

Algorithms for Reasoning with graphical models

**Slides Set 11 (part a):
Sampling Techniques for Probabilistic
and Deterministic Graphical models**

Rina Dechter

(Reading” Darwiche chapter 15, related papers)

Sampling Techniques for Probabilistic and Deterministic Graphical models

ICS 276, Spring 2018

Bozhena Bidyuk

Reading” Darwiche chapter 15, related papers

Overview

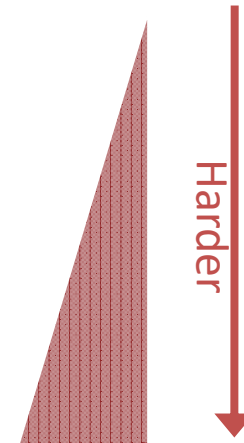
1. Basics of sampling
2. Importance Sampling
3. Markov Chain Monte Carlo: Gibbs Sampling
4. Sampling in presence of Determinism
5. Rao-Blackwellisation, cutset sampling

Overview

1. Basics of sampling
2. Importance Sampling
3. Markov Chain Monte Carlo: Gibbs Sampling
4. Sampling in presence of Determinism
5. Rao-Blackwellisation, cutset sampling

Types of queries

▶ Max-Inference	$f(\mathbf{x}^*) = \max_{\mathbf{x}} \prod_{\alpha} f_{\alpha}(\mathbf{x}_{\alpha})$
▶ Sum-Inference	$Z = \sum_{\mathbf{x}} \prod_{\alpha} f_{\alpha}(\mathbf{x}_{\alpha})$
▶ Mixed-Inference	$f(\mathbf{x}_M^*) = \max_{\mathbf{x}_M} \sum_{\mathbf{x}_S} \prod_{\alpha} f_{\alpha}(\mathbf{x}_{\alpha})$



- **NP-hard**: exponentially many terms
- We will focus on **approximation** algorithms
 - **Anytime**: very fast & very approximate ! Slower & more accurate

Monte Carlo estimators

- Most basic form: empirical estimate of probability

$$\mathbb{E}[u(x)] = \int p(x)u(x) \approx U = \frac{1}{m} \sum_i u(\tilde{x}^{(i)}) \quad \tilde{x}^{(i)} \sim p(x)$$

- Relevant considerations

- Able to sample from the target distribution $p(x)$?
- Able to evaluate $p(x)$ explicitly, or only up to a constant?

- “Any-time” properties $\mathbb{E}[U] = \mathbb{E}[u(x)]$

- Unbiased estimator, $\mathbb{E}[U] \rightarrow \mathbb{E}[u(x)]$ as $m \rightarrow \infty$
or asymptotically unbiased,

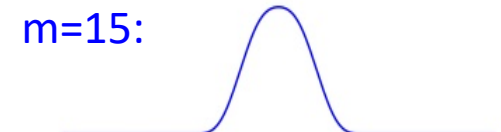
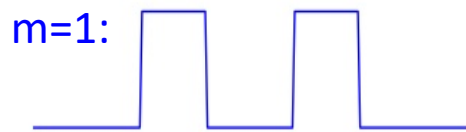
- Variance of the estimator decreases with m

Monte Carlo estimators

- Most basic form: empirical estimate of probability

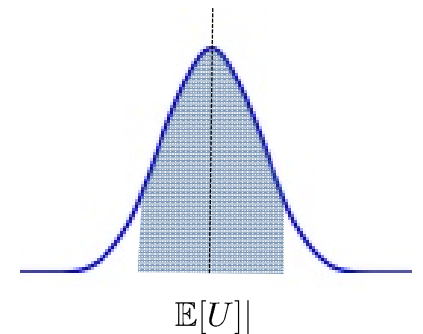
$$\mathbb{E}[u(x)] = \int p(x)u(x) \approx U = \frac{1}{m} \sum_i u(\tilde{x}^{(i)}) \quad \tilde{x}^{(i)} \sim p(x)$$

- Central limit theorem
 - $p(U)$ is asymptotically Gaussian:



- **Finite sample confidence intervals**
 - If $u(x)$ or its variance are bounded, e.g., $u(x^{(i)}) \in [0, 1]$ probability concentrates rapidly around the expectation:

$$\Pr[|U - \mathbb{E}[U]| > \epsilon] \leq O(\exp(-m\epsilon^2))$$



Estimating an Expectation: Monte Carlo Simulation

Since the sample mean is a function of the sample space, it has its own expectation and variance

Let $A_{V_n}(f)$ be a sample mean, where the function f has expectation μ and variance σ^2 . The expectation of the sample mean $A_{V_n}(f)$ is μ and its variance is σ^2/n .

The estimate $A_{V_n}(f)$ is said to be **unbiased**

since the expectation of the estimate equals the quantity we are trying to estimate.

The variance of this estimate is inversely proportional to the sample size n

Estimating an Expectation: Monte Carlo Simulation

Central Limit Theorem

Let $A_{V_n}(f)$ be a sample mean, where the function f has expectation μ and variance σ^2 . As the sample size n tends to infinity, the distribution of $\sqrt{n}(A_{V_n}(f) - \mu)$ converges to a Normal with mean 0 and variance σ^2 . We say in this case that the estimate $A_{V_n}(f)$ is **asymptotically Normal**.

Continues to hold if we replace σ^2 by the **sample variance**:

$$S^2_n(f) \stackrel{\text{def}}{=} \frac{1}{n-1} \sum_{i=1}^n (f(\mathbf{x}^i) - A_{V_n}(f))^2$$

Allows us to compute confidence intervals, even when we do not know the value of variance σ^2 .

A Sample

- Given a set of variables $X=\{X_1,\dots,X_n\}$, a sample, denoted by S^t is an instantiation of all variables:

$$S^t = (x_1^t, x_2^t, \dots, x_n^t)$$

How to Draw a Sample ?

Univariate Distribution

- Example: Given random variable X having domain $\{0, 1\}$ and a distribution $P(X) = (0.3, 0.7)$.
- Task: Generate samples of X from P .
- How?
 - draw random number $r \in [0, 1]$
 - If $(r < 0.3)$ then set $X=0$
 - Else set $X=1$

How to Draw a Sample?

Multi-Variate Distribution

- Let $X=\{X_1,...,X_n\}$ be a set of variables
- Express the distribution in product form

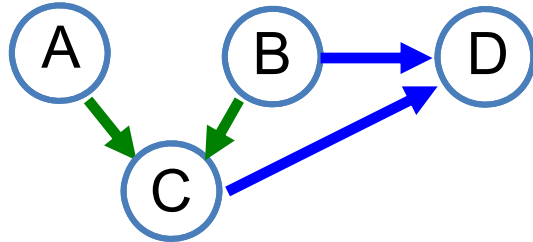
$$P(X) = P(X_1) \times P(X_2 | X_1) \times \dots \times P(X_n | X_1, \dots, X_{n-1})$$

- Sample variables one by one from left to right, along the ordering dictated by the product form.
- Bayesian network literature: *Logic sampling* or *Forward Sampling*.

Sampling in Bayes nets (Forward Sampling)

- No evidence: “causal” form makes sampling easy
 - Follow variable ordering defined by parents
 - Starting from root(s), sample downward
 - When sampling each variable, condition on values of parents

$$p(A, B, C, D) = p(A) p(B) p(C | A, B) p(D | B, C)$$



Sample:

$$a \sim p(A)$$

$$b \sim p(B)$$

$$c \sim p(C | A = a, B = b)$$

$$d \sim p(D | C = c, B = b)$$

Forward Sampling: No Evidence (Henrion 1988)

Input: Bayesian network

$X = \{X_1, \dots, X_N\}$, N - #nodes, T - # samples

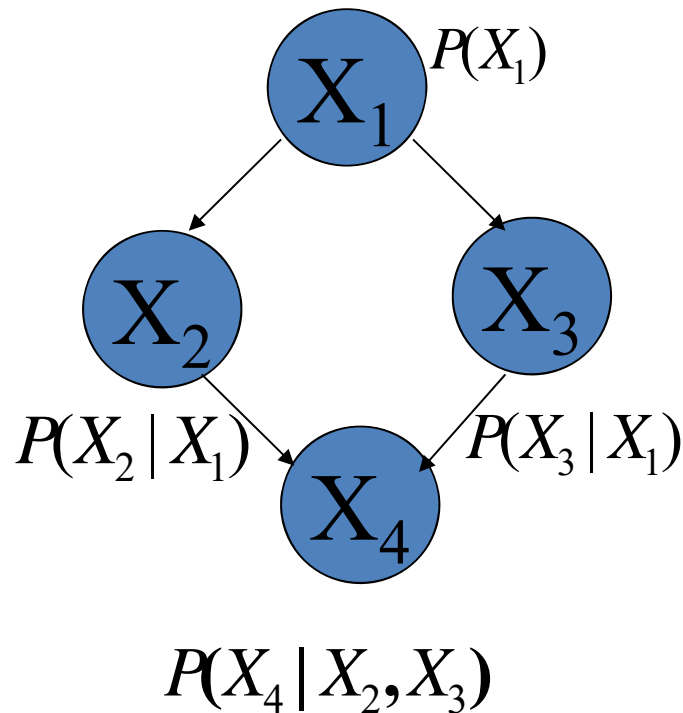
Output: T samples

Process nodes in topological order – first process the ancestors of a node, then the node itself:

