6.882: Bayesian Modeling and Inference **Lecture Date: March 8, 2018**

Hamiltonian Monte Carlo

Lecturers:
Sushruth Redd, Rohan Banerjee, Isaac Kontomah

Scribes:
Nick Hoernle, Brian Trippe, Brandon Zeng

1 Outline

- Introduction
- Physics
  - Hamiltonian Mechanics
  - Statistical Mechanics
- MCMC using Hamiltonian Dynamics
  - Leapfrog Method
  - Algorithm
  - Illustration
- Tuning
  - Tuning ϵ
  - Tuning L
- Results

2 Introduction

We have now studied a range of sampling techniques. Each of these techniques struggle to adapt to higher dimensional problems:

- Random Walk Metropolis-Hastings
  - Use the current sample x to make a proposal x' that is a draw from some user defined transition function $q(x' \mid x)$.
  - Metropolis algorithm defines an acceptance probability A for accepting the move to a new state. $A = \min\left(\frac{p(x')q(x|x')}{p(x)q(x'|x)}\right)$
  - If we let the proposal distribution $q(x' \mid x) = \mathcal{N}(x'; x, 1)$, we can study the distribution of the location after t steps. Let $d^{(t)}(x)$ denote this location where $d^{(t)}(x) = \sum_{i=1}^{t} x^{(i)}$. It can now be noted that $E[d^{(t)}(x)] = 0$ and $Var[d^{(t)}(x)] = t\sigma^2$, meaning the amount of time t to travel a distance of L is $\approx (L/\sigma)^2$. The random walk behavior of this algorithm results in unnecessary computation to traverse length L.

- Gibbs Sampling
  - Uses *full conditionals* to update a subset of the variables (dependent on the variables that are not in that subset) at each step.
  - Calculating the full conditionals can be analytically intractable or can require undesirable computation.
- Slice Sampling
  - Finding suitable hyper-rectangles that include the slice becomes non-trivial in higher dimensions.
  - Recall that we desire the smallest hyper-rectangle that contains the slice (such that we don't waste samples outside the slice and such that we can explore the **entire** support of the slice). If these rectangles are too large, we reject too many samples (it is also not clear how to perform efficient shrinkage in higher dimensions). If the rectangles are too small, we return to a RW behaviour with problems of stinted exploration of the target distribution.

The primary reason for studying Hamiltonian Monte Carlo (HMC) is to target the random walk behaviour of the previously seen MCMC algorithms. We desire an algorithm that explores the target distribution in a systematic manner and performs consistently with an increase of dimensionality of the target distribution. See Section 6 for an example of the performance of Random Walk Metropolis-Hastings and HMC on a high dimensional Gaussian with varied standard deviations in the different dimensions.

3 Physical Intuition and Motivation

The intuition behind the formulation of HMC arises from a series of observations from physics. First is that a system of one or more interacting particles at a constant temperature environment will occupy a particular microstate, m, with probability proportional to the exponential of the negative energy, E_m, of that microstate:

$$p(m) = \frac{1}{Z} e^{\frac{-E_m}{T*k}}$$

Where T is the temperature of the system, k is Boltzman's constant, and Z is a normalizing constant, which depends on the energies of all other accessible states. In particular, we consider a system consisting of a single particle whose state is entirely defined by its position and it momentum. In this case, the particle's time evolution can uniquely be described by Hamiltonian dynamics. A complete discussion of how the states of systems evolve such that they lead sampling this exact distribution over states is in the realm of statistical mechanics and is outside of the scope of this course. However, the observation that the dynamics of physical system lead particles to sample from enormously complex distributions motivates the simulation of these dynamics in silico to sample from distributions of interest. This observation had been made prior to the introduction of HMC to the statistics community, in applications such as molecular dynamics simulations. Before diving into the application of these methods to statistical models, we briefly review the Hamiltonian dynamics simulated in HMC and touch on the core statistical mechanics which underlies the connection to sampling from the distribution of microstates. As we will see, these dynamics have the useful property that they lead particles to coherently traverse space, diverging from the random walk behaviour of the random-walk Metropolis algorithm.

3.1 Hamiltonian Dynamics

The time evolution of particles as described above is uniquely described by Hamiltonian dynamics. We now make these dynamics explicit.

