Lecture 17.

Metropolis Algorithm

Note Title 5/7/2006

$\pi(\underline{x}) = \frac{1}{Z} e^{-E(\underline{x})}$ ← sisp

Neighborhood $N(\underline{x})$, s.t. if $\underline{y} \in N(\underline{x})$ then $\underline{x} \in N(\underline{y})$. $|N(\underline{x})|$ indep. of $\underline{x}$

Transition kernel:

$$K(\underline{x}^{t+1} | \underline{x}^t) = \frac{1}{|N(\underline{x}^t)|} \min\left\{1, \frac{\pi(\underline{x}^{t+1})}{\pi(\underline{x}^t)}\right\}, \quad \underline{x}^{t+1} \in N(\underline{x}^t),\ \underline{x}^{t+1} \neq \underline{x}^t$$

propose a move $\underline{x}^t \to \underline{x}^{t+1} \in N(\underline{x}^t)$ with uniform probability

accept move with prob $\min\left\{1, \frac{\pi(\underline{x}^{t+1})}{\pi(\underline{x}^t)}\right\}$

$$K(\underline{x}^t | \underline{x}^t) = 1 - \sum_{\underline{x}^{t+1} \in N(\underline{x}^t)} K(\underline{x}^{t+1} | \underline{x}^t)$$

→ prob you stay at $\underline{x}^t$

Intuition for Metropolis:

$$\min\left\{1, \frac{\pi(\underline{x}^{t+1})}{\pi(\underline{x}^t)}\right\} = \min\left\{1, e^{E(\underline{x}^t) - E(\underline{x}^{t+1})}\right\}$$

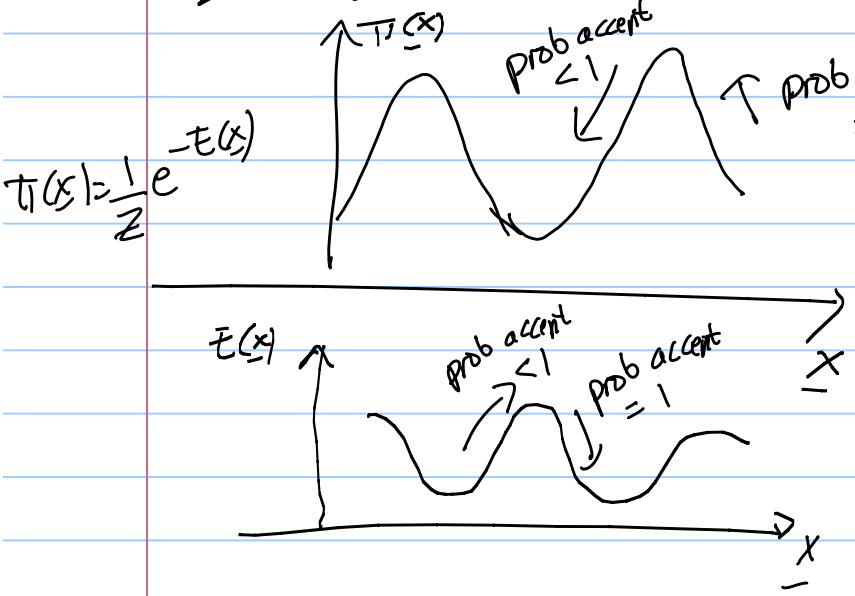
If $E(\underline{x}^{t+1}) \leq E(\underline{x}^t)$, the move $\underline{x}^t$ to $\underline{x}^{t+1}$ will always be accepted.

If $E(\underline{x}^{t+1}) > E(\underline{x}^t)$, the move $\underline{x}^t$ to $\underline{x}^{t+1}$ will be accepted with probability $e^{E(\underline{x}^t) - E(\underline{x}^{t+1})} < 1$.

Hence proposed moves to lower energy / higher probability states will always be accepted.

But moves to higher energy / lower probability states have a probability of being accepted.

