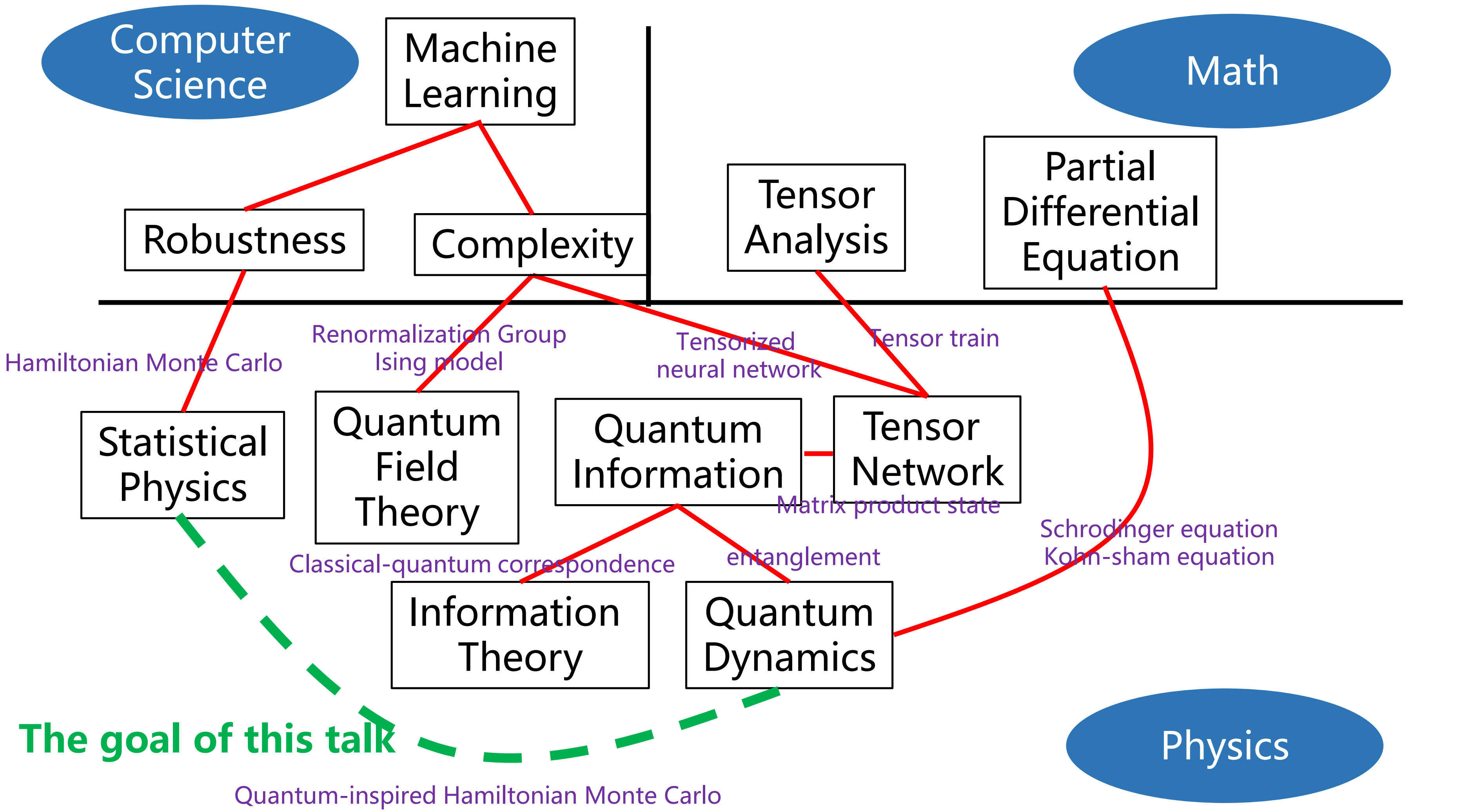
Langevin-type sampling methods

Based on summer literature review
School of Physics, Peking University
Ziming Liu
Advisor: Zheng Zhang, UCSB ECE

The big party



Overview

1. Introduction to Bayesian models
2. Introduction to Langevin dynamics (1st,2nd,3rd-order)
3. Hamiltonian Monte Carlo (HMC)

1. Introduction to Bayesian models

Maxwell-Boltzmann distribution

Description: for isothermal system

Single-particle system (temperature T)

Link between pdf and energy func

For state x, energy $E(x)$, then probability density $p(x) \propto \exp(-\frac{E(x)}{k_B T})$

Theory

- Maximum entropy principle for isolated system (canonical ensemble)
- Minimum free energy principle for isothermal system

Ideal gas	$E = \frac{q^2}{2m}$	$p(q) \propto \exp(-\frac{q^2}{2mk_B T})$
Static isothermal atmosphere	$E = U(x)$	$p(x) \propto \exp(-\frac{U(x)}{k_B T})$
Gas in a well	$E = U(x) + \frac{q^2}{2m}$	$p(x, q) \propto \exp(-\frac{U(x) + \frac{q^2}{2m}}{k_B T})$

Bayesian model

What ?

θ: model parameters

$$p(\theta|D) \propto p(D|\theta)p(\theta)$$

posterior likelihood prior

Link to regression models

$$p(\theta|D) = \exp\big(-U(\theta)\big)$$

$$U(\theta) = -\log\big(p(D|\theta)\big) - \log(p(\theta))$$

Regression error Regularization term

$$\theta^* = \underset{\theta}{argmin}\, U(\theta) \Longleftrightarrow \theta^* = \underset{\theta}{argmax}\, p(\theta|D)$$

Global Optima Maximum Posterior Estimation

Bayesian model

Why ?

$$\theta^* = \underset{\theta}{argmin}\, U(\theta) \iff \theta^* = \underset{\theta}{argmax}\, p(\theta|D)$$

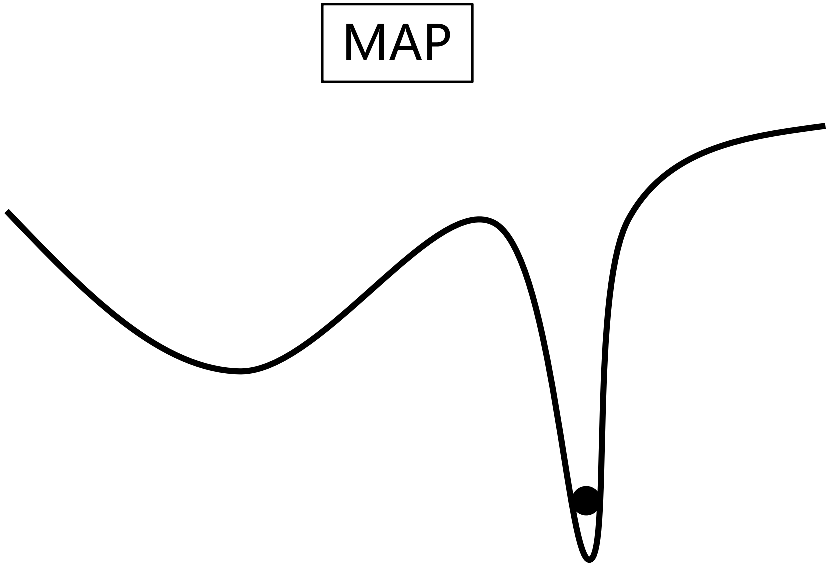
Global Optima

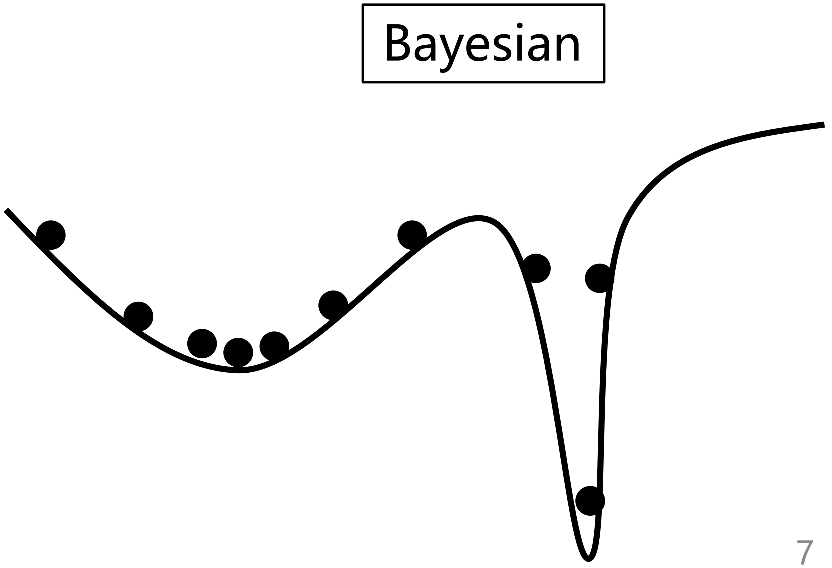
Maximum a Posterior Estimation

Limitations of MAP

1 No uncertainty quantification (point estimation)

2 Risk of overfitting !





优化算法新观点

2prime

applied math, data science, numerical analysis

已关注

luckystarufo、张旦波、王磊、侯霁开等 208 人赞同了该文章

然后随机的分析方法就是 https://zhuanlan.zhihu.com/p/33563623

基本的motivation是最简单的梯度下降

$$x_{k+1} = x_k - \Delta t \nabla f(x_k) + \sqrt{\Delta t}\eta_k, \eta_k \sim N(0; I)$$

