Introduction to Monte Carlo

Astro 542

Princeton University

Shirley Ho

Agenda

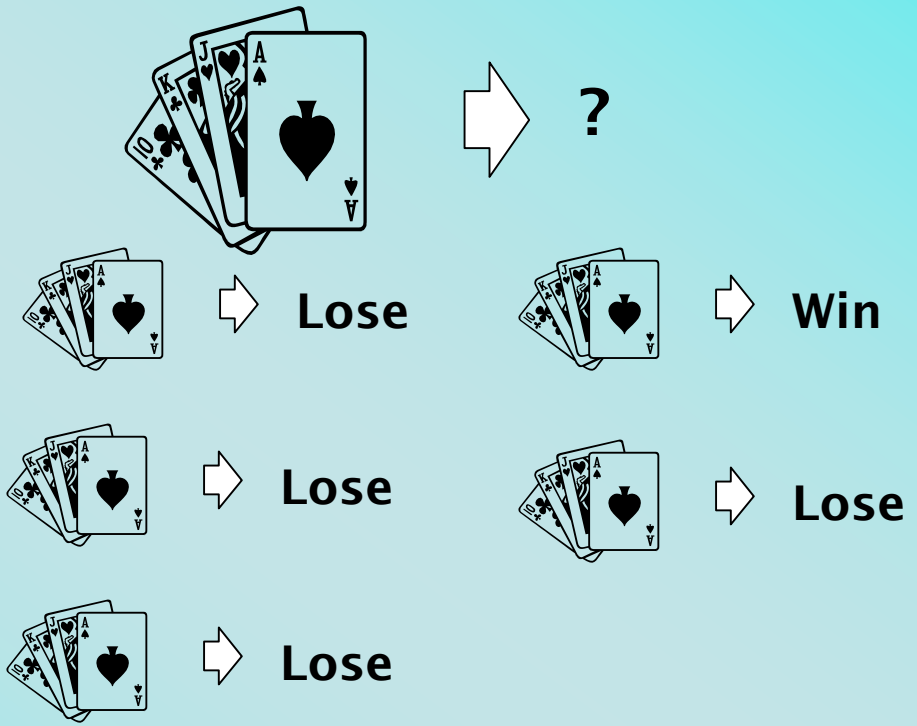
- Monte Carlo -- definition, examples
- Sampling Methods (Rejection, Metropolis, Metropolis-Hasting, Exact Sampling)
- Markov Chains -- definition,examples
- Stationary distribution
- Markov Chain Monte Carlo -- definition and examples

Monte Carlo -- a bit of history

- Credit for inventing the Monte Carlo method often goes to Stanislaw Ulam, a Polish born mathematician who worked for John von Neumann on the United States Manhattan Project during World War II.
- Ulam is primarily known for designing the hydrogen bomb with Edward Teller in 1951.
- He invented the Monte Carlo method in 1946 while pondering the probabilities of winning a card game of solitaire.
- (Rumors: That's why it is called Monte Carlo (referred to the city of Monte Carlo in Monaco where lots of gambling go on))

Monte Carlo Method

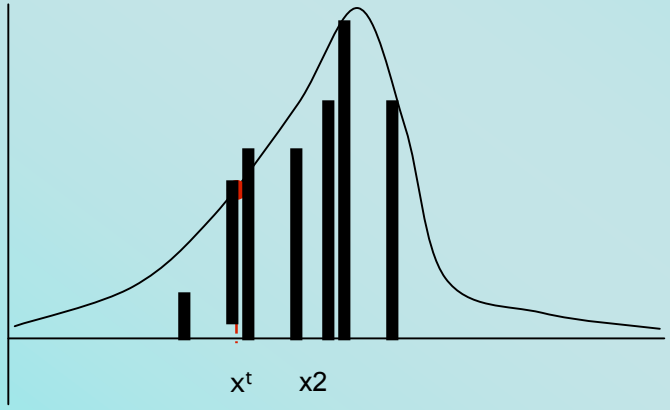
- Consider the game of solitaire: what's the chance of winning with a properly shuffled deck?
- Hard to compute analytically because winning or losing depends on a complex procedure of reorganizing cards
- Insight: why not just *play a few hands*, and see empirically how many do in fact win?
- More generally, can approximate a probability density function using only samples from that density



Chance of winning is 1 in 5!

Monte Carlo principle

- Given a very large set *X* and a distribution p(*x*) over it
- We draw i.i.d. (independent and identically distributed) a set of N samples
- We can then approximate the distribution using these samples



$$\mathrm{p}_N(x) = \frac{1}{N}\sum_{i=1}^{N} 1(x^{(i)} = x) \quad \underset{N\to\infty}{\longrightarrow} \mathrm{p}(x)$$

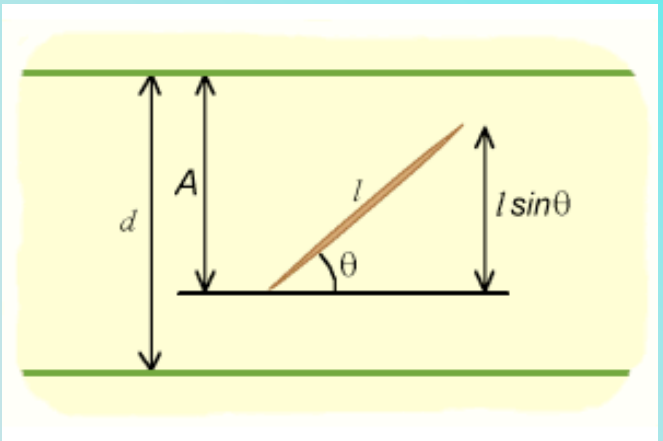
Monte Carlo Principle

We can also use these samples to compute expectations

$$E_N(f) = \frac{1}{N}\sum_{i=1}^{N} f(x^{(i)}) \underset{N\to\infty}{\rightarrow} E(f) = \sum_x f(x)\,\mathrm{p}(x)$$

Examples: Buffon needles

- More than 200 years before Metropolis coined the name Monte Carlo method, George Louis Leclerc, Comte de Buffon, proposed the following problem.
- If a needle of length *l* is dropped at random on the middle of a horizontal surface ruled with parallel lines a distance *d* > *l* apart, what is the probability that the needle will cross one of the lines?



