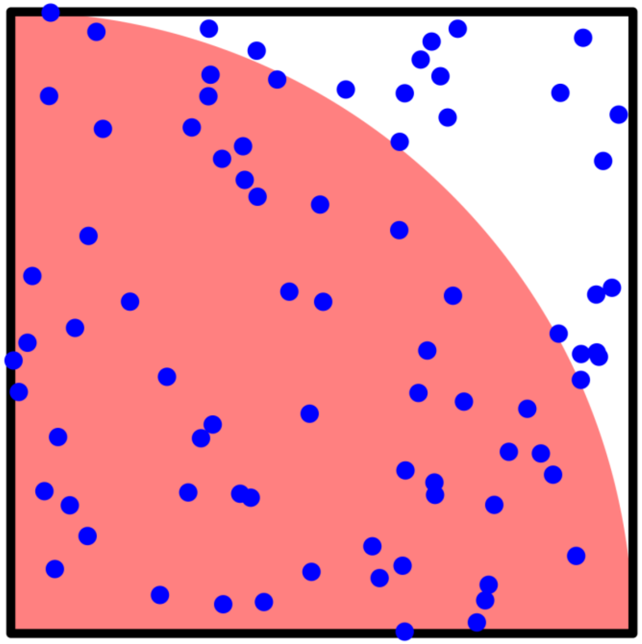
The 9th HuaDa QCD School

Sampling by MC and MCMC

Hao Wu (吴昊)
hwu@tongji.edu.cn

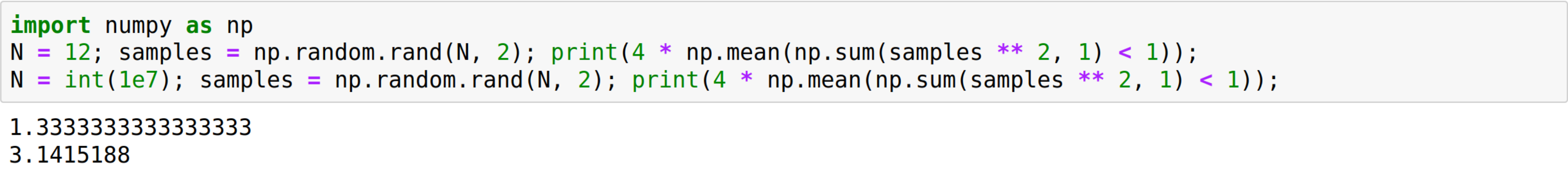
13-Oct-21

A dumb approximation of π



$$P(x,y) = \begin{cases} 1 & 0<x<1 \text{ and } 0<y<1 \\ 0 & \text{otherwise} \end{cases}$$

$$\pi = 4 \iint \mathbb{I}\left((x^2+y^2) < 1\right) P(x,y) \, \mathrm{d}x \, \mathrm{d}y$$



```python
import numpy as np
N = 12; samples = np.random.rand(N, 2); print(4 * np.mean(np.sum(samples ** 2, 1) < 1));
N = int(1e7); samples = np.random.rand(N, 2); print(4 * np.mean(np.sum(samples ** 2, 1) < 1));
```

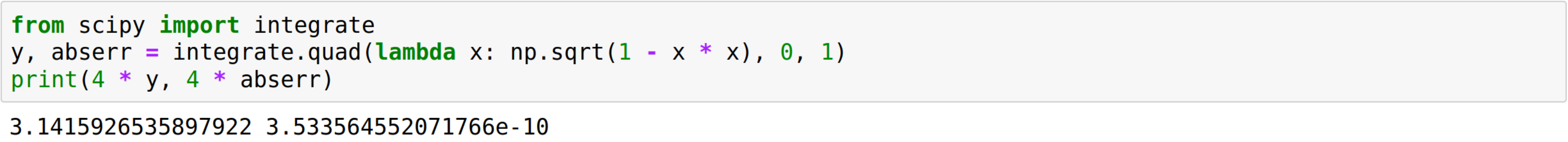
```
1.3333333333333333
3.1415188
```

Why sampling?

“*Monte Carlo is an extremely bad method; it should be used only when all alternative methods are worse.*”

— Alan Sokal, 1996

Example: numerical solutions to (nice) 1D integrals are fast.



```python
from scipy import integrate
y, abserr = integrate.quad(lambda x: np.sqrt(1 - x * x), 0, 1)
print(4 * y, 4 * abserr)
```

```
3.1415926535897922 3.533564552071766e-10
```

Numerical analysis lecturers are covering alternatives for higher dimensions.

But, no approx. integration method always works. Sometimes Monte Carlo is the best.

Probabilistic approximation taxonomy

Many problems of interest in probabilistic approximation can be written as an integral of type:

$$\int f(x,y)\mathrm{d}x$$

Examples:

- Free energy: $$-\log \int e^{-U(x|y)}\mathrm{d}x$$
- Thermodynamics/posterior expectations: $$\int f(x)P(x|y)\mathrm{d}x$$
- Evidence and model selection: $$\int P(y|x)P(x)\mathrm{d}x$$
- Prediction: $$\int P(y_{\mathrm{future}}|x)P(x|y_{\mathrm{past}})\mathrm{d}x$$

In practice, these integrals can rarely be evaluated exactly.

Probabilistic approximation taxonomy

$$\int f(x,y)\mathrm{d}x \approx \sum_i w_i f(x^i, y)$$

- Replace hard integrals with **summations**.
- Sampling methods
- Central problem: how to sample x^i
- Monte Carlo, MCMC, Gibbs, etc.

Probabilistic approximation taxonomy

$$\int f(x,y)\mathrm{d}x \approx \int g(x,y)\mathrm{d}x$$

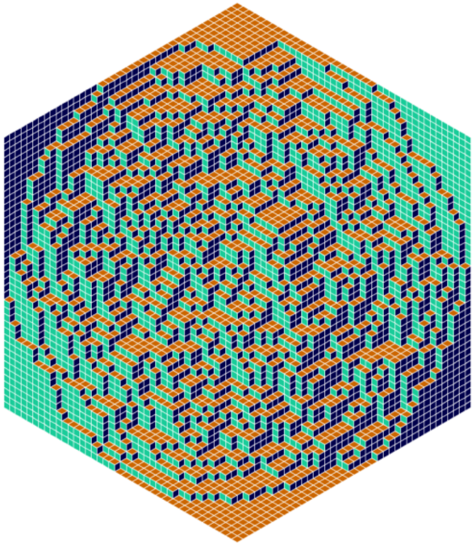
- Replace hard integrals with **easier integrals**.
- Message passing on factor graph
- Central problem: how to find $g \in \mathcal{G}$
- VB, EP, etc.

Probabilistic approximation taxonomy

$$\int f(x, y)\mathrm{d}x \approx \int h(x, y; x^*)\mathrm{d}x$$

- Replace hard integrals with **estimators**.
- "Non-Bayesian" methods
- Central problem: how to find x^*
- MAP, ML, Laplace, etc.

Eye-balling samples



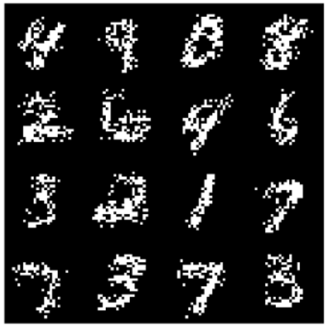
Sometimes samples are pleasing to look at:

(if you're into geometrical combinatorics)

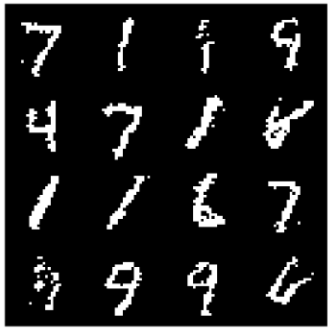
Figure by Propp and Wilson. Source: MacKay textbook.

Sanity check probabilistic modelling assumptions:

Data samples



MoB samples



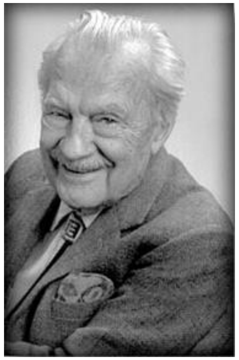
RBM samples

A brief history

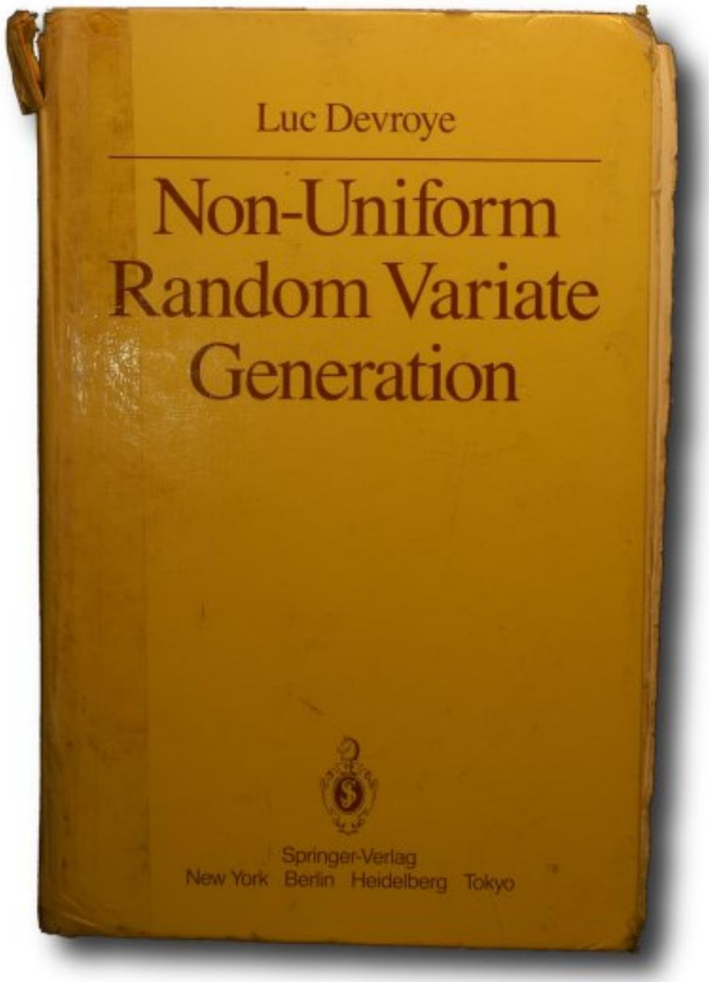
Buffon (1707 - 1788): Needle problem.