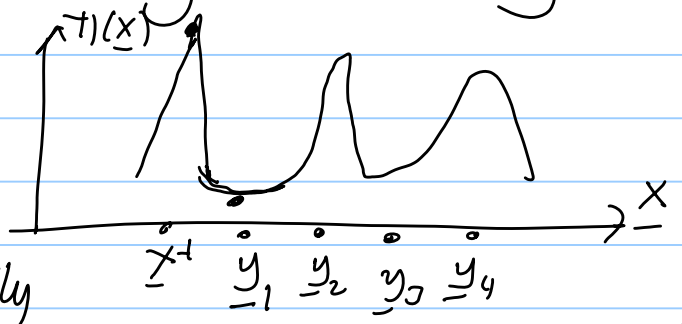
Intuitively, Metropolis encourages movement to states of high probability, but also allows movement to low probability states.



This ability to move in directions of lower probability (or higher energy) is needed to ensure that the MCMC explores the space of all $\underline{x}$.

Metropolis obeys detailed balance (last lecture) and so is guaranteed to converge eventually to samples from $\pi(\underline{x})$.

But - WARNING - Convergence may take a long time.



Suppose after t iterations the chain is at $\underline{x}^t$

Then the probability to moving downhill in probability from $\pi(\underline{x}^t)$ to $\pi(\underline{y}_1)$ is small. It requires a sequence of moves which decrease the probability and which all have prob < 1 of being accepted.

Metropolis-Example: Ising Model in 1-D.

Note: we already know how to sample a 1-D Ising model by sequential methods. This enables us to compare MCMC to sequential sampling.

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

Choose neighborhood structure:

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

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

i.e. flip the state $x_k \to \bar{x}_k$, of one node k.

Metropolis: At time t, state $\underline{x}^t = (x_0^t, \ldots, x_d^t)$

Select node k with prob $1/(d+1)$

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

Calculate: $E(\underline{x}^t_{,k}) - E(\underline{x}^t)$

$= 2\gamma x_k^t c_k + 2\beta x_k^t (x_{k+1}^t + x_{k-1}^t)$, for $k \neq 0, d$

(for $k=0$, or $k=d$ we modify the second term).

If $E(\underline{x}^t_{,k}) - E(\underline{x}^t) \leq 0$, set $\underline{x}^{t+1} = \underline{x}^t_{,k}$

If $E(\underline{x}^t_{,k}) - E(\underline{x}^t) > 0$,

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

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

The book gives examples (Liu, p 109) with $\gamma = 0$. Sequential sampling can be used — ie express $\pi(\underline{x}) = \pi(x_d)\,\pi(x_{d-1}|x_d)\ldots\pi(x_0|x_1)$ — to get samples directly for comparison.

Suppose we want to estimate $\sum_{\underline{x}} \left(\sum_{i=0}^{d} x_i\right)\pi(\underline{x})$.

This is the Magnetization M.

