Outline for Probabilitstic numerics/Laplace approximation proof-of-concept paper

Shaun McDonald, Dave Campbell

April 28, 2020

Abstract

In many statistical models, we need to integrate functions that may be high-dimensional. Such integrals may be impossible to compute exactly, or too expensive to compute numerically. Instead, we can use the *Laplace approximation* for the integral. This approximation is exact if the function is proportional to the density of a normal distribution; therefore, its effectiveness depends intimately on the true shape of the function. To assess the quality of the approximation, we use *probabilistic numerics*: recasting the approximation problem in the framework of probability theory. In this probabilistic approach, uncertainty and variability don't come from a frequentist notion of randomness, but rather from the fact that the function may only be partially known. We will use this framework to develop a diagnostic tool for the Laplace approximation, modelling the function and its integral as a Gaussian process and devising some kind of test by conditioning on a finite number of function values. We will discuss approaches for optimizing the tool, choosing various parameters to maximize our ability to detect non-Gaussianity. Our methods have been demonstrated on a variety of distributions, showing the effects of shape and dimension on the optimization. We will also discuss the possibility of using statistical ideas like hypothesis testing in this unconventional non-random paradigm.

1 Framework

- We have the **full function** $f(\theta, t)$, with structural parameters $\theta \in \mathbb{R}^d$ and nuiscance parameters $t \in \mathbb{R}^d$ (typically f is some probability density)
- We are interested in the **target function** $\int f(\theta, t)\mathrm{d}t$ (function of θ only)
- In what follows, we will (mostly) ignore θ for simplicity (i.e. simply view full f as function of t)
- Assumptions on f:
    1. Even if we can't easily evaluate its integral or describe it analytically, we can easily evaluate it at any finite number of points.

2. It has a maximum (mode) at some $\hat{t}$. We can easily determine $\hat{t}$.
3. We can easily evaluate the second derivative (or the Hessian matrix, for multivariate t) of f at the mode $\hat{t}$.
4. f is log-concave - or at the very least, log-concave at $\hat{t}$.

- Then the **Laplace approximation** for the target is (based on a Taylor expansion for $\log f$)

$$\int f(t)\mathrm{d}t \approx \hat{L}(f) := f(\hat{t})\sqrt{2\pi\left|-\frac{\mathrm{d}^2}{\mathrm{d}t^2}\log f(\hat{t})\right|^{-1}}, \tag{1}$$

  which is exact if f is (proportional to) the pdf of a Normal r.v. with mean $\hat{t}$ and variance $\left|-\frac{\mathrm{d}^2}{\mathrm{d}t^2}\log f(\hat{t})\right|^{-1}$ (this is why we need the log-concavity assumption).

- Let $F(t) = \int_a^t f(z)\mathrm{d}z$ for some a (want to avoid improper integrals for practical reasons).

- **Main goal**: some type of probabilistic numerics-based inference (somewhat Bayesian, in a sense) to assess the quality of approximation 1, as in Zhou [2]

2 Probabilistic numerics

For inference, jointly model (f, F) as (μ_t, μ) with a **Gaussian process prior**: for two sets of points $\boldsymbol{t}_j, \boldsymbol{t}_k$ (where e.g. $\boldsymbol{t}_k = [t_{k1}, \ldots, t_{kN}]'$ is a set of N points in $\mathbb{R}^d$),

$$\begin{bmatrix}\mu_t(\boldsymbol{t}_j)\\ \mu(\boldsymbol{t}_k)\end{bmatrix} \sim \mathcal{GP}\left(\begin{bmatrix}m_t^0(\boldsymbol{t}_j)\\ m^0(\boldsymbol{t}_k)\end{bmatrix}, \begin{bmatrix}C_t^0(\boldsymbol{t}_j, \boldsymbol{t}_j) & \int_a^{\boldsymbol{t}_k} C_t^0(\boldsymbol{t}_j, \boldsymbol{z})\mathrm{d}\boldsymbol{z}\\ \int_a^{\boldsymbol{t}_k} C_t^0(\boldsymbol{z}, \boldsymbol{t}_j)\mathrm{d}\boldsymbol{z} & C^0(\boldsymbol{t}_k, \boldsymbol{t}_k)\end{bmatrix}\right), \tag{2}$$

where

1. Terms in 2 are simply vectors or matrices of elementwise evaluations (e.g. $m_0(\boldsymbol{t}_k) = [m_0(t_{k1}), \ldots, m_0(t_{kmN})]'$)
2. The **prior mean** m_t^0 is the kernel of a Gaussian with mean $\hat{t}$ and covariance $-\frac{d^2}{dt^2}\log f(\hat{t})^{-1}$, scaled by $f(\hat{t})$
3. $m^0(t) = \int_a^t m_t^0(z)\mathrm{d}z$. Note that $\lim_{t\to\infty, a\to-\infty} m^0(t) = \hat{L}(f)$, and integrals with respect to $z \in \mathbb{R}^d$ are defined in this manuscript as d-tuple integrals w.r.t. each component of z.