Buffon asked what was the probability that the needle would fall across one of the lines, marked here in green. That outcome will occur only if A < l sin (theta)

Buffon's needle continued…

- The positioning of the needle relative to nearby lines can be described with a random vector which has components A and θ. The random vector (A,θ) is uniformly distributed on the region [0,d) [0, π). Accordingly, it has probability density function 1/(dπ).
- The probability that the needle will cross one of the lines is given by the integral

$$\int_0^{\pi}\int_0^{l\sin\theta} \frac{1}{d\pi}\, dA\, d\theta = \frac{2l}{d\pi}$$

- Suppose Buffon's experiment is performed with the needle being dropped *n* times. Let *M* be the random variable for the number of times the needle crosses a line, then the probability of the needle crossing the line will be: E(M)/n
- Thus :

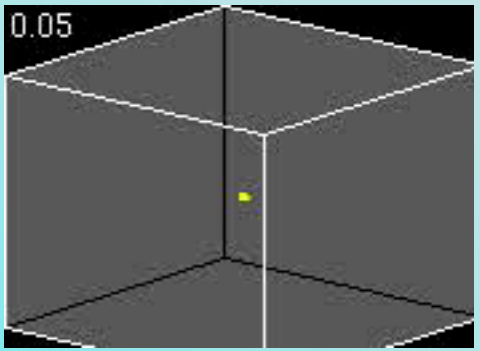
$$\pi = \frac{n}{E(M)}\frac{2l}{d}$$

Applications of Monte Carlo

- Example1: To understand the behavior of electrons in a semi-conductor materials, we need to solve Boltzmann Transport equation:

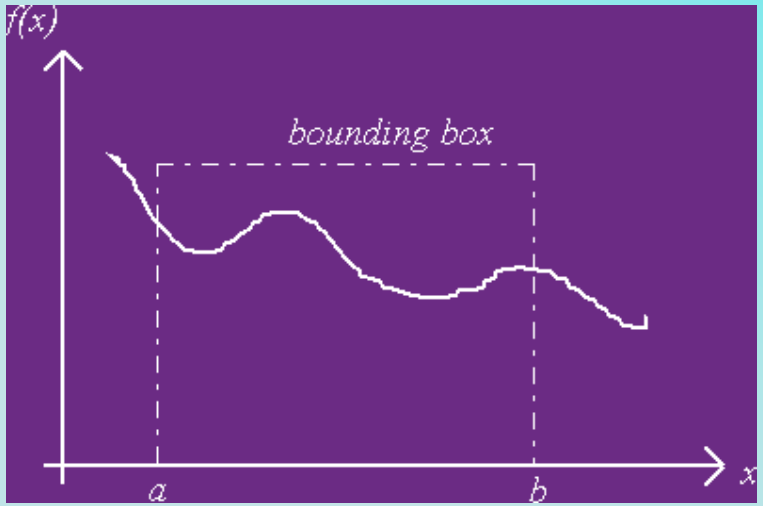
$$\frac{\partial f}{\partial t} + \bar{v} \cdot \nabla_{\bar{r}} f + \bar{F} \cdot \nabla_{\bar{p}} f = s(\bar{r}, \bar{p}, t) + \left.\frac{\partial f}{\partial t}\right|_{collision}$$

- To do this, we need to integrate some complicated functions and that's where Monte Carlo methods come in. But before doing the hard stuff, let's watch the outcome of using Monte Carlo method to understand the electrons in a pure silicon crystal at 300K



How did we integrate using Monte Carlo method then?

- Pairs of random numbers can be transformed into coordinates uniformly distributed within the box. The fraction of coordinates that falls below the function multiplied with the area of the limiting box, gives the solution of the integral.
- The accuracy of the solution depends on the number of random numbers used.
- The exact solution will be found within some interval around the result obtained by the Monte Carlo method. For an infinite number of coordinates the solution will be exact.



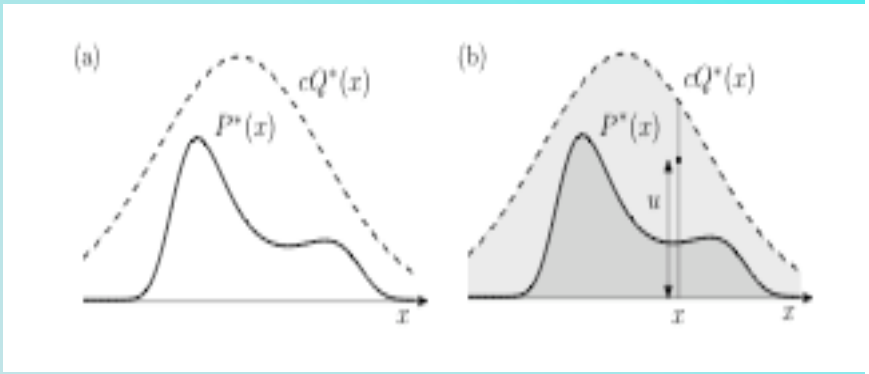
Sampling Methods

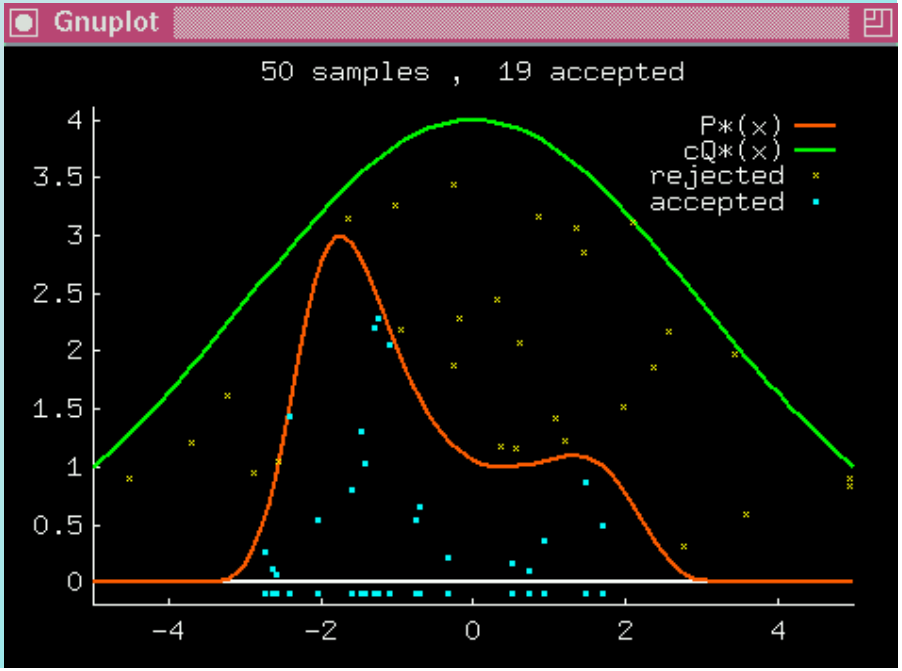
- Here, we will talk about the sampling methods: Rejection, Metropolis and exact sampling.
- Why do we need to know about sampling?
- Correct samples from P(x) will by definition tend to come from places in x-space where P(x) is big; how can we identify those places where P(x) Is big, without evaluating P(x) everywhere (which takes a lot of time especially in higher dimension systems)?