Definition 1. The Hamiltonian H is defined as $H = K(p) + U(q)$, where p is momentum and q is location. $K(p)$ is the kinetic energy, defined as $K(p) = \frac{p^2}{2m}$. As will cover in Section 4, this motivates drawing momentum from a Gaussian proposal in the HMC algorithm. $U(q)$ is known as the potential energy, and in HMC this is tied to he log probability.

The Hamiltonian dynamics are described by the pair of differential equations below:

$$\dot{p} = -\frac{\partial H}{\partial q} \tag{1}$$

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

We can consider a harmonic oscillator (e.g. a mass connected to a spring) as an example, where $U(q) = \frac{q^2}{2}$ and $H = \frac{p^2}{2} + \frac{q^2}{2}$. The Hamiltonian (H) describes the total energy in the system which we know is conserved. Plotting the (p, q) evolution over time gives Figure 1. Considering only the position axis will lead to the motion that we would observe (observing the change in position through time).

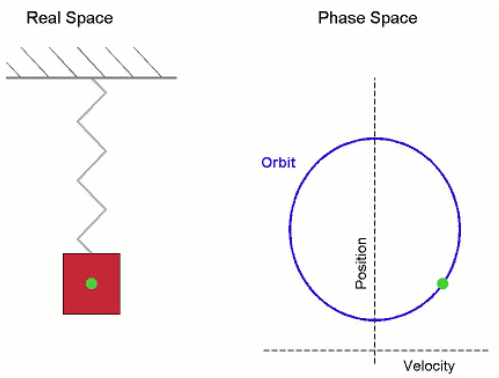

Figure 1: Example of a Hamiltonian system with a mass connected to a spring and the position and momentum (velocity) evolve along the plotted circle.

Following the physical theory, we identify desirable characteristics for deriving Hamiltonian Monte Carlo:

1. Energy is conserved.

2. Dynamics are reversible. The condition requires that reversing the momentum means the position variable will retrace any steps that have already been traversed.

3. Phase space volume is conserved. Evolving a slice of volume V for time t will result in a transformation that will also have volume V.

3.2 Statistical Mechanics

Statistical mechanics provides a perspective and language for describing how the dynamics of particles and the interactions between them lead equilibrium distributions over states, which is often useful for understanding the thermodynamic properties of materials and large systems. Here, we review two ideas in statistical mechanics which will provide intuition into how and why HMC converges to the correct stationary distribution.

Microcanonical Ensemble. Consider one particle in a system with constant energy such that it evolves according to Hamilton's equations. If we consider a system with a position vector of three dimensions, then there is an associated momentum vector of three dimensions. The particles dynamics can be described as a 5 dimensional hypersurface in a 6 dimensional space (as the constraint of constant energy uses one of these degrees of freedom). The stationary distribution, corresponding to the Hamiltonian dynamics is uniform over the hyper-surface. We therefore may consider the following representation for the Hamiltonian:

$$H = \frac{p_1^2 + p_2^2 + p_3^2}{2m} + \frac{x_1^2 + x_2^2 + x_3^2}{2} \tag{3}$$

Canonical Ensemble. Here we describe a system that has a constant temperature T. The distributions over possible position and momentum vectors is given by:

$$p(p, q) = \frac{\exp \frac{-H(p,q)}{K}}{Z} \tag{4}$$

We have two conditions on the system:

- p and q evolve via Hamiltonian dynamics
- the momentum can be randomized as a result of interacting with the environment. Notice in Section 4 when we introduce resampling the momentum coordinate (typically from a Gaussian proposal), this has the connection to these randomized interactions with the environment.

These physical examples provide intuition for why Hamiltonian dynamics with periodic re-sampling of momentum will allow us to sample from an unknown distribution. Introducing the momentum variable, allows us to simulate these deterministic dynamics through (p, q) space such that total energy and volume is conserved. The reversibility of the Hamiltonian dynamics is critical for ensuring detailed balance for the sampling. We therefore use Hamiltonian dynamics to make new proposals for the Metropolis algorithm. In contrast to random walk Metropolis, this leads to proposals which can be set by long trajectories which follow the curvature of the space, navigating through regions of lower energy which correspond to higher probability.