可以理解成在simulate这个sde $\dot{X} = -\nabla f(X) + dW_t$

今年nips有工作把加速的框架放进到这个随机的版本里了，这里不提

我更想说的是

这个sde $\dot{X} = -\nabla f(X) + dW_t$ 最后收敛到gibbs分布

所以这里就把贝叶斯采样和优化算法联系起来了，这样也能理解一个著名的结果sgd会“train faster generalize better”，因为贝叶斯会采样到flat的minima，所以会更加robust

Markov Chain Monte Carlo (MCMC)

Markov Chain — Bayesian model

$$P_s(\theta) = P(\theta|D)$$

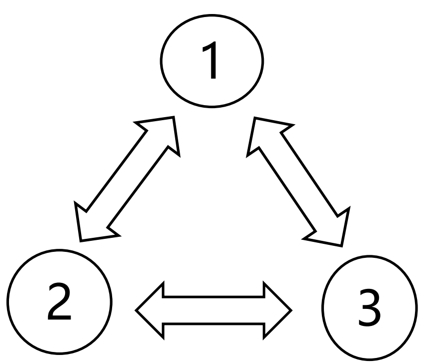
Steady distribution — Posterior distribution

Given a steady distribution, how to construct a Markov Chain?

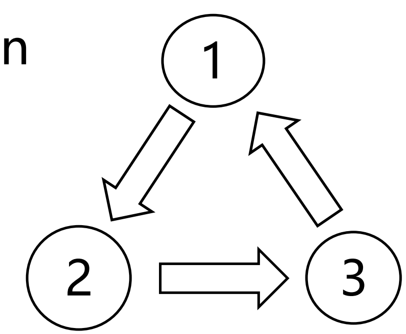
(Simplified to)

Steady distribution $\Rightarrow$ Detailed balance

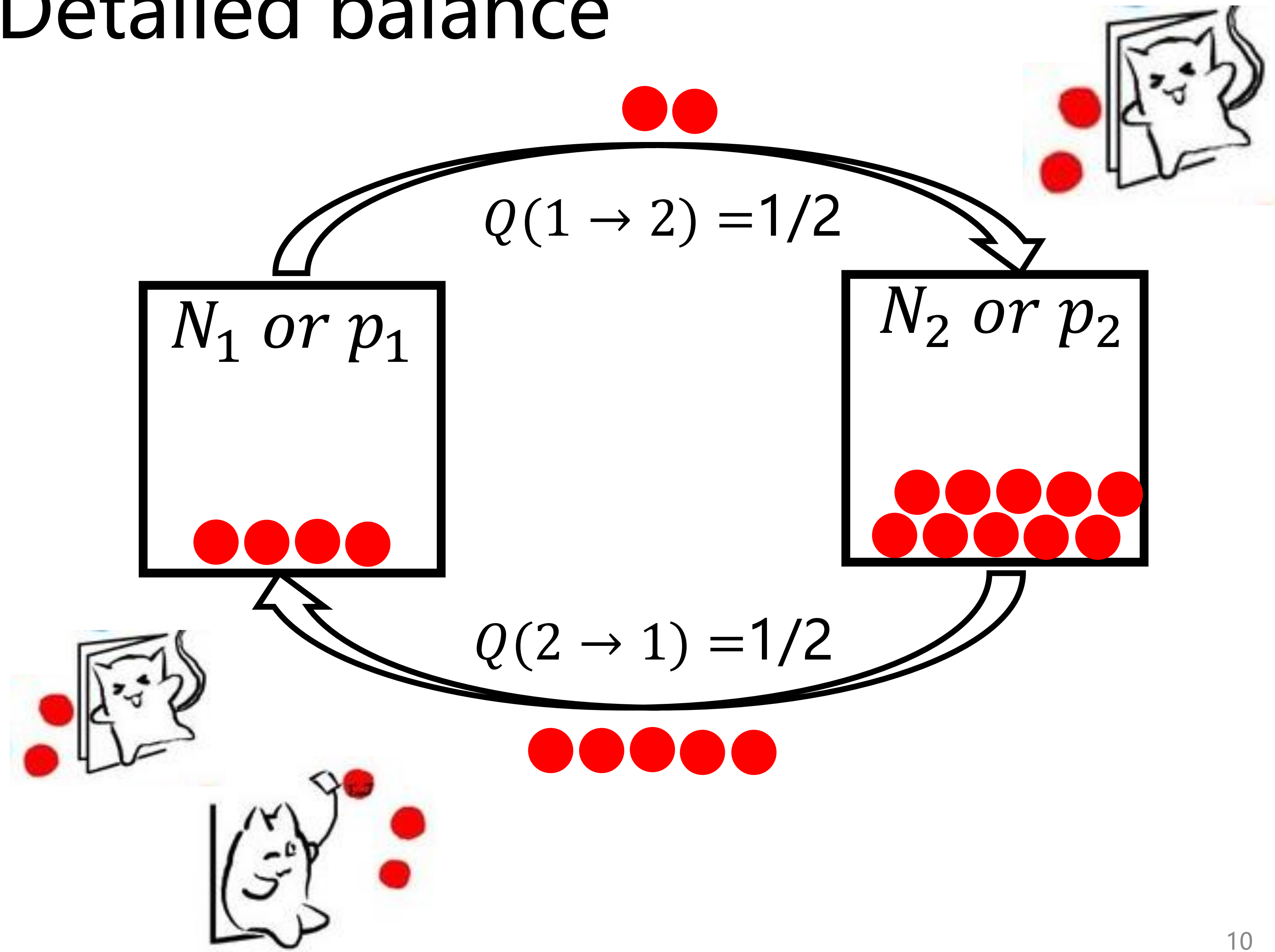
Detailed balance:



Steady distribution
But no db:



Detailed balance



Metropolis-Hastings (MH)

The MH algorithm for sampling from a target distribution $p(x)$, using transition kernel Q, consists of the following steps:

- For $t = 1, 2, \cdots$
  - Sample y from $Q(y|x_t)$. Think of y as a proposed value of x_{t+1}.
  - Compute acceptance probability

$$A(x_t \rightarrow y) = \min(1, \frac{p(y)Q(x_t|y)}{p(x_t)Q(y|x_t)})$$

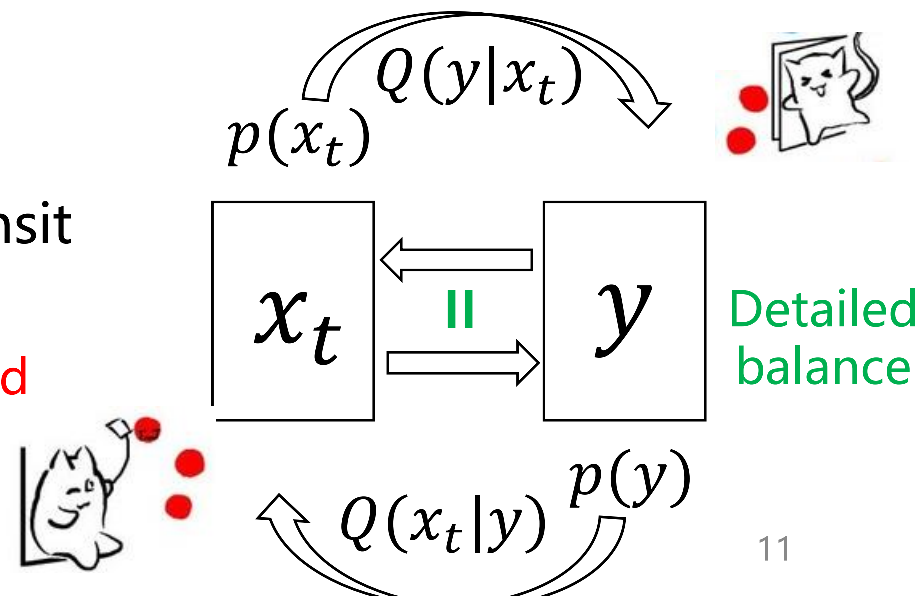
  - With probability A accept the proposed value, and set $x_{t+1} = y$. Otherwise, set $x_{t+1} = x_t$.

