An Introduction to Markov Chain Monte Carlo

Teg Grenager

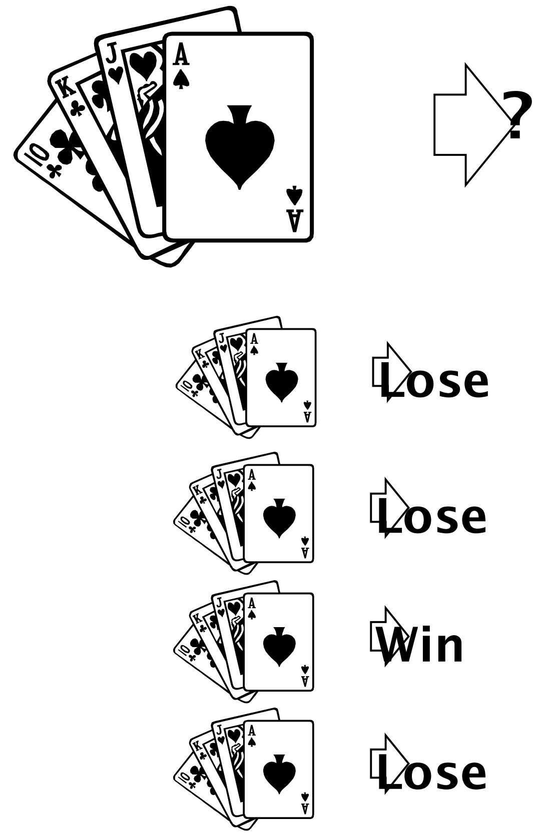
July 1, 2004

Agenda

- Motivation
- The Monte Carlo Principle
- Markov Chain Monte Carlo
- Metropolis Hastings
- Gibbs Sampling
- Advanced Topics

Monte Carlo principle

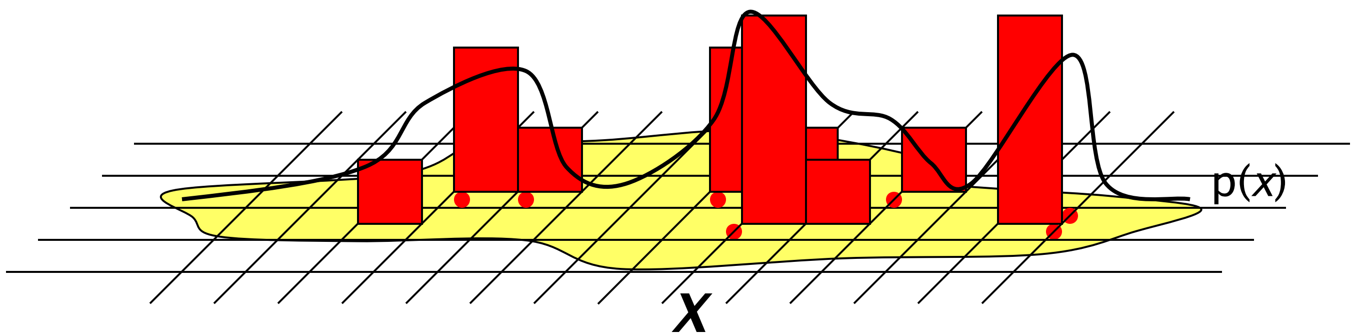
- Consider the game of solitaire: what's the chance of winning with a properly shuffled deck?
- Hard to compute analytically because winning or losing depends on a complex procedure of reorganizing cards
- Insight: why not just *play a few hands*, and see empirically how many do in fact win?
- More generally,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



Chance of winning is 1 in 4!

