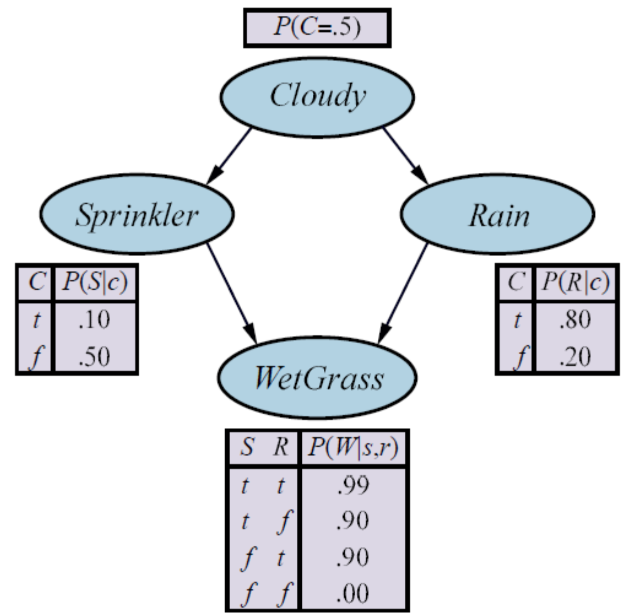
Lecture notes for *Principles of Artificial Intelligence* (Com S 4720/5720)
Yan-Bin Jia, Iowa State University

Metropolis-Hastings Sampling

Outline

I. Correctness of Gibbs sampling

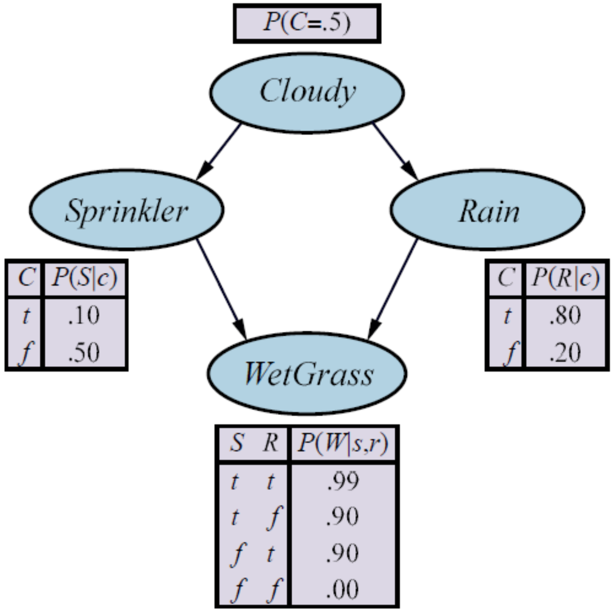
II. Metropolis-Hastings sampling



* Figures are either from the textbook site or drawn by the instructor.

I. Gibbs Sampling (reviewed)

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



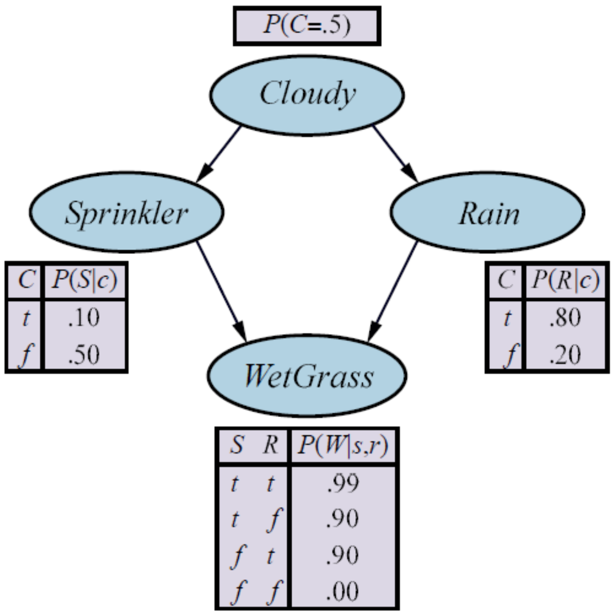
Order: *Cloudy*, *Sprinkler*, *Rain*, *WetGrass*

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

(initial state)

I. Gibbs Sampling (reviewed)

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



- Non-evidence variables are then sampled in random order following some probability distribution $\rho(i)$.

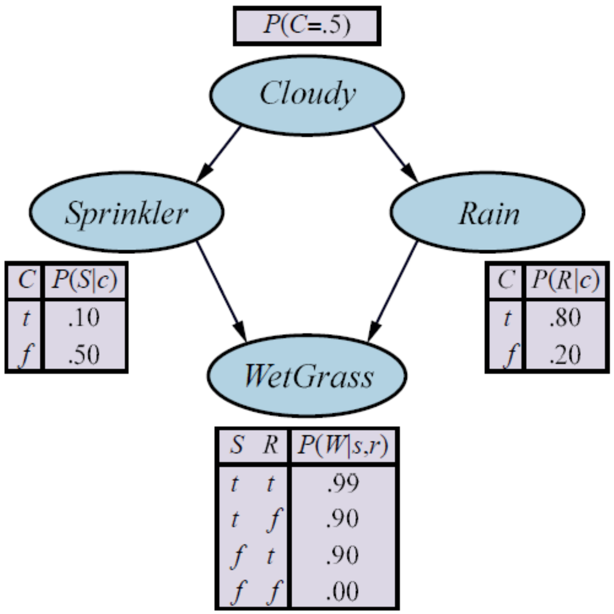
Order: *Cloudy*, *Sprinkler*, *Rain*, *WetGrass*

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

(initial state)

I. Gibbs Sampling (reviewed)

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



- 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*}.

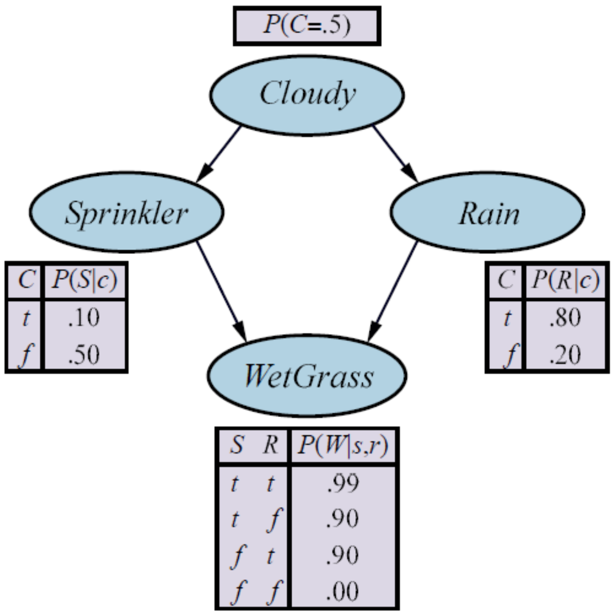
Order: *Cloudy*, *Sprinkler*, *Rain*, *WetGrass*

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

(initial state)

I. Gibbs Sampling (reviewed)

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



- 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:

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

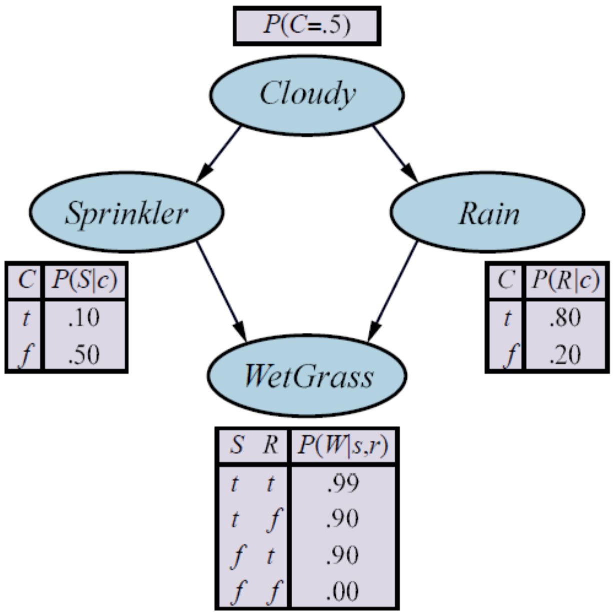
Order: *Cloudy*, *Sprinkler*, *Rain*, *WetGrass*

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

(initial state)

I. Gibbs Sampling (reviewed)

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



- 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*.

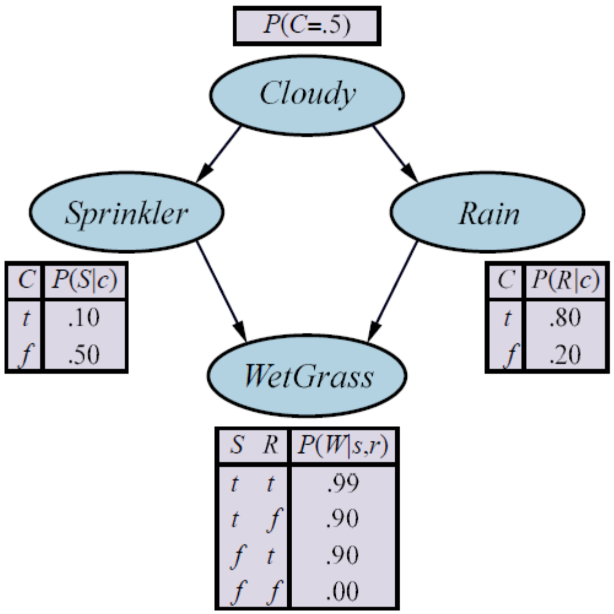
Order: *Cloudy*, *Sprinkler*, *Rain*, *WetGrass*

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

(initial state)

I. Gibbs Sampling (reviewed)

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



- 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: $Cloudy = false$.

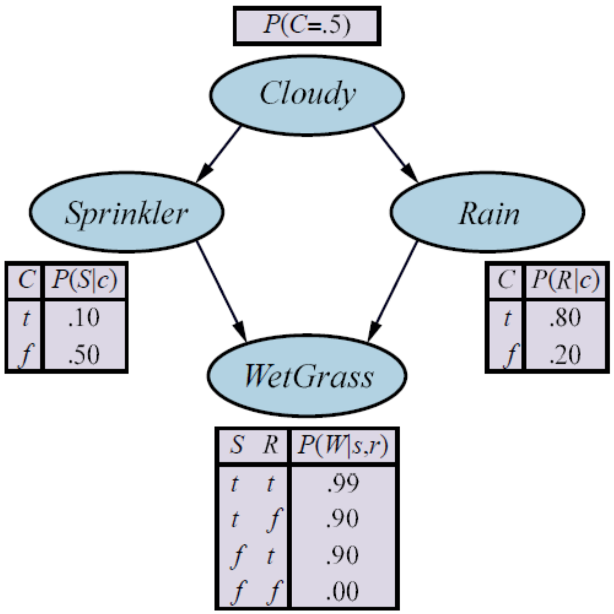
Order: *Cloudy*, *Sprinkler*, *Rain*, *WetGrass*

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

(initial state) (2nd state)

I. Gibbs Sampling (reviewed)

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



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

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

(initial state) (2nd state)

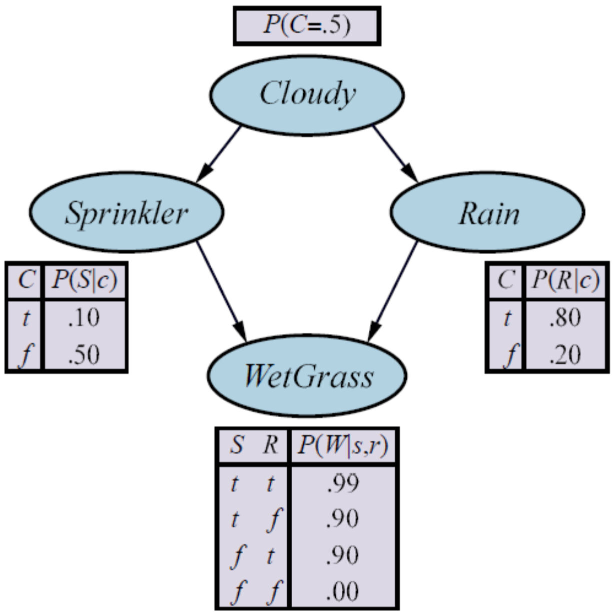
- Non-evidence variables are then sampled in random order following some probability distribution $\rho(i)$.
  - *Cloudy* is chosen and sampled given the current values of its Markov blanket {*Sprinkler*, *Rain*}.
    - Sampling distribution:

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

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

I. Gibbs Sampling (reviewed)

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



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

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

(initial state) (2nd 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:

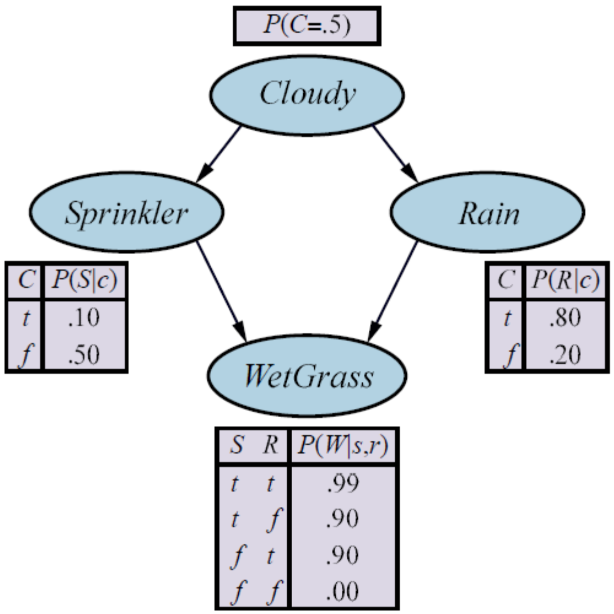
      $\mathbf{P}(Cloudy \mid Sprinkler = true, Rain = false\}$

