Hamiltonian Monte Carlo

Andrew J. Holbrook

UCLA Biostatistics

Bayesian inference

We observe data $y_1, \dots, y_N \stackrel{iid}{\sim} p(y_n|\boldsymbol{\theta})$ and assume $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$. Here,

- $p(\mathrm{y}|\boldsymbol{\theta}) = \prod_{n=1}^{N} p(y_n|\boldsymbol{\theta})$ is the *likelihood*,
- $p(\boldsymbol{\theta})$ is the *prior*,

and the goal of Bayesian inference is to obtain the *posterior*

$$p(\boldsymbol{\theta}|\mathrm{y}) = \frac{p(\mathrm{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathrm{y})} = \frac{p(\mathrm{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_{\Theta} p(\mathrm{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}\,.$$

Bayesian inference

We're usually interested in computing another integral

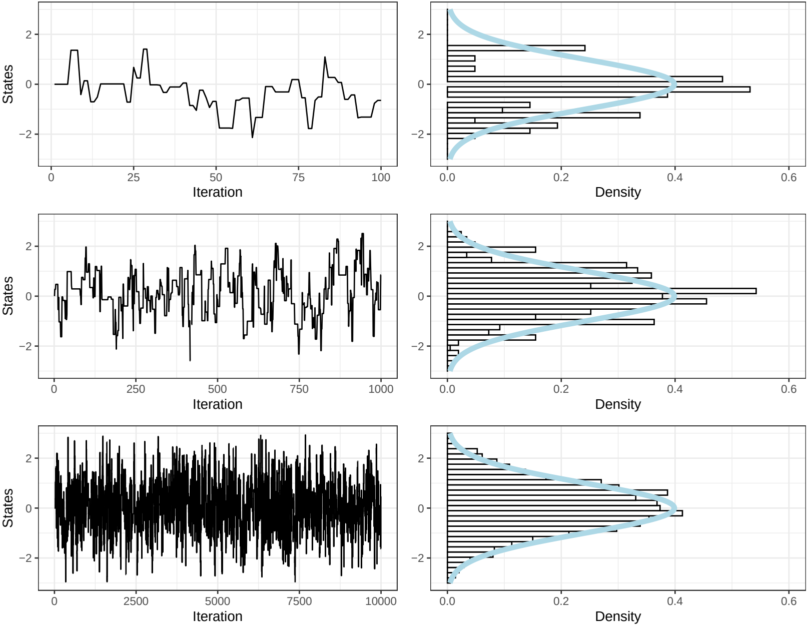
$$\mathbb{E}_{\boldsymbol{\theta}|\mathrm{y}} f(\boldsymbol{\theta}) = \int_{\Theta} f(\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathrm{y}) \mathrm{d}\boldsymbol{\theta} \,,$$

so we do what statisticians have been doing forever. We collect samples and rely on the law of large numbers. Suppose $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_S \overset{iid}{\sim} p(\boldsymbol{\theta}|\mathrm{y})$ ($\mathbb{E}_{\boldsymbol{\theta}|\mathrm{y}}|\boldsymbol{\theta}| < \infty$) and $f(\cdot)$ *a.s.* continuous, then

- (WLLN) $\sum_{s=1}^{S} f(\boldsymbol{\theta}_s)/S \xrightarrow{P} \mathbb{E}_{\boldsymbol{\theta}|\mathrm{y}} f(\boldsymbol{\theta})$
- (SLLN) $\sum_{s=1}^{S} f(\boldsymbol{\theta}_s)/S \xrightarrow{a.s.} \mathbb{E}_{\boldsymbol{\theta}|\mathrm{y}} f(\boldsymbol{\theta})$

But where do we find our samples?

Markov chain Monte Carlo



Rejection sampling

We want to sample from generic $p(\boldsymbol{\theta})$ but only know $p^*(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta})$. We can easily sample from $q(\boldsymbol{\theta})$ and know a number $M > 0$ s.t. $p^*(\boldsymbol{\theta}) < Mq(\boldsymbol{\theta})$.

Algorithm for generating $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$:

1. Draw $\boldsymbol{\theta}^* \sim q(\boldsymbol{\theta})$ and $U \sim U(0, 1)$
2. $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}^*$ if $U < \frac{p*(\boldsymbol{\theta}^*)}{Mq(\boldsymbol{\theta}^*)}$

Funky rejection sampler

We want to sample from generic $p(\boldsymbol{\theta})$ and *we can*. But we also feel the compulsion to use the invertible, differentiable function $T(\cdot)$.

