la Monte Carlo, or chosen via information-theoretic principles or some other scheme. If no warping function is used, as in BMC, then the optimal set of locations in terms of minimizing the posterior variance/entropy of our belief about Z can be precomputed, as the posterior covariance of a GP does not depend on the observed values [16]. However, in the scheme outlined above, the approximate posterior covariance of f, $K_{\mathcal{D}}$, *does* depend on the observed values, as it a function of the mean belief in the transformed space, $\mu_{\mathcal{D}}$; see below for details. Thus, to make use of policies that maximize information gain in this setting, observation locations must be selected sequentially. In WSABI, samples are chosen by greedily maximizing information gain about the integrand, selecting each point to maximize the posterior variance: $x^* = \arg\max_x K_{\mathcal{D}}(x, x)$. Osborne et al. [19] chose samples so as to maximize the expected information gain about an integral Z directly. Both are compatible with our proposed framework.

3.1 Transform selection, moment matching

We briefly pause to discuss the moment-matching step in step (4) of our procedure. Several useful general-purpose transformations admit closed-form expressions for the posterior mean and covariance on f given a GP belief about $g = \xi^{-1}(f)$, $p(g) = \mathcal{GP}(\mu, \Sigma)$. We provide a summary for several notable examples in Table 1; details can be found in the supplemental material.

For a nonnegative function taking values on $\mathcal{Y} = (0, \infty)$, we may use the square root transform $\xi^{-1} = \sqrt{f}$ or the log transform $\xi^{-1} = \log f$. Choosing an appropriate transform for a given scenario will require consideration of the data. For example, when the data has extreme dynamic range, as is often the case for likelihood surfaces, a log transformation may be desired. Figure 1 shows an example *log* likelihood surface for a real-wold astronomical model we will consider in our experiments [7]. Note that computing a model evidence requires integrating the *likelihood* surface, not the log likelihood. The dynamic range of the likelihood is on the rough order of $10^{10\,000}$, and no off-the-shelf GP could reasonably model this function. The square root of the likelihood, as would be used in WSABI, reduces the dynamic range to an equally unmanageable 10^{5000}. The log transformation, however, produces a well-behaved surface that could be reasonably modeled with a GP.

To model a bounded function taking values on the interval $(0, 1)$, we could use a probit transform $\xi = \Phi(f)$; closed-form moments for the induced belief on f are also provided. The covariance requires the bivariate Gaussian CDF, which can be estimated efficiently with high precision [8]. By shifting and scaling appropriately, we can model a function taking values on any interval of the form $(-\infty, a)$,(a, b), or (b, ∞).

For an arbitrary polynomial warping $\xi = a_n f^n + a_{n-1} f^{n-1} \cdots + a_0$, an extension of Isserlis' theorem guarantees that the moments of f will be polynomials

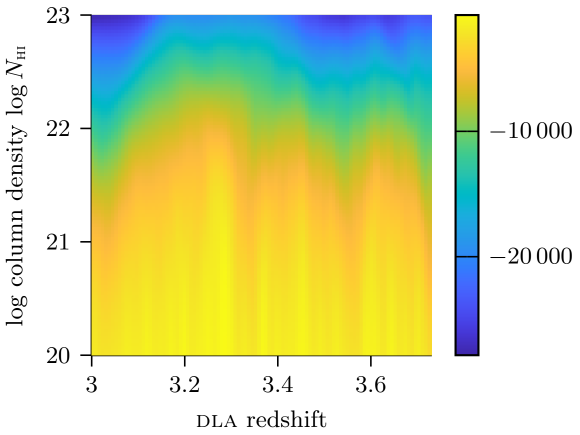


Figure 1: The log-likelihood surface for a real-world astronomical dataset corresponding to a an astronomical model described further in our experiments [7]. The dynamic range is massive, on the order of $\exp(27\,135) \gg 10^{10\,000}$.

in μ and Σ (of degree n for the mean and $2n$ for the covariance), and a simple algorithm can generate these moments on demand [25].

We show a brief demonstration of fitting the bounded function $f(x) = 0.95\exp(-2x^2)$ (scaled to avoid the value of exactly 1 at 0) using a log and probit transformation in Figure 2. The model fit to data directly and unaware of the transformation produces considerable predictive mass on invalid values. The exact posteriors for the log and probit transformations both absolutely respect their respective constraints. The moment-matched GPs are excellent approximations.

