Scalable Markov Chain Monte Carlo

Pavel Temirchev

Deep|Bayes, 2017

Contents

Why we need MCMC?

Why we need MCMC?

- To sample from a complex distribution $p(z)$.
- Possibly known only up to the normalization constant $p = \hat{p}/Z$.
- Evaluate statistics of that distribution, such as $\mathbb{E}_p f(z)$.

Why we need samples?

Example - Supervised learning:
Maximum likelihood approach:

$$p(y_i|x_i, D) = p(y_i|x_i, \theta^{ML})$$

Maximum aposteriory approach:

$$p(y_i|x_i, D) = p(y_i|x_i, \theta^{MAP})$$

But actually we need this:

$$p(y_i|x_i, D) = \mathbb{E}_{p(\theta|D)} p(y_i|x_i, \theta)$$

Why we need samples?

To evaluate this expectation using common Monte Carlo we need samples:

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

But,

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{\int p(D|\theta)p(\theta)d\theta}$$

isn't even tractable.

Reminder: Markov chains

Markov chain model:

$$p(z_1, z_2, ..., z_N) = p(z_1)p(z_2|z_1)...p(z_N|z_{N-1})$$

Markov property:

$$p(z_i|z_1, z_2, ..., z_N) = p(z_i|z_{i-1})$$

What is MCMC?

- The idea is to subsequently sample from a sufficient Markov Chain instead of sampling from p
- Samples will not be i.i.d. anymore
- But they must give us same convergency if we will use them in common Monte Carlo
- They must asymptotically explore all the p distribution

Example: Metropolis algorithm

Our desirable distribution to sample from is $p(z)$
We define a Markov chain using arbitrary proposal distribution q and sample a candidate point from it

$$z_* \sim q(z|z_t)$$

q must be symmetric, that is:

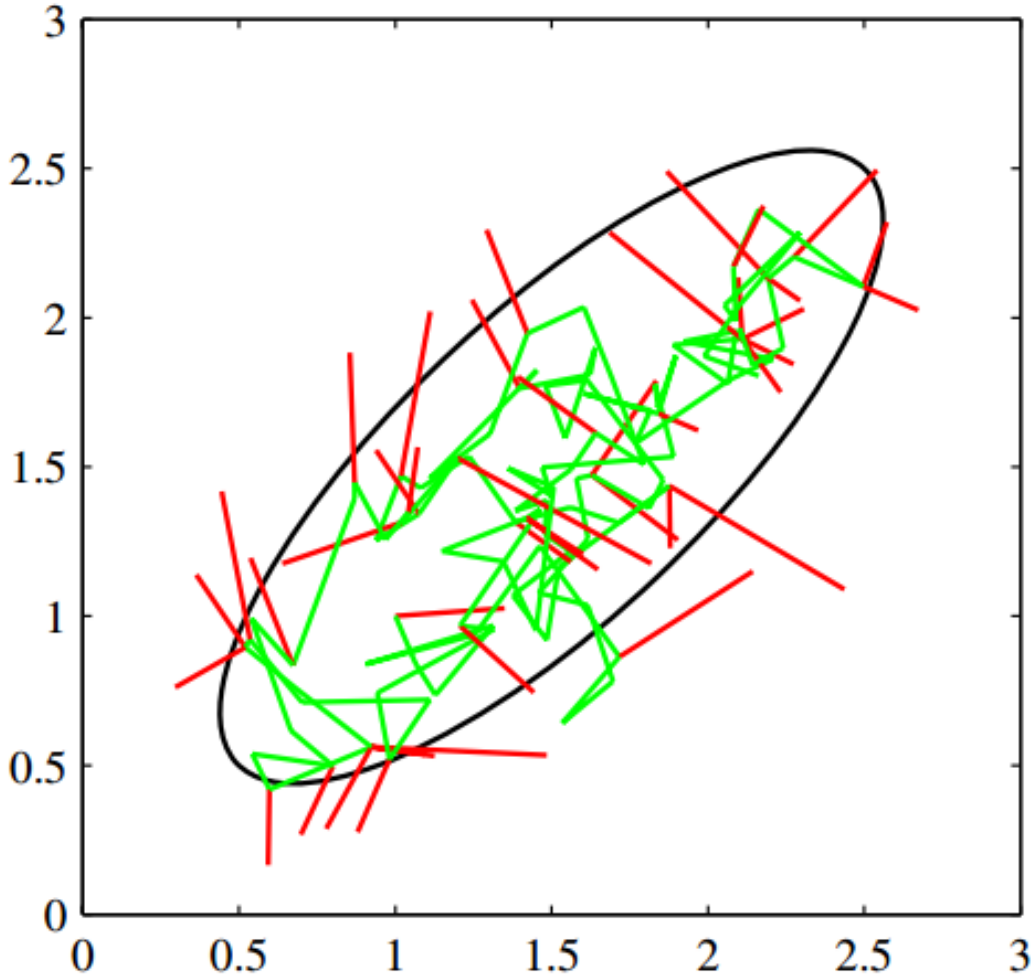
$$q(z_A|z_B) = q(z_B|z_A), \forall A, B$$