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

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

I. Gibbs Sampling (reviewed)

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



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

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

(initial state) (2nd 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:

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

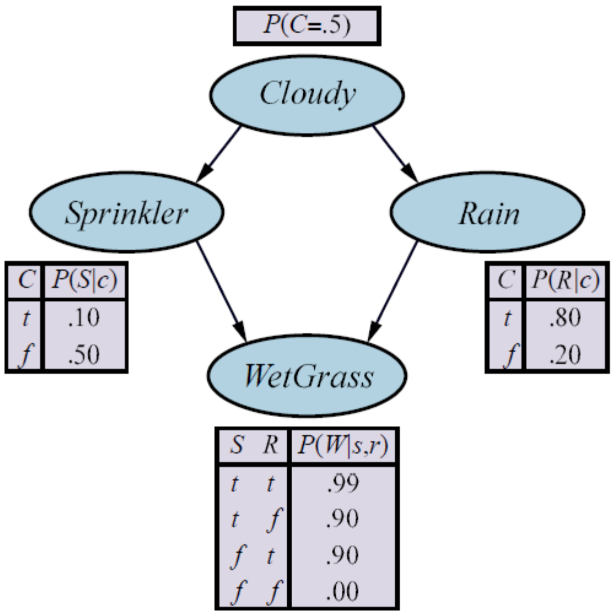
    - Sampling result: *Cloudy* = *false*.
  - *Rain* is chosen next and sampled given the current values of its Markov blanket {*Cloudy, Sprinkler*, *WetGrass*}.
    - Sampling distribution:

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

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

I. Gibbs Sampling (reviewed)

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



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

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

(initial state) (2nd state)

→ [*false*,*true*,*true*,*true*]

(3rd 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:

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

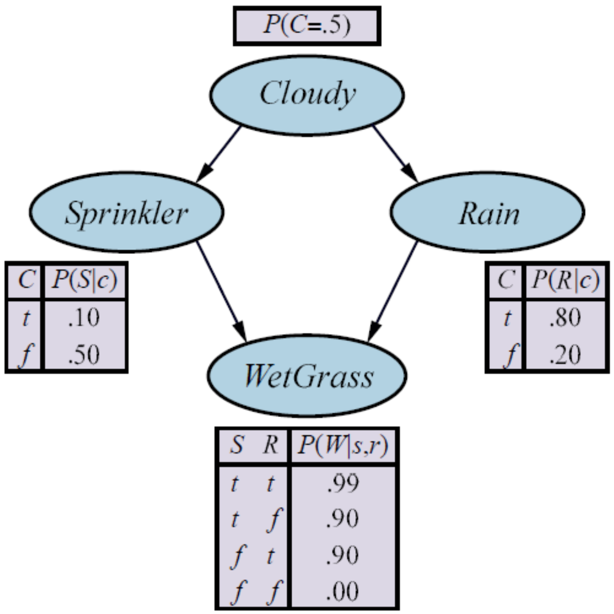
    - Sampling result: *Cloudy* = *false*.
  - *Rain* is chosen next and sampled given the current values of its Markov blanket {*Cloudy, Sprinkler*, *WetGrass*}.
    - Sampling distribution:

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

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

I. Gibbs Sampling (reviewed)

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



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

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

(initial state) (2nd state)

→ [*false*,*true*,*true*,*true*]

(3rd 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:

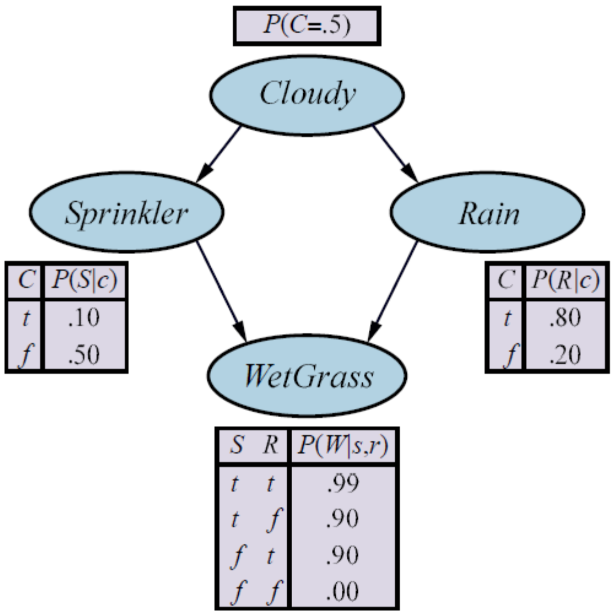
      $\boldsymbol{P}(Rain \mid Cloudy = false, Sprinkler = true, WetGrass = true)$

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

Markov blanket distribution $\boldsymbol{P}(X_i \mid mb(X_i))$

I. Gibbs Sampling (reviewed)

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



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

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

(initial state) (2nd state)

→ [*false*,*true*,*true*,*true*]

(3rd 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:

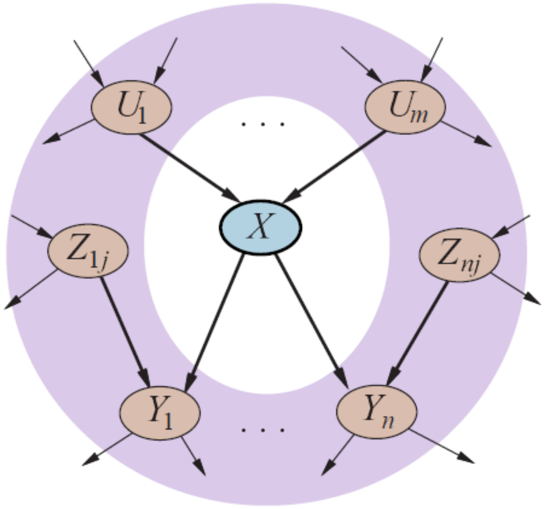
      $$\boldsymbol{P}(Rain \mid Cloudy = false,\ Sprinkler = true,\ WetGrass = true)$$

    - Sampling result: $Rain = true$.

Markov blanket distribution $\boldsymbol{P}\big(X_i \mid mb(X_i)\big)$

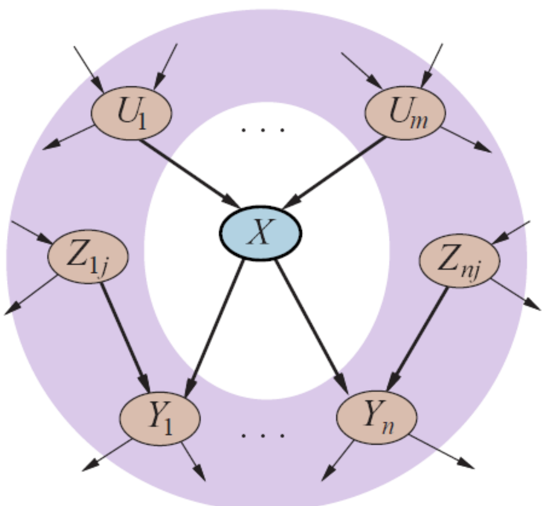
Correctness of Gibbs Sampling

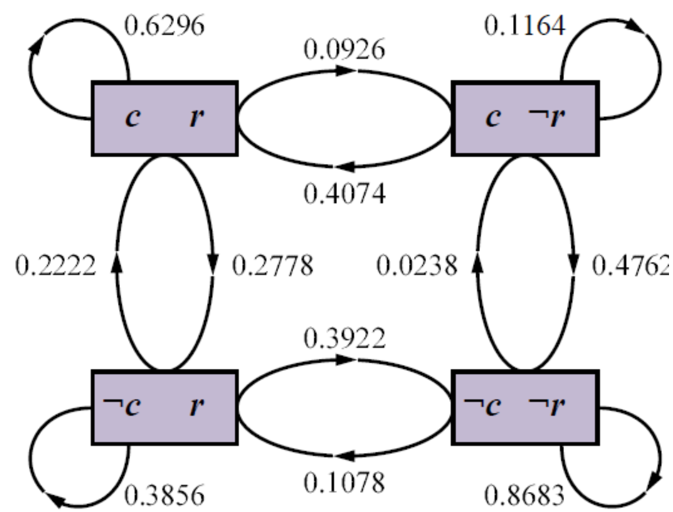
- In Gibbs sampling, a variable X_i is chosen and sampled conditionally
  - on the current values of all the other variables,
  - equivalently, when sampling a Bayes net, on the values of the variables in X_i's Markov blanket.



Correctness of Gibbs Sampling

- 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, on the values of the variables in X_i's Markov blanket.

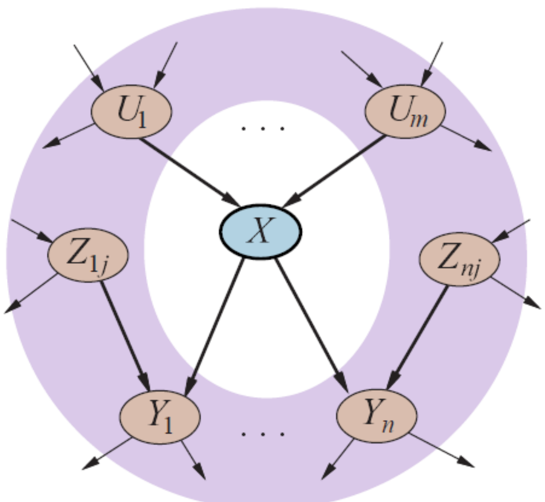


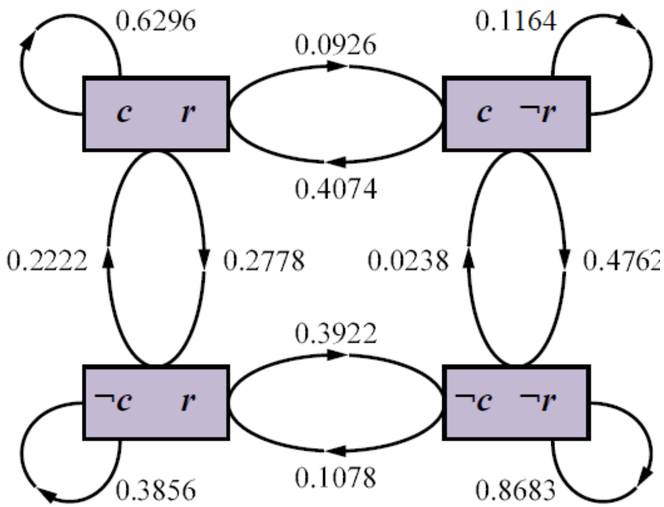


- Gibbs sampling is essentially a random walk along the induced Markov chain.

Correctness of Gibbs Sampling

- 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, on the values of the variables in X_i's Markov blanket.





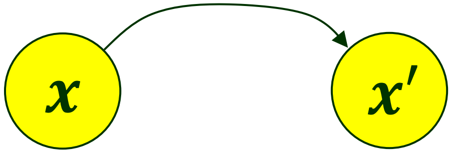
- Gibbs sampling is essentially a random walk along the induced Markov chain.

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

Defining a Transition Kernel

$\overline{\boldsymbol{X}_i}$: variables except X_i and the evidence variables. $\bar{\boldsymbol{x}}_i$: their values.

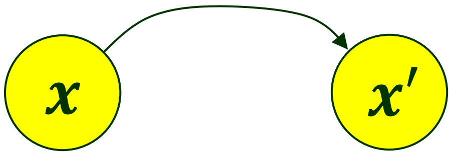
Define $k(\boldsymbol{x} \rightarrow \boldsymbol{x}')$ in the following three cases:



Defining a Transition Kernel

$\overline{\boldsymbol{X}_i}$: variables except X_i and the evidence variables. $\overline{\boldsymbol{x}_i}$: their values.

Define $k(\boldsymbol{x} \to \boldsymbol{x}')$ in the following three cases:



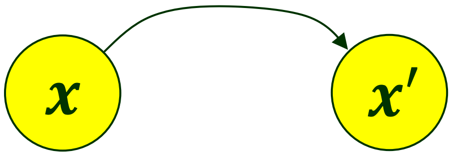
Case 1. The states $\boldsymbol{x}$ and $\boldsymbol{x}'$ differ in ≥ 2 variables. Since Gibbs sampling changes only one variable, we set

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

Defining a Transition Kernel

$\overline{\boldsymbol{X}_i}$: variables except X_i and the evidence variables. $\bar{\boldsymbol{x}}_i$: their values.

