Probabilistic Graphical Models

Lecture 12: Approximate Inference: Sampling

Matthew Walter

TTI-Chicago

May 19, 2020

Some slide content courtesy of Eric Xing

Approaches to Inference

- Exact inference algorithms
  - Variable elimination
  - Belief propagation
    - Sum-product message passing
    - Belief update message passing

- Approximate inference algorithms
  - Variational methods
    - Loopy belief propagation
    - Mean-field approximation
  - Sample-based inference: Markov chain Monte Carlo (MCMC)

Monte Carlo Methods

- Suppose that we have a distribution P over $\mathcal{X}$
- We want to estimate $\mathbb{E}[f(\mathcal{X})]$ for some function $f(\mathcal{X})$
- Generate M samples ("particles") from P $\{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(M)}\}$ and estimate the expectation of $f(\mathcal{X})$

$$\hat{\mathbb{E}}[f(\boldsymbol{X})] = \frac{1}{M} \sum_{m=1}^{M} f(\boldsymbol{x}^{(m)})$$

- Suppose that we want the marginal $P(\boldsymbol{Y} = \boldsymbol{y})$ for some $\boldsymbol{Y} \subseteq \mathcal{X}$

Monte Carlo Methods

- Suppose that we have a distribution P over $\mathcal{X}$
- We want to estimate $\mathbb{E}[f(\mathcal{X})]$ for some function $f(\mathcal{X})$
- Generate M samples ("particles") from P $\{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(M)}\}$ and estimate the expectation of $f(\mathcal{X})$

$$\hat{\mathbb{E}}[f(\boldsymbol{X})] = \frac{1}{M} \sum_{m=1}^{M} f(\boldsymbol{x}^{(m)})$$

- Suppose that we want the marginal $P(\boldsymbol{Y} = \boldsymbol{y})$ for some $\boldsymbol{Y} \subseteq \mathcal{X}$

  $\Rightarrow$ If $f(\mathcal{X}) = \mathbb{1}(\xi\langle \boldsymbol{Y} \rangle = \boldsymbol{y})$, this gives us the marginal $P(\boldsymbol{y})$

$$\hat{P}(\boldsymbol{y}) = \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}(\langle \boldsymbol{Y} \rangle = \boldsymbol{y})$$

  (i.e., the fraction of particles consistent with $\boldsymbol{Y} = \boldsymbol{y}$)

Monte Carlo Methods

- Draw random samples from the desired distribution
- Yields a stochastic representation of a complex distribution
- Exhibits *asymptotic* convergence
- Questions:
  - How do we sample from a given distribution (i.e., not all distributions can be easily sampled)?
  - How can we make better use of samples (not all samples are equal)?
  - How do we know when we've sampled enough?

Sampling from Distributions

- Key idea: Convert this to sampling from a uniform distribution (easy)
- Consider a discrete distribution, i.e., a multinomial $P(X)$ for $\text{Val}(X) = \{x^1, \ldots, x^k\}$ with $P(X = x^i) = \theta_i$
  1. Partition interval $[0, 1]$ into k subintervals $[0, \theta_1), [\theta_1, \theta_1 + \theta_2), \ldots]$
  2. Sample $s \sim \text{Uniform}[0, 1]$ and choose x^i if s is in the i^{th} interval

Sampling from Distributions

- Key idea: Convert this to sampling from a uniform distribution (easy)
- Consider a discrete distribution, i.e., a multinomial $P(X)$ for $\text{Val}(X) = \{x^1, \ldots, x^k\}$ with $P(X = x^i) = \theta_i$
  1. Partition interval $[0, 1]$ into k subintervals $[0, \theta_1), [\theta_1, \theta_1 + \theta_2), \ldots]$
  2. Sample $s \sim \text{Uniform}[0, 1]$ and choose x^i if s is in the i^{th} interval
- Similarly, for continuous distributions $p(x)$ we consider the cumulative distribution $h(y) = \int_{-\infty}^{y} p(x)\, dx$: Sample $s \sim \text{Uniform}[0, 1]$ and return $h^{-1}(s)$ (not always easy: we can't always compute and invert $h(y)$)

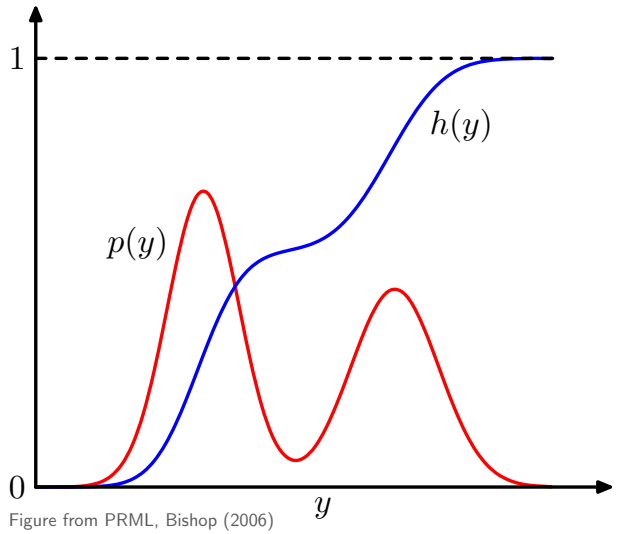
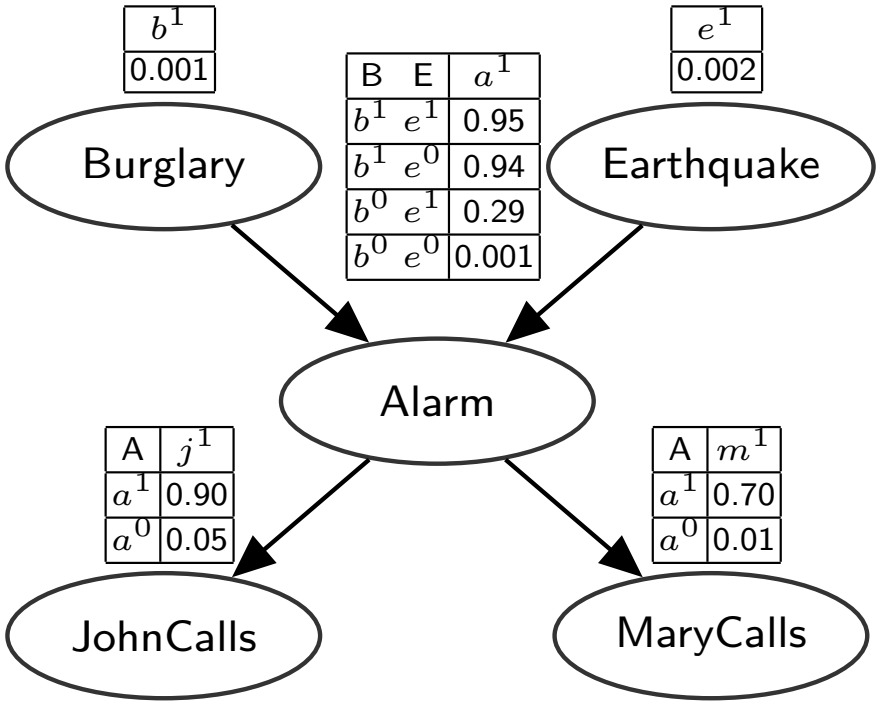


Figure from PRML, Bishop (2006)

Forward Sampling

- Arguably the simplest approach to generating samples (particles)
- Involves *directly* sampling from $P(\mathcal{X})$
- Requires knowledge of the partition function for MRFs
- Primarily restricted to Bayesian networks
- Procedure: Given a ordering of the random variables,
  1. Sample each X_i according to its CPD $P(X_i \,|\, \text{Pa}_{X_i}^G)$ using the current sampled values for its parents
  2. Repeat M times

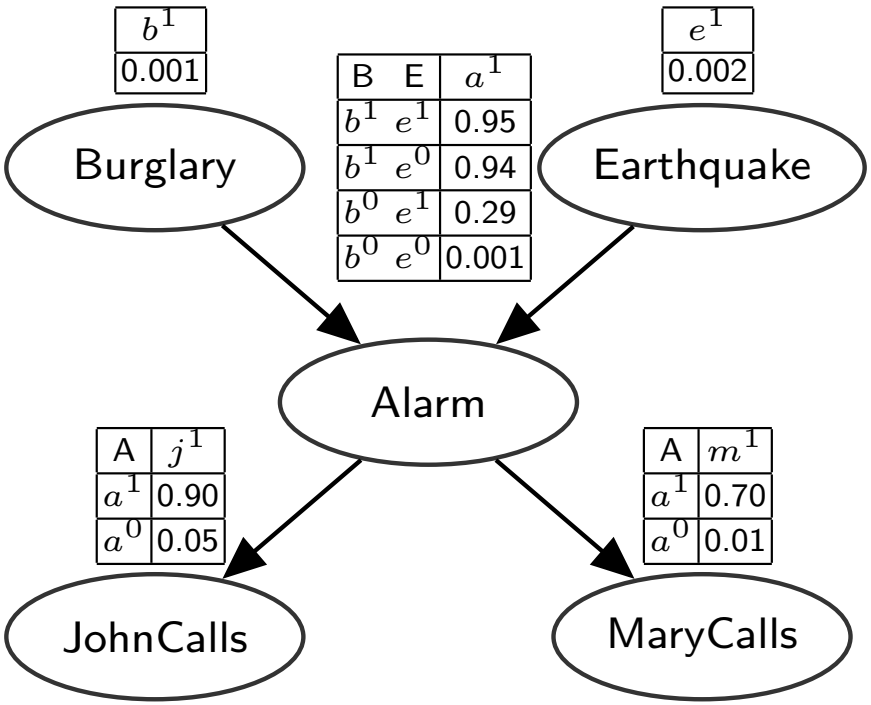
Test


b^1
0.001
Burglary
B E a^1
b^1 e^1 0.95
b^1 e^0 0.94
b^0 e^1 0.29
b^0 e^0 0.001
e^1
0.002
Earthquake
Alarm
A j^1
a^1 0.90
a^0 0.05
JohnCalls
A m^1
a^1 0.70
a^0 0.01
MaryCalls

