Optimization of the Placement of Interrogation Points

February 9, 2018

We will use a Gaussian Process (GP) based probabilistic numerical integrator to integrate a target function f. The goal is to determine whether or not a Laplace approximation was an adequate integral approximation.

In this paper, the goal is to optimize the GP parameters $\{\alpha, \lambda\}$ and the placement of the interrogation points $\{s_1, \ldots, s_N\}$ (presumably with, say, $N = 5$) so as to meet the following desiderata:

- If the target function is Gaussian we want a small value of our deviation metric.
- If the target function is skewed or heavy tailed we want a big value of our deviation metric.
- We wish to control the variance of the estimator.

Those will ensure that our method can accurately characterize whether or not the Laplace approximation is appropriate for the target function. For the Laplace approximation to be exact we need to have a target function which is a Gaussian.

For the sake of simplicity for now, we'll restrict ourselves to the one-dimensional case and assume f, its derivatives, and the location and value of its maximum are all known.

Consider using the probabilistic integrator to obtain the integrated target function over the interval $[t_1, t]$ (following the conventions of Charlie's thesis, we use t and s to respectively denote interrogation and evaluation points, although there may be some overlap as discussed below). As the interval narrows, the variance decreases, but the ability to distinguish between target function shapes will diminish. As the interval widens, the variance will increase, masking the distinction between distinguishing functional features.

Consider the posterior distribution, P, for the integrated target function over the interval $[t_1, t]$. Using our one-shot updater, P is a $N\left(m^1(t), C^1(t,t)\right)$ density. For a Gaussian target, the posterior for the integrated function will be $Q = N\left(m^0(t), C^1(t,t)\right)$ (refer to Charlie's thesis for the relevant formulas).

Maximizing the KL divergence between these two densities would ensure that we can tell the difference between a Gaussian and non-Gaussian target function, noting that it will always be zero for a truly Gaussian target:

$$D_{kl}(P||Q) = \int_{-\infty}^{\infty} P(x) \log\left(\frac{P(x)}{Q(x)}\right) dx \tag{1}$$

$$= \frac{\left(m^1(t) - m^0(t)\right)^2}{2C^1(t,t)}$$

$$= \frac{(\text{Bias})^2}{2 * \text{variance}}$$

$$= \frac{\left[\left(\int_{t_1}^{t} C_t^0(z, \boldsymbol{s})dz\right)\left(C_t^0(\boldsymbol{s}, \boldsymbol{s})\right)^{-1}\left(f(\boldsymbol{s}) - m_t^0(\boldsymbol{s})\right)\right]^2}{2\left[C^0(t,t) - \left(\int_{t_1}^{t} C_t^0(z, \boldsymbol{s})dz\right)\left(C_t^0(\boldsymbol{s}, \boldsymbol{s})\right)^{-1}\left(\int_{t_1}^{t} C_t^0(z, \boldsymbol{s})dz\right)^T\right]}. \tag{2}$$

However this does not consider the quality of the fit to a Gaussian target density. (*Not sure what you mean by this? -Shaun*) **It's possible that we can ignore this because it might not be a problem at all.**

Note that maximizing (1) occurs with respect to all of $\{\alpha, \lambda, s_1, \ldots, s_N\}$ but we could place the mid point $s_{(N+1)/2}$ at the peak of the target function and assume symmetry between the other interrogation points in order to reduce the search space and allow for skew in different directions. Moreover, it may be reasonable to fix $s_1 = t_1$ and $s_N = t$, i.e. set the interrogation endpoints to be equal to the evaluation endpoints (which are chosen to ensure sufficiently high coverage probability for the Laplace approximation, e.g. with a simple application of the 1.96 rule for normal distributions). If we did make our interrogation points symmetric about the maximum of the target, we would then only need to consider the distance between the midpoint and the other interior points (likely in terms of standard deviations of the target).

However, note that equation 2 is **very very ugly**. It would probably be easier to optimize the log of this expression, but either way it will involve heavy matrix calculus (unless we figure out a non-derivative-based method for this) that's largely dependent on (a) the form of f (assuming a member of the exponential family may be a workable amount of generality) and (b) the covariance kernel used (perhaps stick with squared exponential at first and dive into the Lovecraftian nightmare of algebra that is the uniform kernel later). Considering the functions involved with the squared exponential covariance (see the supplementary material from Oksana's paper), doing this analytically will probably not be an option. Some nice things *may* fall out of the uniform kernel (piecewise quadratics, etc.), but there will still be matrix inverses and exponentials (and possibly logarithms) that are challenging. Symbolic differentiation may be our friend here.

The other concern is assuming that there *is* a maximum for equation 2. KL divergence is convex with respect to P and Q, but viewing this as a composition of functions of $\boldsymbol{s}$ (and α and λ) it's not immediately obvious. Possible alternative approach: suppose you were able to use the integrated target function itself as the prior mean for the Gaussian process. Then the posterior would be $R = (F(t) - F(t_1), C_1(t,t))$, where F is the "target CDF" (apologies for overloaded nomenclature). What if we tried to *minimize* $D_{kl}(P||R)$ - i.e. chose our interrogation points and hyperparameters to bring our one-shot updater as close as possible to what it *should* be? Minimizing may be a more attainable goal than maximizing. However, the algebra is not any easier (and indeed may be harder, as F will certainly not be an elementary function).

1 Other options

The *Jensen-Shannon divergence*:

$$D_{js}(P||Q) = \frac{1}{2}\left(D_{kl}(P||Q) + D_{kl}(Q||P)\right) \tag{3}$$

$$= \frac{1}{2}\left(\int P(x)\log P(x)dx + \int Q(x)\log Q(x)dx\right) - \int M(x)\log M(x)dx \tag{4}$$

$$\tag{5}$$

where $M = \frac{1}{2}(P+Q)$ (a mixture of Gaussians, in our case). Using this depends on whether or not there's some nice expression for the log density of a random variable - I suspect there might be, but not sure.

Two other statistical distances are the Bhattacharyya and Hellinger distances. However, in this case of two Gaussians with equal variance, they both boil down to a bias/variance ratio, so are equivalent to KL divergence.

This paper (https://arxiv.org/pdf/1201.0418.pdf) proposes a family of distances that are bounded variants of the Bhattacharyya distance. However, they seem to largely end up as something of the form $\log\left(1+\exp\left(\frac{\text{Bias}^2}{\text{Variance}}\right)\right)$. I'm doubtful that's any easier mathematically, but I could be wrong.

There's also the family of Wasserstein metrics (`https://en.wikipedia.org/wiki/Wasserstein_metric`). The 2-Wasserstein metric for Gaussians of equal variance apparently reduces to a difference of means. This would compromise our ability to control the variance, but note that the bias *does* depend on a few covariance terms so there would be *some* effect. It would also be a lot easier mathematically. Not sure what the other Wasserstein metrics would be like (or if they even have easy forms for Gaussians).

There's also the Cramer-von Mises metric, which is the integral of the squared difference between the CDF's (and is equivalent to the energy metric in one dimension: `https://en.wikipedia.org/wiki/Energy_distance`). However, I don't imagine that will be very easy mathematically either.