# Inference with Markov Chains

## Outline

I. Gibbs sampling

II. Markov chains

III. Metropolis-Hastings sampling

* Figures are either from the textbook site.

# I. Gibbs Sampling

- A Markov Chain Monte Carlo (MCMC) algorithm
  - specifies a value for every variable at the current state.
  - generates a next state by making random changes to the current state.
- *Markov chain* is a random process that generates a sequence of states.

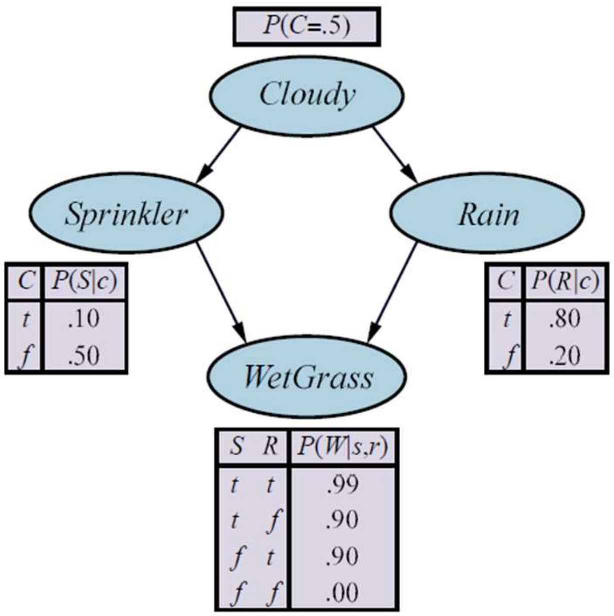
Gibbs sampling (well suited for Bayes nets) is an MCMC algorithm that

- starts with an arbitrary state,
- fix evidence variables at their observed values, and
- generates a next state by randomly sampling a value for a nonevidence variable $X_i$ chosen according to probability distribution $\rho(i)$.

  $X_i$ is independent of all the variables outside of its *Markov blanket* (consisting its parents, children, and children's other parents).

# Example of Gibbs Sampling

Gibbs sampling for $X_i$ is conditioned on the current values of the variables in its Markov blanket.




**Query** $\boldsymbol{P}(Rain \mid Sprinkler = true, WetGrass = true)$

randomly generated values for nonevidence variables *Cloudy* and *Rain*

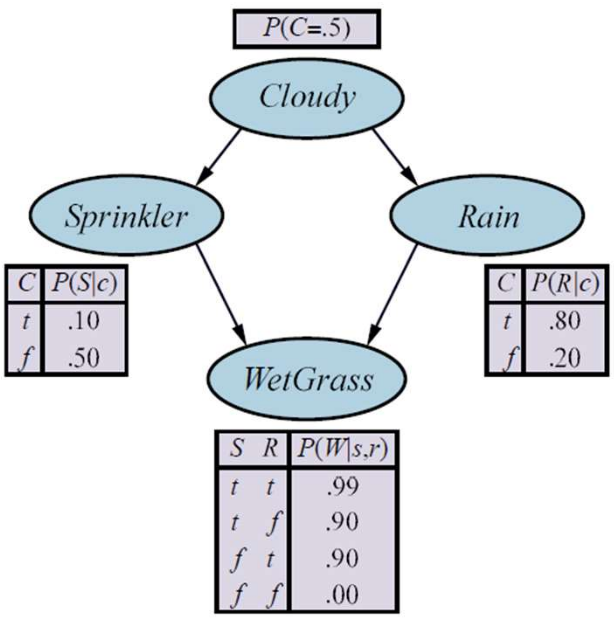
- Initial state [*true*,*true*,*false*,*true*]

evidence variables *Sprinkler* and *WetGrass* fixed to their observed values

Order: *Cloudy*, *Sprinkler*, *Rain*, *WetGrass*

# Example (cont'd)

**Query** $\boldsymbol{P}(Rain \mid Sprinkler = true, WetGrass = true)$




Order: *Cloudy*, *Sprinkler*, *Rain*, *WetGrass*

[*true*,*true*,*false*,*true*]

