# Lecture 6: Hamiltonian Monte Carlo

## STATS305C: Applied Statistics III

Scott Linderman

April 20, 2023

# Last Time...

- Metropolis-Hastings, Gibbs Sampling
- Probabilistic PCA, Factor Analysis, and Friends

# Today...

**Outline:**

- Hamiltonian Monte Carlo

**Reading:**

- MCMC using Hamiltonian dynamics [Neal, 2012]
- Optional: A Conceptual Introduction to Hamiltonian Monte Carlo [Betancourt, 2017]

# How Can We Make Smarter Proposals?

- Metropolis-Hastings with a symmetric Gaussian proposal behaves (kind of) like a random walk.
- Neal [2012] argues that in $D$ dimensions, random walk MH needs $O(D^2)$ iterations to get an independent sample.
- Can we develop more efficient transition distributions?
    - **Yes**! If we have more information about the log probability.
- For example, suppose that the log probability $\log p(\theta)$ is differentiable. We can use the gradient to make proposals that move farther and are more likely to be accepted.

# Metropolis Adjusted Langevin Algorithm (MALA)

The *Metropolis-Adjusted Langevin Algorithm* uses the gradient of the log probability to make asymmetric proposals,

$$q(\boldsymbol{\theta}' \mid \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta} + \tau \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}, \boldsymbol{X}), 2\tau^2 \boldsymbol{I}) \tag{1}$$

**Note:** $q(\boldsymbol{\theta}' \mid \boldsymbol{\theta}) \neq q(\boldsymbol{\theta} \mid \boldsymbol{\theta}')$! To calculate the acceptance probability, you need the gradient at both points.

MALA can be motivated as a discrete-time approximation to the *Langevin* diffusion, a continuous-time stochastic differential equation for modeling molecular dynamics.

In high dimensions, the extra information provided by the gradient can lead to much more efficient chains. Neal argues that MALA needs $O(D^{4/3})$ computation to produce an independent sample.

But why stop at one gradient step?

# Hamiltonian Monte Carlo

**Reference:** Neal [2012] *MCMC using Hamiltonian dynamics.*

**Idea:** *Think of negative log probability as an energy landscape. Now imagine a puck sliding around on this bumpy surface. Give it random kicks; it will tend to slide downhill toward points of low potential energy (high probability). Each kick can displace the puck by a large amount. Done properly, the puck will visit points with probability proportional to the posterior probability.*

# Notation

Following Neal [2012], let

- $\boldsymbol{q} \in \mathbb{R}^D$ denote the *position*; i.e. the current parameters (previously $\theta$)
- $\boldsymbol{p} \in \mathbb{R}^D$ denote the *momentum*; auxiliary variables that we don't care about, but which are necessary for HMC.
- $\boldsymbol{z} = [\boldsymbol{q}, \boldsymbol{p}]^\top \in \mathbb{R}^{2D}$ denote the combined *state of the system.*
- $\boldsymbol{M}$ denote the *mass matrix*, another artificial construct. Typically, this will be $m\boldsymbol{I}$
- $U(\boldsymbol{q})$ denote the *potential energy*
- $K(\boldsymbol{p}) = \frac{1}{2}\boldsymbol{p}^\top \boldsymbol{M}^{-1}\boldsymbol{p}$ denote the *kinetic energy*

# Hamiltonian Dynamics

The *Hamiltonian* is the sum of the potential $H(\boldsymbol{q}, \boldsymbol{p}) = U(\boldsymbol{q}) + K(\boldsymbol{p}) = U(\boldsymbol{q}) + \frac{1}{2}\boldsymbol{p}^\top \boldsymbol{M}^{-1}\boldsymbol{p}$.

The partial derivatives determine how the state evolves over time,

$$\frac{\mathrm{d}q_d}{\mathrm{d}t} = \frac{\partial H}{\partial p_d} = [\boldsymbol{M}^{-1}\boldsymbol{p}]_d \tag{2}$$

$$\frac{\mathrm{d}p_d}{\mathrm{d}t} = -\frac{\partial H}{\partial q_d} = -\frac{\partial U}{\partial q_d} \tag{3}$$

for $d = 1, \ldots, D$.

