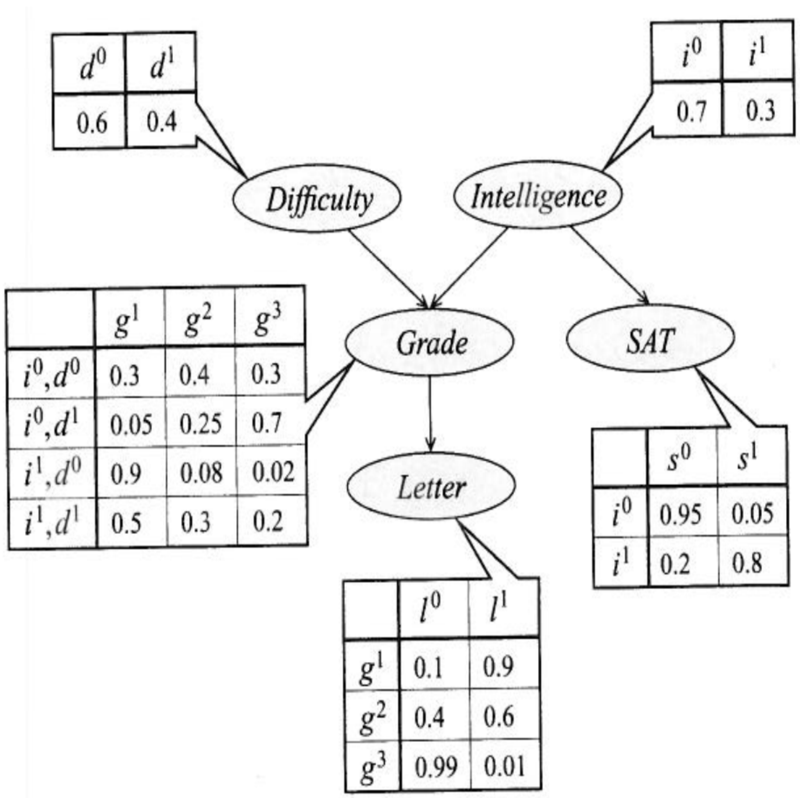
Lecture 18: March 30, 2015
cs 573: Probabilistic Reasoning
Professor Nevatia
Spring 2015

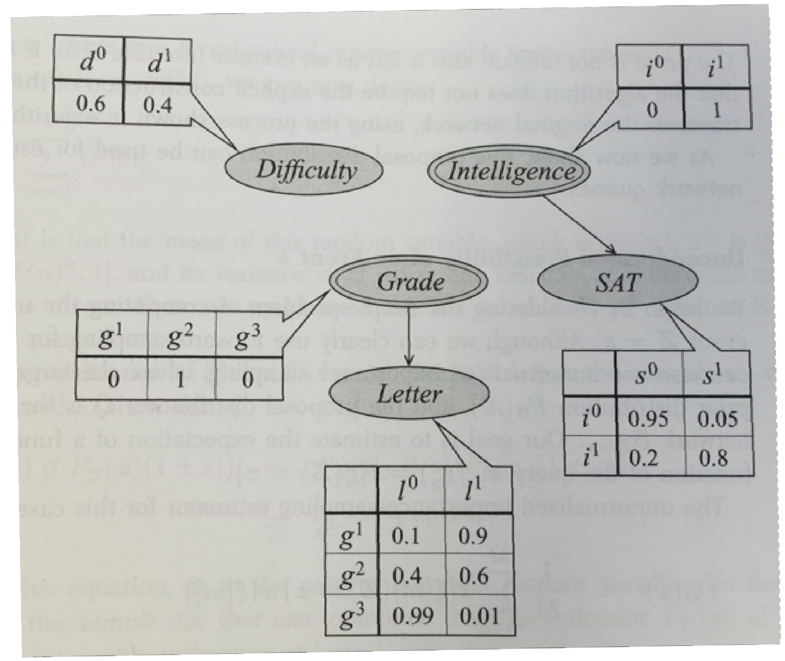
Review

- HW #5 due today
- HW #6 in two parts, one to be posted today, other part later in the week, both due 4/8/15
- Previous Lecture
  - Various sampling approaches
    - Likelihood weighting
    - Unnormalized and normalized importance sampling
    - MCMC Intro
- Today's objective
  - Markov Chain Monte Carlo (MCMC) methods
  - Intro to temporal models