Enrico Fermi (1901 - 1954): Monte Carlo method for neutron diffusion

Stanisław Ulam (1909 - 1984), John von Neumann (1903 - 1957), Nicholas Metropolis (1915 - 1999): Markov Chain Monte Carlo (MCMC)



Sampling from distributions


Luc Devroye
Non-Uniform Random Variate Generation
Springer-Verlag
New York Berlin Heidelberg Tokyo

Use library routines for univariate distributions
(and some other special cases)

This book (free online) explains how some of them work

Sampling from distributions

How to convert samples from a Uniform[0,1] generator:

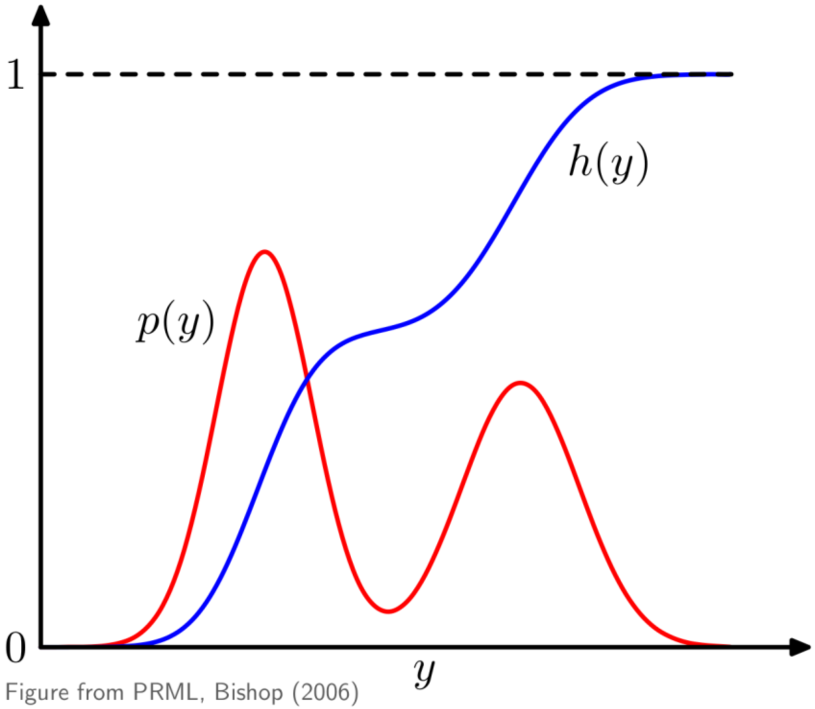


Figure from PRML, Bishop (2006)

$$h(y) = \int_{-\infty}^{y} p(y')\,\mathrm{d}y'$$

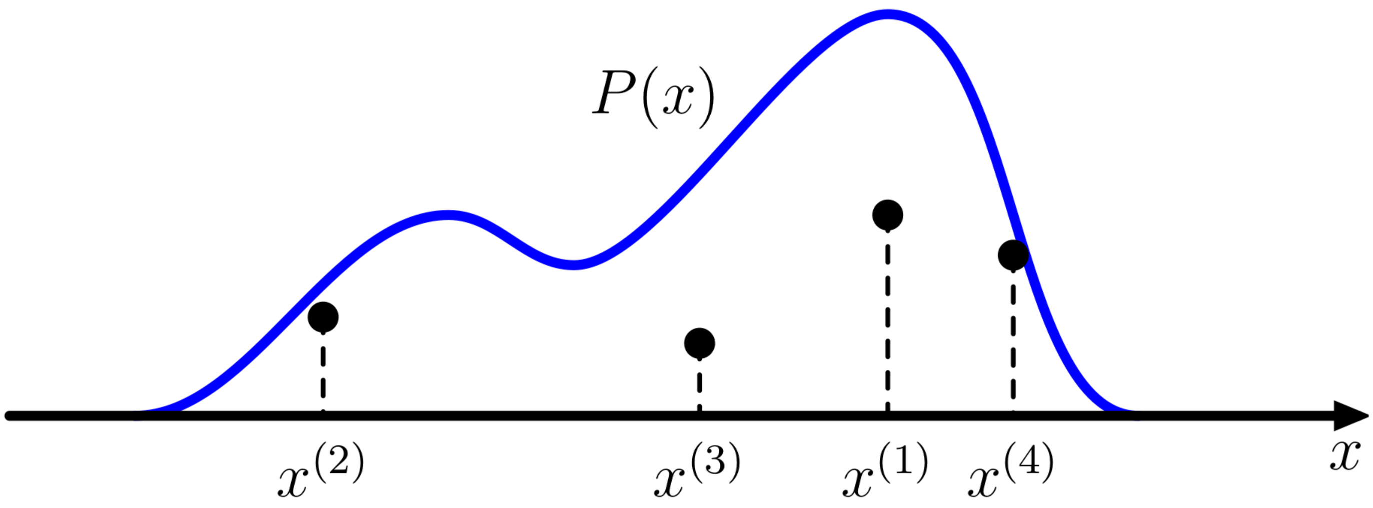
Draw mass to left of point:
$u \sim$ Uniform[0,1]

Sample, $y(u) = h^{-1}(u)$

Although we can't always compute and invert $h(y)$

Sampling from distributions

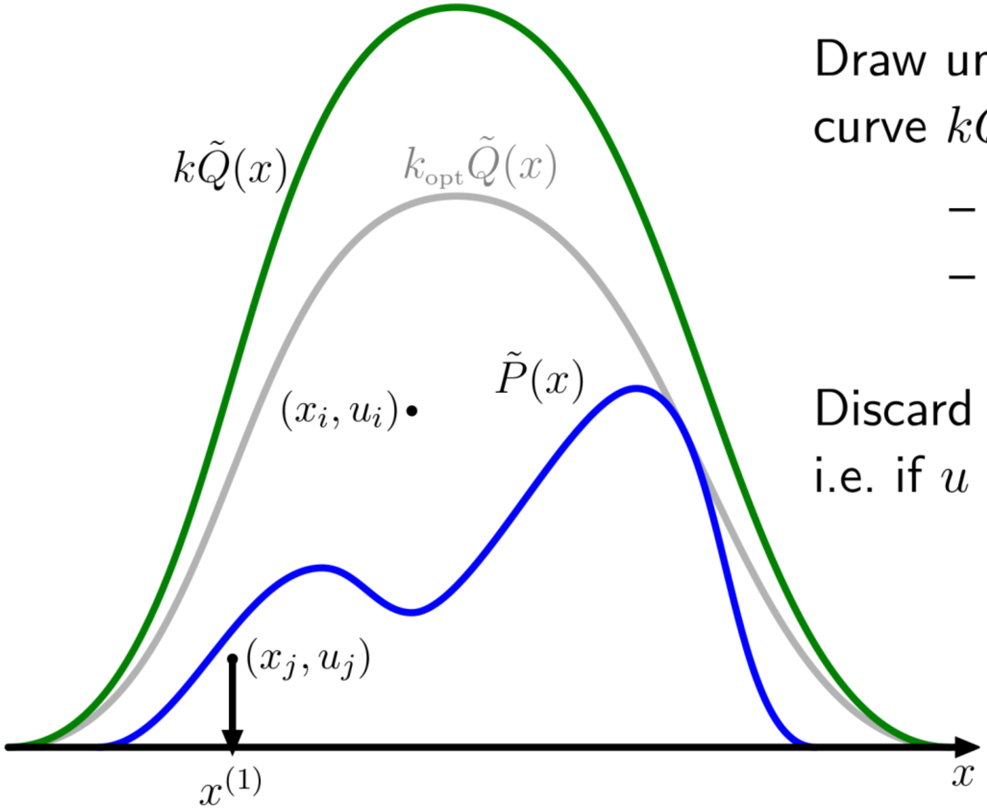
Draw points uniformly under the curve:



Probability mass to left of point $\sim$ Uniform[0,1]

Rejection sampling

Sampling underneath a $\tilde{P}(x) \propto P(x)$ curve is also valid



Draw underneath a simple curve $k\tilde{Q}(x) \geq \tilde{P}(x)$:

- Draw $x \sim Q(x)$
- height $u \sim \mathrm{Uniform}[0, k\tilde{Q}(x)]$

Discard the point if above $\tilde{P}$, i.e. if $u > \tilde{P}(x)$

Importance sampling

Computing $\tilde{P}(x)$ and $\tilde{Q}(x)$, then *throwing* x *away* seems wasteful
Instead rewrite the integral as an expectation under Q:

$$\begin{aligned}\int f(x)P(x)\,\mathrm{d}x &= \int f(x)\frac{P(x)}{Q(x)}Q(x)\,\mathrm{d}x, \qquad (Q(x)>0 \text{ if } P(x)>0)\\ &\approx \frac{1}{S}\sum_{s=1}^{S} f(x^{(s)})\frac{P(x^{(s)})}{Q(x^{(s)})}, \quad x^{(s)}\sim Q(x)\end{aligned}$$

This is just simple Monte Carlo again, so it is unbiased.