1. For $t = 0$ to T
2. For $i = 0$ to N
3. $X_i \leftarrow$ sample x_i^t from $P(x_i \mid \text{pa}_i)$

Forward Sampling (example)

$$P(X_1, X_2, X_3, X_4) = P(X_1) \times P(X_2 | X_1) \times P(X_3 | X_1) \times P(X_4 | X_2, X_3)$$



No Evidence

// generate sample k

1. Sample x_1 from $P(x_1)$

2. Sample x_2 from $P(x_2 | X_1 = x_1)$

3. Sample x_3 from $P(x_3 | X_1 = x_1)$

4. Sample x_4 from $P(x_4 | X_2 = x_2, X_3 = x_3)$

Forward Sampling w/ Evidence

Input: Bayesian network

$X = \{X_1, \dots, X_N\}$, N - #nodes

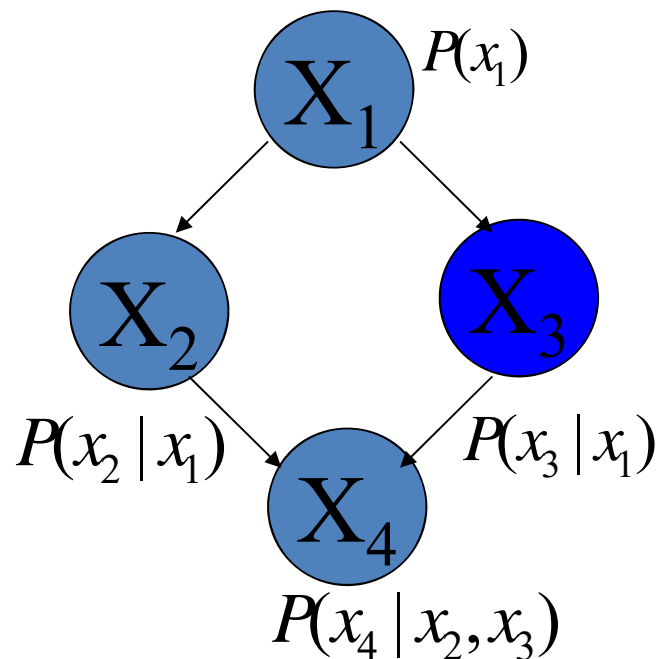
E – evidence, T - # samples

Output: T samples consistent with E

1. For $t=1$ to T
2. For $i=1$ to N
3. $X_i \leftarrow$ sample x_i^t from $P(x_i \mid \text{pa}_i)$
4. If X_i in E and $X_i \neq x_i$, reject sample:
5. Goto Step 1.

Forward Sampling (example)

Evidence : $X_3 = 0$



// generate sample k

1. Sample x_1 from $P(x_1)$
2. Sample x_2 from $P(x_2 | x_1)$
3. Sample x_3 from $P(x_3 | x_1)$
4. If $x_3 \neq 0$, reject sample and start from 1, otherwise
5. Sample x_4 from $P(x_4 | x_2, x_3)$

How to answer queries with sampling?

Expected value and Variance

Many queries can be phrased as computing expectation of some functions

Expected value: Given a probability distribution $P(X)$ and a function $g(X)$ defined over a set of variables $X = \{X_1, X_2, \dots, X_n\}$, the expected value of g w.r.t. P is

$$E_P[g(x)] = \sum_x g(x)P(x)$$

Variance: The variance of g w.r.t. P is:

$$Var_P[g(x)] = \sum_x [g(x) - E_P[g(x)]]^2 P(x)$$

Monte Carlo Estimate

- **Estimator:**

- An estimator is a function of the samples.
- It produces an estimate of the unknown parameter of the sampling *distribution*.

Given i.i.d. samples S^1, S^2, \dots, S^T drawn from P ,
the Monte carlo estimate of $E_P[g(x)]$ is given by :

$$\hat{g} = \frac{1}{T} \sum_{t=1}^T g(S^t)$$

Example: Monte Carlo estimate

- Given:
 - A distribution $P(X) = (0.3, 0.7)$.
 - $g(X) = 40$ if X equals 0
= 50 if X equals 1.
- Estimate $E_p[g(x)] = (40 \times 0.3 + 50 \times 0.7) = 47$.
- Generate k samples from P : 0,1,1,1,0,1,1,0,1,0

$$\begin{aligned}\hat{g} &= \frac{40 \times \# \text{ samples}(X = 0) + 50 \times \# \text{ samples}(X = 1)}{\# \text{ samples}} \\ &= \frac{40 \times 4 + 50 \times 6}{10} = 46\end{aligned}$$

Bayes Nets with Evidence

- Estimating posterior probabilities, $P[A = a \mid E=e]$?
- Rejection sampling
 - Draw $x \sim p(x)$, but discard if $E \neq e$
 - Resulting samples are from $p(x \mid E=e)$; use as before
 - Problem: keeps only $P[E=e]$ fraction of the samples!
 - Performs poorly when evidence probability is small
- Estimate the ratio: $P[A=a, E=e] / P[E=e]$
 - Two estimates (numerator & denominator)
 - Good finite **sample bounds** require low *relative* error!
 - Again, performs poorly when evidence probability is small
 - **What bounds can we get?**

Bounds on the Absolute Error

The **absolute error** of an estimate $A_{V_n}(\alpha)$

is the absolute difference it has with the true probability $\Pr(\alpha)$ we are trying to estimate.

For any $\epsilon > 0$, we have

$$\mathbb{P}\left(|A_{V_n}(\alpha) - \Pr(\alpha)| < \epsilon\right) \geq 1 - \frac{\Pr(\alpha)\Pr(\neg\alpha)}{n\epsilon^2}$$

The estimate $A_{V_n}(\alpha)$ computed by direct sampling will fall within the interval $(\Pr(\alpha) - \epsilon, \Pr(\alpha) + \epsilon)$ with probability at least $1 - \Pr(\alpha)\Pr(\neg\alpha)/n\epsilon^2$

Bounds on the Absolute Error

A sharper bound which does not depend on the probability $\Pr(\alpha)$

Hoeffding's inequality

Let $A_{V_n}(f)$ be a sample mean, where the function f has expectation μ and values in $\{0, 1\}$. For any $\epsilon > 0$, we have:

$$\mathbb{P}\left(|A_{V_n}(f) - \mu| \leq \epsilon\right) \geq 1 - 2e^{-2n\epsilon^2}$$

For any $\epsilon > 0$, we have:

$$\mathbb{P}\left(|A_{V_n}(\alpha) - \Pr(\alpha)| \leq \epsilon\right) \geq 1 - 2e^{-2n\epsilon^2}$$

The estimate $A_{V_n}(\alpha)$ computed by direct sampling will fall within the interval $(\Pr(\alpha) - \epsilon, \Pr(\alpha) + \epsilon)$ with probability at least $1 - 2e^{-2n\epsilon^2}$

Bounds on the Relative Error

For any $\epsilon > 0$, we have:

$$\mathbb{P}\left(\frac{|A_{V_n}(\alpha) - \Pr(\alpha)|}{\Pr(\alpha)} \leq \epsilon\right) \geq 1 - 2e^{-2n\epsilon^2\Pr(\alpha)^2}$$

Require the probability $\Pr(\alpha)$ (or some lower bound on it).

Bounds on the Relative Error

The **relative error** of an estimate $A_{V_n}(\alpha)$

$$\frac{|A_{V_n}(\alpha) - \Pr(\alpha)|}{\Pr(\alpha)}$$

The bound on the absolute error becomes tighter as the probability of an event becomes more extreme. Yet, the corresponding bound on the relative error becomes looser as the probability of an event becomes more extreme.

Example

For an event with probability .5 and a sample size of 10000, there is a 95% chance that the **absolute** error is $\approx 4.5\%$. However, for the same confidence level, the relative error increases to $\approx 13.4\%$ if the event has probability .1, and increases again to $\approx 44.5\%$ if the event has probability .01

Bayes Nets With Evidence

- Estimating the probability of evidence, $P[E=e]$:

$$P[E = e] = \mathbb{E}[\mathbb{1}[E = e]] \approx U = \frac{1}{m} \sum_i \mathbb{1}[\tilde{e}^{(i)} = e]$$

- Finite sample bounds: $u(x) \in [0,1]$ [e.g., Hoeffding]

$$\Pr[|U - \mathbb{E}[U]| > \epsilon] \leq 2 \exp(-2m\epsilon^2)$$

What if the evidence is unlikely? $P[E=e]=1e-6$) could estimate $U = 0$!

- Relative error bounds [Dagum & Luby 1997]

$$\Pr\left[\frac{|U - \mathbb{E}[U]|}{\mathbb{E}[U]} > \epsilon\right] \leq \delta \quad \text{if} \quad m \geq \frac{4}{\mathbb{E}[U]\epsilon^2} \log \frac{2}{\delta}$$

So, if U , the probability of evidence is very small we would need many many samples that are not rejected.