3.2 Hyperparameter optimization

When GPs are used for inference, an important consideration is how to set the associated hyperparameters. One commonly used method is to optimize the marginal likelihood of the observed data using gradient-based methods as the gradient of the marginal likelihood w.r.t. hyperparameters is readily available for this model class. The motivation for fitting hyperparameters by maximizing the marginal likelihood is to explain the observed data as well as possible. However, when performing inference 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. Previous related approaches (e.g., [9, 19]) have ignored this fact and fit the hyperparameters of the warped GP in the warped space. We will show this can lead to poor behavior.

We propose setting hyperparameters by maximizing the marginal likelihood of the untransformed data using the (approximate) posterior belief on f; we will refer to optimizing the hyperparameters in this manner as

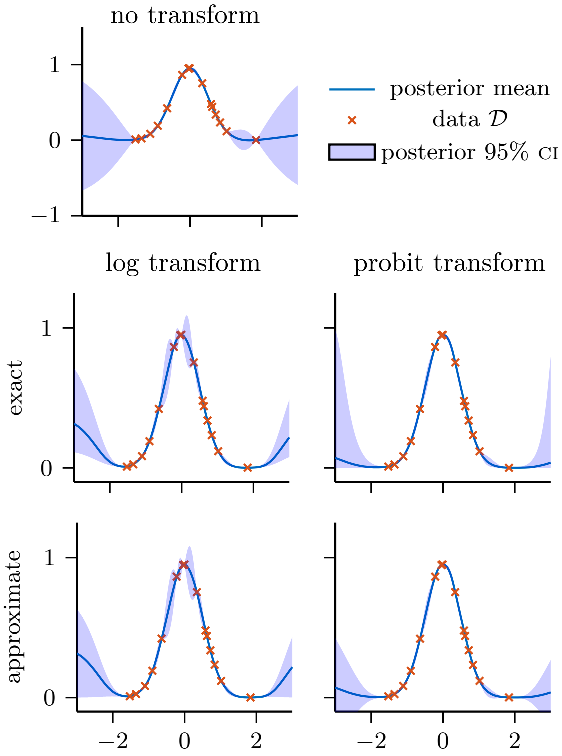
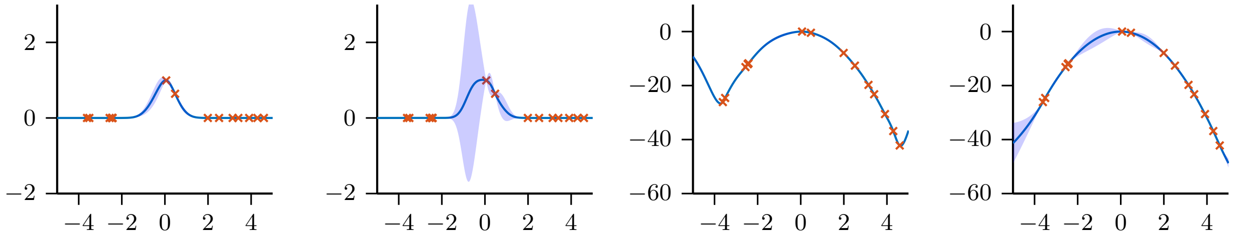


Figure 2: A demonstration of fitting a simple function $f(x) = 0.95\exp(-2x^2)$ on the interval $[-3, 3]$ using a log and probit transformation in our framework. Each column shares an x axis and each row shares a y axis.

“fitting in f-space” as opposed to “fitting in g-space.”

Formally, if $p(g) = \mathcal{GP}\big(\mu(\theta), \Sigma(\theta)\big)$ (where dependence on hyperparameters θ has been written explicitly), our framework approximates $p(f)$ with $p(f) \approx \mathcal{GP}\Big(m\big(\mu(\theta), \Sigma(\theta)\big), K\big(\mu(\theta), \Sigma(\theta)\big)\Big)$. The exact relationship between θ and the mean/covariance of f depends on the transformation ξ. For many natural choices, the partial derivatives ${}^{\partial m}/_{\partial \mu}$, ${}^{\partial m}/_{\partial \Sigma}$, ${}^{\partial K}/_{\partial \mu}$ and ${}^{\partial K}/_{\partial \Sigma}$ will be available. Thus, we can evaluate the partial derivative of f w.r.t. to θ and use the same gradient-based methods used to fit hyperparameters in g-space to fit hyperparameters in f-space; for the transformations found in Table 1, the relevant partial derivatives can be found in the supplementary material.

