Introduction to Hamiltonian Monte Carlo methods

Minwoo Chae

Department of Industrial and Management Engineering
Pohang University of Science and Technology

Seminar at Department of Statistics
Seoul National University

Outline

Outline

Introduction

- The Hamiltonian Monte Carlo (HMC) is an MCMC method using the Hamiltonian dynamics.
- It is first introduced by Duane et al. (1987) for lattice field theory simulations of quantum chromodynamics.
- They called it as "Hybrid Monte Carlo".
- In statistical community, Neal (1996) firstly applied HMC to neural network models.
- MacKay (2003) used the term "Hamiltonian Monte Carlo".
- We start with a brief introduction to the Hamiltonian dynamics.

Hamiltonian dynamics

Notations

- $q \in \mathbb{R}^d$: position
- $p \in \mathbb{R}^d$: momentum $(= mv)$
- $U(q)$: potential energy $(= mgh)$
- $K(p)$: kinetic energy $(= |p|^2/(2m))$
- $H(q,p)$: Hamiltonian

$$H(q,p) = U(q) + K(p)$$

Hamilton's equations

- Equations of motion: For $i = 1, \ldots, d$,

$$\dot{q}_i = \frac{dq_i}{dt} = \frac{\partial H}{\partial p_i}$$
$$\dot{p}_i = \frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i}$$

- These equations define a mapping

$$T_s : (q(t), p(t)) \mapsto (q(t+s), p(t+s)).$$

Potential and kinetic energy for HMC

- For HMC, we usually use $H(q,p)$ of the form

$$H(q,p) = U(q) + K(p)$$

with

$$K(p) = \frac{1}{2}p^T M^{-1} p,$$

where M is SPD.

- $U(q)$ is the negative log probability density of interest.

1-dim example

EXAMPLE Consider the Hamiltonian $H(q,p) = U(q) + K(p)$ with

$$U(q) = \frac{q^2}{2}, \quad K(p) = \frac{p^2}{2}.$$

Then,

$$\frac{dq}{dt} = p, \quad \frac{dp}{dt} = -q.$$

The solution is, for some constant r and a,

$$q(t) = r\cos(a+t), \quad p(t) = -r\sin(a+t).$$

Properties of $H(p,q)$: Reversibility

- The Hamiltonian dynamics is reversible in the sense that the map $T_s : (q(t), p(t)) \mapsto (q(t+s), p(t+s))$ has an inverse T_{-s}.
- If $H(p,q) = U(q) + K(p)$ and $K(p) = K(-p)$, the inverse T_{-s} can be obtained by
  1. negating p,
  2. applying T_s, and
  3. negating p again.
- The reversibility will play an important role to prove that HMC updates leave the distribution invariant.

Properties of $H(p,q)$: Invariance

- The dynamics keeps Hamiltonian invariant:

$$\begin{aligned}\frac{dH}{dt} &= \sum_{i=1}^{d}\left[\frac{dq_i}{dt}\frac{\partial H}{\partial q_i} + \frac{dp_i}{dt}\frac{\partial H}{\partial p_i}\right] \\ &= \sum_{i=1}^{d}\left[\frac{\partial H}{\partial p_i}\frac{\partial H}{\partial q_i} - \frac{\partial H}{\partial q_i}\frac{\partial H}{\partial p_i}\right] = 0\end{aligned}$$

Properties of $H(p, q)$: Volume preservation

- (q, p) space is often called the phase space.
- Hamiltonian dynamics preserve volume in phase space (Liouville's theorem).
- Equivalently, the determinant of the $2d \times 2d$ Jacobian matrix of T_s has absolute value one.