Overview

1. Basics of sampling
2. Importance Sampling
3. Markov Chain Monte Carlo: Gibbs Sampling
4. Sampling in presence of Determinism
5. Rao-Blackwellisation, cutset sampling

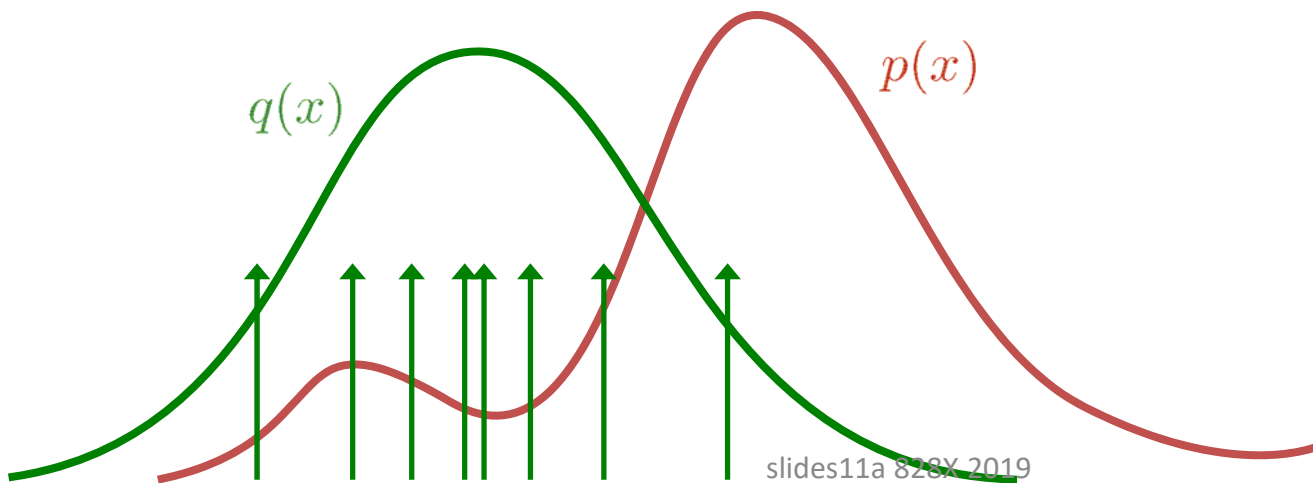
Importance Sampling: Main Idea

- Express query as the expected value of a random variable w.r.t. to a distribution Q .
- Generate random samples from Q .
- Estimate the expected value from the generated samples using a monte carlo estimator (average).

Importance Sampling

$$\bullet \mathbb{E}[u(x)] = \int p(x)u(x) \approx \hat{u} = \frac{1}{m} \sum_i u(\tilde{x}^{(i)}) \quad \tilde{x}^{(i)} \sim p(x)$$

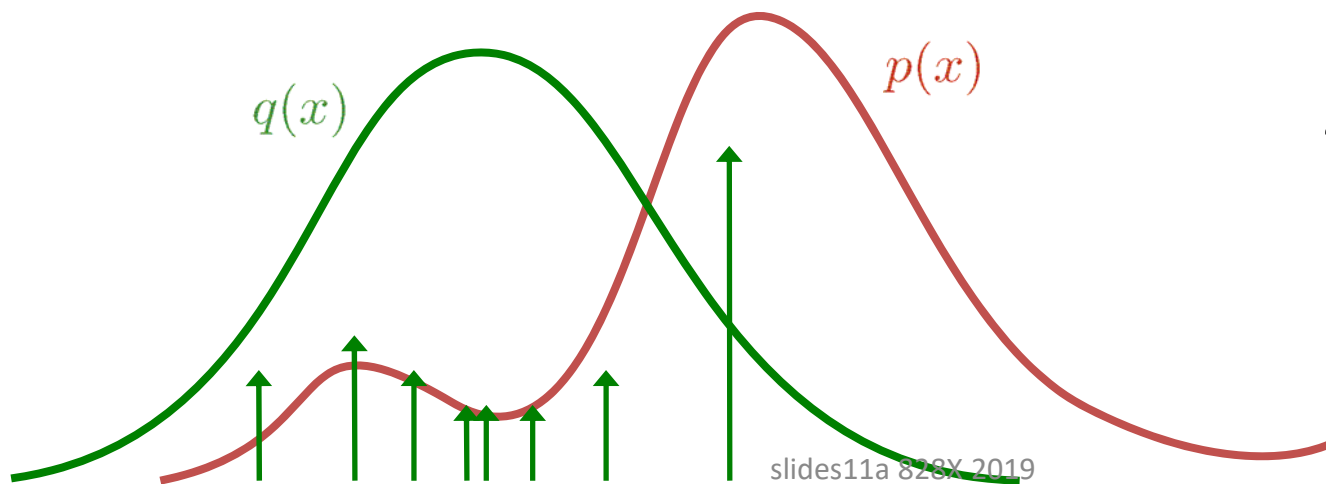
$$\int p(x)u(x) = \int q(x) \frac{p(x)}{q(x)} u(x) \approx \frac{1}{m} \sum_i \frac{p(\tilde{x}^{(i)})}{q(\tilde{x}^{(i)})} u(\tilde{x}^{(i)}) \quad \tilde{x}^{(i)} \sim q(x)$$



Importance Sampling

$$\bullet \mathbb{E}[u(x)] = \int p(x)u(x) \approx \hat{u} = \frac{1}{m} \sum_i u(\tilde{x}^{(i)}) \quad \tilde{x}^{(i)} \sim p(x)$$

$$\int p(x)u(x) = \int q(x) \frac{p(x)}{q(x)} u(x) \approx \frac{1}{m} \sum_i \frac{p(\tilde{x}^{(i)})}{q(\tilde{x}^{(i)})} u(\tilde{x}^{(i)}) \quad \tilde{x}^{(i)} \sim q(x)$$



“importance weights”

$$w^{(i)} = \frac{p(\tilde{x}^{(i)})}{q(\tilde{x}^{(i)})}$$

Estimating $P(E)$ and
 $P(X|e)$

Importance Sampling For $P(e)$

Let $Z = X \setminus E$,

Let $Q(Z)$ be a (proposal) distribution, satisfying

$$P(z, e) > 0 \Rightarrow Q(z) > 0$$

Then, we can rewrite $P(e)$ as :

$$P(e) = \sum_z P(z, e) = \sum_z P(z, e) \frac{Q(z)}{Q(z)} = E_Q \left[\frac{P(z, e)}{Q(z)} \right] = E_Q[w(z)]$$

Monte Carlo estimate :

$$\hat{P}(e) = \frac{1}{T} \sum_{t=1}^T w(z^t), \text{ where } z^t \leftarrow Q(Z)$$

Properties of IS Estimate of $P(e)$

- **Convergence:** by law of large numbers

$$\hat{P}(e) = \frac{1}{T} \sum_{i=1}^T w(z^i) \xrightarrow{a.s.} P(e) \text{ for } T \rightarrow \infty$$

- **Unbiased.**

$$E_Q[\hat{P}(e)] = P(e)$$

- **Variance:**

$$Var_Q[\hat{P}(e)] = Var_Q\left[\frac{1}{T} \sum_{i=1}^T w(z^i)\right] = \frac{Var_Q[w(z)]}{T}$$

Properties of IS Estimate of $P(e)$

- Mean Squared Error of the estimator

$$MSE_Q[\hat{P}(e)] = E_Q \left[\left(\hat{P}(e) - P(e) \right)^2 \right]$$

$$= \left(P(e) - E_Q[\hat{P}(e)] \right)^2 + Var_Q[\hat{P}(e)]$$

$$= Var_Q[\hat{P}(e)]$$

$$= \frac{Var_Q[w(x)]}{T}$$

This quantity enclosed in the brackets is zero because the expected value of the estimator equals the expected value of $g(x)$

Estimating $P(E)$ and $P(X|e)$

Estimating $P(X_i | e)$

Let $\delta_{x_i}(z)$ be a dirac - delta function, which is 1 if z contains x_i and 0 otherwise.

$$P(x_i | e) = \frac{P(x_i, e)}{P(e)} = \frac{\sum_z \delta_{x_i}(z) P(z, e)}{\sum_z P(z, e)} = \frac{E_Q \left[\frac{\delta_{x_i}(z) P(z, e)}{Q(z)} \right]}{E_Q \left[\frac{P(z, e)}{Q(z)} \right]}$$

Idea : Estimate numerator and denominator by IS.

$$\text{Ratio estimate : } \bar{P}(x_i | e) = \frac{\hat{P}(x_i, e)}{\hat{P}(e)} = \frac{\sum_{k=1}^T \delta_{x_i}(z^k) w(z^k, e)}{\sum_{k=1}^T w(z^k, e)}$$

Estimate is biased : $E[\bar{P}(x_i | e)] \neq P(x_i | e)$

Properties of the IS estimator for $P(X_i | e)$

- Convergence: By Weak law of large numbers

$$\bar{P}(x_i | e) \rightarrow P(x_i | e) \text{ as } T \rightarrow \infty$$

- Asymptotically unbiased

$$\lim_{T \rightarrow \infty} E_P[\bar{P}(x_i | e)] = P(x_i | e)$$

- Variance

- Harder to analyze

- Liu suggests a measure called “Effective sample size”

Effective Sample Size

$$P(x_i | e) = \sum_z g_{x_i}(z) P(z | e)$$

Given samples from $P(z | e)$, we can estimate $P(x_i | e)$ using :

$$\hat{P}(x_i | e) = \frac{1}{T} \sum_{j=1}^T g_{x_i}(z^j) \quad \longrightarrow \quad \text{Ideal estimator}$$

$$\text{Define : } ESS(Q, T) = \frac{T}{1 + \text{var}_Q[w(z)]} \quad \longrightarrow \quad \text{Measures how much the estimator deviates from the ideal one.}$$

$$\frac{\text{Var}_P[\hat{P}(x_i | e)]}{\text{Var}_Q[\bar{P}(x_i | e)]} \approx \frac{T}{ESS(Q, T)}$$

Thus T samples from P are worth $ESS(Q, T)$ samples from Q .