Figure 3 shows the impact of fitting the hyperparameters in f-space as opposed to fitting in g-space using our toy function $f(x) = 0.95\exp(-2x^2)$. The hyperparameters learned in f-space result in a model that fits the f-space data well but do a poor job explaining the data in g-space; the learned mean is much higher than the mean of the transformed data and the learned output scale is very small, leading to unreasonably little uncertainty in the model. However, these learned hyperparameters make sense in the context of the f-space data, where most of the observations are



(a) $p(f)$, fit in f-space (b) $p(f)$, fit in g-space (c) $p(g)$, fit in f-space (d) $p(g)$, fit in g-space

Figure 3: Fitting in f-space vs. fitting in g-space. We model the function $f(x) = 0.95 \exp(-2x^2)$ on the interval $[-5, 5]$, conditioning on 15 observations at locations sampled uniformly at random. We place a GP prior on $g = \log f$ with constant mean and Matérn covariance with $\nu = {}^3\!/\!_2$. This model has three hyperparameters: a mean, an output scale, and a length scale. These were fit in f-space ((a) and (c)) and g-space ((b) and (d)). See the legend in Figure 2.

effectively zero and the maximum observed value is slightly less than one. Conversely, the hyperparameters learned in g-space fit the g-space data very cleanly, with a well-scaled uncertainty. However, this translates to a poorly-behaved model in f-space; the region from $[-2, -0]$ has what appears to be a very reasonable variance in g-space, but this corresponds to a massive variance in f-space that strongly defies the nonnegativity constraint.

We offer two practical notes about fitting in f-space in the case of a log transform learned through our experiments. First, we suggest shifting the g-space data so that the maximum observed value is exactly zero, as this places the observations into a regime where the inverse transformation is well-behaved. We are free to make such a shift as doing so simply scales the f-space data by a constant. Second, initializing the hyperparameter optimization procedure must be done carefully when fitting in f-space. If one is using a constant mean, we recommend avoiding naïvely initializing the prior mean to be the mean of the transformed data. Instead, we initialized the mean to one of -1, -2, -5, and -10 and initialized the output scale of the covariance function to the mean initialization divided by -2. We believe this set of initializations to be sufficient after shifting the data because the relevant portions of the f-space data should be well-described by a hyperparameter setting reachable from these initializations. Lower means may result in undesirable behavior, as the corresponding output scales would need to be large to explain the shifted observation at zero.

3.3 Approximating the posterior on Z

For some combinations of linear functionals and warping functions, the posterior belief on Z (2), may be intractable, i.e., either $L[m]$ or $L^2[K]$ cannot be expressed in closed form. This is the case for quadrature with the log transformation and most common choices of covariance function, including the Matérn and squared exponential kernels, as the posterior belief contains a term of the form $\int \exp \exp x \, \mathrm{d}x$.

Various approximation techniques can be used to estimate these intractable quantities. Osborne et al. [19] use BQ itself, a somewhat unsatisfying approach as it leads to infinite regress. Briol et al. [2] provide a theoretical justification for the use of Monte Carlo based methods when estimating intractable posterior means. We propose an alternative approximation scheme that makes use of a Taylor series expansion to approximate the f-space moments $m(x)$ and $K(x, x')$. The exact nature of the Taylor series will depend on the warping function ξ; for $\xi = \exp f$, the following approximations follow from the expressions in Table 1:

$$\begin{aligned} m(x) &\approx 1 + \mu(x) + {}^1\!/\!_2\Sigma(x, x) \\ &\quad + \big(\mu(x) + {}^1\!/\!_2\Sigma(x, x)\big)^2/2 + \dots \end{aligned} \tag{3}$$

$$\begin{aligned} K(x, x') &\approx 1 + \Sigma(x, x') + {}^1\!/\!_2\Sigma(x, x')^2 \\ &\quad + \Sigma(x, x')\big(\mu(x') + {}^1\!/\!_2\Sigma(x', x') \\ &\quad + \mu(x) + {}^1\!/\!_2\Sigma(x, x)\big) + \dots \end{aligned} \tag{4}$$