Forward Sampling: Example

Procedure: Given a ordering of the random variables,

1. Sample each X_i according to it's CPD $P(X_i \mid \text{Pa}_{X_i}^G)$ using the current sampled values for its parents
2. Repeat M times



E0	B0	A0	M0	J0
E0	B0	A0	M0	J0
E0	B0	A0	M0	J1
E0	B0	A0	M0	J0
E0	B0	A0	M0	J0
E0	B0	A0	M0	J0
E1	B0	A1	M1	J1
E0	B0	A0	M0	J0
E0	B0	A0	M0	J0
E0	B0	A0	M0	J0

$$P(J \mid a^0) = \frac{P(J, a^0)}{P(a^0)} = \{1/9, 8/9\}$$

Forward Sampling: Example

Procedure: Given a ordering of the random variables,

1. Sample each X_i according to it's CPD $P(X_i \mid \text{Pa}_{X_i}^G)$ using the current sampled values for its parents
2. Repeat M times

Problem: It is difficult to get sufficient samples of rare events

- What about $P(J \mid A^1)$? We only have one sample: $P(J \mid a^1) = \frac{P(J,a^1)}{P(a^1)} = \{0, 1\}$
- What about $P(J \mid b^1)$? We don't have any samples!

E0	B0	A0	M0	J0
E0	B0	A0	M0	J0
E0	B0	A0	M0	J1
E0	B0	A0	M0	J0
E0	B0	A0	M0	J0
E0	B0	A0	M0	J0
E1	B0	A1	M1	J1
E0	B0	A0	M0	J0
E0	B0	A0	M0	J0
E0	B0	A0	M0	J0

How Many Samples Are Necessary?

- Suppose that we are interested in a particular event $\boldsymbol{Y} = \boldsymbol{y}$
- $\mathbb{1}(\langle \boldsymbol{Y} \rangle = \boldsymbol{y})$ is a Bernoulli random variable with success $P(\boldsymbol{y})$
  - M samples define M independent Bernoulli trials

How Many Samples Are Necessary?

- Suppose that we are interested in a particular event $\boldsymbol{Y} = \boldsymbol{y}$
- $\mathbb{1}(\langle \boldsymbol{Y} \rangle = \boldsymbol{y})$ is a Bernoulli random variable with success $P(\boldsymbol{y})$
  - M samples define M independent Bernoulli trials
- **Hoeffding bound**: Consider a sequence $\mathcal{D} = \{x^{(1)}, x^{(2)}, \ldots, x^{(M)}\}$ of M independent Bernoulli trials with probability p. Letting $T_{\mathcal{D}} = \frac{1}{M} \sum_m x^{(m)}$ (e.g., fraction of heads), for some $\epsilon > 0$

$$P_{\mathcal{D}}(T_{\mathcal{D}} > p + \epsilon) \leq e^{-2M\epsilon^2}$$

$$P_{\mathcal{D}}(T_{\mathcal{D}} < p - \epsilon) \leq e^{-2M\epsilon^2}$$

How Many Samples Are Necessary?

- Suppose that we are interested in a particular event $\boldsymbol{Y} = \boldsymbol{y}$
- $\mathbb{1}(\langle \boldsymbol{Y} \rangle = \boldsymbol{y})$ is a Bernoulli random variable with success $P(\boldsymbol{y})$
  - M samples define M independent Bernoulli trials
- **Hoeffding bound**: Consider a sequence $\mathcal{D} = \{x^{(1)}, x^{(2)}, \ldots, x^{(M)}\}$ of M independent Bernoulli trials with probability p. Letting $T_{\mathcal{D}} = \frac{1}{M} \sum_m x^{(m)}$ (e.g., fraction of heads), for some $\epsilon > 0$

$$P_{\mathcal{D}}(T_{\mathcal{D}} > p + \epsilon) \leq e^{-2M\epsilon^2}$$

$$P_{\mathcal{D}}(T_{\mathcal{D}} < p - \epsilon) \leq e^{-2M\epsilon^2}$$

- Using the Hoeffding bound, we get

$$P_{\mathcal{D}}(\hat{P}_{\mathcal{D}} \notin [P(\boldsymbol{y}) - \epsilon, P(\boldsymbol{y}) + \epsilon]) \leq 2e^{-2M\epsilon^2}$$

How Many Samples Are Necessary?

- Suppose that we are interested in a particular event $\boldsymbol{Y} = \boldsymbol{y}$
- $\mathbb{1}(\langle \boldsymbol{Y} \rangle = \boldsymbol{y})$ is a Bernoulli random variable with success $P(\boldsymbol{y})$
  - M samples define M independent Bernoulli trials
- **Hoeffding bound**: Consider a sequence $\mathcal{D} = \{x^{(1)}, x^{(2)}, \ldots, x^{(M)}\}$ of M independent Bernoulli trials with probability p. Letting $T_{\mathcal{D}} = \frac{1}{M} \sum_m x^{(m)}$ (e.g., fraction of heads), for some $\epsilon > 0$

$$P_{\mathcal{D}}(T_{\mathcal{D}} > p + \epsilon) \leq e^{-2M\epsilon^2}$$

$$P_{\mathcal{D}}(T_{\mathcal{D}} < p - \epsilon) \leq e^{-2M\epsilon^2}$$

- Using the Hoeffding bound, we get

$$P_{\mathcal{D}}(\hat{P}_{\mathcal{D}} \notin [P(\boldsymbol{y}) - \epsilon, P(\boldsymbol{y}) + \epsilon]) \leq 2e^{-2M\epsilon^2}$$

- Thus, to achieve an estimate within ϵ of the true marginal (absolute error) with probability at least $1 - \delta$

$$M \geq \frac{\ln(2/\delta)}{2\epsilon^2}$$

How Many Samples Are Necessary?

- **Chernoff bound**: Consider a sequence $\mathcal{D} = \{x^{(1)}, x^{(2)}, \ldots, x^{(M)}\}$ of M independent Bernoulli trials with probability p. For $T_{\mathcal{D}} = \frac{1}{M} \sum_m x^{(m)}$ we have

$$P_{\mathcal{D}}(T_{\mathcal{D}} > p(1+\epsilon)) \leq e^{-Mp\epsilon^2/3}$$

$$P_{\mathcal{D}}(T_{\mathcal{D}} < p(1-\epsilon)) \leq e^{-Mp\epsilon^2/2}$$

- Using the Chernoff bound, we can compute the likelihood associated with the *relative error* (suitable for disparate likelihoods)

$$P_{\mathcal{D}}\left(\hat{P}_{\mathcal{D}} \notin P(\boldsymbol{y})(1 \pm \epsilon)\right) \leq 2e^{-MP(\boldsymbol{y})\epsilon^2/3}$$

How Many Samples Are Necessary?

- **Chernoff bound**: Consider a sequence $\mathcal{D} = \{x^{(1)}, x^{(2)}, \ldots, x^{(M)}\}$ of M independent Bernoulli trials with probability p. For $T_{\mathcal{D}} = \frac{1}{M} \sum_m x^{(m)}$ we have

$$P_{\mathcal{D}}(T_{\mathcal{D}} > p(1+\epsilon)) \leq e^{-Mp\epsilon^2/3}$$

$$P_{\mathcal{D}}(T_{\mathcal{D}} < p(1-\epsilon)) \leq e^{-Mp\epsilon^2/2}$$

- Using the Chernoff bound, we can compute the likelihood associated with the *relative error* (suitable for disparate likelihoods)

$$P_{\mathcal{D}}\left(\hat{P}_{\mathcal{D}} \notin P(\boldsymbol{y})(1 \pm \epsilon)\right) \leq 2e^{-MP(\boldsymbol{y})\epsilon^2/3}$$

- Thus, to achieve a relative error no more than ϵ with probability $1 - \delta$

$$M \geq 3\frac{\ln(2/\delta)}{P(\boldsymbol{y})\epsilon^2}$$

How Many Samples Are Necessary?