Importance sampling applies when the integral is not an expectation.
Divide and multiply any integrand by a convenient distribution.

Importance sampling

Previous slide assumed we could evaluate $P(x) = \tilde{P}(x)/\mathcal{Z}_P$

$$\int f(x)P(x)\,\mathrm{d}x \approx \frac{\mathcal{Z}_Q}{\mathcal{Z}_P}\frac{1}{S}\sum_{s=1}^{S} f(x^{(s)})\underbrace{\frac{\tilde{P}(x^{(s)})}{\tilde{Q}(x^{(s)})}}_{\tilde{r}^{(s)}}, \quad x^{(s)} \sim Q(x)$$

$$\approx \cancel{\frac{1}{S}}\sum_{s=1}^{S} f(x^{(s)})\frac{\tilde{r}^{(s)}}{\cancel{\frac{1}{S}}\sum_{s'}\tilde{r}^{(s')}} \equiv \sum_{s=1}^{S} f(x^{(s)})w^{(s)}$$

This estimator is **consistent** but **biased**

Exercise: Prove that $\mathcal{Z}_P/\mathcal{Z}_Q \approx \frac{1}{S}\sum_s \tilde{r}^{(s)}$ (which leads to the Free Energy Perturbation).

Summary so far

- Sums and integrals, often expectations, occur frequently in statistics

- **Monte Carlo** approximates expectations with a sample average

- **Rejection** sampling draws samples from complex distributions

- **Importance** sampling applies Monte Carlo to ‘any’ sum/integral

Application to large problems

Rejection & importance sampling scale badly with dimensionality:

Example:

$$P(x) = \mathcal{N}(0,\, \mathbb{I}), \quad Q(x) = \mathcal{N}(0,\, \sigma^2 \mathbb{I})$$

Rejection sampling:

Requires $\sigma \geq 1$. Fraction of proposals accepted $= \sigma^{-D}$

Importance sampling:

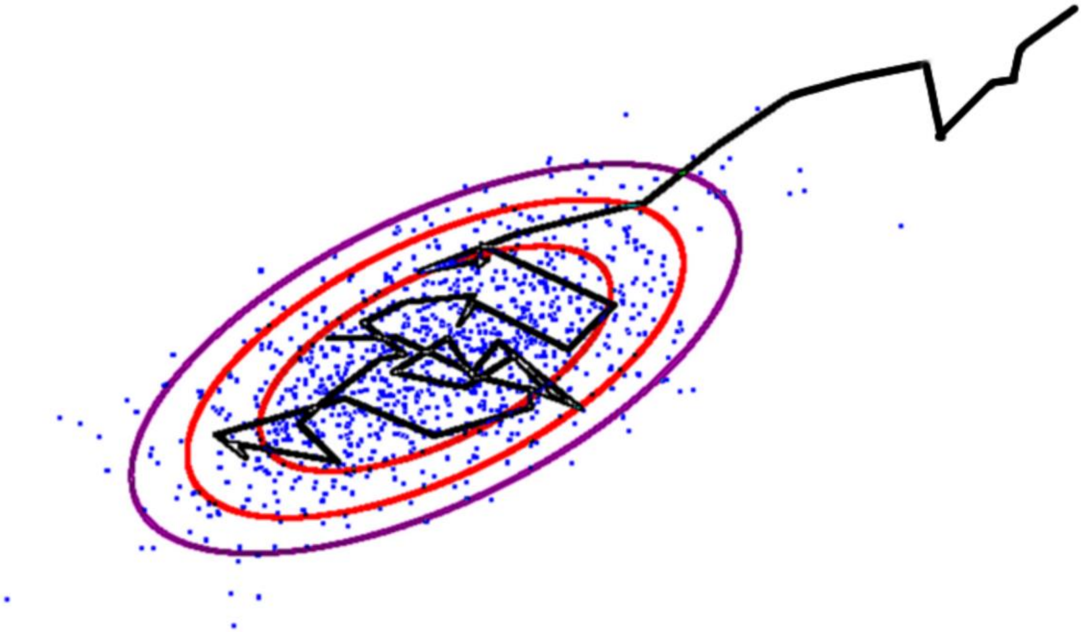
Variance of importance weights $= \left(\frac{\sigma^2}{2-1/\sigma^2}\right)^{D/2} - 1$

Infinite / undefined variance if $\sigma \leq 1/\sqrt{2}$

Markov chain Monte Carlo

Construct a biased random walk that explores target dist $P^*(x)$

Markov steps, $x_t \sim T(x_t \leftarrow x_{t-1})$



MCMC gives approximate, correlated samples from $P^\star(x)$

Transfer operators

Discrete example

$$P^* = \begin{pmatrix} 3/5 \\ 1/5 \\ 1/5 \end{pmatrix}^\top, \qquad T = \begin{pmatrix} 2/3 & 1/6 & 1/6 \\ 1/2 & 0 & 1/2 \\ 1/2 & 1/2 & 0 \end{pmatrix}, \qquad T_{ij} = T(x_j \leftarrow x_i)$$

P^* is an invariant distribution of T because $P^*T = P^*$, i.e.

$$\sum_x T(x' \leftarrow x) P^*(x) = P^*(x')$$

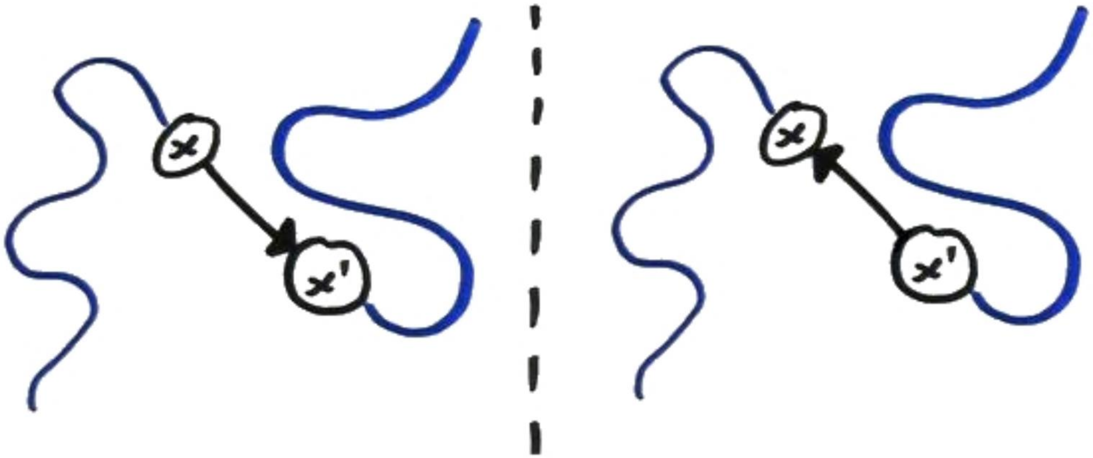
Also P^* is the *equilibrium distribution* of T:

$$(1, 0, 0)T^{100} = (3/5, 1/5, 1/5) = P^* (\text{to machine precision})$$

Ergodicity requires: Elements of P^*, T^K are positive for some K.

Detailed balance

Detailed balance means → x → x' and → x' → x are equally probable:



$$T(x' \leftarrow x)P^{\star}(x) = T(x \leftarrow x')P^{\star}(x')$$

Exercise: Prove detailed balance wrt P^* ⇒ P^* is the equilibrium distribution of T

Enforcing detailed balance is easy: it only involves isolated pairs

Metropolis–Hastings

Transfer operator:

- Propose a move from the current state $Q(x';x)$, e.g. $\mathcal{N}(x, \sigma^2)$
- Accept with probability $\min\left(1, \frac{P(x')Q(x;x')}{P(x)Q(x';x)}\right)$
- Otherwise next state in chain is a copy of current state

Notes:

- Can use $\tilde{P} \propto P(x)$; normalizer cancels in acceptance ratio
- Satisfies detailed balance (Exercise: Prove this.)
- Q must be chosen to fulfill the other technical requirements

Solution

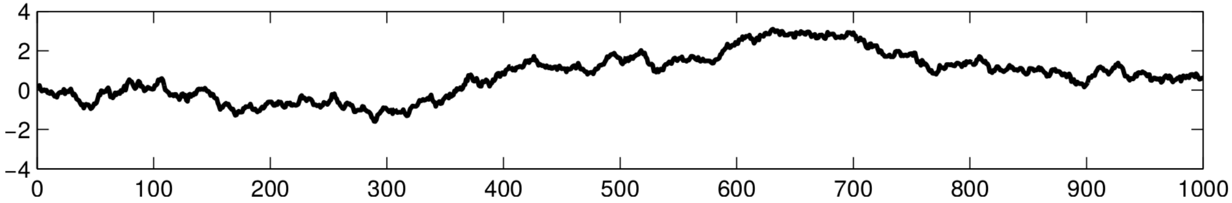
$$
\begin{aligned}
P(x)\cdot T(x' \leftarrow x) = P(x)\cdot Q(x';x)\min\left(1,\ \frac{P(x')Q(x;x')}{P(x)Q(x';x)}\right) &= \min\Big(P(x)Q(x';x),\ P(x')Q(x;x')\Big)\\
&= P(x')\cdot Q(x;x')\min\left(1,\ \frac{P(x)Q(x';x)}{P(x')Q(x;x')}\right) = P(x')\cdot T(x \leftarrow x')
\end{aligned}
$$