Given these approximations, the posterior mean and variance for quadrature are tractable for certain covariance functions, including the squared exponential kernel [10]. Indeed, for reasonably well-behaved warpings ξ, the associated approximations will be polynomial functions of μ and Σ, and thus tractable for integrating against standard covariance functions. This last result follows directly from Isserlis' theorem (see § 3.1). Unfortunately, computing this approximation is expensive for higher-order terms: computing the dth order term in either Taylor series after making n function evaluations takes $\Theta(n^{2d})$ time.

4 Experiments

We perform experiments in a variety of settings to evaluate our proposed framework and demonstrate the importance of our proposed improvements. We begin by exploring the effect of fitting in f-space using different transformations on a simple regression task. Then we apply our framework to quadrature of nonnegative integrands using a moment-matched log transformation (MMLT). We compare these results against WSABI and BMC as well as Monte Carlo methods. If not otherwise specified, all GP priors were chosen to have constant mean and Matérn covariance with $\nu = 3/2$, all sample locations were selected iteratively using uncertainty sampling in f-space [9], all hyperparameters were fit in f-space when applicable, and all intractable posteriors were estimated using quasi-Monte Carlo [3].

4.1 Hyperparameter tuning

To assess the impact of modeling constrained functions using transformations, we consider three regression tasks using the standard benchmarks of the HPOlib package [5]: online LDA, SVM, and logistic regression (LR). For each benchmark, Eggensperger et al. [5] provide a list of hyperparameter settings for the eponymous machine learning algorithm along with the associated observations of some relevant, machine learning quantity: for the online LDA benchmark, the observed values are per-word perplexities (which are nonnegative), whereas for the SVM and LR benchmarks the observed values are prediction error rates (which are bounded between 0 and 1). The online LDA, SVM, and LR datasets contain 289, 1400, and 9680 observations, respectively.

For each benchmark, we ran the following experiment 100 times: we randomly select some percentage of the dataset to be a training set (20% for online LDA, 5% for the other two) and designate the remaining observations to be a test set. We fit a moment-matched GP to the training set using both the log and square root transformations for online LDA and a probit transformation for both SVM and LR. We compare our framework against a standard, constraint-unaware GP and a moment-matched GP where the hyperparameters were fit in g-space as opposed to in f-space. We consider two metrics: the root mean squared error (RMSE) on the test set and the mean predictive log likelihood (MLL) of observations in the test set, $\mathbb{E}\big[\log p\big(f(x) \mid x, \mathcal{D}\big)\big]$.

The results are shown in Table 2. We can extract a few trends. Using a transformation that respects the *a priori* knowledge about the target function leads to an improvement in accuracy; for the online LDA benchmark, the difference between the RMSE of the constraint-agnostic GP and the RMSEs of all methods using a transformation is significant at the 1% significance level according to a one-sided paired t-test. In general, our proposed hyperparameter optimization methodology does not lead to a significant difference in the RMSE. All methods tend to learn similar predictive means in f-space for these datasets, which do not reflect extreme behavior. The impact of our proposed methodology can be seen in the mean predictive log likelihoods, however. In terms of this metric, fitting in f-space is preferable to fitting in g-space for both transforms as it leads to better-scaled uncertainties.

Table 2: Regression experiment results.

dataset	transform	RMSE	MLL
LDA	none	153	-1.0×10^{10}
	square root (g-space)	142	-2.1×10^{6}
	square root (f-space)	142	-6.1×10^{5}
	log (g-space)	134	-4.1×10^{6}
	log (f-space)	133	-4.8×10^{5}
SVM	none	0.015	2.83
	probit (g-space)	0.015	2.82
	probit (f-space)	0.015	2.91
LR	none	0.036	1.98
	probit (g-space)	0.036	2.06
	probit (f-space)	0.035	2.07
IM	none	0.281	−0.110
	probit (g-space)	0.266	−0.324
	probit (f-space)	0.256	0.319

The gains of fitting in f-space are reduced when using the probit transformation on these particular benchmarks because the dynamic range is not very large: observations of the per-word perplexity in the LDA benchmark range from roughly 1000 to 5000, whereas observations of the error rates for the SVM and LR benchmarks only range from 0.24 to 0.50 and from 0.07 to 0.91, respectively. Although the range of observations for the LR benchmark may seem large, this translates to observations between −1.5 and 1.5 in the transformed space.