4. C_t^0 is a covariance operator defined by the self-convolution of some kernel R_λ (e.g. squared exponential or uniform):

$$C_t^0(t,u) = \alpha^{-1} \int_{\mathbb{R}} R_\lambda(t,z) R_\lambda(z,u) \mathrm{d}z$$

parameterized by **length-scale** λ and **precision** α - the hyperparameters of the GP

5. $C^0(t,u) = \int_a^t \int_a^u C_t^0(y,z) \mathrm{d}y \mathrm{d}z$

Inference about approximation 1 is based on $F(t)$ for some large t, via the "posterior distribution" of $\mu(t)$ conditional on some evaluations of f. Specifically, we obtain values for f at a grid of **interrogation points** $\boldsymbol{s} = (s_1, \ldots, s_n)$, where we'd typically take $s_1 = a$ and $s_n = t$ in the univariate case. Then we obtain the "posterior" [1]

$$\mu(t) \mid [\mu_t(\boldsymbol{s}) = f(\boldsymbol{s})] \sim \mathcal{GP}\left(m^1(t), C^1(t,t)\right), \tag{3}$$

$$m^1(t) = m^0(t) + \int_a^t C_t^0(z,\boldsymbol{s}) dz \left(C_t^0(\boldsymbol{s},\boldsymbol{s})\right)^{-1} \left(f(\boldsymbol{s}) - m_t^0(\boldsymbol{s})\right), \tag{4}$$

$$C^1(t,t) = C^0(t,t) - \int_a^t C_t^0(z,\boldsymbol{s}) \mathrm{d}z \left(C_t^0(\boldsymbol{s},\boldsymbol{s})\right)^{-1} \int_a^t C_t^0(\boldsymbol{s},z) \mathrm{d}z. \tag{5}$$

Remark: note that the posterior covariance update does *not* depend on f. On the other hand, if f is truly Gaussian (in which case the L.A. is exact), then $m^1 \equiv m^0$. Thus, one idea for inference is to compare 3 to the "null" $\mathcal{GP}\left(m^0(t), C^1(t,t)\right)$

3 Initial work: point placement for small grid

Our initial work on this (aside from Charlie's thesis [2]) focused mainly on the univariate case for simplicity, and considered a small grid of interrogation points:

$$\begin{aligned} \boldsymbol{s} &= [s_1, s_2, s_3] \\ &= \left[\hat{t} - \epsilon \sqrt{\left(-\frac{d^2}{dt^2} \log f(\hat{t})\right)^{-1}},\ \hat{t},\ \hat{t} + \epsilon \sqrt{\left(-\frac{d^2}{dt^2} \log f(\hat{t})\right)^{-1}}\right]. \end{aligned} \tag{6}$$

- One interrogation point at the mode, two equidistant boundary points
- ϵ = number of standard deviations (of the density corresponding to $\hat{L}(f)$) between interrogation points
- Range of integration is $a = s_1$, $t = s_3$ in the univariate case

- R_λ = squared exponential kernel (nice mathematical properties, and we don't need the sparsity of the uniform kernel with only three points)
- Following advice in [1, 2], hyperparameters set to $\lambda = 1.5(s_2 - s_1)$ and $\alpha = 1/(s_2 - s_1)$ (i.e. λ and α^{-1} proportional to distance between neighbouring points)
- **Idea**: optimize ϵ (and integration range and hyperparameters, indirectly) to be maximally sensitive to f deviating from Gaussian
  - The goal is to have the most powerful diagnostic - something that measures deviations from Gaussianity as effectively as possible
  - In view of the remark before this section, choose ϵ to maximize our ability to differentiate $\mathcal{N}\left(m^1(t), C^1(t,t)\right)$ from the "null posterior" $\mathcal{N}\left(m^0(t), C^1(t,t)\right)$
- Arguably the most natural way to do this is to maximize KL divergence as a function of ϵ. For two Gaussians with equal variance,