Rejection Sampling

- We would like to sample from p(x), but it's easier to sample from a *proposal distribution* q(x)
- A proposal distribution is a simpler distribution that we sample from
- q(x) satisfies p(x) ≤ M q(x) for some $M<\infty$
- Procedure:
  - Sample $x^{(i)}$ from q(x)
  - Accept with probability p($x^{(i)}$) / Mq($x^{(i)}$)
  - Reject otherwise
- The accepted $x^{(i)}$ are sampled from p(x)!
- Problem: it works well only if p(x) and q(x) are similar!

  and we have no easy (and fast) way to make sure that these two distributions are similar.
- If M is too large, we will rarely accept samples !
- In high dimensional space, you will have too much to sample from



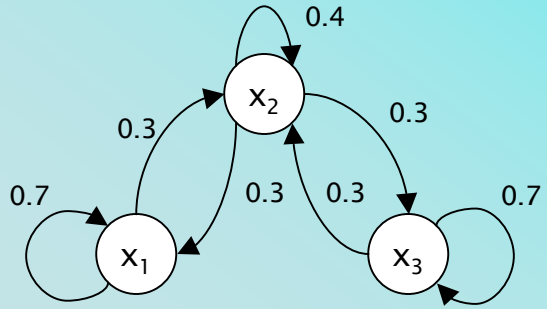


Transition…

- Since we will need to understand state diagrams and transition between states to talk about the following two sampling methods (Metropolis, Metroplis-Hasting and exact sampling)
- I will switch gear here to introduce Markov Chains first before we come back to the sampling methods

Markov Chains

$$T = \begin{pmatrix} 0.7 & 0.3 & 0 \\ 0.3 & 0.4 & 0.3 \\ 0 & 0.3 & 0.7 \end{pmatrix}$$



- Markov chain on a space ***X*** with transitions ***T*** is a random process (infinite sequence of random variables) $(x^{(0)}, x^{(1)}, \ldots x^{(t)}, \ldots) \in \boldsymbol{X}^{\infty}$ that satisfy

$$p(x^{(t)} \mid x^{(t-1)}, \ldots, x^{(1)}) = T(x^{(t-1)}, x^{(t)})$$

- T is the transition probability matrix, P is the probability for x to be in state $x^{(t)}$ given the history of the state.
- That is, the probability of being in a particular state at time *t* given the state history depends only on the state at time *t*-1 --> Memoryless

Markov chain for sampling

- In order for a Markov chain to useful for sampling p(x), we require that for any starting state $x^{(1)}$

$$\mathrm{p}^{(t)}_{x^{(1)}}(x) \underset{t\to\infty}{\to} \mathrm{p}(x)$$

- Stationary distribution (π) : for any pairs of state i,j : $\pi_i T_{ij} = \pi_j T_{ji}$
- Equivalently, the stationary distribution of the Markov chain must be p(x)

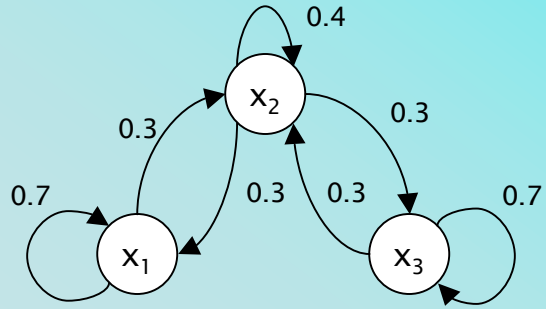
$$[\mathrm{p}\,\mathbf{T}](x) = \mathrm{p}(x)$$

- If this is the case, we can start in an arbitrary state, use the Markov chain to do a random walk for a while, and stop and output the current state $x^{(t)}$
- The resulting state will be sampled from p(x)!

Stationary distribution example:

Consider the Markov chain given above:

$$T = \begin{pmatrix} 0.7 & 0.3 & 0 \\ 0.3 & 0.4 & 0.3 \\ 0 & 0.3 & 0.7 \end{pmatrix}$$



- The stationary distribution is

$$\begin{pmatrix} 0.33 & 0.33 & 0.33 \end{pmatrix} \times \begin{pmatrix} 0.7 & 0.3 & 0 \\ 0.3 & 0.4 & 0.3 \\ 0 & 0.3 & 0.7 \end{pmatrix} = \begin{pmatrix} 0.33 & 0.33 & 0.33 \end{pmatrix}$$

- Some samples:

1,1,2,3,2,1,2,3,3,**2**
1,2,2,1,1,2,3,3,3,**3**
1,1,1,2,3,2,2,1,1,**1**
1,2,3,3,3,2,1,2,2,**3**
1,1,2,2,2,3,3,2,1,**1**
1,2,2,2,3,3,3,2,2,**2**

Markov chain and sampling

Claim: To ensure that the stationary distribution of the Markov chain is p(x) it is sufficient for p and *T* to satisfy the *detailed balance* (*reversibility*) condition

$$\mathrm{p}(x)T(x,y) = \mathrm{p}(y)T(y,x)$$

Proof: for all *y* we have

$$[\mathrm{p}\,\mathbf{T}](y) = \sum_x \mathrm{p}(x)T(x,y) = \sum_x \mathrm{p}(y)T(y,x) = \mathrm{p}(y)$$

-> stationary distribution!

Once we know that it is a stationary distribution, we can then take the samples from the stationary distribution and it should reflect p(x) if we create the Markov chain correctly.

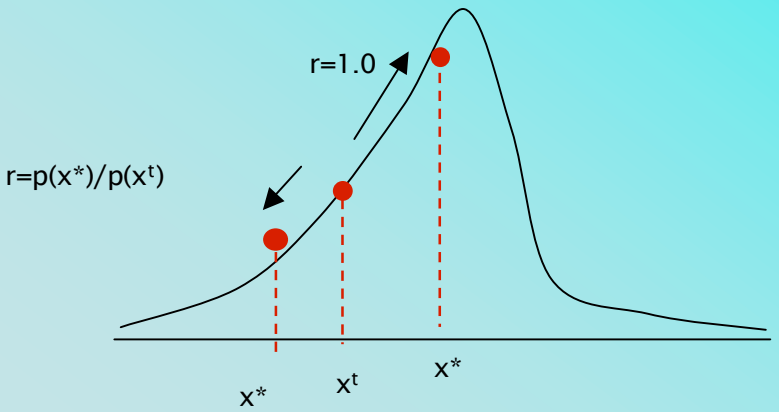
! Recall that we want to integrate efficiently some difficult functions, and we want to use Monte Carlo integration, but we don't want to sample around the regions where the probability of accepting is low, now with Markov chains, we can sample more efficiently!

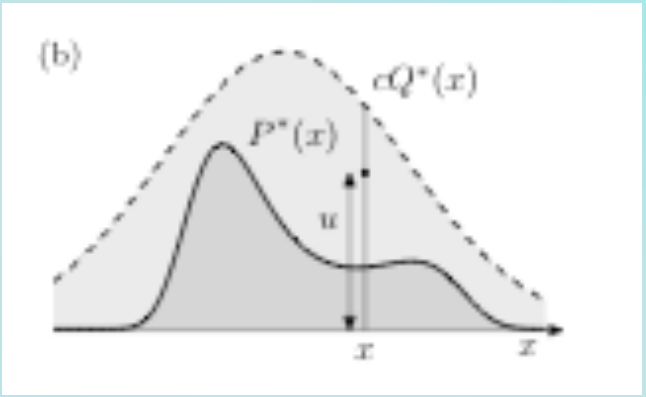
Metropolis algorithm

- Suppose our distribution p(x) is easy to sample, and easy to compute *up to a normalization constant*, but hard to compute exactly
  - We tried using Rejection Sampling to sample p(x), but in high dimensional space, there are too many samples that is being rejected-> BAD
  - So, we can use a Markov Chain with the following algorithm to make sure that when we sample, we stay very closely around where the p(x) is high, thus most of our samples will be accepted (when you sample from the Markov chain).
  - How do we do that?
  - We define a Markov chain with the following process:
  - Sample a candidate point x* from a *proposal distribution* $q(x^*|x^{(t)})$ which is *symmetric*: q(x|y)=q(y|x)
  - Compute the *ratio*: $r = \frac{\mathrm{p}(x^*)}{\mathrm{p}(x^{(t)})}$
  - With probability min(*r*,1) transition to x*, otherwise stay in the same state

How does Metropolis work?