(initial state)

→ [*false*,*true*,*true*,*true*]

(new current state)

- Non-evidence variables are then sampled in random order following some probability distribution $\rho(i)$.
  - *Cloudy* is chosen and sampled given the current values of its Markov blanket $\{Sprinkler, Rain\}$.
    - Sampling distribution:

      $\boldsymbol{P}(Cloudy \mid Sprinkler = true, Rain = false\}$

    - Sampling result: *Cloudy* = *false*.
  - *Rain* is chosen next and sampled given the current values of its Markov blanket $\{Cloudy, Sprinkler, WetGrass\}$.
    - Sampling distribution:

      How to calculate?

      $\boldsymbol{P}(Rain \mid Cloudy = false, Sprinkler = true, WetGrass = true)$

    - Sampling result: *Rain* = *true*.

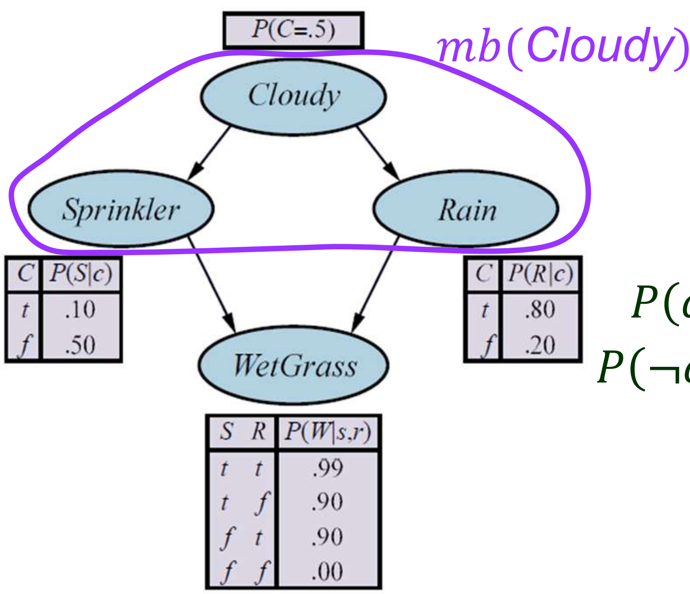
# Markov Blanket Distribution

$MB(X_i)$: variables in the Markov blanket of $X_i$.

$mb(X_i)$: values of the variables in $MB(X_i)$.

$\boldsymbol{P}(X_i \mid mb(X_i)\}$ is determined as follows:

$$P(x_i \mid mb(X_i)) = \alpha P(x_i \mid \textit{parents}(X_i)) \prod_{Y_j \in \textit{Children}(X_i)} P(y_j \mid \textit{parents}(Y_j))$$




- Sampling distribution:

  $\boldsymbol{P}(\textit{Cloudy} \mid \textit{Sprinkler} = \textit{true}, \textit{Rain} = \textit{false}\}$

