Markov Chain Monte Carlo. (MCMC)

Note Title 5/7/2006

MCMC is a way to sample from any distribution. $\pi(\underline{x})$

It proceeds by designing a Markov process that converges to samples from $\pi(\underline{x})$.

The original MCMC is the Metropolis algorithm (Metropolis, Rosenbluth & Rosenbluth, Teller & Teller 1953).

Generalized to the Metropolis-Hastings algorithm

Strictly speaking, Metropolis-Hastings is a design principle for algorithm.

What is a Markov Chain?

Define a function

$$p(\underline{x}^{t+1} | \underline{x}^t) = K(\underline{x}^{t+1} | \underline{x}^t)$$

obeys $\sum_{\underline{x}^{t+1}} K(\underline{x}^{t+1} | \underline{x}^t) = 1, \forall \underline{x}^t$

Sample $\underline{x}^0, \underline{x}^1, \ldots \underline{x}^t, \ldots$

$\underline{x}^1$ from $p(\underline{x}^1 | \underline{x}^0)$, .., $\underline{x}^t$ from $p(\underline{x}^t | \underline{x}^{t-1})$ ...

This is a Markov Chain. E.g. A random walk.

(*) Irreducible, $\forall x, y \; \exists n$ s.t. $\sum_{x_1 \ldots x_n} K(x|x_1) K(x_1|x_2) \ldots K(x_n|y) > 0$.

(*) Aperiodic - almost always satisfied -

MCMC is a special Markov Chain where $K(\underline{x}^{t+1}|\underline{x}^t)$ — the transition probability — is designed to satisfy the condition:

$$\sum_{\underline{y}} K(\underline{x}|\underline{y}) \pi(\underline{y}) = \pi(\underline{x}).$$

In practice, $K(\underline{x}^{t+1}|\underline{x}^t)$ is almost always required to satisfy a stronger condition called "detailed balance"

$$K(\underline{x}|\underline{y}) \pi(\underline{y}) = K(\underline{y}|\underline{x}) \pi(\underline{x})$$

Detailed Balance implies that

$$\sum_{\underline{y}} K(\underline{x}|\underline{y}) \pi(\underline{y}) = \sum_{\underline{y}} K(\underline{y}|\underline{x}) \pi(\underline{x}) = \pi(\underline{x}) \; //$$

Provided $K(\underline{x}|\underline{y})$ obeys additional technical conditions (*), then samples from the Markov Chain will converge to samples from $\pi(\underline{x})$

$$\underline{x}^0, \underline{x}^1, \underline{x}^2, \ldots \underline{x}^t, \ldots$$

For sufficiently large t, $\underline{x}^t$ is a sample from $\pi(\underline{x})$.

Metropolis

Suppose $\pi(\underline{x}) = \frac{1}{Z} e^{-E(\underline{x})}$ (Z hard/impossible to calculate.)

For each $\underline{x}$, define a neighbourhood structure $N(\underline{x})$ s.t. $\underline{y} \in N(\underline{x}) \iff \underline{x} \in N(\underline{y})$ and $|N(\underline{x})|$ is independent of $\underline{x}$.

Loop over t,

At time t with state $\underline{x}^t$:

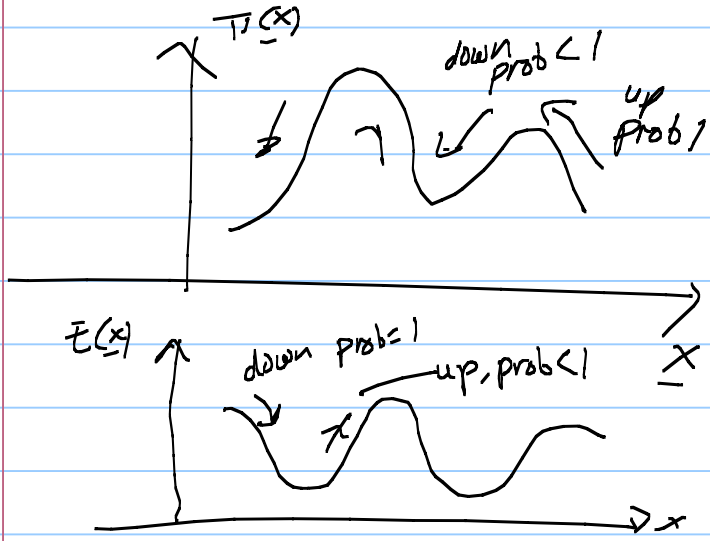
(1) propose a move $\underline{x}^t \to \underline{x}^{t+1} \in N(\underline{x}^t)$ with uniform probability. (i.e. $p(\underline{x}^{t+1}) = \frac{1}{|N(\underline{x}^{t+1})|}$)

