Bayesian Lipschitz Constant Estimation and Quadrature

Jan-Peter Calliess
Dept. of Engineering
University of Cambridge, UK

Abstract

Lipschitz quadrature methods provide an approach to one-dimensional numerical integration on bounded domains. On the basis of the assumption that the integrand is Lipschitz continuous with a known Lipschitz constant, these quadrature rules can provide a tight error bound around their integral estimates and utilise the Lipschitz constant to guide exploration in the context of adaptive quadrature. In this paper, we outline our ongoing work on extending this approach to settings where the Lipschitz constant is probabilistically uncertain. As the key component, we introduce a Bayesian approach for updating a subjectively probabilistic belief of the Lipschitz constant. Combined with any Lipschitz quadrature rule, we obtain an approach for translating a sample into an integral estimate with probabilistic uncertainty intervals. The paper concludes with an illustration of the approach followed by a discussion of open issues and future work.

1 Introduction

The field of quadrature deals with algorithms for the computation of estimates of integrals of functions on the basis of a finite sample. In this work, we will employ Bayesian inference to infer probabilistic bounds on the definite integral $\int_I f(x)\,\mathrm{d}x$ (where domain $I \subset \mathcal{X}$ is bounded) on the basis of a finite sample $\mathcal{D}_n = \{(s_i, f_i) | f_i = f(s_i), i = 1, ..., N_n\}$ of integrand $f : I \subset \mathcal{X} \to \mathbb{R}$.

As with any type of inference, a priori assumptions are key since otherwise, generalisation is impossible. As a first step, one assumes $f \in \mathcal{K}_{prior}$ for some set of functions $\mathcal{K}_{prior}$. For instance, many classical quadrature techniques assume that this a priori class of function is a subset of continuously differentiable functions; Monte-Carlo methods frequently provide error bounds on the basis that K_{prior} is a subset of the set of square-integrable functions. In Lipschitz quadrature [1], one assumes K_{prior} is contained in the set

$$\mathrm{Lip}_I(L) = \left\{\phi : I \subset \mathcal{X} \to \mathbb{R} | \forall x, x' \in I : |\phi(x) - \phi(x')| \leq L \; \mathtt{d}_{\mathcal{X}}(x, x').\right\} \tag{1}$$

of $L-$ Lipschitz continuous functions for a given $\mathtt{d}_{\mathcal{X}}$ and known *Lipschitz constant* L (where in [1] only the one-dimensional case with $\mathcal{X} \subset \mathbb{R}$, $\mathtt{d}_{\mathcal{X}}(x, x') = |x - x'|$ is considered).

Once sample $\mathcal{D}_n$ is known, we can combine the information contained therein with the information available a priori. At the very least, we can define a *posterior hypothesis space* $\mathcal{K}_{post}$ by falsification. That is, we eliminate all hypotheses from $\mathcal{K}_{prior}$ that could not have generated the sample. In the case of noise-free data, this might be done as per $\mathcal{K}_{post}(\mathcal{D}_n) = \mathcal{K}_{prior} \cap \mathcal{K}(\mathcal{D}_n)$ where $\mathcal{K}(D_n) = \Big\{\phi : \mathcal{X} \to \mathcal{Y} | \, \forall i \in \{1, ..., N_n\} : \; \phi(s_i) = f_i\Big\}$ is the set of *sample-consistent* functions. In many cases, forming this posterior set facilitates the derivation of worst-case bounds on the inferences made. For instance, in the aforementioned case of $\mathcal{K}_{prior} = \mathrm{Lip}_I(L)$ being Lipschitz functions with a known Lipschitz constant L, it is possible to compute Lipschitz continuous upper and lower bound functions $\mathfrak{u}_n(\cdot; L) : x \mapsto \sup_{\phi \in \mathcal{K}_{post}} \phi(x)$ and $\mathfrak{l}_n(\cdot; L) : x \mapsto \inf_{\phi \in \mathcal{K}_{post}} \phi(x)$, respectively [2,3,11] that allow bounding the maxima, minima, values and definite integrals of any sample-consistent target function. In this case, $f \in \mathcal{K}_{post} \subseteq \mathcal{E}_{\mathfrak{l}_n}^{\mathfrak{u}_n}(L) = \{\phi | \mathfrak{l}_n(x; L) \leq \phi(x) \leq \mathfrak{u}_n(x; L), \forall x\}$ where the set $\mathcal{E}_{\mathfrak{l}_n}^{\mathfrak{u}_n}(L)$ of all functions bounded by functions $\mathfrak{l}_n(\cdot;L), \mathfrak{u}_n(\cdot;L)$ shall be called the *enclosure*. Since the bounding functions depend on Lipschitz constant L, the enclosure also depends on this constant. In [3], we have shown that both bounding functions uniformly converge to f in the limit of a dense sample.