Step-size demo

Explore standard normal distribution with different step sizes σ

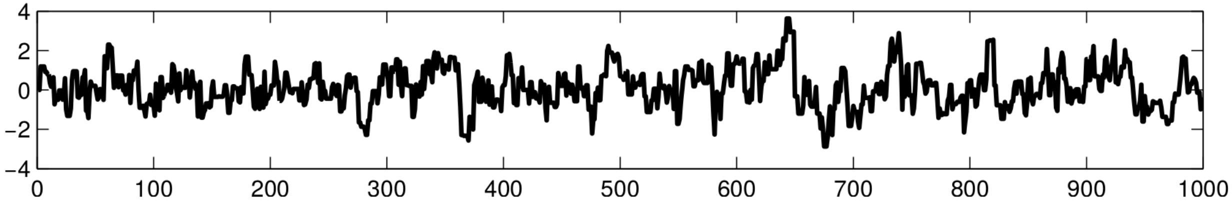
`sigma(0.1)`

99.8% accepts



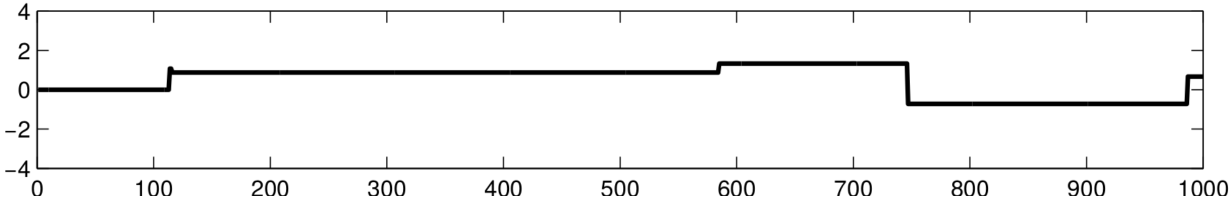
`sigma(1)`

68.4% accepts

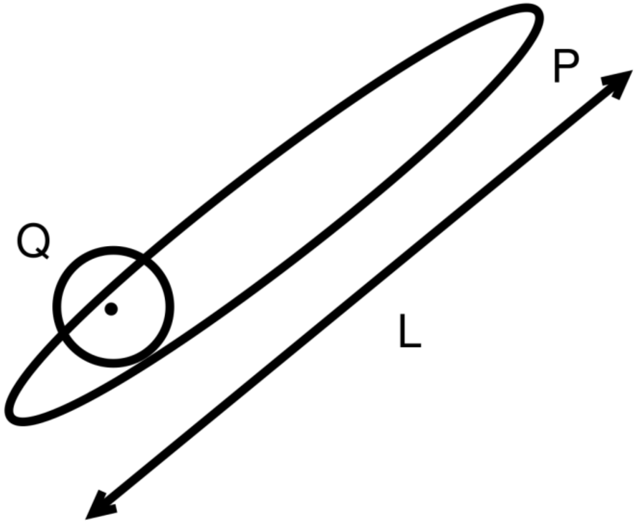


`sigma(100)`

0.5% accepts



Metropolis limitations



Generic proposals use

$$Q(x';x) = \mathcal{N}(x,\sigma^2)$$

σ large → many rejections

σ small → slow diffusion:
$\sim(L/\sigma)^2$ iterations required

Random walk Metropolis

E.g., $Q(x'; x) = \mathcal{N}(x, \sigma^2 I)$

Or $Q(x'; x) \propto 1_{\|x'-x\|_\infty \leq \Delta}$

$$\text{Acceptance prob} = \min\left\{\frac{P^*(x')}{P^*(x)}, 1\right\}$$

How large a step?

Tiny step $\implies$ large $P^*(x')/P^*(x) \implies$ high acceptance

Large step $\implies$ small $P^*(x')/P^*(x) \implies$ low acceptance

We might have wanted high acceptance **and** large moves.

But there's a tradeoff.

0.234 rule

Default advice:

try step sizes until about 23.4% of proposals are accepted. (Wide range ok)

Why?

Gelman, Roberts, Gilks (1996)

Consider exploring a high dimensional unimodal density,

such as $P^* = \mathcal{N}(0, I_d)$ with $Q(x'; x) = \mathcal{N}(x, \sigma_d^2 I_d)$
or $P^* = \mathcal{N}(\mu, \Sigma)$ with $Q(x'; x) = \mathcal{N}(x, \sigma_d^2 \Sigma)$

They find the asymptotically optimal σ_d is $2.38/\sqrt{d}$.

It is hard to scale the problem to make $\Sigma = I$. Easy to monitor acceptance rate.

However the optimal σ_d yields 23.4% acceptance as $d \to \infty$

And close to that for $d \geqslant 5$.

Multimodal problems

Requires larger steps and lower acceptance.

Metropolis-adjusted Langevin algorithm (MALA)

Overdamped Langevin equation:

$$\mathrm{d}x = \nabla \log P^*(x)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t$$

⟹ Euler discretization:

$$x_{k+1} = x_k + \tau\nabla \log P^*(x_k) + \sqrt{2\tau}\xi_k$$

⟹ Metropolis acceptance with proposale density:

$$Q(x';x) = \mathcal{N}(x + \tau\nabla \log P^*(x), 2\tau I)$$

The optimal acceptance rate for this algorithm is 0.574 according to G. O. Roberts and J. S. Rosenthal (1998).

Stochastic gradient Langevin dynamics

Settings: Inference for big data (notations are different here)

Given some parameter vector θ, its prior distribution $p(\theta)$, and a set of data points $X = \{x_i\}_{i=1}^{N}$, Stochastic Gradient Langevin dynamics samples from the posterior distribution

$$p(\theta|X) \propto p(\theta) \prod_{i=1}^{N} p(x_i|\theta)$$

But it is difficult to directly draw samples for an extremely large N.

Stochastic gradient Langevin dynamics

Stochastic optimization: If we are only interested in the MAP estimation

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

where

$$\sum_{t=1}^{\infty} \epsilon_t = \infty \qquad \sum_{t=1}^{\infty} \epsilon_t^2 < \infty$$

Stochastic gradient Langevin dynamics

Stochastic gradient Langevin dynamics:

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

$$\eta_t \sim N(0, \epsilon_t)$$

where

$$\sum_{t=1}^{\infty} \epsilon_t = \infty \qquad \sum_{t=1}^{\infty} \epsilon_t^2 < \infty$$

It can be proved that $\theta_t \rightarrow p(\theta|X)$ as $t \rightarrow \infty$. (Welling and Teh, ICML 2011)

Random batch method for interacting particle systems

Settings: Simulation for a large number of particles

Given a system consisting of N particles $\{x^i\}_{i=1}^N$, the external force $-\nabla V$ and the interacting force K, we hope to draw samples from the equilibrium distribution of

$$\mathrm{d}x^i = -\nabla V(x^i)\mathrm{d}t + \frac{1}{N-1}\sum_{j\neq i} K(x^i - x^j)\mathrm{d}t + \sigma \mathrm{d}W^i$$

But it is difficult to perform direct simulations for an extremely large N.

Random batch method for interacting particle systems

Solution: Perform simulation within a random batch for each step.

Select a random batch $\mathcal{C} \subset \{1, \ldots, N\}$ and perform a simulation step within the batch:

$$\mathrm{d}x^i = -\nabla V(x^i)\mathrm{d}t + \frac{1}{|\mathcal{C}| - 1} \sum_{j \in \mathcal{C} \setminus \{i\}} K(x^i - x^j)\mathrm{d}t + \sigma \mathrm{d}W^i, \text{ for } i \in \mathcal{C}$$

It can be proved that the simulation equilibrium distribution tends to the true one as $\tau \to 0$ and $N \to \infty$. (Jin, JCP 2020)

Combining operators

A sequence of operators, each with $P^\star$ invariant:

$$\begin{array}{lrcl}
x_0 \sim P^\star(x) & & & \\
x_1 \sim T_a(x_1 \leftarrow x_0) & P(x_1) & = \sum_{x_0} T_a(x_1 \leftarrow x_0) P^\star(x_0) & = P^\star(x_1) \\
x_2 \sim T_b(x_2 \leftarrow x_1) & P(x_2) & = \sum_{x_1} T_b(x_2 \leftarrow x_1) P^\star(x_1) & = P^\star(x_2) \\
x_3 \sim T_c(x_3 \leftarrow x_2) & P(x_3) & = \sum_{x_1} T_c(x_3 \leftarrow x_2) P^\star(x_2) & = P^\star(x_3) \\
\quad \ldots & & \ldots &
\end{array}$$

— Combination $T_c T_b T_a$ leaves $P^\star$ invariant

— If they can reach any x, $T_c T_b T_a$ is a valid MCMC operator

— Individually T_c, T_b and T_a need not be ergodic

Gibbs sampling

A method with no rejections:

– Initialize $\mathbf{x}$ to some value

– Pick each variable in turn or randomly and resample $P(x_i|\mathbf{x}_{j\neq i})$

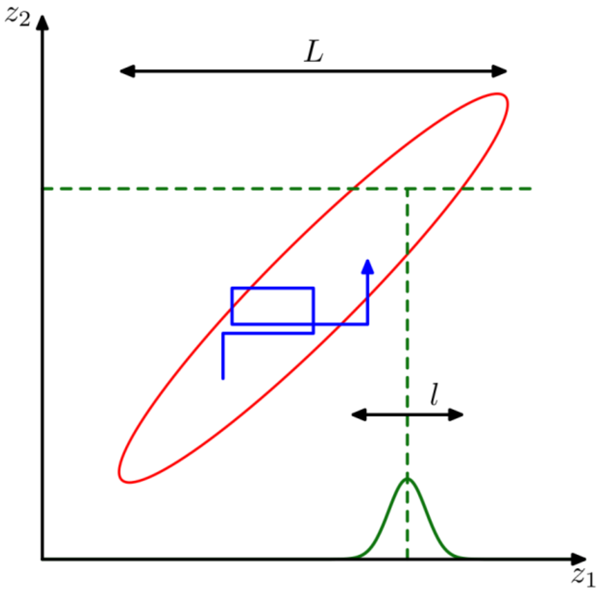


Figure from PRML, Bishop (2006)

Proof of validity: **a)** check detailed balance for component update.
b) Metropolis–Hastings 'proposals' $P(x_i|\mathbf{x}_{j\neq i}) \Rightarrow$ accept with prob. 1
Apply a series of these operators. Don't need to check acceptance.

