STATS 700: Bayesian Inference and Computation **Fall 2017**

Lecture 8: November 1 & 6: Hamiltonian Monte Carlo Sampling

Lecturer: Yang Cheng *Scribes: Wenshi Chen, Yumeng Ma*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

This lecture's notes illustrate some uses of various LaTeX macros. Take a look at this and imitate.

8.1 Do chains congregate?

We need to check whether a chain converge to itself and whether chains generated by the same method congregate. If a chain has different mean for its two subchain, the target distribution may have two or more peaks and the chain goes from one to another. If chains with different initial points don't congregate, there also might be two or more peaks in the target distribution. And our estimation with only one chain is not accurate.

For the left picture in Figure 8.1, we can cut a Markov chain into three pieces. Then compare the mean of the first subchain with the mean of the third subchain. The test is described in R package "coda"'s documentation and "geweke.diag" can run the test for us.

The right picture in Figure 8.1 is an not converging case. Besides drawing pictures for different chains, we can calculate within-chain variance and between-chain variance, which is mentioned in "MCMC.html" Yang posted in Canvas. If within-chain variance is significantly greater than between-chain variance, the two chains are congregating. If between-chain variance is larger than within-chain variance, the two chains are not congregating.

8.2 Hamiltonian Monte Carlo Sampling: Motivation

Monte Carlo Sampling algorithms covered in previous lectures are related to random walk to various extents. We desire for a more heuristic way to sample the distribution. Ideally, we expect the sampler to move faster at low-probability region and towards high-probability region. Such a sampling system is a close analogy to a Hamiltonian system, with probability density analogous to potential energy and sampler update analogous to momentum. Similarly, we can derive a Monte Carlo sampling algorithm based on Hamiltonian mechanics of a particle.

8.3 Hamiltonian Monte Carlo Sampling: Theory

A Hamiltonian system is described by a function of momentum p and position q, known as the Hamiltonian $H(p, q) = K(p)+U(q)$, where $K(p)$ and $U(q)$ represents "kinetic" and "potential" energy respectively. Based

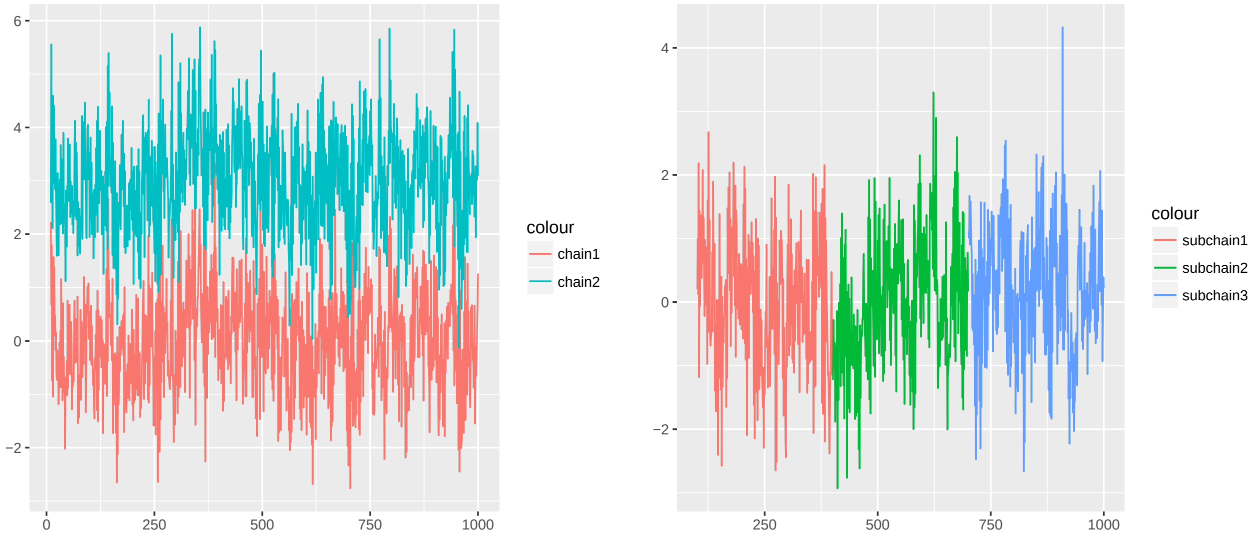

Figure 8.1: Sampling Path

on theories in statistical physics, the canonical distribution of state (p, q) has the joint density of

$$P(p,q) = \frac{1}{Z}\exp\left(\frac{-H(p,q)}{T}\right) = \frac{1}{Z}\exp\left(\frac{-U(q)}{T}\right)\exp\left(\frac{-K(p)}{T}\right). \tag{8.1}$$