Discretization: Euler's method

- Euler's method updates

$$p_i(t+\epsilon) \approx p_i(t) + \epsilon \frac{dp_i}{dt}(t) = p_i(t) - \epsilon \frac{\partial U}{\partial q_i}(q(t))$$
$$q_i(t+\epsilon) \approx q_i(t) + \epsilon \frac{dq_i}{dt}(t) = q_i(t) + \epsilon \frac{\partial K}{\partial p_i}(p(t))$$

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

Discretization: A modified Euler's method

- A modified Euler's method updates

$$
\begin{aligned}
p_i(t+\epsilon) &\approx p_i(t) - \epsilon \frac{\partial U}{\partial q_i}(q(t)) \\
q_i(t+\epsilon) &\approx q_i(t) + \epsilon \frac{\partial K}{\partial p_i}(p(t+\epsilon))
\end{aligned}
$$

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

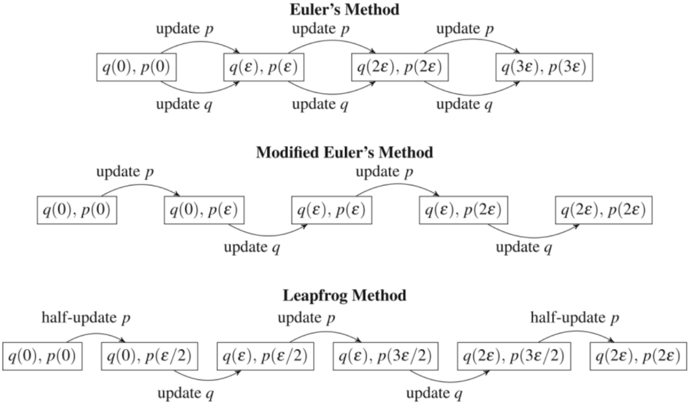
Discretization: The leapfrog method

- The leapfrog method updates

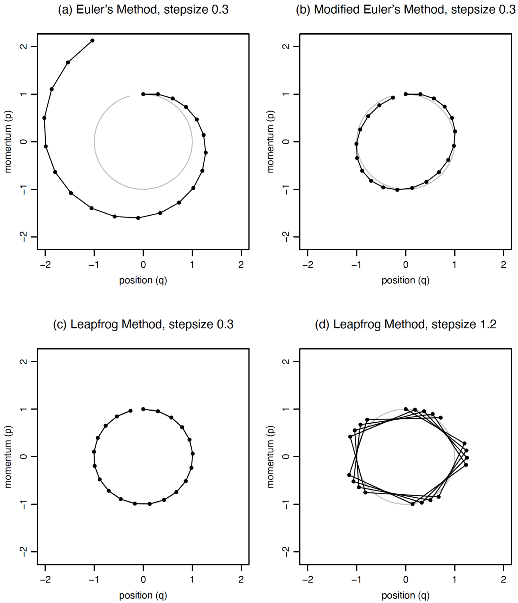
$$
\begin{aligned}
p_i(t+\epsilon/2) &\approx p_i(t) - \frac{\epsilon}{2}\frac{\partial U}{\partial q_i}(q(t)) \\
q_i(t+\epsilon) &\approx q_i(t) + \epsilon\frac{\partial K}{\partial p_i}(p(t+\epsilon/2)) \\
p_i(t+\epsilon) &\approx p_i(t+\epsilon/2) - \frac{\epsilon}{2}\frac{\partial U}{\partial q_i}(q(t+\epsilon))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

  for $i = 1, \ldots, d$.
- L-step leapfrog updates are similar to a modified Euler's method except for the first and last steps.

Discretization: Summary



Discretization: 1-dim illustration



Remark

- In the previous example, if a smaller ϵ is considered for Euler's method, the divergence to infinity is slower, but not eliminated.
- The better performance of modified Euler and leapfrog methods is related to the volume preservation.
- The leapfrog method is reversible by
  1. negating p,
  2. applying the same number of steps again, and
  3. negating p again.
- If ϵ exceeds a certain threshold for the leapfrog method, the trajectory will diverge.

1-dim example

- Recall that $H(q,p) = q^2/2\sigma^2 + p^2/2$.
- A one-step leapfrog update is linear:

$$\begin{bmatrix} q(t+\epsilon) \\ p(t+\epsilon) \end{bmatrix} = \underbrace{\begin{bmatrix} 1-\epsilon^2/2\sigma^2 & \epsilon \\ -\epsilon/\sigma^2 + \epsilon^3/4\sigma^4 & 1-\epsilon^2/2\sigma^2 \end{bmatrix}}_{=A} \begin{bmatrix} q(t) \\ p(t) \end{bmatrix}$$

- If $\epsilon > 2\sigma$, $\lambda_{\max}(A) > 1$ and the trajectory will be unstable.
- If $\epsilon < 2\sigma$, both eigenvalues are complex with absolute value 1, so the the trajectory will be stable.