Compactly,

$$\frac{\mathrm{d}\boldsymbol{z}}{\mathrm{d}t} = \boldsymbol{J}\nabla H(z) \tag{4}$$

where

$$\boldsymbol{J} = \begin{bmatrix} \boldsymbol{0}, \boldsymbol{I} \\ -\boldsymbol{I}, \boldsymbol{0} \end{bmatrix} \tag{5}$$

# One Dimensional Example

Consider the case where $D = 1$ and $U(q) = \frac{1}{2}q^2$ and $K(p) = \frac{1}{2}p^2$.

The partial derivatives are

$$\frac{\partial H}{\partial p} = p \tag{6}$$

$$-\frac{\partial H}{\partial q} = -q \tag{7}$$

so

$$\frac{\mathrm{d}\boldsymbol{z}}{\mathrm{d}t} = \boldsymbol{J}\boldsymbol{z}. \tag{8}$$

This is a linear dynamical system, and the state at time $t + \Delta t$ is $\boldsymbol{z}(t + \Delta t) = e^{\boldsymbol{J}\Delta t}\boldsymbol{z}(t)$.

Since $\boldsymbol{J}\Delta t$ is skew-symmetric, the matrix exponential $e^{\boldsymbol{J}\Delta t}$ is orthogonal. More precisely, $\boldsymbol{z}(t + \Delta t)$ is a rotation about the origin of $\boldsymbol{z}(t)$.

# Properties of Hamiltonian Dynamics

1. **Reversibility:** The mapping from $z(t) \rightarrow z(t + \Delta t)$ is one-to-one and invertible. To go from $z(t + \Delta t)$ to $z(t)$, negate $p(t + \Delta t)$, apply the the Hamiltonian dynamics for $\Delta t$ time, and negate the momentum again.

2. **Conservation of energy:** The Hamiltonian (which is the total energy in a closed system) is conserved,

$$\frac{\mathrm{d}H}{\mathrm{d}t} = \sum_{d=1}^{D} \frac{\mathrm{d}q_d}{\mathrm{d}t}\frac{\partial H}{\partial q_d} + \frac{\mathrm{d}p_d}{\mathrm{d}t}\frac{\partial H}{\partial p_d} \tag{9}$$

$$= \sum_{d=1}^{D} \frac{\partial H}{\partial p_d}\frac{\partial H}{\partial q_d} - \frac{\partial H}{\mathrm{d}q_d}\frac{\partial H}{\partial p_d} = 0. \tag{10}$$

# Properties of Hamiltonian Dynamics II

3. **Volume preserving:** A set in $(q, p)$ space will have the same volume after being mapped through Hamiltonian dynamics. This follows from the fact that the divergence of the vector field is zero everywhere:

$$\operatorname{div}\frac{\mathrm{d}\boldsymbol{z}}{\mathrm{d}t} = \sum_{d=1}^{D}\frac{\partial}{\partial q_d}\frac{\mathrm{d}q_d}{\mathrm{d}t} + \frac{\partial}{\partial p_d}\frac{\mathrm{d}p_d}{\mathrm{d}t} = \sum_{d=1}^{D}\frac{\partial}{\partial q_d}\frac{\partial H}{\partial p_d} - \frac{\partial}{\partial p_d}\frac{\partial H}{\partial q_d} = \sum_{d=1}^{D}\frac{\partial^2 H}{\partial q_d \partial p_d} - \frac{\partial^2 H}{\partial q_d \partial p_d} = 0. \tag{11}$$

4. **Sympleticness** Let $B$ be the Jacobian of the transformation from $\boldsymbol{z}(t) \to \boldsymbol{z}(t + \Delta t)$. It turns out that,

$$\boldsymbol{B}^\top \boldsymbol{J}^{-1}\boldsymbol{B} = \boldsymbol{J}^{-1} \tag{12}$$

which implies that $|\boldsymbol{B}^\top||\boldsymbol{J}^{-1}||\boldsymbol{B}| = |\boldsymbol{J}^{-1}|$ and thus $|\boldsymbol{B}| = 1$. I.e. the dynamics preserve volume.