Here Z is normalizing constant and T is the "temperature" of the system. We can express the posterior distribution using potential energy function

$$U(q) = -\log\left[\pi(q)L(q|D)\right] \tag{8.2}$$

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

Each step of HMC sampling has two steps. In the first step, new values for the momentum variables are randomly drawn from their Gaussian distribution, independently of the current values of the position variables. For the kinetic energy of Equation 5.23, the d momentum variables are independent, with p_i having zero mean and variance m_i. Since q is not changed, and p is drawn from its correct conditional distribution given q (the same as its marginal distribution, due to independence), this step obviously leaves the canonical joint distribution invariant.

8.4 Hamiltonian Monte Carlo Sampling: Numerical Scheme

The evolution of Hamiltonian is governed by Hamiltonian equations

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i}, \frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i}. \tag{8.3}$$

Euler's method is a commonly used numerical scheme for a system of differential equations. Applying Euler's method to the Hamiltonian equations, we obtain:

$$p_i(t+\epsilon) = p_i(t) + \epsilon\frac{dp_i}{dt} = p_i(t) - \epsilon\frac{\partial U}{\partial q_i}(q(t)), \tag{8.4}$$

$$q_i(t+\epsilon) = q_i(t) + \epsilon\frac{dq_i}{dt}(t) = q_i(t) + \epsilon\frac{p_i(t)}{m_i}. \tag{8.5}$$

