Skolkovo Institute of Science and Technology

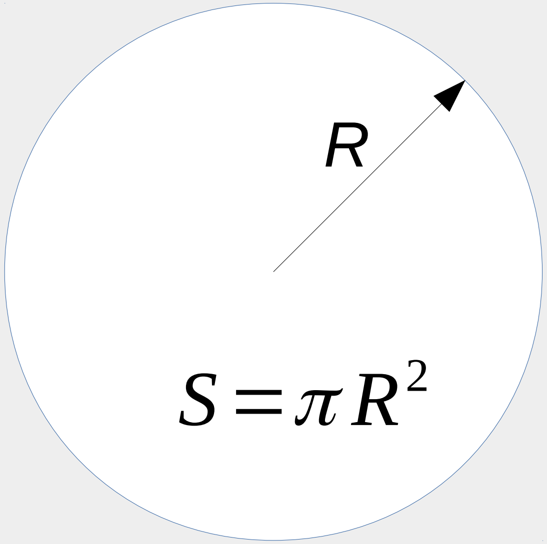
Stochastic methods in Mathematical Modelling

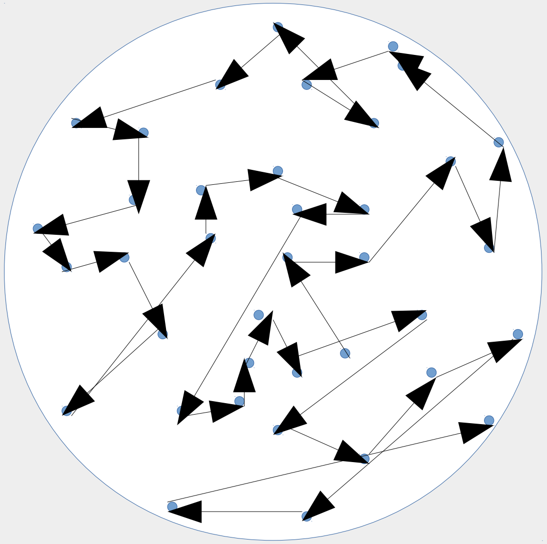
Lecture 17. Markov Chain Monte Carlo methods.

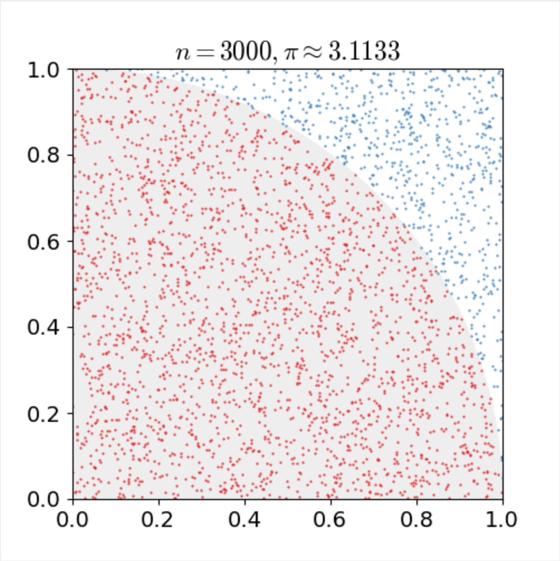
Monte Carlo. The main idea

Random sampling instead of exact solution

Alternatively we could do a random walk and sample the space in a correlated way







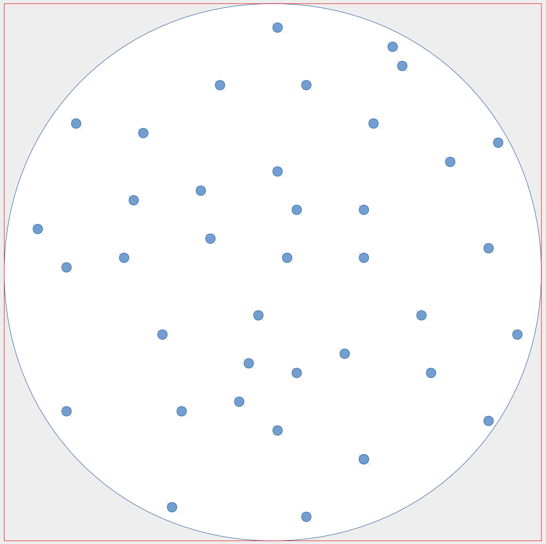
Skoltech
Skolkovo Institute of Science and Technology

Monte Carlo. Two ways

Usually the goal is to sample some complicated distribution f(x) or find E[f(x)]

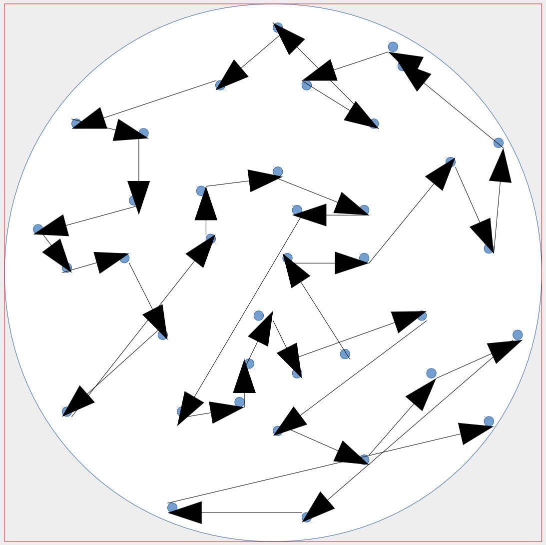
Direct sampling

Independent samples from a distribution



Markov Chain Monte Carlo (MCMC)

Draws are correlated according to a Markov chain



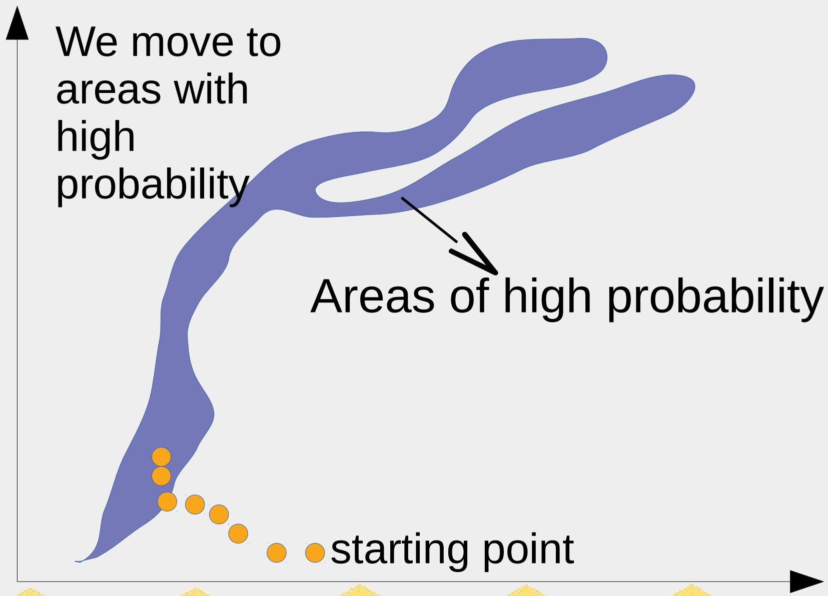
Intuitive idea of MC Monte Carlo

Skolkovo Institute of Science and Technology

Usually the goal is to sample some complicated distribution $f(x)$ or find $E_f[h(X)]$ or an integral $\int f(x)\,dx$

In low dimensional problems rejection sampling (RS) and importance sampling (IS) are sufficiently good because it is possible to find the proposal density P(x) similar to f(x). However, in high dimensional spaces, for large and complex problems it is difficult to create a single density Q(x) that has this property.