Monte Carlo principle

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



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

Monte Carlo principle

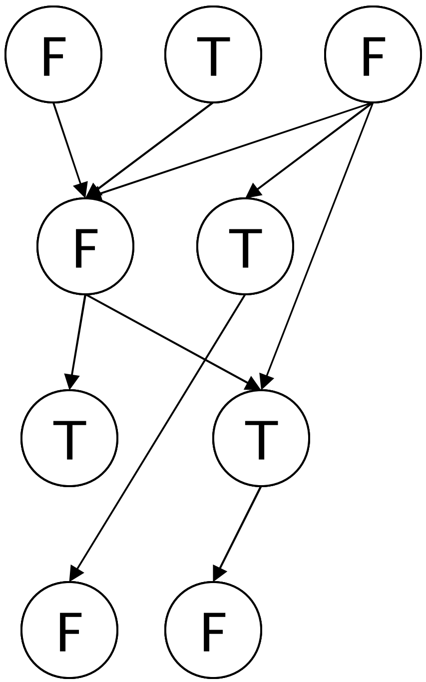
- We can also use these samples to compute expectations

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

- And even use them to find a maximum

$$\hat{x} = \underset{x^{(i)}}{\arg\max}[\mathrm{p}(x^{(i)})]$$

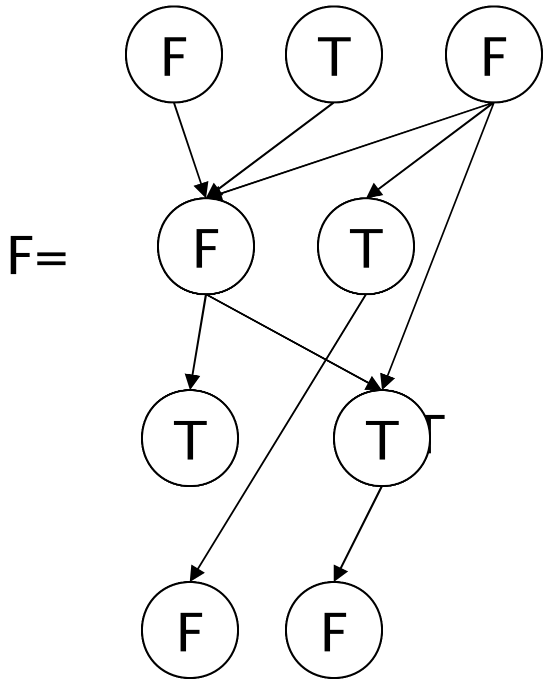
Example: Bayes net inference



Sample 1: FTFTTTFFT
Sample 2: FTFFTTTFF
etc.

- Suppose we have a Bayesian network with variables ***X***
- Our state space is the set of all possible assignments of values to variables
- Computing the joint distribution is in the worst case NP-hard
- However, note that you can draw a sample in time that is linear in the size of the network
- Draw *N* samples, use them to approximate the joint

Rejection sampling



Sample 1: FTFTTTFFT reject
Sample 2: FTFFTTTFF accept
etc.

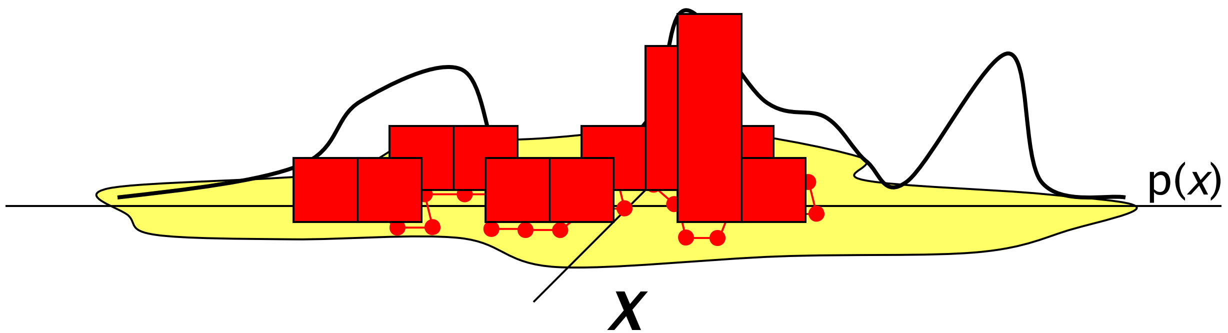
- Suppose we have a Bayesian network with variables ***X***
- We wish to condition on some evidence $\boldsymbol{Z} \subseteq \boldsymbol{X}$ and compute the posterior over $\boldsymbol{Y}=\boldsymbol{X}-\boldsymbol{Z}$
- Draw samples, rejecting them when they contradict the evidence in ***Z***
- Very inefficient if the evidence is itself improbable, because we must reject a large number of samples