Define $k(\boldsymbol{x} \to \boldsymbol{x}')$ in the following three cases:



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

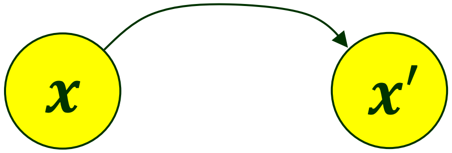
$$k(\boldsymbol{x} \to \boldsymbol{x}') = 0$$

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, $\boldsymbol{x} = (x_i, \bar{\boldsymbol{x}}_i)$ and $\boldsymbol{x}' = (x_i', \bar{\boldsymbol{x}}_i)$.

Defining a Transition Kernel

$\overline{\boldsymbol{X}_i}$: variables except X_i and the evidence variables. $\overline{\boldsymbol{x}_i}$: their values.

Define $k(\boldsymbol{x} \to \boldsymbol{x}')$ in the following three cases:



Case 1. The states $\boldsymbol{x}$ and $\boldsymbol{x}'$ differ in ≥ 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 $\boldsymbol{x}' = (x_i', \overline{\boldsymbol{x}_i})$.

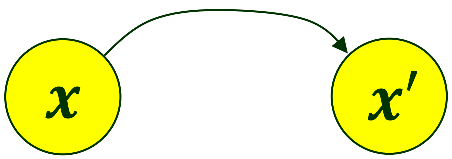
$$k(\boldsymbol{x} \to \boldsymbol{x}') = k((x_i, \overline{\boldsymbol{x}_i}) \to (x_i', \overline{\boldsymbol{x}_i})) = \rho(i) P(x_i' \mid \overline{\boldsymbol{x}_i})$$

Probability of choosing X_i (out of all the nonevidence variables)

Defining a Transition Kernel

$\overline{\boldsymbol{X}}_i$: variables except X_i and the evidence variables. $\overline{\boldsymbol{x}}_i$: their values.

Define $k(\boldsymbol{x} \to \boldsymbol{x}')$ in the following three cases:



Case 1. The states $\boldsymbol{x}$ and $\boldsymbol{x}'$ differ in ≥ 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 $\boldsymbol{x}' = (x_i', \overline{\boldsymbol{x}}_i)$.

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

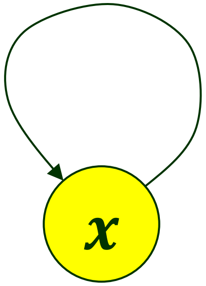
Probability of choosing X_i (out of all the nonevidence variables)

Probability of selecting x_i' (out of all the values of X_i).

$P(x_i' \mid \overline{\boldsymbol{x}}_i) = P(x_i' \mid mb(X_i))$

Completing the Definition

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



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

Correctness of Gibbs Sampling

Theorem 1 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.

Correctness of Gibbs Sampling

Theorem 1 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 kernel k is in detailed balance:

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

Correctness of Gibbs Sampling

Theorem 1 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 kernel k is in detailed balance:

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

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

Correctness of Gibbs Sampling

Theorem 1 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 kernel k is in detailed balance:

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

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

- In both the first case ($\boldsymbol{x}$ and $\boldsymbol{x}'$ differ in the values of ≥ 2 variables such that $k(\boldsymbol{x} \to \boldsymbol{x}') = k(\boldsymbol{x}' \to \boldsymbol{x}) = 0$) and the third case ($\boldsymbol{x} = \boldsymbol{x}'$), detailed balance is trivialy satisfied.

Correctness of Gibbs Sampling

Theorem 1 The previously defined kernel $k(\boldsymbol{x}' \rightarrow \boldsymbol{x})$ 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 kernel k is in detailed balance:

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

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

- In both the first case ($\boldsymbol{x}$ and $\boldsymbol{x}'$ differ in the values of ≥ 2 variables such that $k(\boldsymbol{x} \rightarrow \boldsymbol{x}') = k(\boldsymbol{x}' \rightarrow \boldsymbol{x}) = 0$) and the third case ($\boldsymbol{x} = \boldsymbol{x}'$), detailed balance is trivialy satisfied.

- In the second case, where $\boldsymbol{x}$ and $\boldsymbol{x}'$ differ in the value of one variable X_i

$$\pi(\boldsymbol{x})k(\boldsymbol{x} \rightarrow \boldsymbol{x}') = P(\boldsymbol{x} \mid \boldsymbol{e})\ \rho(i)P(x_i' \mid \overline{\boldsymbol{x}}_i, \boldsymbol{e})$$

Correctness of Gibbs Sampling

Theorem 1 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 kernel k is in detailed balance:

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

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

- In both the first case ($\boldsymbol{x}$ and $\boldsymbol{x}'$ differ in the values of ≥ 2 variables such that $k(\boldsymbol{x} \to \boldsymbol{x}') = k(\boldsymbol{x}' \to \boldsymbol{x}) = 0$) and the third case ($\boldsymbol{x} = \boldsymbol{x}'$), detailed balance is trivialy satisfied.

- In the second case, where $\boldsymbol{x}$ and $\boldsymbol{x}'$ differ in the value of one variable X_i

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

Correctness of Gibbs Sampling

Theorem 1 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 kernel k is in detailed balance:

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

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

- In both the first case ($\boldsymbol{x}$ and $\boldsymbol{x}'$ differ in the values of ≥ 2 variables such that $k(\boldsymbol{x} \to \boldsymbol{x}') = k(\boldsymbol{x}' \to \boldsymbol{x}) = 0$) and the third case ($\boldsymbol{x} = \boldsymbol{x}'$), detailed balance is trivialy satisfied.

- In the second case, where $\boldsymbol{x}$ and $\boldsymbol{x}'$ differ in the value of one variable X_i

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

Correctness of Gibbs Sampling

Theorem 1 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 kernel k is in detailed balance:

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

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

- In both the first case ($\boldsymbol{x}$ and $\boldsymbol{x}'$ differ in the values of ≥ 2 variables such that $k(\boldsymbol{x} \to \boldsymbol{x}') = k(\boldsymbol{x}' \to \boldsymbol{x}) = 0$) and the third case ($\boldsymbol{x} = \boldsymbol{x}'$), detailed balance is trivialy satisfied.
- In the second case, where $\boldsymbol{x}$ and $\boldsymbol{x}'$ differ in the value of one variable X_i

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

Correctness of Gibbs Sampling

Theorem 1 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 kernel k is in detailed balance:

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

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

- In both the first case ($\boldsymbol{x}$ and $\boldsymbol{x}'$ differ in the values of ≥ 2 variables such that $k(\boldsymbol{x} \to \boldsymbol{x}') = k(\boldsymbol{x}' \to \boldsymbol{x}) = 0$) and the third case ($\boldsymbol{x} = \boldsymbol{x}'$), detailed balance is trivialy satisfied.

- In the second case, where $\boldsymbol{x}$ and $\boldsymbol{x}'$ differ in the value of one variable X_i

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

Correctness of Gibbs Sampling

Theorem 1 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 kernel k is in detailed balance:

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

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

- In both the first case ($\boldsymbol{x}$ and $\boldsymbol{x}'$ differ in the values of ≥ 2 variables such that $k(\boldsymbol{x} \to \boldsymbol{x}') = k(\boldsymbol{x}' \to \boldsymbol{x}) = 0$) and the third case ($\boldsymbol{x} = \boldsymbol{x}'$), detailed balance is trivialy satisfied.
- In the second case, where $\boldsymbol{x}$ and $\boldsymbol{x}'$ differ in the value of one variable X_i

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

Correctness of Gibbs Sampling

Theorem 1 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 kernel k is in detailed balance:

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

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

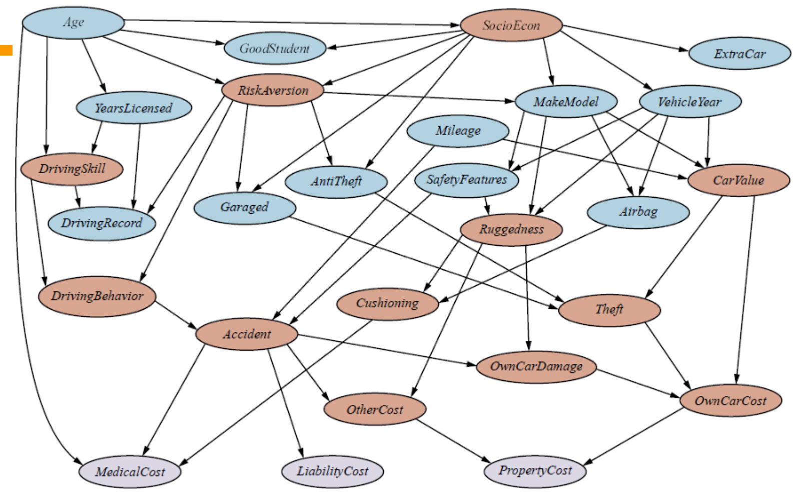
- In both the first case ($\boldsymbol{x}$ and $\boldsymbol{x}'$ differ in the values of ≥ 2 variables such that $k(\boldsymbol{x} \to \boldsymbol{x}') = k(\boldsymbol{x}' \to \boldsymbol{x}) = 0$) and the third case ($\boldsymbol{x} = \boldsymbol{x}'$), detailed balance is trivialy satisfied.

- In the second case, where $\boldsymbol{x}$ and $\boldsymbol{x}'$ differ in the value of one variable X_i

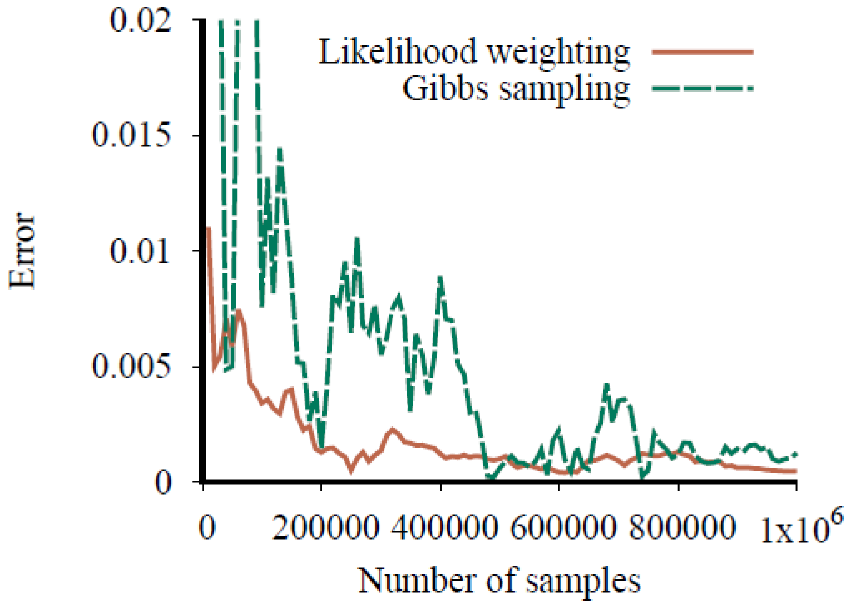
$$\begin{aligned}
\pi(\boldsymbol{x})k(\boldsymbol{x} \to \boldsymbol{x}') &= P(\boldsymbol{x} \mid \boldsymbol{e})\ \rho(i)P(x_i' \mid \bar{\boldsymbol{x}}_i, \boldsymbol{e}) = \rho(i)\ P(x_i, \bar{\boldsymbol{x}}_i \mid \boldsymbol{e})P(x_i' \mid \bar{\boldsymbol{x}}_i, \boldsymbol{e}) \\
&= \rho(i)P(x_i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

Performance of Gibbs Sampling

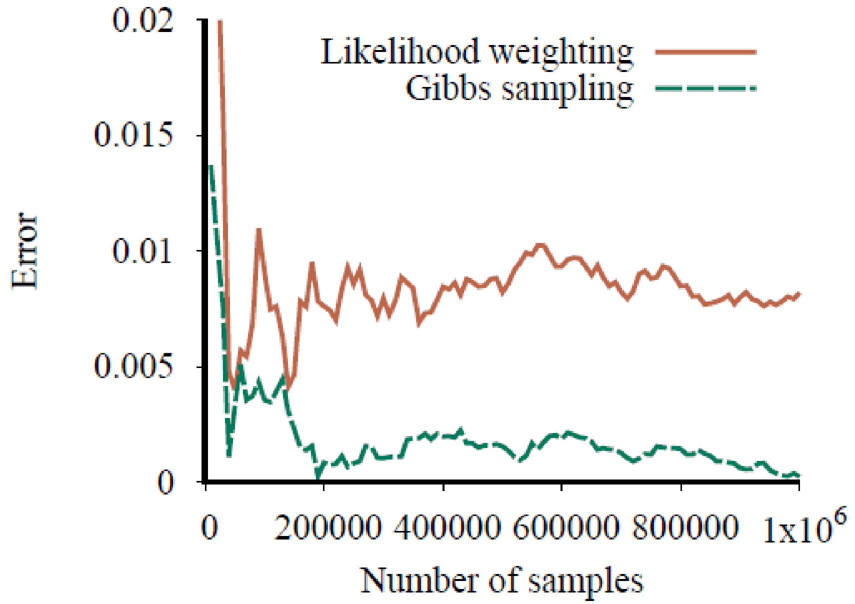
Gibbs sampling is expected to outperform likelihood weighting when evidence is downstream.