Gibbs sampling

Alternative explanation:

Chain is currently at $\mathbf{x}$

At equilibrium can assume $\mathbf{x} \sim P(\mathbf{x})$

Consistent with $\mathbf{x}_{j\neq i} \sim P(\mathbf{x}_{j\neq i}), \quad x_i \sim P(x_i \,|\, \mathbf{x}_{j\neq i})$

Pretend x_i was never sampled and do it again.

“Routine” Gibbs sampling

Gibbs sampling benefits from few free choices and convenient features of conditional distributions:

- Conditionals with a few discrete settings can be explicitly normalized:

$$\begin{aligned} P(x_i|\mathbf{x}_{j\neq i}) &\propto P(x_i, \mathbf{x}_{j\neq i}) \\ &= \frac{P(x_i, \mathbf{x}_{j\neq i})}{\sum_{x_i'} P(x_i', \mathbf{x}_{j\neq i})} \leftarrow \text{this sum is small and easy} \end{aligned}$$

- Continuous conditionals only univariate
  $\Rightarrow$ amenable to standard sampling methods.

“Routine” Gibbs sampling

Gibbs sampling benefits from few free choices and convenient features of conditional distributions:

- Conditionals with a few discrete settings can be explicitly normalized:

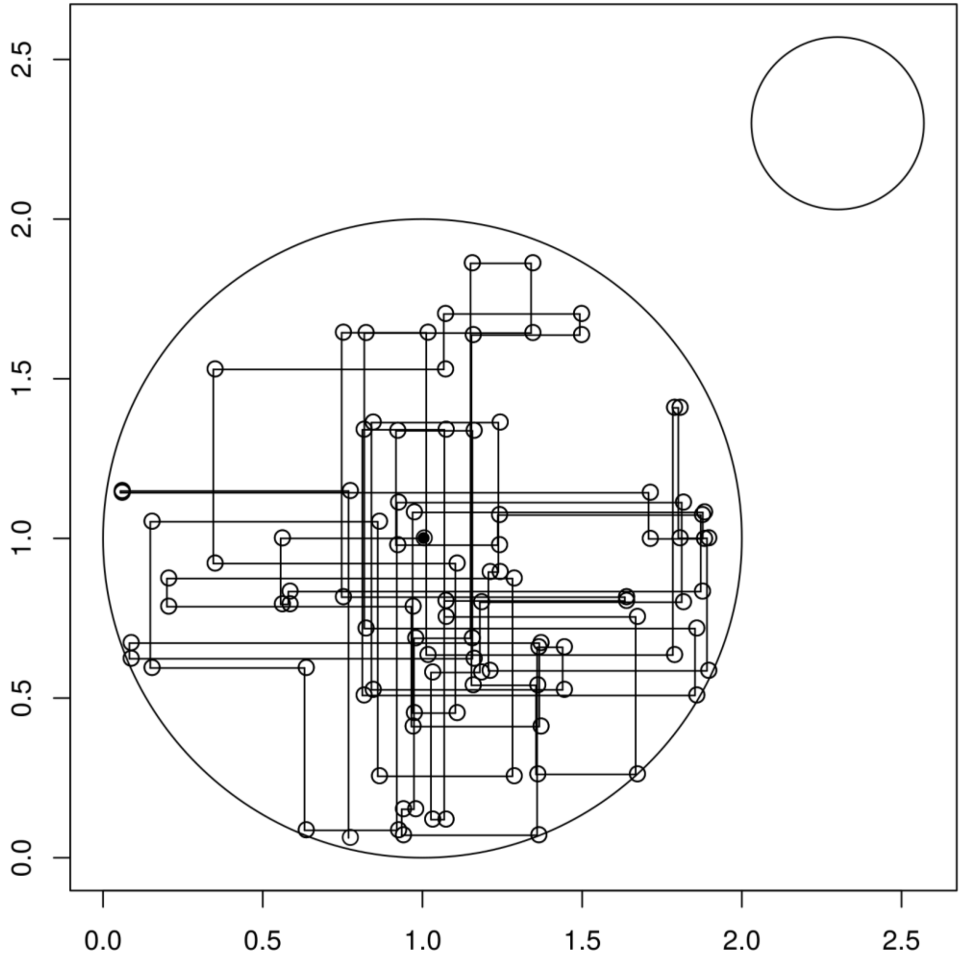
$$P(x_i|\mathbf{x}_{j\neq i}) \;\propto\; P(x_i, \mathbf{x}_{j\neq i})$$

$$= \frac{P(x_i, \mathbf{x}_{j\neq i})}{\sum_{x'_i} P(x'_i, \mathbf{x}_{j\neq i})} \;\leftarrow \text{this sum is small and easy}$$

- Continuous conditionals only univariate
  ⇒ amenable to standard sampling methods.

Metropolis sampling can also be used for each Gibbs sampling step.

Reducible Gibbs

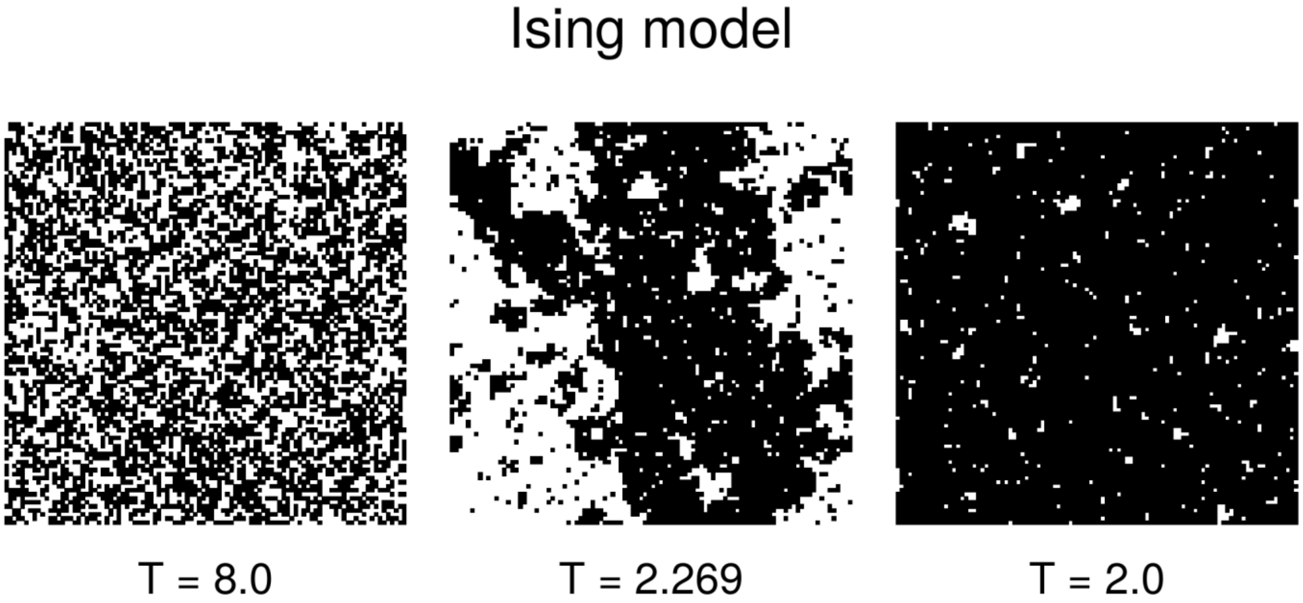


- Uniform in two circles
- Update horizontal then vertical etc.
- We get stuck on Earth
- Never sample the Moon

Ising model

Image $\boldsymbol{x} \in \{-1, 1\}^{R\times C}$ with $\pi(\boldsymbol{x}) = \exp(-H(\boldsymbol{x})/T)$ temperature $T > 0$

