# Monte Carlo methods in Artificial Intelligence and Machine Learning

## Hybrid Monte Carlo

Chun Yuan, Sameh Bilto, Yang Chen

July 24, 2018

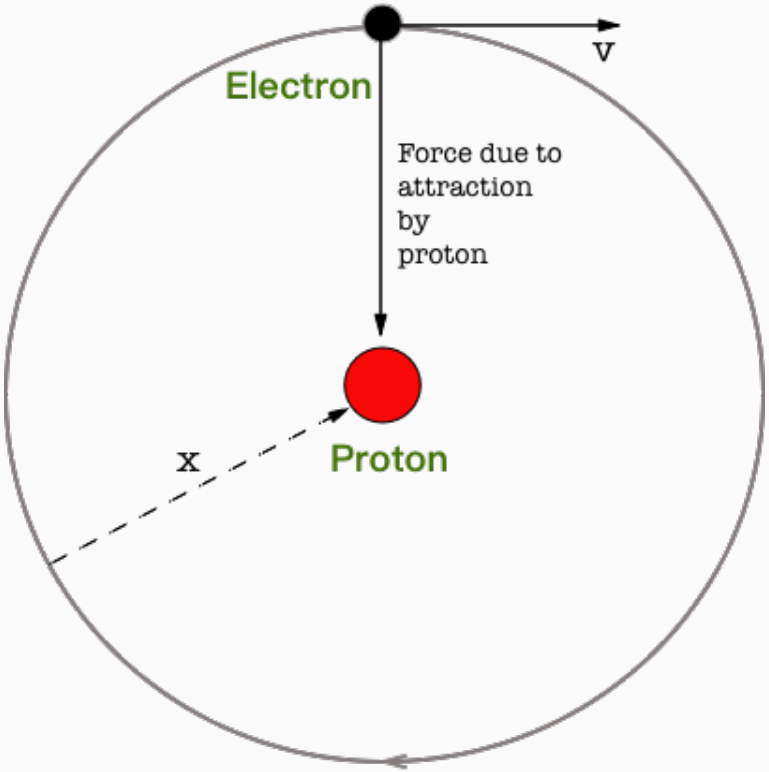
# Outline

1. Background Knowledges
2. Hamiltonian System
3. Hybrid Monte Carlo(HMC)
4. Comparison with Random Walk and Langevin Sampling
5. Conclusion

# Background Knowledges

# Conservation of mechanical energy

For an isolated system, its mechanical energy remains constant in time.[1]

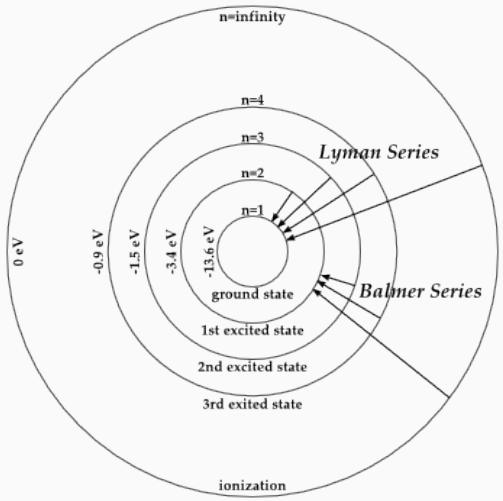


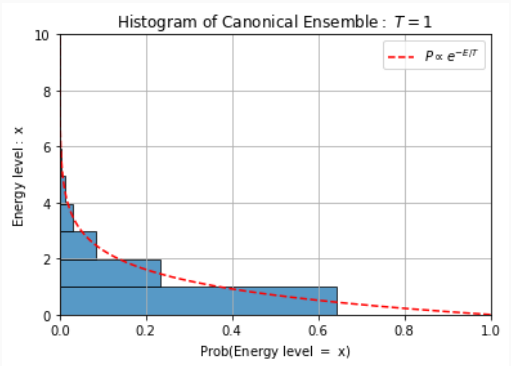

$$E(x, v) = U(x) + K(v) = const.$$

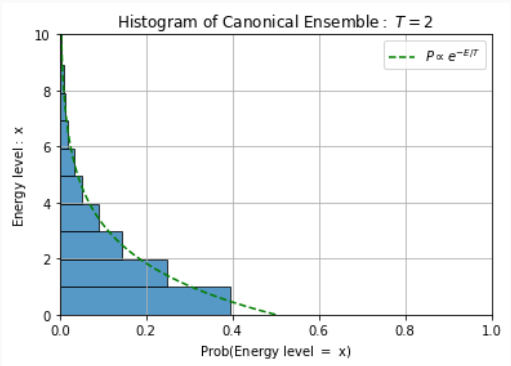
$U(x)$ is potential energy
$K(v)$ is kinetic energy

# Canonical Ensemble

In statistical mechanics, the **canonical ensemble** is the statistical ensemble that represents **the possible states** of a mechanical system at a **fixed** temperature.[2]

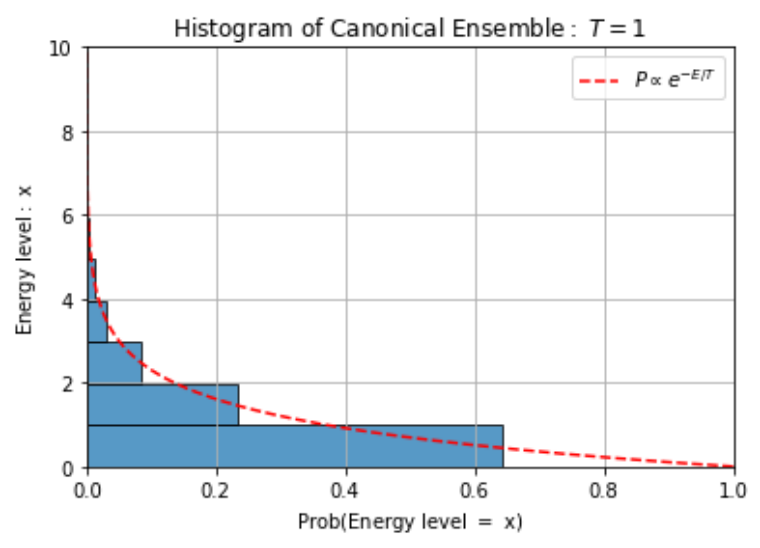


# Canonical Ensemble

The histogram below describes the amount of hydrogen's electrons which stay at different possible energy levels. The canonical ensemble could be defined as:




$$\Pi(x, v) = \frac{1}{Z} e^{-\frac{E(x,v)}{T}} = \frac{1}{Z} e^{-\frac{U(x)+K(v)}{T}}$$