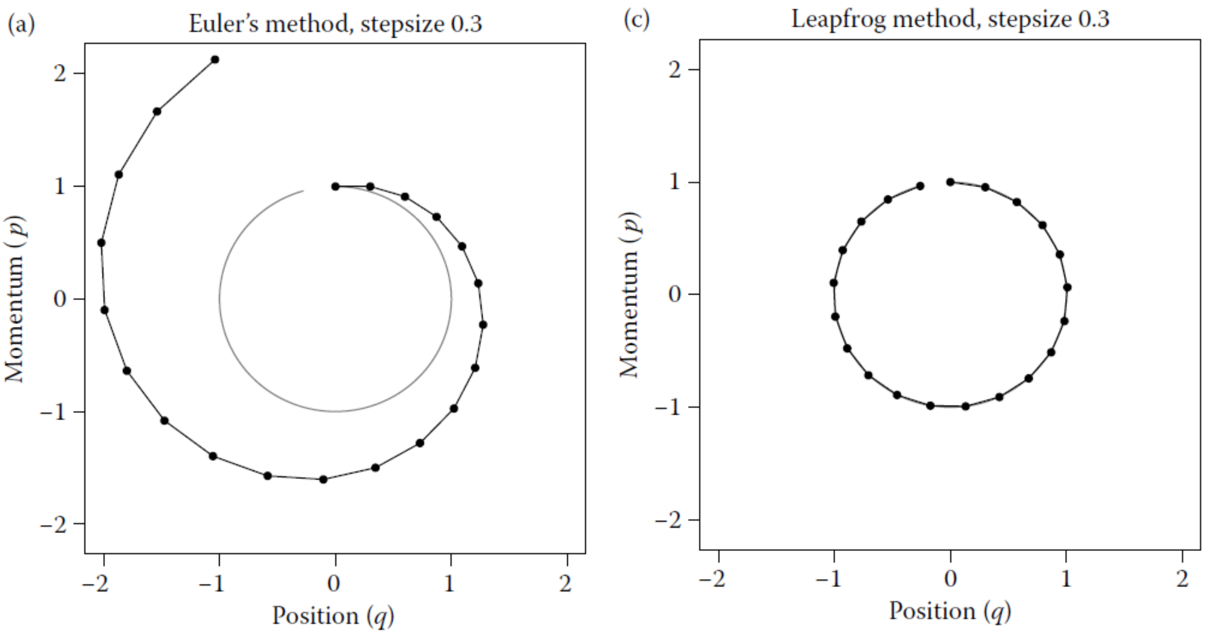

Figure 8.2: Trajectory of Hamiltonian Mechanics by Two Methods

where ϵ is the time step size. However, Hamiltonian is not conserved under Euler's method. The left plot in Figure 8.2 shows the trajectory of a Hamiltonian system $H(p, q) = \frac{p^2}{2} + \frac{q^2}{2}$ with initial condition $p = 1, q = 0$ and time step size $\epsilon = 0.3$. Trajectory in the state space of (p, q) simulated from Euler's method diverges quickly from the true trajectory that is a circle. Leapfrog method is preferred in this case for its higher accuracy and conservation of Hamiltonian. The Leapfrog method solves the Hamiltonian equations as follows:

$$p_i(t + \epsilon/2) = p_i(t) - (\epsilon/2)\frac{\partial U}{\partial q_i}(q(t)), \tag{8.6}$$

$$q_i(t + \epsilon) = q_i(t) + \epsilon\frac{p_i(t + \epsilon 2)}{m_i}, \tag{8.7}$$

$$p_i(t + \epsilon) = p_i(t + \epsilon/2) - (\epsilon/2)\frac{\partial U}{\partial q_i}(q(t + \epsilon)). \tag{8.8}$$

Combining the Leapfrog method with Hamiltonian analogy, we obtain the Hamiltonian Monte Carlo algorithm updating $(\mathbf{p}^{\text{current}}, \mathbf{q}^{\text{current}})$ to $(\mathbf{p}^{\text{next}}, \mathbf{q}^{\text{nextt}})$ as follows:

```
$n \leftarrow$ Dimensionality of the Target Distribution
$L \leftarrow$ Size of Discretization of the Trajectory
$\mathbf{q} \leftarrow \mathbf{q}^{\text{current}}$
for $i \in \{1, \ldots, n\}$ do
    Sample $p_i$ from $N(0, 1)$
end for
$\mathbf{p} \leftarrow \mathbf{p} - \epsilon\frac{\nabla U(\mathbf{q})}{2}$
$L \leftarrow$ Maximum Number of Sampling
for $i \in \{1, \ldots, L\}$ do
    $\mathbf{q} \leftarrow \mathbf{q} + \epsilon\mathbf{p}$
    if $i \neq L$ then
```