$p(x)$: number of particles at state x

$Q(y|x)$: transition rate from x to y

$p(x)Q(y|x)$: number of particles transit from x to y

$p(x)Q(y|x)A(x \rightarrow y)$: number of accepted particles transit from x to y



Metropolis algorithm

The Metropolis algorithm for sampling from a target distribution $p(x)$,transit through random walk, consists of the following steps:

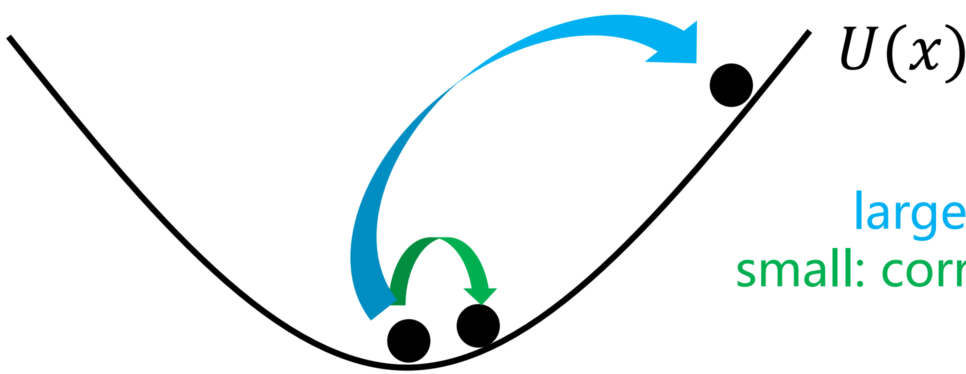
- For $t = 1, 2, \cdots$
  - Random walk to y from x_t. Think of y as a proposed value of x_{t+1}.
  - Compute acceptance probability

$$A(x_t \to y) = \min(1, \frac{p(y)}{p(x_t)})$$

  - With probability A accept the proposed value, and set $x_{t+1} = y$. Otherwise, set $x_{t+1} = x_t$.

Comment: random walk is symmetric, so $Q(y|x) = Q(x|y)$

In thermodynamics models, $P(x) \sim \exp(-U(x)/T)$ (Boltzmann distribution)



Step size
large: low acceptance rate
small: correlated & not independent

2. Introduction to Langevin Dynamics

Zoo of Langevin dynamics

Stochastic Gradient Langevin Dynamics (cite=718)

1^{st}order, general

$$\Delta\theta_t = \frac{\epsilon_t}{2}\left(\nabla \log p(\theta_t) + \frac{N}{n}\sum_{i=1}^{n} \nabla \log p(x_{ti}|\theta_t)\right) + \eta_t$$

$$\eta_t \sim N(0, \epsilon_t) \tag{4}$$

Welling, Max, and Yee W. Teh. "Bayesian learning via stochastic gradient Langevin dynamics." *Proceedings of the 28th international conference on machine learning (ICML-11)*. 2011.

Stochastic sampling using Fisher information (cite=207)

1^{st}order, gaussian

$$\theta_{t+1} \leftarrow \theta_t + \frac{\epsilon C}{2}\{-I_N(\theta_t - \theta_0)\} + \omega$$

$$\text{where} \quad \omega \sim \mathcal{N}(0, \epsilon C - \frac{\epsilon^2}{4} C I_N C)$$

Ahn, Sungjin, Anoop Korattikara, and Max Welling. "Bayesian posterior sampling via stochastic gradient Fisher scoring." *arXiv preprint arXiv:1206.6380* (2012).

Stochastic Gradient Hamiltonian Monte Carlo (cite=300)

2^{nd}order

$$\begin{cases} d\theta = M^{-1} r\, dt \\ dr = -\nabla U(\theta)\, dt - CM^{-1} r dt \\ \qquad + \mathcal{N}(0, 2(C - \hat{B})dt) + \mathcal{N}(0, 2B dt) \end{cases}$$

Chen, Tianqi, Emily Fox, and Carlos Guestrin. "Stochastic gradient hamiltonian monte carlo." *International conference on machine learning*. 2014.

Stochastic sampling using Nose-Hoover thermostat (cite=140)

3^{rd}order

$$d\boldsymbol{\theta} = \mathbf{p}\, dt, \quad d\,\mathbf{p} = \tilde{\mathbf{f}}(\boldsymbol{\theta})dt - \xi\, \mathbf{p}\, dt + \sqrt{2A}\,\mathcal{N}(0, dt)$$

$$d\xi = (\frac{1}{n}\mathbf{p}^\top \mathbf{p} - 1)dt.$$

Ding, Nan, et al. "Bayesian sampling using stochastic gradient thermostats." *Advances in neural information processing systems*. 2014.

1^{st} order Langevin dynamics

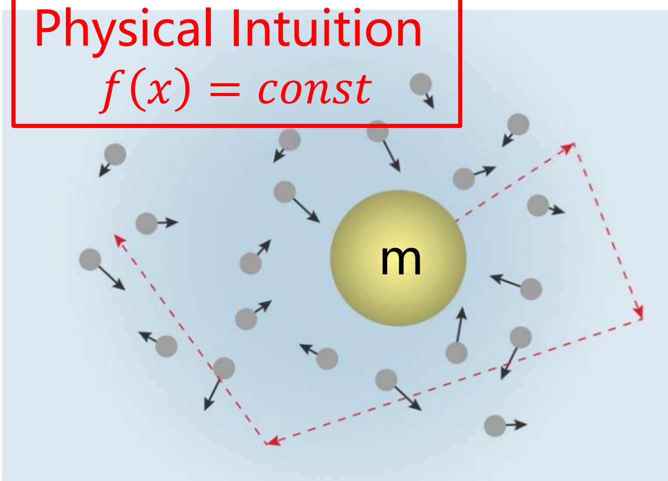
(also known as Brownian motion or Wiener Process)

$$dx = -\nabla f(x)dt + \beta^{-\frac{1}{2}} dW(t) \qquad \rho_s \propto \exp(-\beta f(x))$$

Energy function (bayesian) / loss function (optimization)

Physical Intuition

$f(x) = const$



The properties of the medium

A heat bath (temperature T)

Hit the ball every t_0 (憋大招)

transfer momentum $p \sim exp(-\frac{p^2}{2Tt_0})$

Overdamped (coefficient γ large)

Small relaxation time

①The ball gains a momentum p from particles (fluctuating) around it.

② It travels in the damping medium

$$m\ddot{x} = -\gamma\dot{x}$$

$$\rightarrow \dot{x} = \frac{p}{m}\exp\left(-\frac{\gamma}{m}t\right), x = \frac{p}{\gamma}(1 - \exp(-\frac{\gamma}{m}t))$$

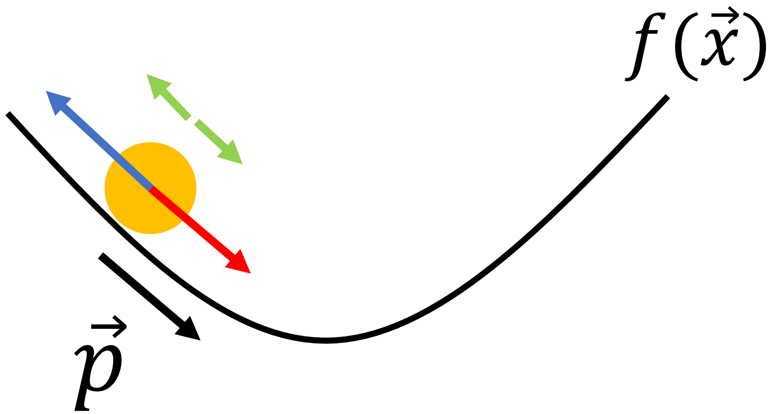
③Overdamped condition, then $\exp\left(-\frac{\gamma}{m}t_0\right) \rightarrow 0.$