# Discretizing Hamilton's Equations

The properties above apply to the *continuous time* Hamiltonian dynamics. Can we maintain them in practice?

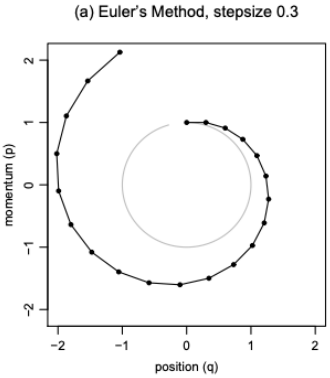
**Idea:** In practice, to simulate $\Delta t$ elapsed time, we break it down into steps of size $\Delta t/\epsilon$.

**Euler's method:** Update the state as,

$$\boldsymbol{z}(t+\epsilon) = \boldsymbol{z}(t) + \epsilon \frac{\mathrm{d}\boldsymbol{z}}{\mathrm{d}t}\bigg|_{\boldsymbol{z}(t)} \tag{13}$$

$$\Rightarrow p_d(t+\epsilon) = p_d(t) - \epsilon \frac{\partial U}{\partial q_d}\bigg|_{\boldsymbol{q}(t)} \tag{14}$$

$$q_d(t+\epsilon) = q_d(t) + \epsilon \frac{p_d(t)}{m_d} \tag{15}$$



Simple Euler integration does not preserve volume: trajectories eventually diverge, even with small $\epsilon$.

# The Leapfrog Integrator

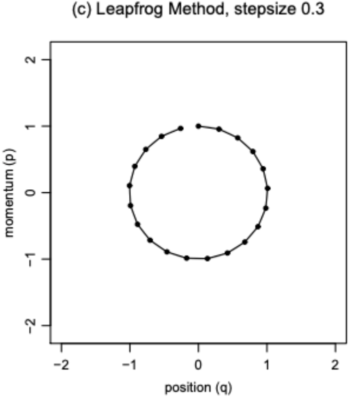
Instead, alternate updates of $p$ and $q$

$$p_d(t+\tfrac{\epsilon}{2}) = p_d(t) - \frac{\epsilon}{2}\frac{\partial U}{\partial q_d}\bigg|_{\boldsymbol{q}(t)} \tag{16}$$

$$q_d(t+\epsilon) = q_d(t) + \epsilon\frac{p_d(t+\tfrac{\epsilon}{2})}{m_d} \tag{17}$$

$$p_d(t+\epsilon) = p_d(t+\tfrac{\epsilon}{2}) - \frac{\epsilon}{2}\frac{\partial U}{\partial q_d}\bigg|_{\boldsymbol{q}(t+\epsilon)} \tag{18}$$

$$\tag{19}$$



Each update is a *shear transformation* in which only some variables change, by amounts that depend on the other, fixed variables. The determinant of such a transformation is one, so it preserves volume.

# Using Hamiltonian Dynamics for Posterior Inference

Define a joint distribution on positions and momenta as,

$$p(\boldsymbol{q}, \boldsymbol{p}) \propto \exp\left\{-H(\boldsymbol{q}, \boldsymbol{p})\right\} \propto \exp\left\{-U(\boldsymbol{q}) - K(\boldsymbol{p})\right\}. \tag{20}$$

Now let $U(\boldsymbol{q}) = -\log p(\boldsymbol{\theta} = \boldsymbol{q}, \boldsymbol{X})$ be the *negative* log joint probability. Then,

$$p(\boldsymbol{q}, \boldsymbol{p}) = p(\boldsymbol{\theta} = \boldsymbol{q} \mid \boldsymbol{X}) \times p(\boldsymbol{p}) \tag{21}$$

Samples of $\boldsymbol{q}$ will be marginally distributed according to the posterior $p(\boldsymbol{\theta} = \boldsymbol{q} \mid \boldsymbol{X})$.

