Probabilistic Graphical Models

Markov Chain Monte Carlo Inference

Siamak Ravanbakhsh

Fall 2019

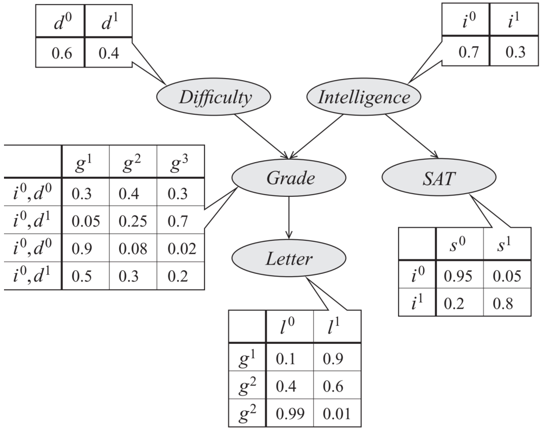
Learning objectives

- Markov chains
- the idea behind Markov Chain Monte Carlo (MCMC)
- two important examples:
    - Gibbs sampling
    - Metropolis-Hastings algorithm

Problem with **likelihood weighting**

Recap

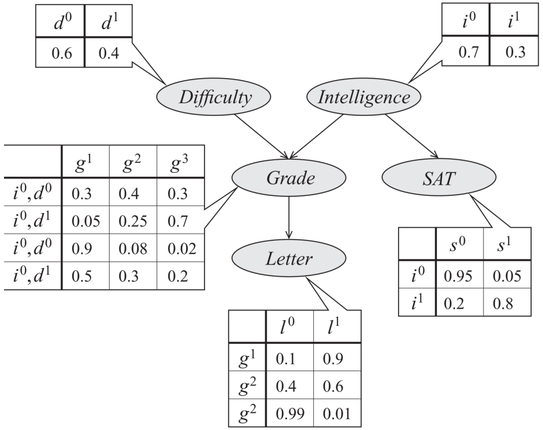
- use a topological ordering
- sample conditioned on the parents
- if observed:
  - keep the observed value
  - update the weight



Problem with likelihood weighting

Recap

- use a topological ordering
- sample conditioned on the parents
- if observed:
  - keep the observed value
  - update the weight



Issues

- observing the child does not affect the parent's assignment
- only applies to Bayes-nets

Gibbs sampling

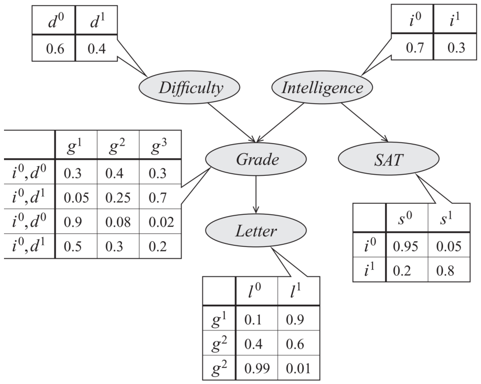
Idea

- iteratively sample each var. condition on its Markov blanket

$$X_i \sim p(x_i \mid X_{MB(i)})$$

- if X_i is observed: keep the observed value

- after many Gibbs sampling iterations $X \sim P$

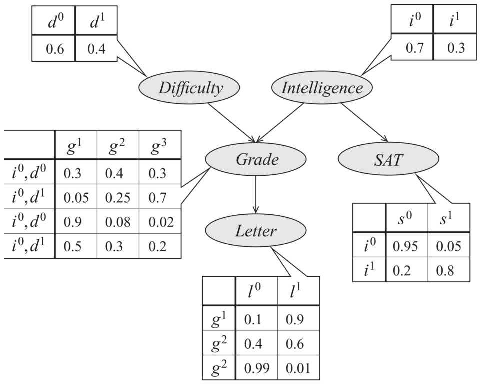


Gibbs sampling

Idea

- iterativelly sample each var. condition on its Markov blanket

$$X_i \sim p(x_i \mid X_{MB(i)})$$



- if X_i is observed: keep the observed value

equivalent to

  - first simplifying the model by removing observed vars
  - sampling from the simplified Gibbs dist.

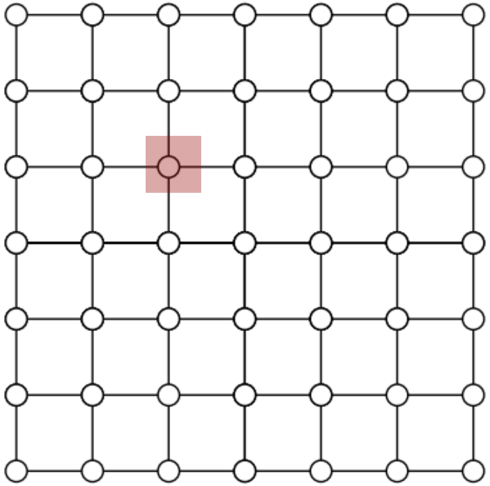
- after many Gibbs sampling iterations $X \sim P$

Example: Ising model

recall the Ising model:

$$p(x) \propto \exp(\sum_i x_i h_i + \sum_{i,j \in \mathcal{E}} x_i x_j J_{i,j})$$

$$x_i \in \{-1, +1\}$$



Example: Ising model

recall the Ising model:

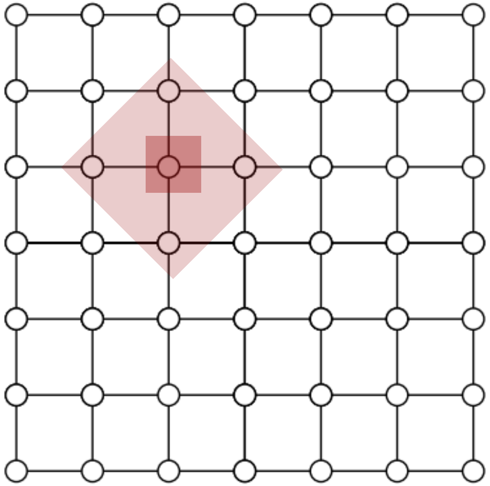
$$p(x) \propto \exp(\sum_i x_i h_i + \sum_{i,j \in \mathcal{E}} x_i x_j J_{i,j})$$

$$x_i \in \{-1, +1\}$$

sample each node i:

$$p(x_i = +1 \mid X_{MB(i)}) =$$

$$\frac{\exp(h_i + \sum_{j \in Mb(i)} J_{i,j} X_j)}{\exp(h_i + \sum_{j \in Mb(i)} J_{i,j} X_j) + \exp(-h_i - \sum_{j \in Mb(i)} J_{i,j} X_j)} =$$



Example: Ising model

recall the Ising model:

$$p(x) \propto \exp(\sum_i x_i h_i + \sum_{i,j \in \mathcal{E}} x_i x_j J_{i,j})$$

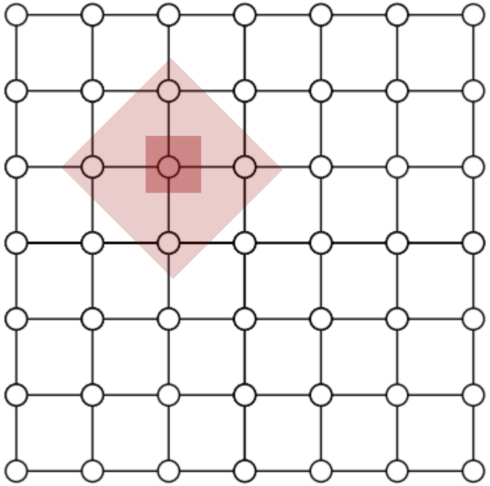
$$x_i \in \{-1, +1\}$$

sample each node i:

$$p(x_i = +1 \mid X_{MB(i)}) =$$

$$\frac{\exp(h_i + \sum_{j \in Mb(i)} J_{i,j} X_j)}{\exp(h_i + \sum_{j \in Mb(i)} J_{i,j} X_j) + \exp(-h_i - \sum_{j \in Mb(i)} J_{i,j} X_j)} =$$