$E$ is electron's mechanical energy

$T$ is surrounding temperature

$Z$ is the normalisation part

# Hamiltonian System

# Background of Hamiltonian System

*A Hamiltonian system is a dynamical system governed by Hamilton's equations. In physics, this dynamical system describes the evolution of a physical system such as a planetary system or an electron in an electromagnetic field.*[3]

Energies in Hamiltonian system also follows the conservation of mechanical energy, and could be represented as canonical ensemble as following:

$$\Pi(x, v) \propto e^{-\frac{U(x)+K(v)}{T}} = const.$$

where, $\underline{x}$ indicates the position information and $\underline{v}$ is the velocity.

# System Dynamics

Random walk sampler utilizes:

$$\underline{x}(t+\epsilon) = \underline{x}(t) + \sqrt{\rho}\underline{z}$$

where $z \sim \mathcal{N}(0, \underline{I})$

Langevin sampling adds some extra information(gradient information) into the dynamics:

$$\underline{x}(t+\epsilon) = \underline{x}(t) + \frac{\epsilon}{2}\nabla \ln p(\underline{x}(t)) + \eta\sqrt{\epsilon}$$

where $\eta \sim \mathcal{N}(0, \underline{I})$

# Hamiltonian Equations

In Hybrid Monte Carlo, in order to sampling from the joint density, we make use of Hamiltonian Equations:

$$\frac{dp}{dt} = -\frac{\partial H}{\partial x}$$

$$\frac{dx}{dt} = +\frac{\partial H}{\partial p}$$

Simple explanation:[4]

- The first Hamilton equation means that the force equals the negative gradient of potential energy.
- The second Hamilton equation means that the particles velocity equals the derivative of its kinetic energy with respect to its momentum.

# Numerical Methods for Hamiltonian Equations

Hamiltonian equations are two differential equations. Therefore, to update our position information $\underline{x}$ and momentum $\underline{p}$, we need methods to compute these differential equation. There exist various algorithm to achieve this:

- Euler's method
- Euler's modified method
- Leapfrog method

# Numerical Methods for Differential Equations

- Euler's Method:

$$\begin{cases} p(t+\epsilon) = p(t) + \epsilon \frac{dp}{dt}(t) = p(t) - \epsilon \frac{\partial H}{\partial x}|_{x(t)} \\ x(t+\epsilon) = x(t) + \epsilon \frac{dx}{dt}(t) = x(t) + \epsilon \frac{\partial H}{\partial p}|_{p(t)} \end{cases}$$

- Euler's Modified Method:

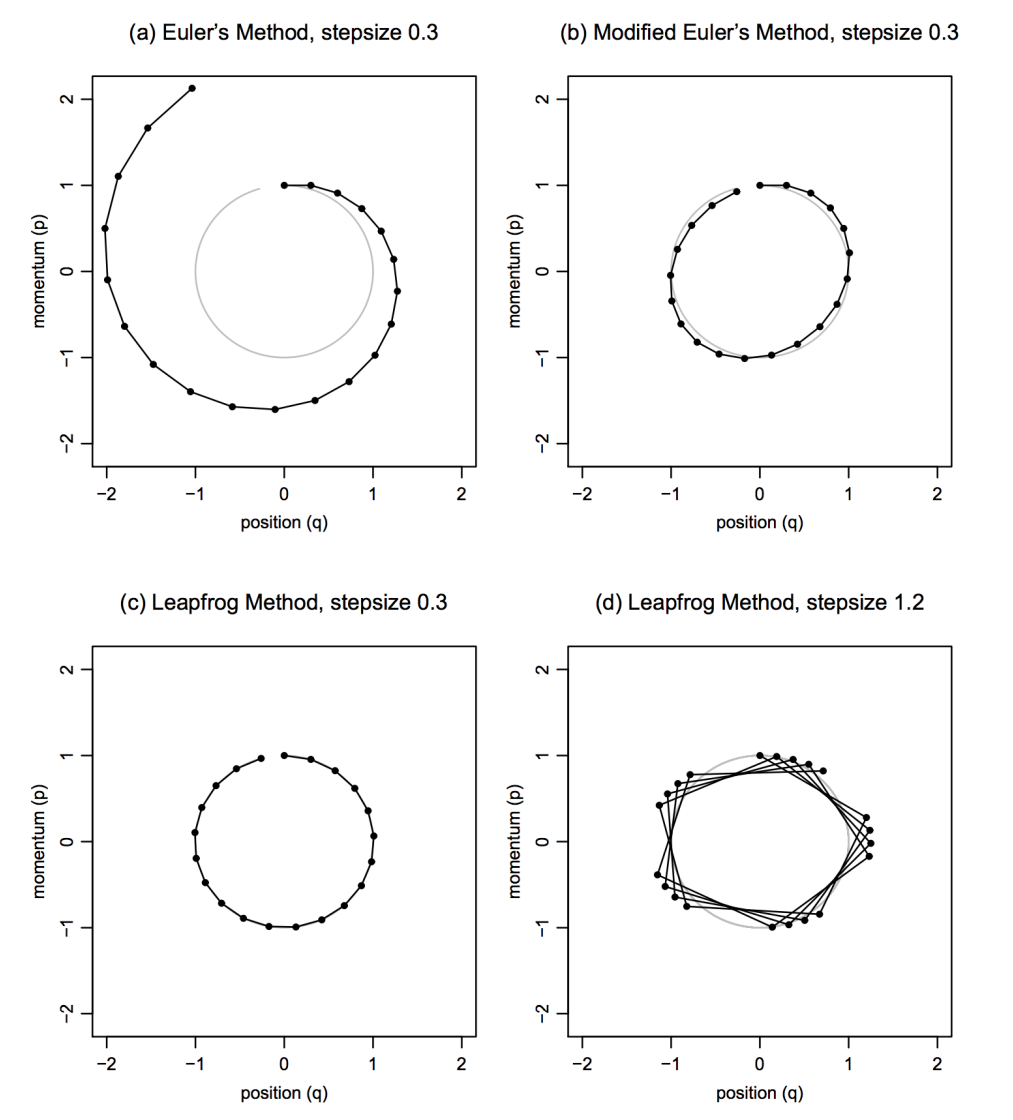
$$\begin{cases} p(t+\epsilon) = p(t) - \epsilon \frac{\partial H}{\partial x}|_{x(t)} \\ x(t+\epsilon) = x(t) + \epsilon p(t+\epsilon) \end{cases}$$