Algorithm for generating $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$:

1. Draw $\boldsymbol{\theta}^{\dagger} \sim p(\boldsymbol{\theta})$ and $U \sim U(0,1)$
2. Obtain $\boldsymbol{\theta}^{*} = T(\boldsymbol{\theta}^{\dagger})$
3. Recognize that $q(\boldsymbol{\theta}^{*}) = p(T^{-1}(\boldsymbol{\theta}^{*}))\,|\nabla T^{-1}(\boldsymbol{\theta}^{*})|$
4. $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}^{*}$ if

$$\begin{aligned} U < \frac{p(\boldsymbol{\theta}^{*})}{M\,q(\boldsymbol{\theta}^{*})} &= \frac{p(\boldsymbol{\theta}^{*})}{M\,p(T^{-1}\,(\boldsymbol{\theta}^{*}))\,|\nabla T^{-1}(\boldsymbol{\theta}^{*})|} \\ &= \frac{p(\boldsymbol{\theta}^{*})}{M\,p(\boldsymbol{\theta}^{\dagger})}|\nabla T(\boldsymbol{\theta}^{\dagger})| \end{aligned}$$

Metropolis-Hastings-Green

An MCMC algorithm for sampling from $p(\boldsymbol{\theta})$. Let $\boldsymbol{\theta} = \boldsymbol{\theta}^{(s-1)}$ be the former Markov chain state.

1. Draw $\boldsymbol{\psi} \sim q(\boldsymbol{\theta}, \cdot)$, $U \sim U(0, 1)$;
2. Obtain $(\boldsymbol{\theta}^*, \boldsymbol{\psi}^*) = T(\boldsymbol{\theta}, \boldsymbol{\psi})$;
3. Obtain

$$r(\boldsymbol{\theta}, \boldsymbol{\psi}) = \frac{p(\boldsymbol{\theta}^*) q(\boldsymbol{\theta}^*, \boldsymbol{\psi}^*)}{p(\boldsymbol{\theta}) q(\boldsymbol{\theta}, \boldsymbol{\psi})} \cdot |\nabla T(\boldsymbol{\theta}, \boldsymbol{\psi})| \, ;$$

4. if $U < r(\boldsymbol{\theta}, \boldsymbol{\psi})$, then $\boldsymbol{\theta}^{(s)} \leftarrow \boldsymbol{\theta}^*$,
   else $\boldsymbol{\theta}^{(s)} \leftarrow \boldsymbol{\theta}$.

HMC main idea: let $T(\cdot, \cdot)$ describe the evolution of a Hamiltonian system. But why?

Change of notation

1. $\boldsymbol{\theta} \mapsto \mathsf{q}$

2. $\boldsymbol{\psi} \mapsto \mathsf{p}$

3. $p(\boldsymbol{\theta}|y) \mapsto \pi(\mathsf{q})$

Hamiltonian dynamics

- $\mathsf{q} \in \mathbb{R}^D$ is the position of an object
- $\mathsf{p} \in \mathbb{R}^D$ is the momentum of an object
- $\mathsf{M} \in \mathbb{R}^{D \times D}$ is the mass matrix of an object
- $U(\mathsf{q}) = -\log \pi(\mathsf{q})$ is the system's potential energy
- $K(\mathsf{p}) = \mathsf{p}^T \mathsf{M}^{-1} \mathsf{p}/2$ is the kinetic energy
- $H(\mathsf{q}, \mathsf{p}) = U(\mathsf{q}) + K(\mathsf{p})$ is the total energy
- System governed by Hamiltonian equations:

$$
\begin{aligned}
\frac{\mathrm{d}\mathsf{q}}{\mathrm{d}t} &= \frac{\partial H}{\partial \mathsf{p}} = \mathsf{M}^{-1}\mathsf{p} \\
\frac{\mathrm{d}\mathsf{p}}{\mathrm{d}t} &= -\frac{\partial H}{\partial \mathsf{q}} = \nabla \log \pi(\mathsf{q}) \; .
\end{aligned}
$$

- The unique function $T_t : (\mathsf{q}_0, \mathsf{p}_0) \longmapsto (\mathsf{q}_t, \mathsf{p}_t)$ satisfies these equations.