Outline

Hamiltonian Monte Carlo

- Consider the target density of the form

$$\pi(q) \propto e^{-U(q)}.$$

- We will construct a Markov chain with stationary distribution

$$\pi(q,p) \propto e^{-H(q,p)} = e^{-U(q)-K(p)},$$

where $K(p) = p^T M^{-1} p/2$ for a SPD matrex M.

Algorithm

1. Set tuning parameters (ϵ, L, M).
2. Initialize $q^{(1)}$.
3. For $t = 1, 2, \ldots$
   1. Sample $p^{(t)} \sim N(0, M)$.
   2. Starting from $(q^{(t)}, p^{(t)})$, simulate Hamiltonian dynamics with L-step leapfrog method with step size ϵ to obtain (q^*, p^*).
   3. Negate p^*, that is, $p^* \leftarrow -p^*$ (not necessary in practice).
   4. Accept (q^*, p^*) with probability

      $$\min\left\{1, e^{-H(q^*,p^*)+H(q^{(t)},p^{(t)})}\right\} = \min\left\{1, e^{-U(q^*)+U(q^{(t)})-K(p^*)+K(p^{(t)})}\right\}$$

   5. Set $q^{(t+1)} = q^*$ if accepted, otherwise $q^{(t+1)} = q^{(t)}$.

Remark

- Obviously, updating p leave $\pi(q, p)$ invariant.
- Without this step, $H(q, p)$ will be (nearly) constant.
- It can be shown that the MH step also leaves $\pi(q, p)$ invariant.
- Negation of p^* makes the Metropolis proposal symmetrical.
- If the simulation of Hamiltonian dynamics is exact, the acceptance probability is 1.
- The performance of HMC is sensitive to the choice of (ϵ, L, M).
- One can choose ϵ or L (or both) randomly.

Idea of proof for invariance

$$\pi(q,p) = \frac{1}{C} e^{-H(q,p)}$$

- Let (A_k) be a partition of the phase space with small sets.
- Let K be the transition kernel obtained by
  - operating L leapfrog steps,
  - negating the momentum, and
  - accept/reject the proposal.
- Let B_k be the image of A_k w.r.t. leapfrog updates and negation.
- Then, (B_k) is also a partition of the phase space.

Idea of proof for invariance (cont.)

- Roughly, it suffices to prove the detailed balance:

$$\Pi(A_i)K(B_j \mid A_i) = \Pi(B_j)K(A_i \mid B_j)$$

- Note that $K(B_j \mid A_i) = K(A_i \mid B_j) = 0$ for $i \neq j$.
- Also, $\mathrm{vol}(A_k) = \mathrm{vol}(B_k) \stackrel{\text{def}}{=} V$ and $H \approx$ const. in small regions.
- For $i = j = k$, the above equation reduces

$$\frac{V}{C} e^{-H_{A_k}} \min\left\{1, e^{-H_{B_k}+H_{A_k}}\right\} = \frac{V}{C} e^{-H_{B_k}} \min\left\{1, e^{-H_{A_k}+H_{B_k}}\right\}.$$

Langevin Monte Carlo

- If $L = 1$ in HMC, it is called the Langevin MC (LMC).
  - Firstly proposed in Rossky, Doll and Friedman (1978).
  - Widely spread by Roberts and Stramer (2003).
- For simplicity, suppose that $K(p) = p^T p/2$.
- Then, the one-step HMC proposal (q^*, p^*) is given as

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

with the acceptance probability

$$
\min\left\{1, \exp\left[-\{U(q^*) - U(q)\} - \frac{1}{2}\sum_{i=1}^{d}\left\{(p_i^*)^2 - p_i^2\right\}\right]\right\}.
$$

Langevin Monte Carlo (cont.)