- Leapfrog Method:

$$\begin{cases} p(t+\frac{\epsilon}{2}) = p(t) - \frac{\epsilon}{2} \frac{\partial H}{\partial x}|_{x(t)} \\ x(t+\epsilon) = x(t) + \epsilon p(t+\frac{\epsilon}{2}) \\ p(t+\epsilon) = p(t+\frac{\epsilon}{2}) - \frac{\epsilon}{2} \frac{\partial H}{\partial x}|_{x(t+\epsilon)} \end{cases}$$

# Numerical Methods for Differential Equations

The results using three methods for approximating Hamiltonian dynamics: [5](section 3.2)



# Local and Global Error of Discretization Methods

It's easy to get:

$$\lim_{\epsilon \to 0} \mathcal{E}_{Approx} = 0$$

Therefore, we have to apply upper bound to the error function.

We define:

- The local error is the error after one step;
- The global error is the error after simulating for some fixed time interval $s$.

If the local error is order $\epsilon^a$, in this time interval $s$, frog will jump $s/\epsilon$ steps:

$$O(\mathcal{E}_g) = O(\mathcal{E}_{loc}) * \frac{s}{\epsilon} = O(\epsilon^{a-1})$$

As shown by Neal (2011, section 2.3), The Euler method and its modication above have order $\epsilon^2$ local error and order $\epsilon$ global error. The leapfrog method has order $\epsilon^3$ local error and order $\epsilon^2$ global error.[5]

# So far ...

- The joint distribution of the system energies could be formulated as:

$$\Pi(x, p) = e^{-\frac{U(x)+K(p)}{T}}$$

- System's dynamics:

$$\begin{cases} p(t+\frac{\epsilon}{2}) = p(t) - \frac{\epsilon}{2}\frac{\partial H}{\partial x}|_{x(t)} \\ x(t+\epsilon) = x(t) + \epsilon p(t+\frac{\epsilon}{2}) \\ p(t+\epsilon) = p(t+\frac{\epsilon}{2}) - \frac{\epsilon}{2}\frac{\partial H}{\partial x}|_{x(t+\epsilon)} \end{cases}$$

- our goal is: to obtain samples $x$ from the joint distribution $\underline{\Pi(x, p)}$ regardless of samples $\underline{p}$

# Hybrid Monte Carlo(HMC)

# Basic Assumptions in HMC

Assumed that system's potential energy is $\underline{U(x) = -\log \pi(x)}$ and kinetic energy has the form $\underline{K(p) = \frac{p^2}{2m}}$.

Then, the canonical ensemble for this system is:

$$\Pi(x, p) \propto e^{-\frac{U(x)+K(p)}{T}} \propto \pi(x) e^{-\frac{p^2}{2mT}} \doteq \pi(x)\pi'(p)$$

This indicates that $\underline{x}$ and $\underline{p}$ are **independent** with each other.

# Basic Assumptions in HMC

Since the distribution of momentum could be represented as:

$$\pi'(p) \propto e^{-\frac{p^2}{2mT}}$$

which is similar with a Gaussian density with zero mean and unit variance.

Therefore, we suppose that the distribution of momentum is:

$$\pi'(p) \sim \mathcal{N}(0, 1)$$

# Metropolis-Hastings Algorithm in HMC

The acceptance ratio of Metropolis-Hastings method in MCMC is:

$$A(x'|x) = \min \left\{ \frac{p(x')q(x|x')}{p(x)q(x'|x)} \ , \ 1 \right\}$$

where $q(\cdot)$ is a proposal distribution.

In Hybrid Monte Carlo, the canonical ensemble of system energies is the corresponding proposal distribution: $\Pi(x, p) = \pi(x)\pi'(p)$. Hence, the acceptance ratio for HMC is:

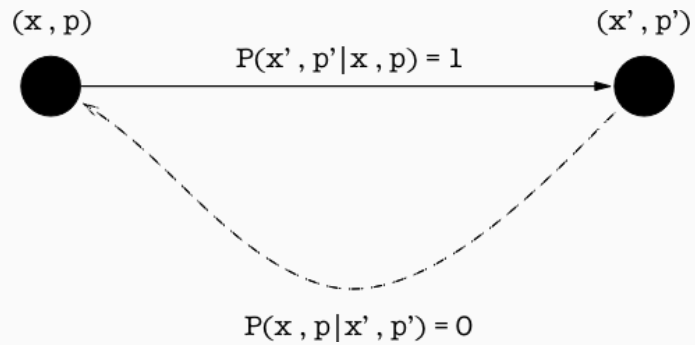
$$A(x', p'|x, p) = \min \left\{ \frac{\Pi(x', p')}{\Pi(x, p)} \frac{p(p, x|p', x')}{p(p', x'|p, x)} \ , \ 1 \right\}$$

# Metropolis-Hastings Algorithm in HMC

- The first term is the ratio of Hamiltonian energy after updating, more specifically, it is just the difference of system energies in two different states:

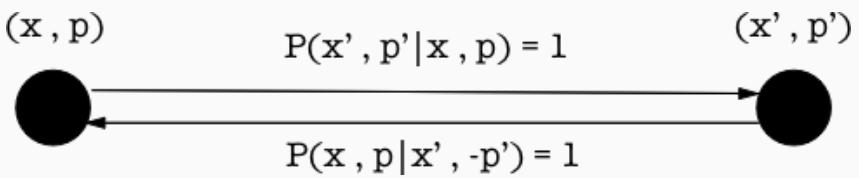
$$\frac{\Pi(x', p')}{\Pi(x, p)} = exp\{-\frac{E_{new} - E_{old}}{T}\}$$

- The second term is about the path of particle. However, it is deterministic, according to our system dynamics.In this case, the acceptance ratio will decrease to $\underline{0}$.



