Hamiltonian Monte Carlo using a nested Laplace approximation

Charles C. Margossian

June 18, 2019

This report goes over theoretical and computational details for the implementation of a nested Laplace approximation in Stan. The here described method is still at an experimental stage. It is adapted from a technical appendix I wrote for a class project, and updated. The purpose of sharing this document is to pool ideas and get feedback. Test and code can be found in the following gitHub branches: (i) for the `math` repo, see `feature/issue-755-laplace`, (ii) for the `stan` repo, see `feature/issue-2522-laplace`, and (iii) for the computer experiment `charlesm93/laplace_approximation` (private repo).

Appendix A Implementing the nested Laplace approximation

A.1 General mathematical implementation

Latent Gaussian models (LGMs) are a popular class of models and typically have the following hierarchical structure:

$$\begin{aligned} \phi &\sim \pi \\ \theta_i &\sim \text{Normal}(\mu(\phi), \Sigma(\phi)) \\ y_{j \in g(i)} &\sim p(\theta_i, \phi) \end{aligned} \tag{1}$$

Lemma 1. *In the latent Gaussian model setting, given by equation 1,*

$$p(\phi|y) \propto \frac{p(y|\theta, \phi) p(\theta|\phi) p(\phi)}{p(\theta|y, \phi)}$$

where the proportionality relationship is with respect to ϕ.

Proof.

$$\begin{aligned} p(\phi|y) &= \frac{p(\phi, y)}{p(y)} \\ &= \frac{p(\phi, y) p(\theta|\phi, y)}{p(y) p(\theta|\phi, y)} \\ &= \frac{p(\phi, \theta, y)}{p(\theta|y, \phi) p(y)} \\ &\propto \frac{p(y|\theta, \phi) p(\theta|\phi) p(\phi)}{p(\theta|y, \phi)} \end{aligned}$$

□

The Laplace approximation is

$$p(\theta|y,\phi) \approx p_{\mathcal{G}}$$

where $p_{\mathcal{G}}$ is found by matching the curvature and mode of $p(\theta|y,\phi)$. (Tierney & Kadane, 1986) work out the asymptotic properties of this approximation and show that, under certain regularity conditions, the error of the approximation is:

$$p(\theta|y,\phi) = p_{\mathcal{G}}(\theta)(1+\mathcal{O}(n^{-\frac{3}{2}}))$$

where n is the number of observations. Note the error is relative, and furthermore the rate of convergence is a factor n larger than what we get from the central limit theorem. The above-mentioned regularity conditions apply, among other cases, when y follows a normal, Poisson, binomial, or negative-binomial distribution.

The ingredients required to construct a nested Laplace approximation sampler are:

1. the mode, θ^*, of $p(\theta|y,\phi)$
2. the curvature of $p(\theta|y,\phi)$
3. the resulting approximate log joint distribution for the LGM

We derive these ingredients, first in a general setting, then in the specific case of a Poisson LGM.

The mode, θ^*, is found with a numerical solver and gives us the mean of the approximating Gaussian, $p_{\mathcal{G}}$. For the curvature, we use the following lemma:

Lemma 2. *The precision matrix of an approximating Gaussian that matches the curvature of $p(\theta^*|y,\phi)$ is*

$$\Sigma_{\mathcal{G}}^{-1} = -\log \nabla^2 p_{\theta|y,\phi}(\theta^*)$$

Proof. It suffices to show the precision matrix gives us the negative of the curvature for a Gaussian distribution.

$$\begin{aligned}\nabla \log p_{\mathcal{G}}(x;\mu) &= \nabla\left(\log\left(\frac{1}{(2\pi)^n \det|\Sigma|}\right) - \frac{1}{2}(x-\mu)^T\Sigma_{\mathcal{G}}^{-1}(x-\mu)\right)\\ &= -\Sigma_{\mathcal{G}}^{-1}(x-\mu)\end{aligned}$$

where the gradient is with respect to x. Thus

$$\nabla^2 \log p_G(x;\mu) = -\Sigma_{\mathcal{G}}^{-1}$$

The wanted result follows. □

Let $H = \log \nabla^2 p_{\theta|y,\phi}(\theta^*)$. Our Laplace approximation is then $p_{\mathcal{G}}(\theta) = \text{Normal}(\theta^*, H^{-1})$.

For ease of notation, let $\Sigma_\phi := \Sigma(\phi)$.