So at time t, the total displacement is $\frac{p}{\gamma} \propto p.$

i.e. $dx \propto \frac{1}{\gamma}\exp\left(-\frac{p^2}{2Tt_0}\right) \propto \sqrt{T}dW(t_0) \quad (\beta = \frac{1}{T})$

2nd order Langevin dynamics

$$\begin{cases} d\vec{x} = \vec{p}dt \\ d\vec{p} = \underbrace{-\nabla f(\vec{x})dt}_{\text{Conservative Force}} - \underbrace{A\vec{p}dt}_{\text{Damping Force}} + \underbrace{\sqrt{2AT}dW}_{\text{Thermal "Force"}} \end{cases}$$



Invariant measure:

$$P_s(\vec{x},\vec{p}) \propto \exp(-(\frac{p^2}{2} + U(\vec{x}))/T)$$

Fokker Planck Eq for 2nd order LD

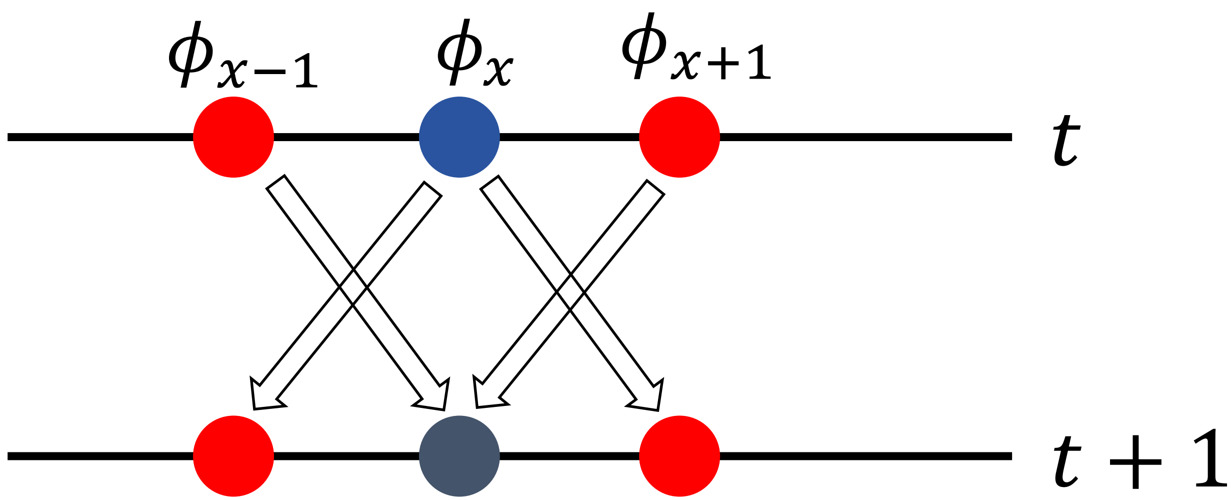
Dynamical Equations

$$\mathrm{d}x = M^{-1}p\mathrm{d}t, \quad \mathrm{d}p = [-\nabla U(x) - \gamma p]\mathrm{d}t + \sigma M^{1/2}\mathrm{d}W$$

Fokker-Planck Equations

$$\phi_t = -M^{-1}p \cdot \nabla_x \phi + \nabla U(x) \cdot \nabla_p \phi + \gamma \nabla_p \cdot (p\phi) + \frac{\sigma^2}{2}\Delta_p \phi$$

One-dim random walk (不变原理)



$$\frac{\partial \phi}{\partial t} = \frac{1}{2}(\phi_{x-1} + \phi_{x+1} - 2\phi_x) \approx \frac{1}{2}\frac{\partial^2 \phi}{\partial x^2}$$

3rd order Langevin dynamics (special)

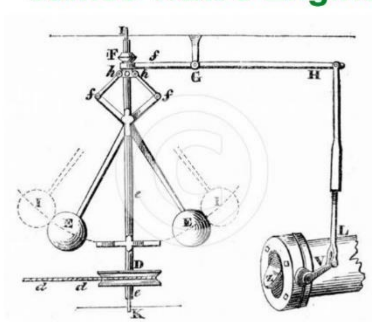
$$\begin{cases} d\vec{\theta} = \vec{p}dt \\ d\vec{p} = -\nabla U\left(\vec{\theta}\right)dt - \zeta\vec{p}dt + \sqrt{2AT}dW \\ d\zeta = \left(\frac{p^2}{n} - T_0\right)dt \end{cases}$$

Thermal term (thermostat)

When $p\uparrow \rightarrow \frac{p^2}{n} \sim T > T_0 \rightarrow \zeta\uparrow \rightarrow$ **more friction on $\vec{p}$** $\rightarrow p\downarrow$

Negative feedback loop

James Watt's Engine



Too fast: balls move to outside, opening valve, releasing steam, reducing pressure, reducing speed

Too slow: balls fall to inside, closing valve, leading to an increase in pressure, increasing speed

3^{rd} order Langevin dynamics (general)

$$\boxed{dq = M^{-1}dp}$$

$$\boxed{dp = -\nabla U(q)dt + \sigma_F\sqrt{\Delta t}M^{\frac{1}{2}}dW - \zeta p dt + \sigma_A M^{\frac{1}{2}}dW}$$

$$\boxed{d\zeta = \frac{1}{\mu}[p^T M^{-1}p - N_d k_B T]dt - \gamma\zeta dt + \sqrt{2k_B T\gamma}dW}$$

Invariant measure: $\exp(-\beta U(q))\exp(-\beta p^T M^{-1}p/2)\exp\left(-\mu(\zeta-\hat{\gamma})^2/2\right)$

$$\hat{\gamma} = \beta(\sigma_F^2 + \sigma_A^2)/2$$

Additivity

The thermostats can be **combined** in most cases without altering their effectiveness (often improving it).

$$\dot{x} = f(x) + g(x)$$

$$\left.\begin{array}{l} \mathscr{L}_f^\dagger \rho = 0 \\ \mathscr{L}_g^\dagger \rho = 0 \end{array}\right. \Rightarrow \mathscr{L}_{f+g}^\dagger \rho = 0$$

Works for **SDEs** too…

Slides from:
https://ergodic.org.uk/~bl/Data/Slides/MD4.pdf

3rd order Langevin dynamics

$$\boxed{dq = M^{-1}dp}$$

$$\boxed{dp = -\nabla U(q)dt + \sigma_F\sqrt{\Delta t}M^{\frac{1}{2}}dW - \zeta p dt + \sigma_A M^{\frac{1}{2}}dW}$$

$$\boxed{d\zeta = \frac{1}{\mu}[p^T M^{-1}p - N_d k_B T]dt - \gamma\zeta dt + \sqrt{2k_B T\gamma}dW}$$

Invariant measure: $\exp(-\beta U(q))\exp(-\beta p^T M^{-1}p/2)\exp\left(-\mu(\zeta-\hat{\gamma})^2/2\right)$

$\hat{\gamma} = \beta(\sigma_F^2 + \sigma_A^2)/2$

Building Blocks

Hamiltonian dynamics	$\exp(-\beta U(q))\exp(-\beta p^T M^{-1}p/2)$
Thermostat	$\exp(-\beta p^T M^{-1}p/2)\exp(-\mu\zeta^2/2)$
OU process for ζ	$\exp(-\mu\zeta^2/2)$
Noise for p	$\exp\left(-\frac{\mu\zeta^2}{2}\right) \to \exp\left(-\frac{\mu(\zeta-\hat{\gamma})^2}{2}\right)$

3. Hamiltonian Monte Carlo (HMC)

Neal, Radford M. "MCMC using Hamiltonian dynamics." *Handbook of markov chain monte carlo* 2.11 (2011): 2.

Betancourt, Michael. "A conceptual introduction to Hamiltonian Monte Carlo." *arXiv preprint arXiv:1701.02434* (2017).