4 MCMC using Hamiltonian Dynamics

As mentioned earlier, we let q represent position and p represent momentum. In the context of posterior sampling, we can interpret q as the variables of interest that makes up our posterior and p as an auxiliary variable that allows us to make use of Hamiltonian dynamics. We will also use a canonical function to serve as a link between energy functions and distributions. More concretely, given $H(p, q) = U(q) + K(p)$, we have the joint distribution

$$P(p, q) = \frac{1}{z} \exp(-H(p, q)) = \frac{1}{Z} \exp(-U(q)) \exp(-K(p)).$$

In the context of Bayesian statistics, we can express the posterior distribution using

$$U(q) = \log[\pi(q)L(q|D)]$$

where $\pi(q)$ is the prior, and $L(q|D)$ the likelihood function given data D, and

$$K(p) = \sum_{i=1}^{d} \frac{p_i^2}{2m_i},$$

where each summand represents individual points i with variances m_i.

4.1 Leapfrog method

Before explaining the algorithm in more detail, we first discuss the leapfrog method, which we use to approximate Hamiltonian dynamics. The Hamiltonian dynamics are defined by a pair of differential equations and thus in general we require a numerical approximation to simulate from these differential equations. Euler's method is presented to simulate the dynamics but the energy conservation is clearly seen to not hold. The leapfrog integration method is therefore introduced to find a numerically stable approximation that conserves energy through the simulation. This integration scheme tailored for Hamiltonian dynamics is also known as a symplectic integration technique.

The leapfrog method adjusts Euler's method to make the following discrete updates to the position and momentum variables respectively.

$$\begin{aligned} p_i(t+\epsilon/2) &= p_i(t) - (\epsilon/2)\frac{\partial U}{\partial q_i}(q(t)) \\ q_i(t+\epsilon) &= q_i(t) + \epsilon\frac{p_i(t+\epsilon/2}{m_i}) \\ p_i(t+\epsilon) &= p_i(t+\epsilon/2) - (\epsilon/2)\frac{\partial U}{\partial q_i}(q(t+\epsilon)) \end{aligned}$$

4.2 Algorithm

We now have sufficient background for presenting the algorithm. The algorithm, starting from an initial state (q_0, p_0), is as follows:

1. Sample a new value for p.
2. Propose a new state $(q*, p*)$ by simulating the Hamiltonian dynamics for length L using step size ϵ. Accept or reject the state based on standard Metropolis criteria.

In the first step, we resample p in order to encourage exploration of different areas of the sample space of q. More formally, we can see that as $H(p, q) = U(q) + K(p)$, if $K(p)$ is fixed, then $U(q)$ is also fixed and thus limited.

In the second step, we identify our proposal state $(q*, p*)$ by starting with the current state, (q, p), and running the Leapfrog method for L steps with a stepsize of ϵ, before negating the momentum variables

(which is needed for a symmetric proposal probability). L and ϵ are thus the parameters of this algorithm, and tuning them is discussed below. Now, given this proposal state, we accept it with probability

$$\min\left(1, \frac{\exp(-H(q*, p*)}{\exp(-H(q,p))}\right) = \min\left(1, \frac{P(q*, p*)}{P(q,p)}\right).$$

To provide some intuition for this decision rule, we note that if the Hamiltonian dynamics are perfectly simulated, we should always accept a new proposal, but as we are only approximating Hamiltonian dynamics, there is room for error. The worse the numerical approximation, the less likely we are to accept the proposal.

4.3 Illustration

Here, we consider sampling from a bivariate Guassian centered at the origin with covariance $\Sigma = \left[\begin{smallmatrix} 1 & 0.95 \\ 0.95 & 1 \end{smallmatrix}\right]$, and introduce two "momentum" variables that are independent normally distributed with zero mean and standard deviation 1. We can then set up the Hamiltonian as