On the car insurance network:



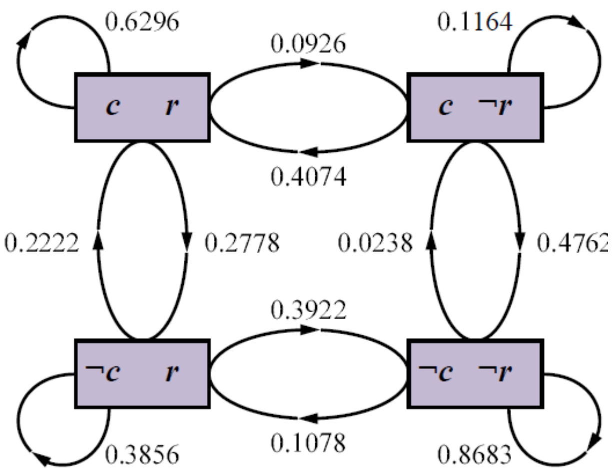
Query on *PropertyCost*



Query on *Age* (with output observed)

Quick Summary

Query $\mathbf{P}(\mathbf{Y} \mid \mathbf{e})$



MC for $\mathbf{P}(Rain \mid Sprinkler = true, WetGrass = true)$

$\mathbf{Y}$: query variables (≥ 1)

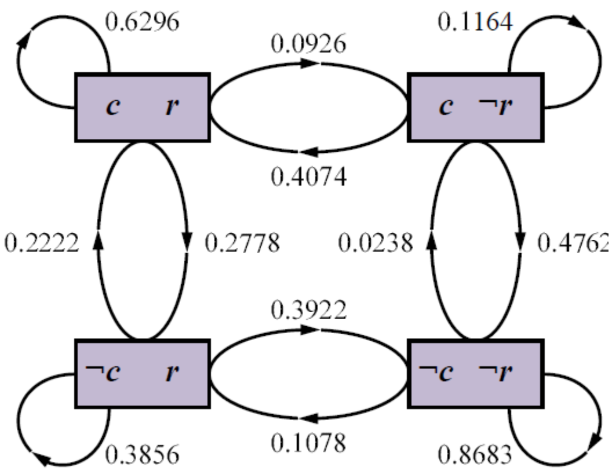
$\mathbf{X}$: nonevidence variables ($\mathbf{Y} \subseteq \mathbf{X}$)

$\mathbf{E}$: evidence variables

$\mathbf{e}$: values of $\mathbf{E}$

Quick Summary

Query $\boldsymbol{P}(\boldsymbol{Y} \mid \boldsymbol{e})$



MC for $\boldsymbol{P}(Rain \mid Sprinkler = true, WetGrass = true)$

$\boldsymbol{Y}$: query variables (≥ 1)

$\boldsymbol{E}$: evidence variables

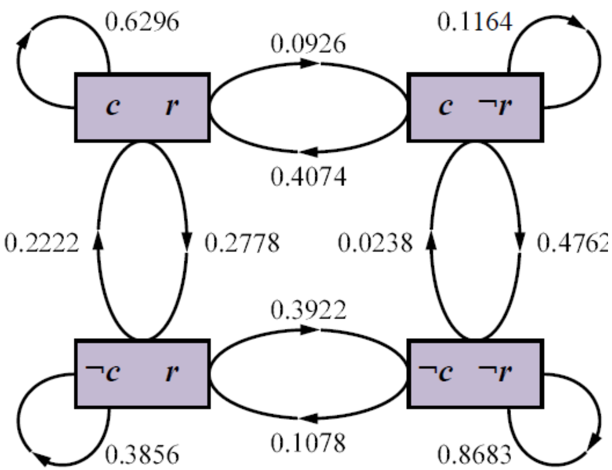
$\boldsymbol{X}$: nonevidence variables ($\boldsymbol{Y} \subseteq \boldsymbol{X}$)

$\boldsymbol{e}$: values of $\boldsymbol{E}$

- The topology of the Markov Chain (as a graph not including edge weights) is determined by $\boldsymbol{X}$, or equivalently, by $\boldsymbol{E}$.

Quick Summary

Query $\boldsymbol{P}(\boldsymbol{Y} \mid \boldsymbol{e})$



MC for $\boldsymbol{P}(Rain \mid Sprinkler = true, WetGrass = true)$

$\boldsymbol{Y}$: query variables (≥ 1)

$\boldsymbol{X}$: nonevidence variables ($\boldsymbol{Y} \subseteq \boldsymbol{X}$)

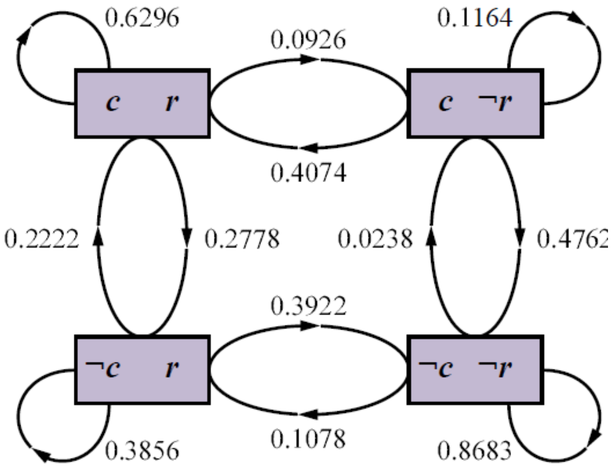
$\boldsymbol{E}$: evidence variables

$\boldsymbol{e}$: values of $\boldsymbol{E}$

- The topology of the Markov Chain (as a graph not including edge weights) is determined by $\boldsymbol{X}$, or equivalently, by $\boldsymbol{E}$.
  - Every vertex corresponds to a different value $\boldsymbol{x}$ of $\boldsymbol{X}$, namely, a different combination of values of the variables in $\boldsymbol{X}$.

Quick Summary

Query $\boldsymbol{P}(\boldsymbol{Y} \mid \boldsymbol{e})$



MC for $\boldsymbol{P}(Rain \mid Sprinkler = true, WetGrass = true)$

$\boldsymbol{Y}$: query variables (≥ 1)

$\boldsymbol{X}$: nonevidence variables ($\boldsymbol{Y} \subseteq \boldsymbol{X}$)

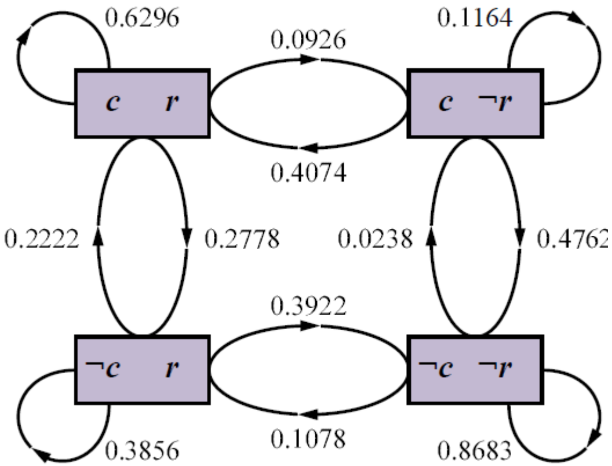
$\boldsymbol{E}$: evidence variables

$\boldsymbol{e}$: values of $\boldsymbol{E}$

- ♦ The topology of the Markov Chain (as a graph not including edge weights) is determined by $\boldsymbol{X}$, or equivalently, by $\boldsymbol{E}$.
  - Every vertex corresponds to a different value $\boldsymbol{x}$ of $\boldsymbol{X}$, namely, a different combination of values of the variables in $\boldsymbol{X}$.
  - Every two vertices sharing an edge differ in the value of *exactly one* variable from $\boldsymbol{X}$.

Quick Summary

Query $\boldsymbol{P}(\boldsymbol{Y} \mid \boldsymbol{e})$



MC for $\boldsymbol{P}(Rain \mid Sprinkler = true, WetGrass = true)$

$\boldsymbol{Y}$: query variables (≥ 1)

$\boldsymbol{E}$: evidence variables

$\boldsymbol{X}$: nonevidence variables ($\boldsymbol{Y} \subseteq \boldsymbol{X}$)

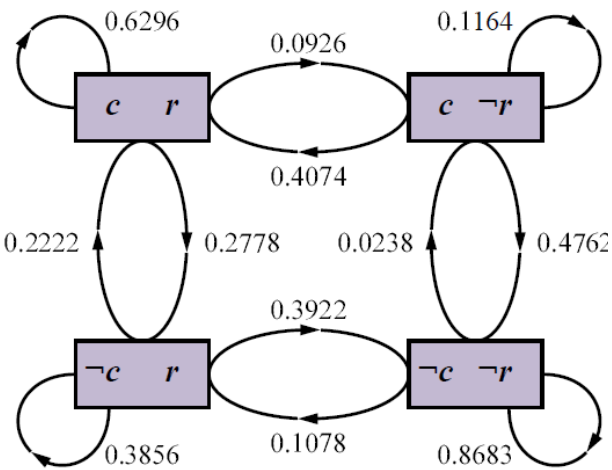
$\boldsymbol{e}$: values of $\boldsymbol{E}$

- The topology of the Markov Chain (as a graph not including edge weights) is determined by $\boldsymbol{X}$, or equivalently, by $\boldsymbol{E}$.
  - Every vertex corresponds to a different value $\boldsymbol{x}$ of $\boldsymbol{X}$, namely, a different combination of values of the variables in $\boldsymbol{X}$.
  - Every two vertices sharing an edge differ in the value of *exactly one* variable from $\boldsymbol{X}$.

- The weight assigned to an edge between two different vertices, which changes the value of one variable $X_i \in \boldsymbol{X}$ to a different value x_i', is the probability $\rho(i)P\left(x_i' \mid mb(X_i)\right)$.

Quick Summary

Query $\boldsymbol{P}(\boldsymbol{Y} \mid \boldsymbol{e})$



MC for $\boldsymbol{P}(Rain \mid Sprinkler = true, WetGrass = true)$

$\boldsymbol{Y}$: query variables (≥ 1) $\boldsymbol{E}$: evidence variables

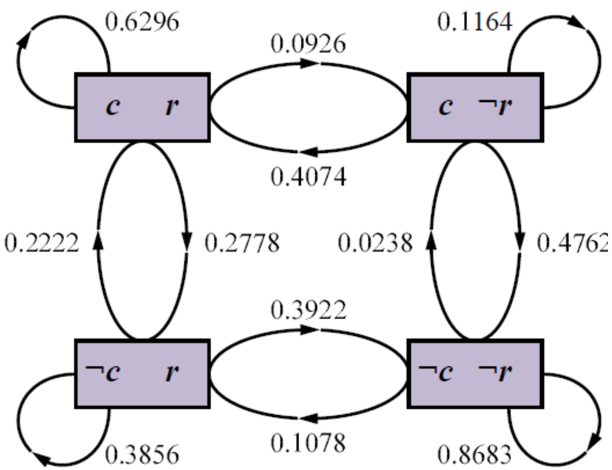
$\boldsymbol{X}$: nonevidence variables ($\boldsymbol{Y} \subseteq \boldsymbol{X}$) $\boldsymbol{e}$: values of $\boldsymbol{E}$

- The topology of the Markov Chain (as a graph not including edge weights) is determined by $\boldsymbol{X}$, or equivalently, by $\boldsymbol{E}$.
  - Every vertex corresponds to a different value $\boldsymbol{x}$ of $\boldsymbol{X}$, namely, a different combination of values of the variables in $\boldsymbol{X}$.
  - Every two vertices sharing an edge differ in the value of *exactly one* variable from $\boldsymbol{X}$.

- The weight assigned to an edge between two different vertices, which changes the value of one variable $X_i \in \boldsymbol{X}$ to a different value x_i', is the probability $\rho(i)P\left(x_i' \mid mb(X_i)\right)$.
- The weight assigned to a loop at a vertex $\boldsymbol{x}$ is the probability $\sum_{X_j \in \boldsymbol{X}} \rho(j)P\left(x_j \mid mb(X_j)\right)$, where x_j is the current value of X_j, for all $X_j \in \boldsymbol{X}$.

Quick Summary

Query $\boldsymbol{P}(\boldsymbol{Y} \mid \boldsymbol{e})$



MC for $\boldsymbol{P}(Rain \mid Sprinkler = true, WetGrass = true)$

$\boldsymbol{Y}$: query variables (≥ 1)

$\boldsymbol{E}$: evidence variables

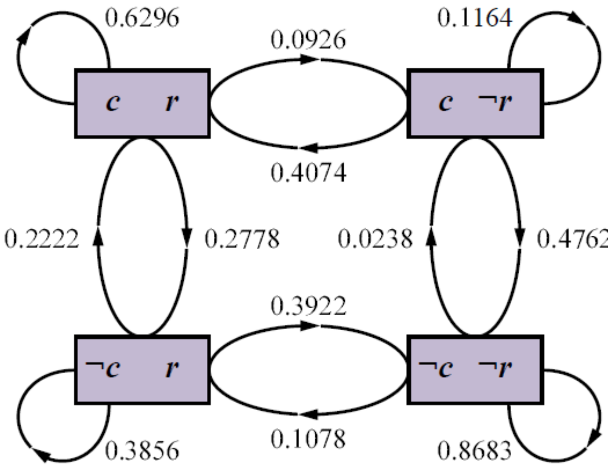
$\boldsymbol{X}$: nonevidence variables ($\boldsymbol{Y} \subseteq \boldsymbol{X}$)

$\boldsymbol{e}$: values of $\boldsymbol{E}$

- The topology of the Markov Chain (as a graph not including edge weights) is determined by $\boldsymbol{X}$, or equivalently, by $\boldsymbol{E}$.
  - Every vertex corresponds to a different value $\boldsymbol{x}$ of $\boldsymbol{X}$, namely, a different combination of values of the variables in $\boldsymbol{X}$.
  - Every two vertices sharing an edge differ in the value of *exactly one* variable from $\boldsymbol{X}$.
- The weight assigned to an edge between two different vertices, which changes the value of one variable $X_i \in \boldsymbol{X}$ to a different value x_i', is the probability $\rho(i)P\left(x_i' \mid mb(X_i)\right)$.
- The weight assigned to a loop at a vertex $\boldsymbol{x}$ is the probability $\sum_{X_j \in \boldsymbol{X}} \rho(j)P\left(x_j \mid mb(X_j)\right)$, where x_j is the current value of X_j, for all $X_j \in \boldsymbol{X}$.
- The kernel function k assigning weights as above will reach a unique stationary distribution π such that $\pi(\boldsymbol{x}) = P(\boldsymbol{x} \mid \boldsymbol{e})$, for every state $\boldsymbol{x}$ of the Markov Chain.