New candidate is accepted with probability:

$$A(z_*, z_t) = \min(1, \frac{p(z_*)}{p(z_t)})$$

In the rejection case $z_{t+1} = z_t$, else $z_{t+1} = z_*$
Samples are not discarded!

Example: Metropolis algorithm



Why it converges to p?

Theorem (Ergodic theorem)

If $(z_1, z_2, ..., z_N)$ *is an* **irreducible (homogeneous)** *discrete Markov Chain with* **stationary dist.** p *then,*

$$\frac{1}{N}\sum_{i=1}^{N} f(z_i) \to \mathbb{E}_p f(z)$$

almost surely when $N \to \infty$
for any bounded $f : \mathbf{Z} \to \mathbb{R}$

Stationary distributions and Irreducibility

A Markov Chain with transitional probabilities $T_m(z_{m+1}|z_m) \equiv p_m(z_{m+1}|z_m)$ is called **irreducible (homogeneous)** if they are the same for all m:

$$T_1(\cdot) = T_2(\cdot) = ... = T_m(\cdot) = ... = T(\cdot)$$

A distribution p is called to be **stationary** for a homogeneous Markov Chain with trans. prob. T if it leaves that distribution invariant:

$$p(z) = \int T(z|z')p(z')dz'$$

Detailed Balance

Detailed Balance property:

$$p(z)\,T(z'|z) = p(z')\,T(z|z')$$

Detailed Balance is sufficient (but not necessary) condition for p being stationary distribution for the Markov Chain
It is easy to proof (??)

Metropolis-Hastings algorithm

This is the generalization of the Metropolis algorithm for cases when q is not symmetric: $q(z_A|z_B) \neq q(z_B|z_A)$.
All the same, except of the following:

$$A(z_*, z_m) = \min(1, \frac{p(z_*)q(z_m|z_*)}{p(z_m)q(z_*|z_m)})$$

The detailed balance holds:

$$p(z)q(z'|z)A(z', z) = \min(p(z)q(z'|z), p(z')q(z|z')) = ...$$

$$... = p(z')q(z|z')A(z, z')$$

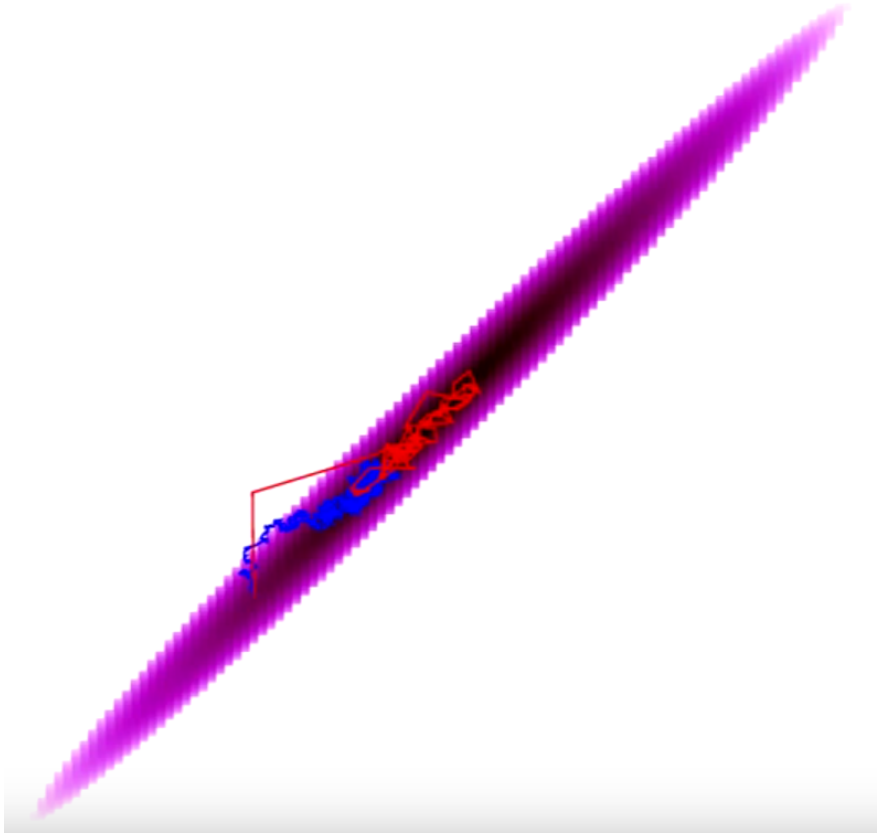
Is it a Scalable approach? (No)

