Relativistic Hamiltonian Monte Carlo

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

1 Introduction

2 Relativistic Hamiltonian Monte Carlo

Consider a target density $f(\theta)$ that can be written as $f(\theta) \propto e^{-U(\theta)}$. Hamiltonian Monte Carlo operates by introducing auxiliary variables p so that $f(\theta, p) \propto e^{-H(\theta,p)}$, where

$$H(\theta, p) = U(\theta) + \frac{1}{2m}p^T p \tag{1}$$

so that p is marginally normally distributed. This Hamiltonian lends itself to the interpretation of a particle with position θ and momentum p moving in a system with potential energy $U(\theta)$ and according to the classical kinetic energy $\frac{1}{2m}p^T p$. We can derive simple update equations for simulating from these dynamics using Hamilton's equations:

$$\dot{\theta} = \frac{\partial H}{\partial p}\dot{p} = -\frac{\partial H}{\partial \theta} \tag{2}$$

giving one possible set of updates (the leapfrog integrator):

$$p_{t+1/2} \leftarrow p_t - \frac{1}{2}\epsilon \nabla U(\theta_t) \tag{3}$$

$$\theta_{t+1} \leftarrow \theta_t + \epsilon \frac{p_{t+1/2}}{m} \tag{4}$$

$$p_{t+1} \leftarrow p_{t+1/2} - \frac{1}{2}\epsilon \nabla U(\theta_{t+1}) \tag{5}$$

which is then followed by a Metropolis Hastings accept/reject step. This choice of update is chosen so that the Hamiltonian H is left approximately invariant, so that as the acceptance probability approaches 1. One consequence of these updates is that, when applying HMC to problems where is very peaked, the momentum p can become very large, resulting in large updates for θ, and thus a very fine discretization is needed. Consider if, instead of the classical kinetic energy were used for the Hamiltonian, the relativistic kinetic energy were used instead:

$$K(p) = mc^2 \left(\frac{p^T p}{m^2 c^2} + 1 \right)^{\frac{1}{2}} \tag{6}$$

where c is the "speed of light" which bounds the speed of any particle. This gives the Hamiltonian:

$$H(\theta, p) = U(\theta) + mc^2 \left(\frac{p^T p}{m^2 c^2} + 1 \right)^{\frac{1}{2}} \tag{7}$$

Submitted to 29th Conference on Neural Information Processing Systems (NIPS 2016). Do not distribute.

The update equations then become

$$p_{t+1/2} \qquad \leftarrow p_t - \tfrac{1}{2}\epsilon \nabla U(\theta_t) \tag{8}$$

$$\theta_{t+1} \quad \leftarrow \theta_t + \epsilon \tfrac{p_{t+1/2}}{m}\left(\tfrac{p_{t+1/2}^T p_{t+1/2}}{m^2c^2} + 1\right)^{-\frac{1}{2}} \tag{9}$$

$$p_{t+1} \qquad \leftarrow p_{t+1/2} - \tfrac{1}{2}\epsilon \nabla U(\theta_{t+1}) \tag{10}$$

Here the momentum is still unbounded and may become very large in the presence of large gradients in the potential energy. However, the size of the θ update is bounded by c, thus the behavior of the proposed samples can be more easily controlled in the presence of large gradients. The marginal distribution of p is no longer normal, but its density is log-concave and can be sampled using Adaptive Rejection Sampling.

3 Stochastic Gradient Relativistic Hamiltonian Monte Carlo

Hamiltonian Monte Carlo algorithms are also of particular interest for "stochastic gradient" style algorithms where mini-batches are used to form noisy estimates of the gradients. One motivation for this is that the momentum serves as a reservoir of previous gradient information; a large gradient will result in a large p, which may stay large for a while unless met with another large gradient, thus retaining the memory of a strong signal on prior batches of data. However, due to potentially large variability in the gradient computed in these algorithms, stochastic gradient Hamiltonian algorithms may still result in overly large updates, requiring very small values of and thus potentially slow convergence. This motivates the use of the Relativistic Hamiltonian in a stochastic gradient sampler; the inherent bound in the update size allows the sampler to more easily smooth out the noise in the gradient over multiple steps.