Quick Summary

Query $\boldsymbol{P}(\boldsymbol{Y} \mid \boldsymbol{e})$



MC for $\boldsymbol{P}(\mathit{Rain} \mid \mathit{Sprinkler} = \mathit{true}, \mathit{WetGrass} = \mathit{true})$

$\boldsymbol{Y}$: query variables (≥ 1)

$\boldsymbol{X}$: nonevidence variables ($\boldsymbol{Y} \subseteq \boldsymbol{X}$)

$\boldsymbol{E}$: evidence variables

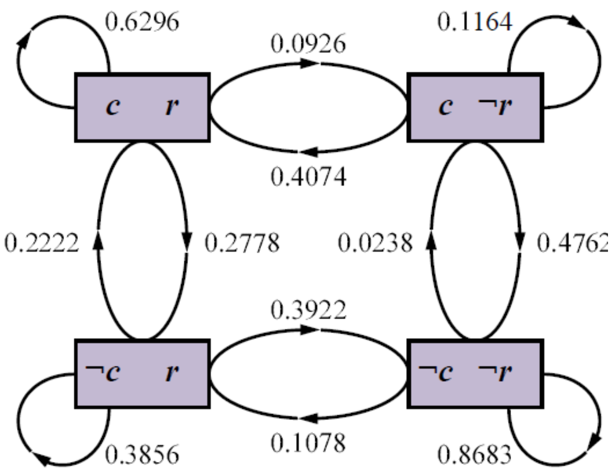
$\boldsymbol{e}$: values of $\boldsymbol{E}$

- The topology of the Markov Chain (as a graph not including edge weights) is determined by $\boldsymbol{X}$, or equivalently, by $\boldsymbol{E}$.
  - Every vertex corresponds to a different value $\boldsymbol{x}$ of $\boldsymbol{X}$, namely, a different combination of values of the variables in $\boldsymbol{X}$.
  - Every two vertices sharing an edge differ in the value of *exactly one* variable from $\boldsymbol{X}$.
- The weight assigned to an edge between two different vertices, which changes the value of one variable $X_i \in \boldsymbol{X}$ to a different value x_i', is the probability $\rho(i)P\big(x_i' \mid mb(X_i)\big)$.
- The weight assigned to a loop at a vertex $\boldsymbol{x}$ is the probability $\sum_{X_j \in \boldsymbol{X}} \rho(j)P\left(x_j \mid mb(X_j)\right)$, where x_j is the current value of X_j, for all $X_j \in \boldsymbol{X}$.
- The kernel function k assigning weights as above will reach a unique stationary distribution π such that $\pi(\boldsymbol{x}) = P(\boldsymbol{x} \mid \boldsymbol{e})$, for every state $\boldsymbol{x}$ of the Markov Chain.

Answer a query: $P(\boldsymbol{y} \mid \boldsymbol{e})$ for every value $\boldsymbol{y}$ of $\boldsymbol{Y}$.

Quick Summary

Query $\boldsymbol{P}(\boldsymbol{Y} \mid \boldsymbol{e})$



MC for $\boldsymbol{P}(Rain \mid Sprinkler = true, WetGrass = true)$

$\boldsymbol{Y}$: query variables (≥ 1)

$\boldsymbol{E}$: evidence variables

$\boldsymbol{X}$: nonevidence variables ($\boldsymbol{Y} \subseteq \boldsymbol{X}$)

$\boldsymbol{e}$: values of $\boldsymbol{E}$

- The topology of the Markov Chain (as a graph not including edge weights) is determined by $\boldsymbol{X}$, or equivalently, by $\boldsymbol{E}$.
  - Every vertex corresponds to a different value $\boldsymbol{x}$ of $\boldsymbol{X}$, namely, a different combination of values of the variables in $\boldsymbol{X}$.
  - Every two vertices sharing an edge differ in the value of *exactly one* variable from $\boldsymbol{X}$.
- The weight assigned to an edge between two different vertices, which changes the value of one variable $X_i \in \boldsymbol{X}$ to a different value x_i', is the probability $\rho(i)P\left(x_i' \mid mb(X_i)\right)$.
- The weight assigned to a loop at a vertex $\boldsymbol{x}$ is the probability $\sum_{X_j \in \boldsymbol{X}} \rho(j)P\left(x_j \mid mb(X_j)\right)$, where x_j is the current value of X_j, for all $X_j \in \boldsymbol{X}$.
- The kernel function k assigning weights as above will reach a unique stationary distribution π such that $\pi(\boldsymbol{x}) = P(\boldsymbol{x} \mid \boldsymbol{e})$, for every state $\boldsymbol{x}$ of the Markov Chain.

Answer a query: $P(\boldsymbol{y} \mid \boldsymbol{e}) = \sum_{\boldsymbol{x} \models \boldsymbol{y}} \pi(\boldsymbol{x})$ for every value $\boldsymbol{y}$ of $\boldsymbol{Y}$.

II. Metropolis-Hastings (MH) Sampling

- The most broadly applicable Markov chain Monte Carlo algorithm.

- MH generates samples $\boldsymbol{x}$ (eventually) according to a target probability distribution $\pi(\boldsymbol{x})$ (in a BN, $\pi(\boldsymbol{x}) = P(\boldsymbol{x} \mid \boldsymbol{e})$).

- MH combines the efficiency of Markov chain sampling with the global reach of direct sampling.

- MH is a general sampling algorithm not just limited to Bayes nets.

Transition Kernel for MH

Define the transition kernel $k(\boldsymbol{x} \rightarrow \boldsymbol{x}')$:

Transition Kernel for MH

Define the transition kernel $k(\boldsymbol{x} \to \boldsymbol{x}')$:

- At the current state $\boldsymbol{x}$, sample a new state $\boldsymbol{x}'$ from a *proposal distribution* $q(\boldsymbol{x}' \mid \boldsymbol{x})$. // $\boldsymbol{x}'$ can be any state in the Markov chain.

Transition Kernel for MH

Define the transition kernel $k(\boldsymbol{x} \to \boldsymbol{x}')$:

- At the current state $\boldsymbol{x}$, sample a new state $\boldsymbol{x}'$ from a *proposal distribution* $q(\boldsymbol{x}' \mid \boldsymbol{x})$. // $\boldsymbol{x}'$ can be any state in the Markov chain.

- Accept $\boldsymbol{x}'$ according to the *acceptance probability*:

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

Transition Kernel for MH

Define the transition kernel $k(\boldsymbol{x} \to \boldsymbol{x}')$:

- At the current state $\boldsymbol{x}$, sample a new state $\boldsymbol{x}'$ from a *proposal distribution* $q(\boldsymbol{x}' \mid \boldsymbol{x})$. // $\boldsymbol{x}'$ can be any state in the Markov chain.

- Accept $\boldsymbol{x}'$ according to the *acceptance probability*:

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

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

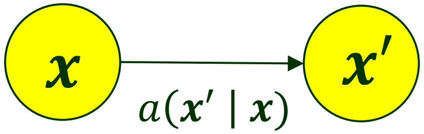
Transition Kernel for MH

Define the transition kernel $k(\boldsymbol{x} \to \boldsymbol{x}')$:

- At the current state $\boldsymbol{x}$, sample a new state $\boldsymbol{x}'$ from a *proposal distribution* $q(\boldsymbol{x}' \mid \boldsymbol{x})$. // $\boldsymbol{x}'$ can be any state in the Markov chain.

- Accept $\boldsymbol{x}'$ according to the *acceptance probability*:

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



State transitions with probabilities after selection of $\boldsymbol{x}'$

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

Transition Kernel for MH

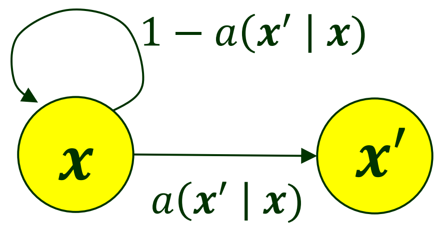
Define the transition kernel $k(\boldsymbol{x} \to \boldsymbol{x}')$:

- At the current state $\boldsymbol{x}$, sample a new state $\boldsymbol{x}'$ from a *proposal distribution* $q(\boldsymbol{x}' \mid \boldsymbol{x})$. // $\boldsymbol{x}'$ can be any state in the Markov chain.

- Accept $\boldsymbol{x}'$ according to the *acceptance probability*:

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

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



State transitions with probabilities after selection of $\boldsymbol{x}'$

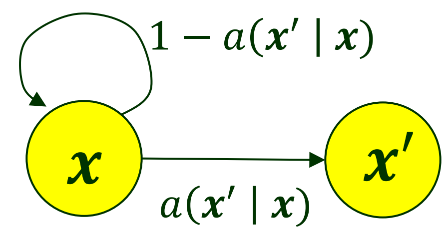
Transition Kernel for MH

Define the transition kernel $k(\boldsymbol{x} \to \boldsymbol{x}')$:

- At the current state $\boldsymbol{x}$, sample a new state $\boldsymbol{x}'$ from a *proposal distribution* $q(\boldsymbol{x}' \mid \boldsymbol{x})$. // $\boldsymbol{x}'$ can be any state in the Markov chain.

- Accept $\boldsymbol{x}'$ according to the *acceptance probability*:

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



State transitions with probabilities after selection of $\boldsymbol{x}'$

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

Probability of a transition to $\boldsymbol{x}'$: $k(\boldsymbol{x} \to \boldsymbol{x}') = q(\boldsymbol{x}' \mid \boldsymbol{x})a(\boldsymbol{x}' \mid \boldsymbol{x})$

Proposal Distribution

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 (importance sampling) to generate $\boldsymbol{x}'$.

Proposal Distribution

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 (importance sampling) 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.

Proposal Distribution

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

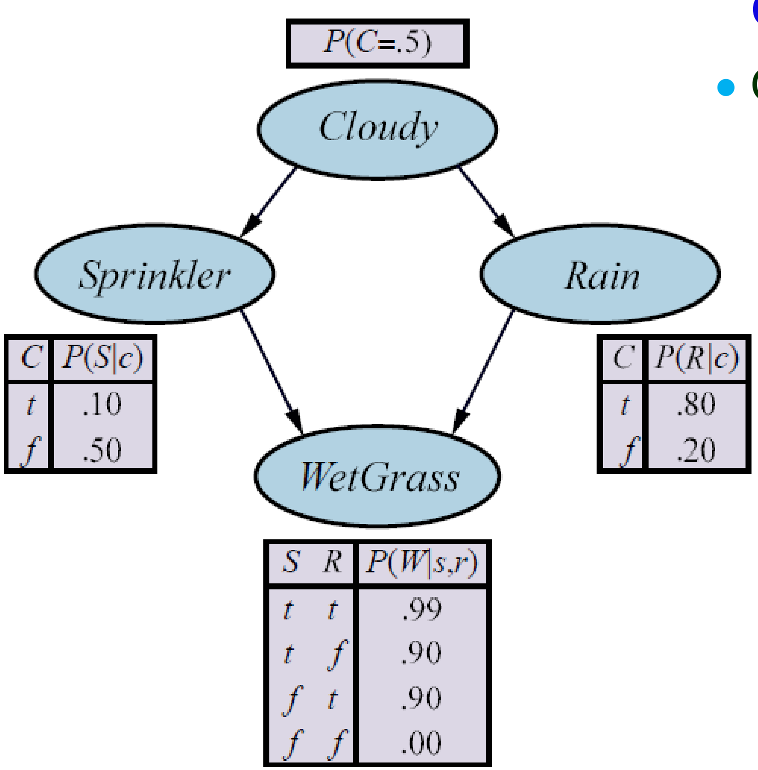
Example $q(\boldsymbol{x}' \mid \boldsymbol{x})$ could be defined as follows:

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

Importance Sampling (reviewed)

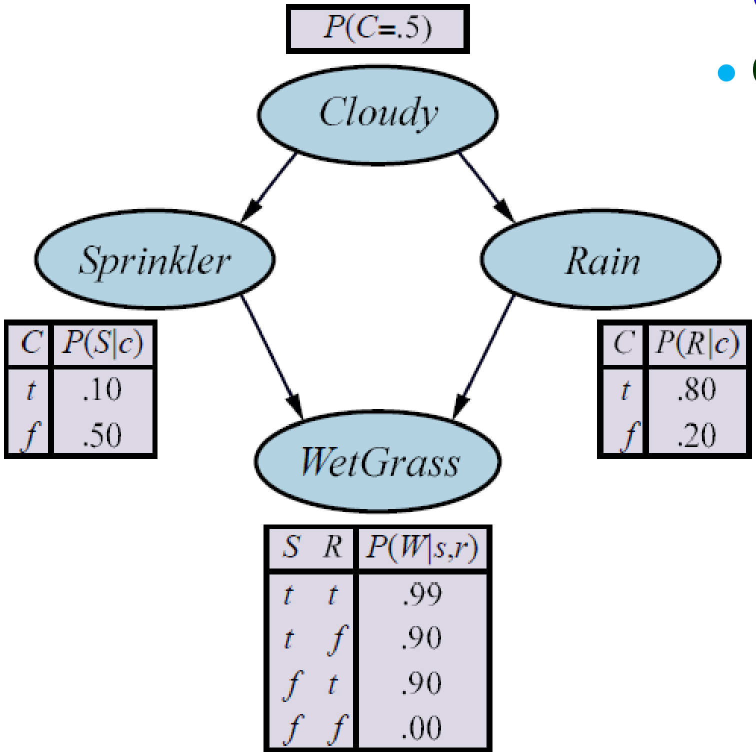


Query $\mathbf{P}(Rain \mid Cloudy = true, WetGrass = true)$

- Generate an event in the chosen topological order.