(2) accept the move with probability

$$\min\left\{1, \frac{\pi(\underline{x}^{t+1})}{\pi(\underline{x}^t)}\right\}.$$ (No need to compute Z!)

Note: you accept the move for certain if $\pi(\underline{x}^{t+1}) > \pi(\underline{x}^t)$ (i.e. $E(\underline{x}_{t+1}) < E(\underline{x}_t)$) but you also accept it with prob. $\frac{\pi(\underline{x}^{t+1})}{\pi(\underline{x}^t)} = e^{E(\underline{x}_t) - E(\underline{x}_{t+1})}$ if $\pi(\underline{x}^{t+1}) < \pi(\underline{x}^t)$ (i.e. $E(\underline{x}_{t+1}) > E(\underline{x}_t)$).

Intuitively, Metropolis encourages movement to states of high probability, but also allows movement to low probability states.



This ability to move in directions of lower probability (or higher energy) is needed to ensure that the MC explores the full space.

Check that Metropolis obeys Detailed Balance.

$$K(x|y)\pi(y) = 0 \quad \text{if } x \notin N(y)$$

otherwise
$$= \frac{1}{|N(y)|}\pi(y)\min\left\{1, \frac{\pi(x)}{\pi(y)}\right\}$$

$$= \frac{1}{|N(y)|}\min\{\pi(y), \pi(x)\} \quad \text{symmetry}$$

$$= K(y|x)\pi(x)$$

Recall $x \in N(y) \Leftrightarrow y \in N(x)$ & $|N(x)|$ indep of x

Also the MC must be able to explore every part of the space (Irreducible)

Metropolis Example

Ising Model:

$$p(\underline{x}) = \frac{1}{Z} e^{\beta \sum_{i=0}^{d-1} x_i x_{i+1} + \gamma \sum_{i=0}^{d} x_i c_i}$$

$x_i \in \{\pm 1\}$

Define neighbourhood structure.

$$N(\underline{x}) = \{ (x_0, x_1, \ldots, \bar{x}_k, \ldots x_d) : k = 0, \ldots d \},$$

where $\bar{x}_k = -x_k$.

At time t: $\underline{x}^t$

select node k at random (prob $1/d+1$)

Let $\underline{x}^t_{,k} = (x_0^t, \ldots \bar{x}_k^t, \ldots x_d^t)$.

$$E(\underline{x}^t_{,k}) - E(\underline{x}^t) = 2\gamma x_k^t c_k + 2\beta x_k^t \{x_{k+1}^t + x_{k-1}^t\}, \quad k \neq 0, d$$

Set $\underline{x}^{t+1} = \underline{x}^t_{,k}$ with prob 1 if $E(\underline{x}^t_{,k}) - E(\underline{x}^t) < 0$

$\underline{x}^{t+1} = \underline{x}^t_{,k}$ with prob $e^{E(\underline{x}^t) - E(\underline{x}^t_{,k})}$ if $E(\underline{x}^t) - E(\underline{x}^t_{,k}) < 0$

otherwise $\underline{x}^{t+1} = \underline{x}^t$.

Metropolis Example (cont)

The book gives examples (with $\gamma = 0$), p 109.