Lemma 3. *If we set $p(\theta|y,\phi) = p_{\mathcal{G}}(\theta)$ and $\theta = \theta^*$, then*

$$\begin{aligned}\log \hat{p}(\phi|y) &= \log p(\phi) + \log p(y|\theta^*, \phi) \\ &\quad -\frac{1}{2}\left(\log\det|\Sigma_\phi| + \log\det|H| + [\theta^* - \mu(\phi)]^T\Sigma_\phi^{-1}[\theta^* - \mu(\phi)]\right)\end{aligned}$$

where the "hat" in $\hat{p}(\phi|y)$ emphasizes that the posterior we are computing is an estimate of $p(\phi|y)$.

Proof. From equation 1, $\theta \sim \text{Normal}(\mu(\phi), \Sigma_\phi)$.
Then

$$p(\theta|\phi) = \left(\frac{1}{(2\pi)^n\det|\Sigma_\phi|}\right)^{\frac{1}{2}}\exp\left(-\frac{1}{2}[\theta^* - \mu(\phi)]^T\Sigma_\phi^{-1}[\theta^* - \mu(\phi)]\right)$$

Next

$$\begin{aligned}p_{\mathcal{G}}(\theta^*) &= \left(\frac{1}{(2\pi)^n\det|H^{-1}|}\right)^{\frac{1}{2}}\exp\left(-\frac{1}{2}(\theta^* - \theta^*)^T H(\theta^* - \theta^*)\right) \\ &= \left(\frac{1}{2\pi}\det|H|\right)^{\frac{1}{2}}\end{aligned}$$

where we used the fact that, for an invertible matrix A, $\det|A^{-1}| = (\det|A|)^{-1}$. Combining all our results, the approximate posterior becomes

$$\begin{aligned}\hat{p}(\phi|y) &= p(\phi)p(y|\theta^*,\phi)\frac{p(\theta^*|\phi)}{p_G(\theta^*)} \\ &= p(\phi)p(y|\theta^*,\phi)\left(\frac{1}{\det|\Sigma_\phi|\det|H|}\right)^{\frac{1}{2}}\exp\left(-\frac{1}{2}[\theta^* - \mu(\phi)]^T\Sigma_\phi^{-1}[\theta^* - \mu(\phi)]\right)\end{aligned}$$

or on the log scale:

$$\begin{aligned}\log \hat{p}(\phi|y) &= \log p(\phi) + \log p(y|\theta^*, \phi) \\ &\quad -\frac{1}{2}\left(\log\det|\Sigma_\phi| + \log\det|H| + [\theta^* - \mu(\phi)]^T\Sigma_\phi^{-1}[\theta^* - \mu(\phi)]\right)\end{aligned}$$

□

A.2 Mathematical implementation for a Poisson LGM

We now refine our results from the previous section to the case in which we have a Poisson LGM:

$$\begin{aligned}\phi &\sim \pi \\ \theta_i &\sim \text{Normal}(0, \Sigma_\phi) \\ y_{j\in g(i)} &\sim \text{Poisson}(e^{\theta_i})\end{aligned} \tag{2}$$

The following calculations allow us to construct a specialized implementation for a model with a log Poisson likelihood. However, we make no assumption on the structure of Σ_ϕ, other that it is a proper covariance matrix.

Remark. *We do not assume that we have the same number of observations for each group. We even allow for the scenario in which we have no observation for a particular group. Note however that we know from which group each observation comes.*

To do a Laplace approximation, we first need to find the mode of $p(\theta|y,\phi)$. Applying Bayes' rule:

$$p(\theta|y,\phi) \propto p(y|\theta,\phi)p(\theta|\phi)$$

By equation 2, the right hand side is a product of poisson and normal distributions. Since our goal is to find the mode, i.e. optimize the function for θ, we can ignore normalizing constants. Let

$$\begin{aligned} m_i &= \sum_{j\in g(i)} 1 \\ S_i &= \sum_{j\in g(i)} y_j \end{aligned}$$

respectively the total number of terms and the total number of counts in the i^{th} group. Let $\mathcal{O}$ denote the set of groups for which $m_i \geq 1$, i.e. the groups for which we have at least one observation. Then, on the log scale, an appropriate objective function is:

$$f(\theta) = \left\{\sum_{i\in\mathcal{O}} S_i\theta_i - m_i e^{\theta_i}\right\} - \tfrac{1}{2}\theta^T\Sigma^{-1}\theta \tag{3}$$