Let's assume $q(z|z_m) = \mathcal{N}(z|z_m, s^2 I)$

- High s values implies big steps, but high rejection rates
- Low s values implies short steps with high acceptance rates.

If p has a covariance Σ, then optimal s^2 is of order $\min_i \Sigma_{ii}$
The number of steps requied to achieve independent sample is then of order $O(\sigma^2_{max}/\sigma^2_{min})$ if p is also Gaussian
With high dimensionality d things became worse.

Is it a Scalable approach? (No)



Hamiltonian Monte Carlo

- Hamiltonian Dynamics has a mechanical interpretation, where z denotes the position of the point.
- We define a momentum of the point, denoted by r, of the same dimensionality d.
- Now, we can define a potential energy $U(z)$ of the point and its kinetic energy $K(r) = r^T M^{-1} r/2$, where M is a mass matrix.
- The movement of the point in a frictionless mechanical system then may be evaluated using Newton's law.

Equations of motion

The system is described by a function of z and r - Hamiltonian:

$$H(z, r) = U(z) + K(r)$$

The partial derivatives of H determines how z and r changes over time:

$$\frac{dz^{(i)}}{dt} = \frac{\delta H}{\delta r^{(i)}}$$

$$\frac{dr^{(i)}}{dt} = -\frac{\delta H}{\delta z^{(i)}}$$

Potential and Kinetic energy

Usually the kinetic energy is chosen to be:

$$K(r) = r^T M^{-1} r/2$$

Where M is a symmetric positive-definite mass matrix. For simplisity of notation we will use $M = diag(m_i)$
For the purposes of MCMC, potential energy must be chosen as

$$p(z) = \frac{1}{Z} \exp(-U(z))$$

Equations of motion

Hence, equations of motion may be rewriten as

$$\frac{dz^{(i)}}{dt} = \frac{r^{(i)}}{m_i}$$

$$\frac{dr^{(i)}}{dt} = -\frac{\delta U}{\delta z^{(i)}}$$

Properties of Hamiltonian dynamics

- **Reversibility**
- **Conservation of the Hamiltonian**

$$\frac{dH}{dt} = \sum_{i=1}^{d} \left[\frac{dz^{(i)}}{dt} \frac{\delta H}{\delta z^{(i)}} + \frac{dr^{(i)}}{dt} \frac{\delta H}{\delta r^{(i)}} \right] = \sum_{i=1}^{d} \left[\frac{\delta H}{\delta r^{(i)}} \frac{\delta H}{\delta z^{(i)}} - \frac{\delta H}{\delta z^{(i)}} \frac{\delta H}{\delta r^{(i)}} \right] = 0$$

- **Volume preservation** since the divergency of the vector field equals to zero:

$$\sum_{i=1}^{d} \left[\frac{\delta}{\delta z^{(i)}} \frac{dz^{(i)}}{dt} + \frac{\delta}{\delta r^{(i)}} \frac{dr^{(i)}}{dt} \right] = \sum_{i=1}^{d} \left[\frac{\delta^2 H}{\delta z^{(i)} \delta r^{(i)}} - \frac{\delta^2 H}{\delta r^{(i)} \delta z^{(i)}} \right] = 0$$

