Kinetic energy choice in Hamiltonian Monte Carlo

Samuel Livingstone, University of Bristol

Hamiltonian Monte Carlo is a Markov chain Monte Carlo method which is now both widely used in Bayesian inference, and increasingly studied and developed. The idea of the approach is to use the deterministic measure-preserving dynamics of Hamiltonian flow to promote fast exploration of a parameter space of interest. To achieve this the space is augmented with a *momentum* variable, and the Markov chain evolves by switching between re-sampling this momentum and solving Hamilton's equations for a prescribed amount of time. Typically the equations cannot be solved exactly and so a time-reversible and volume-preserving numerical integrator is used, with discretisation errors controlled for using a Metropolis step. The method is attractive in the big data setting, in two different ways. First, often large datasets are highly structured and a complex model with many parameters is required to accurately capture dependencies (e.g. [3]). In such cases the $O(p^{1/4})$ mixing time scales with dimension more favourably than random walk-based Markov chains which are $O(p)$ [1]. Second, in the tall data setting, the availability of stochastic gradients can facilitate a reduction in computational cost [2].

Here we consider what choice of momentum distribution $\nu(\cdot)$ is appropriate to draw samples from a distribution of interest $\pi(\cdot)$. We will assume that $\nu(\cdot)$ possesses a density which is proportional to $\exp\{-K(p)\}$ for some $K(p)$, which we call the *kinetic energy* (and that $\pi(\cdot)$ has a density proportional to $\exp\{-U(x)\}$). The standard choice is $K(p) = p^T p/2$, with the resulting $\nu(\cdot)$ a Gaussian. The general requirements are simply that $K(p)$ is differentiable, $\mathbb{E}(p) = 0$ and that $\nu(\cdot)$ can be sampled from. Here we consider how different choices affect the algorithm as a whole, in terms of stability and convergence, and develop guidelines for practitioners.

Our key findings are that the robustness and efficiency of the method can to some degree be controlled through the quantity $\nabla K \circ \nabla U(x)$, which we term *the composite gradient*. In particular, we propose that balancing the tails of the kinetic energy with those of the potential to make this approximately linear in x should give good performance, and that no faster than linear growth is necessary for algorithm stability. When $\pi(\cdot)$ is very light-tailed, this can be done by choosing $\nu(\cdot)$ to be heavier-tailed. There are, however, serious disadvantages to choosing any heavier tails than those of a Laplace distribution, which can result in the sampler moving very slowly in certain regions of the space. We introduce a *negligible moves* property for Markov chains to properly characterize this behaviour. We also find that in practice considering the distribution of $\nabla K(p)$, which we term *the implicit noise*, is important, since this governs the behaviour of the sampler in regions where $\|\nabla U(x)\|$ is small. We suggest various choices for controlling these two quantities, and support our findings with a numerical study. We also comment on how changing the kinetic energy can affect behaviour when $\nabla U(x)$ is approximated, either through subsampling or within an exact-approximate Monte Carlo scheme.

References

[1] A. Beskos, N. Pillai, G. Roberts, J.-M. Sanz-Serna, A. Stuart, et al. Optimal tuning of the hybrid monte carlo algorithm. *Bernoulli*, 19(5A):1501–1534, 2013.

[2] T. Chen, E. B. Fox, and C. Guestrin. Stochastic gradient hamiltonian monte carlo. In *ICML*, pages 1683–1691, 2014.

[3] W. Wang, D. Rothschild, S. Goel, and A. Gelman. Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3):980–991, 2015.