To showcase the power of the probit transformation with more-extreme data, we ran the following in-model (IM) experiment 100 times. We randomly sampled a draw from a two-dimensional GP prior, which we then pushed through the inverse-probit transformation to generate a function bounded between 0 and 1. The output scale and length scales of the GP were set such that samples range roughly from −5 to 5 over the domain. We then sampled 200 points from the draw, fit a moment-matched GP using the probit transform (in both f-space and g-space) to 20% of the points, and predicted the values of the remaining 80%. The results are shown in Table 2. All differences in performance are significant at the 1% significance level according to one-sided paired t-tests. As the results indicate, in this setting, it becomes important to fit hyperparameters in f-space rather than in g-space to achieve reasonably scaled uncertainties.

4.2 Detecting DLAs via model selection

We consider a real-world quadrature application of our framework, a model selection problem from astrophysics. We wish to infer whether a damped Lyman-α absorber (DLA) exists along the line of sight between a quasar and earth given spectrographic observations. DLAs are large gaseous clouds containing neutral hydrogen at high densities. Their location and size can be inferred from observations of quasar spectra as they cause distinctive dips in the observed flux at well-defined wavelengths. The distribution of DLAs throughout the universe is important as it provides insight into models of galaxy formation. Garnett et al. [7] developed a model that specifies the likelihood that a given emission spectrum contains a putative DLA. The model is parameterized by two physical features of a candidate DLA: its column density, which roughly corresponds to its size, and its redshift, which roughly corresponds to its distance from earth. Garnett et al. [7] also specified a data-driven prior distribution over these two parameters, which must be integrated against to calculate the model evidence and derive a posterior distribution of DLA presence. The model evidence of this DLA model is an (intractable) integral of the likelihood over the domain of these two model parameters. Here we will consider computing the model evidence of 2000 spectra gathered from phase III of the Sloan Digital Sky Survey (SDSS–III) [6]. For a complete description of the problem, data, and model, see [7].

A sample log-likelihood surface for this model corresponding to a particular quasar spectrum is shown in Figure 1. These functions are highly multimodal and have a massive dynamic range. These features make computing the model evidence a difficult task for alternative methods such as BMC and WSABI. One convenient feature of this experimental setting is that the dimensionality of the intractable integral can be scaled up to any even number simply by calculating the model evidence for the existence of n DLAs, resulting in a $2n$-dimensional integral [7].

We conducted an experiment comparing the accuracy of BQ methods for estimating model evidence in this setting, including BMC, WSABI, and MMLT. We considered the latter two fitting both in f-space and in g-space. We also compared with sequential Monte Carlo (SMC) and quasi-Monte Carlo (QMC) estimation. We estimate model evidences for a single DLA and three DLAs in 2000 quasar spectra, entailing two- and six-dimensional integrals, respectively. Each method was allotted 5 seconds of wall-clock time for estimating the two-dimensional integrals and 60 seconds for the six-dimensional integrals. Monte Carlo methods drew or constructed samples from the prior specified by Garnett et al. [7].

Table 3: Mean $\log p(Z^* \mid \mathcal{D})$ at termination.

transform	2d	6d
none (BMC)	−0.79	1.93
square root (WSABI) (g-space)	3.67	3.40
square root (WSABI) (f-space)	3.89	3.43
log (MMLT) (g-space)	−266	−505
log (MMLT) (f-space)	10.3	7.57