- **Chernoff bound**: Consider a sequence $\mathcal{D} = \{x^{(1)}, x^{(2)}, \ldots, x^{(M)}\}$ of M independent Bernoulli trials with probability p. For $T_{\mathcal{D}} = \frac{1}{M} \sum_m x^{(m)}$ we have

$$P_{\mathcal{D}}(T_{\mathcal{D}} > p(1+\epsilon)) \leq e^{-Mp\epsilon^2/3}$$

$$P_{\mathcal{D}}(T_{\mathcal{D}} < p(1-\epsilon)) \leq e^{-Mp\epsilon^2/2}$$

- Using the Chernoff bound, we can compute the likelihood associated with the *relative error* (suitable for disparate likelihoods)

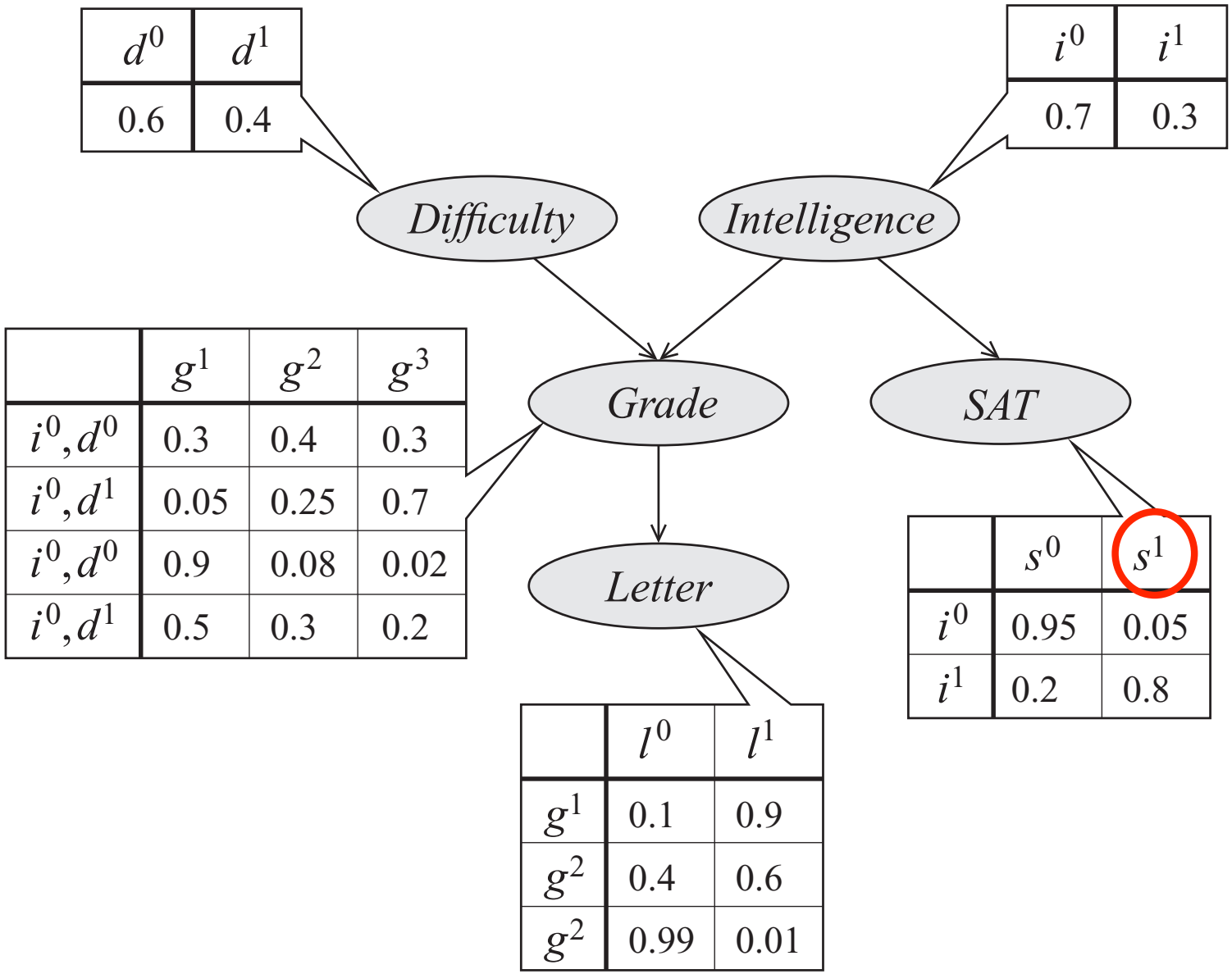
$$P_{\mathcal{D}}\left(\hat{P}_{\mathcal{D}} \notin P(\boldsymbol{y})(1 \pm \epsilon)\right) \leq 2e^{-MP(\boldsymbol{y})\epsilon^2/3}$$

- Thus, to achieve a relative error no more than ϵ with probability $1 - \delta$

$$M \geq 3\frac{\ln(2/\delta)}{P(\boldsymbol{y})\epsilon^2}$$

- The number of samples is inversely proportional to $P(\boldsymbol{y})$
  - Low probability events require more samples
  - Oftentimes, we don't know $P(\boldsymbol{y})$

Dealing with Evidence: Example



Dealing with Evidence

- How can we use sampling to estimate $P(\boldsymbol{y} \,|\, \boldsymbol{E} = \boldsymbol{e})$
- One approach is to sample from $P(\boldsymbol{y} \,|\, \boldsymbol{E} = \boldsymbol{e})$ via *rejection sampling*
  - Draw $\tilde{M}$ samples $\boldsymbol{x} \sim P(\boldsymbol{X})$ as before
  - Discard (reject) samples that are inconsistent with $\boldsymbol{E} = \boldsymbol{e}$
  - Resulting samples are from the posterior $P(\boldsymbol{X} \,|\, \boldsymbol{e})$

Dealing with Evidence

- How can we use sampling to estimate $P(\boldsymbol{y} \,|\, \boldsymbol{E} = \boldsymbol{e})$
- One approach is to sample from $P(\boldsymbol{y} \,|\, \boldsymbol{E} = \boldsymbol{e})$ via *rejection sampling*
  - Draw $\tilde{M}$ samples $\boldsymbol{x} \sim P(\boldsymbol{X})$ as before
  - Discard (reject) samples that are inconsistent with $\boldsymbol{E} = \boldsymbol{e}$
  - Resulting samples are from the posterior $P(\boldsymbol{X} \,|\, \boldsymbol{e})$
- Expected number of particles that are kept is $\tilde{M} P(\boldsymbol{e})$
- Requires $\tilde{M} = M / P(\boldsymbol{e})$ samples to get M that are kept

Dealing with Evidence

- How can we use sampling to estimate $P(\boldsymbol{y} \,|\, \boldsymbol{E} = \boldsymbol{e})$
- One approach is to sample from $P(\boldsymbol{y} \,|\, \boldsymbol{E} = \boldsymbol{e})$ via *rejection sampling*
  - Draw $\tilde{M}$ samples $\boldsymbol{x} \sim P(\boldsymbol{X})$ as before
  - Discard (reject) samples that are inconsistent with $\boldsymbol{E} = \boldsymbol{e}$
  - Resulting samples are from the posterior $P(\boldsymbol{X} \,|\, \boldsymbol{e})$
- Expected number of particles that are kept is $\tilde{M} P(\boldsymbol{e})$
- Requires $\tilde{M} = M / P(\boldsymbol{e})$ samples to get M that are kept
- Unfortunately, $P(\boldsymbol{e})$ is often small in practice (i.e., rare events)

Dealing with Evidence

- How can we use sampling to estimate $P(\boldsymbol{y} \,|\, \boldsymbol{E} = \boldsymbol{e})$
- One approach is to sample from $P(\boldsymbol{y} \,|\, \boldsymbol{E} = \boldsymbol{e})$ via *rejection sampling*
  - Draw $\tilde{M}$ samples $\boldsymbol{x} \sim P(\boldsymbol{X})$ as before
  - Discard (reject) samples that are inconsistent with $\boldsymbol{E} = \boldsymbol{e}$
  - Resulting samples are from the posterior $P(\boldsymbol{X} \,|\, \boldsymbol{e})$
- Expected number of particles that are kept is $\tilde{M} P(\boldsymbol{e})$
- Requires $\tilde{M} = M / P(\boldsymbol{e})$ samples to get M that are kept
- Unfortunately, $P(\boldsymbol{e})$ is often small in practice (i.e., rare events)
- Alternatively, separately estimate $P(\boldsymbol{y}, \boldsymbol{e})$ & $P(\boldsymbol{e})$ and use Bayes' rule
  - Low relative error requires a number of samples that grows linearly with $1/P(\boldsymbol{e})$
  - A bound on the absolute error for $P(\boldsymbol{e})$, which does not depend on $P(\boldsymbol{e})$, does not provide any bound on $P(\boldsymbol{y}, \boldsymbol{e})/P(\boldsymbol{e})$

Likelihood Weighting

- With forward sampling, evidence only affects sampling of descendents
- Forward sampling would reject particles that are inconsistent with evidence, but does not result in true posterior (e.g., 30% of the samples include (i^1, s^1), which is the same as the prior over I)
- Basic idea: Weight each particle by the probability of the evidence variables taking the observed values (e.g., upweight (i^1, s^1) and downweight (i^0, s^1))
- Now, we can estimate the conditional as

$$\hat{P}_{\mathcal{D}}(\boldsymbol{y} \mid \boldsymbol{e}) = \frac{\sum_{m=1}^{M} w^{(m)} \mathbb{1}(\boldsymbol{y}^{(m)} = \boldsymbol{y})}{\sum_{m=1}^{M} w^{(m)}}$$

where $w^{(m)}$ is the product of probabilities of the evidence

Likelihood Weighting

For $m = 1, 2, \ldots, M$:

1. Initialize $w^{(m)} = 1$
2. For each X_i in topological order:
   - If X_i is an evidence variable
     - $x_i^{(m)} \leftarrow e_i$
     - $w^{(m)} \leftarrow w^{(m)} \cdot P(x_i^{(m)} \mid \boldsymbol{u}_i^{(m)})$ where $\boldsymbol{u}_i^{(m)}$ are sampled parents of $x_i^{(m)}$
   - Else:
     - Sample $x_i^{(m)}$ from $P(x_i^{(m)} \mid \boldsymbol{u}_i^{(m)})$

Likelihood Weighting

For $m = 1, 2, \ldots, M$:

1. Initialize $w^{(m)} = 1$
2. For each X_i in topological order:
   - If X_i is an evidence variable
     - $x_i^{(m)} \leftarrow e_i$
     - $w^{(m)} \leftarrow w^{(m)} \cdot P(x_i^{(m)} \mid \boldsymbol{u}_i^{(m)})$ where $\boldsymbol{u}_i^{(m)}$ are sampled parents of $x_i^{(m)}$
   - Else:
     - Sample $x_i^{(m)}$ from $P(x_i^{(m)} \mid \boldsymbol{u}_i^{(m)})$

- In the previous example, while 70% of the particles would involve (i^0, s^1), their weight would only be 0.05
- Forward sampling is a special case of likelihood weighting with $w^{(m)} = 1$

Importance Sampling

- Often, we may not be able to sample from the *target distribution* P
  - P may not be known
  - Sampling from P may be computationally expensive (e.g., a posterior distribution for a Bayesian network)
- Instead, consider sampling from a (simpler) *proposal distribution* Q, s.t. $Q > 0$ whenever $P > 0$ (i.e., support of Q contains support of P)
- This gives rise to an equivalent expression for the expectation:

$$\mathbb{E}_P(f(\boldsymbol{X})) = \mathbb{E}_Q\left[f(\boldsymbol{X})\frac{P(\boldsymbol{X})}{Q(\boldsymbol{X})}\right]$$

Importance Sampling

- Often, we may not be able to sample from the *target distribution* P
  - P may not be known
  - Sampling from P may be computationally expensive (e.g., a posterior distribution for a Bayesian network)
- Instead, consider sampling from a (simpler) *proposal distribution* Q, s.t. $Q > 0$ whenever $P > 0$ (i.e., support of Q contains support of P)
- This gives rise to an equivalent expression for the expectation:

$$\mathbb{E}_P(f(\boldsymbol{X})) = \mathbb{E}_Q\left[f(\boldsymbol{X})\frac{P(\boldsymbol{X})}{Q(\boldsymbol{X})}\right]$$

- Gives rise to the *unnormalized importance sampling* estimator: Given M samples from Q, weight each one by $\frac{P(\boldsymbol{x}^{(m)})}{Q(\boldsymbol{x}^{(m)})}$, yielding the estimate

$$\hat{\mathbb{E}}_{\mathcal{D}}[f] = \frac{1}{M}\sum_{m=1}^{M} f(\boldsymbol{x}^{(m)})\frac{P(\boldsymbol{x}^{(m)})}{Q(\boldsymbol{x}^{(m)})}$$

Unnormalized Importance Sampling

- (Admittedly, the name is confusing)
- Estimator is *unbiased*, i.e., for D sampled data sets

$$\mathbb{E}_D\left[\hat{\mathbb{E}}_D[f]\right] = \mathbb{E}_Q[f(\boldsymbol{X})w(\boldsymbol{X})] = \mathbb{E}_P[f(\boldsymbol{X})]$$

- The variance of the estimator becomes

$$\begin{aligned}\mathbb{V} &= \frac{1}{M}\mathbb{E}_Q[(f(\boldsymbol{X})w(\boldsymbol{X}))^2] - \mathbb{E}_Q[(f(\boldsymbol{X})w(\boldsymbol{X}))]^2 \\ &= \frac{1}{M}\mathbb{E}_Q[(w(\boldsymbol{X}))^2] - (\mathbb{E}_Q[(w(\boldsymbol{X}))])^2 \quad \text{when } f(\boldsymbol{X}) = 1\end{aligned}$$

  (i.e., the variance of $w(\boldsymbol{X}) = P(\boldsymbol{X})/Q(\boldsymbol{X})$)
- Variance of the estimate decreases with
  - A larger number of samples
  - Choosing Q to be closer to P, specifically $Q \propto |f(\boldsymbol{X})|P(\boldsymbol{X})$

Normalized Importance Sampling

- However, we may not be able to evaluate $P(\boldsymbol{x}^{(m)})$
  - May involve a posterior distribution over a Bayesian network
  - May require calculating a partition function (i.e., for an MRF)
- Suppose that we know $\tilde{P}$, i.e., P up to a normalizing constant
- What if we use $\tilde{P}$ in place of P when computing weights?
- Intuition: the weight is a random variable with expected value

$$\mathbb{E}_Q[w(\boldsymbol{X})] = \sum_{\boldsymbol{x}} Q(\boldsymbol{x}) \frac{\tilde{P}(\boldsymbol{x})}{Q(\boldsymbol{x})} = \sum_{\boldsymbol{x}} \tilde{P}(\boldsymbol{x}) = Z$$

Normalized Importance Sampling

- We can write the objective as

$$\mathbb{E}_P[f(\boldsymbol{X})] = \frac{\mathbb{E}_Q[f(\boldsymbol{X})w(\boldsymbol{X})]}{\mathbb{E}_Q[w(\boldsymbol{X})]}$$

where

$$w(\boldsymbol{X}) = \frac{\tilde{P}(\boldsymbol{X})}{Q(\boldsymbol{X})}$$

- Gives rise to *normalized importance sampling estimator*

$$\hat{\mathbb{E}}_{\mathcal{D}}[f] = \frac{\sum_{m=1}^{M} f(\boldsymbol{x}^{(m)})w(\boldsymbol{x}^{(m)})}{\sum_{m=1}^{M} w(\boldsymbol{x}^{(m)})} \qquad w(\boldsymbol{x}^{(m)}) = \frac{\tilde{P}(\boldsymbol{x}^{(m)})}{Q(\boldsymbol{x}^{(m)})}$$

(*normalized* importance sampling uses the *unnormalized* distribution)

Normalized Importance Sampling

- Not unbiased, but bias decreases as $\frac{1}{M}$
- Typically, the variance is lower than that of the unnormalized estimator (but not always), outweighing added bias

$$\mathbb{V}_P[\hat{\mathbb{E}}_{\mathcal{D}}(f(\mathbf{X}))] \approx \frac{1}{M}\mathbb{V}_P[f(\boldsymbol{X})](1 + \mathbb{V}_Q[w(\boldsymbol{X})])$$

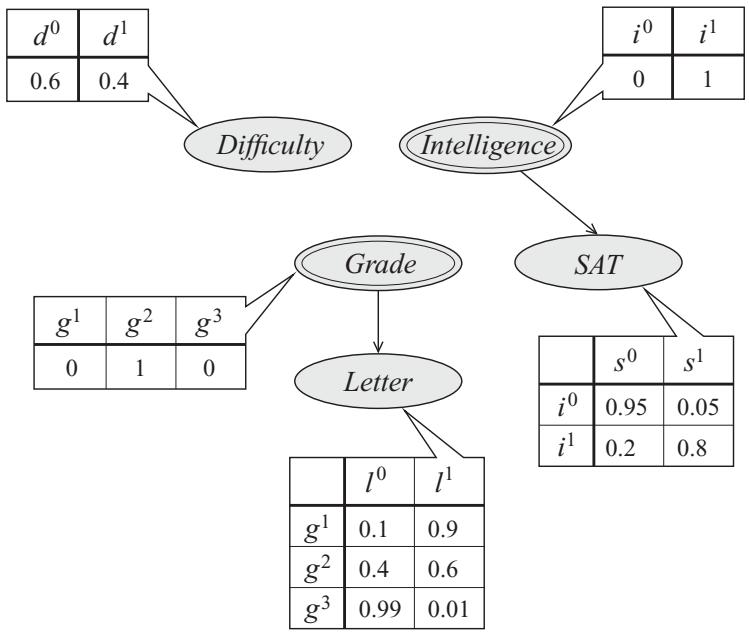
- We can use the above expression to measure the *effective sample size*, which is useful in deciding whether to keep sampling, i.e., for a set $\mathcal{D}$ of M samples:

$$M_{\text{eff}} = \frac{M}{1 + Var[\mathcal{D}]}$$

$$\mathbb{V}[\mathcal{D}] = \sum_{m=1}^{M} w(\boldsymbol{x}^{(m)})^2 - \left(\sum_{m=1}^{M} w(\boldsymbol{x}^{(m)})\right)^2$$