Intuitive idea of MCMC



(a) the region of high probability tends to be "connected", i.e. one can get from one point to another without going through a low-probability region

(b) we tend to be interested in the expectations of functions that are relatively smooth and have lots of "symmetries", hence need only a small number of representative points

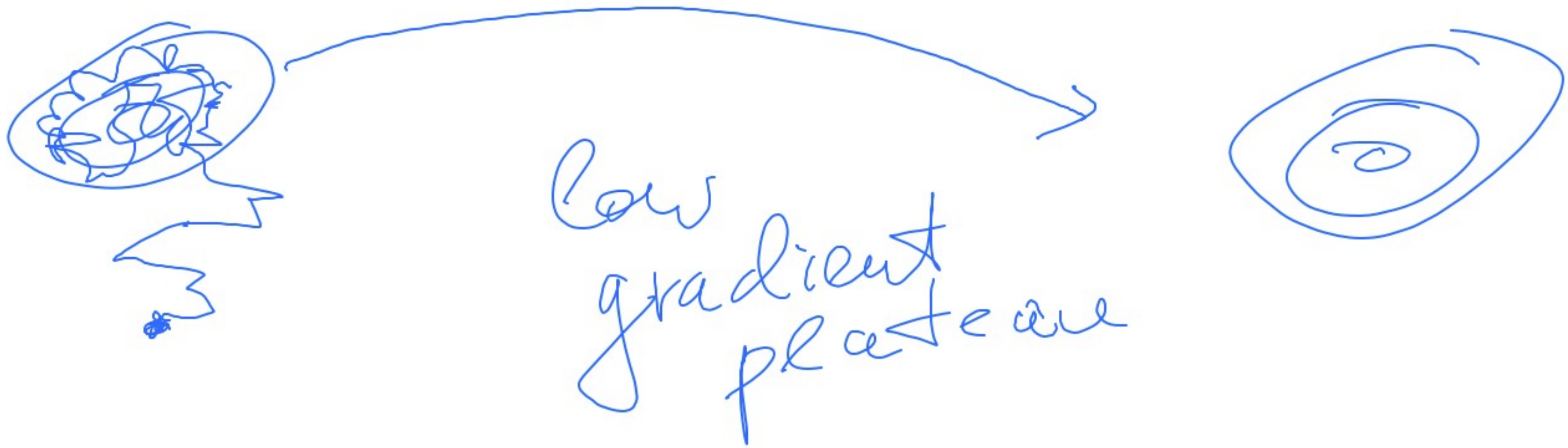
Advantages/Disadvantages of MCMC:

Advantages:

- applicable even when we can’t directly draw samples
- works for complicated distributions in high-dimensional spaces, even when we don’t know where the regions of high probability are
- relatively easy to implement
- fairly reliable

Disadvantages:

- the initial samples may follow a very different distribution, especially if the starting point is in a region of low density. A burn-in (thermalisation) period is typically necessary, where an initial number of samples (e.g. the first 1000 or so) are thrown away.
- can be very difficult to assess accuracy and evaluate convergence, even empirically
- Does not work with multimodal distributions


low
gradient
plateau

MCMC methods

- Metropolis-Hastings method
- Gibbs Sampling
- Slice sampling
- Hamiltonian (or Hybrid) Monte Carlo (HMC)
- etc.

MCMC. Metropolis–Hastings method

we start $X(0)$

our goal is to approach f $\quad X(t) \xrightarrow{t\to\infty} \underline{f}$

we want $X(t+s)$ to be independent of $X(t)$

$A = E_f[V(x)]$ → function of „interest"

$$\hat{A} = \frac{1}{L} \sum_{t=1}^{L} V(x(t))$$

If ergodicity holds for the Markov chain

$\forall X(0) \quad \hat{A} \to A$ for $L \to \infty$

Let's consider set of discrete states S:

$x_1, \dots, x_d$

$f = (f_1, \dots, f_d)$, $f_k \equiv f(X_k)$

Transition probability: $P_{xy} = P(x \leftarrow y) = P(X(t+1)=x \mid X(t)=y)$

we would like to make f to be stationary distr. of Markov chain

$$f(x) = \sum_{y \in S} f(y) P_{xy}$$

Detailed balance condition

$$f(x) P_{yx} = f(y) P_{xy} \quad \forall x, y$$

Metropolis method

1) P_{xy} have to follow detailed balance (DB)

2) Let's build RW assuming $f(x)$ are given

$$a(x) = P(x \to x+1)$$

$$b(x) = P(x \to x)$$

$$c(x) = P(x \to x-1)$$

$$a(x) + b(x) + c(x) = 1$$

$$0 \le a(x) \le 1, \quad 0 \le b(x) \le 1, \quad 0 \le c(x) \le 1 \quad \forall x$$

Let's also assume $P(d \to d+1) = 0 = a(d) = 0$

$a(x), c(x) > 0$ $\quad P(1 \to 0) = c(1) = 0$

DB condition: $\quad f(x)\, a(x) = f(x+1)\, c(x+1)$

$$f(x)\,a(x) = f(x+1)\,c(x+1)$$

Case (1) $f(x) \le f(x+1)$

$$c(x+1) = a(x)\,\frac{f(x)}{f(x+1)}$$

if $a(x) = 1$ $\qquad c(x+1) = \frac{f(x)}{f(x+1)} \le 1$

(2) $f(x) \ge f(x+1)$

$c(x+1) = 1$ $\qquad a(x) = \frac{f(x+1)}{f(x)} \le 1$

$\Downarrow$ we could find a, b, c

BUT: $a + b + c = 1 \Rightarrow$ we could get $b(x) < 0$

Hence we need to adjust it: $\longrightarrow$

Let's try to be less greedy:

$a(x) \le 1/2$, $c(x) \le 1/2$ then $\underline{b(x) \ge 0}$

(i) $f(x) \ge f(x+1)$ $\quad c(x+1) = \frac{1}{2}$ $\quad a(x) = \frac{f(x+1)}{f(x)} \cdot \frac{1}{2}$

(ii) $f(x) \le f(x+1)$ $\quad a(x) = 1/2$ $\quad c(x+1) = \frac{f(x)}{f(x+1)} \cdot \frac{1}{2}$

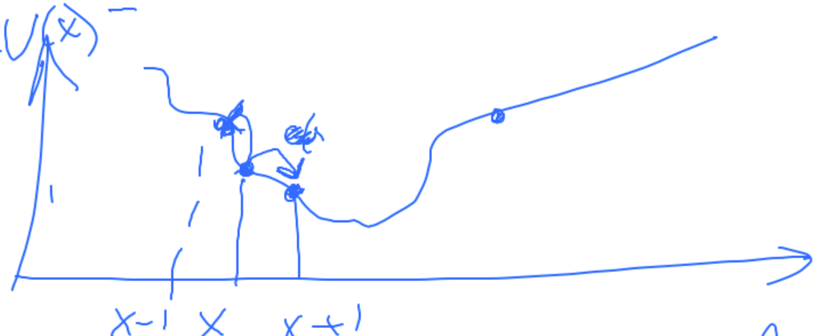
$\left.\begin{matrix} a(x) = \frac{f(x+1)}{f(x)} \cdot \frac{1}{2} \\ c(x+1) = \frac{f(x)}{f(x+1)} \cdot \frac{1}{2} \end{matrix}\right\} \le \frac{1}{2}$

the prob to stay

$a(d) = 0 = c(1)$

$b(x) = 1 - a(x) - c(x)$

$f(x) = U(x)$



(i): we propose a move with $p = \frac{1}{2}$

if $x \to x-1$ we accept it with $p = 1$

$x \to x+1$ we accept it with $p = \frac{f(x+1)}{f(x)} \le 1$

1) proposal stage

2) acceptance/rejection stage

In general: 1) $K_{xy} \geq 0$ the proposal

$$\sum_x K_{xy} = 1$$

2) The acceptance rates R_{xy} to fulfil DB $\Rightarrow$

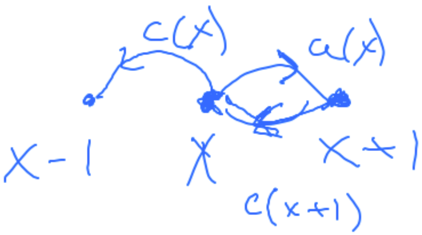
$$f(x) K_{yx} R_{yx} = f(y) K_{yx} R_{xy}$$

$$0 \leq R_{yx}, K_{yx} \leq 1$$

we would like R_{xy} to be close to 1:

$$\boxed{R_{xy} = \min\left(1, \frac{f(y) K_{yx}}{f(x) K_{xy}}\right)}$$

Convergence speed
crude estimate



$$f(x) \le f(x+1)$$

from the theory of RWs:

after T steps the distance (the number of visited states)

grows as $\sqrt{T}\,\varepsilon$ $(\sim d)$

ε – length of the move

If we have d states

$$\boxed{T \sim \left(\frac{d}{\varepsilon}\right)^2}$$

a rough estimate of the number of states to produce independent samples.

Metropolis-Hastings method

1. Initialisation:

Pick an initial state x_0

2. Iteration:

(i) Generate a random candidate state *x’* according to *K*(*x’*,*x*).

(ii) Calculate the acceptance probability $R_{x',x} = min\left(1, \frac{f(x) K_{x,x'}}{f(x') K_{x',x}}\right)$

3. Accept or reject:

(i) generate a uniform random number *u* from *U*[0,1]

if $u \leq R_{x'x}$, then accept the new state and set $x_{t+1}=x'$

if $u>R_{x',x}$, then reject the new state, and copy the old state forward $x_{t+1}=x_t$.

4. Increment: set t=t+1

Note: It is not the same as acceptance-rejection algorithm. In the latter when we *reject* a proposed value the number of samples does not increase. In MH method when the rejection happens the chain still takes a time step

Metropolis-Hastings method

3. Accept or reject:

(i) generate a uniform random number u from U[0,1]

if $u \le R_{x'x}$, then accept the new state and set $x_{t+1}=x'$

if $u>R_{x',x}$, then reject the new state, and copy the old state forward $x_{t+1}=x_t$

Comment: It is not the same as acceptance-rejection algorithm. In the latter when we *reject* a proposed value the number of samples *does not increase*. In MH method when the rejection happens the chain *still* takes a time step! In that case two consecutive states are the same