For this problem, you can do exact sampling

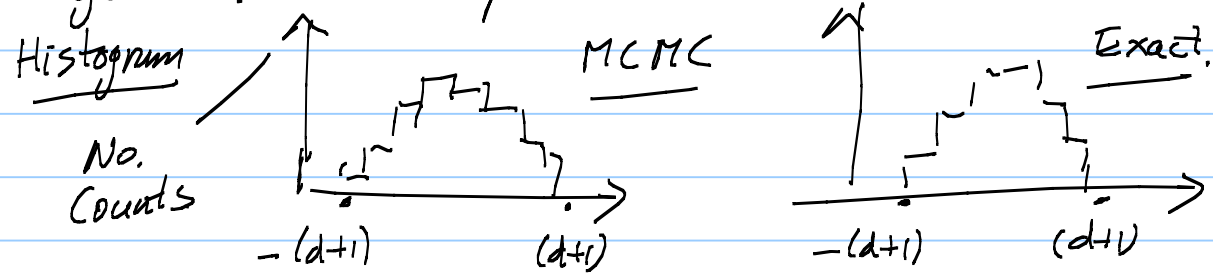
- express $\pi(\underline{x}) = \pi(x_d)\,\pi(x_{d-1}|x_d)\,..\,\pi(x_0|x_1)$ (using DP), then sample x_d from $\pi(x_d)$, $x_{d-1} \sim \pi(x_{d-2}|x_{d-1})$ etc.

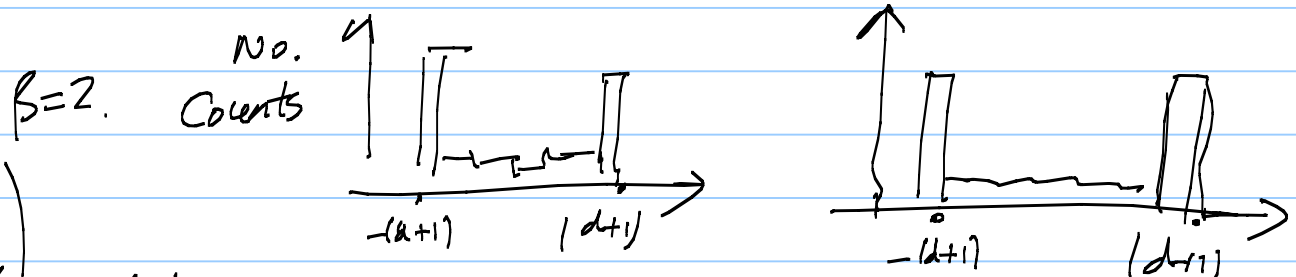
Summary Statistic "magnetization" $M^{(t)} = \sum_{i=0}^{d} x_i^{(t)}$

Run Metropolis for 1,000,000 iterations

Choose $x^{50}, x^{100}, \ldots x^{50k}, \ldots$ $k = 1, \ldots 20{,}000$

to get 2,000 samples.





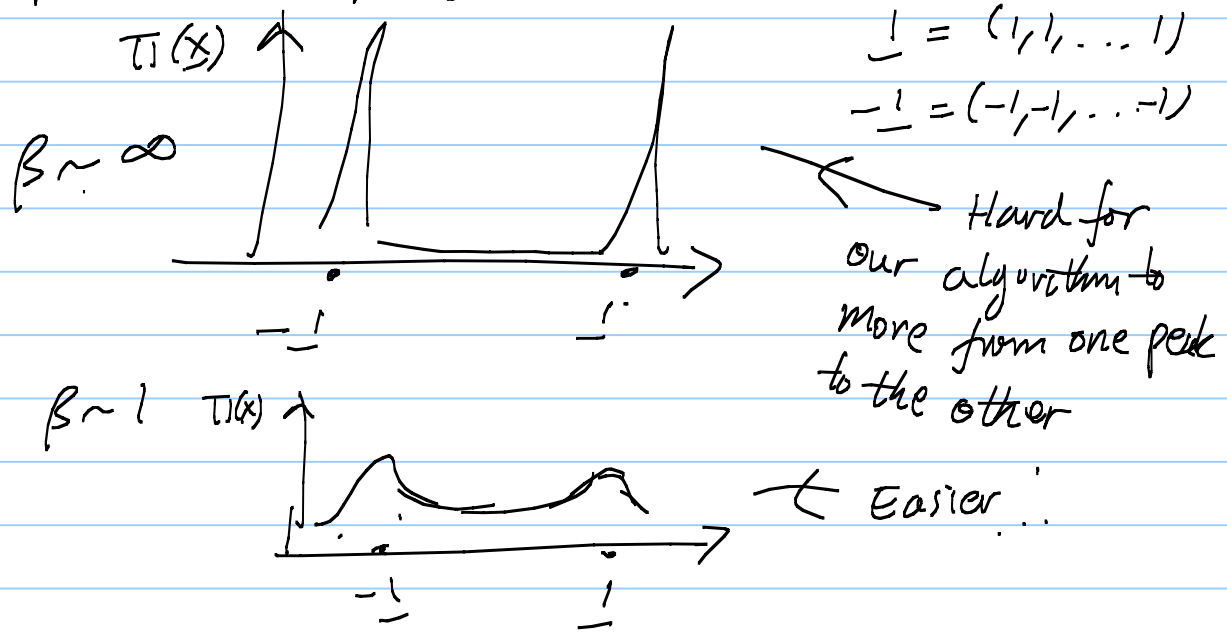
(Better figs in Book p 109)