$$\sigma(2h_i + 2\sum_{j \in Mb(i)} J_{i,j} X_j)$$

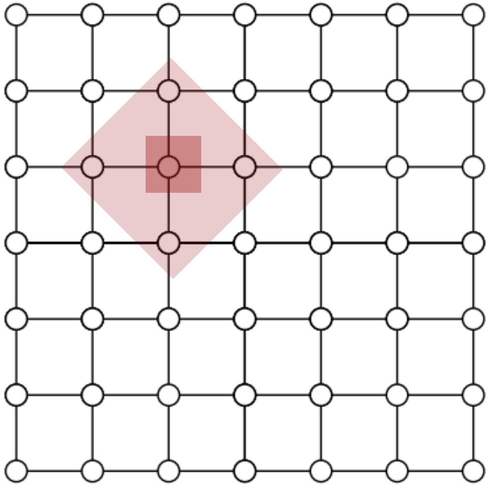


Example: Ising model

recall the Ising model:

$$p(x) \propto \exp(\sum_i x_i h_i + \sum_{i,j \in \mathcal{E}} x_i x_j J_{i,j})$$

$$x_i \in \{-1, +1\}$$



sample each node i:

$$p(x_i = +1 \mid X_{MB(i)}) =$$

$$\frac{\exp(h_i + \sum_{j \in Mb(i)} J_{i,j} X_j)}{\exp(h_i + \sum_{j \in Mb(i)} J_{i,j} X_j) + \exp(-h_i - \sum_{j \in Mb(i)} J_{i,j} X_j)} =$$

$$\sigma(2h_i + 2\sum_{j \in Mb(i)} J_{i,j} X_j)$$

compare with mean-field

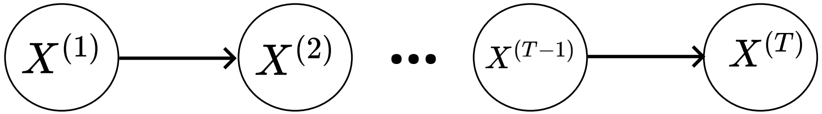
$$\sigma(2h_i + 2\sum_{j \in Mb(i)} J_{i,j} \mu_j)$$

Markov Chain

a sequence of random variables with Markov property

$$P(X^{(t)}|X^{(1)},\ldots,X^{(t-1)}) = P(X^{(t)}|X^{(t-1)})$$

its graphical model



many applications:

- language modeling: X is a word or a character
- physics: with correct choice of X, the world is Markov

Transition model

we assume a homogeneous chain: $P(X^{(t)}|X^{(t-1)}) = P(X^{(t+1)}|X^{(t)}) \quad \forall t$

cond. probabilities remain the same across time-steps

notation: conditional probability $P(X^{(t)} = x|X^{(t-1)} = x') = T(x', x)$

is called the **transition model**

think of this as a matrix T

Transition model

we assume a homogeneous chain: $P(X^{(t)}|X^{(t-1)}) = P(X^{(t+1)}|X^{(t)}) \quad \forall t$

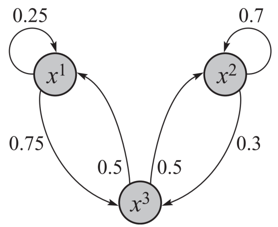
cond. probabilities remain the same across time-steps

notation: conditional probability $P(X^{(t)} = x|X^{(t-1)} = x') = T(x', x)$

is called the **transition model**

think of this as a matrix T

state-transition diagram



its transition matrix

$$T = \begin{bmatrix} .25 & 0 & .75 \\ 0 & .7 & .3 \\ .5 & .5 & 0 \end{bmatrix}$$

Transition model

we assume a homogeneous chain: $P(X^{(t)}|X^{(t-1)}) = P(X^{(t+1)}|X^{(t)}) \quad \forall t$

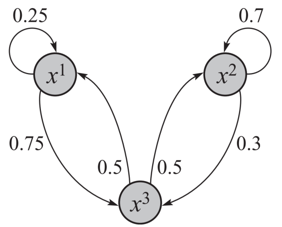
cond. probabilities remain the same across time-steps

notation: conditional probability $P(X^{(t)} = x|X^{(t-1)} = x') = T(x', x)$

is called the **transition model**

think of this as a matrix T

state-transition diagram



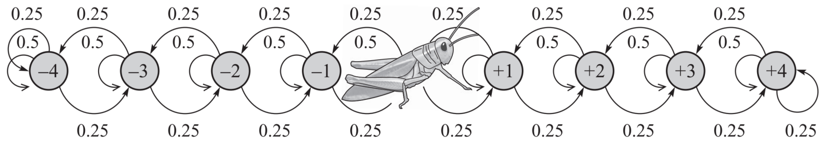
its transition matrix

$$T = \begin{bmatrix} .25 & 0 & .75 \\ 0 & .7 & .3 \\ .5 & .5 & 0 \end{bmatrix}$$

evolving the distribution $P(X^{(t+1)} = x) = \sum_{x' \in Val(X)} P(X^{(t)} = x')T(x', x)$

Markov Chain Monte Carlo (MCMC)

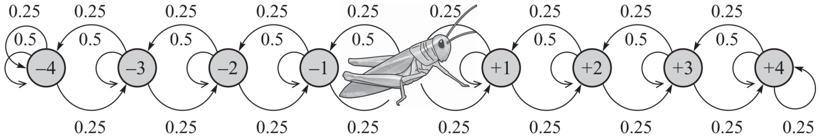
Example state-transition diagram for grasshopper random walk



initial distribution $P^{(0)}(X = 0) = 1$

Markov Chain Monte Carlo (MCMC)

Example state-transition diagram for grasshopper random walk

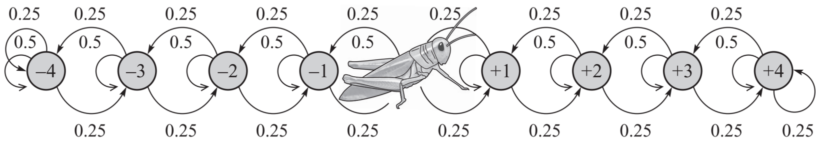


initial distribution $P^{(0)}(X = 0) = 1$

after t=50 steps, the distribution is almost uniform $P^t(x) \approx \frac{1}{9} \quad \forall x$

Markov Chain Monte Carlo (MCMC)

Example state-transition diagram for grasshopper random walk



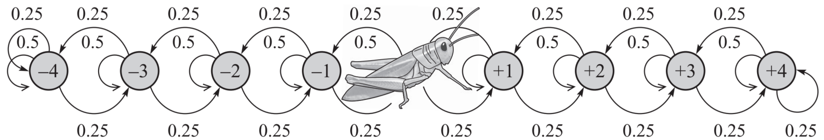
initial distribution $P^{(0)}(X = 0) = 1$

after t=50 steps, the distribution is almost uniform $P^t(x) \approx \frac{1}{9} \quad \forall x$

use the chain to sample from the uniform distribution $P^t(X) \approx \frac{1}{9}$

Markov Chain Monte Carlo (MCMC)

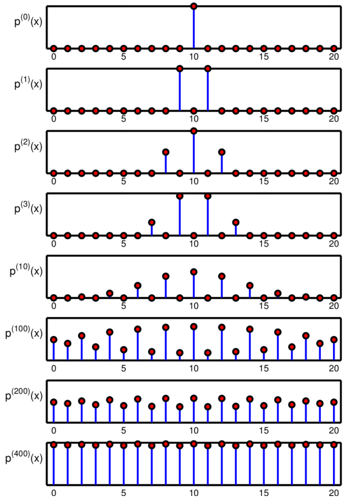
Example state-transition diagram for grasshopper random walk



initial distribution $P^{(0)}(X = 0) = 1$

after t=50 steps, the distribution is almost uniform $P^t(x) \approx \frac{1}{9} \quad \forall x$

use the chain to sample from the uniform distribution $P^t(X) \approx \frac{1}{9}$

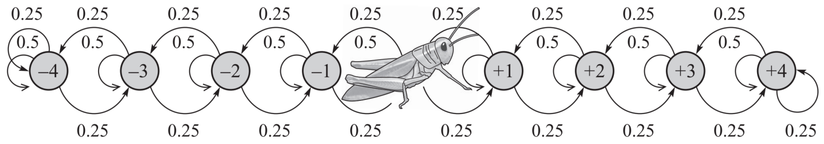


why is it uniform?