- One can derive the LMC without explicit mention of momentum variables, by performing an MH with the proposal

$$q_i^* \mid q_i \sim N\left(q_i - \frac{\epsilon^2}{2}\frac{\partial U}{\partial q_i}(q), \epsilon^2\right).$$

- In this case, the MH acceptance probability is

$$\min\left\{1, \prod_{i=1}^{d} \frac{\exp[-\{q_i - q_i^* + (\epsilon^2/2)[\partial U/\partial q_i](q^*)\}^2/2\epsilon^2]}{\exp[-\{q_i^* - q_i + (\epsilon^2/2)[\partial U/\partial q_i](q)\}^2/2\epsilon^2]}\right\}$$

  which is the same to that from the one-step HMC.
- Note that the LMC returns a reversible MC.

Illustration: 2-dim example 1

$$H(q,p) = \frac{1}{2}q^T\Sigma^{-1}q + \frac{1}{2}p^Tp, \text{ with } \Sigma = \begin{bmatrix} 1 & 0.95 \\ 0.95 & 1 \end{bmatrix}$$

- Leapfrog updates with $\epsilon = 0.25$ and $L = 25$:

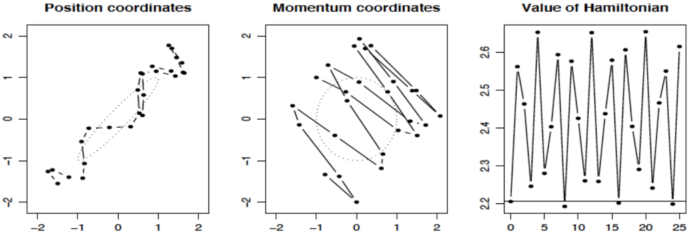


Illustration: 2-dim example 2

$$H(q,p) = \frac{1}{2}q^T\Sigma^{-1}q + \frac{1}{2}p^Tp, \text{ with } \Sigma = \begin{bmatrix} 1 & 0.98 \\ 0.98 & 1 \end{bmatrix}$$

- Comparison of HMC and random walk MH:
  - HMC with $\epsilon = 0.18$ and $L = 20$
    - Rejection rate: 0.09
  - Random walk with Gaussian proposal with $\sigma = 0.18$ and $\rho = 0$
    - Every 20th state from 400 iterations are recorded.
    - Rejection rate: 0.037

Illustration: 2-dim example 2 (cont.)

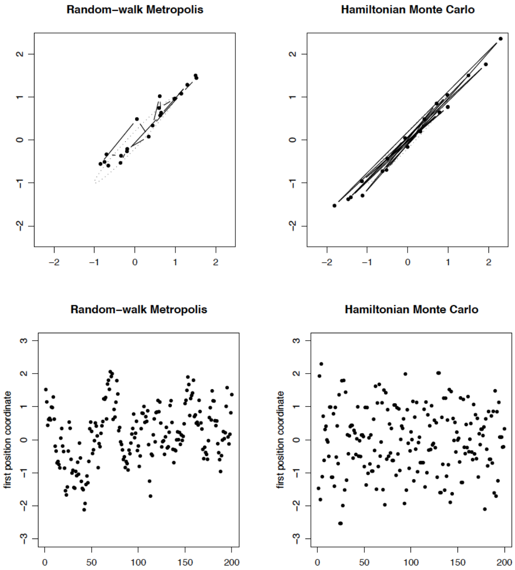


Illustration: 100-dim example

- $U(q) = q^T \Sigma^{-1} q/2$ with

$$\Sigma^{1/2} = \mathrm{diag}(0.01, 0.02, \ldots, 1.00).$$

- $K(p) = p^T p/2$.
- The leapfrog updates operate independently for each (q_i, p_i).
- The acceptance probability depends on the total error in the Hamiltonian.
- $\epsilon \asymp 0.01$ is required to keep this error small.

Illustration: 100-dim example (cont.)

- Comparison:
  - HMC
    - $L = 150$
    - $\epsilon \sim \text{Unif}(0.013 \pm 20\%)$
    - Rejection rate: 0.13
  - Random walk MH
    - Independent Gaussian proposal
    - $\sigma = 0.022 \pm 20\%$
    - 150 updates as one iteration
    - Rejection rate: 0.75
- Nearly optimal settings for both.
- Randomization of ϵ is necessary for avoiding
  - periodicity, and
  - danger caused by different stability limits.