Mutilated Network

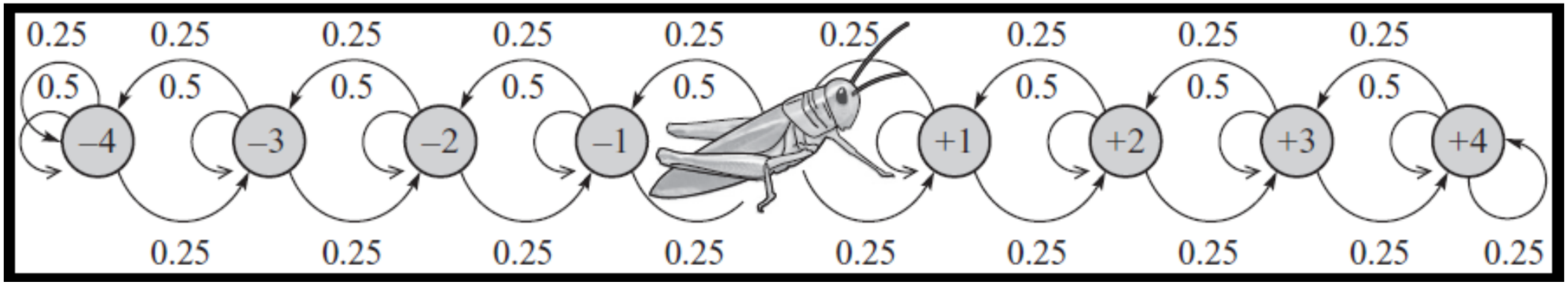
- Note: error in figure in last lecture (error in earlier edition of book)





Markov Chains

- Defined by a transition function T ($\mathbf{x} \rightarrow \mathbf{x}'$) between a pair of states ($\mathbf{x}$, $\mathbf{x}'$) which defines the probability of going from current state $\mathbf{x}$ to new state $\mathbf{x}'$. A state is given by assignments to variables.
  - Note T will have n^2 entries if $\mathbf{X}$ can take n values
  - Can be viewed as a matrix
- Homogeneous Markov Chain
  - Transition probability does not change over time
- Grasshopper Example
  - State: 9 integers from -4 to +4
  - Initial position: 0
  - At each instance, T(i→i) =.5, T(i→i-1) = .25, T (i→i+1) = .25
  - At two ends, can not jump beyond (stays in the same state)
    - T(4 → 4)= .75
  - Write as a transition matrix



$$P^{(t+1)}(\boldsymbol{X}^{(t+1)} = \boldsymbol{x}') = \sum_{\boldsymbol{x} \in Val(\boldsymbol{X})} P^{(t)}(\boldsymbol{X}^{(t)} = \boldsymbol{x}) \mathcal{T}(\boldsymbol{x} \rightarrow \boldsymbol{x}').$$

At t=0, $P(X^0 = 0) = 1$

At t =1, $P(X^1 = 0) = .5$, $P(X^1 = 1) = .25$, $P(X^1 = -1) = .5$

At t =2, $P(X^2 = 0) = .5x.5 + .25\ x.25 + .25x.25 = .375$

$P(X^2 = 1 \text{ or } -1) = .5 \text{ x } .25 + .25 \text{ x } .5 = .25$

$P(X^2 = 2 \text{ or } -2) = ..25 \text{ x } .25 = .0625$

Position probability converges to a nearly uniform distribution with time for this example

Stationary Distribution

- At convergence, we expect:

$$P^{(t)}(\boldsymbol{x}') \approx P^{(t+1)}(\boldsymbol{x}') = \sum_{\boldsymbol{x} \in Val(\boldsymbol{X})} P^{(t)}(\boldsymbol{x}) \mathcal{T}(\boldsymbol{x} \to \boldsymbol{x}')..$$