Importance Sampling for Bayesian Networks

- Suppose that P is represented by a Bayesian network and that we are interested in event $\boldsymbol{Z} = \boldsymbol{z}$
- The proposal distribution Q takes the form of a **mutilated network** $\mathcal{B}_{\mathbf{Z}=\mathbf{z}}$ for instantiation $\mathbf{Z} = \mathbf{z}$, where:
  - Each $Z_i \in \boldsymbol{Z}$ has no parents, and $P(Z_i = z_i) = 1$ and zero otherwise
  - Parents and CPDs of all other variables $X \notin Z_i$ are unchanged



d^0	d^1
0.6	0.4

i^0	i^1
0	1

g^1	g^2	g^3
0	1	0

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

$\mathcal{B}_{I=i^1, G=g^2}$

Ratio and Normalized Likelihood Weighting

- Determine $P(\boldsymbol{Y} = \boldsymbol{y} \,|\, \boldsymbol{E} = \boldsymbol{e})$ by sampling from Q (mutilated network)
- Ratio likelihood weighting: Calculate $P(\boldsymbol{Y} = \boldsymbol{y} \,|\, \boldsymbol{E} = \boldsymbol{e})$ using two runs of likelihood weighting, one for $P(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{E} = \boldsymbol{e})$ and the other for $P(\boldsymbol{E} = \boldsymbol{e})$ (via Bayes' rule)

$$\hat{P}(\boldsymbol{y} \,|\, \boldsymbol{e}) = \frac{\hat{P}(\boldsymbol{y}, \boldsymbol{e})}{\hat{P}(\boldsymbol{e})} = \frac{1/M \sum_{m=1}^{M} w^{(m)}}{1/M' \sum_{m=1}^{M'} \bar{w}^{(m)}}$$

  - Useful for single queries

Ratio and Normalized Likelihood Weighting

- Determine $P(\boldsymbol{Y} = \boldsymbol{y} \,|\, \boldsymbol{E} = \boldsymbol{e})$ by sampling from Q (mutilated network)
- Ratio likelihood weighting: Calculate $P(\boldsymbol{Y} = \boldsymbol{y} \,|\, \boldsymbol{E} = \boldsymbol{e})$ using two runs of likelihood weighting, one for $P(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{E} = \boldsymbol{e})$ and the other for $P(\boldsymbol{E} = \boldsymbol{e})$ (via Bayes' rule)

$$\hat{P}(\boldsymbol{y} \,|\, \boldsymbol{e}) = \frac{\hat{P}(\boldsymbol{y}, \boldsymbol{e})}{\hat{P}(\boldsymbol{e})} = \frac{1/M \sum_{m=1}^{M} w^{(m)}}{1/M' \sum_{m=1}^{M'} \bar{w}^{(m)}}$$

  - Useful for single queries
- Normalized likelihood weighting: Calculate conditional probability for a *full distribution* over $\boldsymbol{Y}$ with $P(\boldsymbol{Y} \,|\, \boldsymbol{E} = \boldsymbol{e})$ as the target using normalized importance sampling evaluated at $\tilde{P}(\boldsymbol{Y}, \boldsymbol{e})$ and the LW estimate

$$\hat{P}(\boldsymbol{y} \,|\, \boldsymbol{e}) = \frac{\sum_{m=1}^{M} \mathbb{1}(\boldsymbol{y}^{(m)} = \boldsymbol{y}) w^{(m)}}{\sum_{m=1}^{M} w^{(m)}}$$

Performance of Importance Sampling

- Efficiency of importance sampling depends on how close proposal Q is to the target P
- If the evidence is at the roots, then $Q = P(\boldsymbol{Y} \mid \boldsymbol{e})$ (the posterior) and all samples have equal weight ($P(\boldsymbol{e})$) (no need to compensate as evidence is sampled)
- If evidence is at the leaves, then $Q = P(\boldsymbol{Y})$ (the prior), relying on the weights to correct for evidence
  - Samples are irrelevant if the prior and posterior differ significantly
  - Samples get small weight if evidence is unlikely

Limitations of Monte Carlo

- In general, it is difficult to sample rare events, particularly in high-dimensions
- Importance sampling works poorly if Q is very different from P
- Constructing Q to be similar to P can be difficult
  - A good proposal often requires knowledge of the analytic form of P
- Instead . . . rather than use a fixed proposal Q, what if we could use an *adaptive* proposal?

Markov Chain Monte Carlo

- Intuition: "Fix" an initial sample by resampling some of the variables generated previously

Markov Chain Monte Carlo

- Intuition: "Fix" an initial sample by resampling some of the variables generated previously
- General idea: generate an initial sample and then iteratively sample from each variable *given* previous samples of other variables
- Generates a *sequence* of samples

Markov Chain Monte Carlo

- Intuition: "Fix" an initial sample by resampling some of the variables generated previously
- General idea: generate an initial sample and then iteratively sample from each variable *given* previous samples of other variables
- Generates a *sequence* of samples
- In the beginning, likely not sampling from P
- Over time, sampling distribution gets closer to P

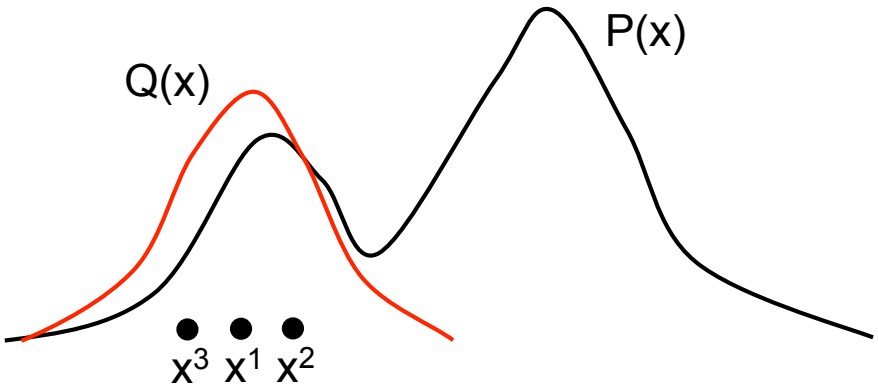
Markov Chain Monte Carlo

- Intuition: "Fix" an initial sample by resampling some of the variables generated previously
- General idea: generate an initial sample and then iteratively sample from each variable *given* previous samples of other variables
- Generates a *sequence* of samples
- In the beginning, likely not sampling from P
- Over time, sampling distribution gets closer to P
- Useful in both directed and undirected models (unlike likelihood weighting and other forward sampling methods)

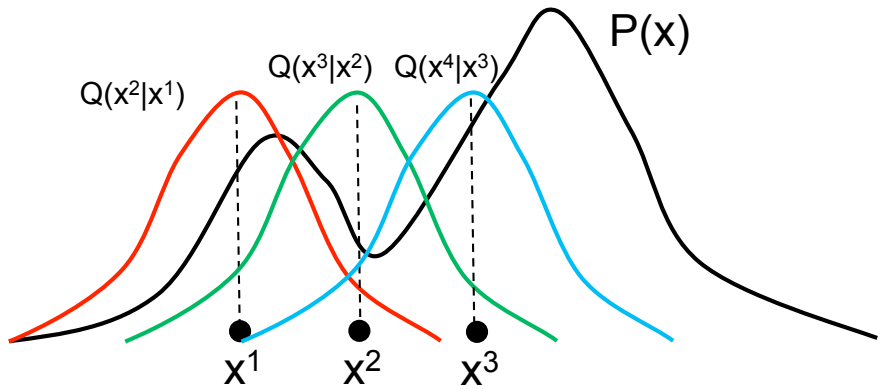
Markov Chain Monte Carlo

- MCMC algorithms employ adaptive proposals
  - Instead of $Q(\boldsymbol{x})$, MCMC uses $Q(x' \mid \boldsymbol{x})$ where $\boldsymbol{x}$ is the previous sample
  - As $\boldsymbol{x}$ changes, so does $Q(x' \mid \boldsymbol{x})$ (as a function of x')

Importance sampling with a (bad) proposal Q(x)



MCMC with adaptive proposal Q(x'|x)