(mixing image: Murphy's book)

Markov Chain Monte Carlo (MCMC)

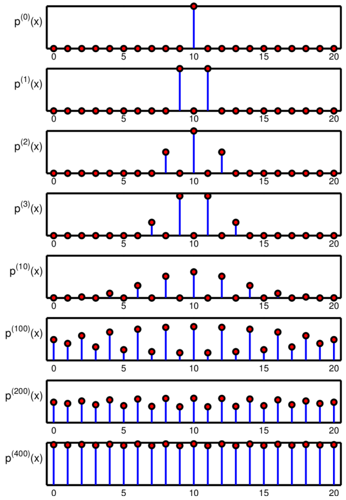
Example state-transition diagram for grasshopper random walk



initial distribution $P^{(0)}(X=0)=1$

after t=50 steps, the distribution is almost uniform $P^t(x) \approx \frac{1}{9} \quad \forall x$

use the chain to sample from the uniform distribution $P^t(X) \approx \frac{1}{9}$



why is it uniform?

MCMC generalize this idea beyond uniform dist.

- we want to sample from P^*
- pick the transition model such that $P^{\infty}(X) = P^*(X)$

(mixing image: Murphy's book)

Stationary distribution

given a transition model $T(x, x')$ if the chain converges:

global balance equation $P^{(t)}(x) \approx P^{(t+1)}(x) = \sum_{x'} P^{(t)}(x')T(x', x)$

Stationary distribution

given a transition model $T(x, x')$ if the chain converges:

global balance equation $P^{(t)}(x) \approx P^{(t+1)}(x) = \sum_{x'} P^{(t)}(x')T(x', x)$

this condition defines the stationary distribution: π

$$\pi(X = x) = \sum_{x' \in Val(X)} \pi(X = x')T(x', x)$$

Stationary distribution

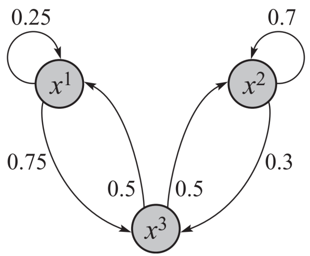
given a transition model $T(x, x')$ if the chain converges:

global balance equation $P^{(t)}(x) \approx P^{(t+1)}(x) = \sum_{x'} P^{(t)}(x')T(x', x)$

this condition defines the stationary distribution: π

$$\pi(X = x) = \sum_{x' \in Val(X)} \pi(X = x')T(x', x)$$

Example finding the stationary dist.

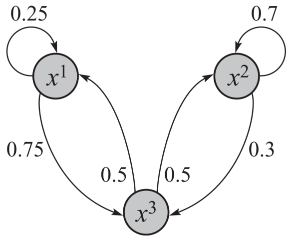


$$\pi(x^1) = .25\pi(x^1) + .5\pi(x^3)$$
$$\pi(x^2) = .7\pi(x^2) + .5\pi(x^3)$$
$$\pi(x^3) = .75\pi(x^1) + .3\pi(x^2)$$
$$\pi(x^1) + \pi(x^2) + \pi(x^3) = 1$$

$$\pi(x^1) = .2$$
$$\pi(x^2) = .5$$
$$\pi(x^3) = .3$$

Stationary distribution as an eigenvector

Example finding the stationary dist.

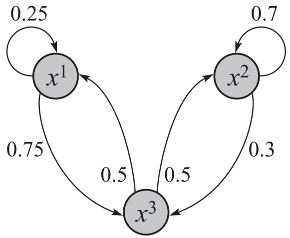


$$\pi(x^1) = .25\pi(x^1) + .5\pi(x^3)$$
$$\pi(x^2) = .7\pi(x^2) + .5\pi(x^3)$$
$$\pi(x^3) = .75\pi(x^1) + .3\pi(x^2)$$
$$\pi(x^1) + \pi(x^2) + \pi(x^3) = 1$$

$$\pi(x^1) = .2$$
$$\pi(x^2) = .5$$
$$\pi(x^3) = .3$$

Stationary distribution as an eigenvector

Example finding the stationary dist.



$$\pi(x^1) = .25\pi(x^1) + .5\pi(x^3)$$
$$\pi(x^2) = .7\pi(x^2) + .5\pi(x^3)$$
$$\pi(x^3) = .75\pi(x^1) + .3\pi(x^2)$$
$$\pi(x^1) + \pi(x^2) + \pi(x^3) = 1$$

$$\pi(x^1) = .2$$
$$\pi(x^2) = .5$$
$$\pi(x^3) = .3$$

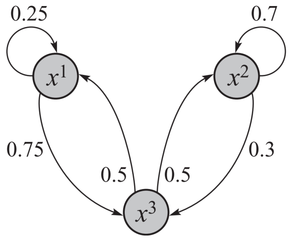
viewing $T(.,.)$ as a matrix and $P^t(x)$ as a vector

- evolution of dist $P^t(x)$: $P^{(t+1)} = T^{\mathsf{T}} P^{(t)}$
- multiple steps: $P^{(t+m)} = (T^{\mathsf{T}})^m P^{(t)}$

$$\underbrace{\begin{bmatrix} .25 & 0 & .5 \\ 0 & .7 & .5 \\ .75 & .3 & 0 \end{bmatrix}}_{T^{\mathsf{T}}} \underbrace{\begin{bmatrix} .2 \\ .5 \\ .3 \end{bmatrix}}_{\pi} = \underbrace{\begin{bmatrix} .2 \\ .5 \\ .3 \end{bmatrix}}_{\pi}$$

Stationary distribution as an eigenvector

Example finding the stationary dist.



$$\pi(x^1) = .25\pi(x^1) + .5\pi(x^3)$$
$$\pi(x^2) = .7\pi(x^2) + .5\pi(x^3)$$
$$\pi(x^3) = .75\pi(x^1) + .3\pi(x^2)$$
$$\pi(x^1) + \pi(x^2) + \pi(x^3) = 1$$

$$\pi(x^1) = .2$$
$$\pi(x^2) = .5$$
$$\pi(x^3) = .3$$

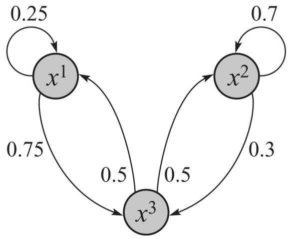
viewing $T(.,.)$ as a matrix and $P^t(x)$ as a vector

- evolution of dist $P^t(x)$: $P^{(t+1)} = T^{\mathsf{T}} P^{(t)}$
- multiple steps: $P^{(t+m)} = (T^{\mathsf{T}})^m P^{(t)}$
- for stationary dist: $\pi = T^{\mathsf{T}} \pi$

$$\underbrace{\begin{bmatrix} .25 & 0 & .5 \\ 0 & .7 & .5 \\ .75 & .3 & 0 \end{bmatrix}}_{T^{\mathsf{T}}} \underbrace{\begin{bmatrix} .2 \\ .5 \\ .3 \end{bmatrix}}_{\pi} = \underbrace{\begin{bmatrix} .2 \\ .5 \\ .3 \end{bmatrix}}_{\pi}$$

Stationary distribution as an eigenvector

Example finding the stationary dist.



$$\pi(x^1) = .25\pi(x^1) + .5\pi(x^3)$$
$$\pi(x^2) = .7\pi(x^2) + .5\pi(x^3)$$
$$\pi(x^3) = .75\pi(x^1) + .3\pi(x^2)$$
$$\pi(x^1) + \pi(x^2) + \pi(x^3) = 1$$

$$\pi(x^1) = .2$$
$$\pi(x^2) = .5$$
$$\pi(x^3) = .3$$

viewing $T(.,.)$ as a matrix and $P^t(x)$ as a vector

- evolution of dist $P^t(x)$: $P^{(t+1)} = T^\top P^{(t)}$
- multiple steps: $P^{(t+m)} = (T^\top)^m P^{(t)}$
- for stationary dist: $\pi = T^\top \pi$
- π is an eigenvector of $T^\top$ with eigenvalue 1