Hamiltonian dynamics are

1. reversible: replace p with $-$p
2. volume preserving ($|\nabla T| = 1$):

$$\nabla \cdot \left(\frac{\mathrm{dq}}{\mathrm{d}t}, \frac{\mathrm{dp}}{\mathrm{d}t}\right) = \sum_{d=1}^{D} \left(\frac{\partial}{\partial \mathrm{q}_d}\frac{\mathrm{dq}_d}{\mathrm{d}t} + \frac{\partial}{\partial \mathrm{p}_d}\frac{\mathrm{dp}_d}{\mathrm{d}t}\right)$$
$$= \sum_{d=1}^{D} \left(\frac{\partial}{\partial \mathrm{q}_d}\frac{\partial H}{\partial \mathrm{p}_d} - \frac{\partial}{\partial \mathrm{p}_d}\frac{\partial H}{\partial \mathrm{q}_d}\right) = 0$$

3. energy conserving:

$$\frac{\mathrm{d}H}{\mathrm{d}t} = \sum_{d=1}^{D} \left(\frac{\partial H}{\partial \mathrm{q}_d}\frac{\mathrm{dq}_d}{\mathrm{d}t} + \frac{\partial H}{\partial \mathrm{p}_d}\frac{\mathrm{dp}_d}{\mathrm{d}t}\right)$$
$$= \sum_{d=1}^{D} \left(\frac{\partial H}{\partial \mathrm{q}_d}\frac{\partial H}{\partial \mathrm{p}_d} - \frac{\partial H}{\partial \mathrm{q}_d}\frac{\partial H}{\partial \mathrm{p}_d}\right) = 0$$

Hamiltonian Monte Carlo (sort of)

Augment parameter space with auxiliary Gaussian variable p and construct a Hamiltonian energy function:

$$\begin{aligned} H(\mathrm{q}, \mathrm{p}) &= -\log(\pi(\mathrm{q}) \times \phi_M(\mathrm{p})) \\ &\propto -\log \pi(\mathrm{q}) + \frac{1}{2}\mathrm{p}^T \mathrm{M}^{-1}\mathrm{p}\,. \end{aligned}$$

Given $\mathrm{q}_0 = \mathrm{q}_t^{(s-1)}$ a new state of the Markov chain is proposed by forward integrating Hamilton's equations for time t.

1. Draw $\mathrm{p}_0 \sim N(0, \mathrm{M})$;
2. Obtain $(\mathrm{q}_t, \mathrm{p}_t) = T_t(\mathrm{q}_0, \mathrm{p}_0)$;
3. Accept q_t with probability

$$1 \wedge \frac{\pi(\mathrm{q}_t)\phi_M(\mathrm{p}_t)}{\pi(\mathrm{q}_0)\phi_M(\mathrm{p}_0)} \cdot |\nabla T| = 1 \wedge \frac{\pi(\mathrm{q}_t)\phi_M(\mathrm{p}_t)}{\pi(\mathrm{q}_0)\phi_M(\mathrm{p}_0)}$$

Hamiltonian Monte Carlo (sort of)

1. Draw $\mathrm{p}_0 \sim N(0, \mathrm{M})$;
2. Obtain $(\mathrm{q}_t, \mathrm{p}_t) = T_t(\mathrm{q}_0, \mathrm{p}_0)$;
3. Accept q_t with probability

$$1 \wedge \frac{\pi(\mathrm{q}_t)\phi_M(\mathrm{p}_t)}{\pi(\mathrm{q}_0)\phi_M(\mathrm{p}_0)} \cdot |\nabla T| = 1 \wedge \frac{\pi(\mathrm{q}_t)\phi_M(\mathrm{p}_t)}{\pi(\mathrm{q}_0)\phi_M(\mathrm{p}_0)}$$

But wait!!

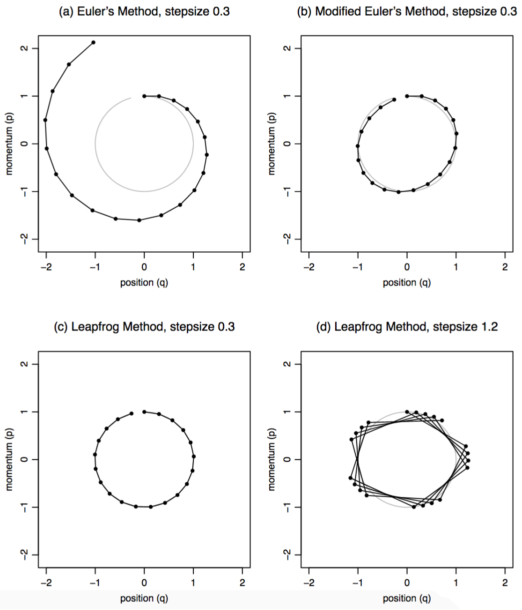
$$\begin{aligned}\frac{\pi(\mathrm{q}_t)\phi_M(\mathrm{p}_t)}{\pi(\mathrm{q}_0)\phi_M(\mathrm{p}_0)} &= \exp\left(\log\left(\frac{\pi(\mathrm{q}_t)\phi_M(\mathrm{p}_t)}{\pi(\mathrm{q}_0)\phi_M(\mathrm{p}_0)}\right)\right) \\ &= \exp\big(H(\mathrm{q}_0, \mathrm{p}_0) - H(\mathrm{q}_t, \mathrm{p}_t)\big) = 1\end{aligned}$$