Gibbs Sampling

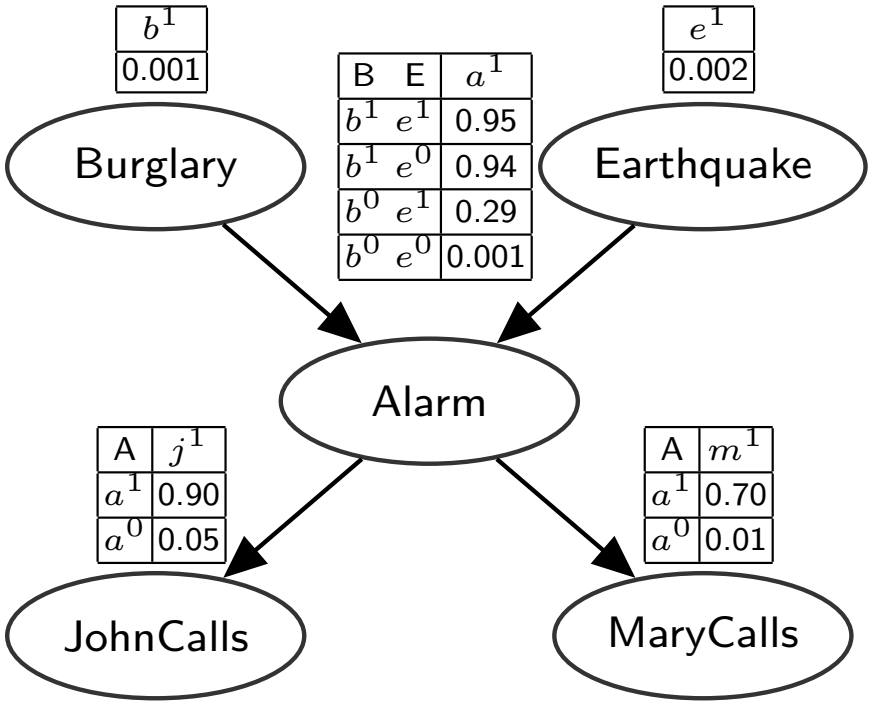
Input: Random variables $\boldsymbol{X}$, factors Φ, initial state distribution $P^{(0)}(\boldsymbol{X})$, and number of time steps T

1. Sample $\boldsymbol{x}^{(0)} \sim P^{(0)}(\boldsymbol{X})$
2. for $t = 1, \ldots, T$
   - $\boldsymbol{x}^{(t)} \leftarrow \boldsymbol{x}^{(t-1)}$
   - For each $X_i \in \boldsymbol{X}$: Sample $x_i^{(t)} \sim P(X_i \,|\, \boldsymbol{x}_{-i})$
3. Return $\boldsymbol{x}^{(0)}, \ldots, \boldsymbol{x}^{(T)}$

Gibbs Sampling

- During each round, we sample one random variable conditioned on other variables fixed to the most recent sampled value
  - This only involves a subset of the factors
  - We only need to look at some values of those factors
- We can consider evidence by first reducing the factors
- Appeal: Unlike forward sampling, each variable is influenced by the others, including downstream evidence

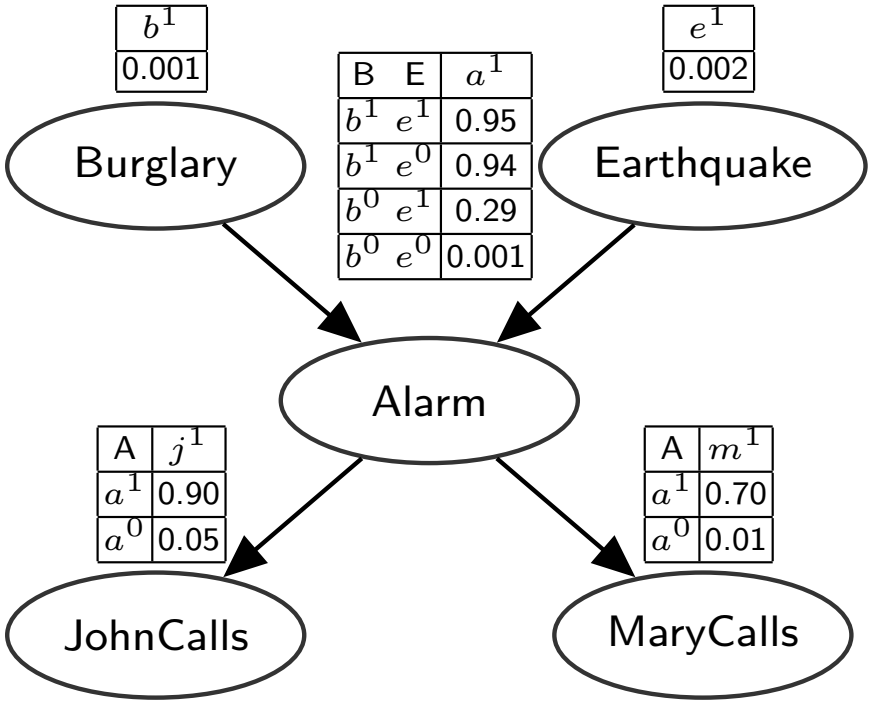
Gibbs Sampling: Example



t	B	E	A	J	M
0	F	F	F	F	F
1					
2					
3					
4					

- Consider the alarm network example
- Assume that we sample variables in the order B, E, A, J, M
- Initialize all variables at $t = 0$ to False

Gibbs Sampling: Example



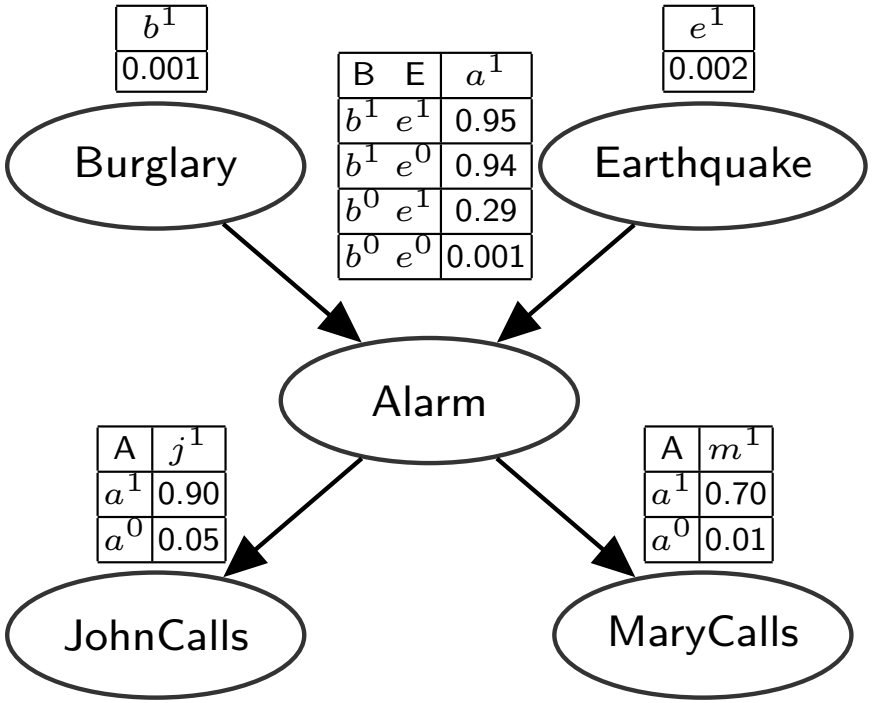
t	B	E	A	J	M
0	F	F	F	F	F
1	F				
2					
3					
4					

- Sample $P(B \mid A, E)$ at $t = 1$ via Bayes rule

$$P(B \mid A, E) \propto P(A \mid B, E)P(B)$$

- $A =$ False, $E =$ False:
  - $P(B = \mathsf{T} \mid A = \mathsf{F}, E = \mathsf{F}) \propto 0.06 \cdot 0.01 = 0.0006$
  - $P(B = \mathsf{F} \mid A = \mathsf{F}, E = \mathsf{F}) \propto 0.999 \cdot 0.999 = 0.9980$

Gibbs Sampling: Example



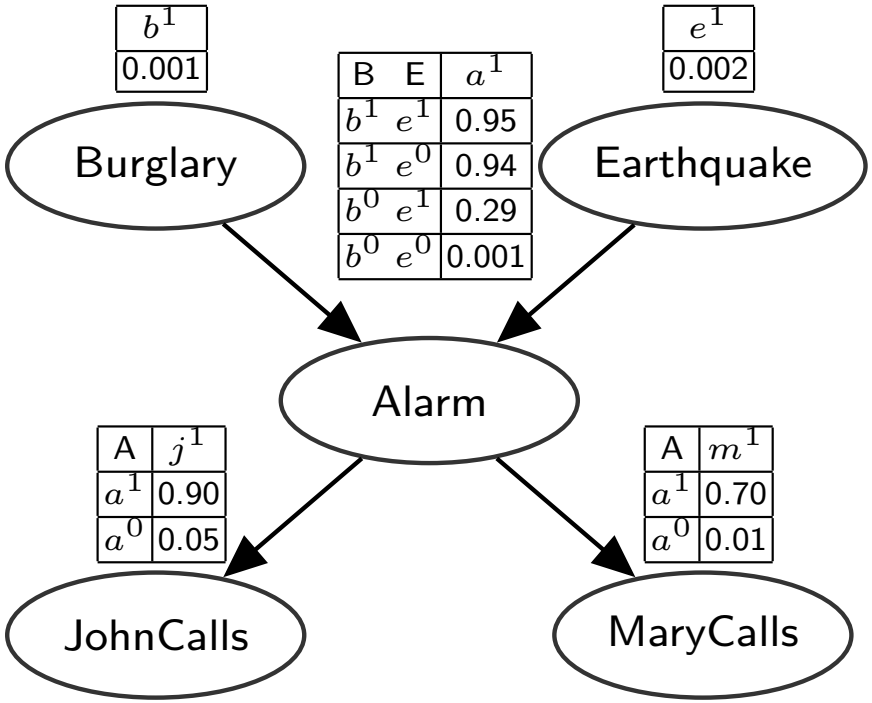
t	B	E	A	J	M
0	F	F	F	F	F
1	F	T			
2					
3					
4					