- Stationary Distribution

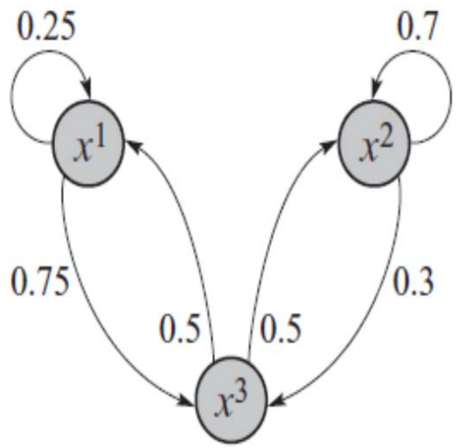
A distribution $\pi(\boldsymbol{X})$ *is a* stationary distribution *for a Markov chain* $\mathcal{T}$ *if it satisfies:*

$$\pi(\boldsymbol{X} = \boldsymbol{x}') = \sum_{\boldsymbol{x} \in Val(\boldsymbol{X})} \pi(\boldsymbol{X} = \boldsymbol{x}) \mathcal{T}(\boldsymbol{x} \to \boldsymbol{x}').$$

A stationary distribution is also called an *invariant distribution.*

- In linear algebra formulation: T $\pi(x) = \pi(x)$; *i.e.* the stationary distribution is an eigenvector of the transition matrix with eigenvalue = 1

Example 12.7



Stationary distribution must satisfy

$$\begin{aligned}\pi(x^1) &= 0.25\pi(x^1) + 0.5\pi(x^3)\\ \pi(x^2) &= 0.7\pi(x^2) + 0.5\pi(x^3)\\ \pi(x^3) &= 0.75\pi(x^1) + 0.3\pi(x^2),\end{aligned}$$

Transition equations

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

Normalize

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

Some chains can oscillate between two distributions: *periodic chains*

In some, there are distinct regions not reachable from others; stationary distribution depends on choice of first sample: *Reducible Markov Chains*

Regular Chain

- We consider only chains with unique, stationary distributions
- *Regular chain*: There exists a k such that probability of going from x to x' in exactly k steps is > 0
  - True for grasshopper (9 steps) and fig 12.4 (2 steps)
- Thm 12.3: If a chain is regular, it has a unique stationary distribution

Multiple Transition Kernels

- Consider grasshopper to be hopping on a 2-D grid
- Define a separate transition model for each dimension (X, Y)
- Each such model is called a *kernel* .
- Cycle through multiple kernels one at a time or by some stochastic choice
- Multiple kernels have stationary distributions if each kernel has a stationary distribution
- Gibbs sampler is a special case (see next slide)

Gibbs Chain

- Gibbs chain follows the formal transition rule:

$$T_i \{(\boldsymbol{x}_{-I}, x_i) \rightarrow (\boldsymbol{x}_{-I}, x_i')\} = P(x_i' \mid \boldsymbol{x}_{-i})$$

- Can be shown that posterior distribution $P_\Phi(X \mid e)$ is a stationary distribution
- Can be shown that only the assignments in the Markov blanket of X_i matter in the equation above

Reversible Chain, Dynamic Balance

- A chain is said to be *reversible* if there exists a unique distribution π such that for all x, x', it satisfies the following *detailed balance* equation

$$\pi(x)\ \mathrm{T}(x \to x') = \pi(x')\ \mathrm{T}(x' \to x)$$

- Pick a random starting state from $\pi(x)$ and a random transition according to transition probability, then probability of a transition from x to x' is same as from x' to x.
- If a chain is regular and satisfies detailed balance according to π then π is the unique stationary distribution of the chain
- Gibbs-chain is a reversible chain
- So is a chain constructed by the Metropolis-Hastings algorithm (next slide)