# The Negation of Momentum

Instead of choosing proposal $(x', p')$, we first flip the direction of momentum $p'$ to $-p'$, and select $(x', -p')$ as our proposal.




This operation will lead the second term $\frac{p(p,x|p',x')}{p(p',x'|p,x)} = 1$

*This negation of momentum need not be done in practice, since* $K(p) = K(-p)$.[5]

# Acceptance of new Proposal in HMC

Now, the acceptance ratio for Hybrid Monte Carlo could be simplified as:

$$A(x', p'|x, p) = \min\left\{ \frac{\Pi(x', p')}{\Pi(x, p)} \ , \ 1 \right\} = \min\left\{ exp^{-\frac{E_{new} - E_{old}}{T}} \ , \ 1 \right\}$$

**Algorithm 1:** Accept New Proposal

**input :** Initial State: $(x_t, p_t)$

**output:** New Proposal: $(x_{t+1}, p_{t+1})$

$(x_{t+1}, p_{t+1}) = leapfrog(x_t, p_t, steps, \epsilon)$;

$A(x_{t+1}, p_{t+1}|x_t, p_t) = exp\{\frac{E_{old} - E_{new}}{T}\}$;

**if** $A(x_{t+1}, p_{t+1}|x_t, p_t) >= u \sim U(0, 1)$ **then**

  $s = x_{t+1}$;

**else**

  $s = x_t$;

**end**

# Hybrid Monte Carlo Sampling

There exists two steps in the HMC sampling at each iteration:[5](section 3.2)

- 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.

- In the second step, a update is performed, using Hamiltonian dynamics to propose a new state.

*Note: the momentum will be replaced before it is used again, in the first step of the next iteration. Therefore, we don't have to inverse its direction.*

# Hybrid Monte Carlo Sampling

**Algorithm 2:** Hybrid Monte Carlo Sampling

**input :** The Number of Samples: $N$, Step Size: $\epsilon$, System Temperature: $T$

**output:** Samples: $\underline{X} = (x_1, ..., x_n)$ from $p(x)$

initialization: choose $x_t$ randomly, $t = 1$;
**while** $t < N$ **do**
    each time choose $p_t$ randomly;
    $(x_{t+1}, p_{t+1}) = leapfrog(x_t, p_t, steps, \epsilon)$;
    **if** $A(x_{t+1}, p_{t+1}|x_t, p_t) >= u \sim U(0, 1)$ **then**
        $s = x_{t+1}$;
    **else**
        $s = x_t$;
    **end**
    Appending $s$ to $\underline{X}$;
    t+=1;
**end**

# Comparison with Random Walk and Langevin Sampling

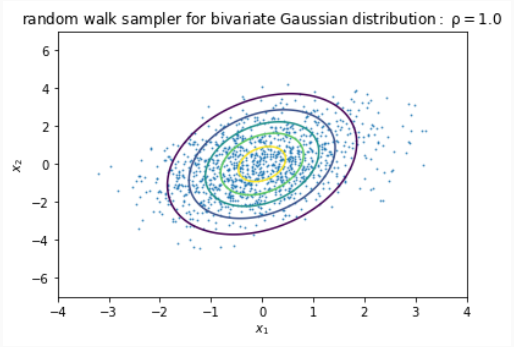
# Bases for the first Experiment

Experiment 1: Given two-dimensional Gaussian distribution with mean vector $\underline{\mu} = (0, 0)^T$ and covariance matrix $\underline{\Sigma} = \begin{pmatrix} 1.0 & 0.6 \\ 0.6 & 2.0 \end{pmatrix}$.

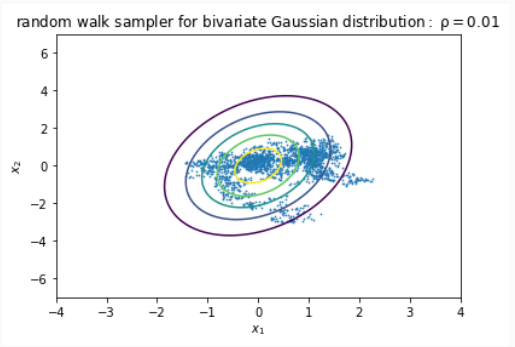
We apply respectively Random Walk, Langevin and Hybrid Monte Carlo sampling methods to sample $N = 2000$ times with different parameters, and calculate the corresponding covariance matrices from samples.

# Random Walk Sampling

The dynamics of Random Walk Sampling is: $\underline{x}' = \underline{x} + \sqrt{\rho}\underline{z}$




**(a)** rho: 1.0




**(b)** rho: 0.01

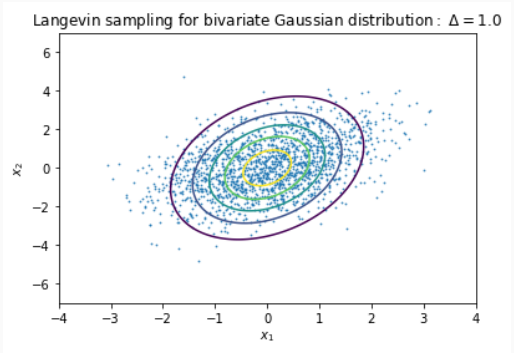
$$\begin{pmatrix} 1.07 & 0.63 \\ 0.63 & 2.22 \end{pmatrix}$$

$$\begin{pmatrix} 0.56 & 0.77 \\ 0.77 & 2.91 \end{pmatrix}$$

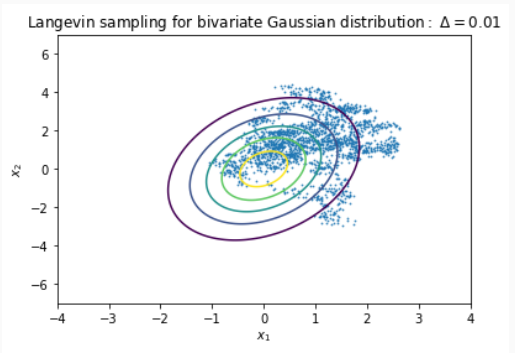
# Langevin Sampling

The dynamics of Langevin Sampling is:

$$\underline{x}(t+\Delta) = \underline{x}(t) + \frac{1}{2}\nabla \ln p(\underline{x}(t))\Delta + \eta\sqrt{\Delta}$$