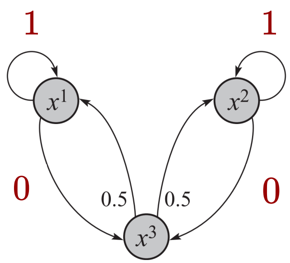
$$\underbrace{\begin{bmatrix} .25 & 0 & .5 \\ 0 & .7 & .5 \\ .75 & .3 & 0 \end{bmatrix}}_{T^\top} \underbrace{\begin{bmatrix} .2 \\ .5 \\ .3 \end{bmatrix}}_{\pi} = \underbrace{\begin{bmatrix} .2 \\ .5 \\ .3 \end{bmatrix}}_{\pi}$$

(produce it by running the chain = power iteration)

Stationary distribution: existance & uniquness

irreducible

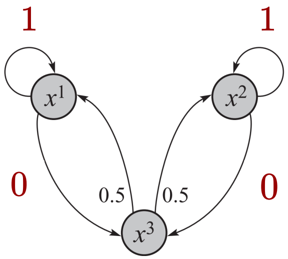
- we should be able to reach any x' from any x
- otherwise, π is not unique



Stationary distribution: existance & uniquness

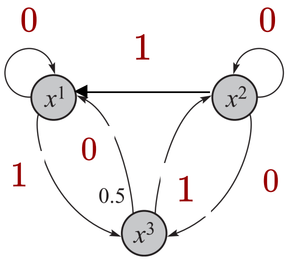
irreducible

- we should be able to reach any x' from any x
- otherwise, π is not unique



aperiodic

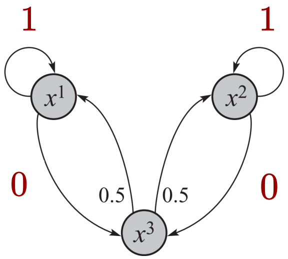
- the chain should not have a fixed cyclic behavior
- otherwise, the chain does not converge (it oscillates)



Stationary distribution: existance & uniquness

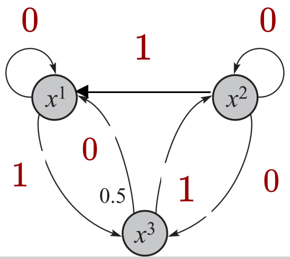
irreducible

- we should be able to reach any x' from any x
- otherwise, π is not unique



aperiodic

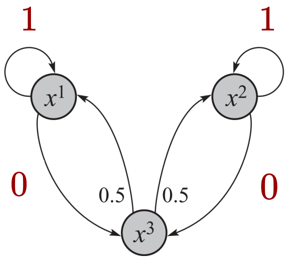
- the chain should not have a fixed cyclic behavior
- otherwise, the chain does not converge (it oscillates)



every aperiodic and irreducible chain (with a finite domain) has a unique limiting distribution π

such that $\pi(X = x) = \sum_{x' \in Val(X)} \pi(X = x')T(x', x)$

Stationary distribution: existance & uniquness

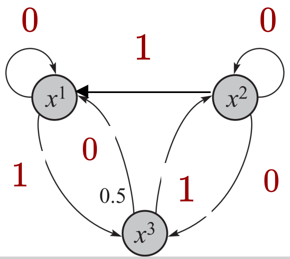


irreducible

- we should be able to reach any x' from any x
- otherwise, π is not unique

aperiodic

- the chain should not have a fixed cyclic behavior
- otherwise, the chain does not converge (it oscillates)



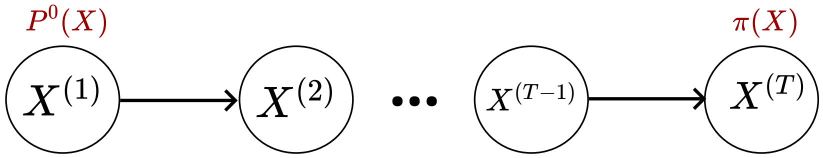
every aperiodic and irreducible chain (with a finite domain) has a unique limiting distribution π

such that $\pi(X = x) = \sum_{x' \in Val(X)} \pi(X = x')T(x', x)$

regular chain a sufficient condition: there exists a K, such that the probability of reaching any destination from any source in K steps is positive (applies to discrete & continuous domains)

MCMC in graphical models

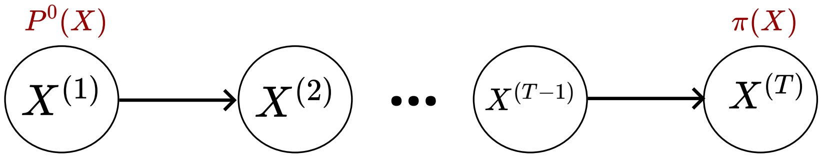
distinguishing the *"graphical models"* involved



1: the Markov chain

MCMC in graphical models

distinguishing the *"graphical models"* involved



1: the Markov chain

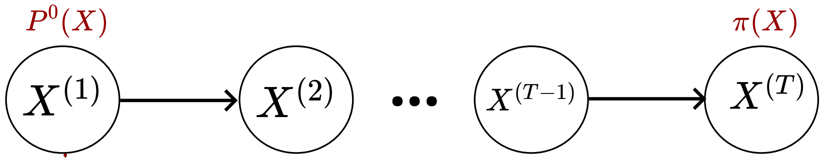
2: state-transition diagram (not shown)

that has exponentially many nodes

$$\#nodes = |Val(X)|$$

MCMC in graphical models

distinguishing the *"graphical models"* involved



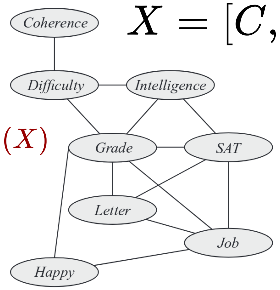
1: the Markov chain

2: state-transition diagram (not shown) that has exponentially many nodes

$$\#nodes = |Val(X)|$$

$$X = [C, D, I, G, S, L, J, H]$$

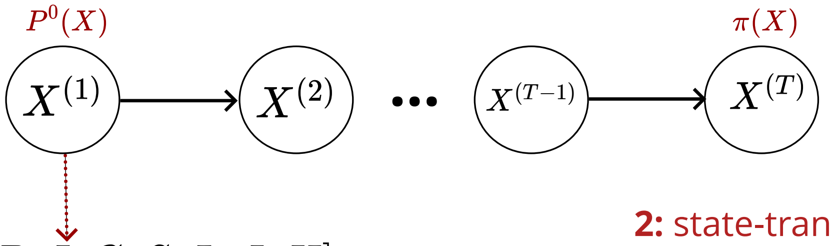
$P^*(X)$



3: the graphical model, from which we want to sample $P^*(X)$

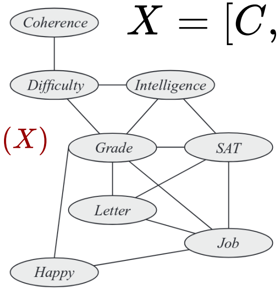
MCMC in graphical models

distinguishing the *"graphical models"* involved



1: the Markov chain

$$X = [C, D, I, G, S, L, J, H]$$



$P^*(X)$

2: state-transition diagram (not shown)

that has exponentially many nodes

$$\#nodes = |Val(X)|$$

3: the graphical model, from which we want to sample $P^*(X)$

objective: design the Markov chain transition so that $\pi(X) = P^*(X)$

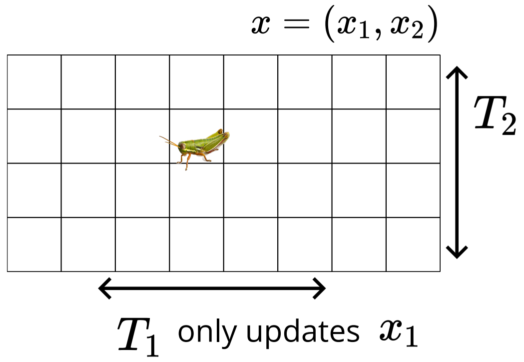
Multiple transition models

idea

aka, **kernels**

have multiple transition models $T_1(x, x'), T_2(x, x'), \ldots, T_n(x, x')$

each making local changes to x



using a single kernel we may not be able to visit all the states while their combination is "ergodic"