2nd order Langevin dynamics

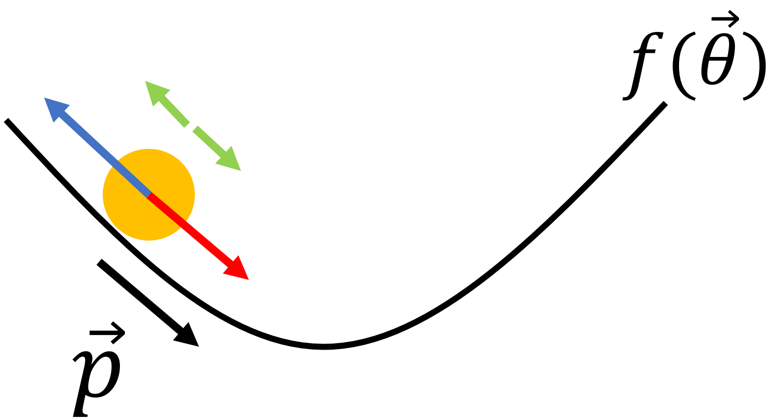
Invariant measure:

$$P_s(\vec{x},\vec{p}) \propto \exp\left(-\left(\frac{p^2}{2}+U(\vec{x})\right)/T\right)$$

$$\begin{cases} d\vec{x} = \vec{p}\,dt \\ d\vec{p} = \underbrace{-\nabla f(\vec{x})\,dt}_{\text{Conservative Force}} - \underbrace{A\vec{p}\,dt}_{\text{Damping Force}} + \underbrace{\sqrt{2AT}\,dW}_{\text{Thermal ``Force''}} \end{cases}$$

$$\Downarrow A = 0$$

$$\begin{cases} d\vec{x} = \vec{p}\,dt \\ d\vec{p} = -\nabla f(\vec{x})\,dt \end{cases}$$



Hamiltonian dynamics

Hamiltonian equations

$$\begin{cases} d\vec{x} = M^{-1}\vec{p}dt & \text{Definition of momentum} \\ d\vec{p} = \underbrace{-\nabla U(\vec{x})}_{f(\vec{x})\text{: conservative force}}dt & \text{Momentum theorem} \end{cases}$$

Hamiltonian $$H(x,p) = \underbrace{\frac{1}{2}\,p^T M^{-1} p}_{\text{Kinetic energy}} + \underbrace{U(x)}_{\text{Potential energy}}$$

Energy conservation

$$dH = p^T M^{-1} dp + dU(x) = -dx^T \nabla U(x) + dU(x) = 0$$

Steady distribution

Hamiltonian Equations

$$\begin{cases} d\vec{x} = M^{-1}\vec{p}dt \\ d\vec{p} = -\nabla U(\vec{x})dt \end{cases}$$

Fokker Planck Equation:

$$\partial_t p + \left(\frac{\partial p}{\partial x}\right)^T \left(\frac{\partial H}{\partial x}\right) + \left(\frac{\partial p}{\partial q}\right)^T \left(\frac{\partial H}{\partial q}\right) = 0$$

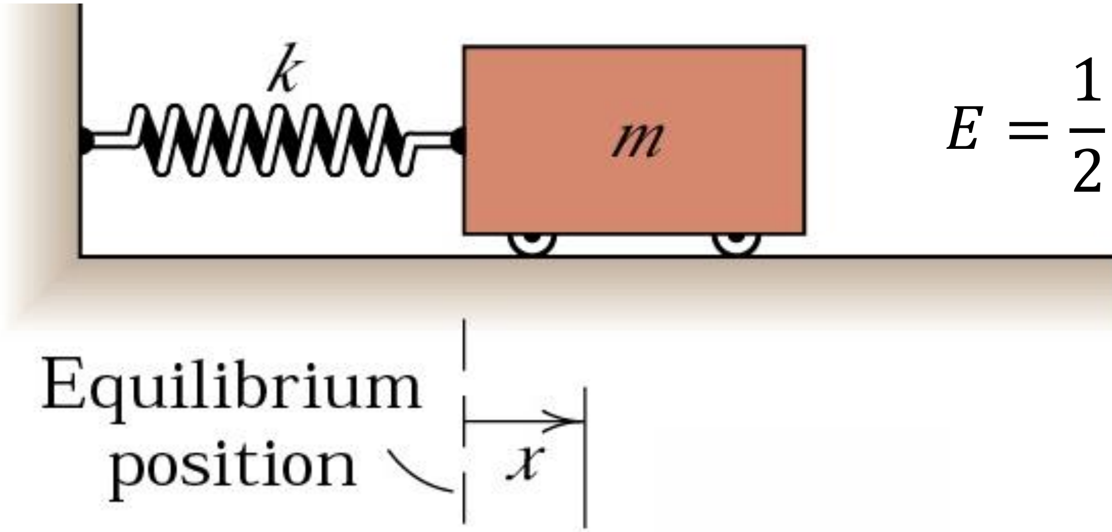
(Also known as Liouville’ s Theorem in physics)

Steady distribution:

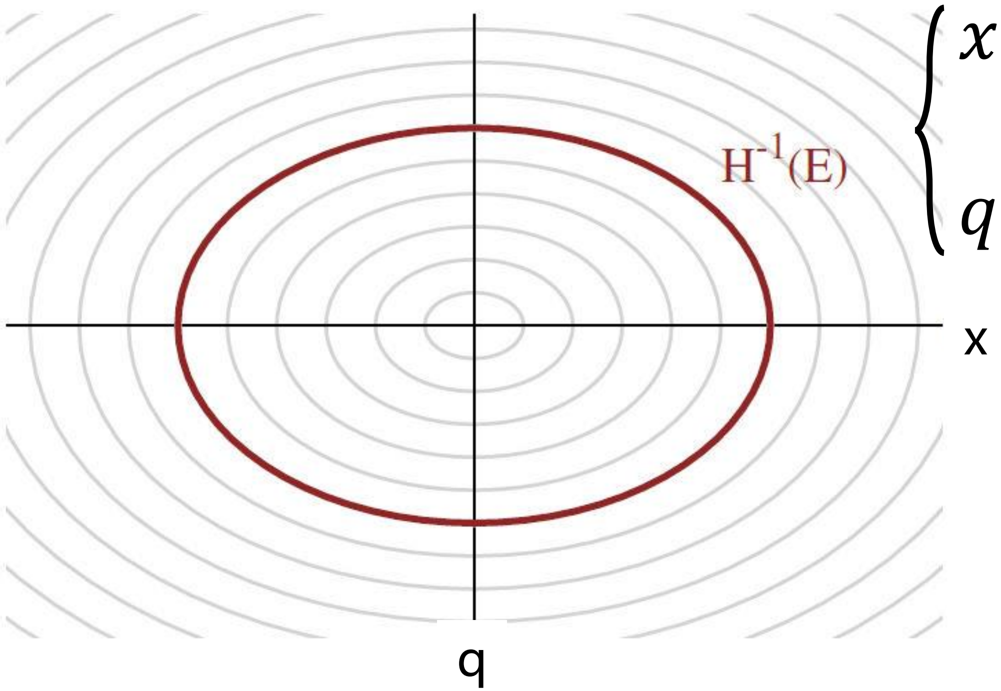
$$p_s(x, q) \propto \exp\left(-U(\vec{x}) - \frac{1}{2} q^T M^{-1} q\right)$$

$$p_s(x) \propto \exp\big(-U(\vec{x})\big) = p(x|D)$$

Example: 1d spring-mass system



$$E = \frac{1}{2}kx^2 + \frac{q^2}{2m}$$



$$\begin{cases} x = A\sin(\omega t + \phi_0) \\ q = m\omega A\cos(\omega t + \phi_0) \end{cases}$$

$$(\omega = \sqrt{\frac{k}{m}})$$

No ergodicity ?