Illustration: 100-dim example (cont.)

Trace plots for the last component

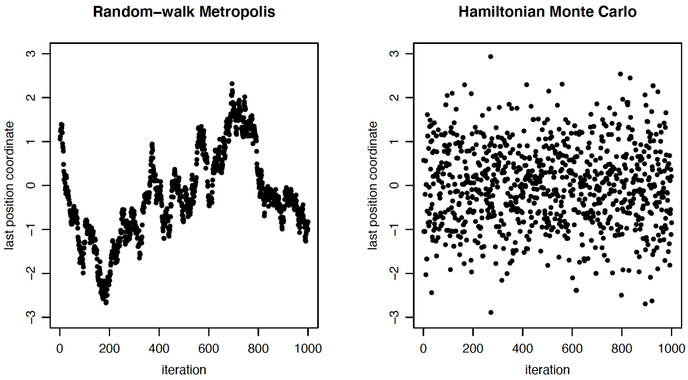
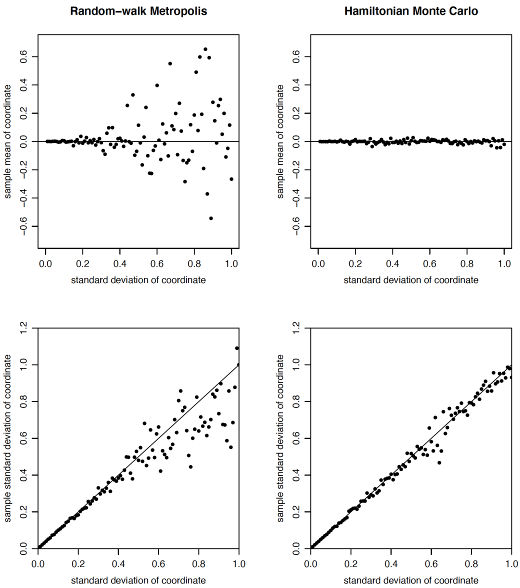


Illustration: 100-dim example (cont.)



Effect of linear transformation

- Recall that the performance of a Gibbs sampler can be significantly improved by a linear transformation.
- For $A \in \mathbb{R}^{d\times d}$, consider the Hamiltonians

$$H(q,p) = U(q) + \frac{1}{2}p^T M^{-1} p$$
$$H'(q',p') = U'(q') + K'(p'),$$

where $q' = Aq$, $p' = (A^T)^{-1}p$,

$$U'(q') = U(A^{-1}q'), \quad K'(p') = \frac{1}{2}(p')^T(M')^{-1}p'$$

and $M' = (AM^{-1}A^T)^{-1}$.

Effect of linear transformation (cont.)

- The dynamics based on H' satisfies

$$\frac{dq}{dt} = M^{-1}p \quad \text{and} \quad \frac{dp}{dt} = -\nabla U(q).$$

- As a consequence, HMCs based on H and H' are the same.
- Practical tips when $\Sigma = \mathrm{Var}(q)$ is known:
  - One may consider HMC with $q' = \Sigma^{-1/2}q$ and $K(p) = p^T p/2$.
  - Equivalently, one may consider HMC with q and $K(p) = p^T \Sigma p/2$.

Remark

- The performance of HMC is very sensitive to the choice of (ϵ, L, M).
- Roughly speaking, the computational complexity of HMC (random walk MH, resp.) for moving to a (nearly) independent state scales as $d^{5/4}$ (d^2, resp.) (in a toy example).
- There are several discretization methods of Hamilton's equations that are reversible, volume-preserving and have a higher order of accuracy than the leapfrog method.
- In practice, however, it is difficult to beat the leapfrog method.