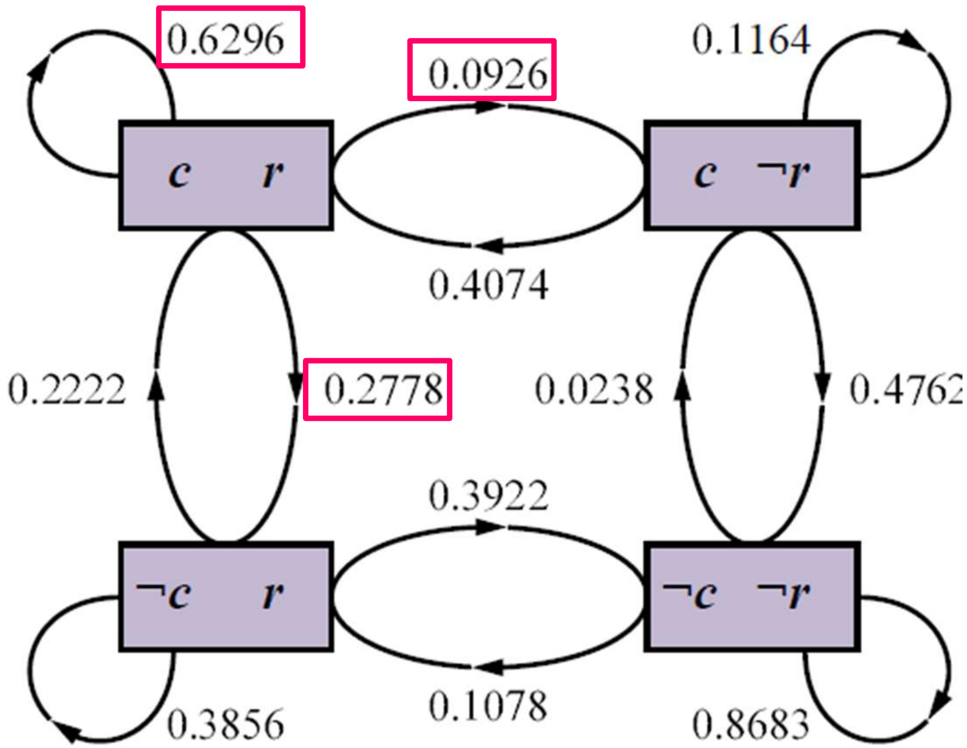
$$P(c \mid s, \neg r) = \alpha P(c) P(s \mid c) P(\neg r \mid c) = \alpha 0.5 \cdot 0.1 \cdot 0.2$$
$$P(\neg c \mid s, \neg r) = \alpha P(\neg c) P(s \mid \neg c) P(\neg r \mid \neg c) = \alpha 0.5 \cdot 0.5 \cdot 0.8$$

$$\boldsymbol{P}(C \mid s, \neg r\} = \alpha \langle 0.001, 0.020 \rangle \approx \langle 0.048, 0.952 \rangle$$

# II. Markov Chain

**Query** $\boldsymbol{P}(Rain \mid Sprinkler = true, WetGrass = true)$

A *state* need only include all nonevidence variables.




Markov chain from uniform choice of the two nonevidence variables ($\rho(Cloudy) = \rho(Rain) = 0.5$)

Probabilities with all the outgoing links of each node sum to 1, e.g., $0.6296 + 0.0926 + 0.2778 = 1$.

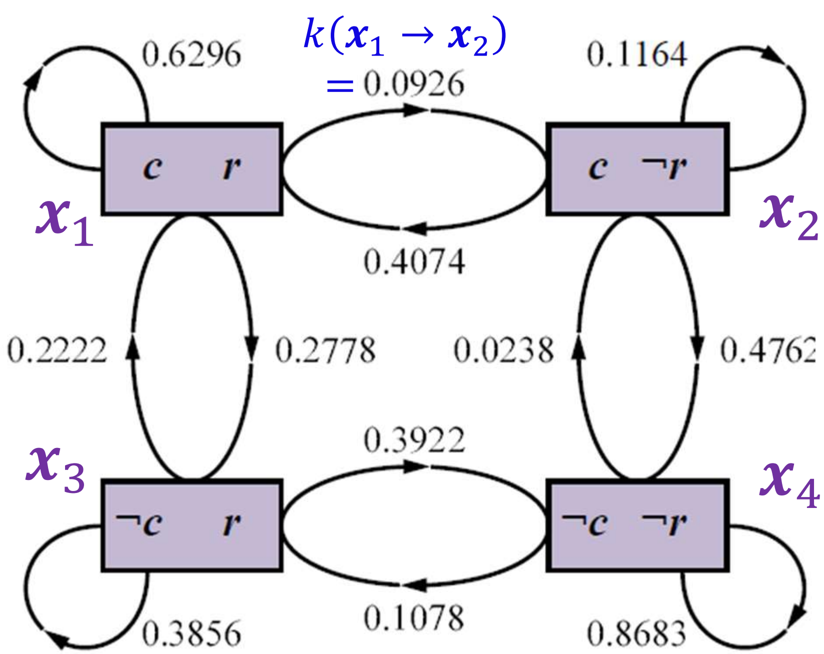
- Gibbs sampling simply wanders around in the graph, following links with probabilities.
- Every state visited is a sample that contributes to the estimate for the query variable *Rain*.