Therefore, the variance of the weights must be as small as possible.

Generating Samples From Q

- No restrictions on “how to”
- Typically, express Q in product form:
 - $Q(Z) = Q(Z_1) \times Q(Z_2 | Z_1) \times \dots \times Q(Z_n | Z_1, \dots, Z_{n-1})$
- Sample along the order Z_1, \dots, Z_n
- Example:
 - $Z_1 \leftarrow Q(Z_1) = (0.2, 0.8)$
 - $Z_2 \leftarrow Q(Z_2 | Z_1) = (0.1, 0.9, 0.2, 0.8)$
 - $Z_3 \leftarrow Q(Z_3 | Z_1, Z_2) = Q(Z_3) = (0.5, 0.5)$

Summary: IS for Common Queries

- Partition function

- Ex: MRF, or BN with evidence

$$Z = \sum_x f(x) = \sum_x q(x) \frac{f(x)}{q(x)} = \mathbb{E}_q \left[\frac{f(x)}{q(x)} \right] \approx \frac{1}{m} \sum w^{(i)} \quad w^{(i)} = \frac{f(\tilde{x}^{(i)})}{q(\tilde{x}^{(i)})}$$

- Unbiased; only requires evaluating unnormalized function $f(x)$

- General expectations wrt $p(x) / f(x)$?

$$\mathbb{E}_p[u(x)] = \sum_x u(x) \frac{f(x)}{Z} = \frac{\mathbb{E}_q[u(x)f(x)/q(x)]}{\mathbb{E}_q[f(x)/q(x)]} \approx \frac{\sum u(\tilde{x}^{(i)})w^{(i)}}{\sum w^{(i)}}$$

Estimate separately

Only asymptotically unbiased...

More on Properties of IS

- Importance sampling:

$$\int p(x)u(x) = \int q(x) \frac{p(x)}{q(x)} u(x) \approx \frac{1}{m} \sum_i \frac{p(\tilde{x}^{(i)})}{q(\tilde{x}^{(i)})} u(\tilde{x}^{(i)}) \quad \tilde{x}^{(i)} \sim q(x)$$

- IS is unbiased and fast if $q(\cdot)$ is easy to sample from
- IS can be lower variance if $q(\cdot)$ is chosen well
 - Ex: $q(x)$ puts more probability mass where $u(x)$ is large
 - Optimal: $q(x) \propto |u(x) p(x)|$
- IS can also give poor performance
 - If $q(x) \ll u(x) p(x)$: rare but very high weights!
 - Then, empirical variance is also unreliable!
 - For guarantees, need to analytically bound weights / variance...



How to get a good proposal?

Outline

- Definitions and Background on Statistics
- Theory of importance sampling
- **Likelihood weighting**
- State-of-the-art importance sampling techniques

Likelihood Weighting

(Fung and Chang, 1990; Shachter and Peot, 1990)

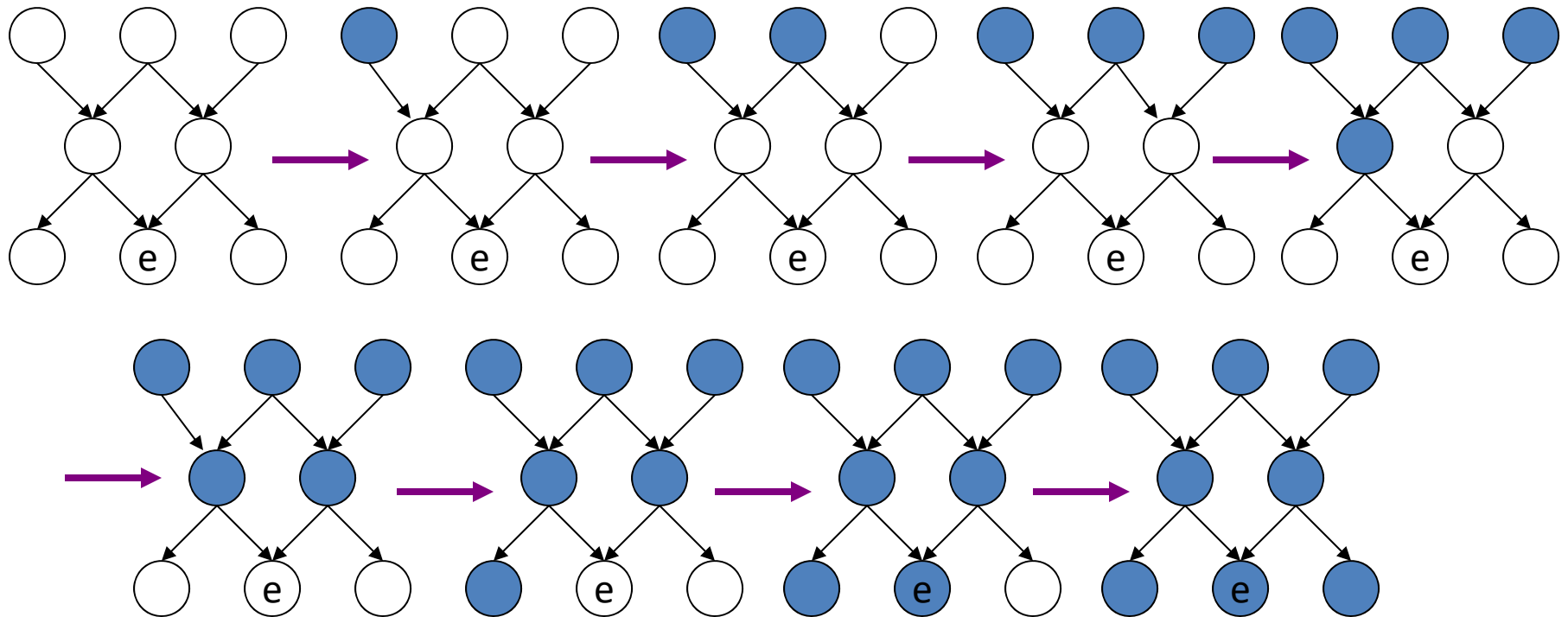
Is an instance of importance sampling!

**“Clamping” evidence+
logic sampling+
weighing samples by evidence likelihood**

Works well for *likely evidence*!

Likelihood Weighting: Sampling

Sample in topological order over **X** !



Clamp evidence, Sample $x_i \leftarrow P(X_i | pa_i)$, $P(X_i | pa_i)$ is a look-up in CPT!

Likelihood Weighting: Proposal Distribution

$$Q(X \setminus E) = \prod_{X_i \in X \setminus E} P(X_i \mid pa_i, e)$$

Notice: Q is another Bayesian network

Example:

Given a Bayesian network : $P(X_1, X_2, X_3) = P(X_1) \times P(X_2 \mid X_1) \times P(X_3 \mid X_1, X_2)$ and Evidence $X_2 = x_2$.

$$Q(X_1, X_3) = P(X_1) \times P(X_3 \mid X_1, X_2 = x_2)$$

Weights :

Given a sample : $x = (x_1, \dots, x_n)$

$$\begin{aligned} w &= \frac{P(x, e)}{Q(x)} = \frac{\prod_{X_i \in X \setminus E} P(x_i \mid pa_i, e) \times \prod_{E_j \in E} P(e_j \mid pa_j)}{\prod_{X_i \in X \setminus E} P(x_i \mid pa_i, e)} \\ &= \prod_{E_j \in E} P(e_j \mid pa_j) \end{aligned}$$

Likelihood Weighting: Estimates

Estimate $P(e)$: $\hat{P}(e) = \frac{1}{T} \sum_{t=1}^T w^{(t)}$

Estimate Posterior Marginals:

$$\hat{P}(x_i | e) = \frac{\hat{P}(x_i, e)}{\hat{P}(e)} = \frac{\sum_{t=1}^T w^{(t)} g_{x_i}(x^{(t)})}{\sum_{t=1}^T w^{(t)}}$$

$g_{x_i}(x^{(t)}) = 1$ if $x_i = x_i^t$ and equals zero otherwise

Properties of Likelihood Weighting

- Converges to exact posterior marginals
- Generates Samples Fast
- Sampling distribution is close to prior (especially if $E \subset \text{Leaf Nodes}$)
- Increasing sampling variance

⇒ Convergence may be slow

⇒ Many samples with $P(x^{(t)})=0$ rejected

Outline

- Definitions and Background on Statistics
- Theory of importance sampling
- Likelihood weighting
- **State-of-the-art importance sampling techniques**

Proposal selection

- One should try to select a proposal that is as close as possible to the posterior distribution.

$$\text{Var}_Q[\hat{P}(e)] = \frac{\text{Var}_Q[w(z)]}{T} = \frac{1}{N} \sum_{z \in Z} \left(\frac{P(z, e)}{Q(z)} - P(e) \right)^2 Q(z)$$

$$\frac{P(z, e)}{Q(z)} - P(e) = 0, \text{ to have a zero - variance estimator}$$

$$\therefore \frac{P(z, e)}{P(e)} = Q(z)$$

$$\therefore Q(z) = P(z | e)$$

Proposal Distributions used in Literature

- AIS-BN (Adaptive proposal)
 - Cheng and Druzdzel, 2000
- Iterative Belief Propagation
 - Changhe and Druzdzel, 2003
- Iterative Join Graph Propagation (IJGP) and variable ordering
 - Gogate and Dechter, 2005

Perfect sampling using Bucket Elimination

- Algorithm:
 - Run Bucket elimination on the problem along an ordering $o=(X_N, \dots, X_1)$.
 - Sample along the reverse ordering: (X_1, \dots, X_N)
 - At each variable X_i , recover the probability $P(X_i | x_1, \dots, x_{i-1})$ by referring to the bucket.