Ma et al. [2015] gives a framework for taking update equations associated with a particular Hamiltonian and constructing asymptotically consistent stochastic gradient samplers. Specifically, Ma et al. [2015] consider a SDE with drift $f(z)$ and diffusion $2D(z)$:

$$dz = f(z)dt + \sqrt{2D(z)}dW_t \tag{11}$$

where $z = (\theta, p)$, W_t is a d-dimensional Wiener process, and

$$f(z) = -\left[D(z) + Q(z)\right]\nabla H(z) + \Gamma(z), \Gamma_i = \sum_{j=1}^{d} \frac{\partial}{\partial z_j}\left(D_{ij}(z) + Q_{ij}(z)\right) \tag{12}$$

where $Q(z)$ is skew-symmetric. Then the update equations

$$z_{t+1} \leftarrow z_t - \epsilon_t \left[\left[D(z_t) + Q(z_t)\right]\nabla \tilde{H}(z_t) + \Gamma(z_t)\right] + \mathcal{N}(0, \epsilon_t(2D(z_t) - \epsilon_t \hat{B}_t)) \tag{13}$$

gives an asymptotically consistent chain when the stepsizes t decrease to zero at the appropriate rate. Here $\tilde{H}(z)$ is the estimate of the Hamiltonian, e.g. using mini-batches, and $\hat{B}$ is an estimate of the variance of the noise of the approximate gradient computation. Note that this estimate need not be unbiased for the chain to be consistent – failing better choices we may choose $\hat{B}_t = 0$. In practice, decreasing the stepsizes t results in progressively slower mixing, and it is often preferable to fix a stepsize and accept that the sampler will incur some asymptotic bias. We can formulate Relativistic Hamiltonian Monte Carlo into this framework by taking

$$H(\theta, p) \quad = U(\theta) + mc^2\left(\tfrac{p^T p}{m^2c^2} + 1\right)^{\frac{1}{2}} \tag{14}$$

$$D(z) \qquad = \begin{bmatrix} 0 & 0 \\ 0 & D \end{bmatrix} \tag{15}$$

$$Q(z) \qquad = \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix} \tag{16}$$

which gives $\Gamma_i(z) = 0$ and

$$f\left(\begin{bmatrix} \theta \\ p \end{bmatrix}\right) = -\begin{bmatrix} 0 & -I \\ I & D \end{bmatrix}\begin{bmatrix} \nabla U(\theta) \\ \frac{p}{m}\left(\frac{p^T p}{m^2c^2} + 1\right)^{-\frac{1}{2}} \end{bmatrix} \tag{17}$$

which gives the SDE

$$d\theta = \frac{p}{m}\left(\frac{p^T p}{m^2c^2}+1\right)^{-\frac{1}{2}} \tag{18}$$

$$dp = \left(-\nabla U(\theta) - D\frac{p}{m}\left(\frac{p^T p}{m^2c^2}+1\right)^{-\frac{1}{2}}\right)dt + \sqrt{2D}dW_t \tag{19}$$

Then (13) gives the updates:

$$p_{t+1} \leftarrow p_t - \epsilon_t \nabla\tilde{U}(\theta_t) - \epsilon_t D\frac{p_t}{m}\left(\frac{p_t^T p_t}{m^2c^2}+1\right)^{-\frac{1}{2}} + \mathcal{N}(0, \epsilon_t(2D - \epsilon_t \hat{B}_t)) \tag{20}$$

$$\theta_{t+1} \leftarrow \theta_t + \epsilon_t \frac{p_{t+1}}{m}\left(\frac{p_{t+1}^T p_{t+1}}{m^2c^2}+1\right)^{-\frac{1}{2}} \tag{21}$$

4 A Stochastic Gradient Nosé-Hoover Thermostat for Relativistic Hamiltonian Monte Carlo

The stochastic gradient version of HMC (SGHMC) introduced by Chen et al. [2014] can be improved by introducing an additional dynamic variable ξ to adaptively increase or decrease the momenta (Ding et al. [2014],Leimkuhler and Shang [2016]). The extended systems has a Hamiltonian of the form

$$H(\theta, p, \xi) = U(\theta) + \frac{1}{2}p^T p + \frac{d}{2}(\xi - D)^2 \tag{22}$$