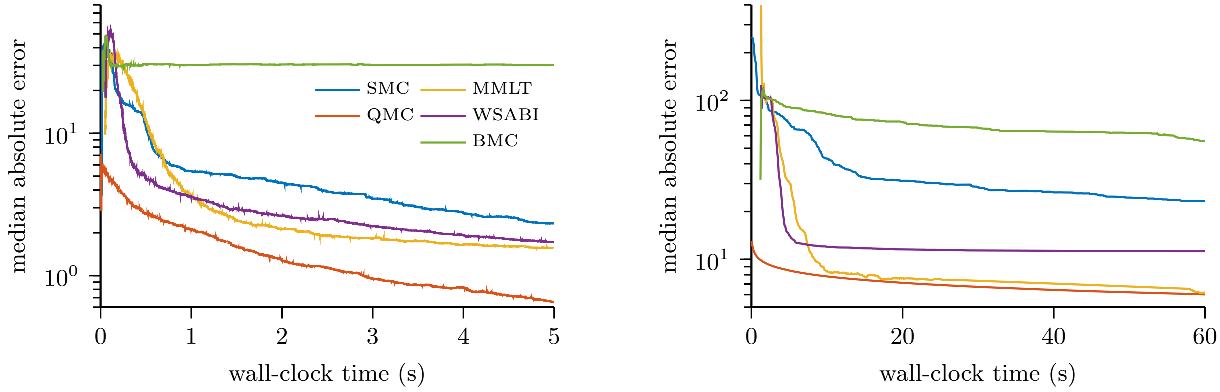
Table 4: Mean MLL at termination.

transform	2d	6d
none (BMC)	−1.66	0.33
square root (WSABI) (g-space)	1.51	1.26
square root (WSABI) (f-space)	1.59	1.51
log (MMLT) (g-space)	−3.87	−7.28
log (MMLT) (f-space)	1.68	1.65

Figure 4 shows the median absolute error over time of each method, using exhaustive QMC sampling as ground truth. MMLT outperforms all other methods except QMC; note that QMC is not necessarily well-suited for model-selection when it is not possible to construct an appropriate low-discrepancy sequence, but we use it to provide a gold-standard baseline. The difference in absolute errors at termination between MMLT and the other BQ methods is significant for the six-dimensional integrals at a 1% significance level according to a one-sided paired t-test.

Tables 3 and 4 show the results of additional experiments performed in this setting that demonstrate the importance of our proposed hyperparameter optimization methodology. Table 3 compares the log-likelihood of the true value of the integral Z^* under each Bayesian method's posterior belief upon termination in these experiments while Table 4 compares the MLL (see § 4.1). Here the MLL is computed by averaging over the log predictive probabilities of the QMC samples used to estimate the model evidence.

MMLT where the hyperparameters are fit in f-space outperforms all alternatives on both metrics in both the two-dimensional and six-dimensional experiments; the differences in Table 3 are significant at a 1% significance level according to a one-sided paired t-test. MMLT where the hyperparameters are fit in g-space significantly *underperforms* the other Bayesian algorithms. The relatively poor performance of fitting in g-space on these metrics is largely due to the high dynamic range of the likelihood surfaces, which forces the output scales learned by fitting in g-space to be high. This in turn causes both the pointwise distributions and the distribution on the value of the integral to have large variances (relative to their means), making the likelihood everywhere low, much like the situation depicted in Figure 3.



(a) 2d model evidence

(b) 6d model evidence

Figure 4: The median absolute predictive error of each method's estimate of the log model evidence over time in the DLA experiments.

The difference between WSABI where the hyperparameters are fit in f-space and WSABI where the hyperparameters are fit in g-space on both metrics is relatively small. This is a consequence of the square root transformation, which barely affects the extreme dynamic range of this data. The likelihood is so extremely small everywhere (on the order of $10^{-10\,000}$) that there is practically no difference between the true values and their square root. Thus, the settings of the hyperparameters arrived at under the two methodologies are very similar; importantly, they have similar output scales, thus explaining their similar uncertainties about both f and Z^*. However, for MMLT, where the transformation does result in a drastic change in the dynamic range of the observations, fitting in f-space is crucial as it ensures that all the benefits of making this more useful transformation can be reaped. Nonetheless, fitting hyperparameters in f-space in general will not decrease performance and can result in significant gains.