- Why does the Metropolis algorithm work?
  - Proposal distribution can propose anything it likes (as long as it can jump back with the same probability)
  - Proposal is always accepted if it's jumping to a more likely state
  - Proposal accepted with the ratio if it's jumping to a less likely state
- The acceptance policy, combined with the reversibility of the proposal distribution, makes sure that the algorithm explores states in proportion to p(*x*)!





Detailed Balance

Looking at Metropolis Algorithm and assume that p(y)>=p(x) without loss of generality

$$\begin{aligned} \mathrm{p}(x)T(x,y) &= \mathrm{p}(x)q(y \mid x) \\ &= \mathrm{p}(x)q(x \mid y) \\ &= \mathrm{p}(y)q(x \mid y)\frac{\mathrm{p}(x)}{\mathrm{p}(y)} \\ &= \mathrm{p}(y)T(y,x) \end{aligned}$$

Detailed balance!

Other Sampling Methods

- Metroplis-Hasting: We do not require the proposal distribution to be symmetric (q(x|y =q(y|x)) that Metropolis needs and instead use:

$$r = \frac{\mathrm{p}(x^*)\,\mathrm{q}(x^{(t)} \mid x^*)}{\mathrm{p}(x^{(t)})\,\mathrm{q}(x^* \mid x^{(t)})}$$

  as the ratio to determine whether we accept or not.
- Gibbs Sampling: A special case of Metropolis-Hasting, but we use the conditional $P(x_j|x_i)$ instead.

Practicalities

- Can we predict how long a Markov chain Monte Carlo simulation will take to equilibrate? (reaching the stationary distribution)

->By considering the random walks involved in a MCMC simulation, we can obtain simple lower bounds on the time required for convergence. (say the length scale of the state space is L (the curvature of the pdf), and step size is s, then you will need T steps = $(L/s)^{1/2}$ before the chain equilibrate)

- But predicting this time more precisely is a difficult problem.
- **Exact sampling** offers a solution to this problem and we will talk about this later.

More practicalities

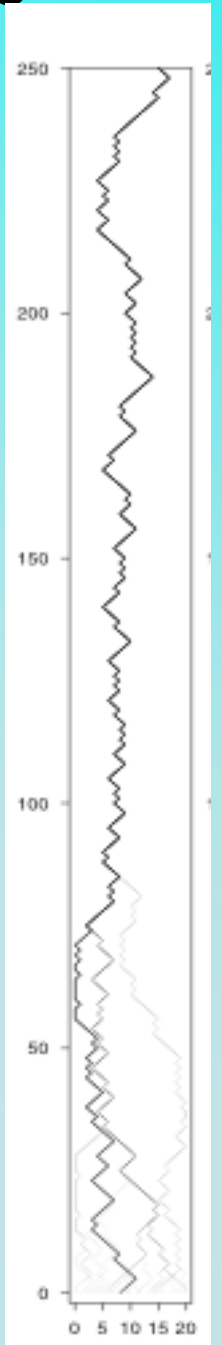
- Can we diagnose or detect convergence in a mcmc ? -> very difficult
- Can we speed up the convergence?

-> Hamiltonian Monte Carlo

-> Overrelaxation

-> Simulated annealing

How do we know the Markov chains have reached the equilibrated state?
--> Exact sampling



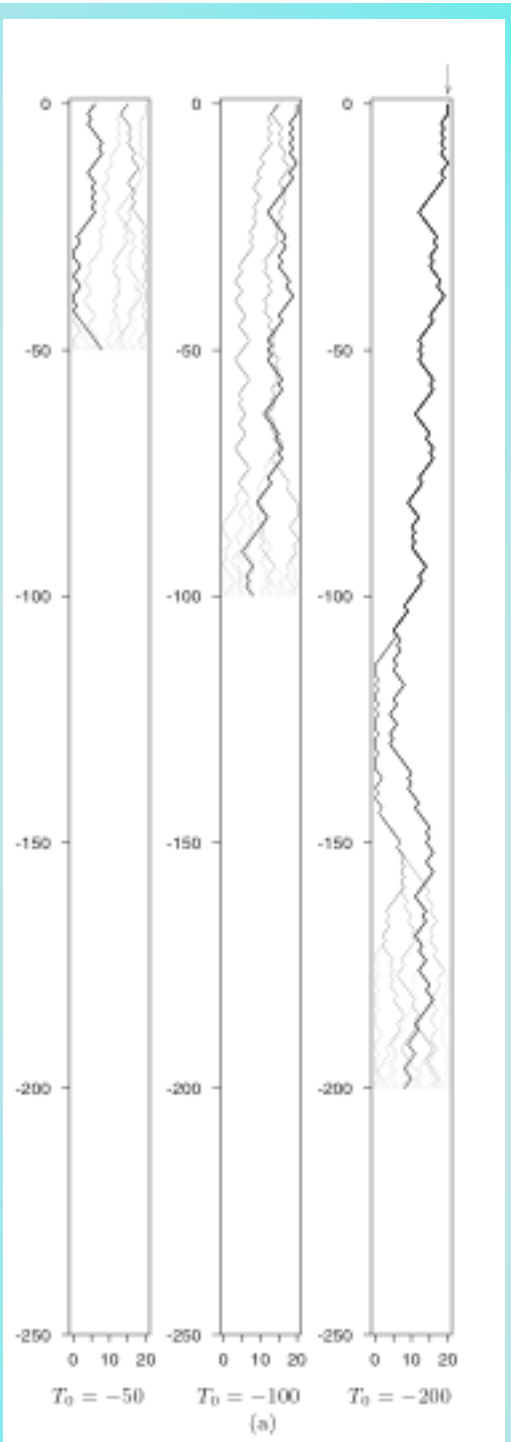
- We will know for sure for some Markov chains that they have reached the stationary states using exact sampling:
- 3 Big ideas:

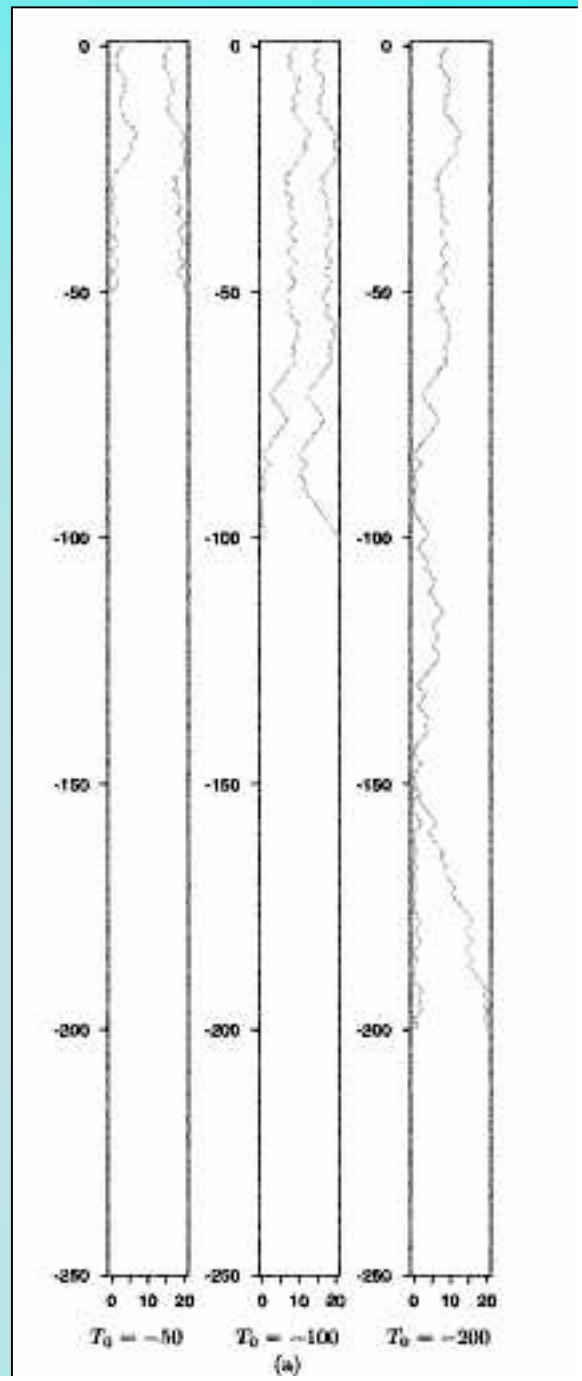
1) When the Markov chains (with the same random number seed) meet, then the chains will not separate anymore

Exact Sampling:

2) Couplings from the past: Let the Markov Chains run and then stop at time t=0, if not all the chains meet together, then start the simulation at t=-T using the same random number seeds (so that the same situation is modeled)

3) For some Markov Chains, they never cross each other, then we just need to look at the maximal and the minimal state and look for the time when they meet.

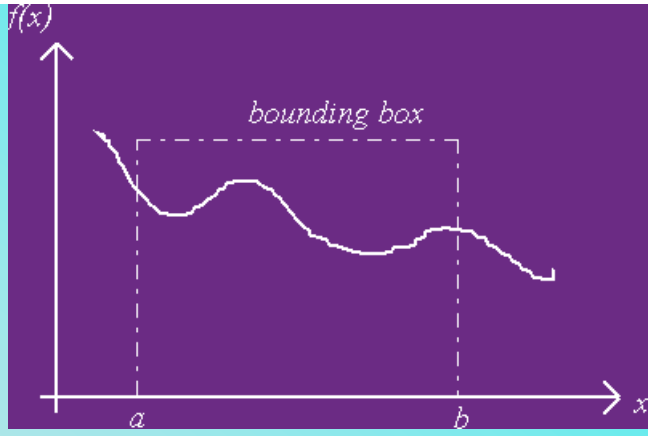




Exact sampling: Ising Model

- Exact sampling is very useful especially for spins.
- Spins can be for example electrons that have simple +1/2 or -1/2 spin.
- Since for some of the spins systems, they can have a maximal and a minimal state, so that we can feasibly use exact sampling.
- Example: Ferromagnetic system (4 spins only):
- You can order the states from ++++ to ----
- One can evolve the spin chain to the final equilibrated state in a way such that we only consider the maximal and minimal state, and then wait until the Markov chain of both of these states converge, then we know that we will get exact samples from that equilibrated spin chains

MCMC in Action:



- WMAP!
- Used Markov Chain Monte Carlo to get the cosmological parameters with their confidence levels.
- To get the confidence levels of parameter a_i, one has to marginalize over all the other parameters in the parameter space (thus doing an integration!! -> Recall how we did Monte Carlo integration!)
- We need to get the area underneath a certain unknown function, thus, sample the space around the curve (in the rectangular box) and in this case, we are going to use Markov Chain (produced using Metropolis Algorithm) Monte Carlo to get the samples.

Markov Chain Monte Carlo Analysis of WMAP

1) Start with a set of cosmological parameters {$\boldsymbol{a_1}$} , compute the likelihood $L_1 = L(a_1 \mid \hat{C}_l)$
   $\hat{C}_l$ denotes the best estimator of the C_l^{sky}

2) Take a random step in parameter space to obtain a new set of cosmological parameters {$\boldsymbol{a_2}$} . The probability distribution of the step is taken to be Gaussian in each direction i with r.m.s given by σ_i .
   We will refer below to σ_i as the "step size". The choice of the step size is important to optimize the chain efficiency.
3) Compute L_2 for the new set of cosmological parameters.

More on WMAP analysis

4.a) If $L_2 > L_1$, “take the step” i.e. save the new set of cosmological parameters $\{a_2\}$ as part of the chain, then go to step 2 after the substitution $\{a_1\} \rightarrow \{a_2\}$.

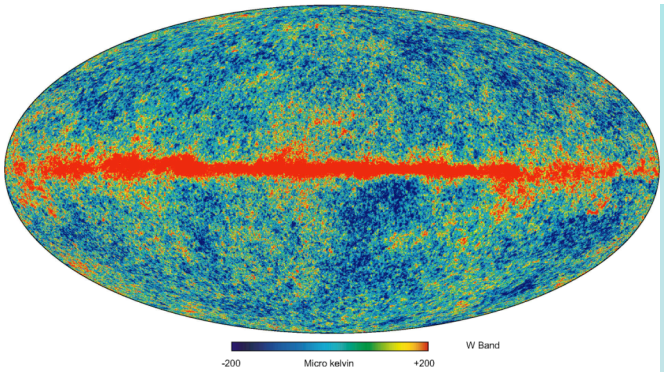
4.b) If $L_2 < L_1$, draw a random number x from a uniform distribution from 0 to 1

-> If $x >= L_2/L_1$ “do not take the step”, i.e. save the parameter set $\{a_1\}$ as part of the chain and return to step 2.

-> If $x < L_2/L_1$, “ take the step”, i.e. do as in 4.a).

5) For each cosmological model run four chains starting at randomly chosen, well-separated points in parameter space.