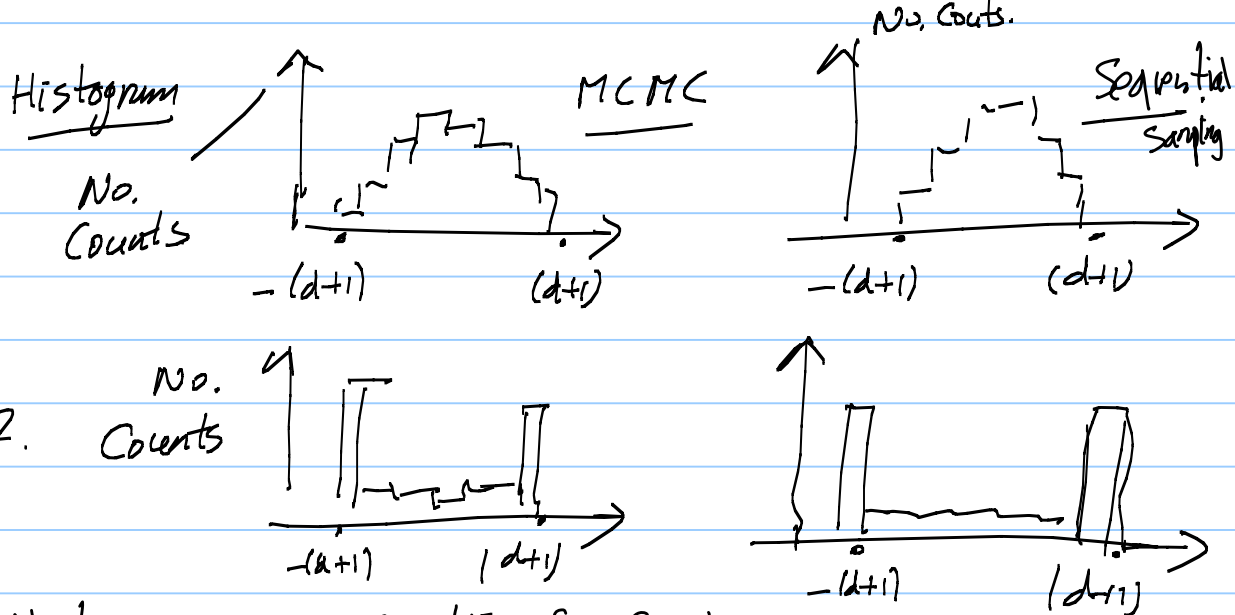
Consider $M^t = \sum_{i=0}^{d} x_i^t$, where $(x_0^t, \ldots x_d^t)$ is the t^{th} sample of the MCMC

note $-(d+1) \le M^t \le (d+1)$.

Run Metropolis for 1,000,000 iterations

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

to get 20,000 samples and plot their histograms.



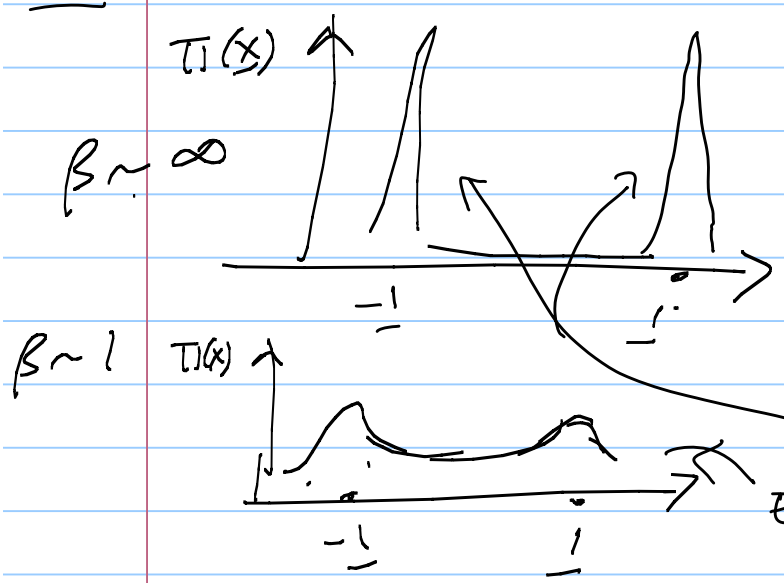
(i) Histograms are similar for $\beta = 1$. MCMC gives similar results to sequential sampling (but slower computation)

Page 5. (ii) Histograms for MCMC and sequential sampling differ more for $\beta = 2$. Also both histograms differ greatly from the histograms at $\beta = 1$.

Why?

First: Why are MCMC and sequential samples similar for $\beta = 1$ but less similar for $\beta = 2$?

Note: The distribution $\Pi(\underline{x})$ gets sharper as β gets larger.