Metropolis-Hastings Algorithm

- Sample not according to P (may be hard to compute) but some other distribution Q
- Let T^Q define a transition model from x to x'
- We accept this transition according to some probability $A(x \to x')$; Effectively, the transition model is:

$$\begin{array}{rcl} \mathcal{T}(x \to x') & = & \mathcal{T}^Q(x \to x')\mathcal{A}(x \to x') \qquad x \neq x' \\ \mathcal{T}(x \to x) & = & \mathcal{T}^Q(x \to x) + \sum_{x' \neq x} \mathcal{T}^Q(x \to x')(1 - \mathcal{A}(x \to x')). \end{array}$$

- Choice of Q is rather arbitrary but resulting chain must be regular: for example, we can choose a uniform distribution over values of X_i or a Gaussian over current state x
- To achieve detailed balance, we must have for x not equal to x'

$$\pi(x)\mathcal{T}^Q(x \to x')\mathcal{A}(x \to x') = \pi(x')\mathcal{T}^Q(x' \to x)\mathcal{A}(x' \to x).$$

- See next slide for solution

Metropolis-Hastings Algorithm

- One solution to previous equation is:

$$\mathcal{A}(x \to x') = \min\left[1, \frac{\pi(x')T^Q(x' \to x)}{\pi(x)T^Q(x \to x')}\right]$$

  - Metropolis-Hastings algorithm
- Consider case where Q is a uniform distribution, T^Q terms cancel in equation above and we get ratio of $\pi(x')$ to $\pi(x)$
  - If first is larger, we always transition to x' (with probability 1), but may also transition when $\pi(x')$ is smaller.
    - Like stochastic hill climbing
- Thm 12.5: For any proposal distribution Q, the Markov chain defined by previous slide with the above acceptance probability:
  - If the resulting chain is regular, it has a stationary distribution π.

MCMC for Graphical Models

$$\mathcal{A}(\boldsymbol{x}_{-i}, x_i \to \boldsymbol{x}_{-i}, x_i') = \min\left[1, \frac{\pi(\boldsymbol{x}_{-i}, x_i')\mathcal{T}_i^{Q_i}(\boldsymbol{x}_{-i}, x_i' \to \boldsymbol{x}_{-i}, x_i)}{\pi(\boldsymbol{x}_{-i}, x_i)\mathcal{T}_i^{Q_i}(\boldsymbol{x}_{-i}, x_i \to \boldsymbol{x}_{-i}, x_i')}\right]$$
$$= \min\left[1, \frac{P_\Phi(x_i', \boldsymbol{x}_{-i})}{P_\Phi(x_i, \boldsymbol{x}_{-i})} \frac{\mathcal{T}_i^{Q_i}(\boldsymbol{x}_{-i}, x_i' \to \boldsymbol{x}_{-i}, x_i)}{\mathcal{T}_i^{Q_i}(\boldsymbol{x}_{-i}, x_i \to \boldsymbol{x}_{-i}, x_i')}\right].$$

$$\frac{P_\Phi(x_i', \boldsymbol{x}_{-i})}{P_\Phi(x_i, \boldsymbol{x}_{-i})} = \frac{P_\Phi(x_i' \mid \boldsymbol{x}_{-i})P_\Phi(\boldsymbol{x}_{-i})}{P_\Phi(x_i \mid \boldsymbol{x}_{-i})P_\Phi(\boldsymbol{x}_{-i})}$$
$$= \frac{P_\Phi(x_i' \mid \boldsymbol{x}_{-i})}{P_\Phi(x_i \mid \boldsymbol{x}_{-i})}.$$

As for Gibbs sampling, we can use the observation that each variable X_i is conditionally independent of the remaining variables in the network given its Markov blanket. Letting $\boldsymbol{U}_i$ denote $\mathrm{MB}_\mathcal{K}(X_i)$, and $\boldsymbol{u}_i = (\boldsymbol{x}_{-i})\langle \boldsymbol{U}_i \rangle$, we have that:

$$\frac{P_\Phi(x_i' \mid \boldsymbol{x}_{-i})}{P_\Phi(x_i \mid \boldsymbol{x}_{-i})} = \frac{P_\Phi(x_i' \mid \boldsymbol{u}_i)}{P_\Phi(x_i \mid \boldsymbol{u}_i)}.$$