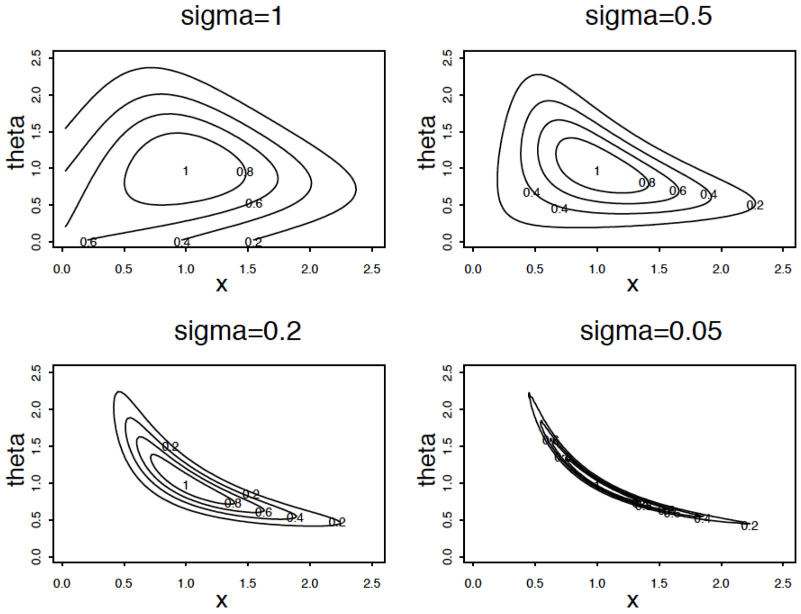

Figure 8.3: Target Distribution under Different σ

```
        $\mathbf{p} \leftarrow \mathbf{p} - \epsilon \nabla U(\mathbf{q})$
      end if
    end for
    $\mathbf{p} \leftarrow \mathbf{p} - \frac{\epsilon}{2} \nabla U(\mathbf{q})$
    $\mathbf{p} = -\mathbf{p}$
    $U^{\text{current}} \leftarrow U(\mathbf{q})$
    $K^{\text{current}} \leftarrow \frac{1}{2}\|\mathbf{p}^{\text{current}}\|_2^2$
    $U^{\text{proposed}} \leftarrow U(\mathbf{q})$
    $K^{\text{proposed}} \leftarrow \frac{1}{2}\|\mathbf{p}\|_2^2$
    Accept $\mathbf{q}$ as $\mathbf{q}^{\text{next}}$ with probability $\min(1, \exp(U^{\text{current}} - U^{\text{proposed}} + K^{\text{current}} - K^{\text{proposed}})$
```

The "potential energy" function $U(\mathbf{q})$ is related to the target distribution $p(\mathbf{x}|y)$ by $p(\mathbf{x}|y) = \exp(-U(\mathbf{x}))$. The second loop in the algorithm is essentially performing integral over the contour of $H(\mathbf{p}, \mathbf{q}) = \text{constant}$.

8.5 Hamiltonian Monte Carlo Sampling: Examples

To demonstrate the power of HMC, we devise a specialized distribution as follows. The observation (x, y) are sampled as $Y \sim N(\theta x, \sigma^2), x \sim N(\theta, 1)$. As a result, the log-posterior distribution has

$$\pi_\sigma(x, \theta) \propto \exp\left(-\frac{(y - x\theta)^2}{2\sigma^2} - \frac{(x-\theta)^2}{2}\right)\pi_0(\theta), \tag{8.9}$$

where $\pi_0(\theta)$ is the prior distribution. When σ is large, the high-probability region is huge and convex. However, when σ gets smaller, the shape of high-probability region will resemble crescent moon, which is difficult to sample from with conventional MC algorithms. Figure 8.3 compare the target distribution under different values of σ.

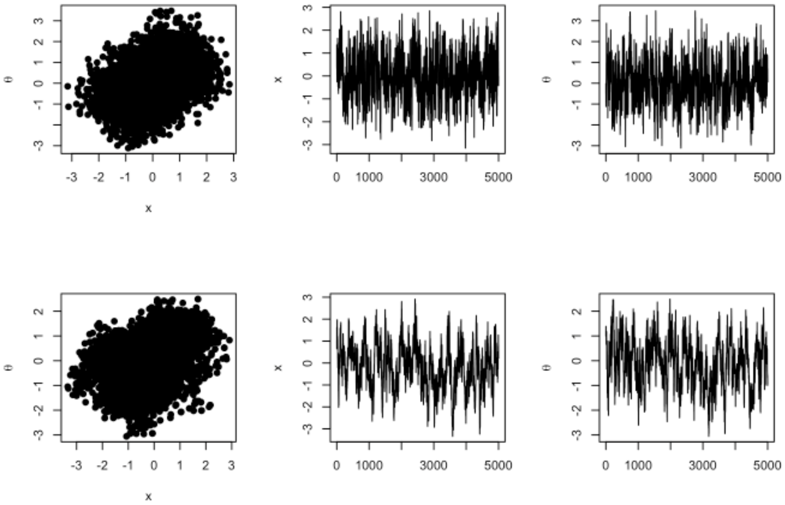

Figure 8.4: Sampling Results by HMC and MHMC

Now we perform Hamiltonian MC and Metropolis-Hastings MC on this distribution respectively. The maximum number of samples is 5000. Figure 8.4 shows the results of both methods. The first row shows the result of HMC and the second row shows the MHMC. Comparing the sampling paths of both methods, MHC converges significantly faster than MHMC. After roughly 500 steps, MHC method converges to the target distribution, while in the case of MHMC, the convergence is not reached even after 3000 steps. Hence, this example successfully demonstrate the superior performance of HMC in convergence rate when compared to MHMC.

References

[1] R. Neal, "Chapter 5: MCMC Using Hamiltonian Dynamics," *Handbook of Markov Chain Monte Carlo*, pp. 113–162.

[2] Yang Chen., "Special Topics in Applied Statistics," *Canvas Slides*, University of Michigan, 2017.