Rejection sampling

- More generally, we would like to sample from $p(x)$, but it's easier to sample from a *proposal distribution* $q(x)$
- $q(x)$ satisfies $p(x) \le M\, q(x)$ 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: if M is too large, we will rarely accept samples
  - In the Bayes network, if the evidence ***Z*** is very unlikely then we will reject almost all samples

Markov chain Monte Carlo

- Recall again the set *X* and the distribution p(*x*) we wish to sample from
- Suppose that it is hard to sample p(*x*) but that it is possible to "walk around" in *X* using only local state transitions
- Insight: we can use a "random walk" to help us draw random samples from p(*x*)



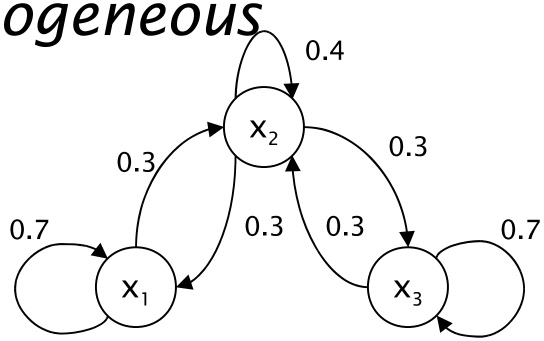
Markov chains

- 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

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

- 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
- If the transition probabilities are fixed for all t, the chain is considered *homogeneous*

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



Markov Chains for sampling

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

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

- Equivalently, the stationary distribution of the Markov chain must be p(x)

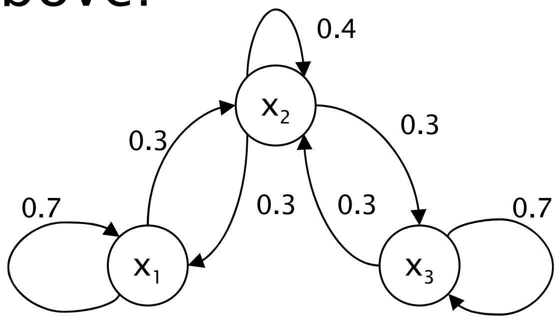
$$[p\mathbf{T}](x) = p(x)$$

- 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

- 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**

Empirical Distribution:

$$\begin{pmatrix} 0.33 & 0.33 & 0.33 \end{pmatrix}$$

Ergodicity

- Claim: To ensure that the chain converges to a unique stationary distribution the following conditions are sufficient:
  - *Irreducibility*: every state is eventually reachable from any start state; for all $x,y \in \boldsymbol{X}$ there exists a t such that

$$p_x^{(t)}(y) > 0$$

  - *Aperiodicity*: the chain doesn't get caught in cycles; for all $x,y \in \boldsymbol{X}$ it is the case that

$$\gcd\{t : p_x^{(t)}(y) > 0\} = 1$$

- The process is *ergodic* if it is both irreducible and aperiodic
- This claim is easy to prove, but involves eigenstuff!