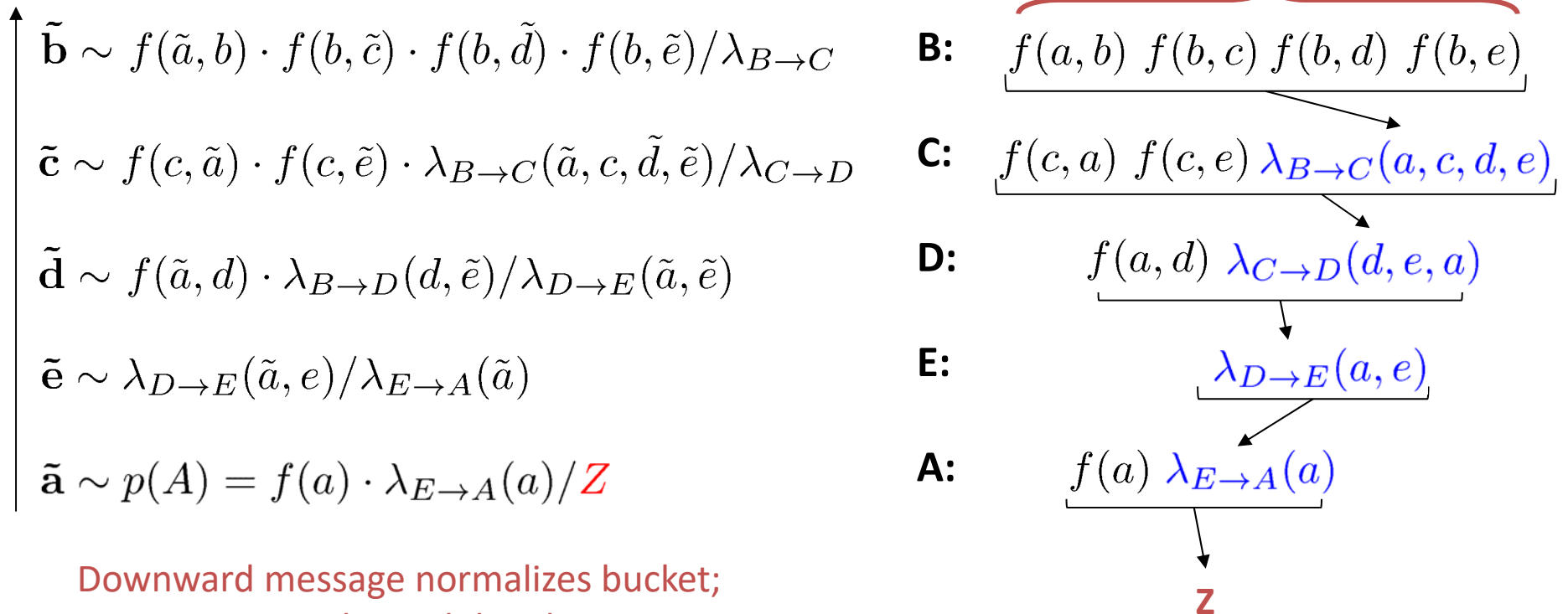
Exact Sampling using Bucket Elimination

- Algorithm:
 - Run Bucket elimination on the problem along an ordering $o=(X_1,\dots,X_N)$.
 - Sample along the reverse ordering
 - At each branch point, recover the edge probabilities by performing a constant-time table lookup!
 - Complexity: $O(\text{Bucket-elimination})+O(M*n)$
 - M is the number of solution samples and n is the number of variables

How to sample from a Markov network?

Exact sampling via inference

- Draw samples from $P[A | E=e]$ directly?
 - Model defines un-normalized $p(A, \dots, E=e)$
 - Build (oriented) tree decomposition & sample

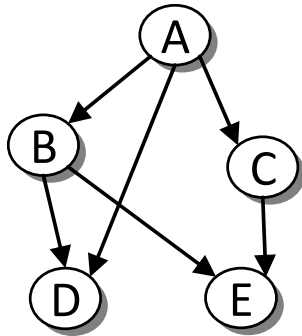


Downward message normalizes bucket;
ratio is a conditional distribution

Work: $O(\exp(w))$ to build distribution

$O(n d)$ to draw each sample

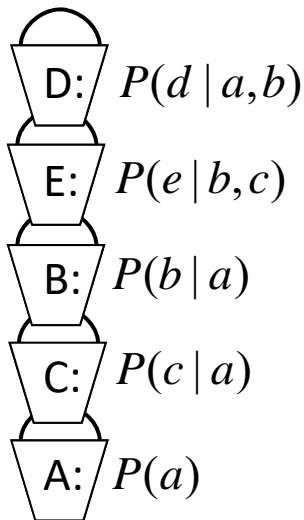
Bucket Elimination



Query: $P(a | e = 0) \propto P(a, e = 0)$ Elimination Order: d,e,b,c

$$\begin{aligned}
 P(a, e = 0) &= \sum_{c,b,e=0,d} P(a)P(b|a)P(c|a)P(d|a,b)P(e|b,c) \\
 &= P(a) \sum_c P(c|a) \sum_b P(b|a) \sum_{e=0} P(e|b,c) \sum_d P(d|a,b)
 \end{aligned}$$

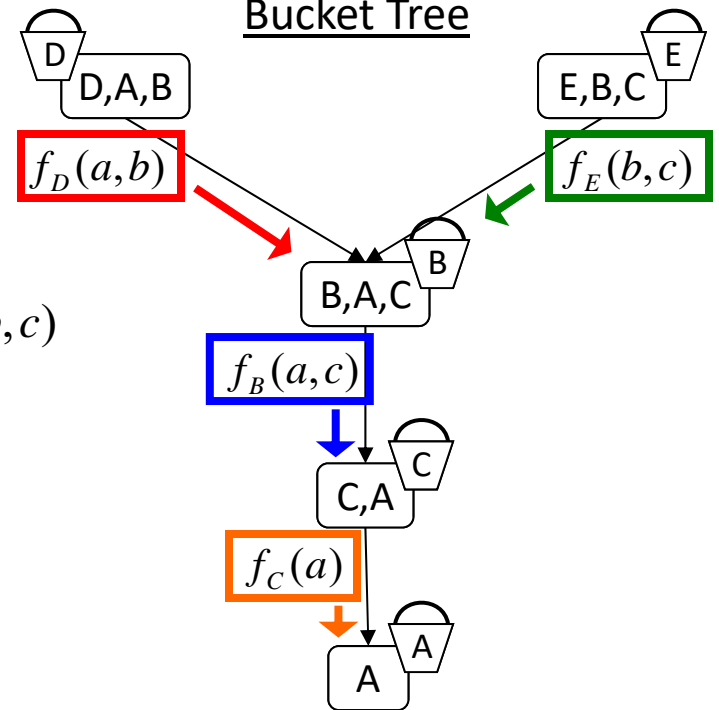
Original Functions

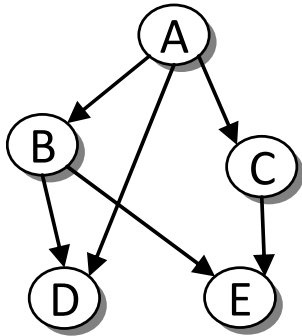


Messages

$$\begin{aligned}
 f_D(a, b) &= \sum_d P(d | a, b) \\
 f_E(b, c) &= P(e = 0 | b, c) \\
 f_B(a, c) &= \sum_b P(b | a) f_D(a, b) f_E(b, c) \\
 f_C(a) &= \sum_c P(c | a) f_B(a, c) \\
 P(a, e = 0) &= p(A) f_C(a)
 \end{aligned}$$

Bucket Tree





Bucket elimination (BE)

$$\sum_b \prod$$

Elimination operator

bucket B:

$$P(B|A) \quad P(D|B,A) \quad P(e|B,C)$$

bucket C:

$$P(C|A) \quad h^B(A, D, C, e)$$

bucket D:

$$h^C(A, D, e)$$

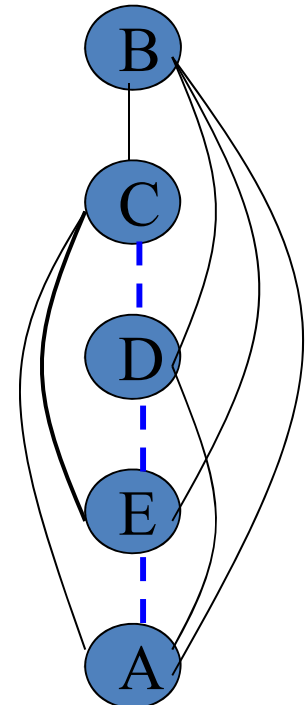
bucket E:

$$h^D(A, e)$$

bucket A:

$$P(a) \quad h^E(a)$$

$$P(e)$$



Sampling from the output of BE

(Dechter 2002)