If the process visits 20 states with *Rain* = *true* and 60 states with *Rain* = *false*, then the answer to the query is $\alpha\langle 20,60\rangle = \langle 0.25,0.75\rangle$.

# Analysis of Markov Chains

Why does Gibbs sampling work? Or, why does its estimates converge to correct values in the limit?

*Transition kernel* $k$ assigns a probability $k(\boldsymbol{x} \to \boldsymbol{x}')$ to a transition from state $\boldsymbol{x}$ to state $\boldsymbol{x}'$.




$\pi_t(\boldsymbol{x})$: probability that the system is in state $\boldsymbol{x}$ after $t$ transitions

$$\boxed{\pi_{t+1}(\boldsymbol{x}') = \sum_{x} \pi_t(\boldsymbol{x}) k(\boldsymbol{x} \to \boldsymbol{x}')}$$

$$\pi_{t+1}(\boldsymbol{x}_2) = 0.0926\, \pi_t(\boldsymbol{x}_1) + 0.1164 \pi_t(\boldsymbol{x}_2) + 0.0238\, \pi_t(\boldsymbol{x}_4)$$

The chain has reached its stationary distribution if $\pi_{t+1}(\boldsymbol{x}) = \pi_t(\boldsymbol{x})$ for all $\boldsymbol{x}$.

# Stationary Distribution

A distribution $\pi$ of the Markov chain is *stationary* if

$$\pi(\boldsymbol{x}') = \sum_{x} \pi(\boldsymbol{x}) k(\boldsymbol{x} \to \boldsymbol{x}') \quad \text{for all } \boldsymbol{x}, \boldsymbol{x}'.$$

Such a probability distribution remains unchanged in the Markov chain as time progresses.

A kernel $k$ is *ergodic* if every state is reachable from every other state and there exists no strictly periodic cycles.

There exists exactly one stationary distribution for every ergodic kernel of the Markov chain.

# Achieving a Stationary Distribution

In a stationary distribution $\pi$, the expected "outflow" from each state is equal to the expected "inflow" from all the other states.

$$\underbrace{\pi(\overset{\text{"population"}}{\boldsymbol{x}'})}_{\substack{\text{expected}\\\text{"outflow"}}} = \underbrace{\sum_{x} \pi(\boldsymbol{x}) k(\boldsymbol{x} \to \boldsymbol{x}')}_{\text{expected "inflow"}}$$

A *detailed balance* $k$ with $\pi$ is a distribution that satisfies

$$\pi(\boldsymbol{x}) k(\boldsymbol{x} \to \boldsymbol{x}') = \pi(\boldsymbol{x}') k(\boldsymbol{x}' \to \boldsymbol{x}) \qquad \text{for all } \boldsymbol{x}, \boldsymbol{x}'.$$

The detailed balance $k$ makes $\pi(\boldsymbol{x})$ stationary because

$$\begin{aligned}\sum_{\boldsymbol{x}} \pi(\boldsymbol{x}) k(\boldsymbol{x} \to \boldsymbol{x}') &= \sum_{\boldsymbol{x}} \pi(\boldsymbol{x}') k(\boldsymbol{x}' \to \boldsymbol{x}) \\ &= \pi(\boldsymbol{x}') \boxed{\sum_{\boldsymbol{x}} k(\boldsymbol{x}' \to \boldsymbol{x})} = 1\end{aligned}$$

# Correctness of Gibbs Sampling

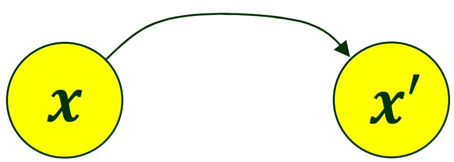
The stationary distribution of the Gibbs sampling process is exactly the posterior distribution for the nonevidence variable conditioned on the evidence.