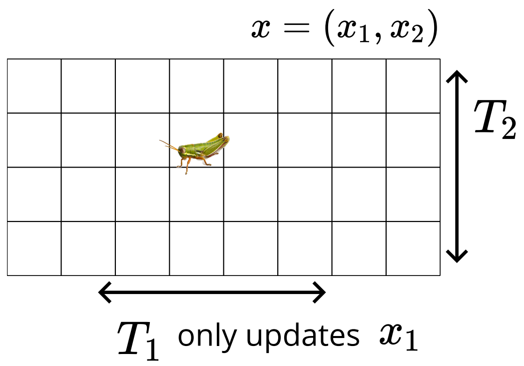
Multiple transition models

idea

aka, **kernels**

have multiple transition models $T_1(x, x'), T_2(x, x'), \ldots, T_n(x, x')$

each making local changes to x

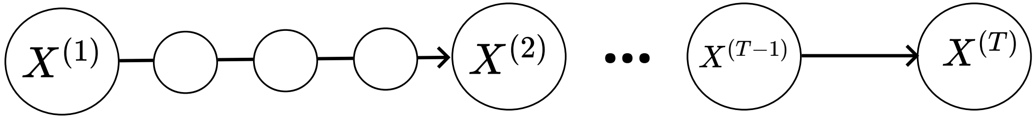


using a single kernel we may not be able to visit all the states while their combination is "ergodic"

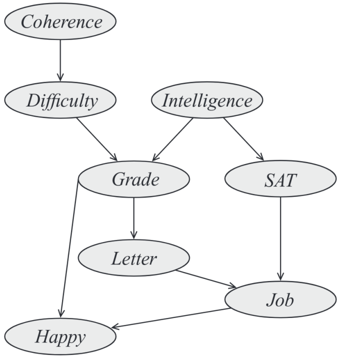
if $\pi(X = x) = \sum_{x' \in Val(X)} \pi(X = x') T_k(x', x) \quad \forall k$

then we can combine the kernels:

- mixing them $T(x', x) = \sum_k p(k) T_k(x', x)$
- cycling them $T(x', x) = \int_{x^{[1]}, x^{[2]}, \ldots, x^{[n]}} T_1(x', x^{[1]}) T_2(x^{[1]}, x^{[2]}), \ldots T_n(x^{[n-1]}, x) \mathrm{d}x^{[1]} \mathrm{d}x^{[2]} \ldots \mathrm{d}x^{[n]}$

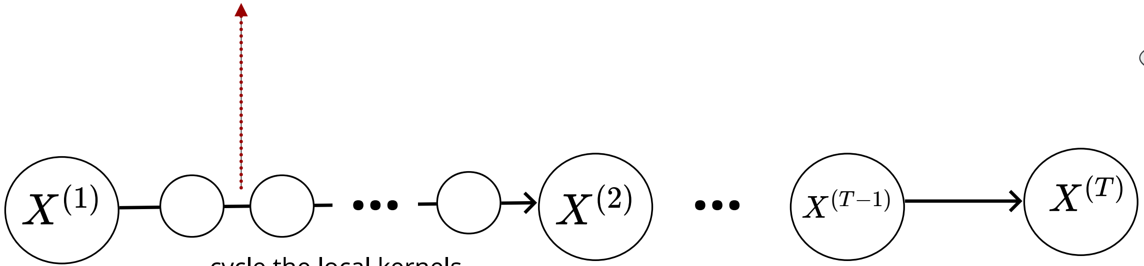


Revisiting Gibbs sampling



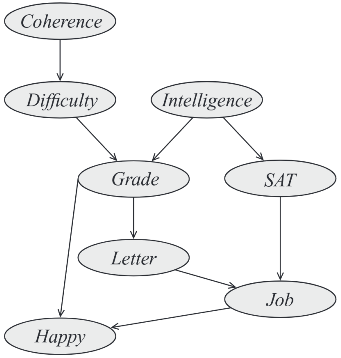
one kernel for each variable
perform local, conditional updates

$$T_i(x^{(t)}, x^{(t+1)}) = P^*(x_i^{(t+1)} | x_{-i}^{(t)}) \mathbb{I}(x_{-i}^{(t+1)} = x_{-i}^{(t)})$$



cycle the local kernels

Revisiting Gibbs sampling

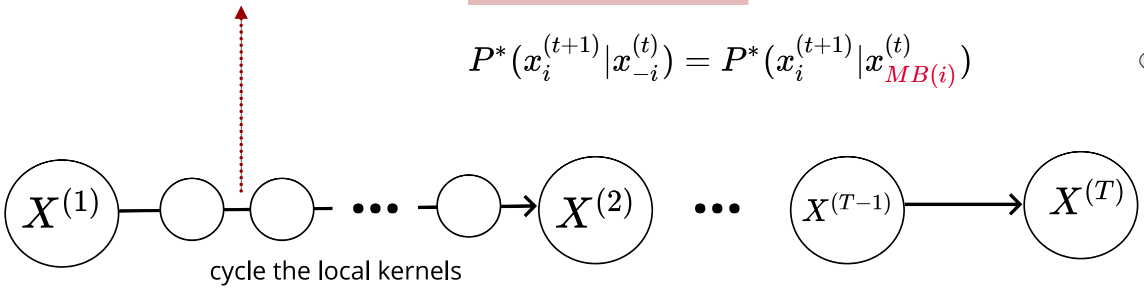


one kernel for each variable

perform local, conditional updates

$$T_i(x^{(t)}, x^{(t+1)}) = P^*(x_i^{(t+1)} | x_{-i}^{(t)}) \mathbb{I}(x_{-i}^{(t+1)} = x_{-i}^{(t)})$$

$$P^*(x_i^{(t+1)} | x_{-i}^{(t)}) = P^*(x_i^{(t+1)} | x_{MB(i)}^{(t)})$$



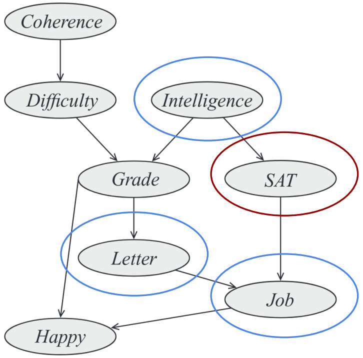
Revisiting Gibbs sampling

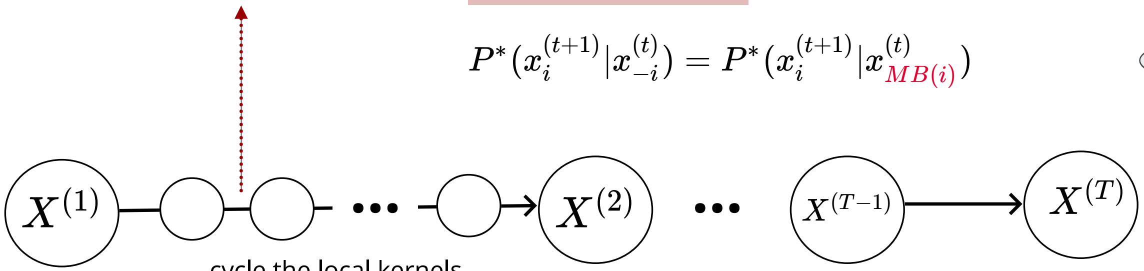
one kernel for each variable

perform local, conditional updates

$$T_i(x^{(t)}, x^{(t+1)}) = P^*(x_i^{(t+1)} | x_{-i}^{(t)}) \mathbb{I}(x_{-i}^{(t+1)} = x_{-i}^{(t)})$$