Set $A = a, D = d, C = c$ in the bucket

Sample : $B = b \leftarrow Q(B \mid a, e, d) \propto P(B \mid a)P(d \mid B, a)P(e \mid b, c)$

bucket B: $P(B \mid A) \ P(D \mid B, A) \ P(e \mid B, C)$

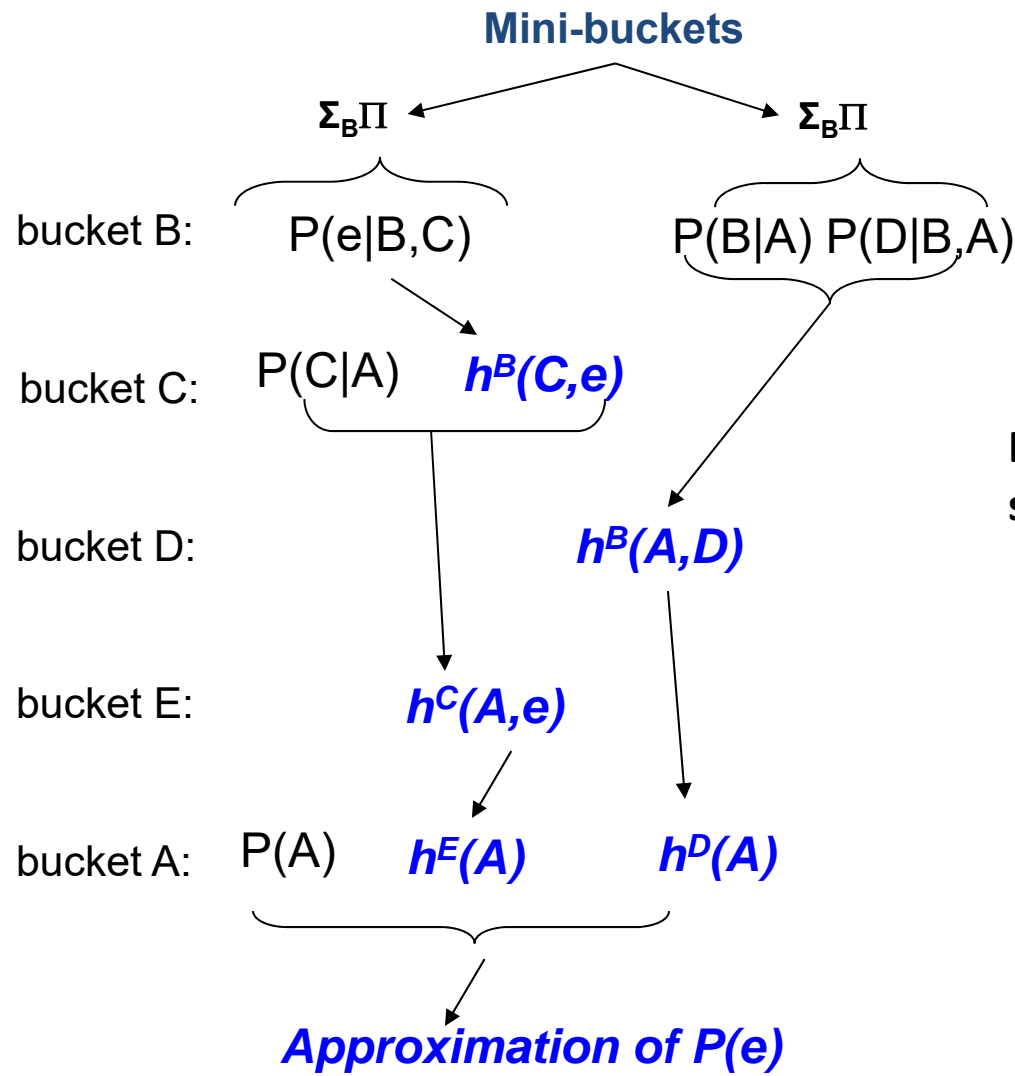
bucket C: $P(C \mid A) \ h^B(A, D, C, e)$ Set $A = a, D = d$ in the bucket
 Sample : $C = c \leftarrow Q(C \mid a, e, d) \propto P(C \mid A) \cdot h^B(a, d, C, e)$

bucket D: $h^C(A, D, e)$ Set $A = a$ in the bucket
 Sample : $D = d \leftarrow Q(D \mid a, e) \propto h^C(a, D, e)$

bucket E: $h^D(A, e)$ Evidence bucket : ignore

bucket A: $P(A) \ h^E(A) \quad Q(A) \propto P(A) \times h^E(A)$
 Sample : $A = a \leftarrow Q(A)$

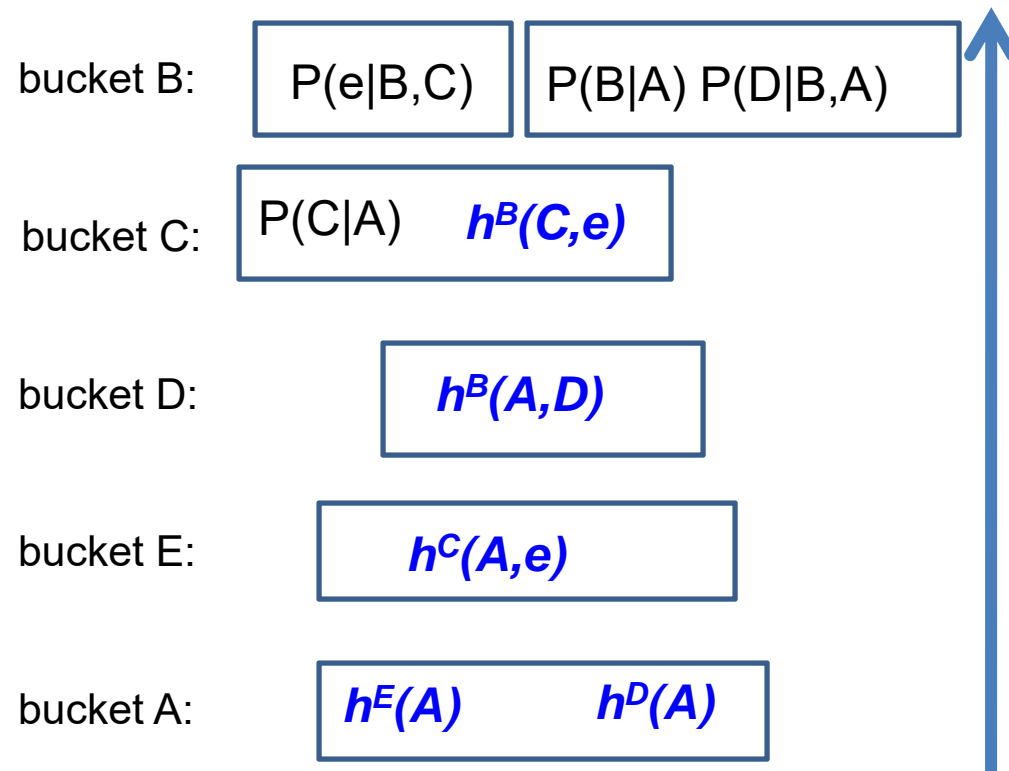
Mini-Bucket Elimination



Space and Time constraints:
Maximum scope size of the new function generated should be bounded by 2

BE generates a function having scope size 3. So it cannot be used.

Sampling from the output of MBE



Sampling is same as in BE-sampling except that now we construct Q from a randomly selected “mini-bucket”