- In Gibbs sampling, a variable $X_i$ is chosen and sampled conditionally on
  - the current values of all the other variables,
  - equivalently, when sampling a Bayes net, the variable's Markov blanket.

# Transition Kernel for Gibbs Sampling

$\overline{\boldsymbol{X}}_i$: variables except $X_i$ and evidence variables.

$\overline{\boldsymbol{x}}_i$: their values.




Case 1. The states $\boldsymbol{x}$ and $\boldsymbol{x}$' differ in $\geq 2$ variables. Since Gibbs sampling changes only one variable, we set

$$k(\boldsymbol{x} \to \boldsymbol{x}') = 0$$

Case 2. The states $\boldsymbol{x}$ and $\boldsymbol{x}$' differ in the value of exactly one variable $X_i$, which changes from $x_i$ to $x_i{}'$. That is, $\boldsymbol{x} = (x_i, \overline{\boldsymbol{x}}_i)$ and $\boldsymbol{x}' = (x_i', \overline{\boldsymbol{x}}_i)$.

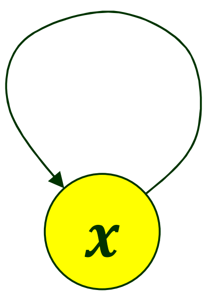
$$k(\boldsymbol{x} \to \boldsymbol{x}') = k((x_i, \overline{\boldsymbol{x}}_i) \to (x_i', \overline{\boldsymbol{x}}_i)) = \boxed{\rho(i)} P(x_i' \mid \overline{\boldsymbol{x}}_i)$$

probability of choosing $X_i$

The transition probability is the product of the probability of selecting the variable $X_i$ (out of all the nonevidence variables) with the probability of selecting $x_i{}'$ (out of all the values of $X_i$).

# Completing the Definition

Case 3. The states are the same $\boldsymbol{x} = \boldsymbol{x}'$. Any variable could be chosen but then the sampling process reproduce the current value of the variable.




$$k(\boldsymbol{x} \to \boldsymbol{x}') = \sum_{\boldsymbol{i}} \rho(i) k((x_i, \bar{\boldsymbol{x}}_i) \to (x_i', \bar{\boldsymbol{x}}_i)) = \sum_{\boldsymbol{i}} \rho(i) P(x_i \mid \bar{\boldsymbol{x}}_{\boldsymbol{i}})$$

# Correctness of Gibbs Sampling

**Theorem** The previously defined kernel $k(\boldsymbol{x}' \to \boldsymbol{x})$ for Gibbs sampling has a stationary distribution equal to $P(\boldsymbol{x} \mid \boldsymbol{e})$, the true posterior distribution on the nonevidence variables.

**Proof** It suffices to show that, with $\pi(\boldsymbol{x}) = P(\boldsymbol{x} \mid \boldsymbol{e})$, the following condition for $k$ in detailed balance is satisfied:

$$\pi(\boldsymbol{x})k(\boldsymbol{x} \to \boldsymbol{x}') = \pi(\boldsymbol{x}')k(\boldsymbol{x}' \to \boldsymbol{x}) \qquad \text{for all states } \boldsymbol{x}, \boldsymbol{x}'.$$

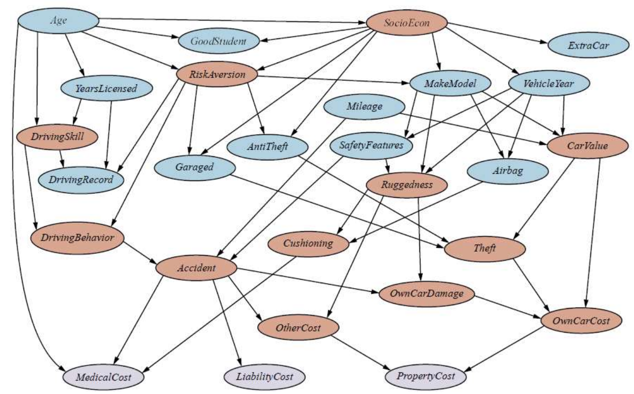
Then it follows that $k$ implies the stationarity distribution $P(\boldsymbol{x} \mid \boldsymbol{e})$.