(1) Histograms similar for $\beta = 1$, MCMC gives similar results to Exact. Similar computation time

Metropolis Example: (cont, cont)

(ii) Histograms differ for $\beta = 2$, also histogram for $\beta = 2$ differ greatly to those for $\beta = 1$. What is going on?

The distribution for $\pi(x)$ gets sharpened as β gets larger



Note: The Metropolis sampler we used is better at sampling from $\pi(\underline{x})$ with $\beta = 1$.

For $\pi(\underline{x})$ with $\beta \geq 2$, it takes a very long time for the algorithm to move from peak at $-\underline{1}$ to peak at $\underline{1}$.

Why do histograms differ from $\beta=1$ & $\beta=2$?

Answer: Phase Transition.

$$Pr(M=z) = \sum_{x_0 \ldots x_d} \delta_{x_0 + \cdots + x_d, z} \prod(\underline{x})$$

x: any value

Kronecker delta (Indicator function)

There is a phase factor, no. of different ways that you can get $M=z$ from $\underline{x}$.

For $M = d+1$, only one way $\underline{x} = (1, \ldots 1)$

$M = -(d+1)$, only one way $\underline{x} = (-1, \ldots -1)$

$$Pr(M = d+1) = Pr(M = -(d+1)) = \prod(\underline{1}) = \prod(\underline{-1})$$

For $M = d$, impossible

$M = d-1$, $(d+1)$ ways, $(-1, 1, 1 \ldots 1)$ $(1, -1, 1 \ldots 1)$ etc.

for $M = 0$, an exponential large no. of ways (provided $d+1$ is even)

In short, small phase factor for $M = \pm d+1$

enormous phase factor for $M = 0$.

Two Conflicting "forces"

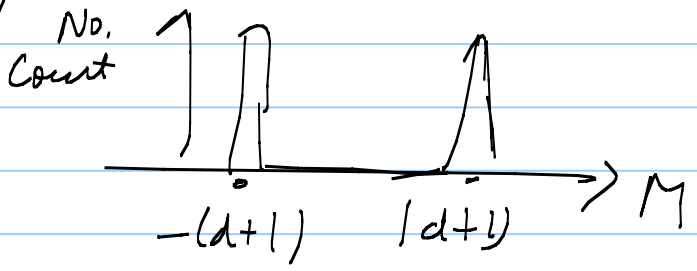
Phase factor wants $M = 0$.

Probability want $M = \pm(d+1)$.

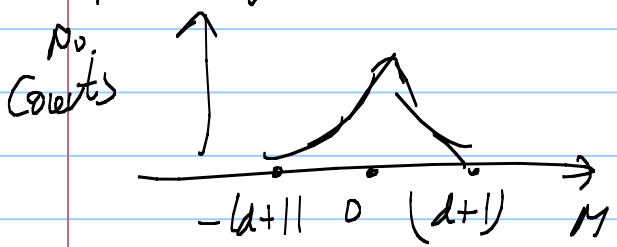
For large β, probability biases strongly to ± 1 (i.e. $M = \pm(d+1)$).

For small β, probability biases weakly to ± 1.