in $\frac{1}{n}\sum_{k=1}^{n} f\left(X_k\right)$ some terms can be the same

Metropolis-Hastings method

An example of slow convergence

Target distribution

$$P(x) = 1/21, x \in \{0,1,\ldots,20\}$$

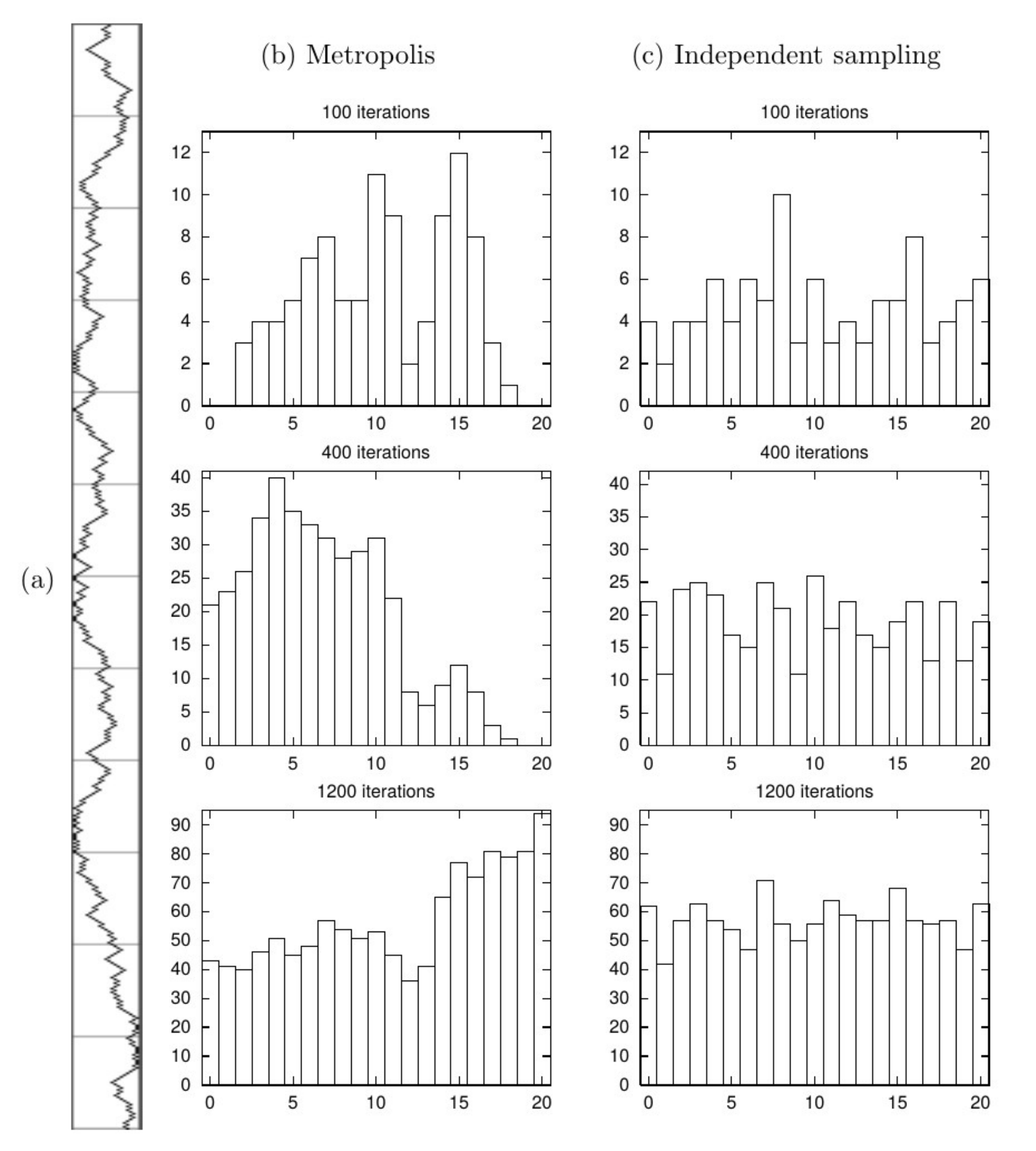
$$P(x) = 0, otherwise$$

Proposal distribution

$$K(x, x') = 1/2, x' = x \pm 1$$

$$K(x, x') = 0, otherwise$$

and rejections at the ends!



D. J. C. Mackay, Information theory, Inference, and learning algorithms. 2003

Metropolis for Ising model

$s_i = \pm 1$ $\uparrow \downarrow$; the set of spins

$$E = -\sum J_{ij} s_i s_j - \sum s_j h_j$$

Simple case: nearest-neighbour interactions & $J_{ij} = 1$

$$E = -\sum_{\langle ij \rangle} s_i s_j$$

Minimum is achieved when all spins ~~are~~ point in 1 direction

Boltzmann distr.: $P_{eq}(s) = \frac{1}{Z} e^{-\beta E(s)}$, $Z = \sum_s e^{-\beta E(s)}$

$$\beta = \frac{1}{T}$$

$T \to 0$ $(\beta \to \infty)$ $P_{eq}(s) \neq 0$ only for a few states close to minimum energy state (ground state)

$T \to \infty$ $(\beta \to 0)$ all states are equiprobable

$\uparrow\uparrow\downarrow\uparrow\uparrow\downarrow\uparrow$

$\uparrow\downarrow\uparrow\uparrow\downarrow\uparrow\uparrow$ > the same energy but diff micro states

<u>Macrostates</u>: W is a number of microstates with the same energy

$$\frac{W}{Z} e^{-\beta E} = \frac{1}{Z} e^{-\beta E + \ln W} =$$

$\ln W = S \to$ entropy

$$= \frac{1}{Z} e^{-\beta(\underbrace{E - TS}_{F})}$$

the most likely state corresponds to minimum of F

F is free energy

N spins $\uparrow\uparrow\downarrow\uparrow$

2^N combinations or states

1). Brute force approach requires $\underline{2^N}$

inefficient

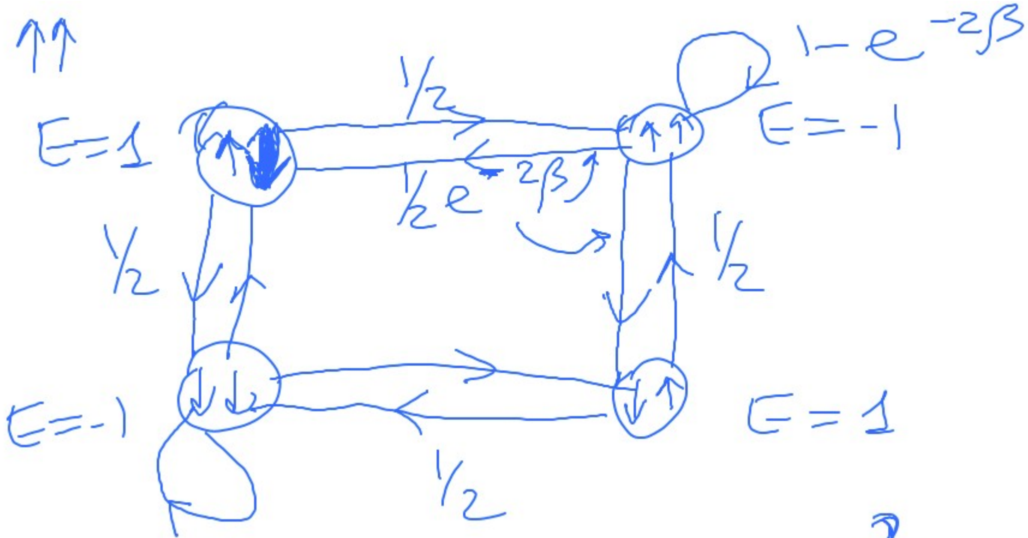
2) Metropolis-Hastings for Ising

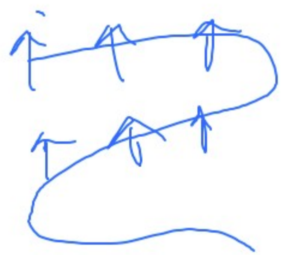
Let's construct the Markov chain on a hypercube with 2^N vertices

a) Choose a state a random

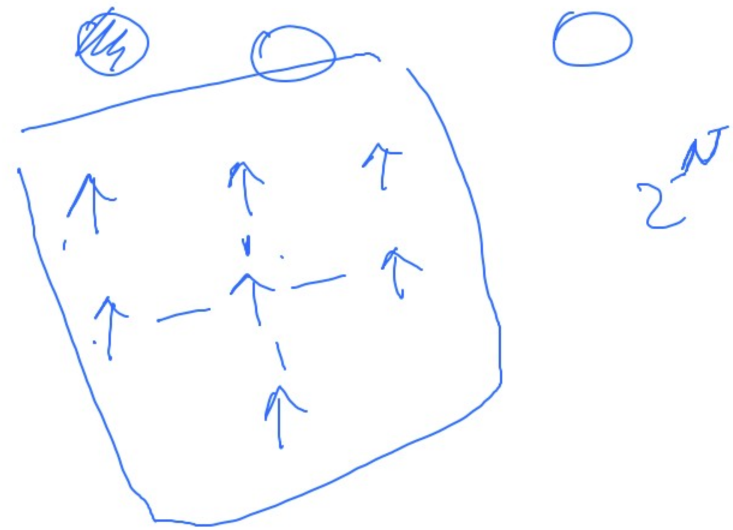
b) Compute the prob. to flip the spin as

$$P = \begin{cases} 1, & \Delta E \le 0 \\ e^{-\beta \Delta E}, & \Delta E \ge 0 \end{cases}$$