$$H(q,p) = q^T\Sigma^{-1}q/2 + p^Tp/2.$$

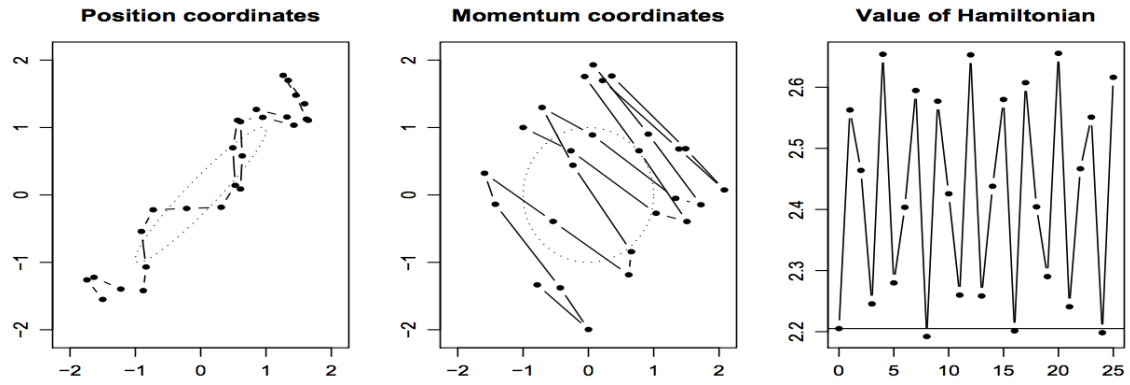

Figure 2: Sampling from a 2D Gaussian using 25 leapfrog steps with a stepsize of 0.25.

Figure 2 shows a possible sampling result from Hamiltonian MCMC, using $L = 25$, $\epsilon = 0.25$, and initial state $q = [-1.50, -1.55]^T, p = [-1, 1]^T$, The first and second plots indicate position coordinates and momentum coordinates, respectively, while the third plot shows the value of the Hamiltonian itself.

Studying these plots, we can see that the trajectory doesn't exhibit random walk behavior; indeed, starting from the lower left corner, the position variables systematically move to the right and up until they reach the upper right corner, after which the trajectory reverses. At the same time, momentum oscillates due to the correlation between the variables, but this does not affect our sampling too much. Finally, we note that the Hamiltonian is not perfectly conserved due to numerical errors, but our trajectory is still stable. We will now explore the conditions required for stability.

5 Tuning

Due to the numerical approximations that are introduced in the leapfrog method, the Hamiltonian dynamics are not exact and we still perform the acceptance check. Increasing the step size ϵ for the numerical

integration allows the algorithm to traverse the space more efficiently but it also introduces more numerical error and thus will increase the probability of rejection of a trajectory. Similarly, the length of the trajectory L is a tunable parameter. Ideally, we wish to simulate the Hamiltonian dynamics to reach new and unexplored parts of the distribution. As was seen in Figure 1, the dynamics can become periodic if the trajectory length is long enough (resulting in taking the position coordinate back to the starting point). This is highly undesirable as a lot of computation is done for little or no progress. Tuning the balance of these two parameters is therefore of paramount importance for this algorithm.

5.1 Tuning ϵ

Studying a simple one-dimension problem can be helpful but note that in general deriving results for appropriate choices of ϵ can be very hard. If we use the following Hamiltonian, we see that a leapfrog step can be understood as a linear mapping.

$$H(q,p) = \frac{q^2}{2^2} + \frac{p^2}{2} \tag{5}$$

The associated linear mapping is:

$$\begin{bmatrix} q(t+\epsilon) \\ p(t+\epsilon) \end{bmatrix} = \begin{bmatrix} 1 - \frac{\epsilon^2}{2\sigma^2} & \epsilon \\ -\frac{\epsilon}{\sigma^2} + \frac{\epsilon^3}{4\sigma^4} & 1 - \frac{\epsilon^2}{2\sigma^2} \end{bmatrix} \begin{bmatrix} q(t) \\ p(t) \end{bmatrix} \tag{6}$$

We can then study the eigenvalues of the linear mapping in Equation 6 to understand the stability when this mapping is repeatedly applied to the (position, momentum) vector. The eigenvalues are:

$$\left(1 - \frac{\epsilon^2}{2\sigma^2}\right) \pm \left(\frac{\epsilon}{\sigma}\right) \sqrt{\frac{\epsilon^2}{4\sigma^2} - 1} \tag{7}$$

If the magnitude of one of the eigenvalues is greater than 1, the mapping is unstable. We therefore require $\epsilon < 2\sigma$ for the trajectory to have stable dynamics.