- Sample $P(E \mid A, B)$ at $t = 1$ via Bayes rule

$$P(E \mid A, B) \propto P(A \mid B, E)P(E)$$

- $A =$ False, $B =$ False:
  - $P(E = \mathsf{T} \mid A = \mathsf{F}, B = \mathsf{F}) \propto 0.71 \cdot 0.02 = 0.0142$
  - $P(E = \mathsf{F} \mid A = \mathsf{F}, B = \mathsf{F}) \propto 0.999 \cdot 0.998 = 0.9970$

Gibbs Sampling: Example



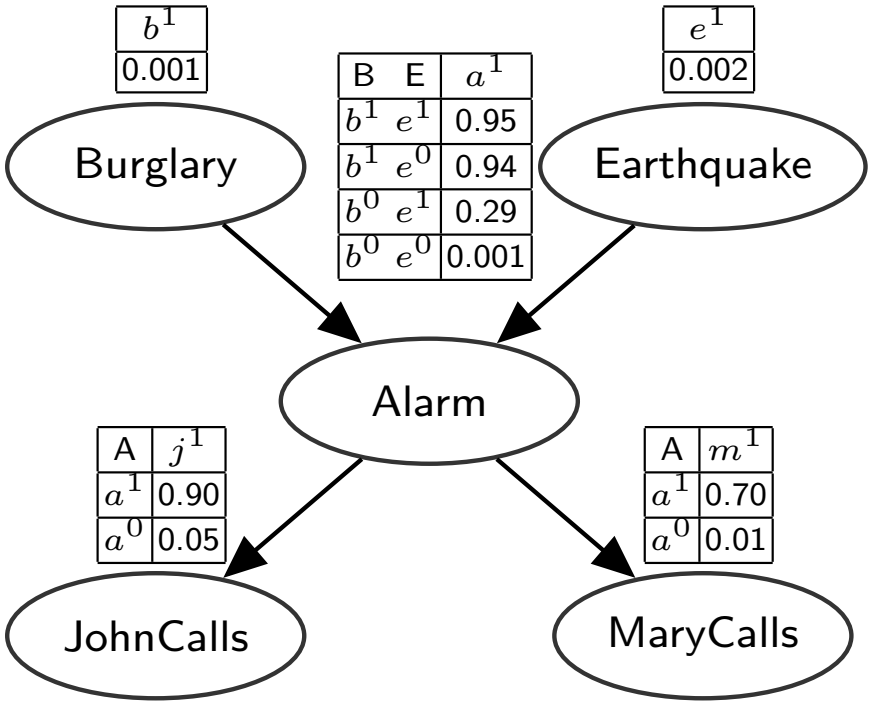
t	B	E	A	J	M
0	F	F	F	F	F
1	F	T	F		
2					
3					
4					

- Sample $P(A \mid B, E, J, M)$ at $t = 1$ via Bayes rule

$$P(A \mid B, E, J, M) \propto P(J \mid A)P(M \mid A)P(A \mid B, E)$$

- $(B, E, J, M) = (\mathsf{F}, \mathsf{T}, \mathsf{F}, \mathsf{F})$:
  - $P(A = \mathsf{T} \mid B = \mathsf{F}, E = \mathsf{T}, J = \mathsf{F}, M = \mathsf{F}) \propto 0.1 \cdot 0.3 \cdot 0.29 = 0.0087$
  - $P(A = \mathsf{F} \mid B = \mathsf{F}, E = \mathsf{T}, J = \mathsf{F}, M = \mathsf{F}) \propto 0.95 \cdot 0.99 \cdot 0.71 = 0.6678$

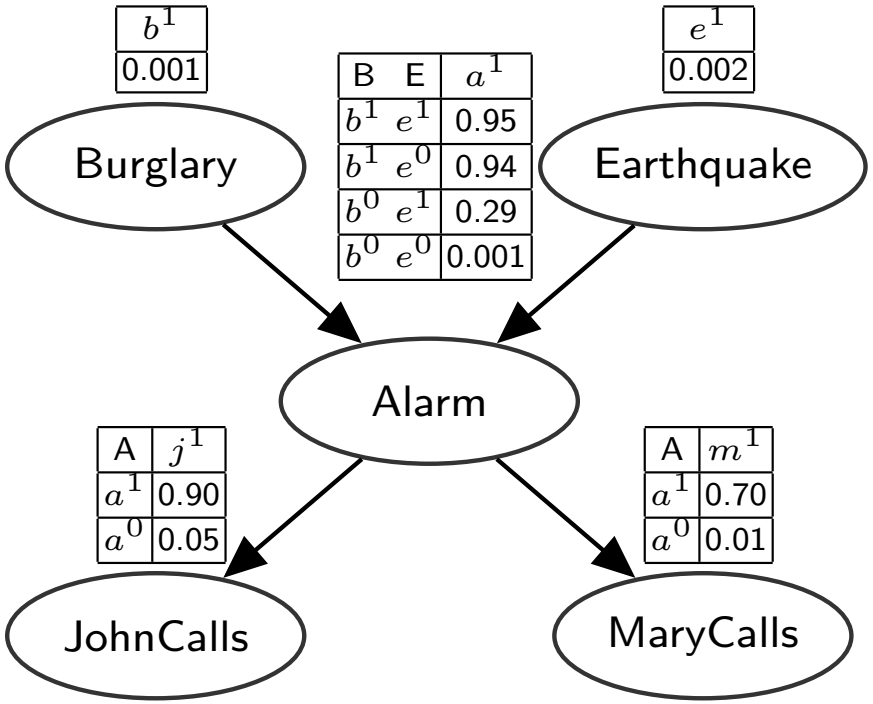
Gibbs Sampling: Example



t	B	E	A	J	M
0	F	F	F	F	F
1	F	T	F	T	
2					
3					
4					

- Sample $P(J \mid A)$ at $t = 1$ directly (no need for Bayes rule)
- $A = \mathsf{F}$:
  - $P(J = \mathsf{T} \mid A = \mathsf{F}) = 0.05$
  - $P(J = \mathsf{F} \mid A = \mathsf{F}) = 0.95$

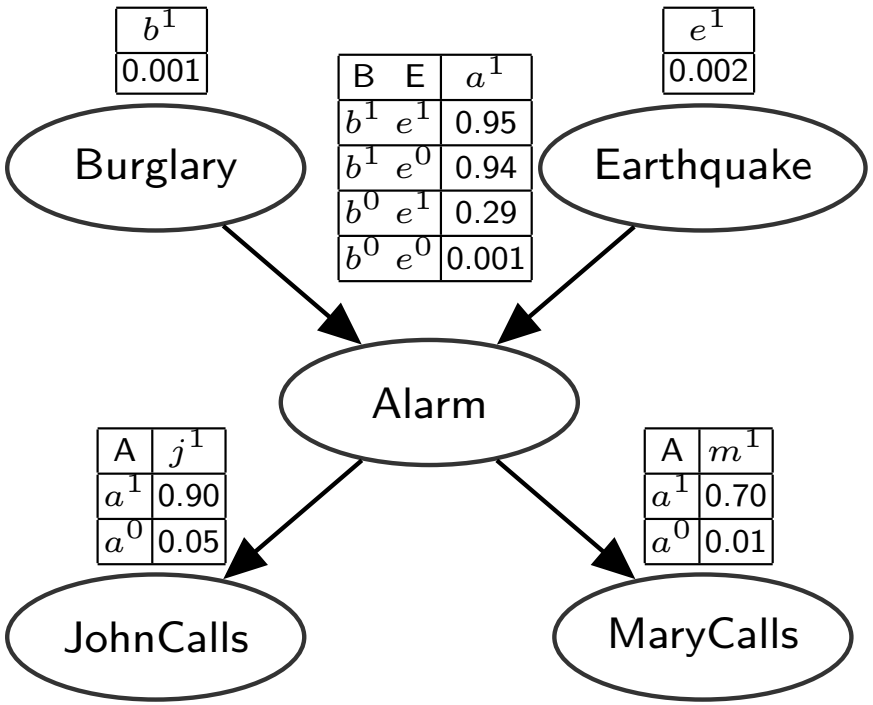
Gibbs Sampling: Example



t	B	E	A	J	M
0	F	F	F	F	F
1	F	T	F	T	F
2					
3					
4					

- Sample $P(M \mid A)$ at $t = 1$ directly (no need for Bayes rule)
- $A = \mathsf{F}$:
  - $P(M = \mathsf{T} \mid A = \mathsf{F}) = 0.01$
  - $P(M = \mathsf{F} \mid A = \mathsf{F}) = 0.99$

Gibbs Sampling: Example



t	B	E	A	J	M
0	F	F	F	F	F
1	F	T	F	T	F
2	F	T	T	T	T
3	T	F	T	F	T
4	T	F	T	F	F

- For $t = 2, 3, \ldots$, we repeat this process

Markov Chains

- A **Markov chain** is defined via a *state space* $\text{Val}(\boldsymbol{X})$ and a *transition model* $\mathcal{T}(\boldsymbol{x} \to \boldsymbol{x}')$ specifying probability of going from $\boldsymbol{x}$ to $\boldsymbol{x}'$
- Defined in terms of a graph of states
  - This graph is different from the original graphical model
  - Nodes are possible assignments to $\boldsymbol{X}$
- Sampler takes a *random walk* through the states