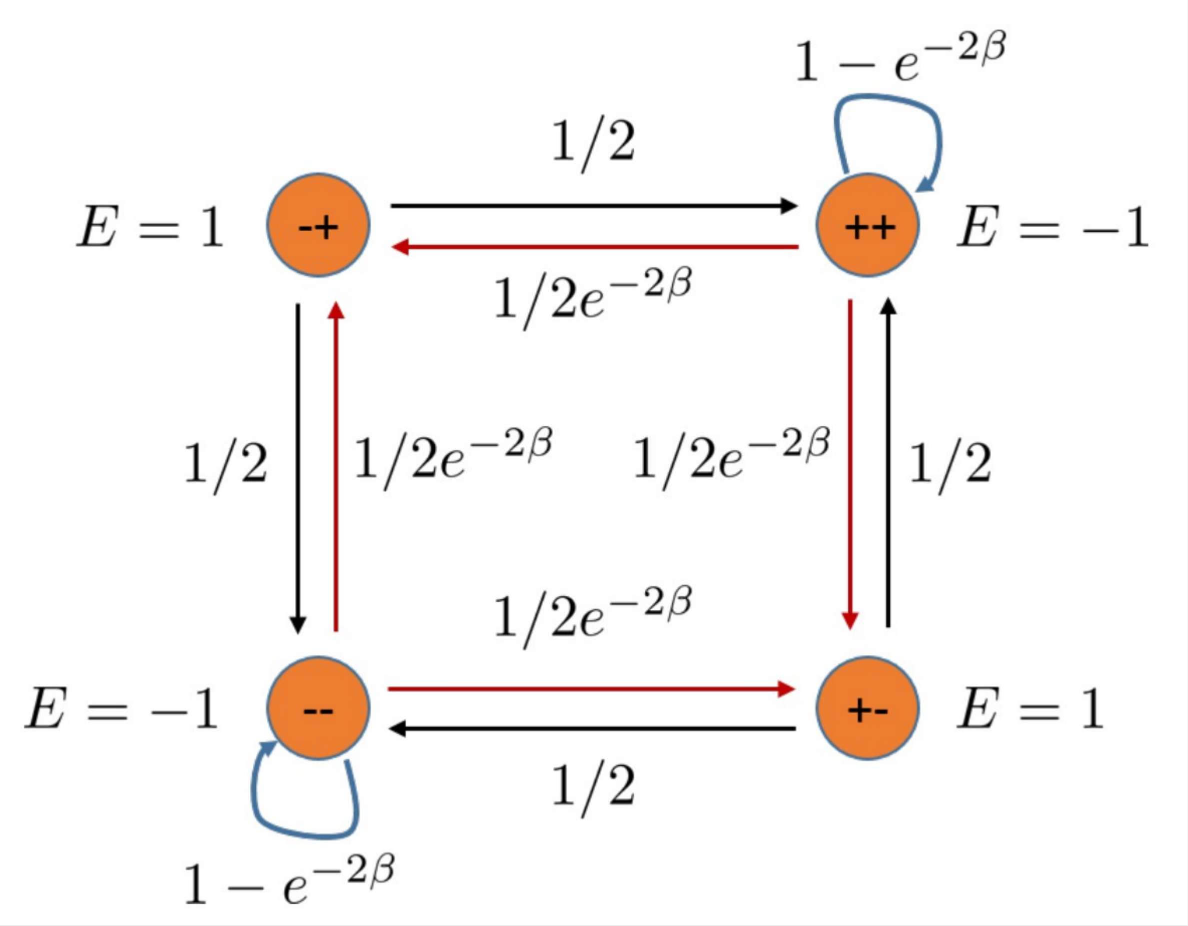


N^2 vs 2^N



Metropolis-Hastings method

Ising model. Metropolis-Hastings Markov chain example for two spins



An alternative

Gibbs (Glauber) sampling

other way to fulfil the DB

N^2

$$\begin{cases} p_{\downarrow} + p_{\uparrow} = 1 \\ \dfrac{p_{\uparrow}}{p_{\downarrow}} = e^{-\beta \Delta E} \end{cases}$$

$p_{\uparrow}$

$p_{\downarrow}$

Probability flux $\uparrow$ to $\downarrow$:

$$J_{\uparrow\downarrow} = \frac{1}{Z} e^{-\beta E(\uparrow)} p_{\downarrow}$$

$$J_{\downarrow\uparrow} = \frac{1}{Z} e^{-\beta E(\downarrow)} p_{\uparrow}$$

$$\boxed{J_{\uparrow\downarrow} = J_{\downarrow\uparrow}}$$

Gibbs sampling method

(also called Glauber dynamics if applied for Ising model)

Gibbs sampling can be viewed as a Metropolis method in which a sequence of proposal distributions *Q* are defined in terms of the conditional distributions of the joint distribution *P*(*x*). Used when the conditional distributions are easier to sample than *P*(*x*)

$$
\begin{aligned}
x_1^{(t+1)} &\sim P(x_1 \mid x_2^{(t)}, x_3^{(t)}, \ldots, x_K^{(t)}) \\
x_2^{(t+1)} &\sim P(x_2 \mid x_1^{(t+1)}, x_3^{(t)}, \ldots, x_K^{(t)}) \\
x_3^{(t+1)} &\sim P(x_3 \mid x_1^{(t+1)}, x_2^{(t+1)}, \ldots, x_K^{(t)}), \quad \text{etc.}
\end{aligned}
$$

For big problems it may be more efficient to sample groups of variables jointly, that is to use several proposal distributions

$$
\begin{aligned}
x_1^{(t+1)}, \ldots, x_a^{(t+1)} &\sim P(x_1, \ldots, x_a \mid x_{a+1}^{(t)}, \ldots, x_K^{(t)}) \\
x_{a+1}^{(t+1)}, \ldots, x_b^{(t+1)} &\sim P(x_{a+1}, \ldots, x_b \mid x_1^{(t+1)}, \ldots, x_a^{(t+1)}, x_{b+1}^{(t)}, \ldots, x_K^{(t)})
\end{aligned}
$$