Samples of $\boldsymbol{p}$ will be marginally distributed $p(\boldsymbol{p}) = \frac{\exp\{-K(\boldsymbol{p})\}}{\int_{\mathbb{R}^D} \exp\{-K(\boldsymbol{p})\} \, \mathrm{d}\boldsymbol{p}}$. These are *auxiliary variables* that we don't really care about—they're just there to help us construct MH proposals.

We choose $K(\boldsymbol{p})$ so $p(\boldsymbol{p})$ is convenient; e.g. if $K(\boldsymbol{p}) = \frac{1}{2}\boldsymbol{p}^\top \boldsymbol{M}^{-1} \boldsymbol{p}$ then

$$p(\boldsymbol{p}) = \mathcal{N}(\boldsymbol{p} \mid \boldsymbol{0}, \boldsymbol{M}). \tag{22}$$

# Hamiltonian Monte Carlo (HMC)

**Hamiltonian Monte Carlo (HMC)** is Metropolis-Hastings on the joint distribution of $(\boldsymbol{q}, \boldsymbol{p})$ with proposals based on Hamiltonian dynamics.

Starting at point $(\boldsymbol{q}', \boldsymbol{p}')$, sample the proposal distribution:

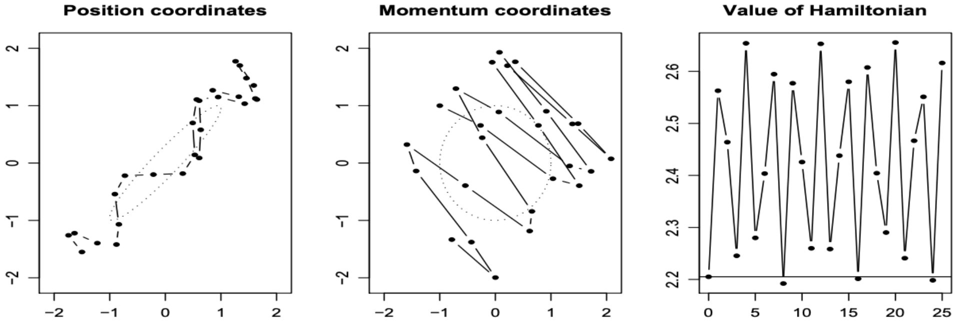
1. Throw away $\boldsymbol{p}'$ and sample new momenta from their marginal distribution $\boldsymbol{p} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{M})$.
2. Approximate Hamiltonian dynamics on $(\boldsymbol{q}, \boldsymbol{p})$ for $\Delta t$ time using $L = \Delta t / \epsilon$ Leapfrog steps each of size $\epsilon$. Call the resulting point $(\boldsymbol{q}, \boldsymbol{p})$.
3. Flip the momentum $\boldsymbol{p} \leftarrow -\boldsymbol{p}$ to make the proposal symmetric.

Then accept the proposed point $(\boldsymbol{q}, \boldsymbol{p})$ with probability,

$$a((\boldsymbol{q}', \boldsymbol{p}') \to (\boldsymbol{q}, \boldsymbol{p})) = \min\left\{1, \frac{\exp\{-H(\boldsymbol{q}, \boldsymbol{p})\}\, q(\boldsymbol{q}', \boldsymbol{p}' \mid \boldsymbol{q}, \boldsymbol{p})}{\exp\{-H(\boldsymbol{q}', \boldsymbol{p}')\} q(\boldsymbol{q}, \boldsymbol{p} \mid \boldsymbol{q}', \boldsymbol{p}')}\right\} = \min\left\{1, \frac{\exp\{-H(\boldsymbol{q}, \boldsymbol{p})\}}{\exp\{-H(\boldsymbol{q}', \boldsymbol{p}')\}}\right\}. \quad (23)$$

If the Hamiltonian dynamics were simulated exactly, HMC would always accept. In practice, differences arise from numerical integration errors.

# HMC Dynamics on a Correlated 2D Gaussian



Position coordinates

Momentum coordinates

Value of Hamiltonian

Figure 3: A trajectory for a 2D Gaussian distribution, simulated using 25 leapfrog steps with a stepsize of 0.25. The ellipses plotted are one standard deviation from the means. The initial state had $q = [-1.50, -1.55]^T$ and $p = [-1, 1]^T$.