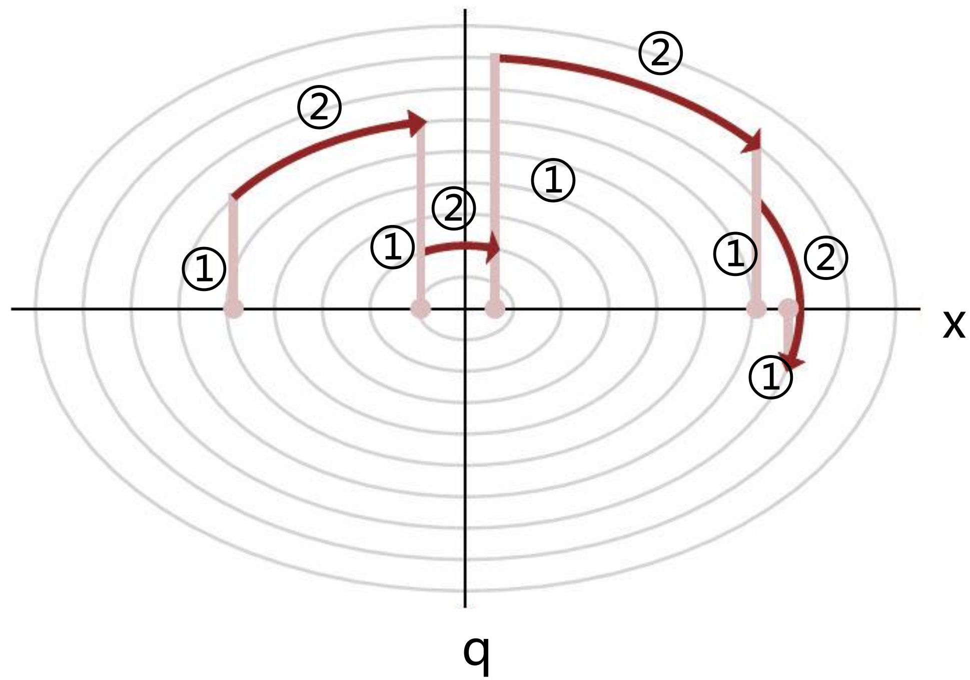
Example: 1d spring-mass system interacting with a heat bath

Ensemble

Maxwell-Boltzmann distribution

$$p(q) \propto \exp(-\frac{q^2}{2mk_BT})$$

Time

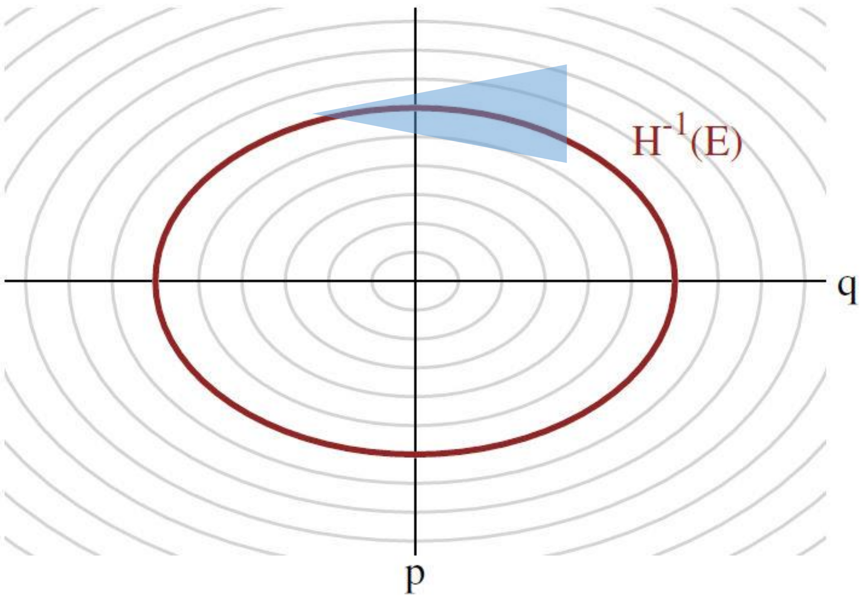


①momentum resampling $(m = k_BT = 1)$

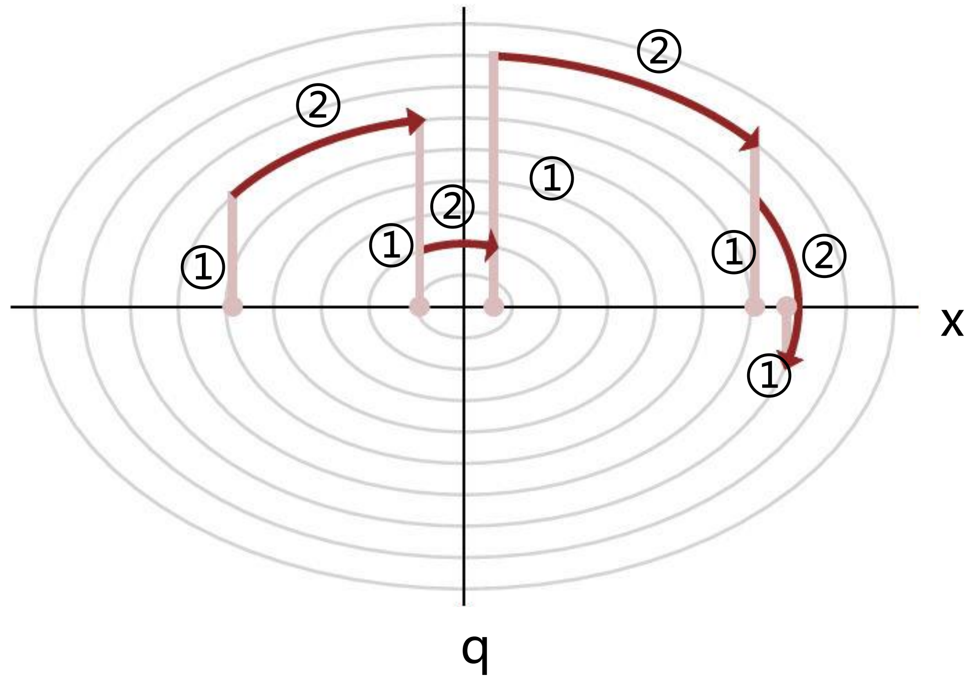
$$q \sim \mathrm{N}(0,1)$$

②travel on an energy level for a certain time (L steps)

2nd LD & HMC



Continuous scattering



Discrete scattering （憋大招）

Algorithm

Algorithm 1: Hamiltonian Monte Carlo

Input: starting point $\mathbf{x_0}$, step size ϵ ;
simulation steps L, mass $\mathbf{M} = m\mathbf{I}$
initialization;
for $j = 1, 2, \cdots$ **do**
  Resample $\mathbf{q} \sim \mathcal{N}(0, m)$;
  $(\mathbf{x}_0, \mathbf{q}_0) = (\mathbf{x}^{(t)}, \mathbf{q}^{(t)})$;
  Simulate dynamics based on Eq. (2);
  $\mathbf{r}_0 \leftarrow \mathbf{r}_0 - \frac{\epsilon}{2}\nabla U(\mathbf{x}_0)$;
  for $i = 1, \cdots, L$ **do**
    $\mathbf{x}_i \leftarrow \mathbf{x}_{i-1} + \epsilon \mathbf{M}^{-1}\mathbf{q}_{i-1}$;
    $\mathbf{q}_i \leftarrow \mathbf{q}_{i-1} - \epsilon \nabla U(\mathbf{x}_i)$
  end
  $\mathbf{q}_L \leftarrow \mathbf{q}_L - \frac{\epsilon}{2}\nabla U(\mathbf{x}_L)$;
  $(\hat{\mathbf{x}}, \hat{\mathbf{q}}) = (\mathbf{x}_m, \mathbf{q}_m)$;
  M-H step: $u \sim \text{Uniform}[0, 1]$;
  $\rho = e^{-H(\hat{\mathbf{x}}, \hat{\mathbf{q}}) + H(\mathbf{x}^{(t)}, \mathbf{q}^{(t)})}$;
  if $u < \min(1, \rho)$ **then**
    $(\mathbf{x}^{(t+1)}, \mathbf{q}^{(t+1)}) = (\hat{\mathbf{x}}, \hat{\mathbf{q}})$
  else
    $(\mathbf{x}^{(t+1)}, \mathbf{q}^{(t+1)}) = (\mathbf{x}^{(t)}, \mathbf{q}^{(t)})$
  end
end

Momentum resampling

Hamiltonian dynamics
(Leap frog scheme)

Metropolis-Hastings

Euler vs leap frog

e.g. $U(x) = \frac{1}{2}kx^2$

Euler's method

$$\begin{cases} q(t+\epsilon) = q(t) - \epsilon \dfrac{\partial U}{\partial x}(x(t)) \\ x(t+\epsilon) = x(t) + \epsilon \dfrac{q(t)}{m} \end{cases} \qquad \begin{bmatrix} x(t+\epsilon) \\ q(t+\epsilon) \end{bmatrix} = \boxed{\begin{bmatrix} 1 & \dfrac{\epsilon}{m} \\ -\dfrac{k\epsilon}{m} & 1 \end{bmatrix}} \begin{bmatrix} x(t) \\ q(t) \end{bmatrix}$$

det>1, not preserving volume !