Outline

Truncated multivariate normal

(Pakman and Paninski, 2014)

- Consider a truncated normal $\pi(q) \propto e^{-q^T q/2}$ subject to

$$f_j^T q + g_j \geq 0 \quad j = 1, \ldots, m.$$

- Standard sampling technique is a Gibbs sampler.
  (Geweke, 1991; Kotecha and Djuric, 1999)
- The above methods boils down to sampling from 1-dim truncated normal. (Robert, 1995; Damien and Walker, 2001)
- The performance of Gibbs sampler is poor when the constrained space is highly correlated.
- A state-of-the-art method relies on exact HMC.
  (Pakman and Paninski, 2014)

Truncated multivariate normal (cont.)

(Pakman and Paninski, 2014)

- Let

$$H(q,p) = \frac{1}{2}q^T q + \frac{1}{2}p^T p.$$

- Since the target is Gaussian, the solution of Hamilton's eq. (without constraints) can be obtained exactly:

$$q_i(t) = a_i \sin(t) + b_i \cos(t)$$
$$a_i = p_i(0), \quad b_i = q_i(0)$$

- The constraint can be regarded as a wall with infinite potential energy.
- Once the particle hits a wall, it will bounce off the wall and the trajectory continues with a reflected velocity.

Truncated multivariate normal (cont.)

(Pakman and Paninski, 2014)

- The hitting time t_h can be calculated with elementary algebra.
- Suppose that the particle hits the hth wall, that is,

$$f_h^T q(t) + g_j = 0.$$

- Decompose the velocity as

$$\dot{q}(t_h) = \dot{q}_\perp(t_h) + \alpha_h f_h,$$

where

$$\alpha_h = \frac{f_h^T \dot{q}(t_h)}{\|f_h\|^2}.$$

Truncated multivariate normal (cont.)

(Pakman and Paninski, 2014)

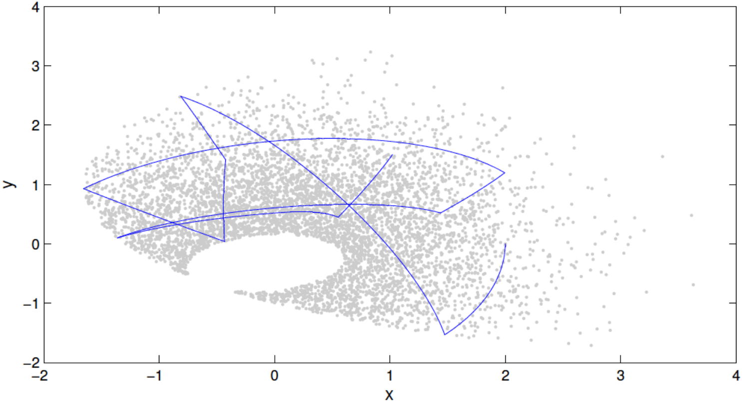
- Then, the reflected velocity is

$$\dot{q}_R(t_h) = \dot{q}_\perp(t_h) - \alpha_h f_h.$$

- This reflection leaves the Hamiltonian invariant.
- The reflected velocity can be used as an initial condition to continue the Hamiltonian dynamics.
- It is only required to determine the travel time.
- $T = \pi/2$ works well in practice.

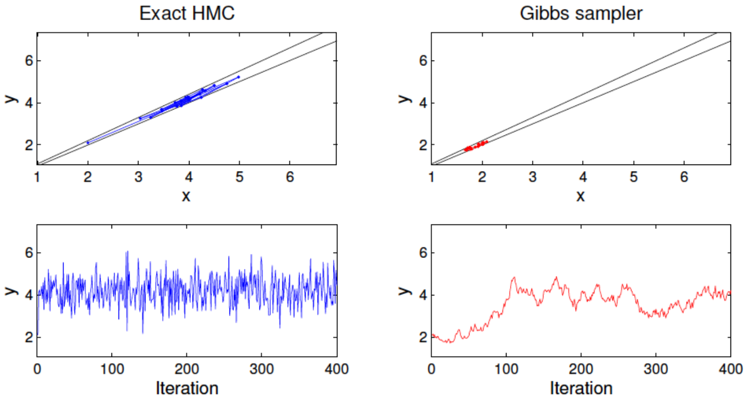
Truncated multivariate normal (cont.)

(Pakman and Paninski, 2014)



Truncated multivariate normal (cont.)