Mixing Time

- How long does it take for a Markov chain to “mix” or “burn in”, *i.e.* distribution is within ε of π
- Analytical derivations are skipped
- Intuitively, highly skewed distributions will mix slowly
  - Hard to transition thru low probability valleys
- In general, mixing times can be rather long
- Data Driven MCMC
  - Can help drive the chain to high probability areas
  - We can use observations (data) to define the Q function

A Computer Vision Example

- Example follows
- For illustration of DDMCMC only; material not included for assignments or exams

Model-based segmentation: A Bayesian Approach

- Problem Statement
  - Given a *foreground blob* (moving pixels) consisting of moving humans, estimate the position, size and the pose of the humans as a secondary objective
- Issues
  - Given a configuration (number, size, position and pose of hypotheses), we can evaluate its goodness (likelihood)
  - However, search space is too large for exhaustive search
  - Gradient ascent and similar methods can easily be locked into local maxima

Prior probabilities and Likelihood Function

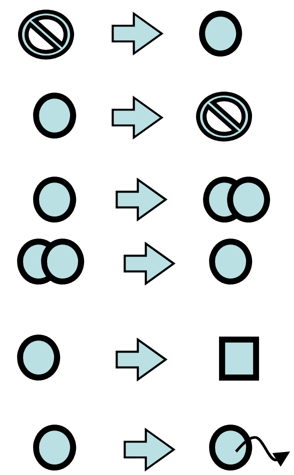
- Define P(X) as a product of several terms: penalize large number of objects, assume some distribution of heights and other parameters
  - Details unimportant for today's discussion
- Likelihood function
  - Given a sample for X= x (*i.e.* given number of humans, their positions and other parameters), we can compute overlap between predicted blob and observed blob
  - Likelihood is a function of this overlap and some other blob properties

Computing the MAP by MCMC

- Define a transition function that creates a regular chain.
- Use an *informed* proposal distribution (probability of this distribution is likely to be higher where the probability of the actual distribution is also high)
- Data driven MCMC (DDMCMC) uses the data to define function Q
  - For this problem, we use various heuristic cues, primarily an estimate of head like shapes present in the image

Reversible Markov chain dynamics

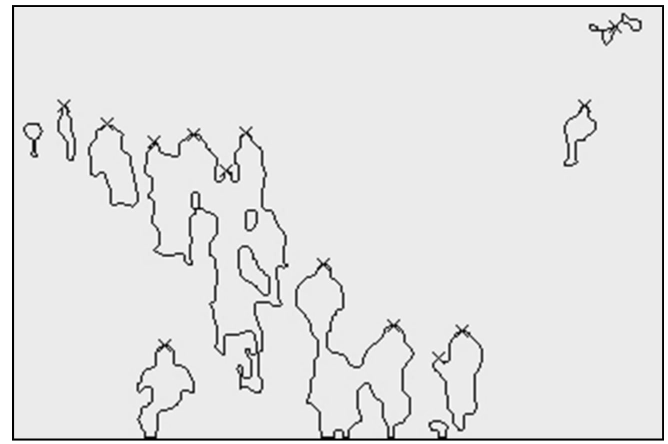
- Dynamics to explore the solution space
  - Adding an object
  - Removing an object
  - Split an object into two
  - Merge two objects into one
  - Switching between different models
  - Stochastic diffusion



- *Jumps* between subspaces of different object number and *diffuses* within each subspace
- In each iteration, one action is chosen randomly
- Results in a Markov chain which is *reversible*, *irreducible* and *aperiodic*.