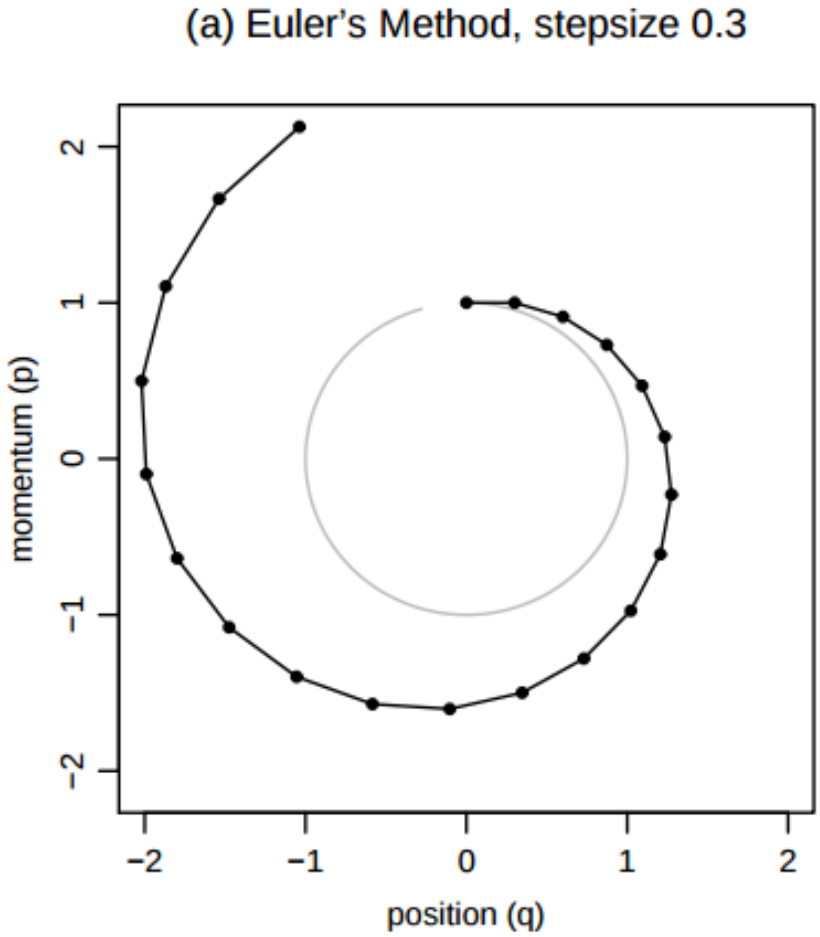
Discretization of Hamilton's equations

Euler's method:

$$r^{(i)}(t+\epsilon) = r^{(i)}(t) + \epsilon \frac{dr^{(i)}}{dt}(t) = r^{(i)}(t) - \epsilon \frac{\delta U}{\delta z^{(i)}}(z(t))$$

$$z^{(i)}(t+\epsilon) = z^{(i)}(t) + \epsilon \frac{dz^{(i)}}{dt}(t) = z^{(i)}(t) + \epsilon \frac{r^{(i)}(t)}{m_i}$$