In the quadrature case, it follows with ease that the definite integral on compact domain $I \subset \mathcal{X}$ of any sample-consistent $L-$ Lipschitz function is contained in the interval $[S_l(L), S_u(L)]$, where $S_l(L) = \int_I \mathfrak{l}_n(x;L) \,\mathrm{d}x$ and $S_u(L) = \int_I \mathfrak{u}_n(x;L) \,\mathrm{d}x$ are the integrals of the upper and lower bound functions, respectively. That is,

$$\forall f \in \mathcal{K}_{post}(L) : \int_I f(x) \,\mathrm{d}x \in \big[S_l(L), S_u(L)\big]. \tag{2}$$

The integrals $S_u(L), S_l(L)$ can be computed in closed-form for one-dimensional compact domains $I \subset \mathbb{R}$ [1]. Therefore, they are suitable bounds for the definite integral of the target function for any target $f \in \mathcal{K}_{post}$ with known Lipschitz constant L.

Apart from such worst-case bounds, probabilistic approaches have become increasingly popular. In Bayesian inference, a priori knowledge is stated via probability measures that express subjective beliefs over the veracity of hypotheses about the ground-truth (e.g. of the integrand being some $\phi \in \mathcal{K}_{prior}$). In Bayesian inference with Gaussian processes (GPs) [9] one typically chooses a GP prior over the integrand [5–8]. In the basic setup, here the a priori set coincides with the closure of the span of eigenfunctions of the kernel operator of the covariance function of the prior [10]. The prior encodes a probabilistic belief over this set K_{prior} of candidate integrands. A sample of the integrand is then used to compute a resulting posterior belief measure over the integrand. This posterior is then utilised in various ways to obtain information about the integral.

In contrast to these works, we assume $\mathcal{K}_{prior} = \bigcup_{L^* \geq 0} \mathrm{Lip}_I(L^*)$ is the set of all Lipschitz functions with any constant. We implicitly place a prior distribution over $\mathcal{K}_{prior}$ by placing a distribution over the smallest Lipschitz constant L^* of the integrand. The posterior is inferred on the basis of the observed data $\mathcal{D}_n$. Connecting to the worst-case approaches in the aforementioned Lipschitz case, we can obtain probabilistic posterior bounds that also apply to multi-dimensional integration-domains. While in general, computation of these bounds might involve the necessity to evaluate intractable integrals, we propose to consider the conjugate pair of uniform likelihoods and Pareto priors. This yields posterior densities and bounds that can be computed in closed-form.

2 Bayesian Lipschitz Constant Inference and Integral Bounds

Inference over the Lipschitz constant of the integrand. In Eq. 2 we provided efficiently computable worst-case bounds around the integral estimates. These bounds hold provided the target is Lipschitz with some known constant $L \in \mathbb{R}_{\geq 0}$. In practice however, one may be uncertain about the constant. In this case, it may be desirable to (i) take this uncertainty into account when utilising the resulting prediction bounds and (ii) ideally, one desires an inference rule that can update one's belief over the Lipschitz constant as more data becomes available.

To address these desiderata we propose a Bayesian approach as follows:

Let $\pi_0 : \mathbb{R}_+ \to \mathbb{R}_+$ be an a priori density over smallest Lipschitz constant $L^* = \min\{L \geq 0 | f \in \mathrm{Lip}_I(L)\}$ of the integrand f. According to Bayes' theorem, we can calculate the posterior density $\pi(L^* = \cdot | \mathcal{D}_n)$ over the smallest Lipschitz constant as per

$$\pi(L^* = \ell | \mathcal{D}_n) = \frac{\pi(\mathcal{D}_n | L^* = \ell)\, \pi_0(L^* = \ell)}{\int_{\mathbb{R}} \pi(\mathcal{D}_n | L^* = \ell)\, \pi_0(L^* = \ell) \,\mathrm{d}\ell}.$$

In general, the posterior density $\ell \mapsto \pi(L^* = \ell | \mathcal{D}_n)$ may not be computable in closed- form and might have to be approximated utilising standard quadrature methods. Instead, we propose to utilise a combination of Pareto priors with uniform likelihoods. Making simplifying assumptions, this allows us to model maximal ignorance over the true best Hölder constant L^* while at the same time allowing for closed-form expressions.