- In the first and third cases, where $\boldsymbol{x}$ and $\boldsymbol{x}$' differ in $\geq 2$ variables and $\boldsymbol{x} = \boldsymbol{x}$', respectively, detailed balance can be easily shown to be satisfied.
- In the second case, where $\boldsymbol{x}$ and $\boldsymbol{x}$' differ in one variable $x_i$, we have

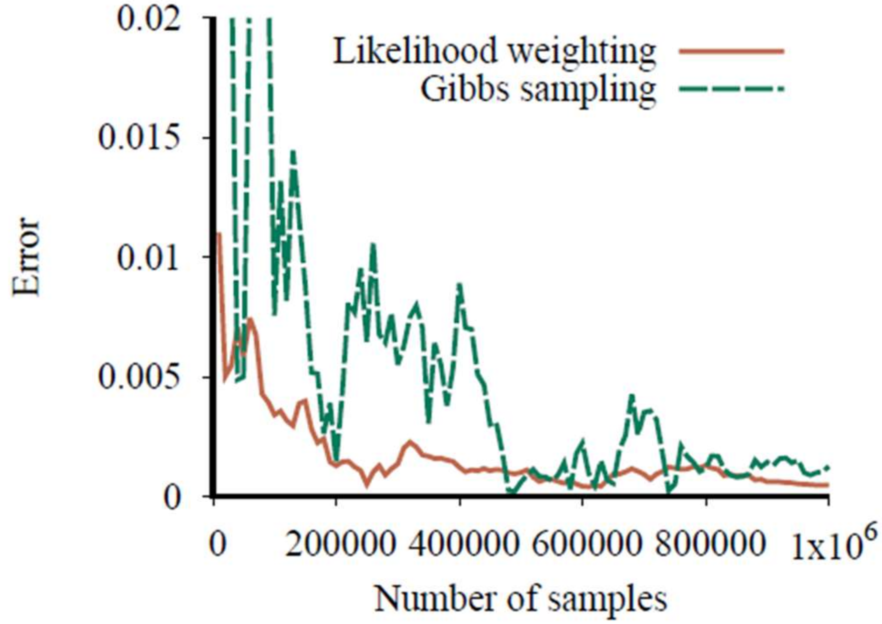
$$\begin{aligned}
\pi(\boldsymbol{x})k(\boldsymbol{x} \to \boldsymbol{x}') &= P(\boldsymbol{x} \mid \boldsymbol{e})\, \rho(i)P(x_i' \mid \overline{\boldsymbol{x}}_i, \boldsymbol{e}) = \rho(i)\ P(x_i, \overline{\boldsymbol{x}_i} \mid \boldsymbol{e})P(\boldsymbol{x}_i' \mid \overline{\boldsymbol{x}}_i, \boldsymbol{e}) \\
&= \rho(i)P(x_i \mid \overline{\boldsymbol{x}}_i, \boldsymbol{e})P(\overline{\boldsymbol{x}_i} \mid \boldsymbol{e})P(x_i' \mid \overline{\boldsymbol{x}}_i, \boldsymbol{e}) \\
&= \rho(i)P(x_i \mid \overline{\boldsymbol{x}}_i, \boldsymbol{e})\ P(x_i', \overline{\boldsymbol{x}_i} \mid \boldsymbol{e}) = \rho(i)P(x_i', \overline{\boldsymbol{x}_i} \mid e)P(x_i \mid \overline{\boldsymbol{x}}_i, \boldsymbol{e}) \\
&= \pi(\boldsymbol{x}')k(\boldsymbol{x}' \to \boldsymbol{x})
\end{aligned}$$