$$H(\boldsymbol{x}) = -\sum_{j\sim k} x_j x_k$$



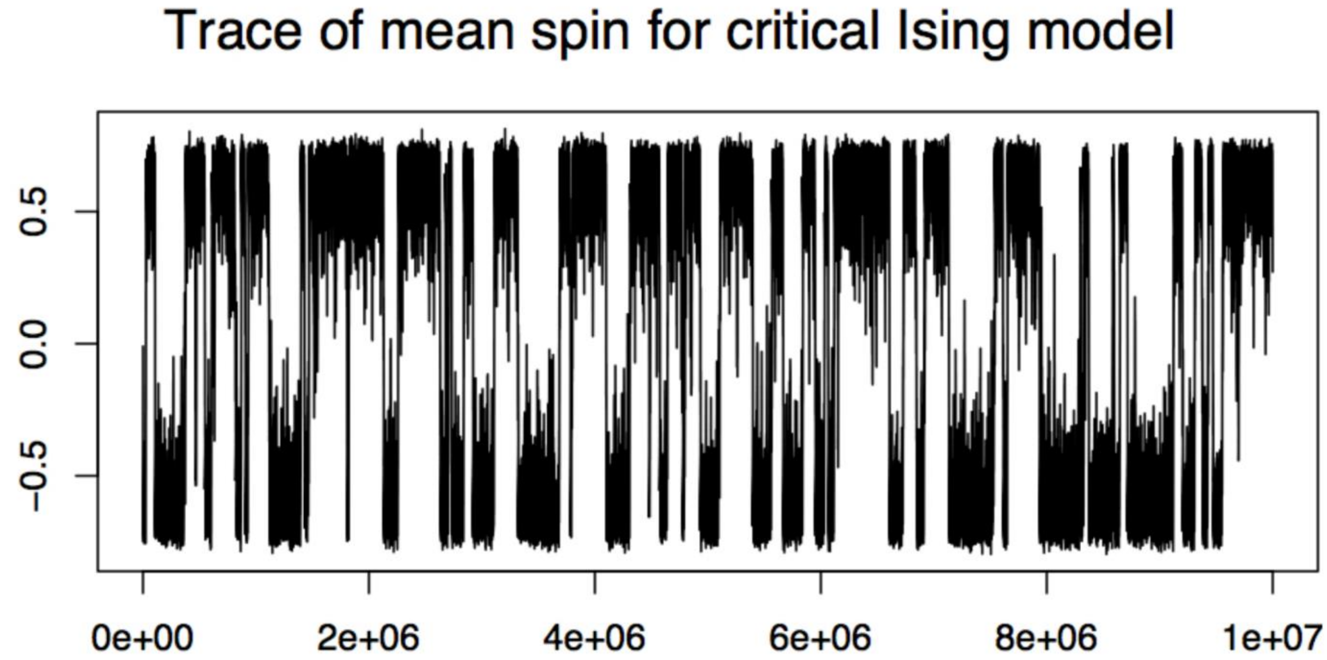
Used in physics (eg magnetism). Besag introduced it to image processing.

Ising model

There are very clever ways to sample the Ising model.

Or we can just flip bits conditionally on their 4 neighbours.

Let's trace mean spin $\frac{1}{RC}\sum_{i=1}^{R}\sum_{j=1}^{C}x_{ij}$



We see it makes a smallish number of round trips.

Summary so far

- We need approximate methods to solve sums/integrals
- Monte Carlo does not explicitly depend on dimension, although simple methods work only in low dimensions
- Markov chain Monte Carlo (MCMC) can make local moves. By assuming less, it's more applicable to higher dimensions
- simple computations ⇒ "easy" to implement (harder to diagnose).

How do we use these MCMC samples?

Burn-in

The law of (Markov chain) large numbers supports:

$$\hat{\mu} = \frac{1}{S} \sum_{i=1}^{S} f(x^{(i)})$$

Burn-in ≡ warmup

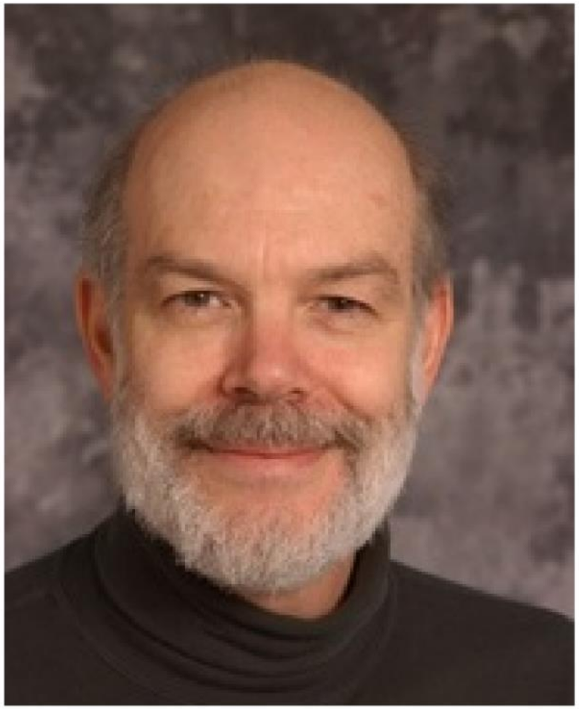
$$\hat{\mu} = \frac{1}{S-B} \sum_{i=B+1}^{S} f(x^{(i)})$$

Skip a few observations. Maybe they're not so close to P^*.

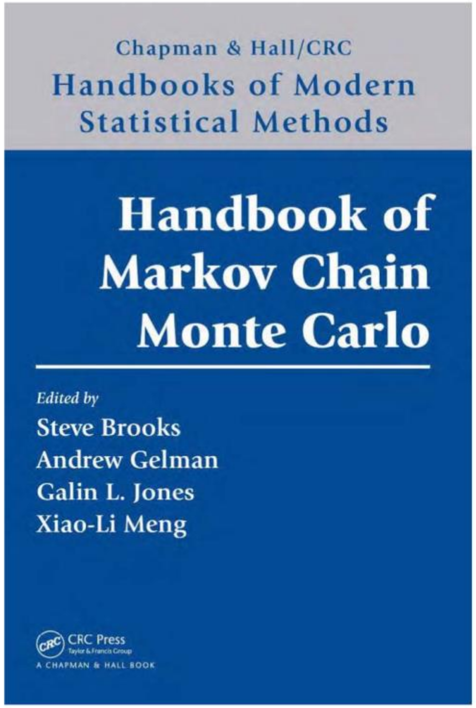
Should we? Yes and no.

Burn-in

Charlie Geyer



Won’t throw out any data.



Chapman & Hall/CRC
Handbooks of Modern
Statistical Methods

Handbook of
Markov Chain
Monte Carlo

Edited by
Steve Brooks
Andrew Gelman
Galin L. Jones
Xiao-Li Meng

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

← In this book. →

Andrew Gelman

Likes to use B = S / 2

Thinning

Approximately independent samples can be obtained by *thinning*.
However, **all the samples can be used.**

Use the simple Monte Carlo estimator on MCMC samples. It is:

— consistent

— unbiased if the chain has “burned in”

The correct motivation to thin: if computing $f(\mathbf{x}^{(s)})$ is expensive

Variance

Assume $x^{(i)} \sim P^*$ (e.g., burn-in) then for $y^{(i)} = f(x^{(i)}) \in \mathbb{R}$,

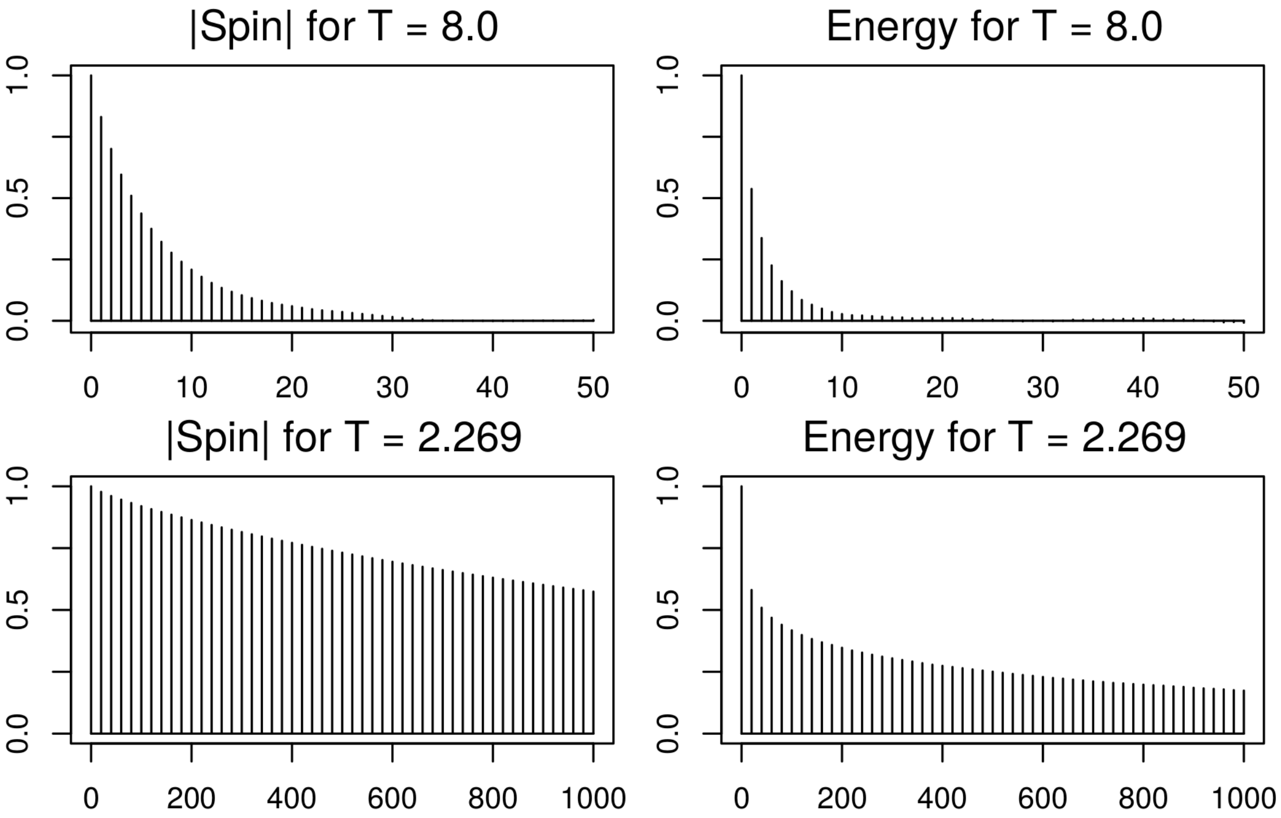
$$\begin{aligned}\mathrm{Var}(\hat{\mu}) &= \tfrac{1}{S^2}\textstyle\sum_{i=1}^{S}\sum_{j=1}^{S}\mathrm{Cov}(y^{(i)}, y^{(j)})\\ &= \frac{\mathrm{Var}(y)}{S^2}\textstyle\sum_{i=1}^{S}\sum_{j=1}^{S}\rho_{|i-j|}\\ &= \frac{\mathrm{Var}(y)}{S}\left(1 + 2\textstyle\sum_{k=1}^{\infty}\rho_k\right)\end{aligned}$$

assuming that the limit of $\sum_{k=1}^{\infty} \rho_k$ exists. Typically they do, like $\rho_k = O(\rho^k)$ for some $\rho < 1$.