For a choice of two-dimensional indices $k \in \mathcal{I}_n \subseteq \{(i,j) | \, \mathtt{d}_{\mathcal{X}}(s_i, s_j) > 0, j < i, i, j = 1, \ldots, N_n\}$ we define the ratios $r_k := \frac{|f_i - f_j|}{\mathtt{d}_{\mathcal{X}}(s_i, s_j)}$ which can be computed from the sample $\mathcal{D}_n$. A priori we are uncertain about them. Hence each a priori uncertain r_k can be modelled as a r.v. R_k. By definition

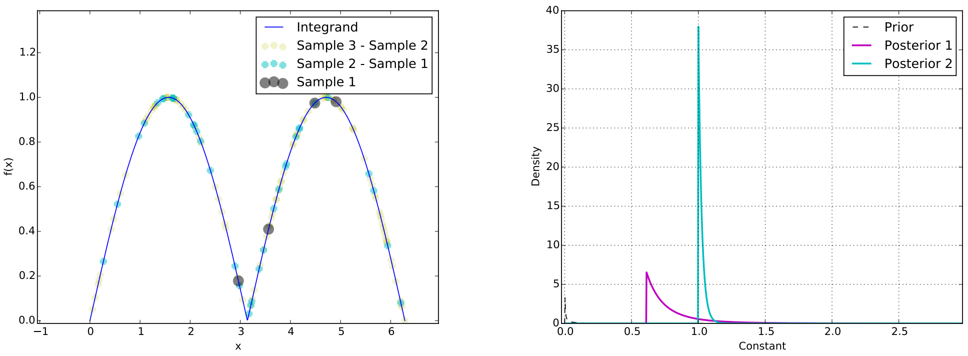

Figure 1: Bayesian Lipschitz constant estimation of the underlying integrand $f : x \mapsto |\sin(x)|$ on domain $I = [0, 2\pi]$. **Left plot:** Depicts f as well as the random samples $\mathcal{D}_1 \subset \mathcal{D}_2 \subset \mathcal{D}_3$ generated i.i.d. uniformly at random. We had $|\mathcal{D}_1| = 4, |\mathcal{D}_2| = 40, |\mathcal{D}_3| = 140$. **Right plot**: The approximately uninformative prior over the best Lipschitz constant of the integrand (blue plot), as well the posterior density functions conditioned on ratio sequences ϱ_1 (magenta plot) and ϱ_2 (cyan plot) obtained by computing all ratios from the data sets $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively. Note, how the posterior densities gravitate towards the true best Lipschitz constant $L^* = 1$ with increasing sample size ($\pi(\cdot|\varrho_3)$ omitted since it was a very high and narrow peak around $\ell = 1$).

of the best Lipschitz constant $L^* = \sup_{s,s'} \frac{\left|f(s)-f(s')\right|}{\mathrm{d}_{\mathcal{X}}(s,s')}$ we have $L^* \geq \max_{k\in\mathcal{I}_n} r_k =: \hat{L}(n)$. We assume that the sample generally does not contain additional information about L^* beyond the chosen ratio sequence $\varrho_n := \Big(r_k\Big)_{k\in\mathcal{I}_n}$ and hence, $\pi(L^* = \cdot|\mathcal{D}_n) = \pi\big(L^* = \cdot|\varrho_n\big)$.[1] In the absence of additional a priori assumptions a suitable ignorance likelihood is the uniform density

$$\pi\big(R_k = r_k|L^* = \ell\big) = \begin{cases} 0, r_k > \ell \\ \frac{1}{\ell}, \text{otherwise.} \end{cases}$$

Assume we choose the prior $\pi_0 : \ell \mapsto \pi_0(L^* = \ell)$ to follow some Pareto distribution $Pa(b, \nu)$. Under the assumption of independence of the chosen ratios in ϱ_n we can reduce our problem to the "taxicab" problem discussed by Minka [4] and appealing to conjugacy, the posterior $\pi\big(L^* = \cdot|\varrho_n\big)$ again follows a Pareto distribution $Pa(m_n, |\mathcal{I}_n| + \nu)$ with density

$$\pi(L^* = \ell|\varrho_n) = \begin{cases} \frac{(|\mathcal{I}_n|+\nu){m_n}^{|\mathcal{I}_n|+\nu}}{\ell^{|\mathcal{I}_n|+\nu+1}}, \ell \geq m_n \\ 0, \;\text{otherwise} \end{cases} \tag{3}$$

where $m_n = \max\{b, \max_{k\in\mathcal{I}_n} r_k\}$. Note, the case of an uninformative prior arises in the limit of $\nu, b \to 0$ [4].