Remark. *If the k^{th} group contains no observation, θ_k only contributes to the quadratic term, which stems from $p(\theta|\phi)$, the prior on θ.*

For ease of notation and implementation in a computer program, we adopt the convention, here and throughout the article, that if the k^{th} group contains no observation, then $S_k = 0$. In such a case, S_k must be treated as a purely mathematical object since we do not claim to observe a total of 0 counts for an unobserved group. With this new convention, the objective function in equation 3 becomes:

$$f(\theta) = \left\{\sum_{i=1}^{M} S_i\theta_i - m_i e^{\theta_i}\right\} - \frac{1}{2}\theta^T\Sigma^{-1}\theta \tag{4}$$

Using the fact Σ^{-1} is symmetric, the gradient is then:

$$\nabla f(\theta) = \mathcal{V} - \Sigma^{-1}\theta \tag{5}$$

where $\mathcal{V}_i = S_i - m_i e^{\theta_i}$.

Noting the normalizing constant can be dropped in the log scale, the Hessian H is easily worked out from equation 5 to be

$$H(\theta) = \mathcal{W} - \Sigma^{-1}$$

where $\mathcal{W}$ is a diagonal matrix with $\mathcal{W}_i = -m_i e^{\theta_i}$. The log posterior is thus:

$$\log p(\phi|y) \approx \log p(\phi) + \log p(y|\theta^*,\phi) - \frac{1}{2}\left(\log\det|\Sigma_\phi| + \log\det|H| + \theta^{*T}\Sigma_\phi^{-1}\theta^*\right)$$

Recall that, in order to couple the nested Laplace approximation with HMC, we need to evaluate and differentiate this density several times per MCMC iteration. It is therefore crucial to perform these tasks efficiently. The most expensive terms are θ^*, which I will discuss in the next section, and Σ^{-1}. Indeed, inverting and differentiating a matrix is an operation we would rather avoid. The term $\theta^{*T}\Sigma_\phi^{-1}\theta^*$ can be obtained without explicitly constructing Σ_ϕ^{-1}, but using a `solve` method. We still need to deal with $\log\det|H| = \log\det|\mathcal{W} - \Sigma^{-1}|$.

Lemma 4. ***Generalized matrix determinant lemma.***
Suppose A is an invertible $n \times n$ matrix, and U, V are $n \times m$ matrices. Then $\det(A + UV^T) = \det(I_m + V^T A^{-1} U)\det(A)$.

The proof is omitted, but can be found in a linear algebra textbook. The following is a consequence of the above lemma:

Lemma 5.
$\log\det|\Sigma_\phi| + \log\det|H| = \log\det|I_n - \Sigma_\phi\mathcal{W}|$, *where I_n is the $n \times n$ identity matrix.*

Proof. Apply the generalized matrix determinant lemma, with $A = -\Sigma^{-1}$, $U = \mathcal{W}$ and $V^T = I_n$:

$$\begin{aligned} & \det|\mathcal{W} - \Sigma^{-1}| = \det|I_n - \Sigma\mathcal{W}|\det|-\Sigma^{-1}| \\ \iff & \log\det|\mathcal{W} - \Sigma^{-1}| = \log\det|I_n - \Sigma\mathcal{W}| - \log\det|\Sigma| \end{aligned}$$

where we used the fact $\det|\Sigma^{-1}| = (\det|\Sigma|)^{-1}$.

$$\therefore \qquad \log\det|\Sigma| + \log\det|H| = \log\det|I_n - \Sigma\mathcal{W}|. \qquad \square$$

Not only does this expression spare us inverting Σ, it also removes the requirement to compute $\log\det|\Sigma|$, which becomes quite expansive as the dimension of θ grows. In conclusion, the target posterior becomes:

$$\log p(\phi|y) \approx \log p(\phi) + \log p(y|\theta^*, \phi) - \frac{1}{2}\left(\log\det|I_n - \Sigma_\phi\mathcal{W}| + \theta^{*T}\Sigma_\phi^{-1}\theta^*\right) \quad (6)$$

Here the computational cost is dominated by the third expression in bracket.