Leapfrog integrator

We need a numerical integrator to solve the Hamiltonian equations. The most popular is the Störmer-Verlet (velocity Verlet) or leapfrog method.

$$
\begin{aligned}
\mathsf{p}(t+\epsilon/2) &= \mathsf{p}(t) + \frac{\epsilon}{2}\nabla \log \pi(\mathsf{q}(t)) \\
\mathsf{q}(t+\epsilon) &= \mathsf{q}(t) + \epsilon\, \mathsf{M}^{-1}\mathsf{p}(t+\epsilon/2) \\
\mathsf{p}(t+\epsilon) &= \mathsf{p}(t+\epsilon/2) + \frac{\epsilon}{2}\nabla \log \pi(\mathsf{q}(t+\epsilon))
\end{aligned}
$$

Leapfrog integrator



Neal (2011)

Leapfrog integrator

Still reversible (flip p) and volume preserving. To see the latter, the Jacobians are

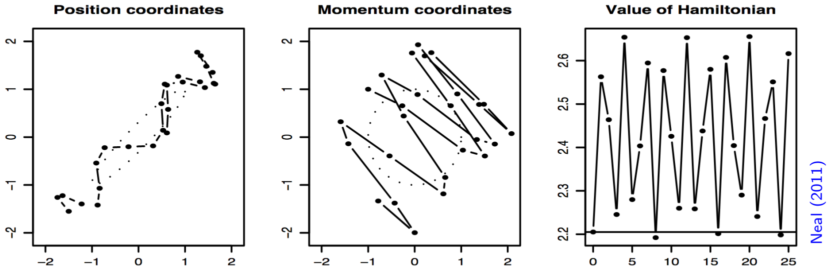
$$\nabla \hat{T}_p(\mathrm{q}, \mathrm{p}) = \begin{pmatrix} 1 & 0 \\ \frac{\epsilon}{2} \nabla^2 \log \pi(\mathrm{q}) & 1 \end{pmatrix}$$

and

$$\nabla \hat{T}_q(\mathrm{q}, \mathrm{p}) = \begin{pmatrix} 1 & \epsilon \mathrm{M}^{-1} \\ 0 & 1 \end{pmatrix} .$$

Unfortunately, we lose energy conservation. The error incurred by a leapfrog trajectory is $O(\epsilon^2)$.

Leapfrog integrator



Hamiltonian Monte Carlo

For $t = L \times \epsilon$:

1. Draw $\mathsf{p}_0 \sim N(0, \mathsf{M})$;
2. Obtain $(\mathsf{q}_t, \mathsf{p}_t) = \hat{T}_\epsilon^L(\mathsf{q}_0, \mathsf{p}_0)$;
3. Accept q_t with probability

$$1 \wedge \frac{\pi(\mathsf{q}_t)\phi_M(\mathsf{p}_t)}{\pi(\mathsf{q}_0)\phi_M(\mathsf{p}_0)} \cdot |\nabla \hat{T}| = 1 \wedge \frac{\pi(\mathsf{q}_t)\phi_M(\mathsf{p}_t)}{\pi(\mathsf{q}_0)\phi_M(\mathsf{p}_0)}$$

Challenges

1. Ill-conditioned target distributions

2. Multimodal target distributions

3. Big data

4. Fast, flexible and friendly software

Things we do in practice

1. Jitter L, the number of leapfrog iterations, at each step.

2. Choose ϵ so that majority (55%, say?) of proposals are accepted (for, say, $L = 100$).

3. Adapt L so that 70% or 80% or 90% of proposals are accepted.

4. Adapt M using the log posterior Hessian or the empirical covariance of p.

5. Parallel implementation of computational bottlenecks, i.e., log likelihood and its gradient.

Neal citation

Some images on these slides were taken from

Neal, R. M. (2011). *MCMC using Hamiltonian dynamics*. Handbook of Markov chain Monte Carlo, 2(11), 2.