# Performance of Gibbs Sampling

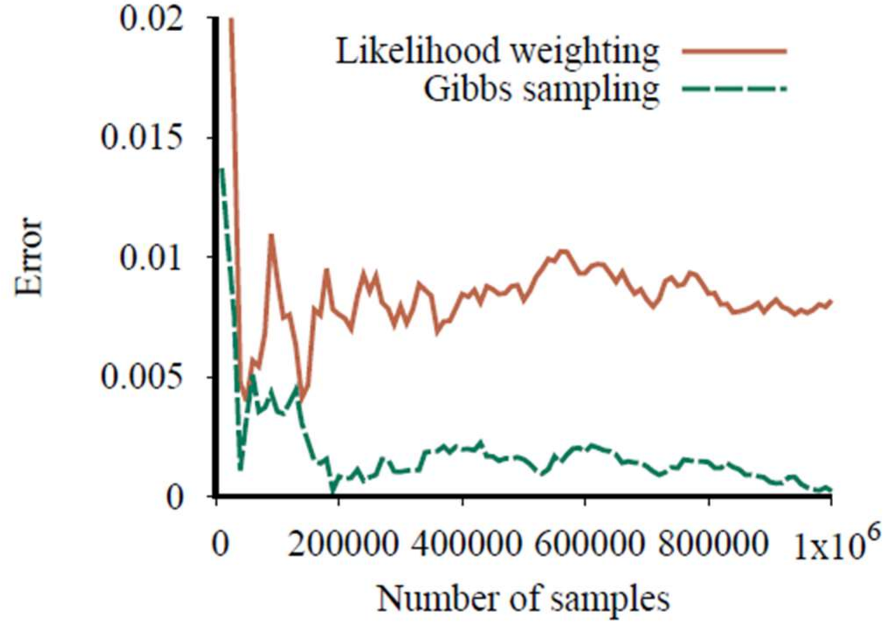
Gibbs sampling is expected to outperform likelihood weighting when evidence is downstream.




On the car insurance network:




Query on *PropertyCost*




Query on *Age* (with output observed)

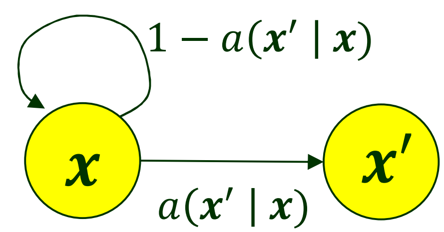
# III. Metropolis-Hastings (MH) Sampling

- The most broadly applicable Markov chain Monte Carlo algorithm.
- MH generates samples $\boldsymbol{x}$ according to a target probability distribution $\pi(\boldsymbol{x})$ (in a BN, $\pi(\boldsymbol{x}) = P(\boldsymbol{x} \mid \boldsymbol{e})$).

The transition kernel $k(\boldsymbol{x} \to \boldsymbol{x}')$ is defined as follows:

- At the current state $\boldsymbol{x}$, sample a new state $\boldsymbol{x}'$ from a proposal distribution $q(\boldsymbol{x}' \mid \boldsymbol{x})$.
- Accept or reject $\boldsymbol{x}'$ according to the acceptance probability:

$$\boxed{a(\boldsymbol{x}' \mid \boldsymbol{x}) = \min\left(1, \frac{\pi(\boldsymbol{x}')q(\boldsymbol{x} \mid \boldsymbol{x}')}{\pi(\boldsymbol{x})q(\boldsymbol{x}' \mid \boldsymbol{x})}\right)}$$




- The state transitions from $\boldsymbol{x}$ to $\boldsymbol{x}'$ in the case of acceptance, and stays at $\boldsymbol{x}$ in the case of rejection.

# Proposal Distribution for MH

The *proposal distribution* $q(\boldsymbol{x}' \mid \boldsymbol{x})$ is responsible for proposing a new state $\boldsymbol{x}'$.