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

Importance Sampling (reviewed)



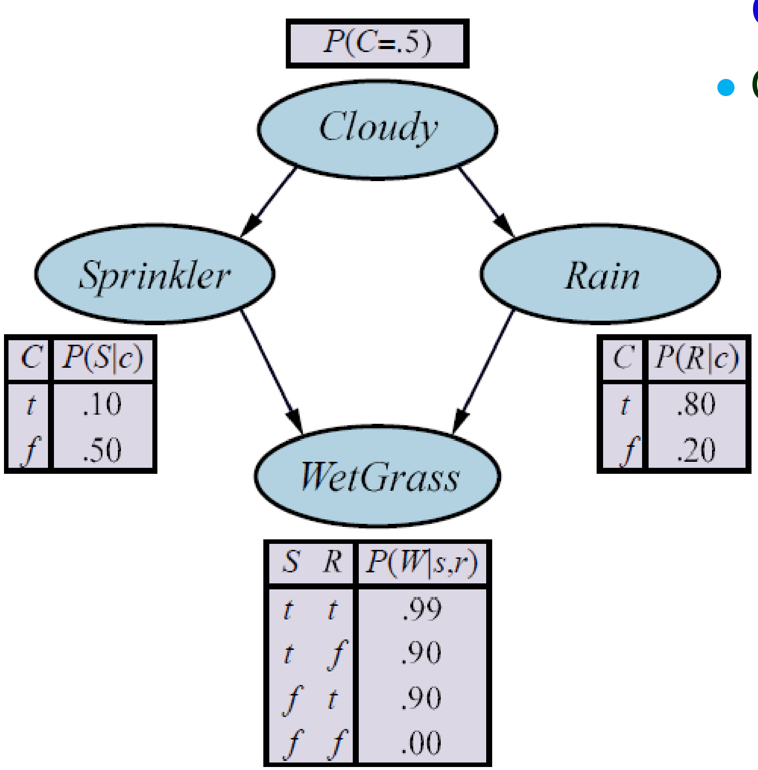
Order: *Cloudy*, *Sprinkler*, *Rain*, *WetGrass*

Query $\mathbf{P}(Rain \mid Cloudy = true, WetGrass = true)$

- Generate an event in the chosen topological order.
  1. *Cloudy* is an evidence variable with value *true*.

$$w \leftarrow w \times P(Cloudy = true) = 0.5$$

Importance Sampling (reviewed)



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

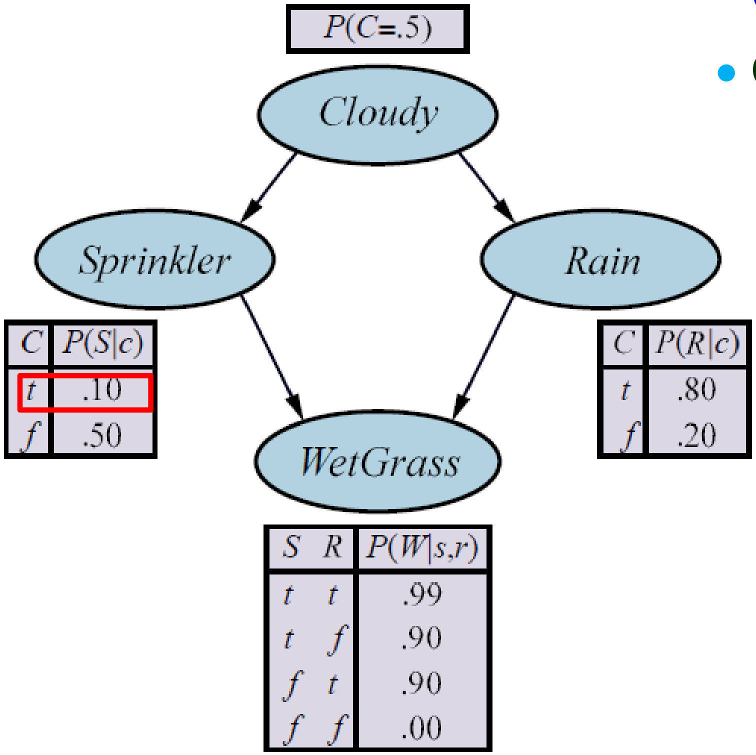
Query $\mathbf{P}(Rain \mid Cloudy = true, WetGrass = true)$

- Generate an event in the chosen topological order.
  1. *Cloudy* is an evidence variable with value *true*.

     $$w \leftarrow w \times P(Cloudy = true) = 0.5$$

  2. *Sprinkler* is not an evidence variable. Sample from $\mathbf{P}(Sprinkler \mid Cloudy = true) = \langle 0.1, 0.9 \rangle$.

Importance Sampling (reviewed)



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

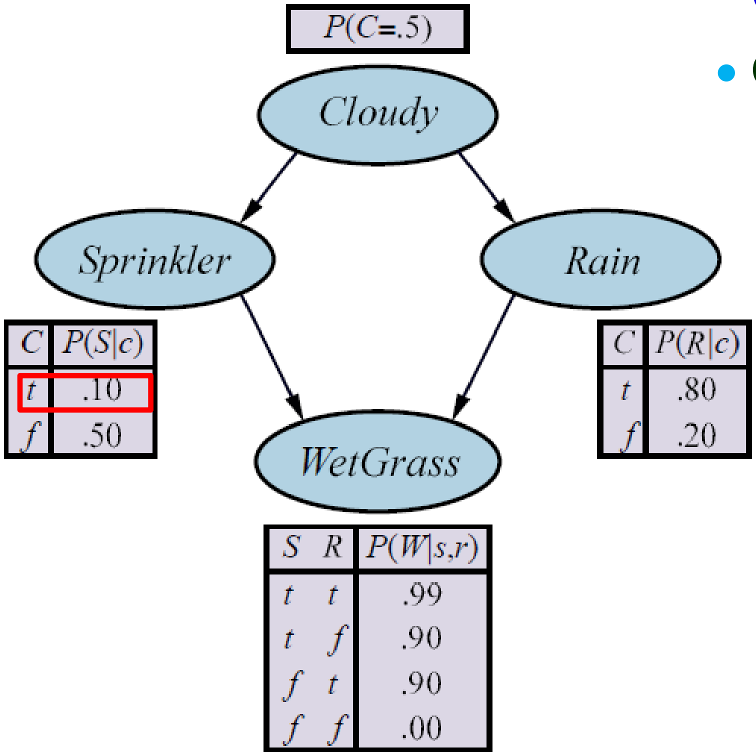
Query $\mathbf{P}(Rain \mid Cloudy = true, WetGrass = true)$

- Generate an event in the chosen topological order.
  1. *Cloudy* is an evidence variable with value *true*.

     $$w \leftarrow w \times P(Cloudy = true) = 0.5$$

  2. *Sprinkler* is not an evidence variable. Sample from $\mathbf{P}(Sprinkler \mid Cloudy = true) = \langle 0.1, 0.9 \rangle$.

Importance Sampling (reviewed)



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

Query $\mathbf{P}(Rain \mid Cloudy = true, WetGrass = true)$

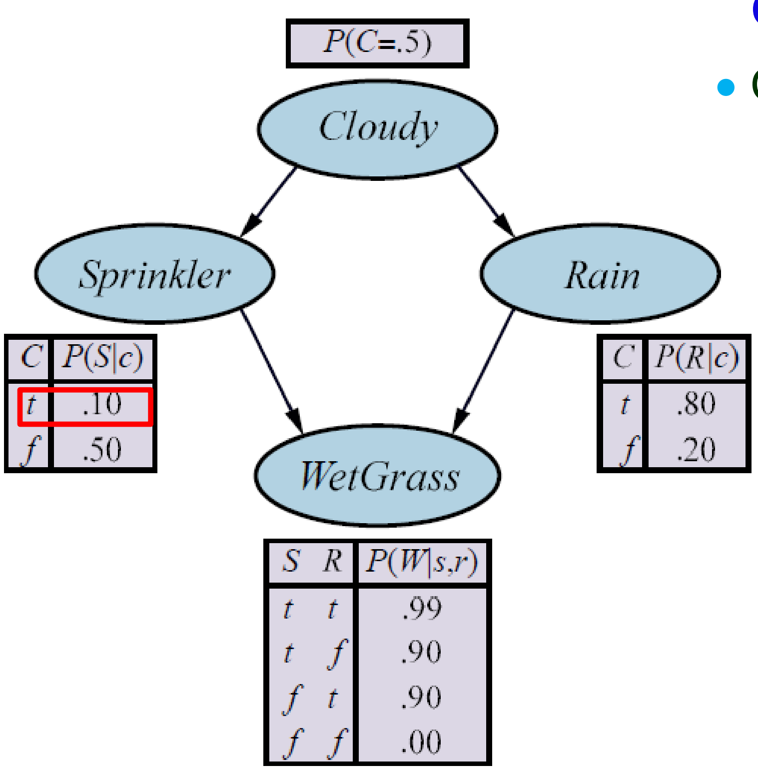
- Generate an event in the chosen topological order.
  1. *Cloudy* is an evidence variable with value *true*.

     $$w \leftarrow w \times P(Cloudy = true) = 0.5$$

  2. *Sprinkler* is not an evidence variable. Sample from $\mathbf{P}(Sprinkler \mid Cloudy = true) = \langle 0.1, 0.9 \rangle$.

     $Sprinkler = false$ (returned value)

Importance Sampling (reviewed)



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

Query $\mathbf{P}(Rain \mid Cloudy = true, WetGrass = true)$

- Generate an event in the chosen topological order.
  1. *Cloudy* is an evidence variable with value *true*.

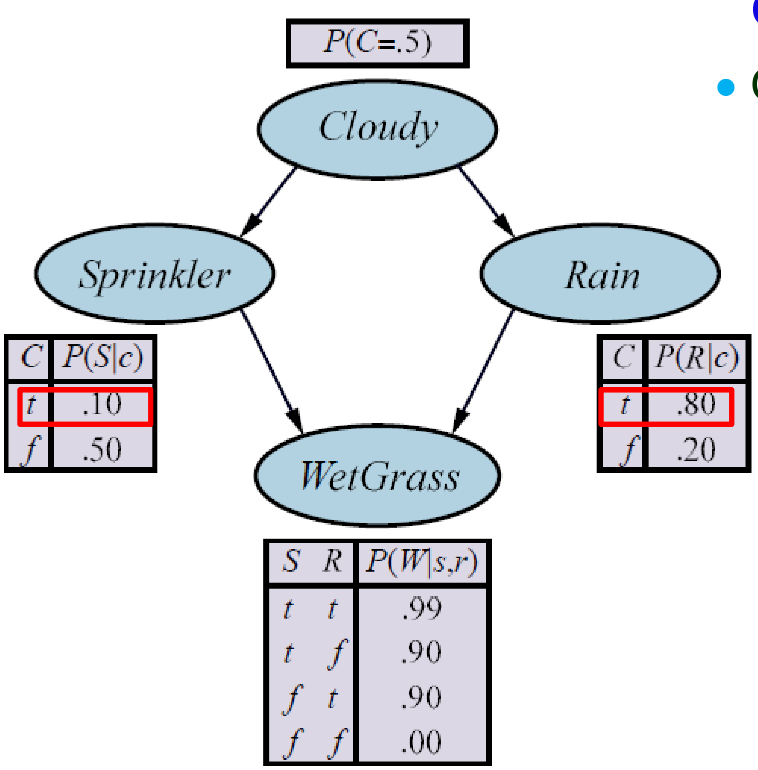
     $$w \leftarrow w \times P(Cloudy = true) = 0.5$$

  2. *Sprinkler* is not an evidence variable. Sample from $\mathbf{P}(Sprinkler \mid Cloudy = true) = \langle 0.1, 0.9 \rangle$.