Discretization of Hamilton's equations



Discretization of Hamilton's equations

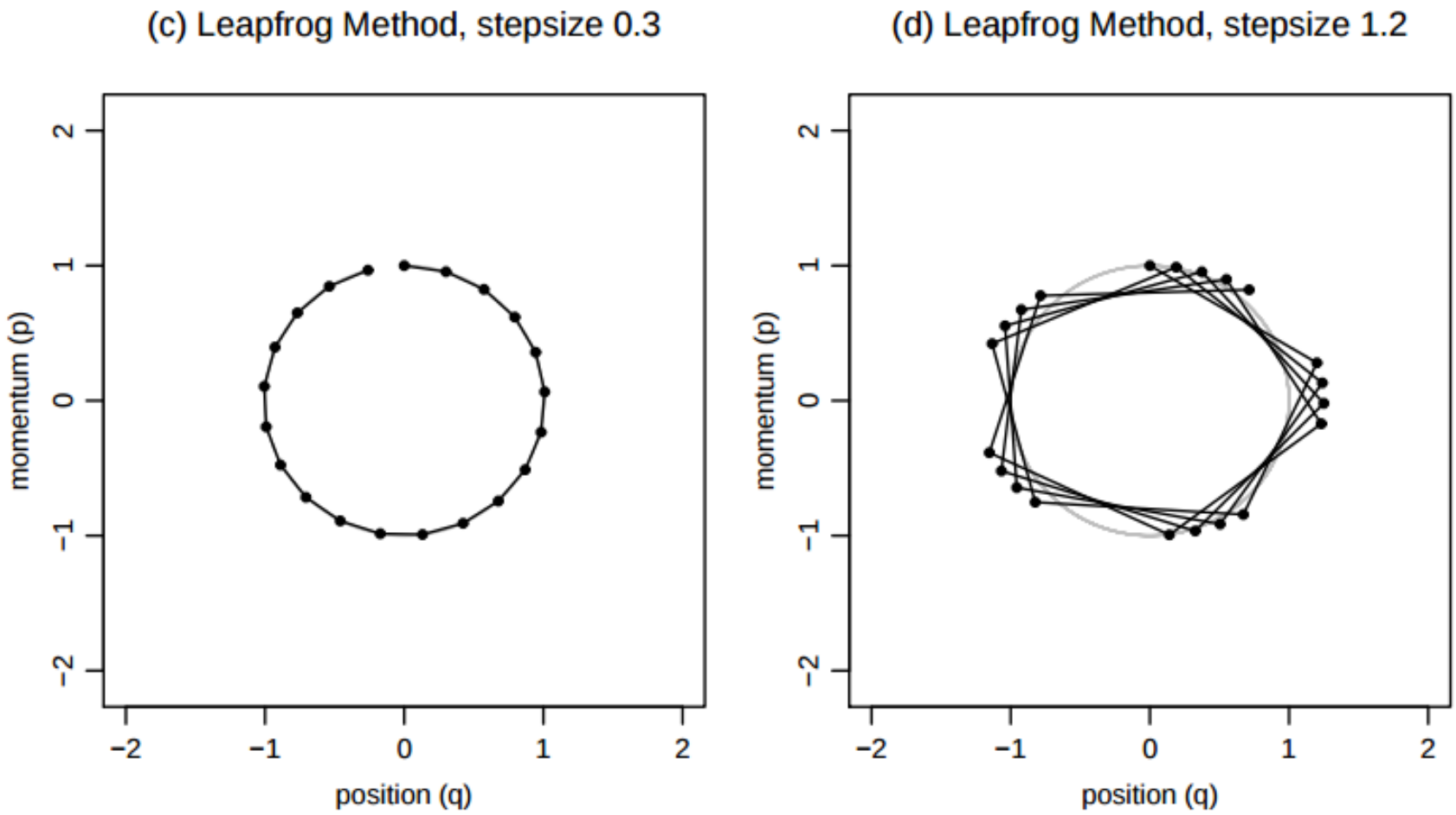
Leapfrog method:

$$r^{(i)}(t+\epsilon/2) = r^{(i)}(t) - (\epsilon/2)\frac{\delta U}{\delta z^{(i)}}(z(t))$$

$$z^{(i)}(t+\epsilon) = z^{(i)}(t) + \epsilon\frac{r^{(i)}(t+\epsilon/2)}{m_i}$$

$$r^{(i)}(t+\epsilon) = r^{(i)}(t+\epsilon/2) - (\epsilon/2)\frac{\delta U}{\delta z^{(i)}}(z(t+\epsilon))$$

Discretization of Hamilton's equations



Hamiltonian Monte Carlo

HMC samples from the distribution for z and r:

$$\frac{1}{Z}\exp(-H(z,r)) = \frac{1}{Z}\exp(-U(z))\exp(-K(r))$$

We can choose the distribution for r as we wish, while it is independent from z

Common choice is a zero-mean factorized Gaussian, such that

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

Hamiltonian Monte Carlo

HMC has three steps:

1. Sample momentum r from its distribution
2. Simulate Hamiltonian dynamics with step-size ϵ for L steps using reversible method. Then the resulted momentum is negated $r' := -r'$ for the symmetricity of a transition probability.
3. Accept new pair (z', r') with probability

$$\min(1, exp(-H(z', r') + H(z, r)))$$

All three steps leaves the distribution of (z, r) invariant

HMC holds Detailed Balance

Let's partition (z, r) space into small regions A_k of volume V.
After leapfrog simulation and negation points from A_k will move to B_k which also partition (z, r) space and has volume V.
Detailed balance holds if

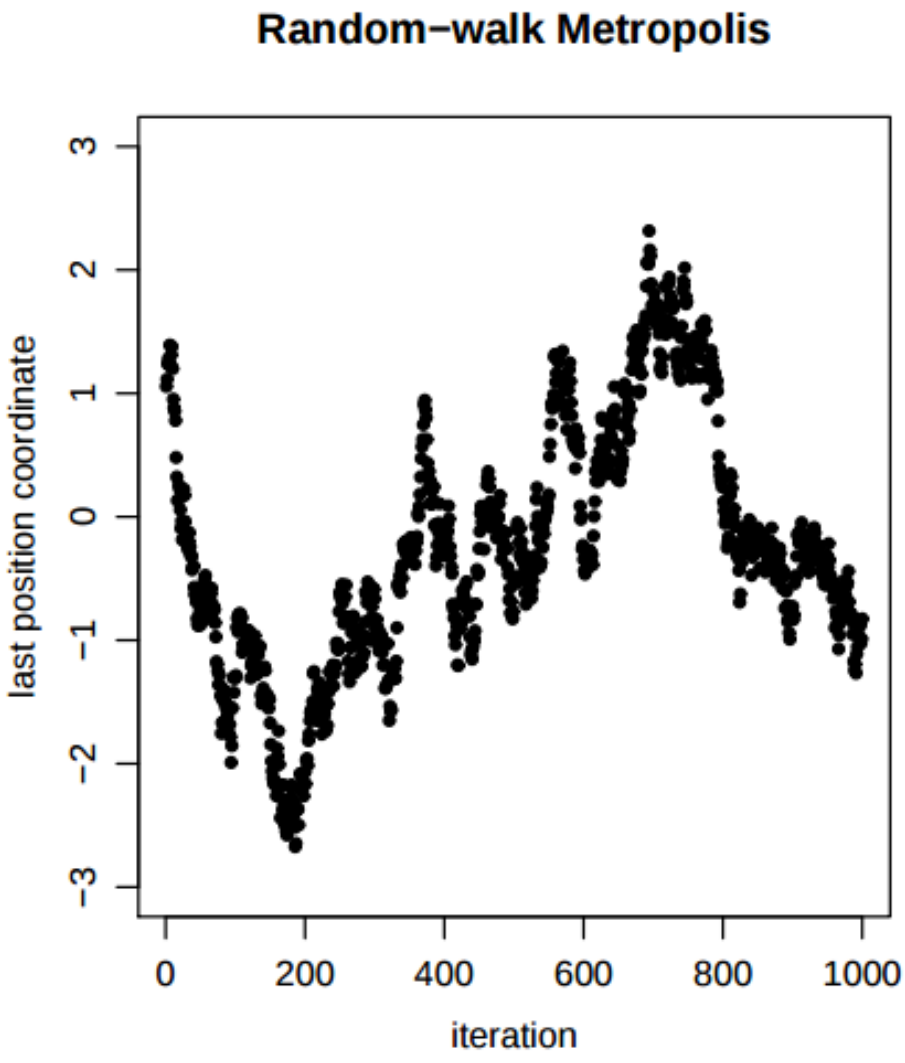
$$p(A_i)T(B_j|A_i) = p(B_j)T(A_i|B_j)$$

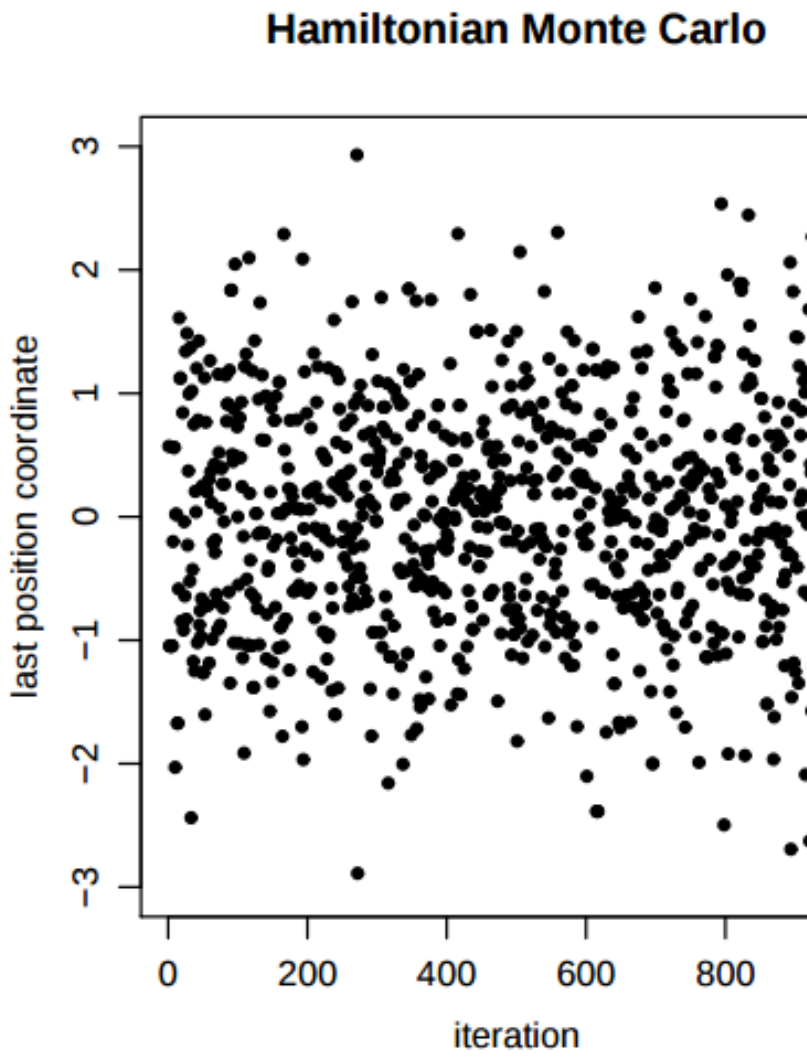
$j \neq i$ implies $T(A_i|B_j) = T(B_j|A_i) = 0$
When $i = j$, say both equal to k, detailed balance may be rewriten as

$$\frac{V}{Z}\exp(-H_{A_k})\min(1, \exp(-H_{A_k} + H_{B_k})) = \ldots$$

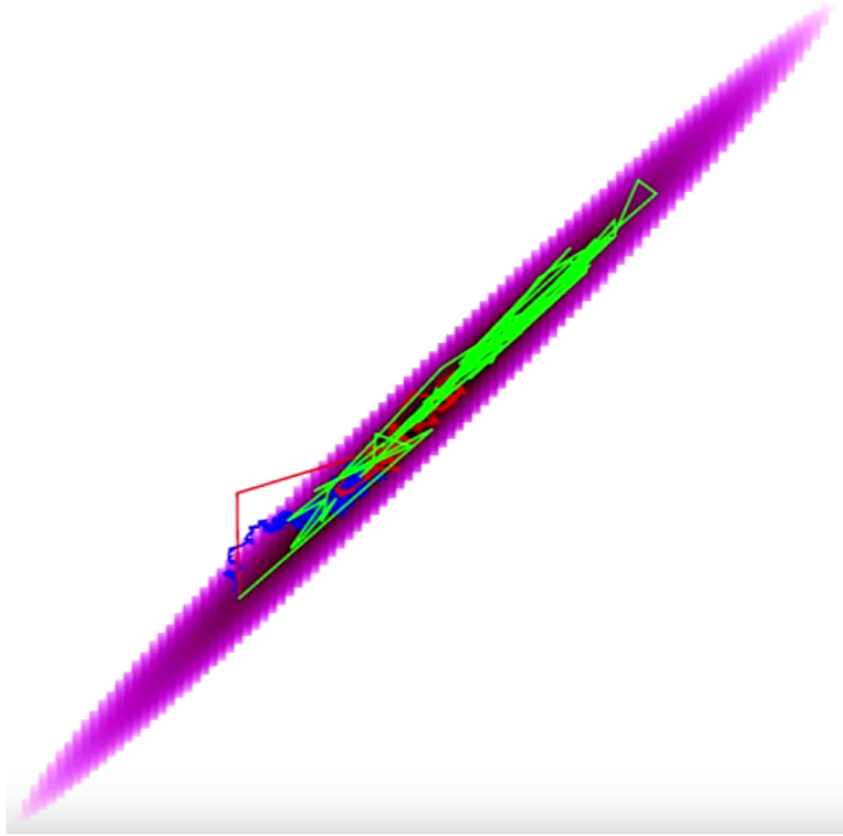
$$\ldots = \frac{V}{Z}\exp(-H_{B_k})\min(1, \exp(-H_{B_k} + H_{A_k}))$$