Skoltech

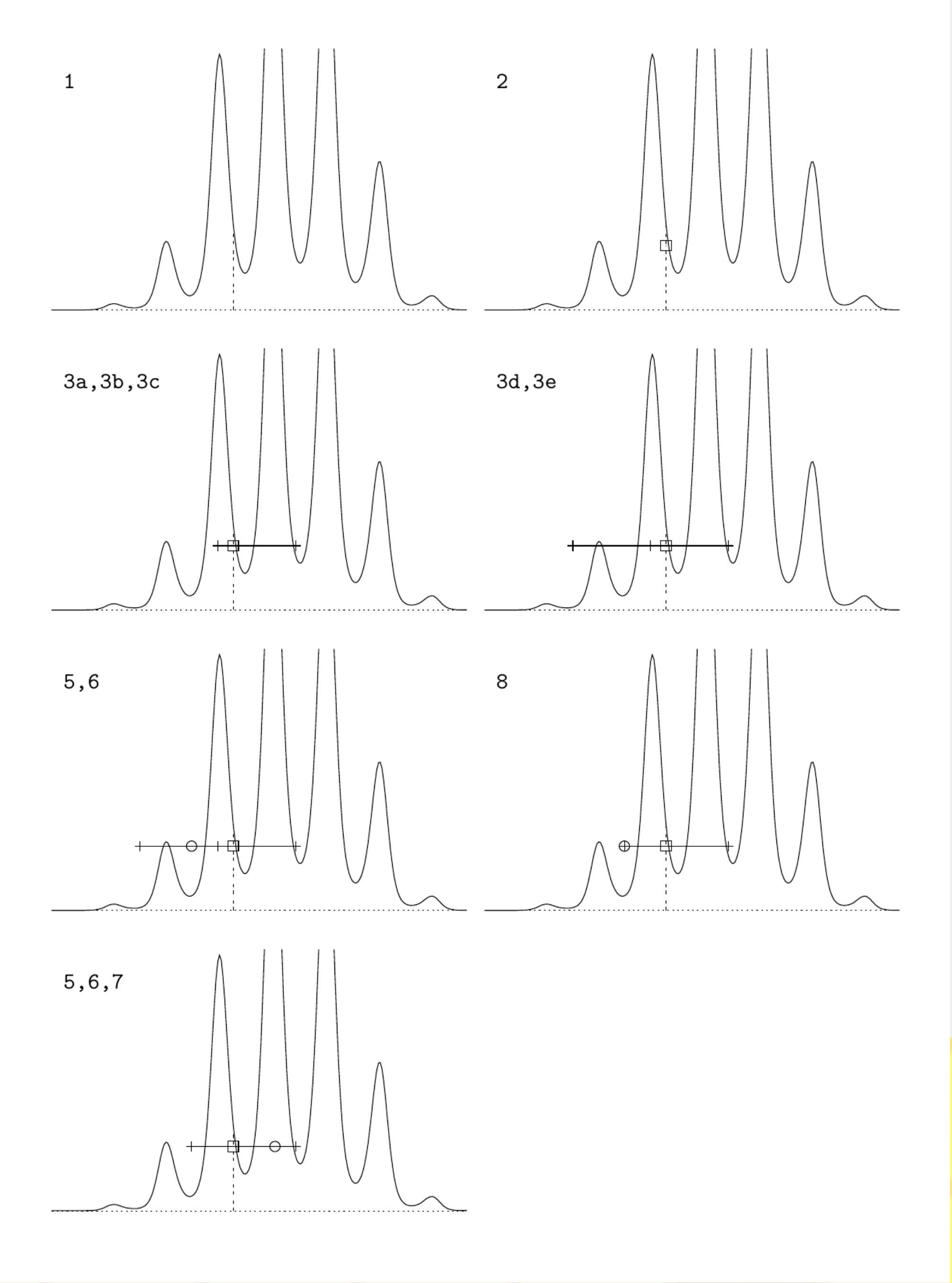
Skolkovo Institute of Science and Technology

Slice sampling method

Can be applied to any distribution
when one can evaluate P(x) at any point x

1: evaluate $P^*(x)$
2: draw a vertical coordinate $u' \sim \text{Uniform}(0, P^*(x))$
3: create a horizontal interval (x_l, x_r) enclosing x
4: loop {
5: draw $x' \sim \text{Uniform}(x_l, x_r)$
6: evaluate $P^*(x')$
7: if $P^*(x') > u'$ break out of loop 4-9
8: else modify the interval (x_l, x_r)
9: }

step size is self-tuning



D. J. C. Mackay, Information theory, Inference, and learning algorithms. 2003

Skoltech
Skolkovo Institute of Science and Technology

Hamiltonian MC method

The Hamiltonian Monte Carlo method is a Metropolis method, applicable to continuous state spaces, that makes use of gradient information to reduce sampling inefficiency of a simple random walk.

$$P(\mathbf{x}) = \frac{e^{-E(\mathbf{x})}}{Z}$$

We can also compute the gradient of E(x). It indicates the direction with states with higher probability. Add momentum variables ***p***

$$H(\mathbf{x}, \mathbf{p}) = E(\mathbf{x}) + K(\mathbf{p})$$

$$P_H(\mathbf{x}, \mathbf{p}) = \frac{1}{Z_H} \exp[-H(\mathbf{x}, \mathbf{p})] = \frac{1}{Z_H} \exp[-E(\mathbf{x})] \exp[-K(\mathbf{p})]$$

$$p(x, v) = \frac{1}{(2\pi)^{\frac{n}{2}} \sqrt{det\Sigma}} exp(-E(x) + \frac{1}{2} v^T \Sigma^{-1} v)$$

Skoltech
Skolkovo Institute of Science and Technology

Hamiltonian MC method

1. A proposal for ***p*** is obtained from a marginal distribution for momenta. This proposal is always accepted

2.
$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{p} \\ \dot{\mathbf{p}} &= -\frac{\partial E(\mathbf{x})}{\partial \mathbf{x}}. \end{aligned}$$

Because of the persistent motion of **x** in the direction of the momentum **p** during each dynamical proposal, the state of the system tends to move a distance that goes *linearly* with the computer time, rather than as the square root (ballistic vs diffusive motion).

The second proposal is accepted in accordance with the Metropolis rule.