     Sprinkler = *false* (returned value)

  3. *Rain* is not an evidence variable. Sample from $\mathbf{P}(Rain \mid Cloudy = true) = \langle 0.8, 0.2 \rangle$.

Importance Sampling (reviewed)



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

Query $\mathbf{P}(Rain \mid Cloudy = true, WetGrass = true)$

- Generate an event in the chosen topological order.
  1. *Cloudy* is an evidence variable with value *true*.

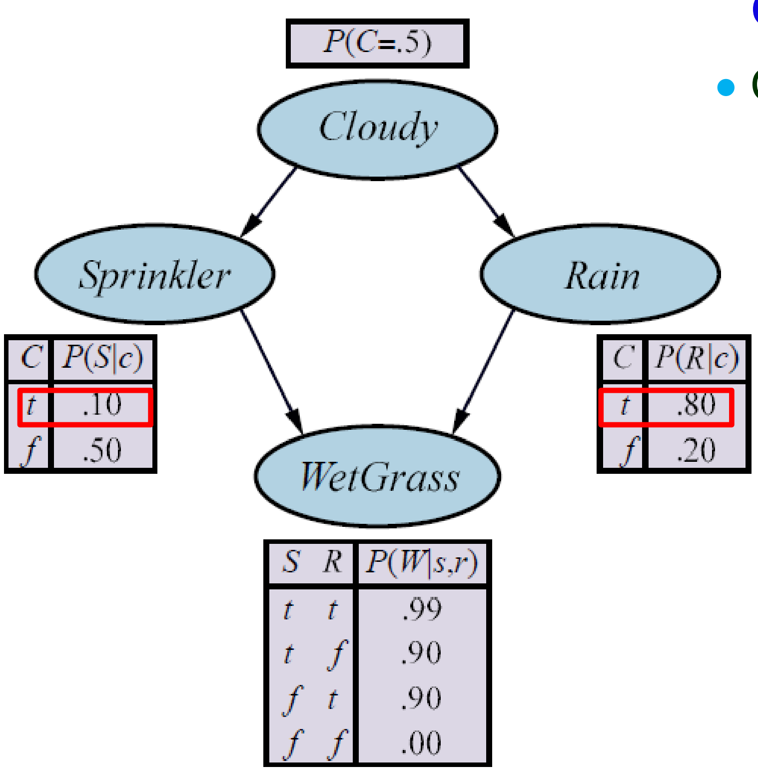
     $$w \leftarrow w \times P(Cloudy = true) = 0.5$$

  2. *Sprinkler* is not an evidence variable. Sample from $\mathbf{P}(Sprinkler \mid Cloudy = true) = \langle 0.1, 0.9 \rangle$.

     $Sprinkler = false$ (returned value)

  3. *Rain* is not an evidence variable. Sample from $\mathbf{P}(Rain \mid Cloudy = true) = \langle 0.8, 0.2 \rangle$.

Importance Sampling (reviewed)



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

Query $\mathbf{P}(\mathit{Rain} \mid \mathit{Cloudy} = \mathit{true}, \mathit{WetGrass} = \mathit{true})$

- Generate an event in the chosen topological order.
  1. *Cloudy* is an evidence variable with value *true*.

     $$w \leftarrow w \times P(\mathit{Cloudy} = \mathit{true}) = 0.5$$

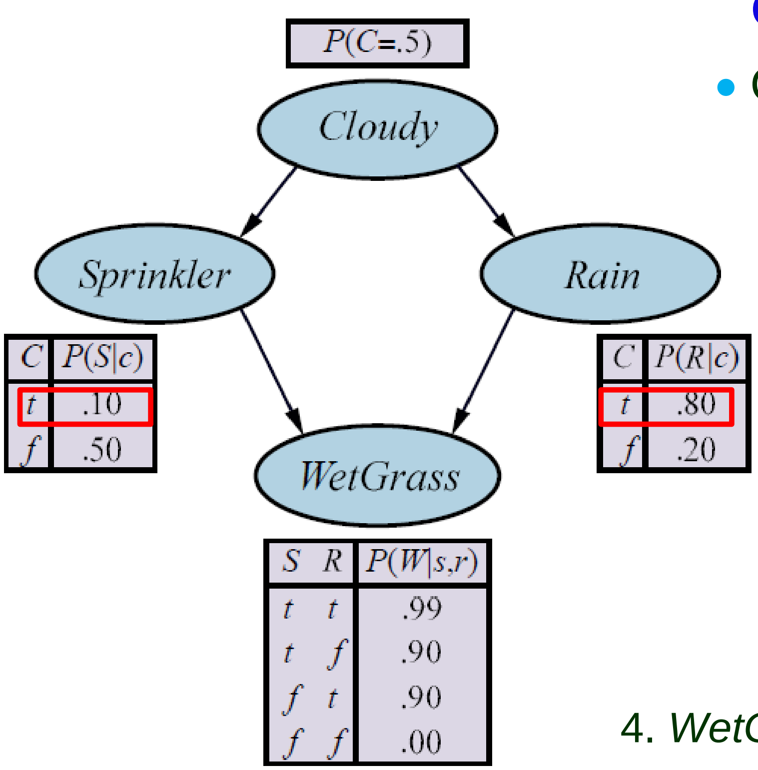
  2. *Sprinkler* is not an evidence variable. Sample from $\mathbf{P}(\mathit{Sprinkler} \mid \mathit{Cloudy} = \mathit{true}) = \langle 0.1, 0.9 \rangle$.

     $\mathit{Sprinkler} = \mathit{false}$ (returned value)

  3. *Rain* is not an evidence variable. Sample from $\mathbf{P}(\mathit{Rain} \mid \mathit{Cloudy} = \mathit{true}) = \langle 0.8, 0.2 \rangle$.

     $\mathit{Rain} = \mathit{true}$

Importance Sampling (reviewed)



Query $\mathbf{P}(Rain \mid Cloudy = true, WetGrass = true)$

- Generate an event in the chosen topological order.

  1. *Cloudy* is an evidence variable with value *true*.

  $$w \leftarrow w \times P(Cloudy = true) = 0.5$$

  2. *Sprinkler* is not an evidence variable. Sample from $\mathbf{P}(Sprinkler \mid Cloudy = true) = \langle 0.1, 0.9 \rangle$.

  $Sprinkler = false$ (returned value)

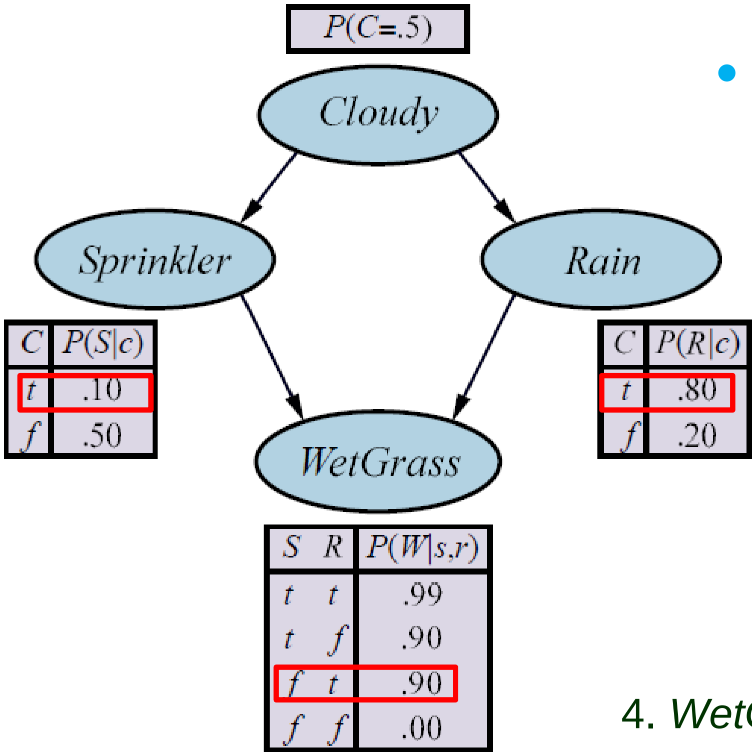
  3. *Rain* is not an evidence variable. Sample from $\mathbf{P}(Rain \mid Cloudy = true) = \langle 0.8, 0.2 \rangle$.

  $Rain = true$

  4. *WetGrass* is an evidence variable with value *true*.

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

Importance Sampling (reviewed)



Query $\mathbf{P}(\textit{Rain} \mid \textit{Cloudy} = \textit{true}, \textit{WetGrass} = \textit{true})$

- Generate an event in the chosen topological order.
  1. *Cloudy* is an evidence variable with value *true*.

     $$w \leftarrow w \times P(\textit{Cloudy} = \textit{true}) = 0.5$$

  2. *Sprinkler* is not an evidence variable. Sample from $\mathbf{P}(\textit{Sprinkler} \mid \textit{Cloudy} = \textit{true}) = \langle 0.1, 0.9 \rangle$.

     $\textit{Sprinkler} = \textit{false}$ (returned value)

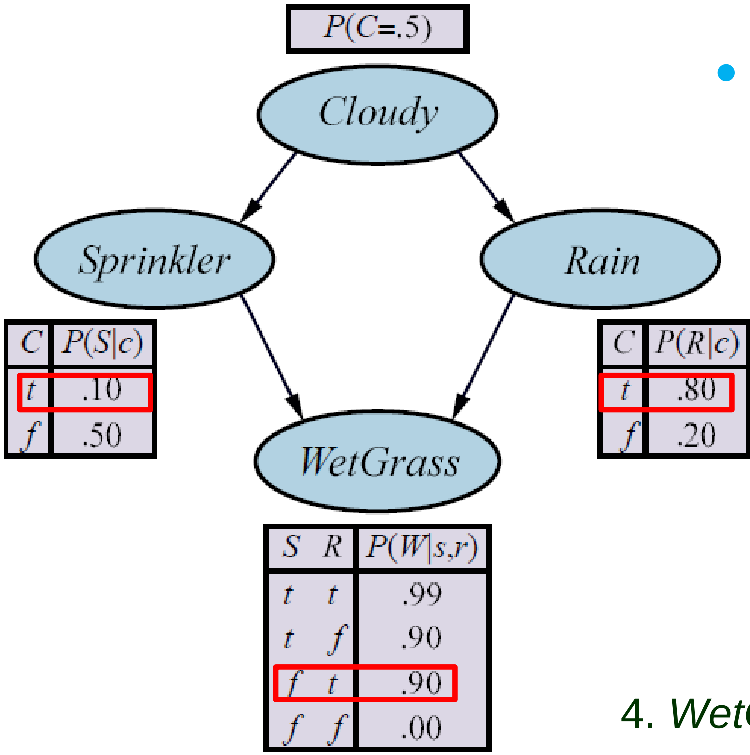
  3. *Rain* is not an evidence variable. Sample from $\mathbf{P}(\textit{Rain} \mid \textit{Cloudy} = \textit{true}) = \langle 0.8, 0.2 \rangle$.

     $\textit{Rain} = \textit{true}$

  4. *WetGrass* is an evidence variable with value *true*.

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

Importance Sampling (reviewed)



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

Query $\mathbf{P}(Rain \mid Cloudy = true, WetGrass = true)$

- Generate an event in the chosen topological order.

1. *Cloudy* is an evidence variable with value *true*.

$$w \leftarrow w \times P(Cloudy = true) = 0.5$$

2. *Sprinkler* is not an evidence variable. Sample from $\mathbf{P}(Sprinkler \mid Cloudy = true) = \langle 0.1, 0.9 \rangle$.

$$Sprinkler = false \quad \text{(returned value)}$$

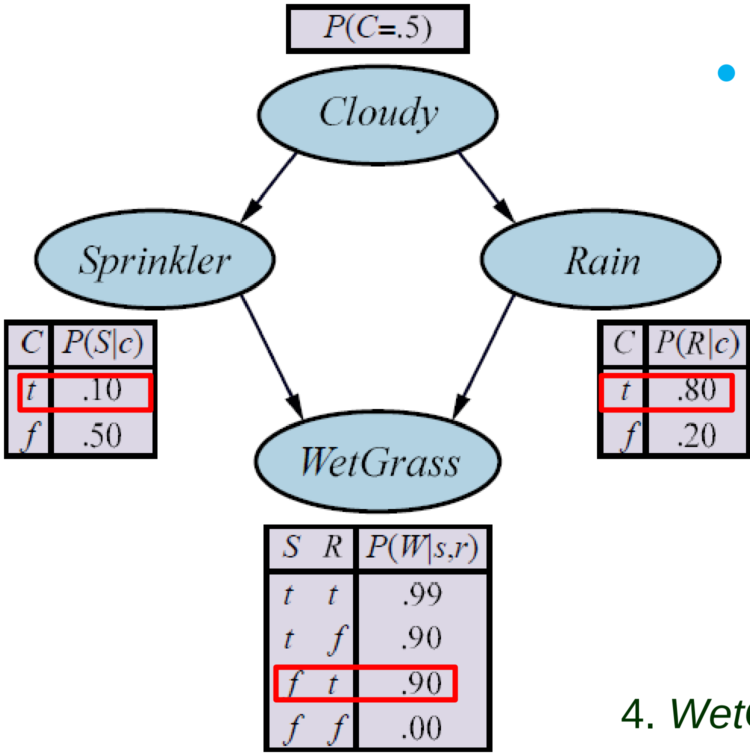
3. *Rain* is not an evidence variable. Sample from $\mathbf{P}(Rain \mid Cloudy = true) = \langle 0.8, 0.2 \rangle$.

$$Rain = true$$

4. *WetGrass* is an evidence variable with value *true*.

$$w \leftarrow w \times P(WetGrass = true \mid Sprinkler = false, Rain = true) = 0.5 \times 0.9 = 0.45$$

Importance Sampling (reviewed)



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

Query $\mathbf{P}(Rain \mid Cloudy = true, WetGrass = true)$

- Generate an event in the chosen topological order.