**(a)** delta: 1.0




**(b)** delta: 0.01

$$\begin{pmatrix} 1.02 & 0.61 \\ 0.61 & 1.99 \end{pmatrix}$$

$$\begin{pmatrix} 1.19 & 0.13 \\ 0.13 & 0.97 \end{pmatrix}$$

# Hybrid Monte Carlo Sampling

The leapfrog algorithm of HMC Sampling is:

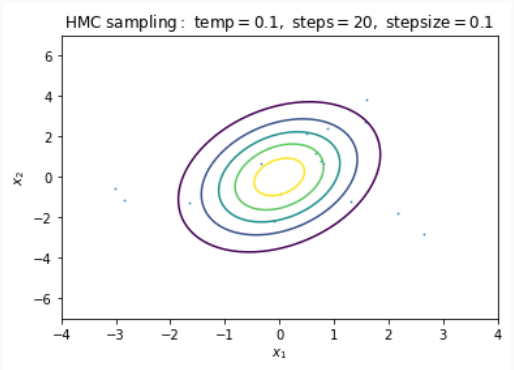
$$\begin{cases} p(t+\frac{\epsilon}{2}) = p(t) - \frac{\epsilon}{2}\frac{\partial H}{\partial x}|_{x(t)} \\ x(t+\epsilon) = x(t) + \epsilon p(t+\frac{\epsilon}{2}) \\ p(t+\epsilon) = p(t+\frac{\epsilon}{2}) - \frac{\epsilon}{2}\frac{\partial H}{\partial x}|_{x(t+\epsilon)} \end{cases}$$

There are several parameters in Hybrid Monte Carlo sampling which we can play with:
The system temperature: $\underline{T}$, the number of jumping step for leap frog: $\underline{N}$, and the size of jumping step: $\underline{\epsilon}$.

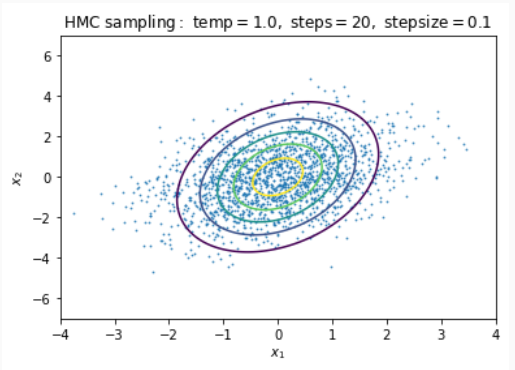
# Hybrid Monte Carlo Sampling

Let the system temperature change only:

$$A(x', p'|x, p) = \min \{ \ exp^{-\frac{E_{new} - E_{old}}{T}} \ , \ 1 \ \}$$




**(a)** temperature: 0.1



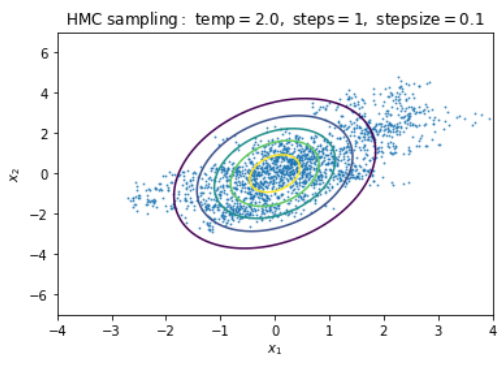

**(b)** temperature: 1

$$\begin{pmatrix} 2.57 & 0.61 \\ 0.61 & 3.51 \end{pmatrix}$$

$$\begin{pmatrix} 1.15 & 0.57 \\ 0.57 & 2.17 \end{pmatrix}$$

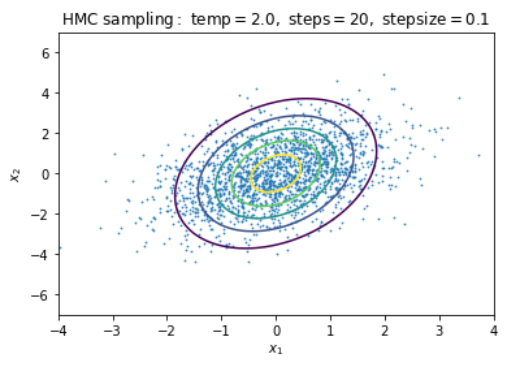
# Hybrid Monte Carlo Sampling

Let the frog's jumping steps change only:



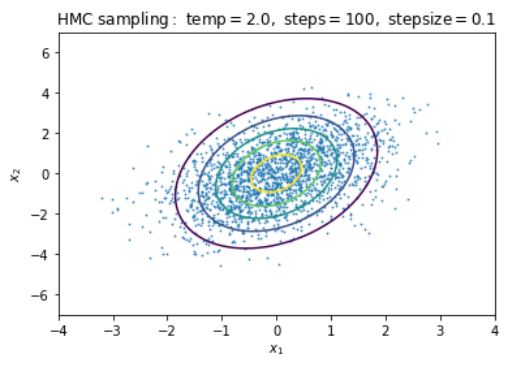

**(a)** steps: 1

$$\begin{pmatrix} 1.46 & 1.34 \\ 1.34 & 2.29 \end{pmatrix}$$




**(b)** steps: 20

$$\begin{pmatrix} 1.15 & 0.57 \\ 0.57 & 2.17 \end{pmatrix}$$

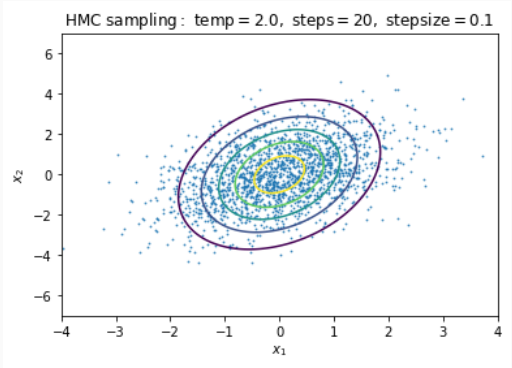



**(c)** steps: 100

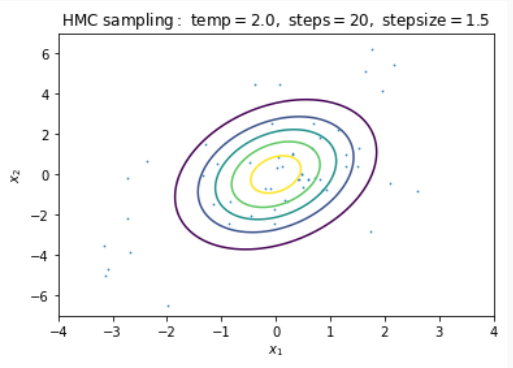
$$\begin{pmatrix} 0.95 & 0.66 \\ 0.66 & 2.20 \end{pmatrix}$$

# Hybrid Monte Carlo Sampling

Let the size of jumping change only:




**(a)** epsilon: 0.1




**(b)** epsilon: 1.5

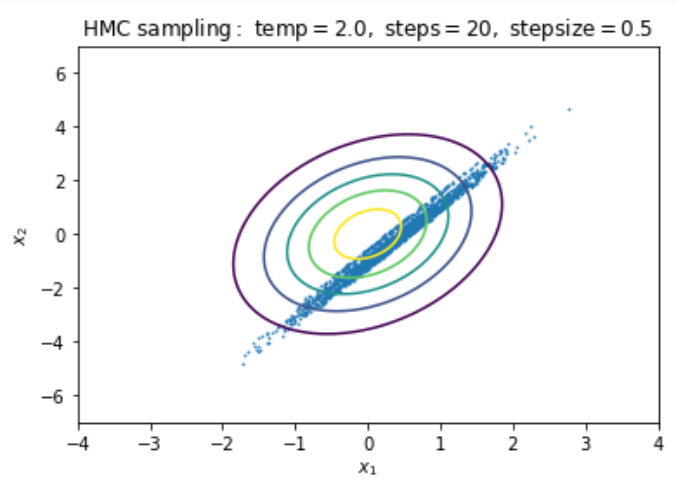
$$\begin{pmatrix} 1.15 & 0.57 \\ 0.57 & 2.17 \end{pmatrix}$$

$$\begin{pmatrix} 3.28 & 3.40 \\ 3.40 & 8.22 \end{pmatrix}$$

# Ergodicity of Hybrid Monte Carlo

Typically, the HMC algorithm will not be trapped in some subset of the state space, and hence will asymptotically converge to its invariant distribution.

However, ergodicity can fail if the L leapfrog steps in a trajectory produce an exact periodicity for some function of state.[5]




Here is an instance we met

This potential problem can be solved by randomly choosing $\epsilon$ or $L$ (or both) from some fairly small interval.[6]

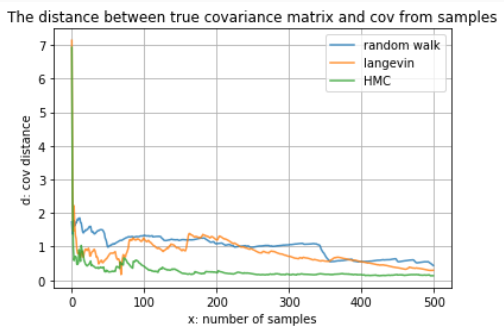
# Bases for the second Experiment

Experiment 2: The same with the first one, given two-dimensional Gaussian distribution with mean vector $\underline{\mu} = (0, 0)^T$ and covariance matrix $\underline{\Sigma} = \begin{pmatrix} 1.0 & 0.6 \\ 0.6 & 2.0 \end{pmatrix}$.

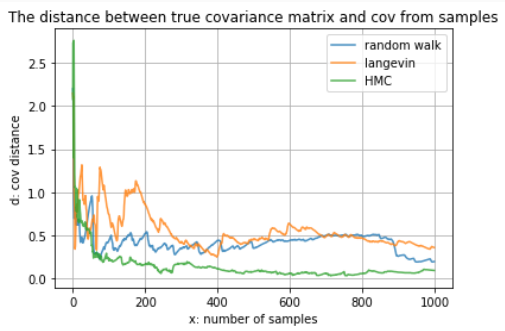
We apply respectively Random Walk, Langevin and Hybrid Monte Carlo sampling methods to obtain the same amount of accepted samples, and calculate the distance between their corresponding covariance matrices and the true one.

# Hybrid Monte Carlo Sampling

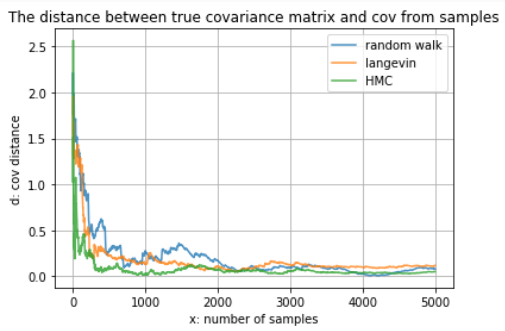
Utilizing Frobenius matrix norm:




**(a)** Accepted: 500




**(b)** Accepted: 1000




**(c)** Accepted: 5000

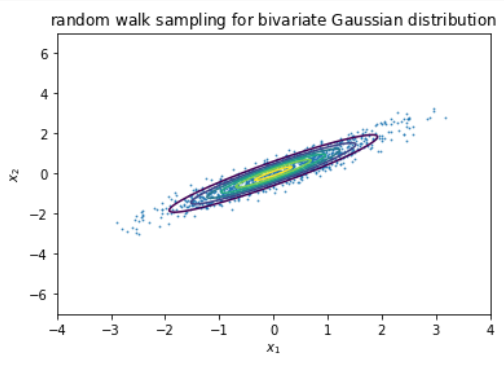
# Bases for the third Experiment

Experiment 3: Given two-dimensional Gaussian distribution with mean vector $\underline{\mu} = (0, 0)^T$ and covariance matrix $\underline{\Sigma} = \begin{pmatrix} 1.0 & 0.95 \\ 0.95 & 1.0 \end{pmatrix}$.