5 Conclusion

We have presented a general Bayesian framework for performing inference about affine transformations of constrained functions. We developed a novel procedure for optimizing the hyperparameters associated with our method whereby the hyperparameters are set to maximize the marginal likelihood of the true data as opposed to the transformed data. Although maximizing the marginal likelihood of the transformed data may seem intuitive, we show that doing so can lead to undesirable behavior, particularly if the target function has a wide dynamic range. We then applied our proposed framework to perform regression on bounded functions and both regression and quadrature on nonnegative functions. This novel BQ algorithm outperforms previously proposed algorithms on synthetic and real-world data, both in terms of accuracy and speed of convergence. In future work, we hope to expand upon step (2) of our framework and explore bespoke sampling mechanisms tailored towards specific inference tasks.

6 Acknowledgements

This work was supported by the National Science Foundation under Award Number IIA-1355406.

References

[1] F. Bach. On the equivalence between kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research*, 18(21): 1–38, 2017.

[2] F.-X. Briol, C. J. Oates, M. Girolami, M. A. Osborne, and D. Sejdinovic. Probabilistic Integration: A Role in Statistical Computation? *arXiv preprint arXiv:1512.00933v6 [stat.ML]*, 2015.

[3] R. E. Caflisch. Monte Carlo and quasi-Monte Carlo methods. *Acta Numerica*, 7:1–49, 1998. doi: 10.1017/S0962492900002804.

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

[5] K. Eggensperger, M. Feurer, F. Hutter, J. Bergstra, J. Snoek, H. Hoos, and K. Leyton-Brown. Towards an empirical foundation for assessing Bayesian optimization of hyperparameters. In *NIPS workshop on Bayesian Optimization in Theory and Practice*, volume 10, page 3, 2013.

[6] D. J. Eisenstein, D. H. Weinberg, E. Agol, H. Aihara, C. Allende Prieto, S. F. Anderson, J. A. Arns, É. Aubourg, S. Bailey, E. Balbinot, and et al. SDSS-III: Massive Spectroscopic Surveys of the Distant Universe, the Milky Way, and Extra-Solar Planetary Systems. *The Astronomical Journal*, 142:72, Sept. 2011. doi: 10.1088/0004-6256/142/3/72.

[7] R. Garnett, S. Ho, S. Bird, and J. Schneider. Detecting Damped Lyman-α Absorbers with Gaussian Processes. *Monthly Notices of the Royal Astronomical Society*, 472(2):1850–1865, 2017.

[8] A. Genz. Numerical computation of rectangular bivariate and trivariate normal and t probabilities. *Statistics and Computing*, 14(3):251–260, 2004.

[9] 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.

[10] P. Hennig and R. Garnett. Exact Sampling from Determinantal Point Processes. *arXiv preprint arXiv:1609.06840 [cs.LG]*, 2016.

[11] M. Kanagawa, B. K. Sriperumbudur, and K. Fukumizu. Convergence guarantees for kernel-based quadrature rules in misspecified settings. In *Advances in Neural Information Processing Systems*, pages 3288–3296, 2016.

[12] M. Kanagawa, B. K. Sriperumbudur, and K. Fukumizu. Convergence analysis of deterministic kernel-based quadrature rules in misspecified settings. *arXiv preprint arXiv:1709.00147*, 2017.

[13] T. Karvonen, C. J. Oates, and S. Särkkä. A Bayes-Sard Cubature Method. *arXiv preprint arXiv:1804.03016*, 2018.

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

[15] 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.

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

[17] R. M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

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

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

[20] C. E. Rasmussen and Z. Ghahramani. Bayesian Monte Carlo. *Advances in Neural Information Processing Systems*, 2003.

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

[22] M. N. Schmidt. Function factorization using warped Gaussian processes. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 921–928. ACM, 2009.

[23] J. Skilling. Nested sampling. *Bayesian inference and maximum entropy methods in science and engineering*, 735:395–405, 2004.

[24] E. Snelson, Z. Ghahramani, and C. E. Rasmussen. Warped Gaussian processes. *Advances in Neural Information Processing Systems*, 2004.

[25] C. S. Withers. The moments of the multivariate normal. *Bulletin of the Australian Mathematical Society*, 32(1):103–107, 1985. doi: 10.1017/S000497270000976X.

[26] Y. Zhang and D.-Y. Yeung. Multi-task warped Gaussian process for personalized age estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2622–2629. IEEE, 2010.