$\underline{1} = (1, 1, \ldots 1)$ (d+1)

$-\underline{1} = (-1, -1, \ldots -1)$

Harder for MCMC to sample for large β. It is difficult to move between the two peaks.

Easier to sample from.

So, for $\beta = 1$, the MCMC has converged to good samples from $\Pi(\underline{x})$ at $N = 50$ and so. sample at $N = 100, 150, 200, \ldots 50k, \ldots$ are also good.

But, for $\beta = 2$, the MCMC needs a longer time to converge because of the difficultly of sampling both of the sharp peaks.

You could get better results by using $N = 1000, 2000, 3000, 4000, \ldots 1000k, \ldots$

Page 6. Second, why are the magnetization histograms very different from $\beta=1$ and $\beta=2$?

Answer: Phase Transition.

$$P_r(M=z) = \sum_{x_0 \cdots x_d} \delta_{x_0 + \cdots + x_d, z} \, \Pi(\underline{x})$$

x: any value

δ: Kronecker delta (Indicator Function)

Phase factor - number of ways you can get $M=z$.

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

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

Hence $P(M=d+1) = P(M=-(d+1)) = \Pi(\underline{1}) = \Pi(\underline{-1})$

For $M = d$, impossible

$M = d-1$, $(d+1)$ ways $(-1,1,1\ldots1)$, $(1,-1,1,\ldots)$ etc.

$$P(M=d-1) = \Pi(-1,1,\ldots1) + \Pi(1,-1,1,\ldots) + \ldots + \Pi(1,1,\ldots,1,-1)$$

For $M=0$, an exponentially large no. of ways provided $d+1$ is an even number.

So small phase factor for $M = \pm(d+1)$

enormous phase factor for $M=0$.

$P(M=0)$ is the sum of an exponential number of terms. Each term may be small, but they can add up to a big number.

Conflicting "forces"

Phase factor wants $M=0$.

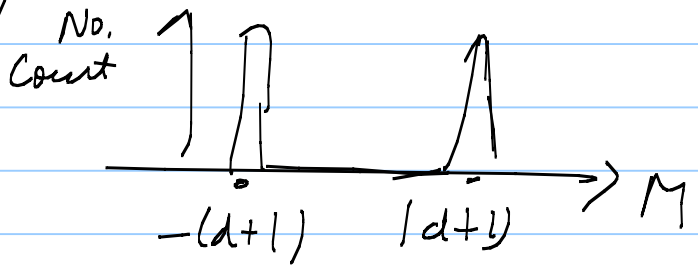
Probability $\Pi(\underline{x})$ wants $M = d+1$ or $-(d+1)$

(Page 7) For large β, the probability $\pi(x)$ is very sharp. The probabilities 'win' and so $P(M)$ is peaked at $M = \pm (d+1)$

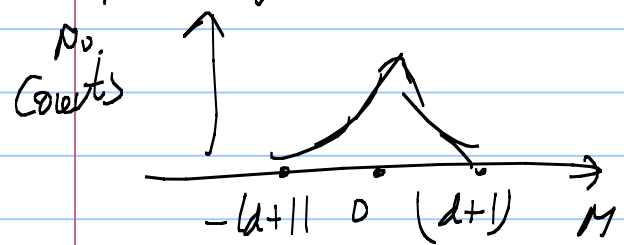
For small β, the probability $\pi(x)$ is smoother. The phase factor 'wins' and so $P(M)$ is peaked at $M = 0$.

This agrees with the histograms of M (our estimate of $P(M)$ make from the samples)

Ex for large β, probabilities win



For small β, "phase factors win.



Summary (1) the number of samples needed for the MCMC to converge will be very depending on the form of the distribution (e.g. differ from $\beta = 1$ to $\beta = 2$)

(2) MCMC can give similar results to sequential sampling, but can be extended to cases where sequential doesn't work.

(3) Probability distributions may have phase transitions. Properties - e.g. $P(M)$ - can change greatly depending on parameters