So for large β, probabilities win



For small β, "phase factors win.



Comments, better MCMC can be designed for case with large β.

MCMC can be as effective as Exact Sampling. But we can use MCMC in cases where exact sampling is impossible

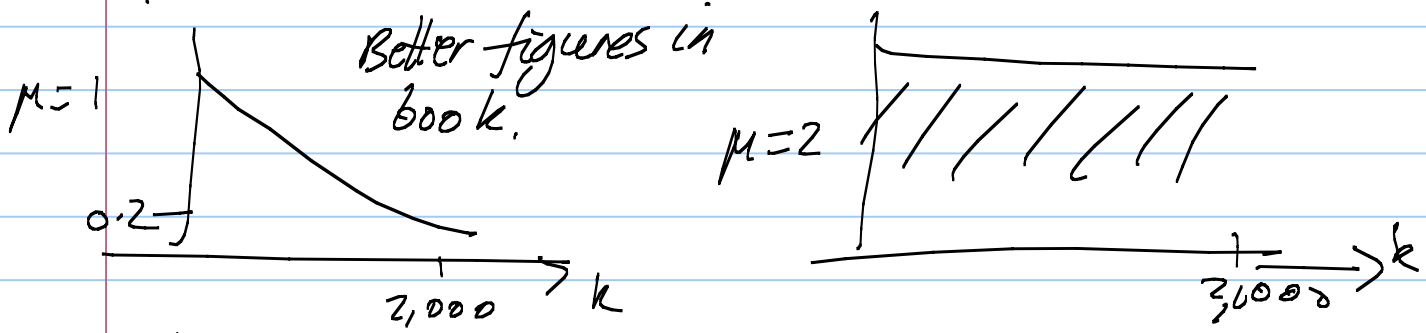
Burn-in and Stickyness

Burn-in is the amount of time for MCMC to produce samples from $\pi(\underline{x})$

Stickyness is a means of how MCMC explores $\pi(\underline{x})$.

Practical Measure: Autocorrelation

ρ_k = correlation $(M^{(i)}, M^{(k+i)})$



In short, this MCMC is sticky for $\mu=2$ since the autocorrelation decreases very slowly with time.

Samples after burn-in will be correlated. Must wait to get i.i.d samples

I.E. instead of samples $X^{50}, X^{100}, X^{150}, \ldots$

need $X^{100}, X^{600}, X^{1100}, \ldots$

(*) To get Metropolis, set $T(y|x)$ to be uniform.

Metropolis-Hastings

This improves the basic Metropolis algorithm by adding a proposal probability.

$T(y|x)$ for $y \in N(x)$.

At time t: state x^t

(i) propose $y \in N(x)$ with probability $T(y|x^t)$.

(ii) accept this proposal with probability $r(x^t, y) = \min\left\{1, \frac{\pi(y)\,T(x^t|y)}{\pi(x)\,T(y|x^t)}\right\}$

If accepted, set $x^{t+1} = y$.

Note: This is equivalent to setting

$$K(y|x) = T(y|x)\cdot \min\left\{1, \frac{\pi(y)\cdot T(x|y)}{\pi(x)\cdot T(y|x)}\right\}$$

Detailed Balance:

$$K(y|x)\pi(x) = \min\{\pi(x)\pi(y|x), \pi(y)\pi(x|y)\},$$

symmetric, so Detailed Balance holds.

Good choice of proposals makes MCMC very fast.

Metropolis Hastings is the most popular way of designing an MCMC. (They are others, see later).

Difficulty is selecting the neighbourhood and the proposal probabilities.

E.6. Proposals for Ising model.

$$N(\underline{x}) = \{ (x_0, \ldots, \overline{x_k}, \ldots x_d) \; ; \; k = 0 \ldots d \}$$

$$T(\underline{x}_{ik} | \underline{x}) = \frac{e^{|c_k|}}{\sum_{i=0}^{d} e^{|c_i|}}$$

, i.e. propose moves which involve flipping states where c_k is largest.

Designing a good MCMC is an art more than a science.

Study of the specific distribution $\pi(\underline{x})$, can suggest good proposal probabilities, good neighbourhood structures, etc.