6)When the convergence criterion is satisfied and the chains have enough points to provide reasonable samples from the a posteriori distributions, stop the chains (we will talk about this step in details the next slide)



Convergence and Mixing (in WMAP analysis)

- How to test convergence?
- Let us consider running M chains, each with 2N elements, and we only consider the last N: $\{y^j_i\}$
- (i runs from 1 to N, j runs from 1 to M)
- Mean of the chain:

$$\bar{y}^j = \frac{1}{N}\sum_{i=1}^{N} y_i^j$$

Mean of Distribution:

$$\bar{y} = \frac{1}{NM}\sum_{ij=1}^{NM} y_i^j$$

- Variance within the chain:

$$W = \frac{1}{M(N-1)}\sum_{ij=1}^{NM}(y_i^j - \bar{y}^j)^2$$

Variance between chains

$$B_n = \frac{1}{M-1}\sum_{j=1}^{M}(\bar{y}^j - \bar{y})^2$$

More on convergence and mixing

- We can define this ratio:

$$R = \frac{\frac{N-1}{N}W + B_n(1+\frac{1}{M})}{W}$$

  numerator-> estimate of the variance that is unbiased if the distribution is stationary (otherwise, overestimate)

  denominator-> underestimate of the variance of target distribution if individual chains have not converged

- When the ratio is nearly 1, that means the chain has basically achieved convergence. WMAP uses R <1.1
- WMAP only samples the chain after it reaches this ratio, thus making sure that it has converged.
- Mixing: It also uses all the points after the first 200 points in the chain and WMAP claims that using at least 30,000 points in each chains is good enough to produce marginalized likelihood for all the parameters.
- This is tested (from David Spergel) by computing the 1-sigma contour of the chains independently and make sure that they agree with each other to within 10%.

Analysis of WMAP cont'd

- Finally, after all the Markov Chains finished running, we need to find the marginalized likelihood for all the parameters!
- First, we have now found a M-dimensional confidence region (say we have M parameters).
- Recall from previous talks: A confidence region (or confidence interval) is just a region of that M-dimensional space (hopefully a small region) that contains a certain (hopefully large) percentage of the total probability distribution.
- Example: You point to a 99% confidence level region and say, e.g., "there is a 99 percent chance that the true parameter values fall within this region around the measured value."

Analysis of WMAP cont'd

Expectation value of each parameter:

$$< a_i >= \int L a_i d\vec{a}$$

How do we do this integral?

->Since the underlying distribution function is so complicated and we want basically an integration over all the other parameters (say you only want 1 parameter 'w')

-> Monte Carlo integration!!

-> And we already have a nicely understood systems of Markov Chains and we can sample from states of the Markov Chains.

Recall that the Markov Chains are designed to walk and stay around where the P(a) is large, so then when we sample the Markov Chain, it will give us a good sampling of the distribution function P(a).

Marginalized Likelihood cont'd

Marginalized Likelihood for one parameter:

$$\int L a_i d\vec{a} = \frac{1}{N} \sum_t a_{t,i}$$

And here N is the number of points in the Markov Chain that has stabilized, $a_{t,i}$ is the value of parameter at the t-th step of the chain

- This can be easily understood as the MCMC gives to each point in the parameter space a "weight"/ probability proportional to the number of steps the chain has spent at that particular location.
- The 100(1-2p)% confidence interval $[c_p, c_{1-p}]$ for a parameter is estimated by setting c_p to the p^{th} quantile of $a_{t,i}$ and c_{p-1} to the $(1-p)^{th}$ quantile of $a_{t,i}$.
- -> DONE !!

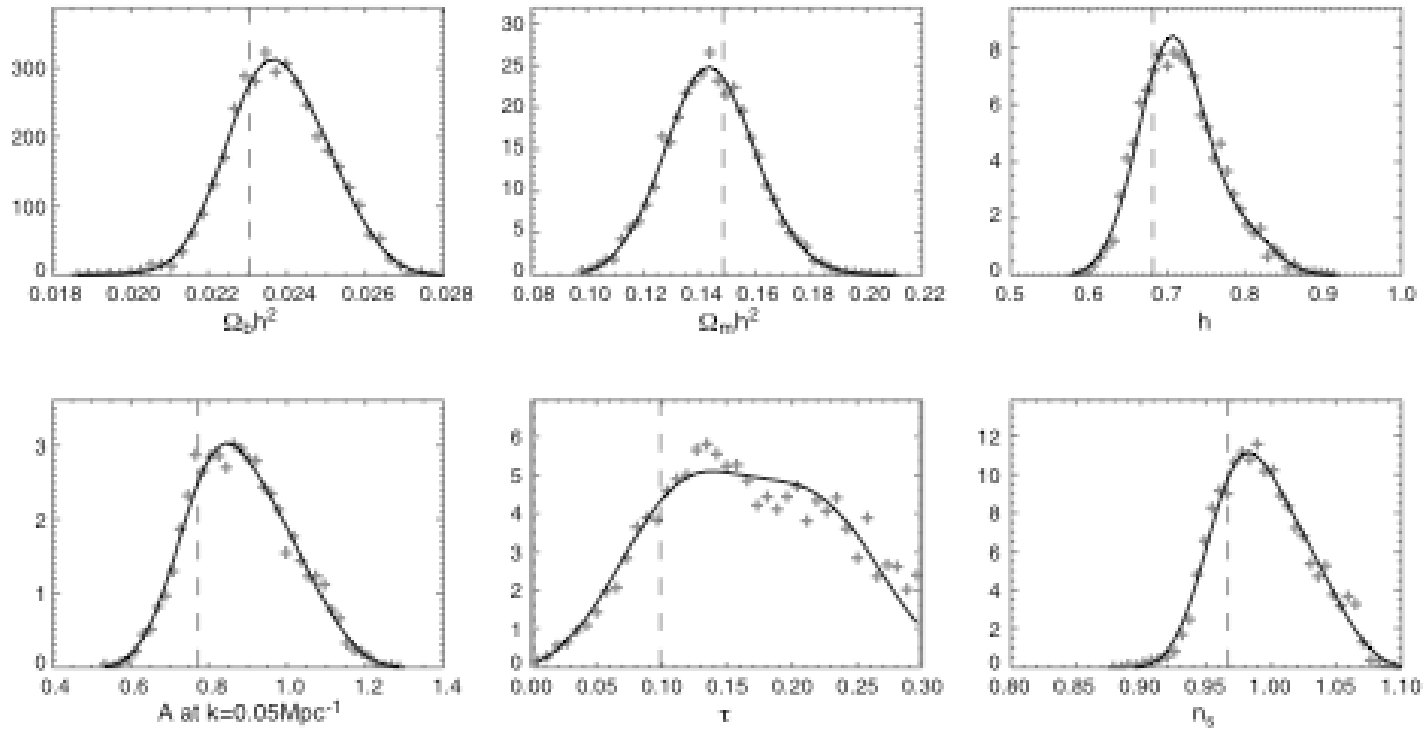


Fig. 3.— This figure shows the likelihood function of the *WMAP* TT + TE data as a function of the basic parameters in the power law ΛCDM *WMAP* model. ($\Omega_b h^2$, $\Omega_m h^2$, h, A, n_s and τ.) The points are the binned marginalized likelihood from the Markov chain and the solid curve is an Edgeworth expansion of the Markov chains points. The marginalized likelihood function is nearly Gaussian for all of the parameters except for τ. The dashed lines show the maximum likelihood values of the global six dimensional fit. Since the peak in the likelihood, x_{ML} is not the same as the expectation value of the likelihood function, $< x >$, the dashed line does not lie at the center of the projected likelihood.

THANK YOU!!! All for coming!