$$\begin{aligned}\mathrm{KL}(\epsilon) &= \text{bias}^2/(2 * \text{variance}) \\ &= \frac{\left[\int_a^t C_t^0(z, \boldsymbol{s})dz \left(C_t^0(\boldsymbol{s}, \boldsymbol{s})\right)^{-1} \left(f(\boldsymbol{s}) - m_t^0(\boldsymbol{s})\right)\right]^2}{2\left[C^0(t,t) - \int_a^t C_t^0(z, \boldsymbol{s})dz \left(C_t^0(\boldsymbol{s}, \boldsymbol{s})\right)^{-1} \int_a^t C_t^0(\boldsymbol{s}, z)dz\right]} \end{aligned} \tag{7}$$

- Can also view this as *power analysis*:
  - Here, the "null hypothesis" is essentially that f is Gaussian
  - Then at the typical 5% significance level, the "power" is

$$\begin{aligned}\mathrm{P}(\epsilon) &= \mathbb{P}\left(\left|\frac{X - m^0(t)}{C^1(t,t)}\right| > 1.96 \mid X \sim N\left(m^1(t), C^1(t,t)\right)\right) \\ &= \Phi\left(1.96 - \frac{\text{bias}}{\sqrt{\text{variance}}}\right) - \Phi\left(-1.96 - \frac{\text{bias}}{\sqrt{\text{variance}}}\right)\end{aligned} \tag{8}$$

- KL and power both depend entirely on bias-variance ratio
- Results *appear* to be equivalent for either, and KL is a bit nicer mathematically, so we focus on that

3.1 Results

Taking $f = f(\theta, t)$ to be some known pdf with parameter(s) θ, optimizing 7 w.r.t. ϵ, and seeing how $\hat{\epsilon} := \arg\max_\epsilon \mathrm{KL}(\epsilon)$ depends on θ (ex. f as a T-distribution with degrees of freedom θ). The examples we tried can lead to some general *conjectures*, but the formulae involved in the optimization are quite complicated so we have yet to devise more abstract/formal proofs. See JSM 2018 poster for more detail. For now, a tl;dr:

- With α and λ defined in terms of step size, scale parameters for f seem to "fall out" of the optimzation. This is good, since ϵ is standardized in terms of standard deviation
- For densities that have a non-differentiable "corner" at the origin (e.g. f a Gamma or Weibull with shape parameter ≤ 2), the optimal ϵ was such that $s_1 = 0$) (i.e. the left endpoint right on the corner). Curiously, $\hat{\epsilon}$ appears continuous as a function of the shape parameter, even as we approach these low values.
- For f a Weibull, the dependence of the KL function on skew is quite intricate. When f is negatively skewed (i.e. large shape parameter), the KL function is bimodal with a larger peak around $\epsilon \approx 2.5$. The relationship between skew and $\hat{\epsilon}$ is asymmetric.
- For f a T-distribution, $\hat{\epsilon}$ varies almost linearly with $f(s_3) - m_t^0(s_3)$. This behaviour has eluded any kind of analytic verification; it could be coincidence.
- Even with symbolic computation, analytic optimization is impossible because of the dependence of 7 on $f - m_t^0$. The most general we can get is to have rule(s) of thumb, possibly dependent on the shape of f. Aside from the aforementioned skew-based findings, taking $\hat{\epsilon}$ between 1.5 and 1.75 seems reasonable
- For multivariate densities, say $f(x, y; \theta)$, we had three 1-D diagnostics per coordinate based on conditionals: looking at $F_{X|Y}(s_3^X \mid y = s_j^Y)$, $j = 1, 2, 3$ (i.e. integrating along one axis while controlling the location of the "slice" on the other)
  - Can assess quality of L.A. *and* dependence between coordinates by comparing the different posteriors
  - The "banana" example from the poster shows that it's hard to differentiate between shift and shape effects when we're assessing the 2-D L.A. This would make it hard for us to see exactly *how/why* the multidimensional L.A. fails

4 Some ideas for further work

- **Note**: although a Gaussian shape for f implies that the approximation 1 is exact, the converse is ***not*** true
  - One example: if f is the pdf of a T-distribution with 2 degrees of freedom truncated to $(-\infty, t^*)$ for some $t^* > 0$, it's possible to choose t^* such that $\int f(t)\mathrm{d}t = \hat{L}(f)$
  - Thus, we should ask what's the more useful thing to diagnose: non-Gaussian f, or failure of the L.A. (where the latter implies the former but not vice-versa)?