The dynamics of this approach, known as stochastic gradient Nosé-Hoover thermostat due to its links to statistical physics, can be expressed as:

$$d\theta = p dt \tag{23}$$

$$dp = -\nabla\tilde{U}dt - \xi p dt + \sqrt{2D}dW_t \tag{24}$$

$$d\xi = \frac{1}{d}\left(p^T p - 1\right)dt \tag{25}$$

Intuitively this approach works because

$$\mathbb{E}\left[\frac{d\xi}{dt}\right] = 0, \text{ when sampling from the target joint distribution} \tag{26}$$

The system adaptively 'heats' or 'cools' to push the system closer to obeying (26). Hence the additional dynamics will move the distribution closer to the equilibrium. In particular this helps to reduce the bias of SGHMC. A natural question is whether these methods can be extended to relativistic HMC. Leimkuhler and Shang [2016] show that for a general kinetic energy $K(p)$, provided that ξ is normally distributed in equilibrium (i.e. using the Hamiltonian in (22)) the ξ dynamics become

$$d\xi = \frac{1}{d}\left(\|\nabla K(p)\|^2 - \nabla^2 K(p)\right)dt \tag{27}$$

Note that these general dynamics can still be interpreted as maintaining an equation like (26) since

$$\mathbb{E}\left[\frac{\partial^2 K}{\partial p_i^2}\right] = \int \frac{\partial^2 K}{\partial p_i^2} e^{-K(p)} dp \tag{28}$$

$$= \underbrace{\int \left[\frac{\partial K}{\partial p_i} e^{-K(p)}\right]_{p_i=\infty}^{\infty} dp_{-i}}_{=0} - \int \frac{\partial K}{\partial p_i}\left(-\frac{\partial K}{\partial p_i} e^{-K(p)}\right) dp = \mathbb{E}\left[\left(\frac{\partial K}{\partial p_i}\right)^2\right] \tag{29}$$

and hence $\mathbb{E}\left[\frac{d\xi}{dt}\right] = 0$. We can fit these ideas into the framework of Ma et al. [2015] by defining:

$$H(\theta, p, \xi) = U(\theta) + K(p) + \frac{d}{2}(\xi - D)^2 \tag{30}$$

$$D(\theta, p, \xi) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & D \cdot I & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{31}$$

$$Q(\theta, p, \xi) = \begin{pmatrix} 0 & -I & 0 \\ I & 0 & \nabla K(p)/d \\ 0 & -\nabla K(p)^T/d & 0 \end{pmatrix} \tag{32}$$

This gives

$$\Gamma = \begin{pmatrix} 0 \\ 0 \\ -\nabla^2 K(p)/d \end{pmatrix} \tag{33}$$

and the dynamics become

$$d\theta = \nabla K(p) dt \tag{34}$$

$$dp = -\nabla \tilde{U} dt - \xi \nabla K(p) dt + \sqrt{2D} dW_t \tag{35}$$

$$d\xi = \frac{1}{d} \left(\|\nabla K(p)\|^2 - \nabla^2 K(p) \right) dt \tag{36}$$

This gives a general recipe for a stochastic gradient Nosé-Hoover thermostat with a general kinetic energy $K(p)$. For the relativistic kinetic energy we find

Add relativistic dynamics and updates from Xiaoyu's notes

5 Experiments

6 Conclusion

References

Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic Gradient Hamiltonian Monte Carlo. In *Proceedings of The 31st International Conference on Machine Learning*, pages 1683–1691, 2014. URL `http://jmlr.org/proceedings/papers/v32/cheni14`.

Nan Ding, Youhan Fang, Ryan Babbush, Changyou Chen, Robert D. Skeel, and Hartmut Neven. Bayesian Sampling Using Stochastic Gradient Thermostats. In *Advances in Neural Information Processing Systems*, pages 3203–3211, 2014. URL `http://papers.nips.cc/paper/5592-bayesian-sampling-using-stochastic-gradient-thermostats`.

Benedict Leimkuhler and Xiaocheng Shang. Adaptive Thermostats for Noisy Gradient Systems. *SIAM Journal on Scientific Computing*, 38(2):A712–A736, mar 2016. ISSN 1064-8275. doi: 10.1137/15M102318X. URL `http://epubs.siam.org/doi/10.1137/15M102318X`.

Yi-An Ma, Tianqi Chen, and Emily B. Fox. A Complete Recipe for Stochastic Gradient MCMC. jun 2015. URL `http://arxiv.org/abs/1506.04696`.