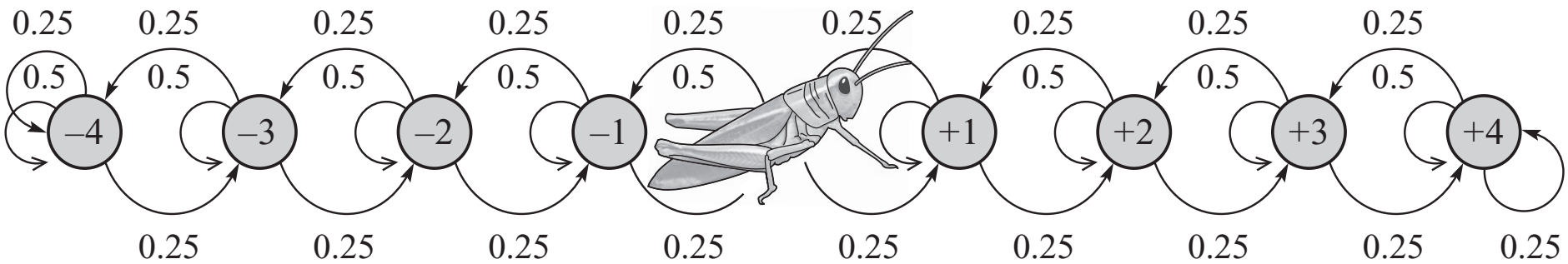


Figure: Markov chain where states are $x \in \{-4, -3, \ldots, 3, 4\}$.

Markov Chains

- Random samples dictated by chain dynamics

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

- Asymptotic behavior: What does the long-term distribution $P^{(t)}$ for $t \gg 1$ look like? (Grasshopper example converges to uniform)
- Asymptotic behavior: We want $P^{(t)}$ to converge to P

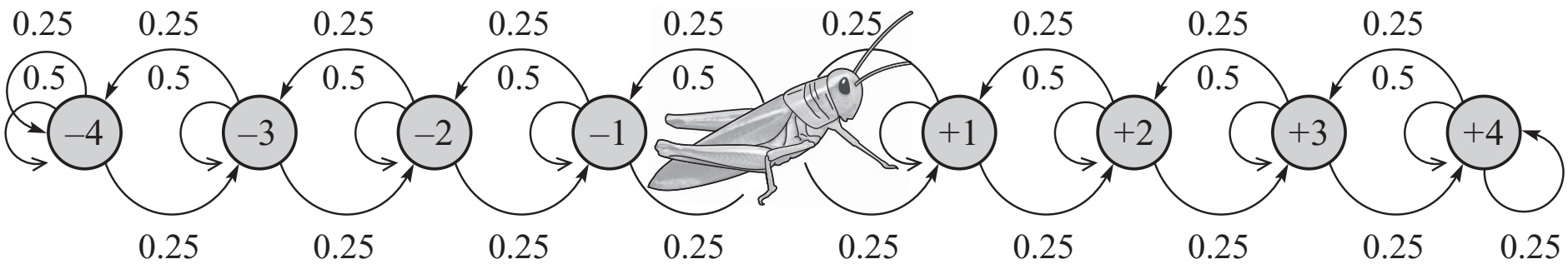


Figure: Markov chain where states are $x \in \{-4, -3, \ldots, 3, 4\}$.

Markov Chains: Asymptotic Behavior

- Distribution $\pi(\boldsymbol{X})$ is a **stationary distribution** for Markov chain $\mathcal{T}$ if

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

  (intuitively, $P^{(t+1)}$ is close to $P^{(t)}$)
- If $A_{i,j} = \mathcal{T}(\boldsymbol{x} \to \boldsymbol{x}')$, then the stationary distribution is an eigenvector of A with eigenvalue 1

Markov Chains: Asymptotic Behavior

- Distribution $\pi(\boldsymbol{X})$ is a **stationary distribution** for Markov chain $\mathcal{T}$ if

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

  (intuitively, $P^{(t+1)}$ is close to $P^{(t)}$)
- If $A_{i,j} = \mathcal{T}(\boldsymbol{x} \to \boldsymbol{x}')$, then the stationary distribution is an eigenvector of A with eigenvalue 1
- A *periodic Markov chain* exhibits cyclic behavior
- A *reducible Markov chain* Has a non-unique stationary distribution that depends on $P^{(0)}$ (i.e., there are sets of states that can not be reached from one another)

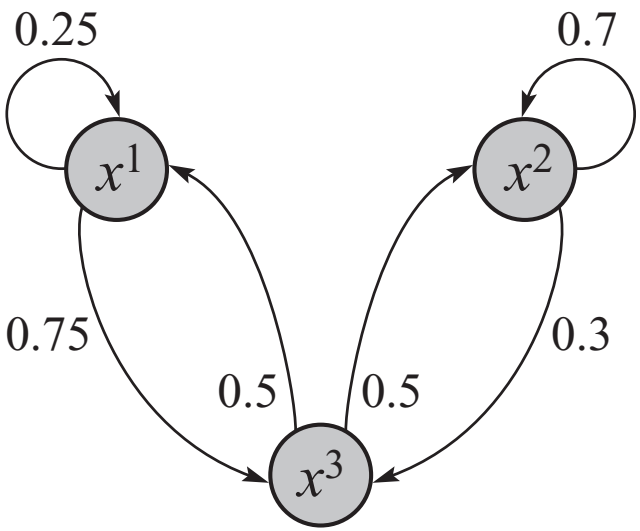
Markov Chains: Asymptotic Behavior

- Distribution $\pi(\boldsymbol{X})$ is a **stationary distribution** for Markov chain $\mathcal{T}$ if

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

- We want a Markov chain with a unique stationary distribution
- A Markov chain is *regular* (*ergodic*) if $\exists\ k$ s.t. $\forall\ \boldsymbol{x}, \boldsymbol{x}' \in \text{Val}(\boldsymbol{X})$, the probability of getting from $\boldsymbol{x}$ to $\boldsymbol{x}'$ in exactly k steps is > 0
  - Sufficient: (i) possible to get from any state to any other state via positive probability path; and (ii) self-loops
- A Markov chain that is regular has a unique stationary distribution

Markov Chains: Asymptotic Behavior



A stationary distribution π satisfies

$$\pi(x^1) = 0.25\pi(x^1) + 0.5\pi(x^3)$$
$$\pi(x^2) = 0.7\pi(x^2) + 0.5\pi(x^3)$$
$$\pi(x^3) = 0.75\pi(x^1) + 0.3\pi(x^2)$$

and

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

yielding

$$\pi(x^1) = 0.2 \quad \pi(x^2) = 0.5 \quad \pi(x^3) = 0.3$$

Markov Chains Monte Carlo (MCMC) Sampling

Markov chain dynamics give rise to the MCMC sampling process

Algorithm 12.5 Generating a Markov chain trajectory

Procedure MCMC-Sample (
  $P^{(0)}(\boldsymbol{X})$, // Initial state distribution
  $\mathcal{T}$, // Markov chain transition model
  T // Number of time steps
)

Sample $\boldsymbol{x}^{(0)}$ from $P^{(0)}(\boldsymbol{X})$
for $t = 1, \ldots, T$
  Sample $\boldsymbol{x}^{(t)}$ from $\mathcal{T}(\boldsymbol{x}^{(t-1)} \rightarrow \boldsymbol{X})$
  return $\boldsymbol{x}^{(0)}, \ldots, \boldsymbol{x}^{(T)}$

Markov Chains: Transition Models

- The state space for graphical models has a factorized structure (each state is an assignment to several variables)
- Rather than transitions between general states, we can consider transitions that update only a single state element at a time
- Gives rise to set of transition models $\{\mathcal{T}_1, \ldots \mathcal{T}_k\}$ (kernels)
- A *multi-kernel Markov chain* cycles through these kernels
- Graphical model: One $\mathcal{T}_i\big((\boldsymbol{x}_{-i}, x_i) \rightarrow (\boldsymbol{x}_{-i}, x_i')\big)$ for each X_i
- While individual kernels may not be regular, the combination may be

Gibbs Sampling and Markov Chains

- A *Gibbs chain* is defined by the following kernel

$$\mathcal{T}_i\big((\boldsymbol{x}_{-i}, x_i) \rightarrow (\boldsymbol{x}_{-i}, x_i')\big) = P(x_i' \,|\, \boldsymbol{x}_{-i})$$

- Each step of Gibbs sampling visits a new state in the chain
- Factors allow efficient calculations at each step
- Over time, sampler converges to the stationary distribution, i.e., $\pi(\boldsymbol{X}) = P(\boldsymbol{X} \,|\, \boldsymbol{e})$, where $\boldsymbol{X} = \mathcal{X} - \boldsymbol{E}$

MCMC Convergence

- It may take MCMC a long time to *mix*, i.e., converge to the stationary distribution
- Often, we throw out early samples that correspond to the "burn-in" period