Hamiltonian Monte Carlo using an adjoint-differentiated Laplace approximation

Bayesian inference for latent Gaussian models and beyond

Charles C. Margossian[1], Aki Vehtari[2], Daniel Simpson[3] and Raj Agrawal[4]

[1]Department of Statistics, Columbia University; [2]Department of Computer Science, Aalto University; [3]Department of Statistical Sciences, Toronto University; [4]CSAIL, Massachusetts Institute of Technology.

Bayesian inference for Latent Gaussian models

Latent Gaussian models (LGMs) are a key class of Bayesian hierarchical models, which observe the following structure:

$$\phi \sim \pi(\phi), \quad \theta \sim \text{Normal}(0, K(\phi)), \quad y \sim \pi(y \mid \theta, \phi),$$

where K is a covariance matrix parameterized by ϕ. We call ϕ the *hyperparameter* and θ the *latent Gaussian variable*. Our goal is to do full Bayesian inference on ϕ and θ.

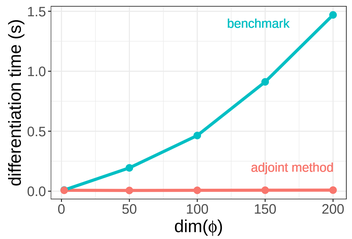
Applications. LGMs encompass a broad range of models with distinct behaviors. Gaussian processes for instance typically have a large dimensional θ and a low dimensional ϕ (e.g. [4, 3]). In other models, ϕ is high dimensional too and $\pi(\phi \mid y)$ is multimodal. This is notably the case for general linear models (GLM) with a sparsity inducing horseshoe prior (e.g. [2]) and sparse kernel interaction models (SKIM) [1]. These models are particularly useful in regimes where we have a large number of covariates, for example in genomics.

Bayesian inference. Performing Bayesian inference on LGMs can be challenging because of the posterior's geometry. The interaction between θ and ϕ often generates highly varying curvatures, which frustrate Markov chains Monte Carlo (MCMC) algorithms. Instead of running MCMC on the joint, $\pi(\theta, \phi \mid y)$, we can use the geometrically better behaved marginal distribution,

$$\pi(\phi \mid y) \propto \pi(\phi)\pi(y \mid \phi) = \pi(\phi) \int_\Theta \pi(y, \theta \mid \phi) \mathrm{d}\theta,$$

to sample ϕ and then draw exact samples from $\pi(\theta \mid \phi, y)$. In most cases, no analytical expressions exist for $\pi(y \mid \phi)$ and $\pi(\theta \mid \phi, y)$, so we resort to a *Laplace approximation*: a normal distribution which matches the mode and curvature of $\pi(\theta \mid \phi, y)$. For several modern problems, $\pi(\phi \mid y)$ is high-dimensional and multimodal. To efficiently sample the hyperparameters, we deploy dynamic Hamiltonian Monte Carlo (HMC), a gradient-based MCMC sampler, which requires the derivative $\nabla_\phi \log \pi(y \mid \phi)$. Our main contribution is a novel, scalable method to differentiate the log marginal likelihood.

Main Contribution: Adjoint method for scalable differentiation



We build on the algorithmic work by Rasmussen & Williams (2006) [3], who use a custom Newton solver to approximate and differentiate $\log \pi(y \mid \phi)$. This differentiation method requires users to pass $K' = \partial K / \partial \phi$. When analytical derivatives are not available, we can resort to automatic differentiation to construct K'. In general, this method scales very poorly when the dimension of ϕ increases.

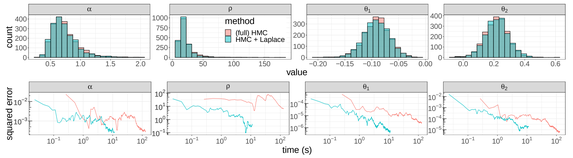
Adjoint method. We derive a novel differentiation algorithm which bypasses altogether the computation of K', by directly evaluating the cotangent-Jacobian product $w^T K'$ for the appropriate cotangent, w^T. The superior scalability of this approach is demonstrated on the SKIM (see left).

Implementation. We build the adjoint method in C++ for an expandable suite of observational models. The code is interfaced with the probabilistic framework Stan. Hence it straightforward to couple dynamic HMC with the adjoint-differentiated Laplace approximation. The code used in the article can be found at https://github.com/charlesm93/laplace_manuscript.

Numerical Experiment

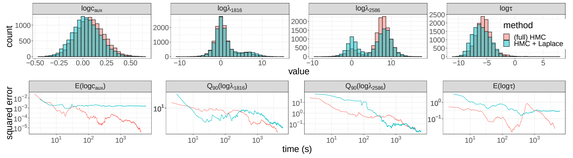
We test our marginalization scheme on classic and cutting-edge problems. Our benchmarks include (i) full HMC run over $\pi(\theta, \phi \mid y)$ and (ii) automatic differentiation variational inference (ADVI). ADVI is strongly biased, while full HMC requires extensive tuning. For clarity most of the analysis on ADVI is relegated to the Supplementary Material.

Gaussian process with a Poisson observational model.
$\dim(\phi) = 2, \quad \dim(\theta) = 100.$



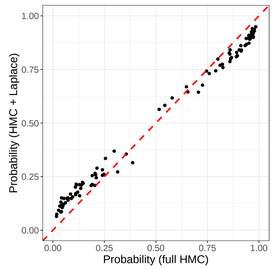
Both samplers are in close agreement. The approximation is about an order of magnitude faster. What is more, two attempts at fitting the model were required in order to tune full HMC.

GLM with a regularized horseshoe prior and Bernoulli observations.
$\dim(\phi) = 5,966, \quad \dim(\theta) = 102.$



We study the probability of developing prostate cancer based on genetic data. The model identifies explanatory covariates, which produce a heavy-tailed and at times multimodal posterior of the hyperparameter λ. The 90^{th} quantile of $\log \lambda$ serves as a soft selection criterion.

Extensive tuning is required in order to run full HMC: it took us more than 4 attempts, each taking several hours to run! By contrast, the embedded Laplace runs smoothly, indicating the geometry of $\pi(\phi \mid y)$ is much better behaved than the geometry of $\pi(\phi, \theta \mid y)$. We however note that the approximation introduces a bias, which is expected when the observational model is Bernoulli. Still, the accuracy is comparable for several quantities of interest.



(full) HMC	2586	1816	4960	4238	4843	3381
HMC + Laplace	2586	1816	4960	4647	4238	3381
ADVI	1816	2416	4284	2586	5279	4940

Top six covariate indices, i, with the highest 90^{th} quantiles of $\log \lambda_i$

Sparse Kernel Interaction Model. The SKIM expands the GLM with a horseshoe prior by accounting for interaction effects. To achieve scalability, the GLM is recast as a Gaussian process using a kernel trick. The results of our experiment on this model are comparable to what we obtained with the GLM and detailed in our article.

References

[1] R. Agrawal, J. H. Huggins, B. Trippe, and T. Broderick. The Kernel interaction trick: Fast Bayesian discovery of pairwise interactions in high dimensions. *Proceedings of the 36th International Conference on Machine Learning*, 97, April 2019.

[2] J. Piironen and A. Vehtari. Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11:5018–5051, 2017.

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

[4] J. Vanhatalo, V. Pietiläinen, and A. Vehtari. Approximate inference for disease mapping with sparse Gaussian processes. *Statistics in Medicine*, 29(15):1580–1607, 2010.

Neural Information Processing Systems, December 2020