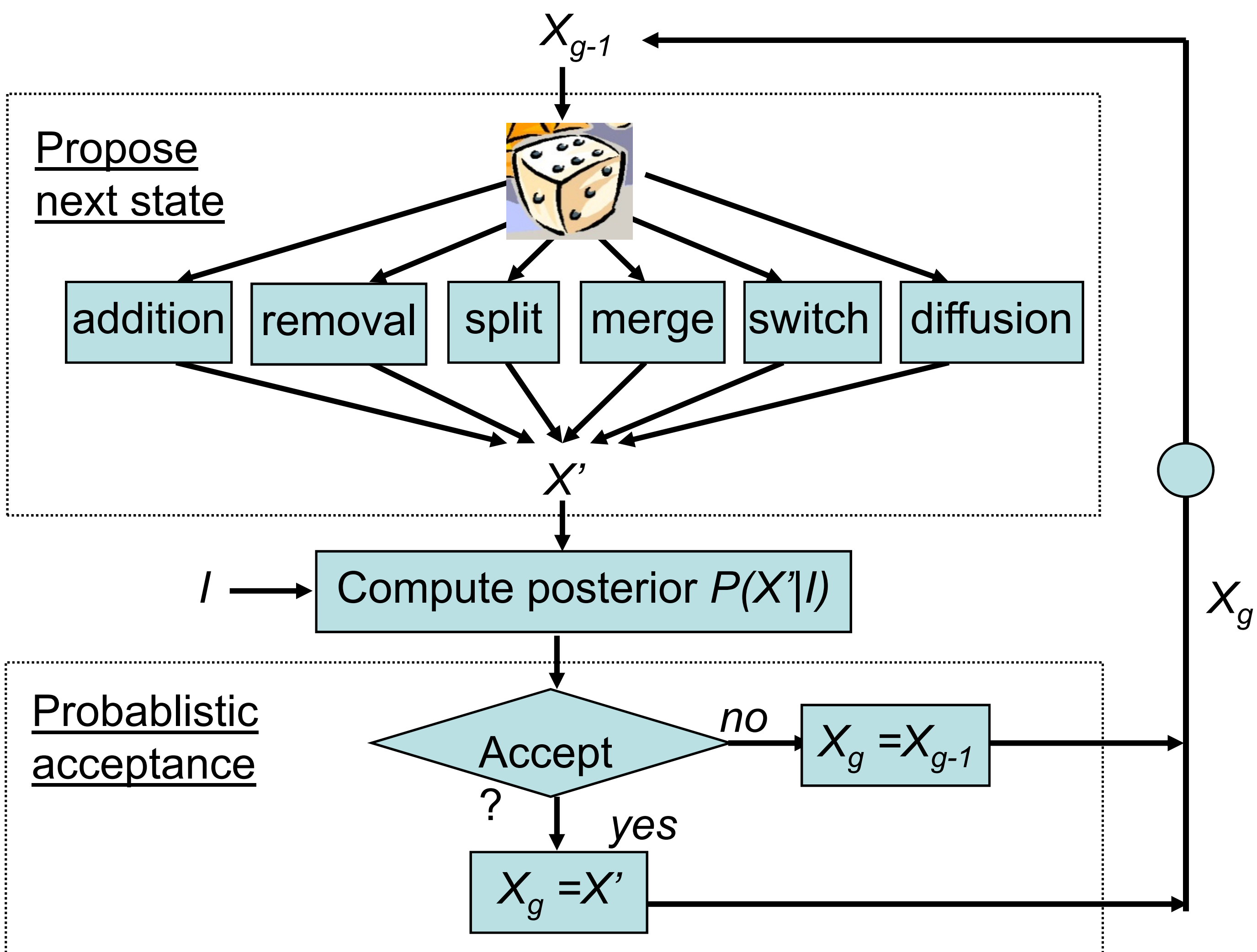
Informed proposals

- Addition 1: head candidates by foreground boundaries
  - Peaks of the foreground boundary
  - Does not work for interior heads
- Addition 2: head candidates by intensity edges



Ω

Summary of the algorithm



- The number of iterations needed depends on the complexity of the data

Result: sequence “Topping”

- 2000 iterations per frame

Next Class

- Read sections 6.2, 15.1 and 15.2 of the KF book