A.3 Finding and differentiating θ^*, the mode of $p(\theta|y, \phi)$

We are dealing with an optimization problem, which for well-behaved systems means finding the root of a derivative. In our experiment, we work out that derivative explicitly (equation 5). Let $\mathcal{D}(\theta, \phi)$ be this derivative. Our task then reduces to solving a system of algebraic equations, that is find θ^* such that

$$\mathcal{D}(\theta^*, \phi) = 0$$

A.3.1 Optimization problem

There exists several approaches to numerically solve such systems. `Stan`'s built-in solver uses Powell's dogleg method (Powell, 1970). However, if the Laplace approximation is reasonable and $p(\theta|y,\phi)$ indeed behaves like a Gaussian, the optimization problem at hand should be (approximatively) convex. Newton's method then becomes a very attractive candidate. Computer experiments show that for the Poisson LGM problem, Newton's method is about an order of magnitude faster.

Newton's method is rather straightforward:

Algorithm 1 Newton's root-finding method

Set a tolerance ϵ and a maximum number of step (if the latter is exceeded, break).

1. Start with an initial guess θ_0.
2. Until $|D(\theta^{(i)},\phi)| \leq \epsilon$, update

$$\theta^{(i)} = \theta^{(i-1)} - [\nabla_\theta \mathcal{D}(\theta^{(i-1)},\phi)]^{-1} \mathcal{D}\left(\theta^{(i-1)},\phi\right)$$

3. Return θ^*.

In the Poisson LGM set up, we derived analytical expressions for both the $\mathcal{D}$ and its derivative, which for convenience we restate here:

$$\begin{aligned} D\left(\theta^{(i-1)},\phi\right) &= \mathcal{V} - \Sigma^{-1}\theta \\ \nabla_\theta \mathcal{D}(\theta^{(i-1)},\phi) &= \mathcal{W} - \Sigma^{-1} \end{aligned}$$

Note that Σ^{-1} does <u>not</u> depend on θ, and can therefore be precomputed, prior to running the solver. This is acceptable because here we do not differentiate through the inverse.

A.3.2 Computing sensitivities

We need to propagate derivatives through the solver in order to simulate Hamiltonian dynamics. Let J_θ^l be the Jacobian of the log density with respect to θ^*. Note that since the log density is a scalar, this Jacobian is a gradient. Applying the chain rule:

$$\begin{aligned} J_\phi^l &= (J_\phi^\theta)^T (J_\theta^l)^T \\ &= -(J_\phi^{\mathcal{D}})^T ([J_\theta^{\mathcal{D}}]^{-1})^T J_\theta^l \\ &= -(J_\phi^{\mathcal{D}})^T ([J_\theta^{\mathcal{D}}]^T)^{-1} J_\theta^l \end{aligned}$$

where we use the implicit function theorem to decompose $(J_\phi^\theta)^T$. The approach used by Stan's algebraic solver is to separately compute $J_\theta^{\mathcal{D}}$ and $J_\phi^{\mathcal{D}}$ using reverse mode automatic differentiation. There exists several ways in which this can be

improved[1]. In our problem, we have an analytical expression for $J_\theta^{\mathcal{D}}$, so our main concern is computing $J_\phi^{\mathcal{D}}$. Doing so with reverse mode automatic differentiation is extremely inefficient because

$$\phi \to \mathcal{D}(\phi)$$

maps from a low dimensional input to a high-dimensional output. Indeed $\dim(\mathcal{D}) = \dim(\theta)$. Forward-mode automatic differentiation is about ~ 100 times faster, when $\dim(\theta) = 500$, and its implementation is straightforward.

Alternatively, we can use one sweep of reverse mode automatic differentiation to compute the directional derivative $J_\phi^{\mathcal{D}} w$, where we use as an initial cotangent vector $w = ([J_\theta^{\mathcal{D}}]^T)^{-1} J_\theta^l$. This approach is more difficult to implement, "programing" wise (it requires doing a reverse automatic differentiation nested inside a `chain` call) and has yet to be tested. In theory, it should scale better than forward-mode automatic differentiation if the dimension of ϕ increases.

[1] See `https://discourse.mc-stan.org/t/better-computation-of-the-jacobian-matrix-for-algebraic-solver/8593/`.

References

Powell, M. J. D. (1970). A hybrid method for nonlinear equations. In P. Rabinowitz (Ed.), *Numerical methods for nonlinear algebraic equations.* Gordon and Breach.

Tierney, L., & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, *81*(393), 82-86. Retrieved from `https://amstat.tandfonline.com/doi/abs/10.1080/01621459.1986.10478240` doi: 10.1080/01621459.1986.10478240