Hamiltonian MC method

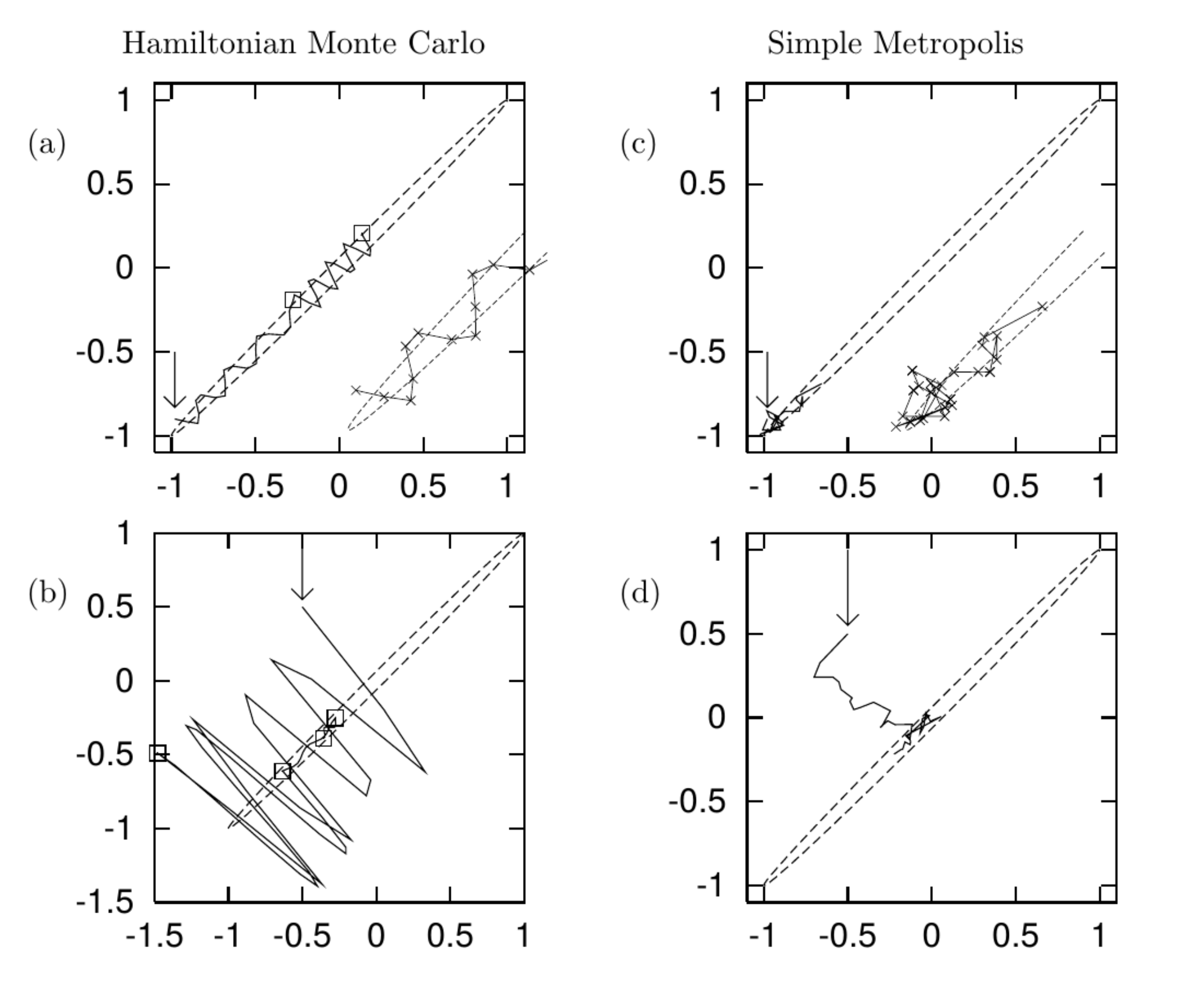


Figure 30.2. (a,b) Hamiltonian Monte Carlo used to generate samples from a bivariate Gaussian with correlation $\rho = 0.998$. (c,d) For comparison, a simple random-walk Metropolis method, given equal computer time.

D. J. C. Mackay, Information theory, Inference, and learning algorithms. 2003

Skoltech
Skolkovo Institute of Science and Technology

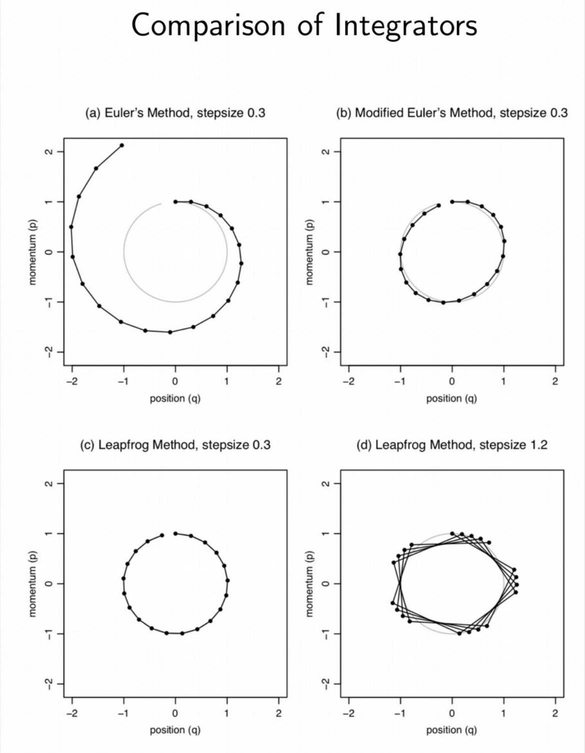
Hamiltonian MC method

- Euler Integrator:

$$\begin{cases} v(t+\epsilon) = \quad p(t) - \epsilon\frac{\partial H}{\partial q(t)} \\ q(t+\epsilon) = q(t) - \epsilon\frac{\partial H}{\partial v(t+\epsilon)} \end{cases}$$