**Example** $q(\boldsymbol{x}' \mid \boldsymbol{x})$ could be defined as follows:

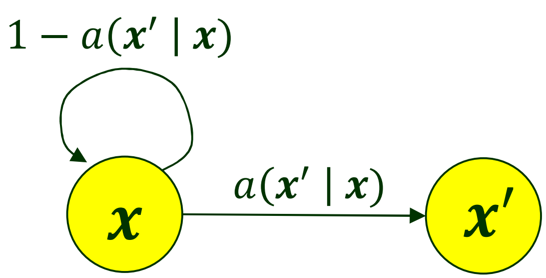
- With probability 0.95, perform a Gibbs sampling step to generate $\boldsymbol{x}'$.
- With probability 0.05, use likelihood weighting to generate $\boldsymbol{x}'$.

♦ This proposal distribution causes MH to do about 19 steps of Gibbs sampling and then generates a new state from scratch.

♦ It gets around the problem of Gibbs sampling getting stuck in one part of the state space.

# Convergence of MH

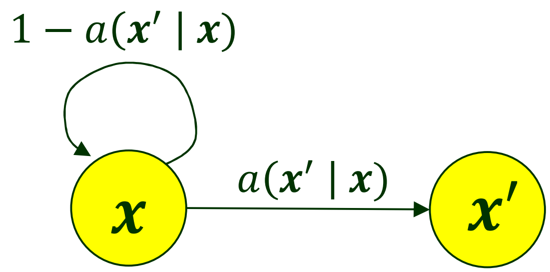
MH converges to the correct stationary distribution for any proposal distribution $q(\boldsymbol{x}' \mid \boldsymbol{x})$, provided it results in an ergodic transition kernel.




- The self-loop with $\boldsymbol{x} = \boldsymbol{x}'$ automatically satisfies the detailed balance condition:

$$\pi(\boldsymbol{x})k(\boldsymbol{x} \to \boldsymbol{x}') = \pi(\boldsymbol{x}')k(\boldsymbol{x}' \to \boldsymbol{x})$$

# Convergence of MH (cont'd)




- In the case $\boldsymbol{x} \neq \boldsymbol{x}'$, the transition can occur only if the proposal of $\boldsymbol{x}'$ is accepted.

$$k(\boldsymbol{x} \to \boldsymbol{x}') = q(\boldsymbol{x}' \mid \boldsymbol{x}) a(\boldsymbol{x}' \mid \boldsymbol{x})$$

We can show that the flow from $\boldsymbol{x}$ to $\boldsymbol{x}'$ equals that from $\boldsymbol{x}'$ to $\boldsymbol{x}$ (i.e., $k(\boldsymbol{x} \to \boldsymbol{x}')$ is in detailed balance with $\pi(\boldsymbol{x})$) as follows:

$$\begin{aligned}
\pi(\boldsymbol{x}) k(\boldsymbol{x} \to \boldsymbol{x}') &= \pi(\boldsymbol{x}) q(\boldsymbol{x}' \mid \boldsymbol{x}) a(\boldsymbol{x}' \mid \boldsymbol{x}) \\
&= \pi(\boldsymbol{x}) q(\boldsymbol{x}' \mid \boldsymbol{x}) \min\left(1, \frac{\pi(\boldsymbol{x}') q(\boldsymbol{x} \mid \boldsymbol{x}')}{\pi(\boldsymbol{x}) q(\boldsymbol{x}' \mid \boldsymbol{x})}\right) \\
&= \min(\pi(\boldsymbol{x}) q(\boldsymbol{x}' \mid \boldsymbol{x}), \pi(\boldsymbol{x}') q(\boldsymbol{x} \mid \boldsymbol{x}')) \\
&= \pi(\boldsymbol{x}') q(\boldsymbol{x} \mid \boldsymbol{x}') \min\left(\frac{\pi(\boldsymbol{x}) q(\boldsymbol{x}' \mid \boldsymbol{x})}{\pi(\boldsymbol{x}') q(\boldsymbol{x} \mid \boldsymbol{x}')}, 1\right) \\
&= \pi(\boldsymbol{x}') k(\boldsymbol{x}' \to \boldsymbol{x})
\end{aligned}$$