Discussion on the paper: Riemann manifold Langevin and Hamiltonian Monte Carlo methods, by Girolami and Calderhead

Iain Murray

School of Informatics, University of Edinburgh, UK

Ryan Prescott Adams

Department of Computer Science, University of Toronto, Canada

This is an exciting piece of work. We agree with the authors that Hamiltonian Monte Carlo (HMC) could be used more broadly. HMC has a reputation as being difficult to tune and we hope that the more robust behaviour demonstrated here will help alleviate this.

Dynamical methods jointly update variables, which may allow larger moves in complex models than updating variables independently. Hierarchical models often contain strong prior dependencies between hyperparameters and parameters and would seem to provide a common case where HMC offers benefits. Unfortunately, updating hyperparameters and parameters jointly using standard HMC does not necessarily work better than updating them individually (Choo, 2000).

Consider a simple hierarchical model (Neal, 2003) where the observations are uninformative about the parameters:

$$\begin{aligned} v &\sim \mathcal{N}(0, \sigma^2), && \text{e.g., } \sigma = 3, \\ x_k &\sim \mathcal{N}(0, e^v), && k = 1 \ldots K = 10 \\ \mathbf{y} &\sim p(\mathbf{y}), && \text{observations are independent of } \boldsymbol{\theta} = (v, \mathbf{x}). \end{aligned}$$

The posterior of this trivial model is equal to the prior, $p(\boldsymbol{\theta})$.

Sampling from $p(\boldsymbol{\theta})$ using simple Markov chain methods can be difficult: 1) sensible step-sizes for $\mathbf{x}$ depend on v, and 2) the hyperparameter v cannot move very much for fixed effects $\mathbf{x}$ (especially for large K). One way to deal with these issues is reparameterisation: rewriting the model in terms of $z_k \sim \mathcal{N}(0, 1)$ with $x_k = z_k e^{v/2}$ gives independent variables $(v, \mathbf{z})$. Most Markov chain methods will now sample effectively from the prior. When the distribution is reweighted by non-trivial likelihood terms, other reparameterisations may be appropriate (Christensen et al., 2006; Murray and Adams, 2010).

Perhaps using a metric can replace the need for careful reparameterisation, although the metric used in this papers' examples would not work: the negative Hessian $-\partial^2 \log p(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}^2$ is not positive definite everywhere. We could use

$$G(\boldsymbol{\theta}) = -E_{\mathbf{y}|\boldsymbol{\theta}}\left[\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \log p(\mathbf{y}|\boldsymbol{\theta})\right] - E_{\mathbf{x}|v}\left[\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \log p(\mathbf{x}|v)\right] - \frac{\partial^2}{\partial \boldsymbol{\theta}^2} \log p(v), \tag{1}$$

which is positive definite. This metric would lead to sensible step-sizes for the random effects x_k: the diffusion terms $\mathbf{z}$ are scaled by the standard deviation $e^{v/2}$. However, the metric is diagonal, which suggests that problems with the dependencies in the prior are not alleviated, at least in the weak data limit.

As highlighted in section 4.2, there is a broader choice of metrics to be considered. We hope that algorithms presented here encourage renewed interest in developing new variable-metric Hamiltonian methods for hierarchical models.

References

Choo, K. (2000). Learning hyperparameters for neural network models using Hamiltonian dynamics. Master's thesis, Department of Computer Science, University of Toronto. Available from `http://www.cs.toronto.edu/~radford/ftp/kiam-thesis.ps`.

Christensen, O. F., G. O. Roberts, and M. Skãld (2006). Robust Markov chain Monte Carlo methods for spatial generalized linear mixed models. *Journal of Computational and Graphical Statistics 15*(1), 1–17.

Murray, I. and R. P. Adams (2010). Slice sampling covariance hyperparameters of latent Gaussian models. In *Advances in Neural Information Processing Systems 23*. To appear, arXiv:1006.0868.

Neal, R. M. (2003). Slice sampling. *Annals of Statistics 31*(3), 705–767.