1. *Cloudy* is an evidence variable with value *true*.

$$w \leftarrow w \times P(Cloudy = true) = 0.5$$

2. *Sprinkler* is not an evidence variable. Sample from $\mathbf{P}(Sprinkler \mid Cloudy = true) = \langle 0.1, 0.9 \rangle$.

$Sprinkler = false$ (returned value)

3. *Rain* is not an evidence variable. Sample from $\mathbf{P}(Rain \mid Cloudy = true) = \langle 0.8, 0.2 \rangle$.

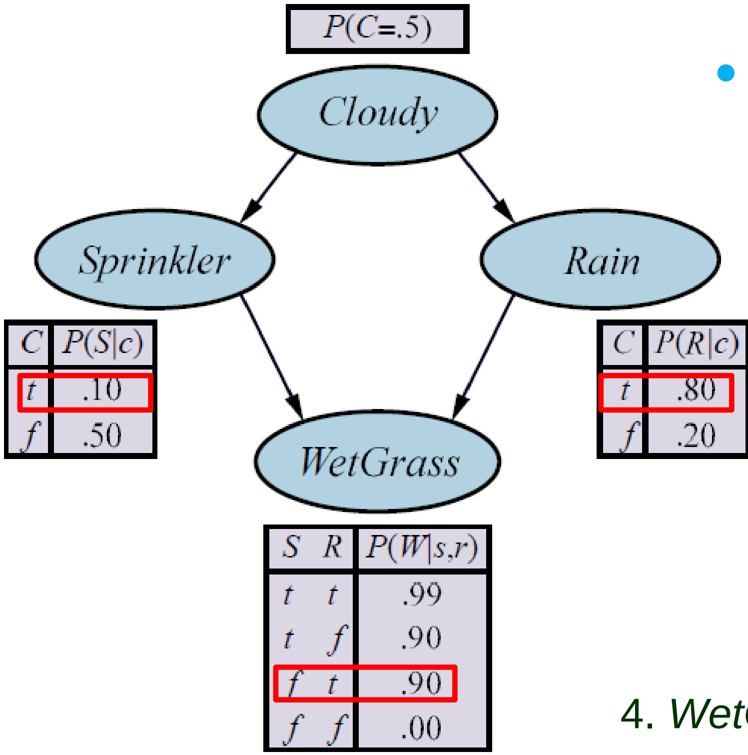
$Rain = true$

4. *WetGrass* is an evidence variable with value *true*.

$$w \leftarrow w \times P(WetGrass = true \mid Sprinkler = false, Rain = true) = 0.5 \times 0.9 = 0.45$$

- This round of sampling returns the event [*true*, *false*, *true*, *true*] with weight 0.45.

Importance Sampling (reviewed)



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

Query $\mathbf{P}(Rain \mid Cloudy = true, WetGrass = true)$

- Generate an event in the chosen topological order.
  1. *Cloudy* is an evidence variable with value *true*.

     $$w \leftarrow w \times P(Cloudy = true) = 0.5$$

  2. *Sprinkler* is not an evidence variable. Sample from $\mathbf{P}(Sprinkler \mid Cloudy = true) = \langle 0.1, 0.9 \rangle$.

     $Sprinkler = false$ (returned value)

  3. *Rain* is not an evidence variable. Sample from $\mathbf{P}(Rain \mid Cloudy = true) = \langle 0.8, 0.2 \rangle$.

     $Rain = true$

  4. *WetGrass* is an evidence variable with value *true*.

     $$w \leftarrow w \times P(WetGrass = true \mid Sprinkler = false, Rain = true) = 0.5 \times 0.9 = 0.45$$

- This round of sampling returns the event [*true*, *false*, *true*, *true*] with weight 0.45.
- This event is tallied under $Rain = true$ in generating the distribution estimate

$$\hat{\mathbf{P}}(Rain \mid Cloudy = true, WetGrass = true)$$

Convergence of MH

Theorem 2 MH converges to the correct stationary distribution $\pi(\boldsymbol{x})$ for any proposal distribution $q(\boldsymbol{x}' \mid \boldsymbol{x})$, provided it defines an *ergodic* transition kernel.

// every state is reachable from every other state,
// and exactly one stationary distribution exists.

Convergence of MH

Theorem 2 MH converges to the correct stationary distribution $\pi(\boldsymbol{x})$ for any proposal distribution $q(\boldsymbol{x}' \mid \boldsymbol{x})$, provided it defines an *ergodic* transition kernel.

// every state is reachable from every other state,
// and exactly one stationary distribution exists.

Proof Consider two cases: $\boldsymbol{x} = \boldsymbol{x}'$ and $\boldsymbol{x} \neq \boldsymbol{x}'$. We show that $\pi(\boldsymbol{x})$ is a *detailed* balance, which will imply it being stationary.

// $\pi(\boldsymbol{x})k(\boldsymbol{x} \to \boldsymbol{x}') = \pi(\boldsymbol{x}')k(\boldsymbol{x}' \to \boldsymbol{x})$ (detailed balance)
// $\pi(\boldsymbol{x}') = \sum_{\boldsymbol{x}} \pi(\boldsymbol{x})k(\boldsymbol{x} \to \boldsymbol{x}')$ (stationary)

Convergence of MH

Theorem 2 MH converges to the correct stationary distribution $\pi(\boldsymbol{x})$ for any proposal distribution $q(\boldsymbol{x}' \mid \boldsymbol{x})$, provided it defines an *ergodic* transition kernel.

// every state is reachable from every other state,
// and exactly one stationary distribution exists.

Proof Consider two cases: $\boldsymbol{x} = \boldsymbol{x}'$ and $\boldsymbol{x} \neq \boldsymbol{x}'$. We show that $\pi(\boldsymbol{x})$ is a *detailed* balance, which will imply it being stationary.

// $\pi(\boldsymbol{x})k(\boldsymbol{x} \to \boldsymbol{x}') = \pi(\boldsymbol{x}')k(\boldsymbol{x}' \to \boldsymbol{x})$ (detailed balance)
// $\pi(\boldsymbol{x}') = \sum_{\boldsymbol{x}} \pi(\boldsymbol{x})k(\boldsymbol{x} \to \boldsymbol{x}')$ (stationary)

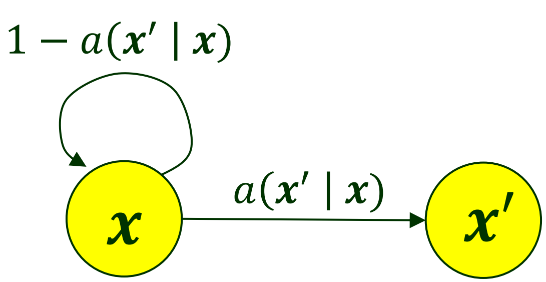
- $\boldsymbol{x} = \boldsymbol{x}'$. The self-loop automatically satisfies the detailed balance condition.

Convergence of MH

Theorem 2 MH converges to the correct stationary distribution $\pi(\boldsymbol{x})$ for any proposal distribution $q(\boldsymbol{x}' \mid \boldsymbol{x})$, provided it defines an *ergodic* transition kernel.

// every state is reachable from every other state,
// and exactly one stationary distribution exists.

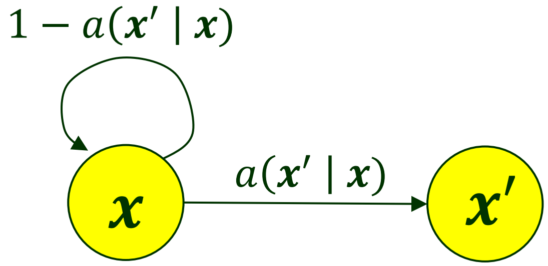
Proof Consider two cases: $\boldsymbol{x} = \boldsymbol{x}'$ and $\boldsymbol{x} \neq \boldsymbol{x}'$. We show that $\pi(\boldsymbol{x})$ is a *detailed* balance, which will imply it being stationary.



// $\pi(\boldsymbol{x})k(\boldsymbol{x} \to \boldsymbol{x}') = \pi(\boldsymbol{x}')k(\boldsymbol{x}' \to \boldsymbol{x})$ (detailed balance)
// $\pi(\boldsymbol{x}') = \sum_{\boldsymbol{x}} \pi(\boldsymbol{x})k(\boldsymbol{x} \to \boldsymbol{x}')$ (stationary)

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

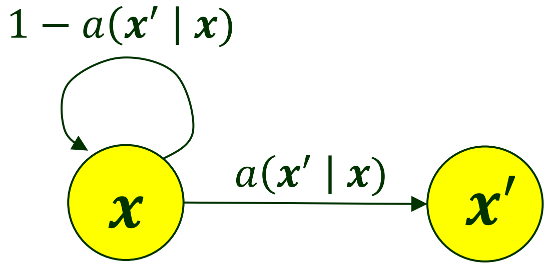
Convergence of MH (cont'd)



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

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

Convergence of MH (cont'd)

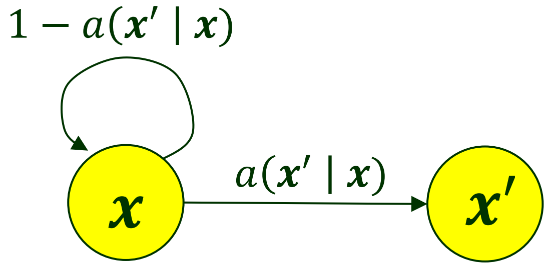


- $\boldsymbol{x} \neq \boldsymbol{x}'$. The transition can occur only if the proposal of $\boldsymbol{x}'$ is accepted.

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

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:

Convergence of MH (cont'd)



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

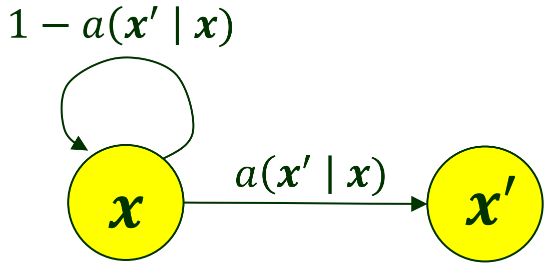
$$k(\boldsymbol{x} \to \boldsymbol{x}') = q(\boldsymbol{x}' \mid \boldsymbol{x})a(\boldsymbol{x}' \mid \boldsymbol{x})$$

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:

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

Convergence of MH (cont'd)

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



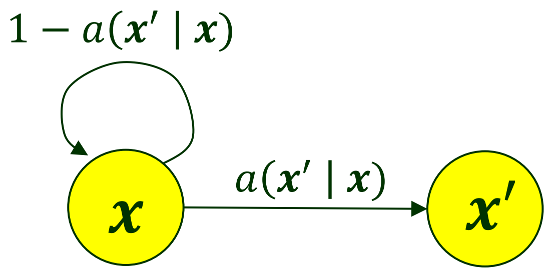
$$k(\boldsymbol{x} \to \boldsymbol{x}') = q(\boldsymbol{x}' \mid \boldsymbol{x})a(\boldsymbol{x}' \mid \boldsymbol{x})$$

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:

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

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

Convergence of MH (cont'd)



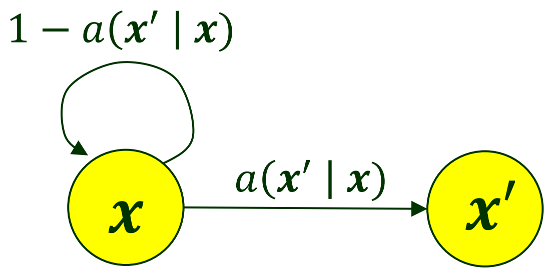
- $\boldsymbol{x} \neq \boldsymbol{x}'$. The transition can occur only if the proposal of $\boldsymbol{x}'$ is accepted.

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

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:

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

Convergence of MH (cont'd)



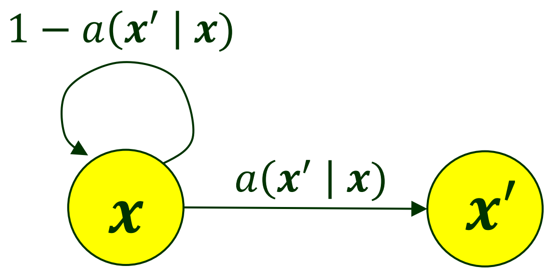
- $\boldsymbol{x} \neq \boldsymbol{x}'$. The transition can occur only if the proposal of $\boldsymbol{x}'$ is accepted.

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

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:

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

Convergence of MH (cont'd)



- $\boldsymbol{x} \neq \boldsymbol{x}'$. 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:

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