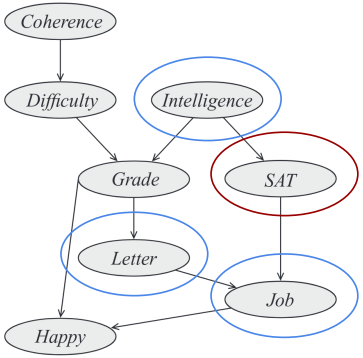
$$P^*(x_i^{(t+1)} | x_{-i}^{(t)}) = P^*(x_i^{(t+1)} | x_{MB(i)}^{(t)})$$





cycle the local kernels

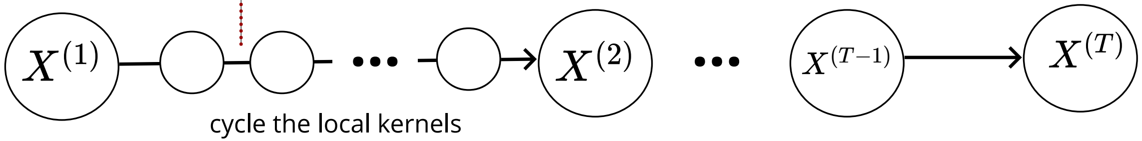
Revisiting Gibbs sampling



one kernel for each variable
perform local, conditional updates

$$T_i(x^{(t)}, x^{(t+1)}) = P^*(x_i^{(t+1)} | x_{-i}^{(t)}) \mathbb{I}(x_{-i}^{(t+1)} = x_{-i}^{(t)})$$

$$P^*(x_i^{(t+1)} | x_{-i}^{(t)}) = P^*(x_i^{(t+1)} | x_{MB(i)}^{(t)})$$



$\pi(X) = P^*(X)$ is the stationary dist. for this Markov chain

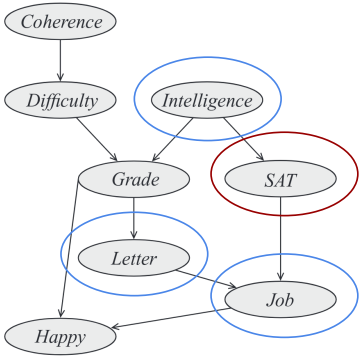
Revisiting Gibbs sampling

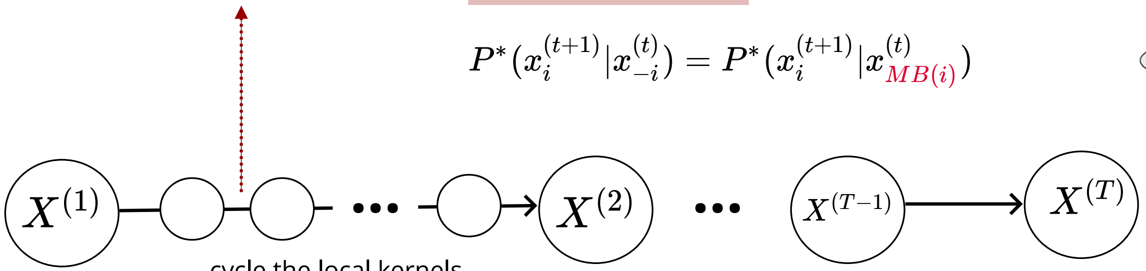
one kernel for each variable

perform local, conditional updates

$$T_i(x^{(t)}, x^{(t+1)}) = P^*(x_i^{(t+1)} | x_{-i}^{(t)})\mathbb{I}(x_{-i}^{(t+1)} = x_{-i}^{(t)})$$

$$P^*(x_i^{(t+1)} | x_{-i}^{(t)}) = P^*(x_i^{(t+1)} | x_{MB(i)}^{(t)})$$





cycle the local kernels

$\pi(X) = P^*(X)$ is the stationary dist. for this Markov chain

if $P^*(x) > 0 \quad \forall x$ then this chain is regular

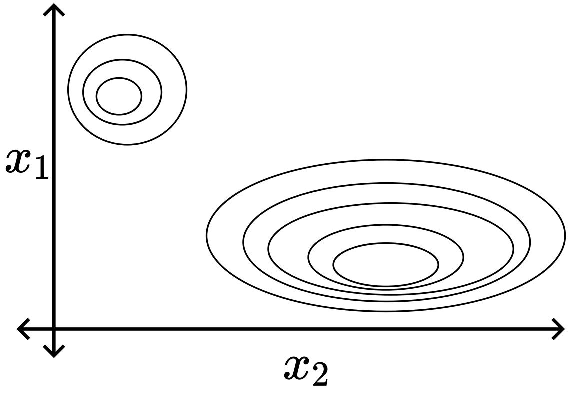
i.e., converges to its unique stationary dist.

Some variations

block Gibbs sampling

local moves can get stuck in modes of $P^*(X)$

updates using $P(x_1 \mid x_2), P(x_2|x_1)$ will have problem exploring these modes



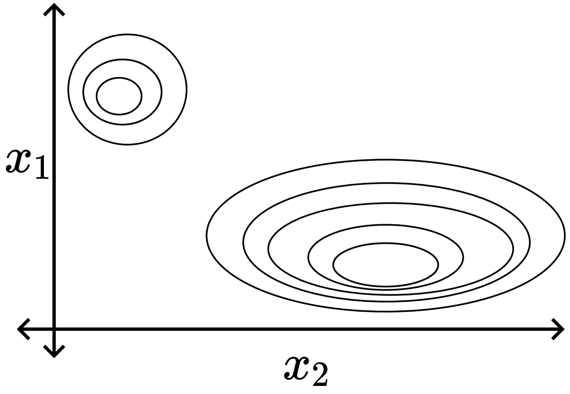
Some variations

block Gibbs sampling

local moves can get stuck in modes of $P^*(X)$

updates using $P(x_1 \mid x_2), P(x_2|x_1)$ will have problem exploring these modes

idea: each kernel updates a block of variables

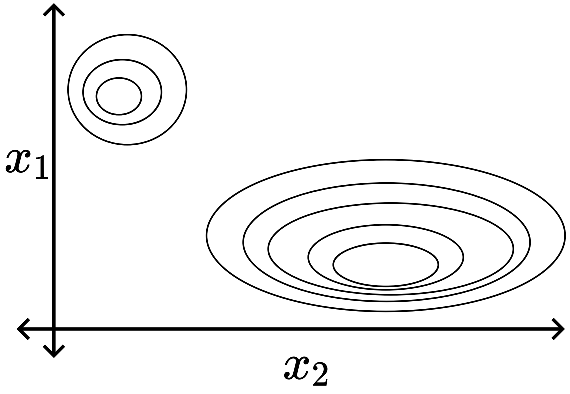


Some variations

block Gibbs sampling

local moves can get stuck in modes of $P^*(X)$

updates using $P(x_1 \mid x_2), P(x_2|x_1)$ will have problem exploring these modes



idea: each kernel updates a block of variables

collapsed Gibbs sampling

marginalize out some variables

ordinary case: $p(X \mid Y, Z), P(Y \mid X, Z), P(Z \mid X, Y)$

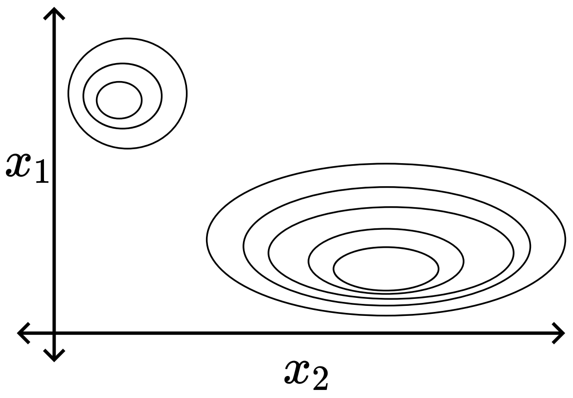
Some variations

block Gibbs sampling

local moves can get stuck in modes of $P^*(X)$

updates using $P(x_1 \mid x_2), P(x_2|x_1)$ will have problem exploring these modes

idea: each kernel updates a block of variables



collapsed Gibbs sampling

marginalize out some variables

ordinary case: $p(X \mid Y, Z), P(Y \mid X, Z), P(Z \mid X, Y)$

marginalize over Y: $P(X \mid Z), P(Z \mid X, Y)$ or $P(X \mid Z), P(Z \mid X)$

involves analytical derivation of collapsed updates

Detailed balance

A Markov chain is reversible if for a unique π

detailed balance $$\pi(x)T(x,x') = \pi(x')T(x',x) \quad \forall x, x'$$

same frequency in both directions

Detailed balance

A Markov chain is reversible if for a unique π

detailed balance $\pi(x)T(x,x') = \pi(x')T(x',x) \quad \forall x, x'$

same frequency in both directions

$$\int_{x'} \pi(x)T(x,x')\mathrm{d}x' = \pi(x)\int_{x'} T(x,x')\mathrm{d}x' = \pi(x) = \int_{x'} \pi(x')T(x',x)\mathrm{d}x'$$

left-hand side

global balance

right-hand side

Detailed balance

A Markov chain is reversible if for a unique π

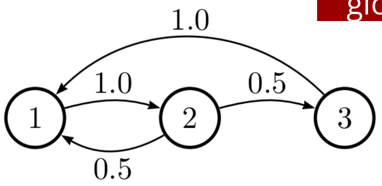
detailed balance $\pi(x)T(x,x') = \pi(x')T(x',x) \quad \forall x, x'$

same frequency in both directions

$$\int_{x'} \pi(x)T(x,x')\mathrm{d}x' = \pi(x)\int_{x'} T(x,x')\mathrm{d}x' = \pi(x) = \int_{x'} \pi(x')T(x',x)\mathrm{d}x'$$

left-hand side **global balance** *right-hand side*

detailed balance is a stronger condition



$\pi = [.4, .4, .2]$

global balance ✔

detailed balance ✖

(example: Murphy's book)