    - Use case for the former: checking the performance of methods based on Gaussian asymptotics for finite samples
    - Use case for the latter: checking anything where we have to marginalize (e.g. hierarchical Bayesian models)
  - This will impact the way we structure inference, etc.

- Probably good to consider increasing number of interrogation points n, rather than just optimizing the two boundary points as in Section 3
  - Adapting proofs from Chkrebtii et al. [1], we should get 3 converging in some sense to the true $F(t)$ (depending on "big-O" behaviour of hyperparameters)
  - How should integration boundary (a, t) be chosen? Can either have fixed boundary (perhaps symmetric about the mode such that $m^0(t) = (1-\delta)\hat{L}(f)$ for some small δ) or allow it to expand as n increases [2]. Depends on inferential goals and asymptotics
  - How to reconcile asymptotics for $n \to \infty$ with the Section 3 idea of optimizing distance between points?

- Testing ideas:
  - Uncertainty here doesn't come from anything truly random *per se*, but rather from unknowns about f itself (sort of Bayesian?)
  - The "null hypothesis" is something like[1] $F(t) = m^0(t)$, with inference for F based on the "sample" $f(\boldsymbol{s})$
  - Based on Section 2, a reasonable "test statistic" for this hypothesis would be something like

$$Z = \frac{m^1(t) - m^0(t)}{\sqrt{C^1(t,t)}} \tag{9}$$

    and then we would reject the null if it doesn't fall within the standard normal critical values
  - Could also use standard Bayesian inference ideas: put a prior probability on the null hypothesis $F(t) = m^0(t)$, use the GP 2 as the prior conditional on the alternative, and look at Bayes factors and/or posterior odds
  - **Potential problem**: rejecting the null even when f is really close to a Gaussian (e.g. a non-truncated T-distribution with high d.f.)
    - Could be an adverse side-effect of the KL-based stuff in Section 3

[1] As per the note at the start of this section, this may change depending on our exact goals. To that end, rather than restricting ourselves to the posterior of $\mu(t)$ we may be able to make use of the entire posterior GP $[\mu_t((\boldsymbol{t}_j), \mu(\boldsymbol{t}_k)] \mid f(\boldsymbol{s})$, which can be easily derived from 2 [e.g. 2].

    - On the asymptotic side, we'd expect that as $n \to \infty$ the numerator of 9 will approach $\int f(t)\mathrm{d}t - \hat{L}(f) \neq 0$, while the denominator (variance) will approach 0. Thus, for a large interrogation grid rejecting the null is inevitable, even if the L.A. is close enough for practical purposes
    - **Idea**: use an *interval null hypothesis*: that $F(t)$ is in some small interval centered around $m^0(t)$
    - This also gets us around the weirdness inherent to Bayesian point hypothesis testing
    - **Challenge**: how to choose the width of the interval there?

- **Things to do, in summary**: asymptotics for size and spacing of interrogation grid, sensible way of extending ideas to multivariate integration, formalizing reasonable test procedures and their properties

- Multivariate stuff:
    - Our JSM approach - looking at multiple diagnositics (centered at n-variate Laplace approximation) for each univariate *conditional* - may be the wrong way to think about it (after hearing from Richard). Does it truly give us information about our quantity of interest?
    - We seem to have decided the integral (i.e. actual Laplace approx) is more important to us than the function shape. To that end, we should probably be conducting inference for the multiple integral, with the appropriate multivariate Gaussian as the prior mean
    - Is it possible to do a multivariate diagnostic in "blocks"?

- Richard points out that for meaningful asymptotics/Gaussian Process theory, we need the covariance operator for the full function to converge. More broadly, the three properties we'd like are

    1. $\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} C_0^t(y,z)\mathrm{d}y\mathrm{d}z < \infty$
    2. $C_t^0(\boldsymbol{s},\boldsymbol{s})$ invertible for any nondegenerate set of interrogation points (i.e. the operator is positive-definite in a "nontrivial" way)
    3. $C_0(t,u) = \int_a^t \int_a^u C_0^t(y,z)\mathrm{d}y\mathrm{d}z$ is in closed form.

    The first one isn't an issue if the domain of t is bounded (as in the methodology of Chkrebtii et al. [1]), but if the idea is that the nuisance parameters are asymptotically Gaussian (the entire premise of the LA), a bounded domain may not make sense. Perhaps it's okay in practice to get around this by assuming a *very* large bounded interval, but for any kind of sensible asymptotics I assume the above condition would still be a requirement. Of course, the covariance kernels considered thus far don't satisfy this - they all lead to covariance shapes (in the 1-D case) that have constant height along the line $x = y$, and therefore, infinite volume. We'd need the height to decay suitably to ensure a finite integral. More on this later.