# HMC vs Random Walk MH

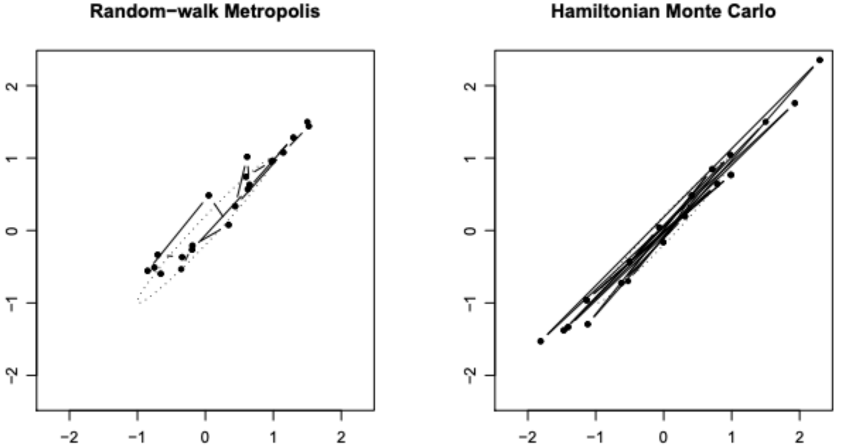


Figure 4: Twenty iterations of the random-walk Metropolis method (with 20 updates per iteration) and of the Hamiltonian Monte Carlo method (with 20 leapfrog steps per trajectory) for a 2D Gaussian distribution with marginal standard deviations of one and correlation 0.98. Only the two position coordinates are plotted, with ellipses drawn one standard deviation away from the mean.

# HMC vs Random Walk MH II

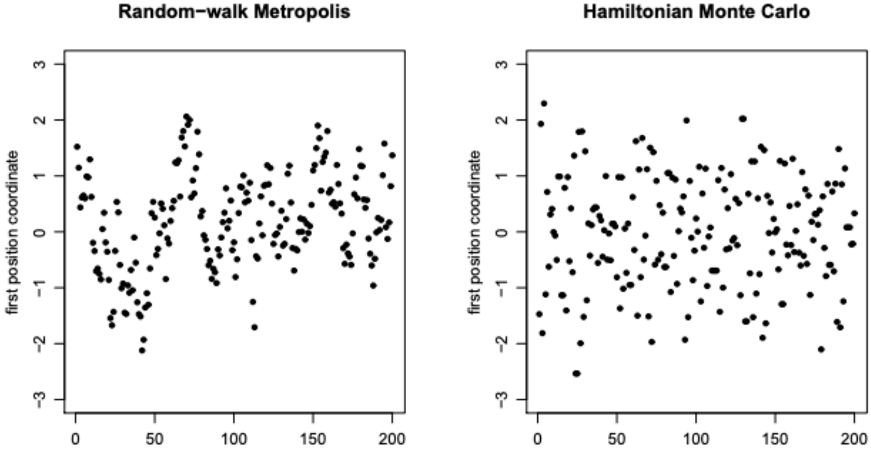


Figure 5: Two hundred iterations, starting with the twenty iterations shown above, with only the first position coordinate plotted.