Detailed balance

A Markov chain is reversible if for a unique π

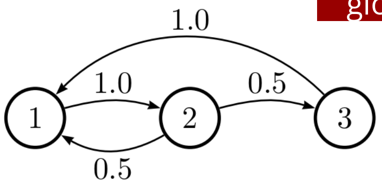
detailed balance $\pi(x)T(x,x') = \pi(x')T(x',x) \quad \forall x, x'$

same frequency in both directions

$$\int_{x'} \pi(x)T(x,x')\mathrm{d}x' = \pi(x)\int_{x'} T(x,x')\mathrm{d}x' = \pi(x) = \int_{x'} \pi(x')T(x',x)\mathrm{d}x'$$

left-hand side **global balance** *right-hand side*

detailed balance is a stronger condition



$\pi = [.4, .4, .2]$

global balance ✔

detailed balance ✖

if Markov chain is regular and π satisfies detailed balance, then π is the unique stationary distribution

(example: Murphy's book)

Detailed balance

A Markov chain is reversible if for a unique π

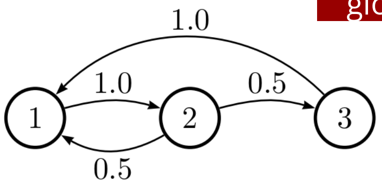
detailed balance $\pi(x)T(x,x') = \pi(x')T(x',x) \quad \forall x, x'$

same frequency in both directions

$$\int_{x'} \pi(x)T(x,x')\mathrm{d}x' = \pi(x)\int_{x'} T(x,x')\mathrm{d}x' = \pi(x) = \int_{x'} \pi(x')T(x',x)\mathrm{d}x'$$

left-hand side global balance *right-hand side*

detailed balance is a stronger condition



$\pi = [.4, .4, .2]$

global balance ✔

detailed balance ✖

if Markov chain is regular and π satisfies detailed balance, then π is the unique stationary distribution

- *analogous to the theorem for global balance*
- *checking for detailed balance is sometimes easier*

(example: Murphy's book)

Detailed balance

A Markov chain is reversible if for a unique π

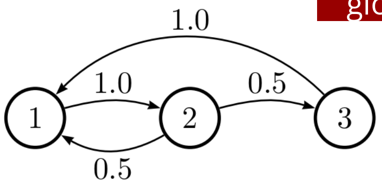
detailed balance $\pi(x)T(x,x') = \pi(x')T(x',x) \quad \forall x, x'$

same frequency in both directions

$$\int_{x'} \pi(x)T(x,x')\mathrm{d}x' = \pi(x)\int_{x'} T(x,x')\mathrm{d}x' = \pi(x) \quad = \quad \int_{x'} \pi(x')T(x',x)\mathrm{d}x'$$

left-hand side global balance *right-hand side*

detailed balance is a stronger condition



$\pi = [.4, .4, .2]$

global balance ✔

detailed balance ✖

if Markov chain is regular and π satisfies detailed balance, then π is the unique stationary distribution

- *analogous to the theorem for global balance*
- *checking for detailed balance is sometimes easier*

what happens if T is symmetric?

(example: Murphy's book)

Using a proposal for the chain

Given P^* design a chain to sample from P^*

idea

Using a proposal for the chain

Given P^* design a chain to sample from P^*

idea

- use a proposal transition $T^q(x, x')$
- we can sample from $T^q(x, \cdot)$
- $T^q(x, x')$ is a regular chain (reaching every state in K steps has a non-zero probability)

Using a proposal for the chain

Given P^* design a chain to sample from P^*

idea

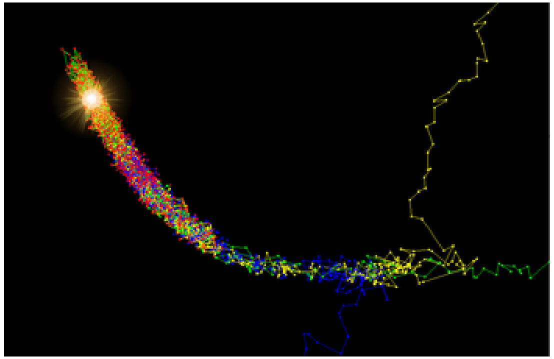
- use a proposal transition $T^q(x, x')$
- we can sample from $T^q(x, \cdot)$
- $T^q(x, x')$ is a regular chain (reaching every state in K steps has a non-zero probability)
- accept the proposed move with probability $A(x, x')$
  - to achieve detailed balance for a desirable P^*

Metropolis algorithm

- use a proposal transition $T^q(x, x')$
- we can sample from $T^q(x, \cdot)$
- $T^q(x, x')$ is a regular chain (reaching every state in K steps has a non-zero probability)
- accept the proposed move with probability $A(x, x')$
  - to achieve detailed balance
- proposal is symmetric $T(x, x') = T(x', x)$

$$A(x, x') \triangleq \min(1, \frac{p(x')}{p(x)})$$

accepts the move if it increases P^*
may accept it otherwise



(image: Wikipedia)

Metropolis-Hastings algorithm

if the proposal is NOT symmetric, then $A(x, x') \triangleq \min(1, \frac{p(x')T^q(x', x)}{p(x)T^q(x, x')})$

Metropolis-Hastings algorithm

if the proposal is NOT symmetric, then $A(x, x') \triangleq \min(1, \frac{p(x')T^q(x', x)}{p(x)T^q(x, x')})$

why does it sample from P^* ?

Metropolis-Hastings algorithm

if the proposal is NOT symmetric, then $A(x,x') \triangleq \min(1, \frac{p(x')T^q(x',x)}{p(x)T^q(x,x')})$

why does it sample from P^* ?

derive the transition kernel:

$T(x,x') = T^q(x,x')A(x,x') \quad \forall x \neq x'$ | move to a different state is accepted

Metropolis-Hastings algorithm

if the proposal is NOT symmetric, then $A(x,x') \triangleq \min(1, \frac{p(x')T^q(x',x)}{p(x)T^q(x,x')})$

why does it sample from P^* ?

derive the transition kernel:

$T(x,x') = T^q(x,x')A(x,x') \quad \forall x \neq x'$ | move to a different state is accepted

$T(x,x) = T^q(x,x) + \sum_{x \neq x'}(1 - A(x,x'))T(x,x')$ | proposal to stay is always accepted
move to a new state is rejected

Metropolis-Hastings algorithm

if the proposal is NOT symmetric, then $A(x,x') \triangleq \min(1, \frac{p(x')T^q(x',x)}{p(x)T^q(x,x')})$

why does it sample from P^* ?

derive the transition kernel:

$T(x,x') = T^q(x,x')A(x,x') \quad \forall x \neq x'$ | move to a different state is accepted

$T(x,x) = T^q(x,x) + \sum_{x \neq x'}(1 - A(x,x'))T(x,x')$ | proposal to stay is always accepted
move to a new state is rejected

substitute this into detailed balance (does it hold?)

$$\pi(x)T^q(x,x')A(x,x') \overset{?}{=} \pi(x')T^q(x',x)A(x',x)$$

$$\min(1, \frac{\pi(x')T^q(x',x)}{\pi(x)T^q(x,x')}) \qquad \min(1, \frac{\pi(x)T^q(x,x')}{\pi(x')T^q(x',x)})$$

this is for ☀ only

Metropolis-Hastings algorithm

if the proposal is NOT symmetric, then $A(x, x') \triangleq \min(1, \frac{p(x')T^q(x', x)}{p(x)T^q(x, x')})$

why does it sample from P^* ?

derive the transition kernel:

$T(x, x') = T^q(x, x')A(x, x') \quad \forall x \neq x'$ | move to a different state is accepted

$T(x, x) = T^q(x, x) + \sum_{x \neq x'}(1 - A(x, x'))T(x, x')$ | proposal to stay is always accepted
move to a new state is rejected

substitute this into detailed balance (does it hold?)

$$\pi(x)T^q(x, x')A(x, x') \stackrel{?}{=} \pi(x')T^q(x', x)A(x', x)$$

$A(x, x') = \min(1, \frac{\pi(x')T^q(x', x)}{\pi(x)T^q(x, x')})$ $\quad$ $A(x', x) = \min(1, \frac{\pi(x)T^q(x, x')}{\pi(x')T^q(x', x)})$

this is for (sun) only

Gibbs sampling is a special case, with $A(x, x') = 1$ all the time!