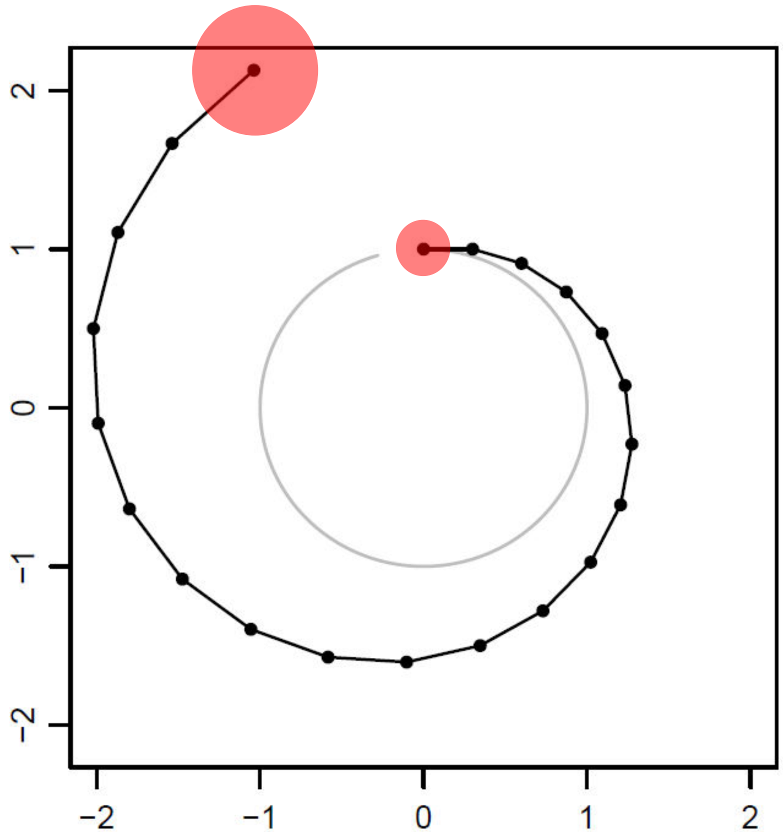
Leapfrog method

$$\begin{cases} q\left(t+\dfrac{\epsilon}{2}\right) = q(t) - \dfrac{\epsilon}{2}\dfrac{\partial U}{\partial x}(x(t)) \\ x(t+\epsilon) = x(t) + \epsilon \dfrac{q(t+\frac{\epsilon}{2})}{m} \\ q(t+\epsilon) = q\left(t+\dfrac{\epsilon}{2}\right) - \dfrac{\epsilon}{2}\dfrac{\partial U}{\partial x}(x(t+\epsilon)) \end{cases} \qquad \begin{bmatrix} x(t+\epsilon) \\ q(t+\epsilon) \end{bmatrix} = \boxed{\begin{bmatrix} 1 & 0 \\ -\dfrac{k\epsilon}{2m} & 1 \end{bmatrix}} \boxed{\begin{bmatrix} 1 & \dfrac{\epsilon}{m} \\ 0 & 1 \end{bmatrix}} \times \boxed{\begin{bmatrix} 1 & 0 \\ -\dfrac{k\epsilon}{2m} & 1 \end{bmatrix}} \begin{bmatrix} x(t) \\ q(t) \end{bmatrix}$$

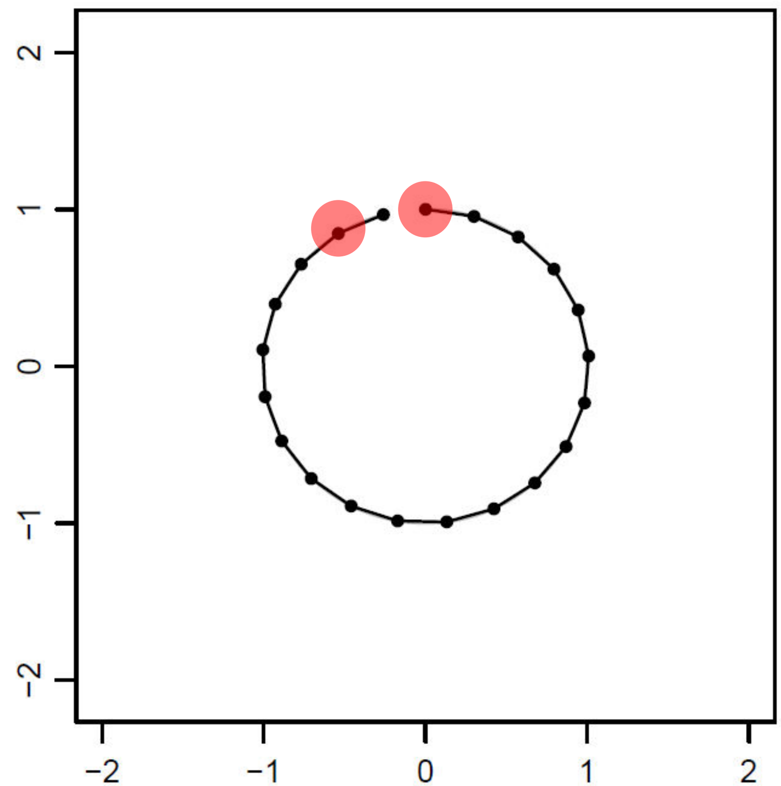
det=1, preserving volume !