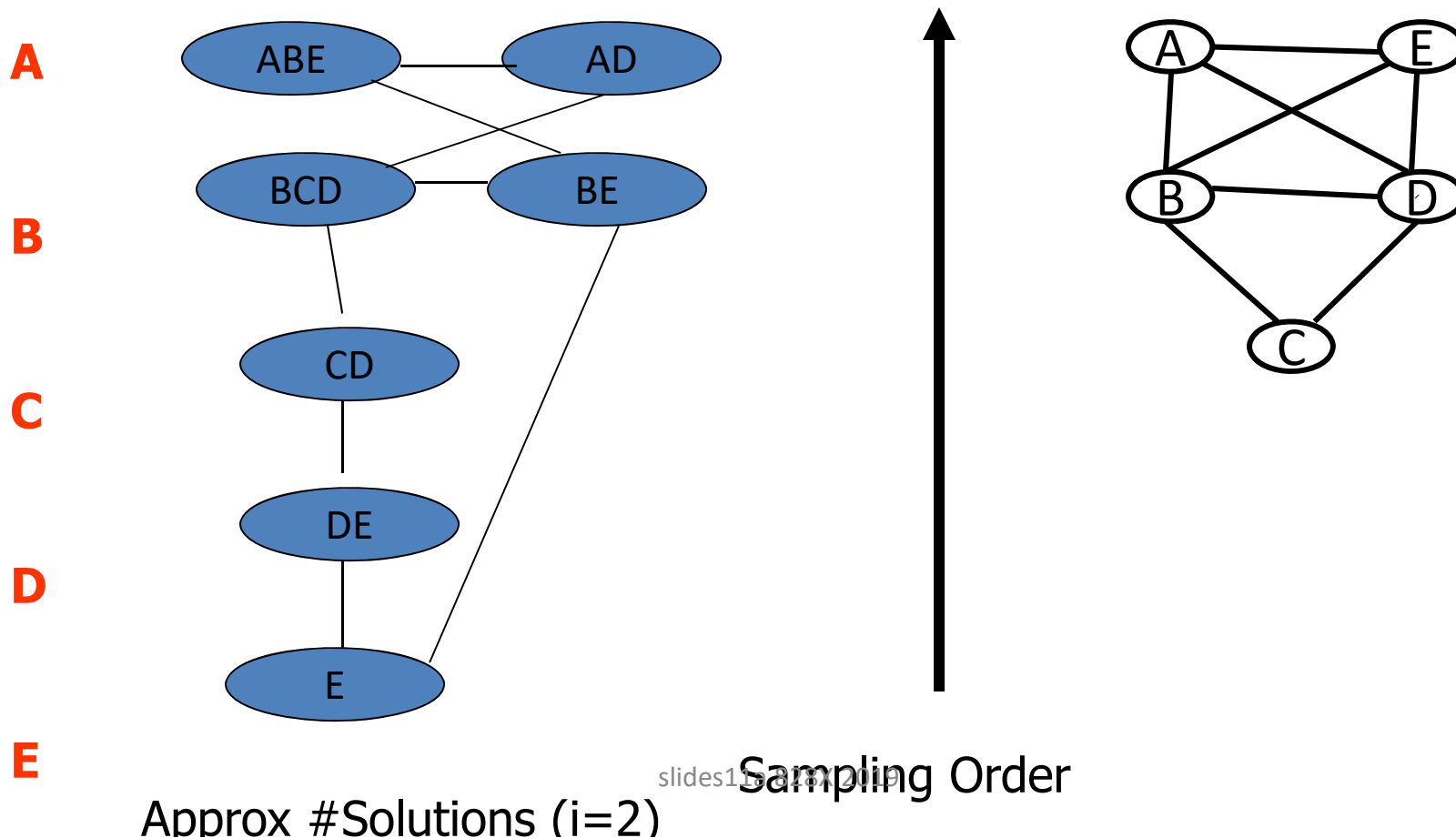
IJGP-Sampling

(Gogate and Dechter, 2005)

- Iterative Join Graph Propagation (IJGP)
 - A Generalized Belief Propagation scheme (Yedidia et al., 2002)
- IJGP yields better approximations of $P(X|E)$ than MBE (Dechter, Kask and Mateescu, 2002)
- Output of IJGP is same as mini-bucket “clusters”
- **Currently one of the best performing IS scheme!**

Example: IJGP-Sampling

- Run IJGP



Current Research Question

- Given a Bayesian network with evidence or a Markov network representing function P , generate another Bayesian network representing a function Q (from a family of distributions, restricted by structure) such that Q is closest to P .
- Current approaches
 - Mini-buckets
 - Ijgp
 - Both
- Experimented, but need to be justified theoretically.

Algorithm: Approximate Sampling

- 1) Run IJGP or MBE
 - 2) At each branch point compute the edge probabilities by consulting output of IJGP or MBE
- Rejection Problem:
 - Some assignments generated are non solutions

Adaptive Importance Sampling

Initial Proposal = $Q^1(Z) = Q(Z_1) \times Q(Z_2 \mid pa(Z_2)) \times \dots \times Q(Z_n \mid pa(Z_n))$

$$\hat{P}(E = e) = 0$$

For $i = 1$ to k do

Generate samples z^1, \dots, z^N from Q^k

$$\hat{P}(E = e) = \hat{P}(E = e) + \frac{1}{N} \sum_{j=1}^N w_k(z^j)$$

$$\text{Update } Q^{k+1} = Q^k + \eta(k)[Q^k - Q']$$

End

$$\text{Return } \frac{\hat{P}(E = e)}{k}$$

Adaptive Importance Sampling

- General case
- Given k proposal distributions
- Take N samples out of each distribution
- Approximate $P(e)$

$$\hat{P}(e) = \frac{1}{k} \sum_{j=1}^k [Avg - weight - jth - proposal]$$

Estimating $Q'(z)$

$$Q'(Z) = Q'(Z_1) \times Q'(Z_2 \mid pa(Z_2)) \times \dots \times Q'(Z_n \mid pa(Z_n))$$

where each $Q'(Z_i \mid Z_1, \dots, Z_{i-1})$

is estimated by importance sampling

Choosing a proposal (wmb-IS)

- Can use WMB upper bound to define a proposal $q(x)$:

$$\tilde{\mathbf{b}} \sim w_1 q_1(b|\tilde{a}, \tilde{c}) + w_2 q_2(b|\tilde{d}, \tilde{e})$$

Weighted mixture:

use minibucket 1 with probability w_1

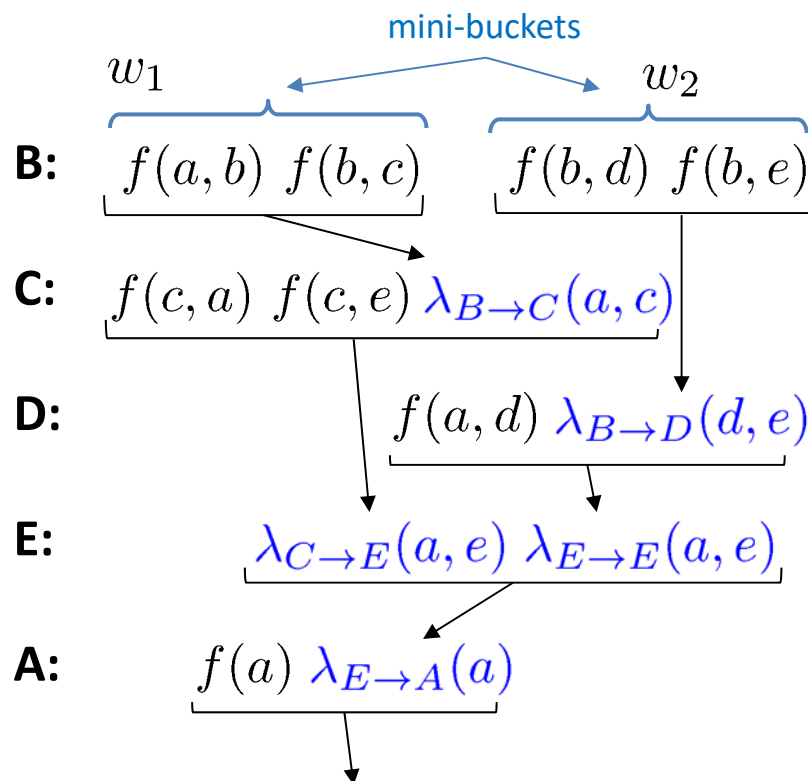
or, minibucket 2 with probability $w_2 = 1 - w_1$

where

$$q_1(b|a, c) = \left[\frac{f(a, b) \cdot f(b, c)}{\lambda_{B \rightarrow C}(a, c)} \right]^{\frac{1}{w_1}}$$

\vdots

$$\tilde{\mathbf{a}} \sim q(A) = f(a) \cdot \lambda_{E \rightarrow A}(a)/U$$



U = upper bound

Key insight: provides bounded importance weights!

$$0 \leq \frac{F(x)}{q(x)} \leq U \quad \forall x$$

WMB-IS Bounds

[Liu, Fisher, Ihler 2015]

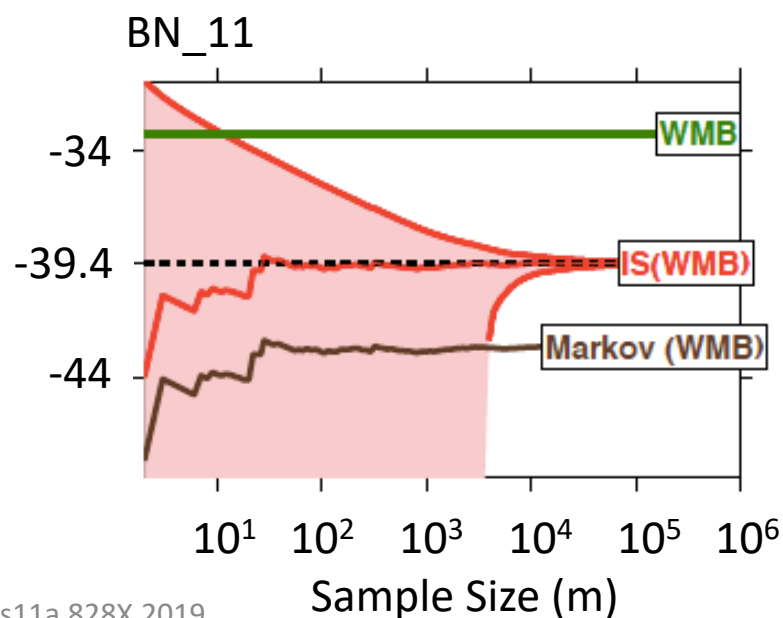
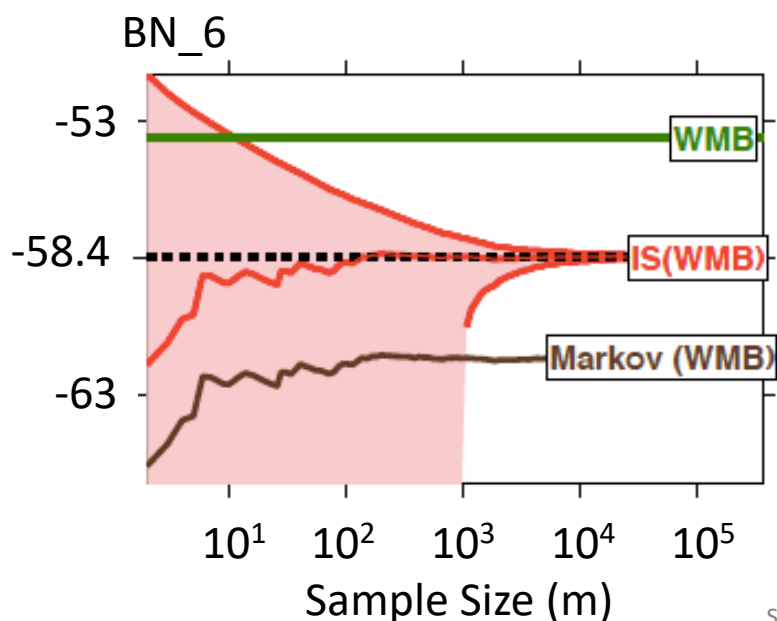
- Finite sample bounds on the average