- References:
- MacKay's book and website
- Mitzenmacher & Upfal, Probability and Computing: Randomized Algorithms and Probabilistic Analysis (not yet published …)
- Verde et al. 2003 , Spergel et al. 2003
- Jokes website: http://www.isye.gatech.edu/~brani/isyebayes/
- A lot of pictures from the web (not including the list here)

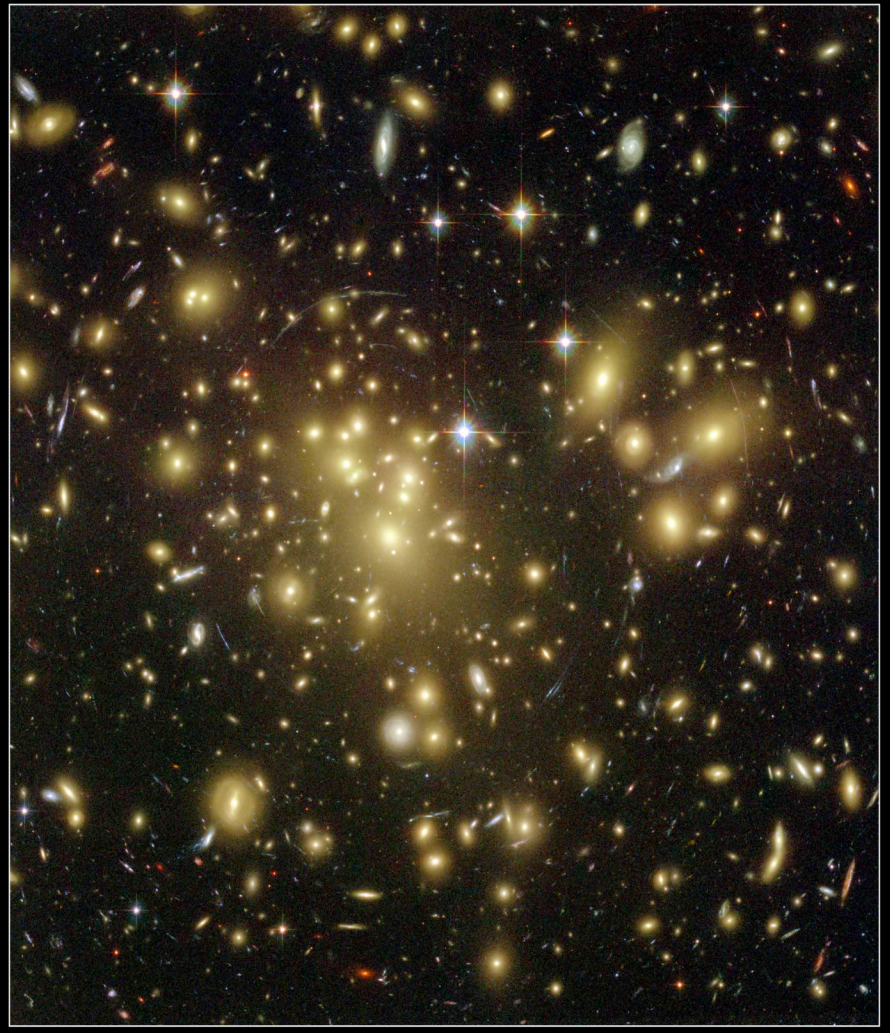
More Monte Carlo application

- Making fake dataset to test the algorithm :

  In high energy physics experiment, one generate a fake dataset using Monte Carlo to test if the algorithm is correct.

- To get the probability of some phenomena:

  In strong lensing simulation, you throw sources repeatedly randomly onto the source plane and ray-traced each source through the lensing cluster to get the probability of how often does the cluster lens.



Galaxy Cluster Abell 1689
Hubble Space Telescope • Advanced Camera for Surveys

Efficient Monte Carlo Methods

- Hamiltonian Monte Carlo
- Overrelaxation (only talk about this when Gibbs sampling is introduced)
- Simulated Annealing

Hamiltonian Monte Carlo

- If we write $P(\mathbf{x})=\exp(-E(\mathbf{x}))/Z$
- If we can augment the state space **x** with **p,** and sample from the joint density:
- $P(\mathbf{x},\mathbf{p})=1/Z_h \exp[-H(\mathbf{x},\mathbf{p})]$
- Algorithm hinges on: when we go to a lower energy state ($dH<0$ or a random number $< \exp(-dH)$), then we move to this new state.

Simulated Annealing

- Introduce a parameter T ~ temperature
- $P_T(x)=1/Z(T)\ \exp[-E_0(x)-E_1(x)/T]$
- So that we will have a well behaved function at high temperature defined by E_0
- Pros: Avoids going into very unlikely region and get stuck there (unrepresentative probability island)
- Cons: does not necessarily give you the sample from the exact distribution.

Gibbs Sampling

- When we can't really draw from P(x) directly since P(x) is too complicated
- But $P(x_j|x_i)$ where i<>j is tractable
- A special case of Metropolis-Hasting, but we use the conditional $P(x_j|x_i)$ instead.