# HMC vs Random Walk MH in 100D

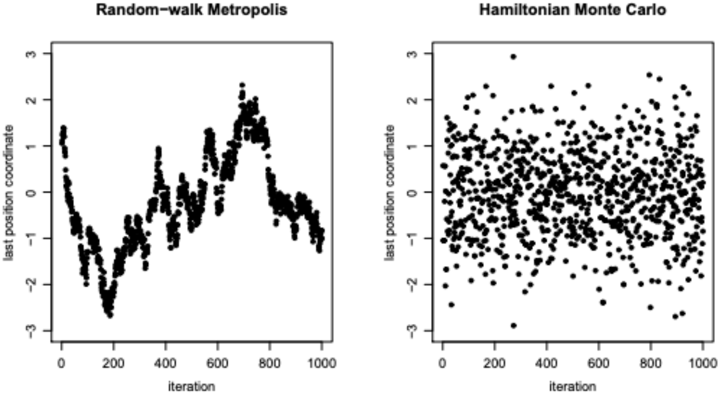
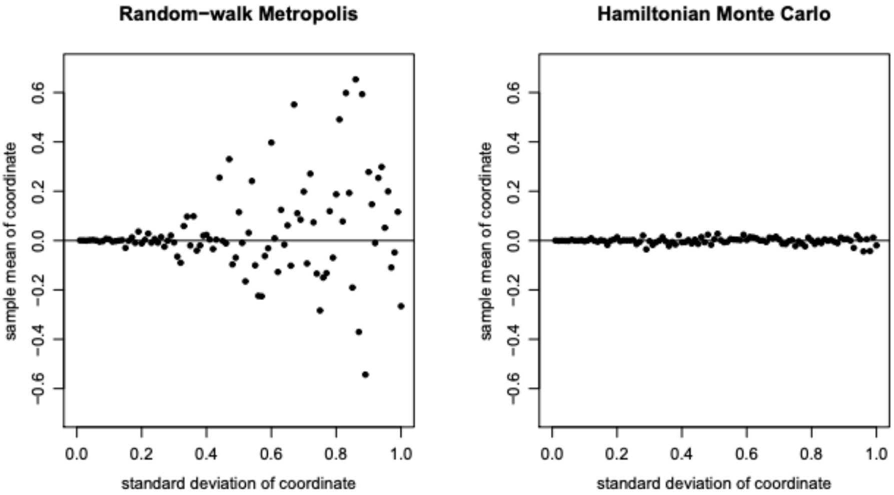


Figure 6: Values for the variable with largest standard deviation for the 100-dimensional example, from a random-walk Metropolis run and an HMC run with $L = 150$. To match computation time, 150 updates were counted as one iteration for random-walk Metropolis.

# HMC vs Random Walk MH in 100D II



# Benefits of Avoiding Random Walks

- To maintain reasonably high acceptance probability, random walk MH needs proposal standard deviation (s.d.) comparable to the s.d. in the most constrained dimension (0.14 in the 2D Gaussian example and 0.01 in the 100D example).

- Num. iterations needed for RW-MH to reach an approximately independent state is proportional to the *square* of the largest standard deviation to the smallest; i.e. to the condition number of the covariance matrix.

- In contrast, integrating the Hamiltonian makes many steps in the same direction. The number of integration steps to reach an independent state is about the ratio of the largest s.d. to the smallest; i.e. the square root of the condition number.

- Neal [2012] argues that the number of leapfrog updates to reach an independent point scales as $O(D^{5/4})$, better than the $O(D^2)$ and $O(D^{4/3})$ estimates for random walk MH and MALA, respectively.

- However, we still need to tune the step size $\epsilon$ to be comparable to the smallest s.d.

# Adapting the step size

- A simple strategy is to tune the step size adaptively during the initial run of the Markov chain.
- For example, set a target acceptance rate (Neal argues that it should be around 0.65), then increase the step size if you're accepting too often and decrease if you're rejecting too often.
- Andrieu and Thoms [2008] proposed a widely-used multiplicative update scheme; it is the default in `tfp.mcmc.SimpleStepSizeAdaptation`. Pyro defaults to a similar "dual averaging" scheme.
- The **No U-Turn Sampler (NUTS)** [Hoffman and Gelman, 2014] adapts the distance traveled in response to the curvature of the target density. Conceptually, it continues until the trajectory turns back on itself (hence the name, "No U-Turn")
- More details can be found in Betancourt [2017].

# Demos

https://chi-feng.github.io/mcmc-demo/app.html

# References I

Radford M Neal. MCMC using Hamiltonian dynamics. June 2012.

Michael Betancourt. A conceptual introduction to Hamiltonian Monte Carlo. January 2017.

Christophe Andrieu and Johannes Thoms. A tutorial on adaptive MCMC. *Stat. Comput.*, 18(4):343–373, December 2008.

Matthew D Hoffman and Andrew Gelman. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.