1 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:

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

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

$$\begin{aligned} p_{t+1/2} &\leftarrow p_t - \frac{1}{2}\epsilon \nabla U(\theta_t) && (3) \\ \theta_{t+1} &\leftarrow \theta_t + \epsilon \frac{p_{t+1/2}}{m} && (4) \\ p_{t+1} &\leftarrow p_{t+1/2} - \frac{1}{2}\epsilon \nabla U(\theta_{t+1}) && (5) \\ & && (6) \end{aligned}$$

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 $\epsilon \to 0$, the acceptance probability approaches 1.

One consequence of these updates is that, when applying HMC to problems where $U(\theta)$ is very peaked, the momentum p can become very large, resulting in large updates for θ, and thus a very fine discretization ϵ is needed.

Consider if, instead of the classical kinetic energy were used for the Hamiltonian, the relativistic kinetic energy were used instead:

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

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

$$H = U(\theta) + mc^2 \left(\frac{p^T p}{m^2 c^2} + 1 \right)^{\frac{1}{2}} \tag{8}$$

The update equations then become

$$\begin{aligned} p_{t+1/2} &\leftarrow p_t - \frac{1}{2}\epsilon \nabla U(\theta_t) && (9) \\ \theta_{t+1} &\leftarrow \theta_t + \epsilon \frac{p_{t+1/2}}{m} \left(\frac{p_{t+1/2}^T p_{t+1/2}}{m^2 c^2} + 1 \right)^{-\frac{1}{2}} && (10) \\ p_{t+1} &\leftarrow p_{t+1/2} - \frac{1}{2}\epsilon \nabla U(\theta_{t+1}) && (11) \end{aligned}$$

Here the momentum is still unbounded and may become very large in the presence of large gradients in the potential energy. However, the update that is incurred onto θ is bounded by ϵc, 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, however its density is log-concave and can be sampled using Adaptive Rejection Sampling.

2 Stochastic Gradient Relativistic Hamiltonian Monte Carlo

Hamiltonian Monte Carlo algorithms are also of particular interest for "stochastic gradient" style algorithms where mini-batches of data are queried for making updates. 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 ϵ 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.

[1] gives a framework for taking update equations associated with a particular Hamiltonian and constructing asymptotically consistent stochastic gradient samplers. Specifically, [1] considers a SDE with drift $f(z)$ and diffusion $\sqrt{2D(z)}$:

$$dz = f(z)dt + \sqrt{2D(z)}dW(t) \tag{12}$$

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

$$f(z) = -[D(z) + Q(z)]\nabla H(z) + \Gamma(z), \quad \Gamma_i(z) = \sum_{j=1}^{d} \frac{\partial}{\partial z_j}\left(D_{ij}(z) + Q_{ij}(z)\right) \tag{13}$$

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

$$z_{t+1} \leftarrow \epsilon_t \left[(D(z_t) + Q(z_t))\nabla \tilde{H}(z_t) + \Gamma(z_t)\right] + \mathcal{N}(0, \epsilon_t(2D(z_t) - \epsilon_t \hat{B}_t)) \tag{14}$$

gives an asymptotically consistent chain when the stepsizes ϵ_t decrease to zero at the appropriate rate. Here $\tilde{H}(z_t)$ is the estimate of the Hamiltonian, e.g. using mini-batches, and $\hat{B}_t$ 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 ϵ_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 fit Relativistic Hamiltonian Monte Carlo into this framework by taking

$$H(\theta, p) = U(\theta) + mc^2 \left[\frac{p^2}{m^2c^2} + 1 \right]^{\frac{1}{2}} \tag{15}$$

$$D(z) = \begin{bmatrix} 0 & 0 \\ 0 & D \end{bmatrix} \tag{16}$$

$$Q(z) = \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix} \tag{17}$$

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

$$f\left(\begin{bmatrix} \theta \\ p \end{bmatrix} \right) = - \begin{bmatrix} 0 & -I \\ I & D \end{bmatrix} \begin{bmatrix} \nabla U(\theta) \\ \frac{p}{m} \left(\frac{p^T p}{m^2c^2} + 1 \right)^{-\frac{1}{2}} \end{bmatrix} \tag{18}$$

which gives the SDE:

$$d\theta = \frac{p}{m} \left(\frac{p^T p}{m^2c^2} + 1 \right)^{-\frac{1}{2}} dt \tag{19}$$

$$dp = \left(-\nabla U(\theta) - D\frac{p}{m} \left(\frac{p^T p}{m^2c^2} + 1 \right)^{-\frac{1}{2}} \right) dt + \sqrt{2D} dW(t) \tag{20}$$

Using (14) gives the updates:

$$p_{t+1} \leftarrow p_t - \epsilon_t \nabla \tilde{U}(\theta_t) - \epsilon_t D \frac{p_t}{m} \left(\frac{p_t^T p_t}{m^2c^2} + 1 \right)^{-\frac{1}{2}} + \mathcal{N}(0, \epsilon_t(2D - \epsilon_t \hat{B}_t)) \tag{21}$$

$$\theta_{t+1} \leftarrow \theta_t + \epsilon_t \frac{p_{t+1}}{m} \left(\frac{p_{t+1}^T p_{t+1}}{m^2c^2} + 1 \right)^{-\frac{1}{2}} \tag{22}$$

References

[1] YA Ma, T Chen, and E Fox. A complete recipe for stochastic gradient MCMC. *Advances in Neural Information Processing ...*, 2015.