Practical Markov chain Monte Carlo
Charles J. Geyer, *Statistical Science*. 7(4):473–483, 1992.
`http://www.jstor.org/stable/2246094`

Autocorrelations

Autocorrelations for the Ising model



Did the chain mix well?

Bad ACF $\Longrightarrow$ No

Good ACF $\Longrightarrow$ Maybe

Recent promising work by Gorham & Mackey using Stein discrepancy can provide a “Yes” (but it’s expensive).

https://arxiv.org/abs/1909.11827
https://arxiv.org/abs/1703.01717

Summary so far

- MCMC algorithms are general and often easy to implement
- Running them *is* a bit messy. . .
  . . . but there are some established procedures.
- Given the samples there might be a choice of estimators

Next question:
Is MCMC research all about finding a good $Q(\mathbf{x})$?

Hamiltonian dynamics

Construct a landscape with gravitational potential energy, E(x):

$$P(x) \propto e^{-E(x)}, \qquad E(x) = -\log P^*(x)$$

Introduce velocity v carrying kinetic energy $K(v) = v^\top v/2$

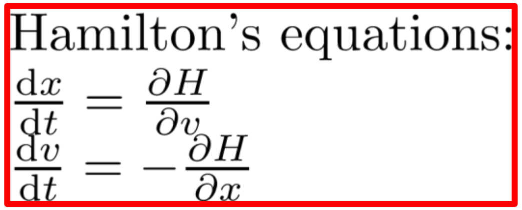
Some physics:

- Total energy or Hamiltonian, $H = E(x) + K(v)$
- Frictionless ball rolling $(x, v) \rightarrow (x', v')$ satisfies $H(x', v') = H(x, v)$
- Ideal Hamiltonian dynamics are time reversible:
  - reverse v and the ball will return to its start point

Hamiltonian Monte Carlo

Define a joint distribution:

- $P(x, v) \propto e^{-E(x)} e^{-K(v)} = e^{-E(x)-K(v)} = e^{-H(x,v)}$
- Velocity is independent of position and Gaussian distributed



Hamilton's equations:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{\partial H}{\partial v}$$

$$\frac{\mathrm{d}v}{\mathrm{d}t} = -\frac{\partial H}{\partial x}$$

Markov chain operators

- Gibbs sample velocity
- Simulate Hamiltonian dynamics then flip sign of velocity
  - Hamiltonian 'proposal' is deterministic and reversible $q(x', v'; x, v) = q(x, v; x', v') = 1$
  - Conservation of energy means $P(x, v) = P(x', v')$
  - Metropolis acceptance probability is 1

Except we can't simulate Hamiltonian dynamics exactly

Leap-frog dynamics

a discrete approximation to Hamiltonian dynamics:

$$
\begin{aligned}
v_i(t+\tfrac{\epsilon}{2}) &= v_i(t) - \frac{\epsilon}{2}\frac{\partial E(x(t))}{\partial x_i} \\
x_i(t+\epsilon) &= x_i(t) + \epsilon\, v_i(t+\tfrac{\epsilon}{2}) \\
v_i\,(t+\epsilon) &= v_i(t+\tfrac{\epsilon}{2}) - \frac{\epsilon}{2}\frac{\partial E(x(t+\epsilon))}{\partial x_i}
\end{aligned}
$$

- H is not conserved
- dynamics are still deterministic and reversible
- Acceptance probability becomes $\min[1, \exp(H(v,x) - H(v',x'))]$

Leap-frog dynamics

a discrete approximation to Hamiltonian dynamics:

$$\begin{aligned}
v_i(t+\tfrac{\epsilon}{2}) &= v_i(t) - \frac{\epsilon}{2}\,\frac{\partial E(x(t))}{\partial x_i} \\
x_i(t+\epsilon) &= x_i(t) + \epsilon\, v_i(t+\tfrac{\epsilon}{2}) \\
v_i\,(t+\epsilon) &= v_i(t+\tfrac{\epsilon}{2}) - \frac{\epsilon}{2}\,\frac{\partial E(x(t+\epsilon))}{\partial x_i}
\end{aligned}$$

- H is not conserved
- dynamics are still deterministic and reversible
- Acceptance probability becomes $\min[1, \exp(H(v,x) - H(v',x'))]$

Why?

MH with deterministic transformation

1. Current sample x
2. Draw a random variable $v \sim g(v)$
3. Perform an invertible and deterministic transformation $(x', v') = h(x, v)$
4. Accept x' as the new sample (i.e., $x := x'$) with probability

$$\alpha(x, x') = \min\left\{1, \frac{p(x')g'(v')}{p(x)g(v)} \left|\frac{\partial(x', v')}{\partial(x, v)}\right|\right\}$$

5. The invariant distribution of of the sampling step is $p(x)$

Green, P. J. (2003). Trans-dimensional Markov chain Monte Carlo, pp. 179–98. OUP, Oxford.

MH with deterministic transformation

1. Current sample x
2. Draw a random variable $v \sim g(v)$
3. Perform an invertible and deterministic transformation $(x', v') = h(x, v)$
4. Accept x' as the new sample (i.e., $x := x'$) with probability

$$\alpha(x, x') = \min \left\{ 1, \frac{p(x')g'(v')}{p(x)g(v)} \left| \frac{\partial(x', v')}{\partial(x, v)} \right| \right\}$$

5. The invariant distribution of of the sampling step is $p(x)$

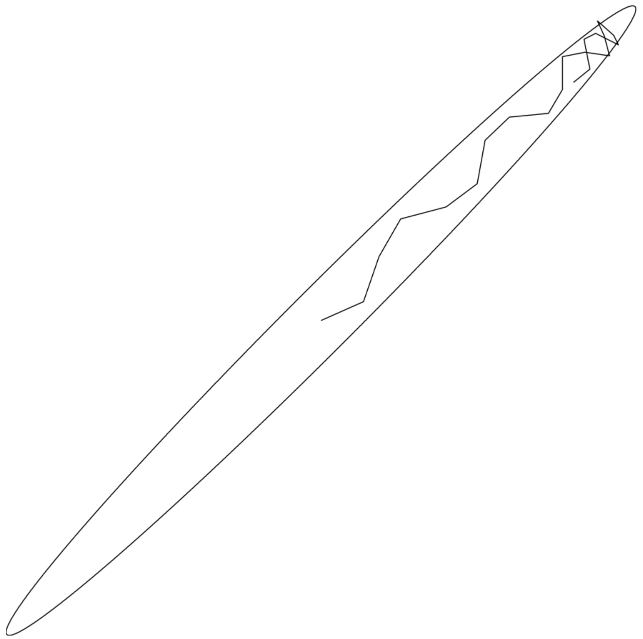
For HMC, $p(x) = \exp(-E(x)), g(v) = g'(v) = \exp(-K(v))$ and $h(x, v)$ is volume preserving.

Green, P. J. (2003). Trans-dimensional Markov chain Monte Carlo, pp. 179–98. OUP, Oxford.

Hamiltonian Monte Carlo

The algorithm:

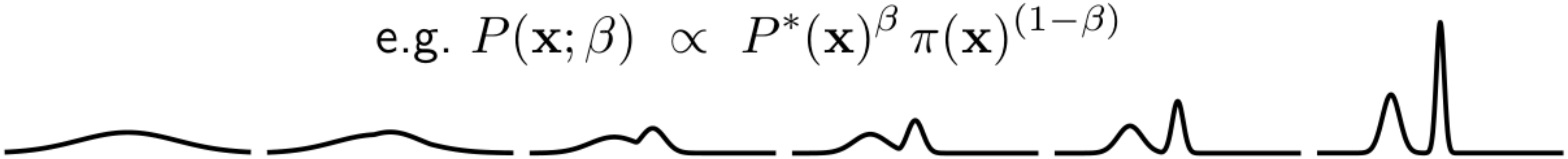
- Gibbs sample velocity $\sim \mathcal{N}(0, \mathbb{I})$
- Simulate Leapfrog dynamics for L steps
- Accept new position with probability $\min[1, \exp(H(v, x) - H(v', x'))]$

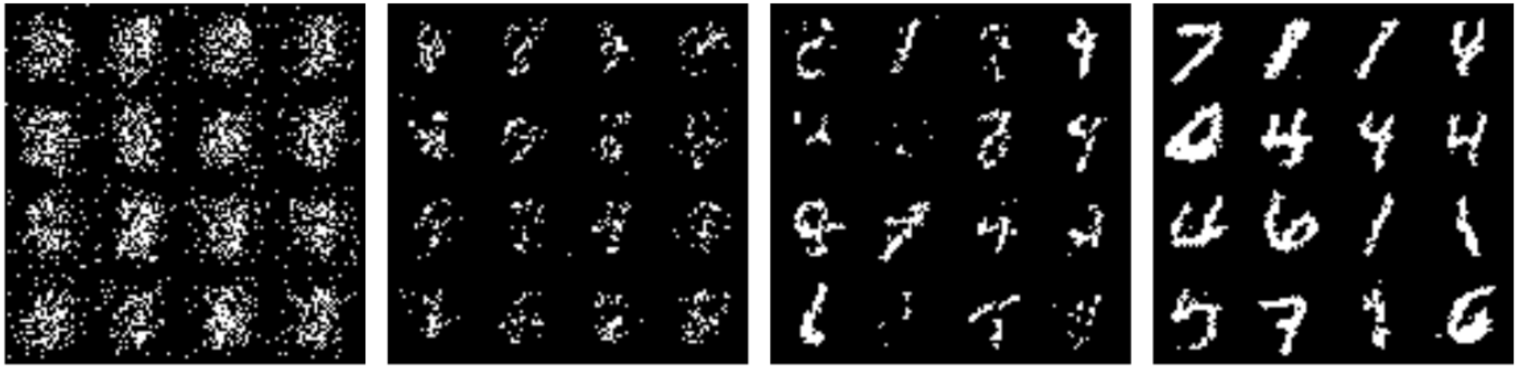


The original name is **Hybrid Monte Carlo**, with reference to the "hybrid" dynamical simulation method on which it was based.