The above stability analysis was for a single dimensional example where the kinetic energy has a quadratic form (and thus a Gaussian proposal is used). In higher dimensional examples, we can use the standard deviation of the most constrained dimension for the calculation of ϵ (again it is worth noting that this standard deviation might not readily be obtained either). Neal et al. (2011) proposes that ϵ can be chosen stochastically. The goal is for the mean of the distribution to be chosen appropriately, but even if this is not the case, if there is sufficient mass in the lower tail, we will still end up selecting values for ϵ that will be small enough to produce stable dynamics.

5.2 Tuning L

We wish to find states in the target distribution that are 'far' away from the current state (resulting in uncorrelated samples and good exploration of the target). Neal suggests a starting value of $L = 100$ and the *auto-correlation* of the samples must be studied to evaluate this choice. If the samples appear to be approximately independent then the value for L can be decreased. However, if there appears to be a large auto-correlation between successive samples (implying the trajectories have not travelled far enough in phase-space), then a longer value for L (say $L = 1000$) can be tried.

We tackle the automatic choice of these parameters in the NUTS section of this class.

6 Results

Recall that we wish to draw i.i.d samples from the posterior distribution. When we use MCMC techniques, the samples are dependent (by construction). Ideally, we desire samples that can be drawn anywhere on the support of the target distribution with probabilities that are representative of the density of the distribution. With RWMH and Gibbs, we require a proposal sample that depends on the position of the current sample. The new sample therefore is 'close' to the current sample (defined by the variance of the proposal distribution). Neal conducts an experiment where he uses RWMH and HMC to draw samples from a 100-dimension Gaussian distribution with means all set to 0 and standard deviations increasing from 0.01 to 1 with steps of 0.01. Figure 3 shows samples from the mean coordinate of the dimension with standard deviation of 1. Note that from MacKay (1998) we know for RWMH to have a reasonable acceptance probability, the standard deviation of the proposal distribution must be $sd \leq 0.01$ (the standard deviation from the dimension with the smallest standard deviation). We would therefore expect RWMH to struggle to explore the full support of the dimension with the largest standard deviation. The samples would be highly correlated and would not look like independent draws from this dimension. We see this exhibited in the trace from the RWMH draw. The samples have a high degree of auto-correlation and do not appear to be independent draws from this dimension. In contrast, the HMC solution appears to draw uncorrelated samples and thus we would expect that the mean estimator that uses these samples would be more effective.

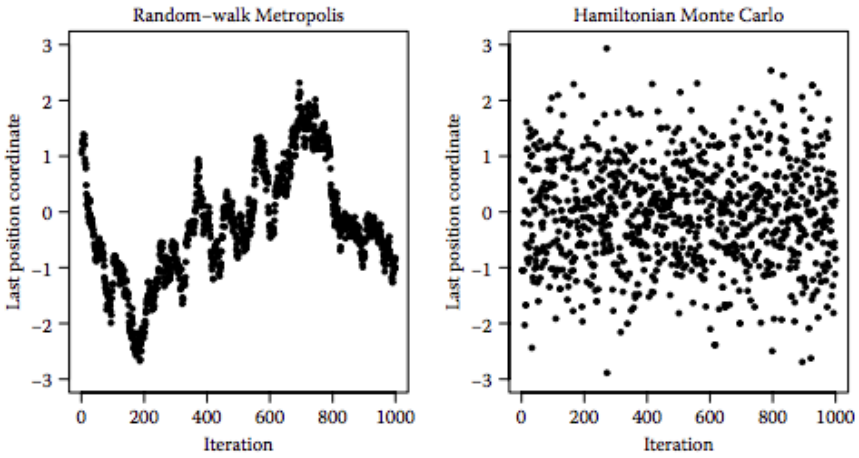

Figure 3: Example sample trace where RWMH exhibits highly correlated samples and HMC appears to obtain uncorrelated samples.