(Pakman and Paninski, 2014)



NUTS: No U-turn sampler

(Hoffman and Gelman, 2014)

- Suppose that $K(p) = p^T p/2$ for simplicity.
- If the dynamics is simulated for long enough, running more leapfrog updates would no longer increase the distance between the proposal q^* and the initial q.
- Thus, too large L would be computationally wasteful.
- One may stop the simulation if

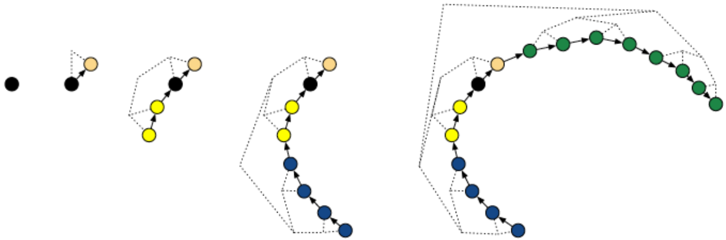
$$\frac{d}{dt}\|q^* - q\|_2^2 = (q^* - q)^T p < 0.$$

- However, this naive stopping rule does not guarantee the convergence to the correct distribution.

NUTS: No U-turn sampler (cont.)

(Hoffman and Gelman, 2014)

- Hoffman and Gelman (2014) developed a “No-U-Turn Sampler” to overcome this issue.



Riemann manifold HMC

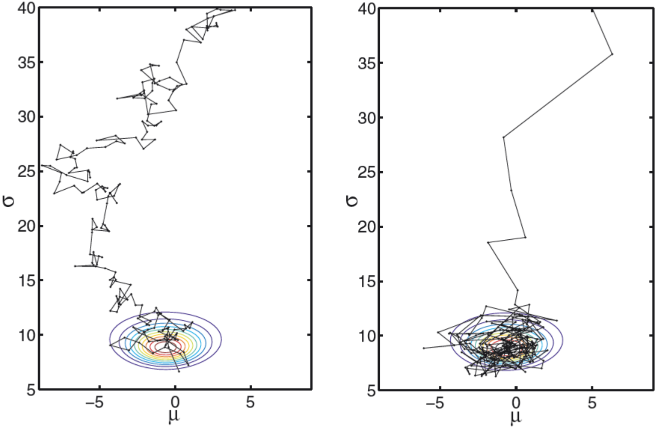
(Girolami and Calderhead, 2011)

- A parametric model $\{f(x \mid q) : q \in \mathcal{Q}\}$ is endowed with a natural Riemann geometry via the Fisher informatin matrix $M(q)$.
- For example, the distance between $N(\mu, \sigma^2)$ and $N(\mu + \delta\mu, \sigma^2 + \delta^2\sigma^2)$ is $(\delta\mu^2 + 2\delta\sigma^2)/\sigma^2$, which decreases as σ^2 increases.
- The Riemann geometry can be utilized in Langevin and Hamiltonian MC methods to determine M.
  - Firstly tried in HMC by Zlochin and Baram (2001).
  - Girolami and Calderhead (2011) developed fundamental methods for RMLMC and RMHMC.

Riemann manifold HMC (cont.)

(Girolami and Calderhead, 2011)

Illustration of Riemann manifold LMC



Riemann manifold HMC (cont.)

(Girolami and Calderhead, 2011)

- Let $M(q)$ be the metric tensor for a given Riemann manifold.
- In a Bayesian framework, one may choose

$$M(q) = -\mathrm{E}_{x|q}\left[\frac{\partial^2}{\partial q^2}\log\{f(x,q)\}\right]$$

  which is the expected Fisher information matrix plus the negative Hessian of the log-prior.

Riemann manifold HMC (cont.)

(Girolami and Calderhead, 2011)

- The Hamiltonian on a Riemann manifold is defined as

$$H(q,p) = U(q) + \frac{1}{2}\log\{(2\pi)^d |M(q)|\} + \frac{1}{2} p^T M(q)^{-1} p.$$

- Note that

$$\int e^{-H(q,p)}\,dp = e^{-U(q)}.$$

- Hamilton's equations:

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i}$$
$$\frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i}$$

Riemann manifold HMC (cont.)