Markov Chains for 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

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

- And thus p must be a stationary distribution of *T*

Metropolis algorithm

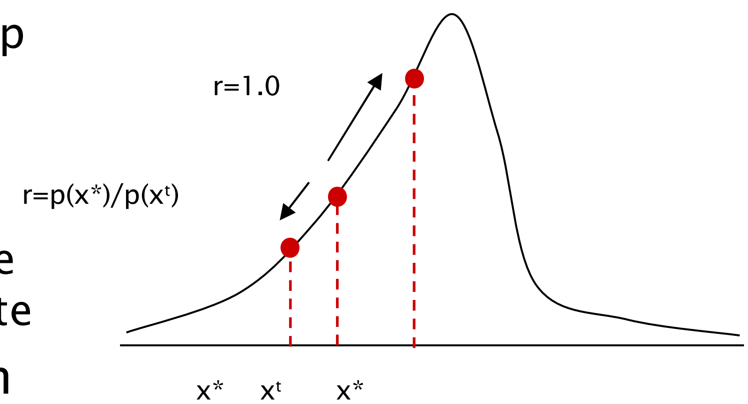
- How to pick a suitable Markov chain for our distribution?
- Suppose our distribution p(x) is easy to sample, and easy to compute *up to a normalization constant*, but hard to compute exactly
  - e.g. a Bayesian posterior $P(M|D) \propto P(D|M)P(M)$
- We define a Markov chain with the following process:
  - Sample a candidate point x* from a *proposal distribution* $q(x^*|x^{(t)})$ which is *symmetric*: $q(x|y)=q(y|x)$
  - Compute the *importance ratio* (this is easy since the normalization constants cancel)

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

  - With probability min(r,1) transition to x*, otherwise stay in the same state

Metropolis intuition

- 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 importance 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*)!
- Now, network permitting, the MCMC demo…



Metropolis convergence

- Claim: The Metropolis algorithm converges to the target distribution p(*x*).
- Proof: It satisfies detailed balance
  - For all $x, y \in \boldsymbol{X}$, wlog assuming $\mathrm{p}(x) \leq \mathrm{p}(y)$

$$
\begin{aligned}
\mathrm{p}(x)T(x,y) &= \mathrm{p}(x)q(y \mid x) && \text{candidate is always accepted b/c p(x)} \leq \text{p(y)} \\
&= \mathrm{p}(x)q(x \mid y) && \text{q is symmetric} \\
&= \mathrm{p}(y)q(x \mid y)\frac{\mathrm{p}(x)}{\mathrm{p}(y)} \\
&= \mathrm{p}(y)T(y,x) && \text{transition prob b/c p(x)} \leq \text{p(y)}
\end{aligned}
$$

Metropolis-Hastings

- 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 importance 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 $x^{(t)}$

MH convergence

- Claim: The Metropolis-Hastings algorithm converges to the target distribution p(x).
- Proof: It satisfies detailed balance
  - For all $x,y \in \boldsymbol{X}$, wlog assume p(x)q(y|x) $\leq$ p(y)q(x|y)

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

candidate is always accepted b/c p(x)q(y|x) $\leq$ p(y)q(x|y)

transition prob b/c p(x)q(y|x) $\leq$ p(y)q(x|y)

Gibbs sampling

- A special case of Metropolis-Hastings which is applicable to state spaces in which we have a factored state space, and access to the full conditionals:

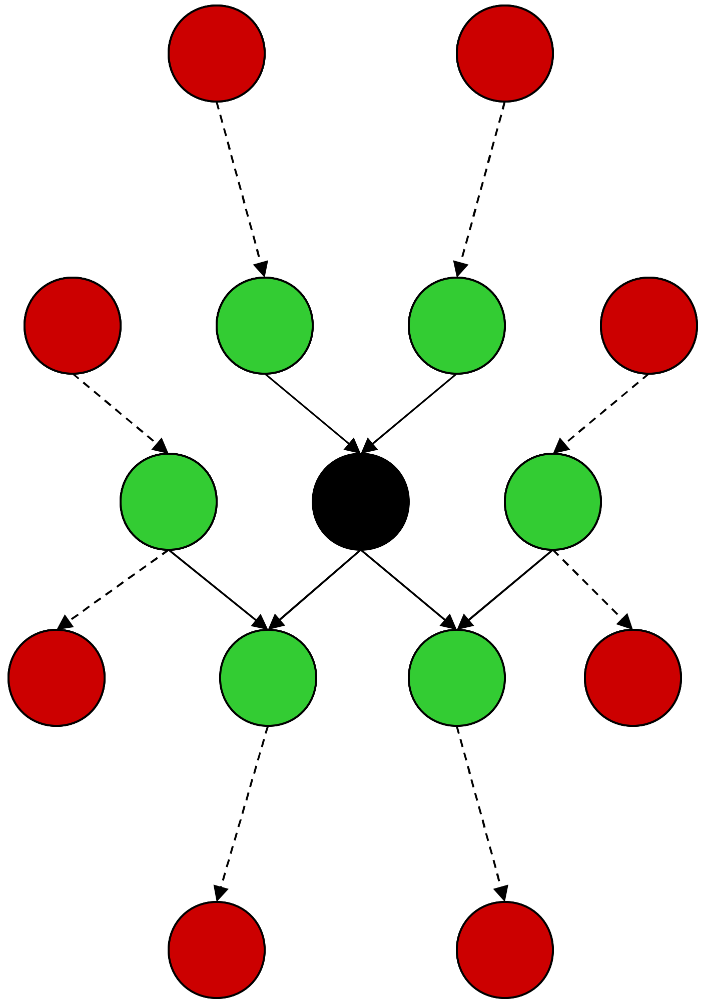
$$\mathrm{p}(x_j \mid x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n)$$

- Perfect for Bayesian networks!
- Idea: To transition from one state (variable assignment) to another,
  - Pick a variable,
  - Sample its value from the conditional distribution
  - That's it!
- We'll show in a minute why this is an instance of MH and thus must be sampling from the full joint

Markov blanket

- Recall that Bayesian networks encode a factored representation of the joint distribution
- Variables are independent of their non-descendents given their parents
- Variables are independent of *everything else in the network* given their *Markov blanket*!
- So, to sample each node, we only need to condition its Markov blanket

$$p(x_j \mid \mathrm{MB}(x_j))$$



Gibbs sampling

- More formally, the proposal distribution is

$$q(x^* \mid x^{(t)}) = \begin{cases} p(x^*_j \mid x^{(t)}_{-j}) & \text{if } x^*_{-j} = x^{(t)}_{-j} \\ 0 & \text{otherwise} \end{cases}$$

- The importance ratio is

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

$$= \frac{\mathrm{p}(x^*)\,\mathrm{p}(x^{(t)}_j \mid x^{(t)}_{-j})}{\mathrm{p}(x^{(t)})\,\mathrm{p}(x^*_j \mid x^*_{-j})}$$

Dfn of proposal distribution

$$= \frac{\mathrm{p}(x^*)\,\mathrm{p}(x^{(t)}_j, x^{(t)}_{-j})\,\mathrm{p}(x^*_{-j})}{\mathrm{p}(x^{(t)})\,\mathrm{p}(x^*_j, x^*_{-j})\,\mathrm{p}(x^{(t)}_{-j})}$$

Dfn of conditional probability

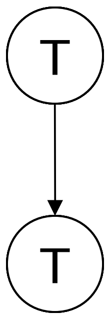
$$= \frac{\mathrm{p}(x^*_{-j})}{\mathrm{p}(x^{(t)}_{-j})} = 1$$

B/c we didn't change other vars

- So we always accept!

Gibbs sampling example

- Consider a simple, 2 variable Bayes net



	a	-a
	0.5	0.5

	b	-b
a	0.8	0.2
-a	0.2	0.8

	b	-b
a 1	1	
-a 1	1	

- Initialize randomly
- Sample variables alternately

Practical issues

- How many iterations?
- How to know when to stop?
- What’s a good proposal function?

Advanced Topics

- Simulated annealing, for global optimization, is a form of MCMC
- Mixtures of MCMC transition functions
- Monte Carlo EM (stochastic E-step)
- Reversible jump MCMC for model selection
- Adaptive proposal distributions

Cutest boy on the planet