The second requirement is necessary for computation, since we have to invert a matrix of covariance values at interrogation points. In particular, it rules out any covariance operators that factorize as $C_0^t(t,u) = g(t)h(u)$ (I did some experimentation with one that unfortunately ended with this realization).

The third one isn't strictly necessary, but would be nice for computational convenience, especially for any optimization. For practical purposes, the important thing is to have C_0 that can be evaluated reasonably easily at high dimensions (to this end, maybe we can broaden "closed-form" to mean "anything for which MATLAB has a built-in function").

4.1 Options for covariance kernels

Below are some options for C_0^t as per the considerations above. It's easier to discuss these for the one-dimensional case here for visualization, etc

- The pdf of a normal distribution with mode at $\hat{t}$ and countours aligned along the line $x = y$. Essentially it's the squared exponential kernel, but modified so that its magnitude decays to 0 as we move farther from $\hat{t}$. Analogously to the squared exponential kernel, it should produce very smooth trajectories. In this case, C_0 will be the cdf a multivariate normal and won't have a closed form, but MATLAB has a function built in for this.

- Multiplying the uniform kernel by some exponential term. $\exp(-x^2 - y^2)$ is a good option for smoothness, but something like $\exp(-|x| - |y|)$ is the only thing I can think of that will have a closed form.

5 New experiments with a finite-volume covariance kernel

In this section, we assume w.l.o.g. that f has a mode at the origin, and the Hessian of $\log f$ (denoted here as H) is the negative of the identity matrix there, i.e. $H(0) = -I$. This ensures that the Gaussian approximation to f, m_0^t, is proportional to a standard d-dimensional Normal pdf, and the axes of the domain correspond to the principal components of H at the mode. In the bivariate case, for example, placing interrogation points at $(2,0)$ and $(0,2)$ therefore amounts to placing them 2 "standard deviations" from the mode along the first and second "principal axes", respectively. If these assumptions are not true, it is always possible to enforce them by replacing the function $f(t)$ with $f\left(V\left[\sqrt{-D}\right]^{-1}t + \hat{t}\right)$, where the matrices V and D arise from the eigendecomposition $H\left(\hat{t}\right) = VDV'$.

As discussed in Section 4.1, here we experiment with a generalized covariance kernel which allows for finite integrals over the whole of $\mathbb{R}^d$. First recall the usual (convoluted) squared exponential covariance as used in Chkrebtii et al. [1]. As

in Zhou [2], we define it on $\mathbb{R}^d$ by componentwise multiplication of the univariate versions:

$$C_t^0(t,u) = \left(\frac{\sqrt{\pi}\lambda}{\alpha}\right)^d \exp\left[-\frac{\|t-u\|^2}{4\lambda^2}\right], \tag{10}$$

where λ is the *length-scale* parameter that controls prior covariance between distinct points (and therefore smoothness), and α is the *precision* parameter that controls the overall scale of prior variance[2].

To ensure finite variance for the integral of μ_t over $\mathbb{R}^d$, we modify (10) as follows.

$$C_t^0(t,u) = \left(\frac{\sqrt{\pi}\lambda}{\alpha}\right)^d \exp\left[-\frac{\|t-u\|^2}{4\lambda^2}\right] \exp\left[-\frac{\|t\|^2+\|u\|^2}{4\gamma^2}\right]. \tag{11}$$

The extra factor ensures that the variance of $\mu_t(t)$, $C_t^0(t,t)$, is largest at the origin (where the mode is assumed to be) and decays for large $|t|$, with the amount of decay controlled by the additional hyperparameter γ. A smaller γ-value causes the variance to shrink more rapidly as t moves away from the origin, whereas $\gamma \to \infty$ returns to (10), in which the variance does not depend on t. For the case $d = 1$, a comparison between (10) and (11) with all hyperparaeters equal to 1 is shown in Figure 1. The left plot highlights that the usual squared exponential kernel is only suitable when we wish to integrate f over a finite domain, as its total volume is infinite. We also see on the right plot that the modified kernel is proportional to a Normal p.d.f. with contours oriented along the main diagonal $x = y$.