Sampling from the chain

at the limit $T \to \infty$, $P^{\infty} = \pi = P^*$

how long should we wait for $D(P^T, \pi) < \epsilon$?

mixing time

$O(\frac{1}{1-\lambda_2} \log(\frac{N}{\epsilon}))$

λ_2: 2nd largest eigenvalue of T

N: #states (exponential)

Sampling from the chain

at the limit $T \rightarrow \infty$, $P^{\infty} = \pi = P^{*}$

how long should we wait for $D(P^{T}, \pi) < \epsilon$?

- run the chain for a burn-in period (T steps)
- collect samples (few more steps)
- multiple restarts can ensure a better coverage

mixing time

$$O\left(\frac{1}{1-\lambda_2} \log\left(\frac{N}{\epsilon}\right)\right)$$

#states (exponential)

2nd largest eigenvalue of T

Sampling from the chain

at the limit $T \to \infty$, $P^{\infty} = \pi = P^*$

how long should we wait for $D(P^T, \pi) < \epsilon$?

- run the chain for a burn-in period (T steps)
- collect samples (few more steps)
- multiple restarts can ensure a better coverage

mixing time

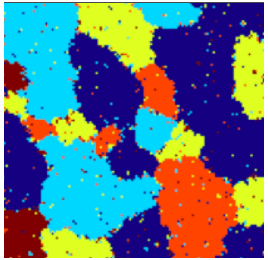
$O(\frac{1}{1-\lambda_2} \log(\frac{N}{\epsilon}))$

λ_2: 2nd largest eigenvalue of T

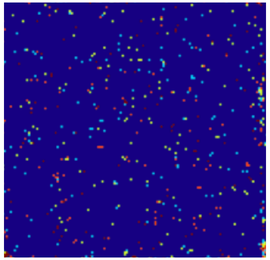
N: #states (exponential)

Example Potts model

- model $p(x) \propto \exp(\sum_i h(x_i) + \sum_{i,j \in \mathcal{E}} .66\mathbb{I}(x_i = x_j))$
- $|Val(X)| = 5$ different colors
- 128x128 grid
- Gibbs sampling



200 iterations



10,000 iterations

image : Murphy's book

Diagnosing convergence

- heuristics for diagnosing non-convergence
- difficult problem
- run multiple chains (compare sample statistics)
- auto-correlation within each chain

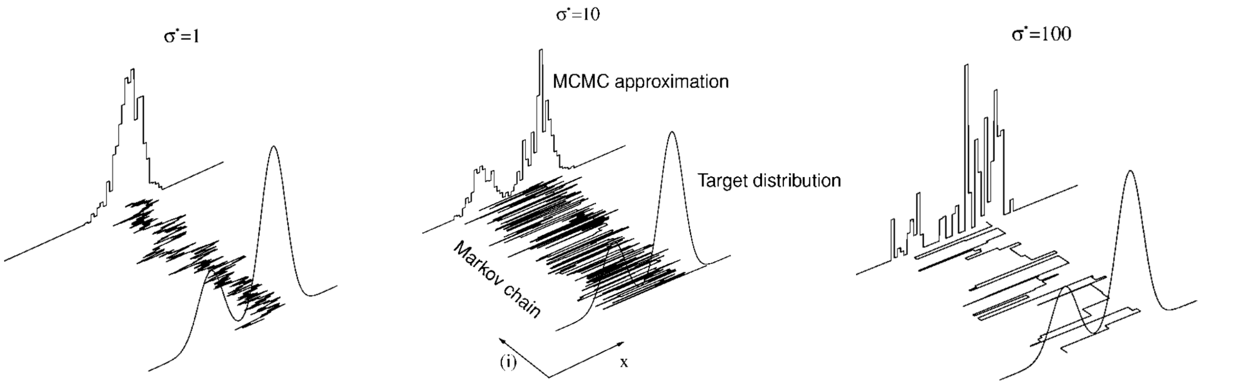
Diagnosing convergence

- heuristics for diagnosing non-convergence
- difficult problem
- run multiple chains (compare sample statistics)
- auto-correlation within each chain

example sampling from a mixture of two 1D Gaussians (3 chains: colors)

metropolis-hastings (MH) with increasing step sizes for the proposal

trace plot



high auto-correlation

step-size is too large (high rejection rate)

image: ANDRIEU et al.'03

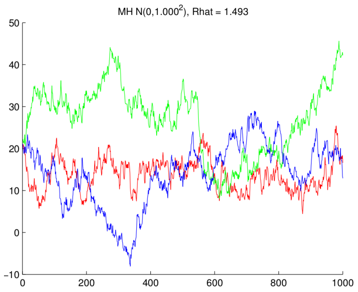
Diagnosing convergence

- heuristics for diagnosing non-convergence
- difficult problem
- run multiple chains (compare sample statistics)
- auto-correlation within each chain

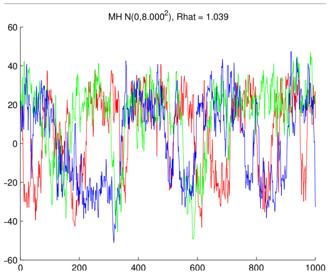
example sampling from a mixture of two 1D Gaussians (3 chains: colors)

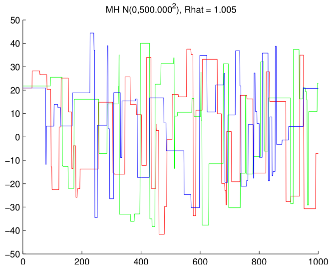
metropolis-hastings (MH) with increasing step sizes for the proposal

trace plot



high auto-correlation





step-size is too large (high rejection rate)

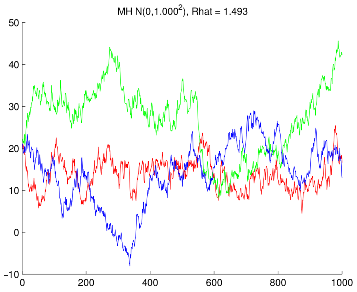
Diagnosing convergence

- heuristics for diagnosing non-convergence
- difficult problem
- run multiple chains (compare sample statistics)
- auto-correlation within each chain

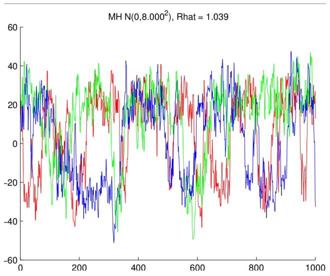
example sampling from a mixture of two 1D Gaussians (3 chains: colors)

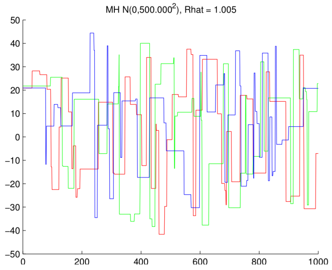
metropolis-hastings (MH) with increasing step sizes for the proposal

trace plot



high auto-correlation





step-size is too large (high rejection rate)

image: Murphy's book

Summary

Markov Chain:

- can model the "evolution" of an initial distribution
- converges to a stationary distribution

Summary

Markov Chain:

- can model the "evolution" of an initial distribution
- converges to a stationary distribution

Markov Chain Monte Carlo:

- design a Markov chain: stationary dist. is what we want to sample
- run the chain to produce samples

Summary

Markov Chain:

- can model the "evolution" of an initial distribution
- converges to a stationary distribution

Markov Chain Monte Carlo:

- design a Markov chain: stationary dist. is what we want to sample
- run the chain to produce samples

Two MCMC methods:

- Gibbs sampling
- Metropolis-Hastings