An example of our Bayesian Lipschitz constant inference method is given in Fig. 1. Here we started with an approximately uninformative Pareto prior $Pa(0.001, 0.01)$ over L^*. Providing three samples of increasing size we documented the resulting posterior densities. As expected, they seemed to converge to the correct Dirac posterior $\delta(\cdot - L^*)$ with increasing sample size.

Application to Lipschitz quadrature. We can now turn to the question of utilising the posterior belief over the best Lipschitz constant in our integration bounds. Statement (2) holds, provided the parameters L is chosen at least as large the smallest Lipschitz constant L^* of the target function. Thus, in the case of Bayesian uncertainty, if we set L to a value L_θ such that the probability of $L_\theta < L^*$ is less than θ, the prediction bounds $S_l(L_\theta), S_u(L_\theta)$ are valid with probability of at least $1 - \theta$. Under the assumptions described above that yield the posterior Pareto density $\pi(\cdot|\varrho_n)$ as per Eq. 3, we could find L_θ as follows: we insert L_θ into the pertaining cdf of the posterior Pareto distribution. Setting the resulting term to less than or equal to θ yields an inequality. Solving for this L_θ yields the desired $(1 - \theta)$ -confidence bound on the Lipschitz constant. For the posterior as per

[1] During the workshop, we have come to believe that this might not be true in general and hence, conditioning on all ratios might affect the inference in a manner that will be subject to further investigation.

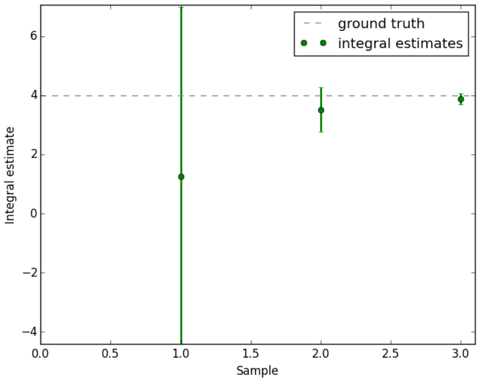

Figure 2: Example quadrature problem of estimating the integral $\int_0^\pi |\sin(x)| \, \mathrm{d}x = 4$. The Bayesian integral estimates pertaining to posterior densities $\pi(\cdot|\varrho_1), \pi(\cdot|\varrho_2)$ and $\pi(\cdot|\varrho_3)$ are the dots at abscissa positions 1.0, 2.0 and 3.0, respectively. The error bars depict the error intervals $[S_l(L_{\theta,i}), S_u(L_{\theta,i})]$ for $\theta = 0.001$ and $i = 1, ..., 3$. Note how the error bars shrink with increasing data and quite accurately estimate the true integral for the largest sample $\mathcal{D}_3$.

Eq. 3 this gives $L_\theta = \max\{\hat{L}(n), \hat{L}(n) \exp(-\frac{\log \theta}{|\mathcal{I}_n|})\}$. This bound can then be used as the Lipschitz parameter in a Lipschitz quadrature rule.

To give a concrete example, we computed three such bounds, $L_{\theta,1}, L_{\theta,2}, L_{\theta,3}$, on the basis of the posterior densities $\pi(\cdot|\varrho_1), \pi(\cdot|\varrho_2)$ and $\pi(\cdot|\varrho_3)$, respectively. Here we opted for high-confidence integration bounds choosing $\theta = 0.001$. We then computed three integral estimates and bounds of the integrand $f(x) = |\sin(x)|$ using a standard Lipschitz quadrature rule [1] parameterised by $L_{\theta,1}, L_{\theta,2}, L_{\theta,3}$, respectively. The resulting integration estimates and bounds are depicted in Fig. 2. The plots illustrate that our method improved in predicting the correct integral with increasing sample size. Furthermore, the ground truth integral $\int_0^{2\pi} f(x) \, \mathrm{d}x = 4$ always was contained in the confidence error intervals $[S_l(L_{\theta,i}), S_u(L_{\theta,i})]$ $(i \in \{1, 2, 3\})$.