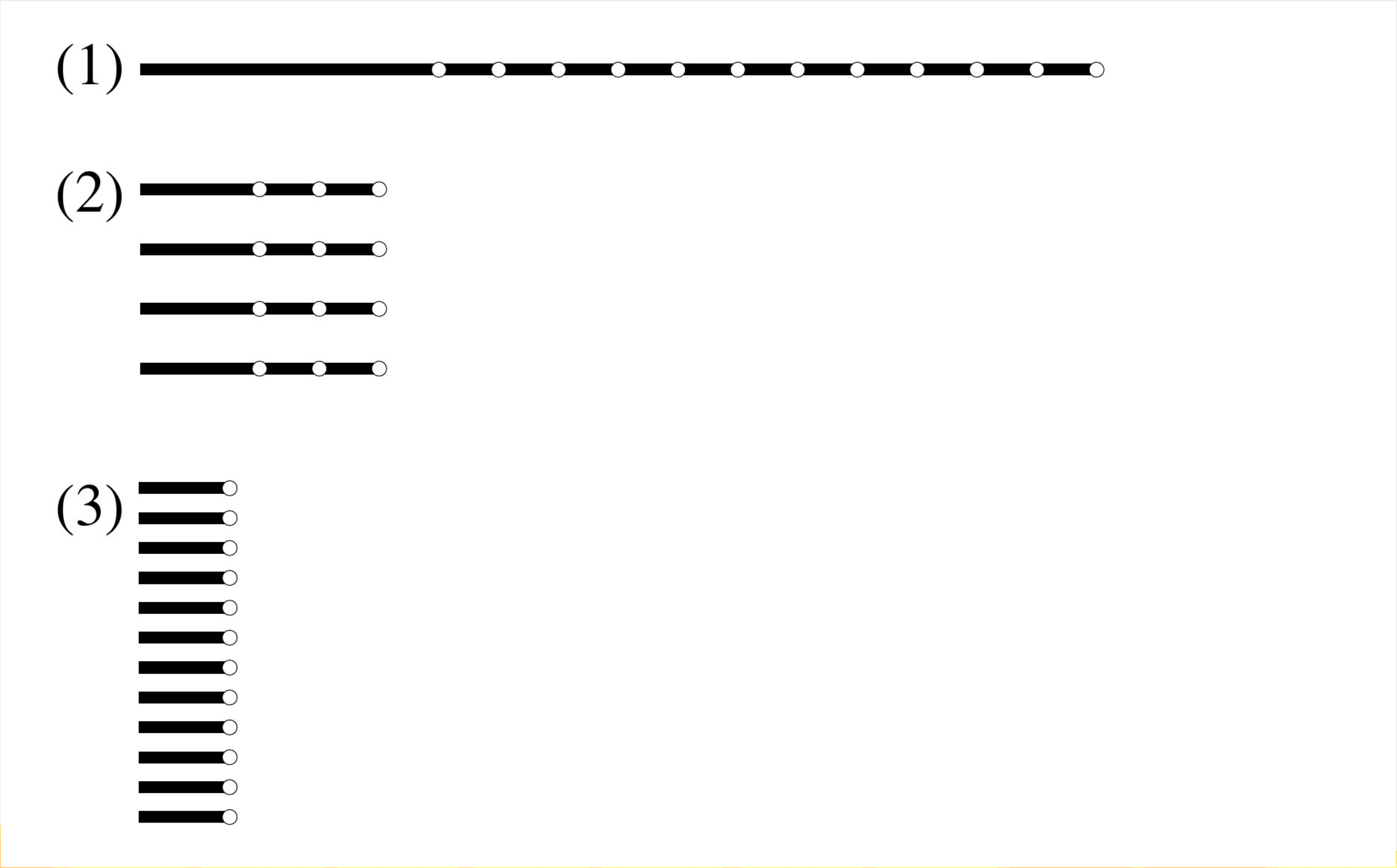
- Leapfrog :

$$\begin{cases} v(t+\frac{\epsilon}{2}) = v(t) - \frac{\epsilon}{2}\frac{\partial H}{\partial q(t)} \\ q(t+\epsilon) = q(t) - \epsilon\frac{\partial H}{\partial v(t+\frac{\epsilon}{2})} \\ v(t+\epsilon) = v(t+\frac{\epsilon}{2}) - \frac{\epsilon}{2}\frac{\partial H}{\partial q(t+\epsilon)} \end{cases}$$



- Problems:
  - Approximate Hamiltonian
  - Discretization error
  - one mode exploration
  - slow mixing
  - slow burn-in
  - not mixing across levels
  - geometric of space

Possible strategies of MCMC



Error sources of MCMC:

1) Burn-in period

2) The variance for correlated samples

$$\mathrm{cov}(X,Y) \equiv E[XY] - E[X]E[Y] = E[(X-\mu_x)(Y-\mu_y)]$$

$$\mathrm{var}\left(\sum_{i=1}^{N} Y_i\right) = \frac{1}{N^2}\sum_{i,j=1}^{N} \mathrm{cov}(x_i, x_j)$$

we get into stationary state & cov depends $|i-j|$

$$\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} f(x_n)$$

$$\mathrm{Var}(\hat{\mu}) = \frac{1}{N^2}\sum_{k,n=1}^{N} \mathrm{cov}(f(x_k), f(x_n))$$

If N is large $\mathrm{var}(f(x_n))$ should be close to $\underline{\sigma_f^2}$

$$\operatorname{cov}(f(X_k), f(X_n)) = [\operatorname{var}(f(X_k)) \operatorname{var}(f(X_n))]^{\frac{1}{2}} \operatorname{corr}(f(X_k) f(X_n))$$

$$\simeq \sigma^2(f) \operatorname{corr}(f(X_k) f(X_n))$$

$$\operatorname{Var}(\hat{\mu}) = \frac{\sigma^2(f)}{N^2} \sum_{k=1}^{N} \sum_{n=1}^{N} \operatorname{corr}(f(X_k) f(X_n))$$

$$\sum_{n=1}^{\infty} \operatorname{corr}(f(X_k) f(X_n)) \simeq \sum_{n=-\infty}^{n=+\infty} \operatorname{corr}(f(X_k) f(X_{k+n})) =$$

$$= 1 + 2 \sum_{n=1}^{\infty} \operatorname{corr}(f(X_k) f(X_{k+n})) = \tau(f)$$

autocorr time

$$\boxed{\operatorname{Var}(\hat{\mu}) \simeq \frac{\sigma^2(f)}{N^2} \sum_{k=1}^{N} \tau(f) = \frac{\sigma^2(f)\, \tau(f)}{N}} \qquad \boxed{\frac{N}{\tau(f)}}$$

size of a batch

Skoltech
Skolkovo Institute of Science and Technology

Convergence

Independence test

$$\lim_{N\to+\infty} \frac{1}{N} \sum_{n=0}^{N} \sum_{n=1}^{N} f(x_n) g(x_{n-1}) \to \mathbb{E}\left[f(x)\right] \mathbb{E}\left[g(x)\right]$$

Asymptotic convergence

$$\lim_{N\to+\infty} \frac{1}{N} \sum_{n=0}^{N} \sum_{n=1}^{N} f(x_n) \to \mathbb{E}\left[f(x)\right]$$

Parametric convergence

The target estimate is reached only in a special limit with respect to a special parameter

$$\lim_{s\to s_*} \lim_{N\to+\infty} \frac{1}{N} \sum_{n=0}^{N} \sum_{n=1}^{N} f_s(x_n) \to \mathbb{E}\left[f_{s_*}(x)\right]$$

Persi Diaconis, Journal of Statistical Physics, 144, 445 (2011),
http://statweb.stanford.edu/~cgates/PERSI/papers/mixing.pdf

Literature

1. D. J. C. Mackay, Information theory, inference, and learning algorithms. Cambridge: Cambridge University Press, 2003. http://www.inference.org.uk/itprnn/book.html
2. https://math.nyu.edu/faculty/goodman/teaching/
3. https://people.eecs.berkeley.edu/~jordan/courses/260-spring10/
4. https://www.math.arizona.edu/~tgk/mc/book.pdf