Euler vs leap frog

Euler’ s method : diverge

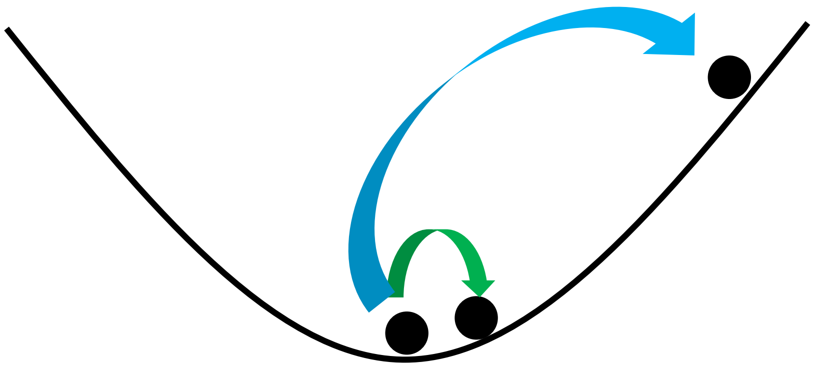


Leapfrog : stable



MCMC & HMC

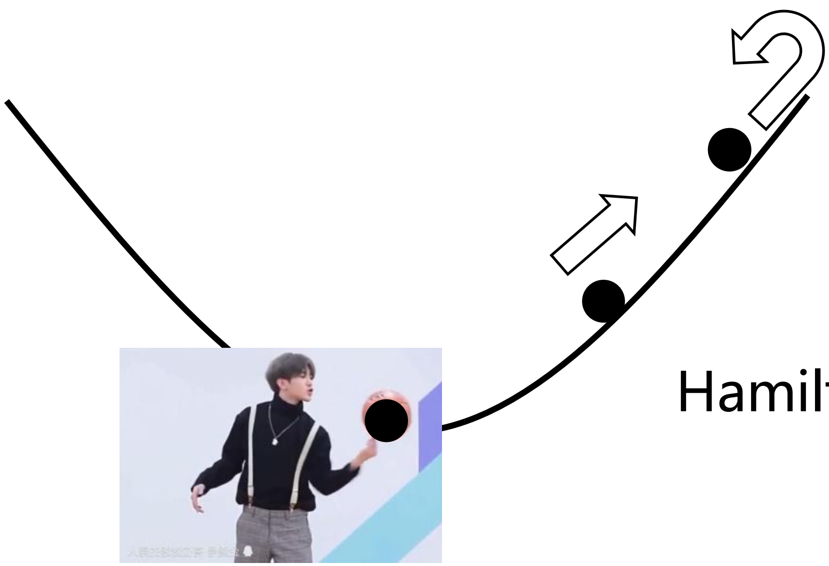
Random walk MCMC: position x



$U(x)$

$$P(x) \sim \exp(-U(x)/T)$$

HMC: position x + momentum q

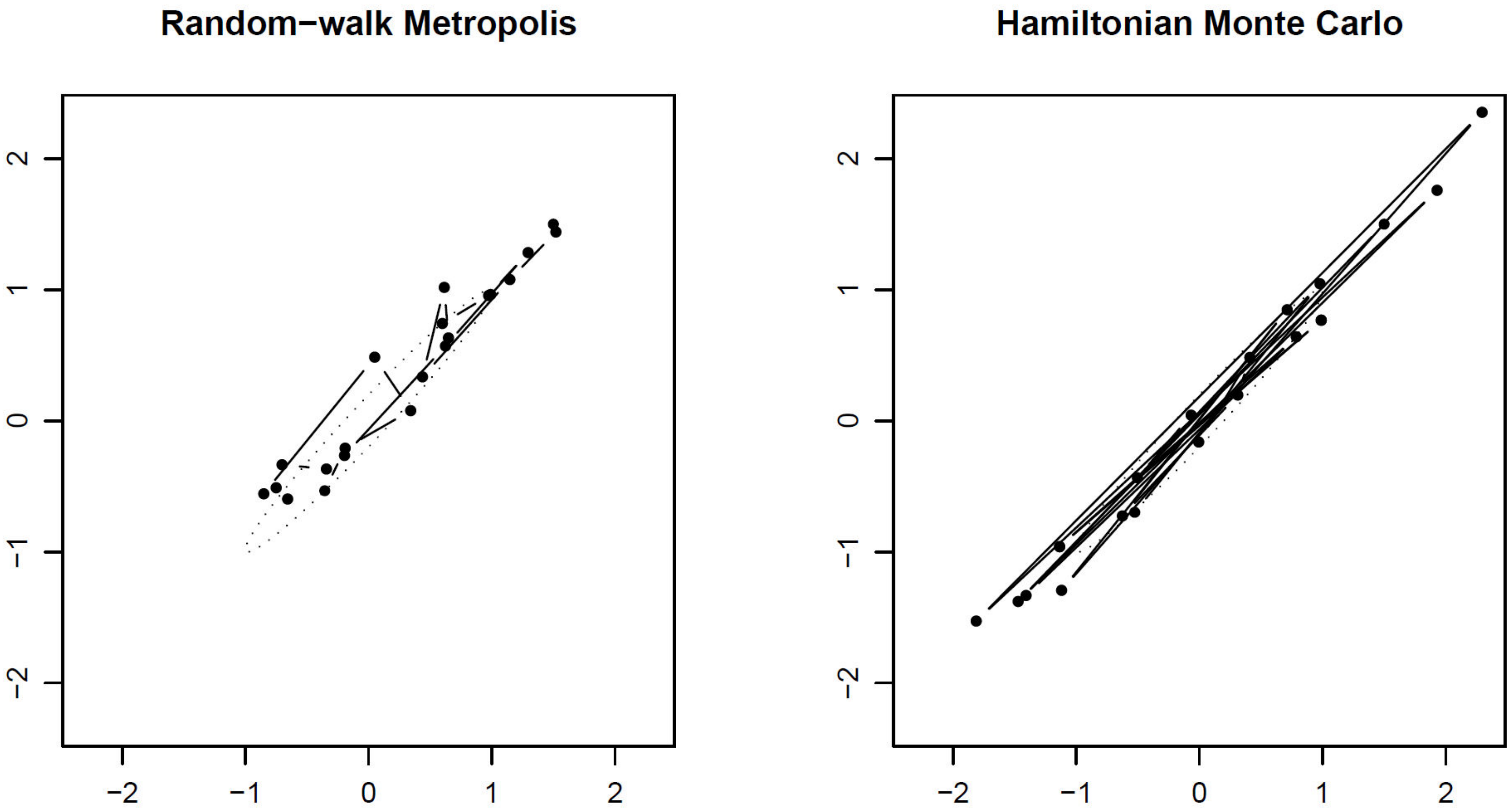


$U(x)$

$$P(x) \sim \exp(-(\underbrace{U(x) + K(q)}_{\text{Hamiltonian dynamics} \rightarrow \text{energy conservation}})/T)$$

Always accept ! (if step size→0)

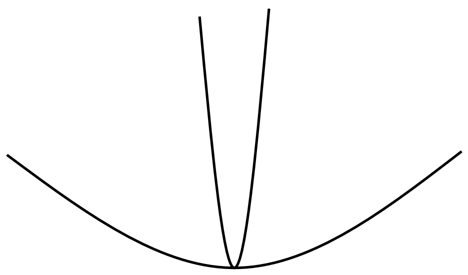
MCMC & HMC



Neal, Radford M. "MCMC using Hamiltonian dynamics." *Handbook of markov chain monte carlo* 2.11 (2011): 2.

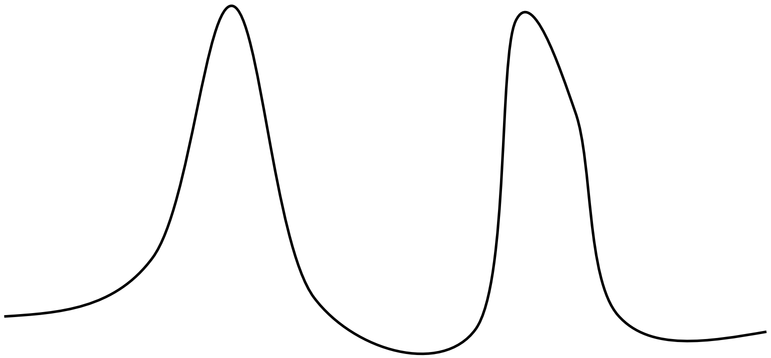
HMC limitations

①Ill-conditioned distributions



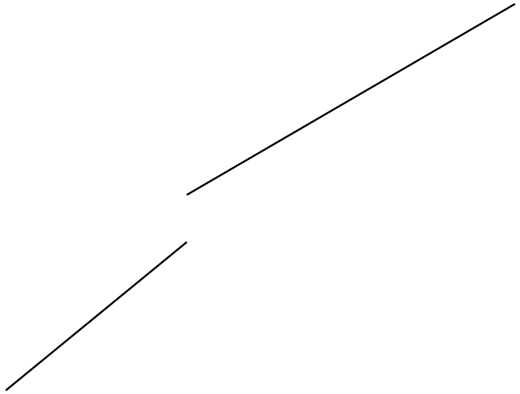
Need different masses
in different directions

②Multimodal distributions



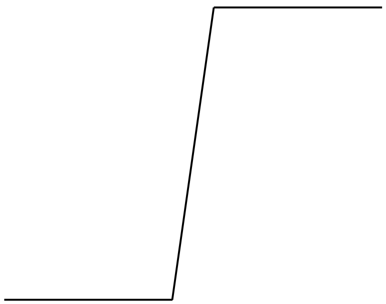
hard to escape from one mode

③Discontinuous



Large energy gap
Acceptance rate low

④spiky



Large gradients
Acceptance rate low

⑤Large training dataset Expensive gradients computation

HMC variants

Table 1: Summary of various HMC methods.

Problems	HMC variants	Physic theory
Ill-conditioned distribution	Riemannian HMC [14]	General relativity
Multimodal distribution	Magnetic HMC [42]	Electromagnetism
	Wormhole HMC [23]	General relativity
	Tempered HMC [15]	Thermodynamics
Large training data set	Stochastic HMC [7]	Langevin dynamics
	Thermostat HMC [8]	Thermodynamics
	Relativistic HMC [25]	Special relativity
Discontinuous energy function	Optics HMC [27]	Optics
Spiky distribution	**Quantum-inspired HMC (this work)**	**Quantum mechanics**

Riemannian HMC
Girolami, Mark, Ben Calderhead, and Siu A. Chin. "Riemannian manifold hamiltonian monte carlo." *arXiv preprint arXiv:0907.1100* (2009).

Magnetic HMC
Tripuraneni, Nilesh, et al. "Magnetic Hamiltonian Monte Carlo." *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017.

Wormhole HMC
Lan, Shiwei, Jeffrey Streets, and Babak Shahbaba. "Wormhole hamiltonian monte carlo." *Twenty-Eighth AAAI Conference on Artificial Intelligence*. 2014.

Continuous tempered HMC
Graham, Matthew M., and Amos J. Storkey. "Continuously tempered hamiltonian monte carlo." *arXiv preprint arXiv:1704.03338* (2017).

Stochastic Gradient HMC
Chen, Tianqi, Emily Fox, and Carlos Guestrin. "Stochastic gradient hamiltonian monte carlo." *International conference on machine learning*. 2014.

Stochastic Gradient Thermostat
Ding, Nan, et al. "Bayesian sampling using stochastic gradient thermostats." *Advances in neural information processing systems*. 2014.

Relativistic Monte Carlo
Lu, Xiaoyu, et al. "Relativistic monte carlo." *arXiv preprint arXiv:1609.04388* (2016).

Optics HMC
Afshar, Hadi Mohasel, and Justin Domke. "Reflection, refraction, and hamiltonian monte carlo." *Advances in Neural Information Processing Systems*. 2015.