Benefits of the HMC





Benefits of the HMC



Tuning of the HMC

- Given the covariance Σ of the z it is usefull to choose $M = \Sigma^{-1}$
- We need preliminary runs to choose ϵ and L
- ϵ must be chosen as a maximal value, that gives us high acceptance rate.
- ϵ must be sampled for each iteration of the outer cycle to avoid periodicity.
- L must be chosen as a smalest value, that gives us uncorelated samples.
- Try to use **PyStan** library. It tunes HMC for you.

Still not Scalable? (Yes)

Let's assume that the desirable dist. is a posterior over parameters of ML model:

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

Then

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

We need full gradient to apply HMC.

Stochastic version?

Using just a mini-batch estimation of the ∇U is possible, but we still need evaluate rejection rate at Metropolis step.
Without this step HMC works purely.
However, there is an article with simular ideas:
`https://arxiv.org/pdf/1402.4102.pdf`

Langevin Monte Carlo

The special case of HMC is a Langevin MC which uses only one leapfrog iteration and mass matrix $M = I$.
It may be shown, that this is similar to just Metropolis-Hastings algorithm with proposal:

$$z' \sim \mathcal{N}\left(z' | z - (\epsilon/2)[\delta U/\delta z](z), \epsilon I\right)$$

Langevin Monte Carlo

Consider ML case, when we want to sample from true posterior over parameters of a model:

$$\nabla U = -(\sum_{i=1}^{N} \nabla \log p(x_i|\theta) + \nabla \log p(\theta)$$

If we assume that acception rate of the LMC is always equal 1, then point update will be following

$$\theta_{t+1} = \theta_t - \frac{\epsilon}{2}(\sum_{i=1}^{N} \nabla \log p(x_i|\theta_t) + \nabla \log p(\theta_t)) + \eta_t$$

$$\eta_t \sim \mathcal{N}(0, \epsilon I)$$

GD vs LMC

Gradient descent:

$$\theta_{t+1} = \theta_t - \frac{\epsilon}{2}(\sum_{i=1}^{N} \nabla \log p(x_i|\theta_t) + \nabla \log p(\theta_t))$$

Langevin MC with 100

$$\theta_{t+1} = \theta_t - \frac{\epsilon}{2}(\sum_{i=1}^{N} \nabla \log p(x_i|\theta_t) + \nabla \log p(\theta_t)) + \eta_t$$

$$\eta_t \sim \mathcal{N}(0, \epsilon I)$$

SGD vs SGLD

Stochastic gradient descent:

$$\theta_{t+1} = \theta_t - \frac{\epsilon_t}{2}\left(\frac{N}{n}\sum_{i=1}^{n} \nabla \log p(x_i|\theta_t) + \nabla \log p(\theta_t)\right)$$

Stochastic Gradient Lagevin Dynamics:

$$\theta_{t+1} = \theta_t - \frac{\epsilon}{2}\left(\frac{N}{n}\sum_{i=1}^{n} \nabla \log p(x_i|\theta_t) + \nabla \log p(\theta_t)\right) + \eta_t$$

$$\eta_t \sim \mathcal{N}(0, \epsilon_t I)$$

SGLD

- Converges if $\sum \epsilon_t \to 0$, $\sum \epsilon_t^2 \to \infty$
- Decreases overfitting
- Allow us to sample from posterior
- Works well for non-convex tusks
- Uncertainty estimation via MCMC sampling

RMSProp vs pSGLD ?

https://arxiv.org/pdf/1512.07666.pdf