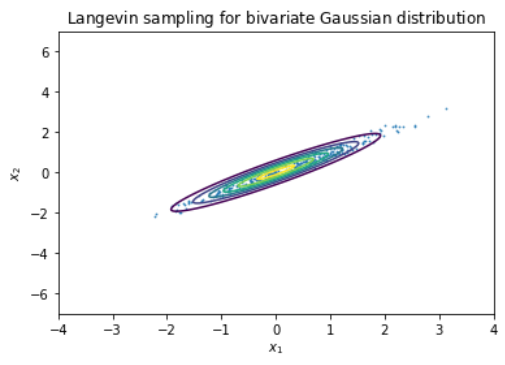
We apply respectively Random Walk, Langevin and Hybrid Monte Carlo methods to sample 5000 times, and compare the amount of accepted samples, in order to prove the high efficiency of HMC.

# Hybrid Monte Carlo Sampling

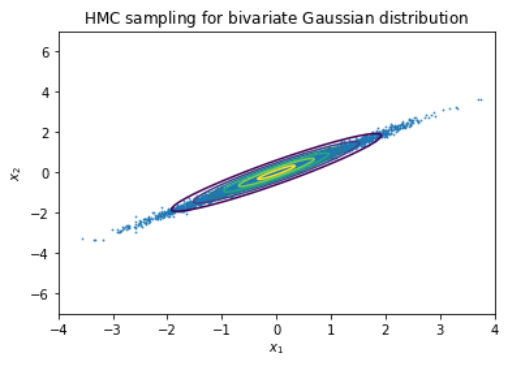
With difference amount of accepted samples:




**(a)** Accepted: 1206




**(b)** Accepted: 326




**(c)** Accepted: 4799

# HMC and Its Benefits

Experiment 3 is a nice example to show the advantages of HMC sampling, comparing with the other two methods.
There is also an illustration shown by Neal (2011, section 3.3)[5]:

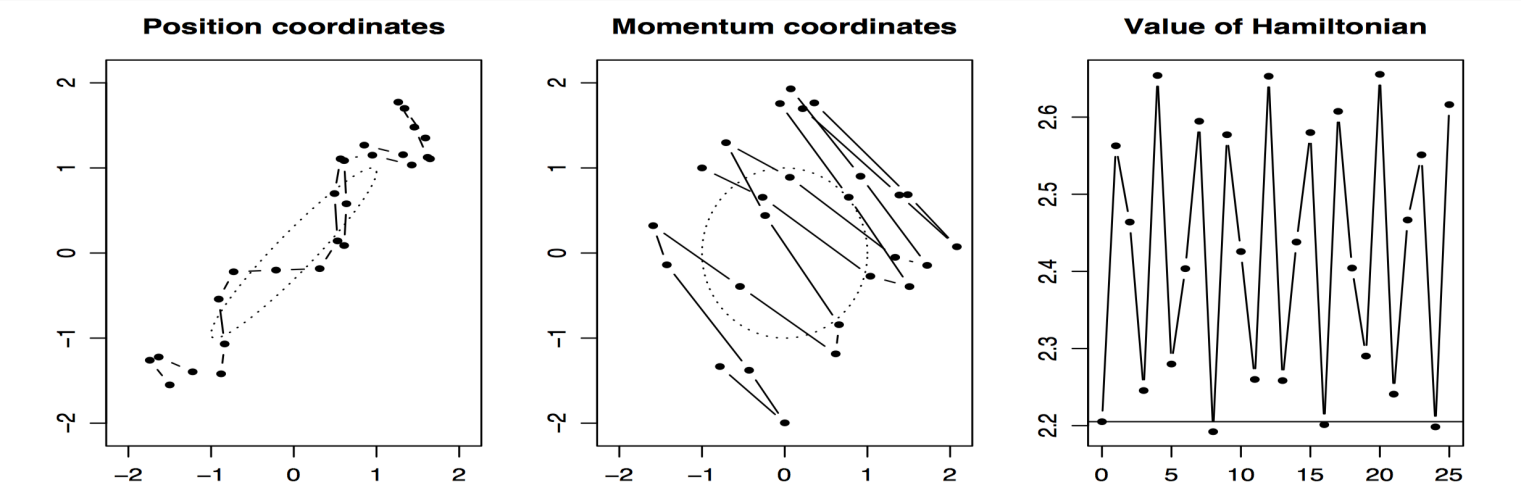



Figure 3: A trajectory for a 2D Gaussian distribution, simulated using 25 leapfrog steps with a stepsize of 0.25. The ellipses plotted are one standard deviation from the means. The initial state had $q = [-1.50, -1.55]^T$ and $p = [-1, 1]^T$.

# HMC and Its Benefits

Avoidance of random-walk behaviour, as illustrated above, is one major benet of Hybrid Monte Carlo.[5]