Annealing / Tempering

e.g. $P(\mathbf{x};\beta) \;\propto\; P^*(\mathbf{x})^{\beta}\,\pi(\mathbf{x})^{(1-\beta)}$



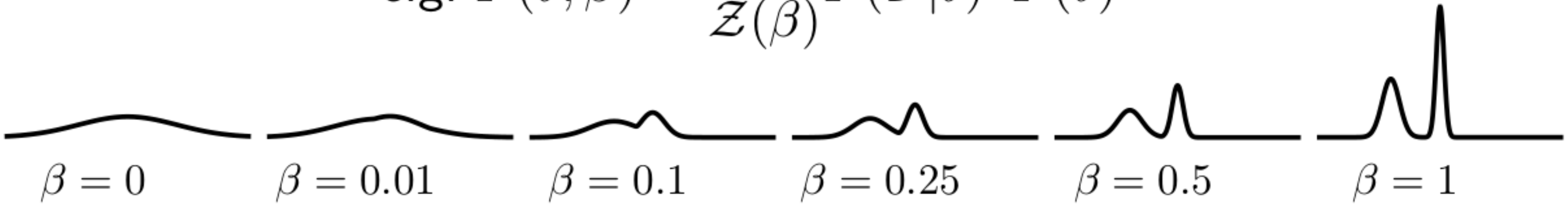


$1/\beta \;=\;$ "temperature"

Using other distributions

Chain between posterior and prior:

$$\text{e.g. } P(\theta;\beta) = \frac{1}{\mathcal{Z}(\beta)} P(\mathcal{D}|\theta)^{\beta} P(\theta)$$



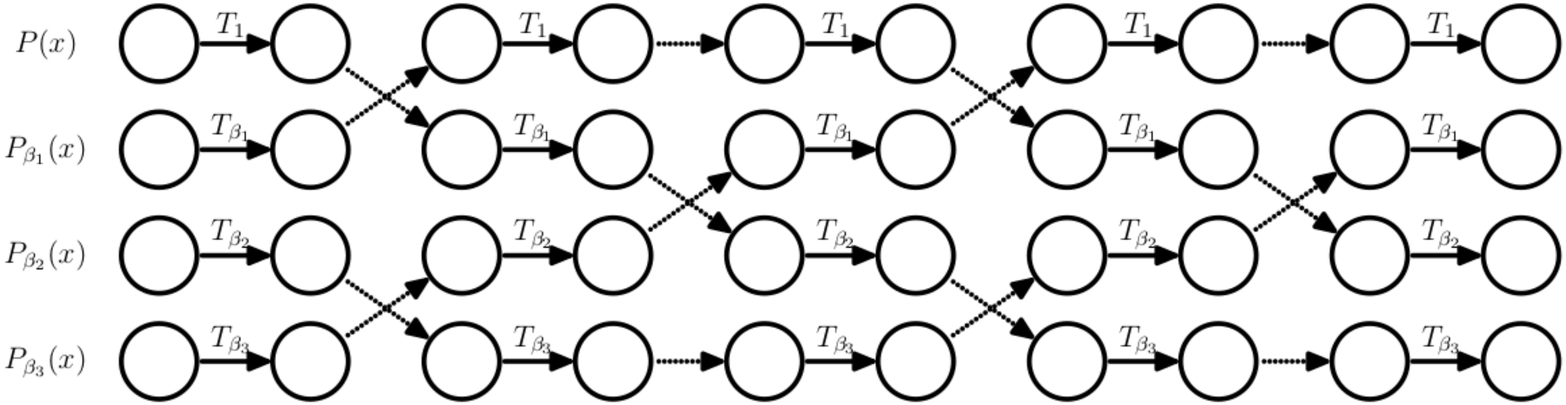
Advantages:

- mixing easier at low β, good initialization for higher β?
- $\dfrac{\mathcal{Z}(1)}{\mathcal{Z}(0)} = \dfrac{\mathcal{Z}(\beta_1)}{\mathcal{Z}(0)} \cdot \dfrac{\mathcal{Z}(\beta_2)}{\mathcal{Z}(\beta_1)} \cdot \dfrac{\mathcal{Z}(\beta_3)}{\mathcal{Z}(\beta_2)} \cdot \dfrac{\mathcal{Z}(\beta_4)}{\mathcal{Z}(\beta_3)} \cdot \dfrac{\mathcal{Z}(1)}{\mathcal{Z}(\beta_4)}$

Related to *annealing* or *tempering*, $1/\beta =$ “temperature”

Parallel tempering

Normal MCMC transitions + swap proposals on $P(X) = \prod_{\beta} P(X;\beta)$



Problems / trade-offs:

- obvious space cost
- need to equilibriate larger system
- information from low β diffuses up by slow random walk

If (X';β) and $(X';\beta)$ are chosen,

they will be exchanged with probability

$$\min\left\{1, \frac{P(X';\beta)P(X;\beta')}{P(X;\beta)P(X';\beta')}\right\}$$

Approx. Bayesian computation

$$\pi(\theta \mid \boldsymbol{x}) \propto \pi(\theta) \times p(\boldsymbol{x} \mid \theta)$$

Sometimes we cannot compute the likelihood $p(\boldsymbol{x} \mid \theta)$.

E.g., θ describes how a colony of bacteria evolves over time, and

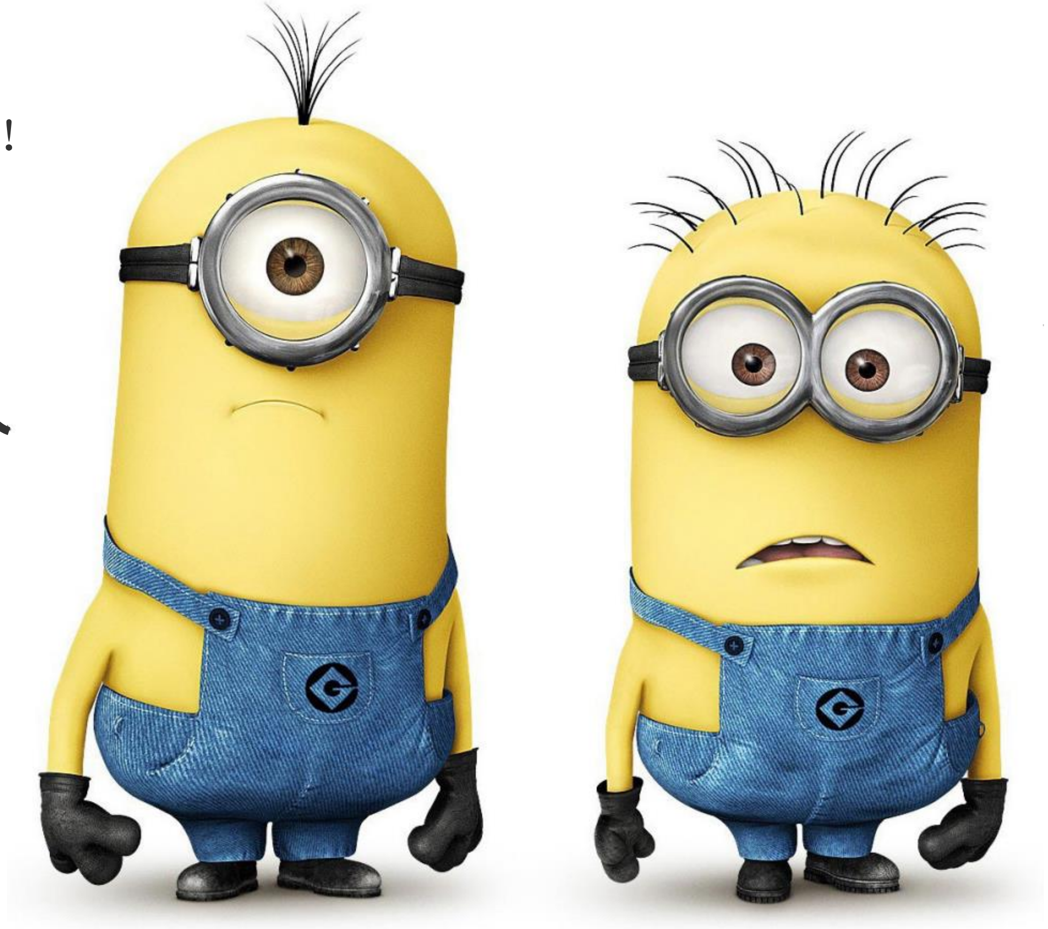
$\boldsymbol{x}$ is how it looks right now

A taste of ABC

```
Loop over $i$
    Sample $\theta_i \sim \pi(\theta)$.
    Sample $\boldsymbol{x}_i \mid \theta_i$
    Keep $\theta_i \iff \|\boldsymbol{x}_i - \boldsymbol{x}\| \leqslant \epsilon$
Use the retained $\theta_i$
```

Many variants. Now a whole handbook.

THE END



Thank you！

Questions？