Similarly, the Neal investigates how the sample estimators (mean and standard deviation) of the coordinates for the different dimensions differ over the dimensions with different standard deviations. Given true i.i.d samples from the 100 dimensional Gaussian, we would expect the mean estimators to be close to 0 (with a variance that does not change with dimension) and standard deviation estimators that reflect the true standard deviation of the dimension. This is not the case for the RWMH implementation. We see in Figure 4 that the accuracy of the mean estimator is highly dependent on the standard deviation of the dimension that is being drawn from. For the same reasons as those presented above, if the sampling algorithm is unable to draw independent samples from the distribution as the standard deviation increases, the resulting estimators are expected to have a large bias associated with them. This is seen as the standard deviation of the dimension increases, the variance of the mean estimators increases drastically. Note that the poor results displayed here are due to the inability of RWMH to draw a useful number of *effective samples* from the high dimensional distribution. We see a similar theme for the sample standard deviation estimators

whose variance (around the true value) increases with the dimensionality that is being approximated. Note that the variance of the estimators appears to be relatively invariant to an increase in dimension for the HMC sampler. This is highly desirable as we are able to draw a useful number of effective samples from the distribution even with a high number of dimensions with a vastly varied standard deviation among dimensions.

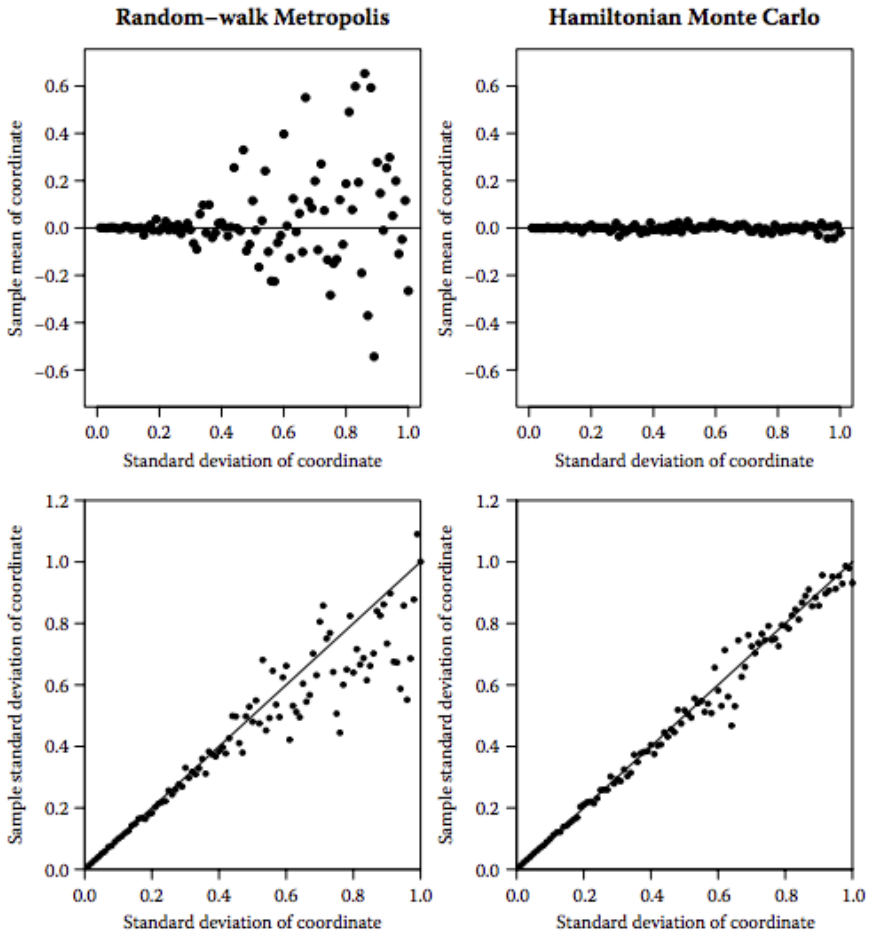

Figure 4: Comparison of true and sampled estimators from RWMH and HMC.

Together these two figures present the main reason for using HMC over the random walk metropolis algorithms in that the authors have tackled the issue that the metropolis algorithms are highly susceptible to high dimensional distributions with highly varied standard deviations.

References

MacKay, D. J. (1998). Introduction to monte carlo methods. In *Learning in graphical models*, pages 175–204. Springer.

Neal, R. M. et al. (2011). Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11).