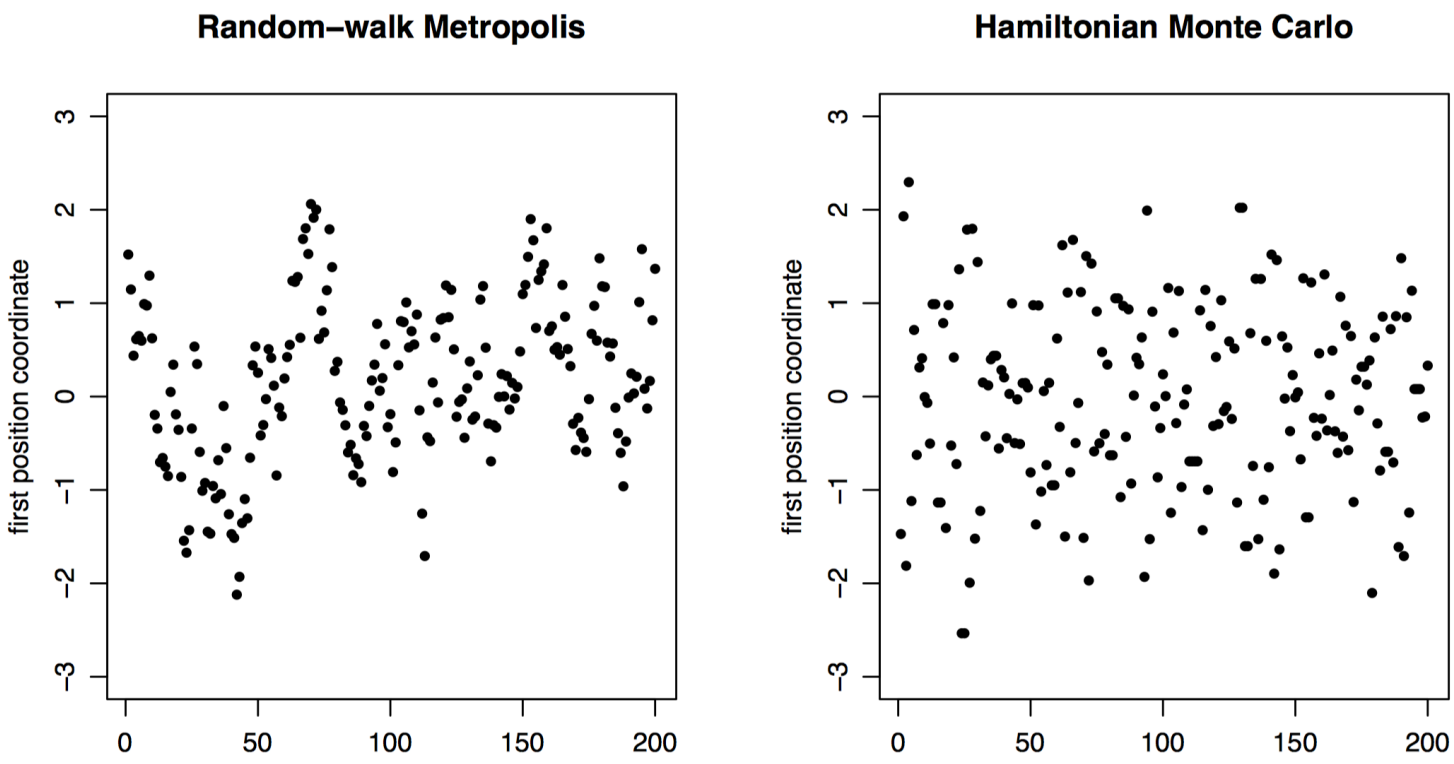



Figure 5: Two hundred iterations, starting with the twenty iterations shown above, with only the first position coordinate plotted.

# Taking One Step at A Time: The Langevin Method

If we express an iteration of HMC with one leapfrog step in the following way:

$$\begin{cases} p(t+\frac{\epsilon}{2}) = p(t) - \frac{\epsilon}{2}\frac{\partial H}{\partial x}|_{x(t)} \\ x(t+\epsilon) = x(t) + \epsilon p(t+\frac{\epsilon}{2}) \\ p(t+\epsilon) = p(t+\frac{\epsilon}{2}) - \frac{\epsilon}{2}\frac{\partial H}{\partial x}|_{x(t+\epsilon)} \end{cases}$$

$$\begin{cases} x(t+\epsilon) = x(t) - \frac{\epsilon^2}{2}\frac{\partial H}{\partial x}|_{x(t)} + \epsilon p(t) \\ p(t+\epsilon) = p(t) - \frac{\epsilon}{2}\frac{\partial H}{\partial x}|_{x(t)} - \frac{\epsilon}{2}\frac{\partial H}{\partial x}|_{x(t+\epsilon)} \end{cases}$$

Again we sample the momentum p from their Gaussian distribution with zero mean and unit variance.

This equation for x update is known in physics as one type of "Langevin equation", and it is a special case of Hybrid Monte Carlo.[7]

# Another Way to Avoid Random Walks

Although LMC can be seen as a special case of HMC, it will explore the distribution via an inefficient random walk, just like random-walk Metropolis updates.

However, there is another way to keep it away from randomly walking, that is, so called "partial momentum refreshment", as proposed by Horowitz (1991)[8].

The main idea is a linear combination between the previous momentum state and a random noise, which has the same mean and covariance matrix with the distribution of momentum:

$$p(t+\epsilon) = \alpha p(t) + \sqrt{(1-\alpha^2)}n, \qquad \textit{where } \alpha \in [-1,1] \;\; \textit{and} \;\; n \sim \mathcal{N}(0, I)$$

# Conclusion

# Pros and Cons of Hybrid Monte Carlo

Pros:

- Hybrid Monte Carlo accepts in most cases new states
- Less iterations to get representative sampling

Cons:

- The gradient of desired distribution $p(x)$ may not exist
- The computational time of gradient may last a long time
- Problems with sampling from distributions with isolated local minimums(*Tempering during a trajectory*)

# Questions?

# References i



Wilczek, Frank.
**Conservation laws (physics), 2014.**
https://www.accessscience.com:443/content/conservation-laws-physics/757423, Last accessed on 2018-07-09.


J Willard Gibbs.
**Elementary principles in statistical mechanics.**
Courier Corporation, 2014.


Wikipedia contributors.
**Hamiltonian system — Wikipedia, the free encyclopedia, 2018.**
[Online; accessed 9-July-2018].

# References ii



Wikipedia contributors.
**Hamiltonian mechanics — Wikipedia, the free encyclopedia, 2018.**
[Online; accessed 9-July-2018].


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


Paul B. Mackenzie.
**An Improved Hybrid Monte Carlo Method.**
*Phys. Lett.*, B226:369–371, 1989.


AD Kennedy.
**The theory of hybrid stochastic algorithms.**
In *Probabilistic methods in quantum field theory and quantum gravity*, pages 209–223. Springer, 1990.

# References iii


Alan M Horowitz.
**A generalized guided monte carlo algorithm.**
*Phys. Lett. B*, 268(CERN-TH-6172-91):247–252, 1991.