Using (11) with $\gamma \in (0,\infty)$, the GP prior (2) from Section 2 can be used to model the target $\int_{\mathbb{R}^d} f$ *without* truncating the integral bounds. In particular, the integration bounds need not be determined by the extreme points of the interrogation grid $\boldsymbol{s}$. If integration *is* over the whole domain, then the Gaussian μ used to model $\int f$ has prior mean $m^0 = \int_{\mathbb{R}^d} m_t^0(z)\mathrm{d}z = \hat{L}(f)$ and prior covariance

$$\begin{aligned} C^0 &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} C_t^0(y,z)\mathrm{d}y\mathrm{d}z \\ &= \left[\frac{4\gamma^2\lambda^2\sqrt{\pi^3}}{\alpha\sqrt{2\gamma^2+\lambda^2}}\right]^d. \end{aligned} \tag{12}$$

If the integration bounds are finite[3], say $(a,a,\ldots,a),(b,b,\ldots,b) \in \mathbb{R}^d$ with $a < b$, it is readily seen from (11) that the prior variance of μ is equal to (12)

[2]Note that λ *also* influences the prior variance scale, since the self-convolution construction [1] causes an additional factor of λ^d. Of course, its effect can be offset by choosing α proportionally.

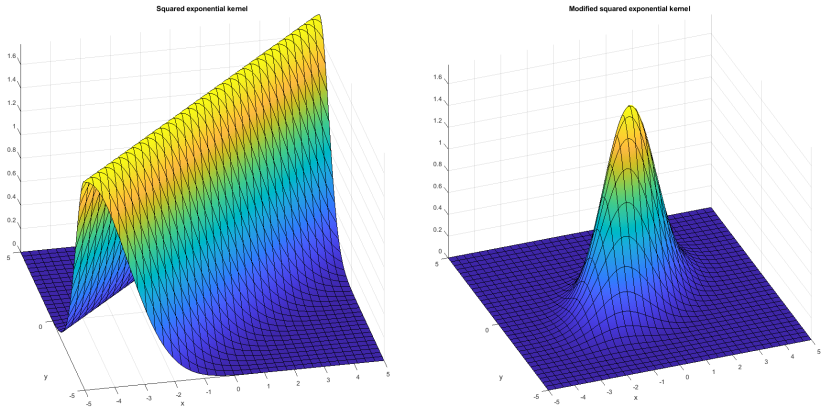

Figure 1: The usual squared exponential kernel with $\lambda = \alpha = 1$ (left) vs. our modified decaying version with $\lambda = \alpha = \gamma = 1$ (right).

times

$$\left[\int_a^b \int_a^b \phi(y,z)\mathrm{d}y\mathrm{d}z\right]^d,$$

where ϕ is the pdf of a bivariate Normal random variable with mean $(0,0)$ and covariance matrix

$$\frac{2\gamma^2}{\lambda^2+2\gamma^2}\begin{bmatrix}\lambda^2+\gamma^2 & \gamma^2 \\ \gamma^2 & \lambda^2+\gamma^2\end{bmatrix}$$

5.1 Results

Here we show some simple experiments conducted with the framework described above. Recalling that the domain is linearly transformed for computations (as described at the beginning of the section), most experiments use "cross-shaped" grids of interrogation points, placing them at half-integer multiples up to 3 "standard deviations" from the mode/orgin along each "principal axis". In mathematical terms, unless otherwise stated the interrogation points are of the form $\pm\frac{m}{2}e_i$, where m is an integer between 0 and 6 and e_i is the i^{th} standard basis vector of $\mathbb{R}^d$. All plots are presented on the *original* domain: the interrogation points $\boldsymbol{s}$ are plotted as $V\left[\sqrt{-D}\right]^{-1}\boldsymbol{s}+\hat{t}$, and the posterior of μ is shown after rescaling by $\left|V\left[\sqrt{-D}\right]^{-1}\right| = \sqrt{|H(0)|^{-1}}$.

[3]Here we assume for computational convenience that all coordinates of the integration bounds are equal. Recalling that f is scaled such that m_t^0 has equal variance along each axis, this is a reasonable assumption.

References

[1] Oksana A Chkrebtii, David A Campbell, Ben Calderhead, and Mark A Girolami. Bayesian Solution Uncertainty Quantification for Differential Equations. *Bayesian Analysis*, 11(4):1239–1267, 2016. doi: 10.1214/16-BA1036. URL `https://projecteuclid.org/download/pdfview{_}1/euclid.ba/1473276259`.

[2] Haoxuan Zhou. *Bayesian Integration for Assessing the Quality of the Laplace Approximation*. PhD thesis, Simon Fraser University, nov 2017. URL `http://summit.sfu.ca/item/17765`.