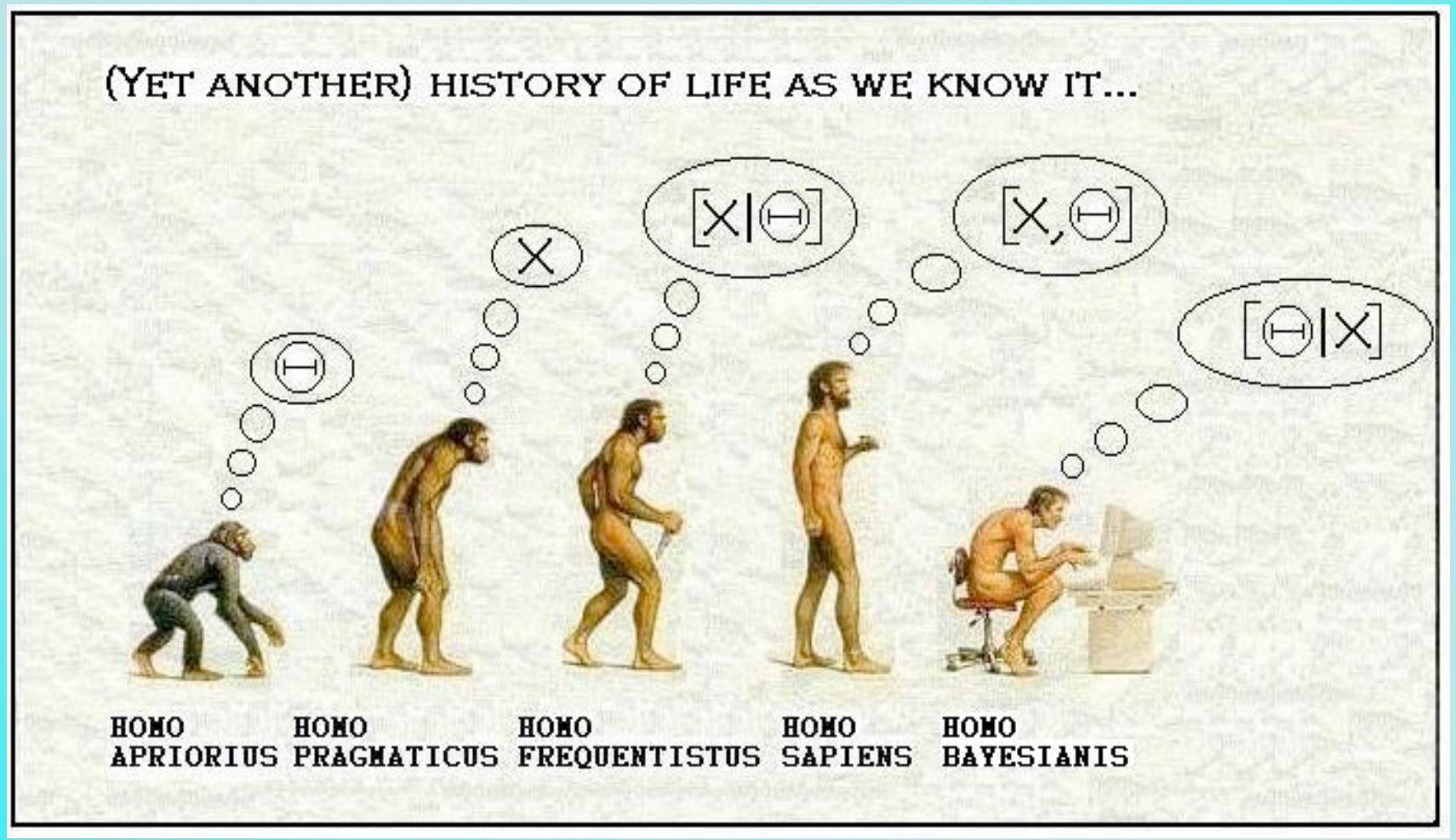
Metropolis-Hasting

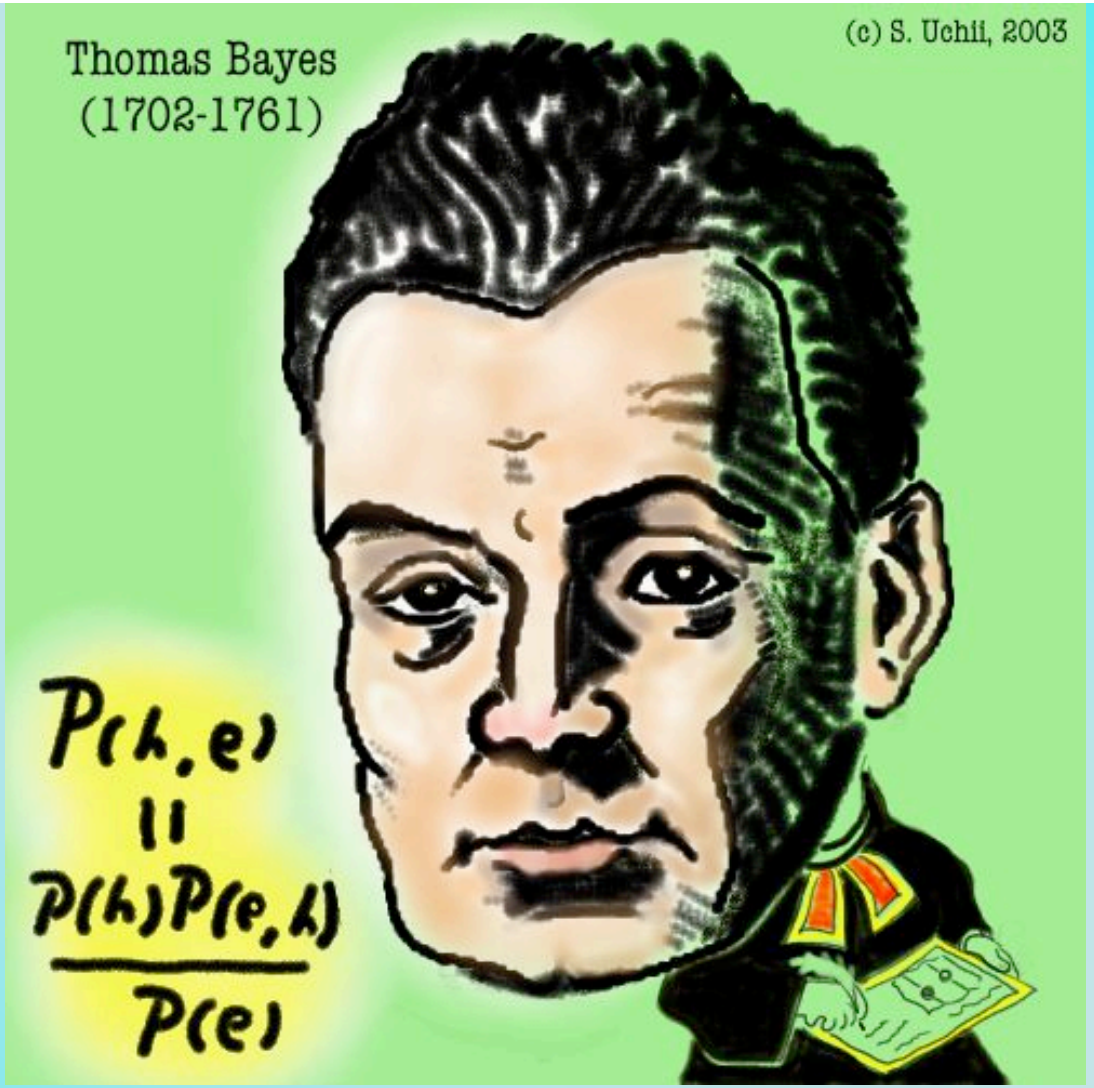
- The symmetry requirement of the Metropolis proposal distribution can be hard to satisfy
- Metropolis-Hastings is the natural generalization of the Metropolis algorithm, and the most popular MCMC algorithm
- We define a Markov chain with the following process:
  - Sample a candidate point x^* from a proposal distribution $q(x^*|x^{(t)})$ which is *not* necessarily symmetric
  - Compute the ratio:

$$r = \frac{p(x^*)\, q(x^{(t)} \mid x^*)}{p(x^{(t)})\, q(x^* \mid x^{(t)})}$$

  - With probability $\min(r,1)$ transition to x*, otherwise stay in the same state $x^{(t)}$
  - One can prove that it satisfy detailed balance too !

Break time again :)





X-> set of observed $\hat{C}_l$

$$P(a \mid x) = \frac{P(x \mid a)P(a)}{\int P(x \mid a)P(a)da}$$

a-> a set of cosmological parameters

Break time ….

From a Frequentist: A Bayesian is one who, vaguely expecting a horse, and catching a glimpse of a donkey, strongly believes he has seen a mule.

From a Bayesian(?):

A Bayesian and a Frequentist were to be executed. The judge asked them what were their last wishes.

The Bayesian replied that he would like to give the Frequentist one more lecture.

The judge granted the Bayesian's wish and then turned to the Frequentist for his last wish.

The Frequentist quickly responded that he wished to hear the lecture again and again and again and again.....