(Girolami and Calderhead, 2011)

- The convergence to the correct distribution is not guaranteed with a naive leapfrog update.
- Generalized leapfrog updates:

$$
\begin{aligned}
p_i(t+\epsilon/2) &\approx p_i(t) - \frac{\epsilon}{2}\frac{\partial H}{\partial q_i}\Big(q(t), p(t+\epsilon/2)\Big) \\
q_i(t+\epsilon) &\approx q_i(t) + \frac{\epsilon}{2}\left\{\frac{\partial H}{\partial p_i}\Big(q(t), p(t+\epsilon/2)\Big)\right. \\
&\qquad \left. + \frac{\partial H}{\partial p_i}\Big(q(t+\epsilon), p(t+\epsilon/2)\Big)\right\} \\
p_i(t+\epsilon) &\approx p_i(t+\epsilon/2) - \frac{\epsilon}{2}\frac{\partial H}{\partial q_i}\Big(q(t+\epsilon), p(t+\epsilon/2)\Big)
\end{aligned}
$$

Riemann manifold HMC (cont.)

(Girolami and Calderhead, 2011)

Algorithm

1. Set tuning parameters (ϵ, L).
2. Initialize $q^{(1)}$.
3. For $t = 1, 2, \ldots$
    1. Sample $p^{(t)} \sim N(0, M(q^{(t)}))$.
    2. Starting from $(q^{(t)}, p^{(t)})$, run the generalized leapfrog steps with parameters (ϵ, L) to obtain (q^*, p^*).
    3. Accept (q^*, p^*) with probability

$$\min\left\{1, e^{-H(q^*,p^*)+H(q^{(t)},p^{(t)})}\right\}$$

    4. Set $q^{(t+1)} = q^*$ if accepted, otherwise $q^{(t+1)} = q^{(t)}$.

Riemann manifold HMC (cont.)

(Girolami and Calderhead, 2011)

- The previous updates are the same to the leapfrog method if $H(q,p) = U(q) + K(p)$.
- Note that updates of $p(t+\epsilon/2)$ and $q(t+\epsilon)$ are defined implicitly.
- In many examples, these implicit equations can be solved by 5-6 fixed point iterations.

Stochastic gradient HMC

(Chen, Fox and Guestrin, 2014)

- Suppose that $K(p) = p^T M^{-1} p/2$ and

$$U(q) = -\log \pi(q) - \sum_{i=1}^{n} \log f(x_i \mid q).$$

- In examples with huge data, $\nabla U(q)$ is expensive to compute.
- At each step of HMC, one may approximate $\nabla U(q)$ as

$$\nabla \widetilde{U}(q) = -\nabla \log \pi(q) - \frac{n}{|I|} \sum_{i \in I} \nabla \log f(x_i \mid q)$$

  with a minibatch I.
- However, this approximation may change the limiting distribution significantly.

Stochastic gradient HMC (cont.)

(Chen, Fox and Guestrin, 2014)

- Roughly, suppose that

$$\nabla \widetilde{U}(q) \approx \nabla U(q) + N(0, V(q)).$$

- The resulting ϵ-discretization of p is

$$p(t+\epsilon) \approx p(t) - \epsilon \nabla U(q) + N(0, \epsilon^2 V(q))$$

- This can be viewed as a discretization of

$$\begin{aligned} dq &= M^{-1} p dt \\ dp &= -\nabla U(q) dt + B dW_t \end{aligned}$$

for some $B = B(q)$, where W_t is the standard Brownian motion.

Stochastic gradient HMC (cont.)

(Chen, Fox and Guestrin, 2014)

- Physically, the additional term $B(q)dW_t$ can be regarded as random wind.
- Chen, Fox and Guestrin (2014) proved that the Hamiltonian is not invariant under the above dynamics.
- As an alternative, they introduce the dynamics

$$
\begin{aligned}
dq &= M^{-1} p dt \\
dp &= -\nabla U(q) dt + B dW_t - B M^{-1} p dt
\end{aligned}
$$

  with which the Hamiltonian is invariant.
- Physically, the additional term $BM^{-1}pdt$ can be interpreted as friction.
- In practice, B is unknown and should be estimated.

Thank you for attention!