Complexity. In general, the Bayesian inference rule requires a numerical approximation method which will dominate the computational effort. In contrast, for our closed-form expressions derived for the Pareto priors, the computational complexity of L_θ scales linearly in the number of ratios $|\varrho_n|$, which, if all ratios are chosen to be conditioned on, is at most quadratic in the sample size. The Lipschitz quadrature method's complexity is linear in the sample size. So, the effort ranges from linear to quadratic, depending on how many ratios we choose to condition on.

3 Conclusions

We started with introducing a Bayesian approach for inferring the Lipschitz constant of a function based on a finite sample. From the posterior densities we can derive confidence bounds on the best Lipschitz constant which we proposed to utilise in conjunction with Lipschitz quadrature methods. This merger yields integral estimates that hold with adjustable subjective probability.

When employing Pareto priors to model one's uncertainty about the best Lipschitz constant one can model ignorance priors. Furthermore, we have given closed-form expressions of the estimates and bounds of the definite integral by conditioning on the chosen slope ratios computed from the sample. The expressions were derived under the assumption that these ratios are independently uniformly distributed given the Lipschitz constant.

Future work will further elucidate these assumptions and the impact of this approach on the conservatism of the resulting bounds. In particular, we need to understand how the resulting confidence bound L_θ relates to the true $1-\theta$-confidence bound of the best Lipschitz constant one would obtain from the full Bayesian posterior distribution conditioned on the sample rather than the ratio sequence. This might require studying the true full likelihood $\pi(\mathcal{D}_n|L^*)$. To this end, a generative model of Lipschitz functions might have to be considered. Answering these question might also shed light on the question of which ratios one should ideally condition on.

Our exposition focussed on the case of definite integration on one-dimensional domains. Extensions to multi-dimensional and unbounded domains, as well as to noisy samples, are in preparation. Moreover, we would like deploy our quadrature method in the context of inference and control. Here, the integration bounds can be useful to avoid underestimating the risk of a control decision. This can become especially important in situations where real-time requirements prohibit the evaluation of a large number of sample points of the integrand.

Acknowledgements

I would like to thank Michael Osborne and Stephen Roberts for useful discussions, as well as anonymous reviewers for helpful feedback. I am grateful to Guillaume Dehaene whose comments pointed me to some of the open issues related to the independence assumption of the ratios and whose investigation opens opportunities for future work. Furthermore, I am grateful for support from EPSRC, grant number NMZR/031 RG64733, received in the context of the *AIS* project.

References

[1] B. Baran, E. D. Demaine, and D. A. Katz. Optimally adaptive integration of univariate Lipschitz functions. *Algorithmica*, 2008.

[2] G. Beliakov. Interpolation of Lipschitz functions. *Journal of Computational and Applied Mathematics*, 2006.

[3] Jan-Peter Calliess. *Conservative decision-making and inference in uncertain dynamical systems*. PhD thesis, University of Oxford, 2014.

[4] T. P. Minka. Bayesian inference of a uniform distribution. http://research.microsoft.com/en-us/um/people/minka/papers/uniform.html, January 2001.

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

[6] A. O'Hagan. Bayes-hermite quadrature. *J. of Stat. Planning and Inference*, 29, 1991.

[7] M.A. Osborne, D.K. Duvenaud, R. Garnett, C.E. Rasmussen, S.J. Roberts, and Z. Ghahramani. Active Learning of Model Evidence Using Bayesian Quadrature. In *Advances in Neural Information Processing Systems (NIPS)*, pages 46–54, 2012.

[8] M.A. Osborne, R. Garnett, S.J. Roberts, C. Hart, S. Aigrain, and N. Gibson. Bayesian quadrature for ratios. In *International Conference on Artificial Intelligence and Statistics*, pages 832–840, 2012.

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

[10] M. Seeger. *Bayesian Gaussian Process Models: PAC-Bayesian Generalisation Error Bounds and Sparse Approximations*. PhD thesis, University of Edinburgh, 2005.

[11] B. Shubert. A sequential method seeking the global maximum of a function. *SIAM J. on Numerical Analysis*, 9, 1972.