$$\Pr\left[|\hat{Z} - Z| > \epsilon\right] \leq 1 - \delta$$

$$\epsilon = \sqrt{\frac{2\hat{V} \log(4/\delta)}{m}} + \frac{7U \log(4/\delta)}{3(m-1)}$$

“Empirical Bernstein” bounds

- Compare to forward sampling



Other Choices of Proposals

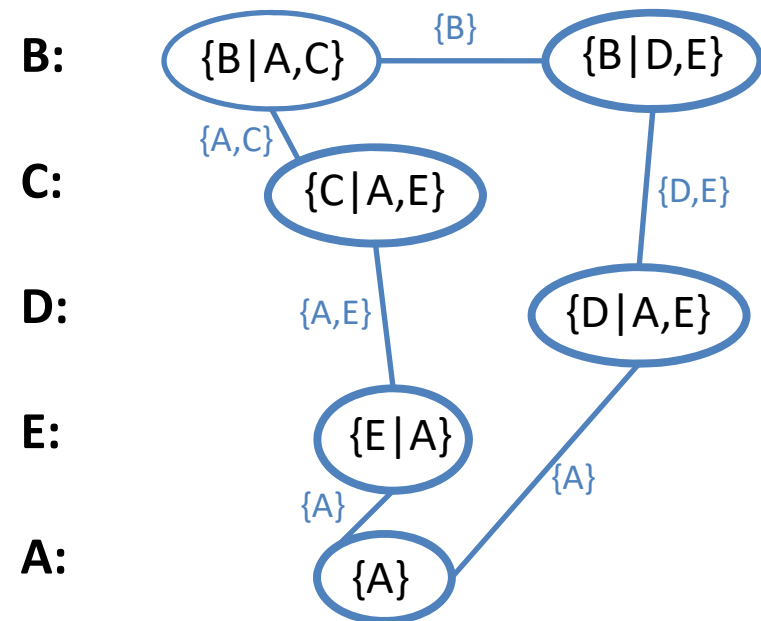
- Belief propagation
 - BP-based proposal
 - Join-graph BP proposal
 - Mean field proposal

[Changhe & Druzdzel 2003]

[Gogate & Dechter 2005]

[Wexler & Geiger 2007]

Join graph:



Other Choices of Proposals

- Belief propagation
 - BP-based proposal [Changhe & Druzdzel 2003]
 - Join-graph BP proposal [Gogate & Dechter 2005]
 - Mean field proposal [Wexler & Geiger 2007]

- Adaptive importance sampling
 - Use already-drawn samples to update $q(x)$
 - Rates v_t and η_t adapt estimates, proposal

– Ex:

[Cheng & Druzdzel 2000]

[Lapeyre & Boyd 2010]

...

– Lose “iid”-ness of samples

Adaptive IS

- 1: Initialize $q_0(x)$
 - 2: **for** $t = 0 \dots T$ **do**
 - 3: Draw $\tilde{X}_t = \{\tilde{x}^{(i)}\} \sim q_t(x)$
 - 4: $U_t = \frac{1}{m_t} \sum f(\tilde{x}^{(i)})/q_t(\tilde{x}^{(i)})$
 - 5: $\hat{U} = (1 - v_t)\hat{U} + v_t U_t$
 - 6: $q_{t+1} = (1 - \eta_t)q_t + \eta_t q^*(X_t)$
-

Overview

1. Probabilistic Reasoning/Graphical models
2. Importance Sampling
- 3. Markov Chain Monte Carlo: Gibbs Sampling**
4. Sampling in presence of Determinism
5. Rao-Blackwellisation
6. AND/OR importance sampling

Outline

- Definitions and Background on Statistics
- Theory of importance sampling
- Likelihood weighting
- **Error estimation**
- State-of-the-art importance sampling techniques

Bounds on the Absolute Error

The **absolute error** of an estimate $A_{V_n}(\alpha)$

is the absolute difference it has with the true probability $\Pr(\alpha)$ we are trying to estimate.

For any $\epsilon > 0$, we have

$$\mathbb{P}\left(|A_{V_n}(\alpha) - \Pr(\alpha)| < \epsilon\right) \geq 1 - \frac{\Pr(\alpha)\Pr(\neg\alpha)}{n\epsilon^2}$$

The estimate $A_{V_n}(\alpha)$ computed by direct sampling will fall within the interval $(\Pr(\alpha) - \epsilon, \Pr(\alpha) + \epsilon)$ with probability at least $1 - \Pr(\alpha)\Pr(\neg\alpha)/n\epsilon^2$

Bounds on the Absolute Error

A sharper bound which does not depend on the probability $\Pr(\alpha)$

Hoeffding's inequality

Let $A_{V_n}(f)$ be a sample mean, where the function f has expectation μ and values in $\{0, 1\}$. For any $\epsilon > 0$, we have:

$$\mathbb{P}\left(|A_{V_n}(f) - \mu| \leq \epsilon\right) \geq 1 - 2e^{-2n\epsilon^2}$$

For any $\epsilon > 0$, we have:

$$\mathbb{P}\left(|A_{V_n}(\alpha) - \Pr(\alpha)| \leq \epsilon\right) \geq 1 - 2e^{-2n\epsilon^2}$$

The estimate $A_{V_n}(\alpha)$ computed by direct sampling will fall within the interval $(\Pr(\alpha) - \epsilon, \Pr(\alpha) + \epsilon)$ with probability at least $1 - 2e^{-2n\epsilon^2}$

Logic Sampling –How many samples?

Theorem: Let $\pi_s(\mathbf{y})$ be the estimate of $\mathbf{P}(\mathbf{y})$ resulting from a randomly chosen sample set \mathbf{S} with \mathbf{T} samples. Then, to guarantee relative error at most ε with probability at least $1-\delta$ it is enough to have:

$$T \geq \frac{c}{P(y) \cdot \varepsilon^2} \bullet \frac{1}{\delta}$$

Derived from *Chebychev's Bound*.

$$\Pr\left(\bar{P}(y) \notin [P(y) - \varepsilon, P(y) + \varepsilon]\right) \leq 2e^{-2N\varepsilon^2}$$

Logic Sampling: Performance

Advantages:

- $P(x_i \mid \text{pa}(x_i))$ is readily available
- Samples are independent !

Drawbacks:

- If evidence **E** is rare ($P(e)$ is low), then we will reject most of the samples!
- Since **P(y)** in estimate of **T** is unknown, must estimate **P(y)** from samples themselves!
- If **P(e)** is small, **T** will become very big!

Bounds on the Relative Error

For any $\epsilon > 0$, we have:

$$\mathbb{P}\left(\frac{|A_{V_n}(\alpha) - \Pr(\alpha)|}{\Pr(\alpha)} \leq \epsilon\right) \geq 1 - 2e^{-2n\epsilon^2\Pr(\alpha)^2}$$

Require the probability $\Pr(\alpha)$ (or some lower bound on it).

Bounds on the Relative Error

The **relative error** of an estimate $A_{V_n}(\alpha)$

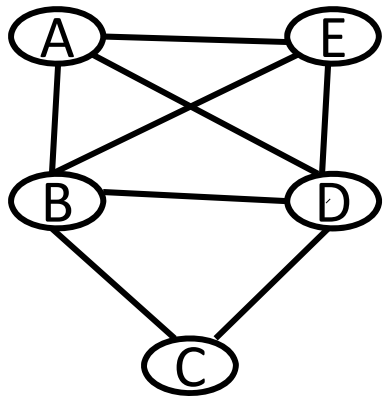
$$\frac{|A_{V_n}(\alpha) - \Pr(\alpha)|}{\Pr(\alpha)}$$

The bound on the absolute error becomes tighter as the probability of an event becomes more extreme. Yet, the corresponding bound on the relative error becomes looser as the probability of an event becomes more extreme.

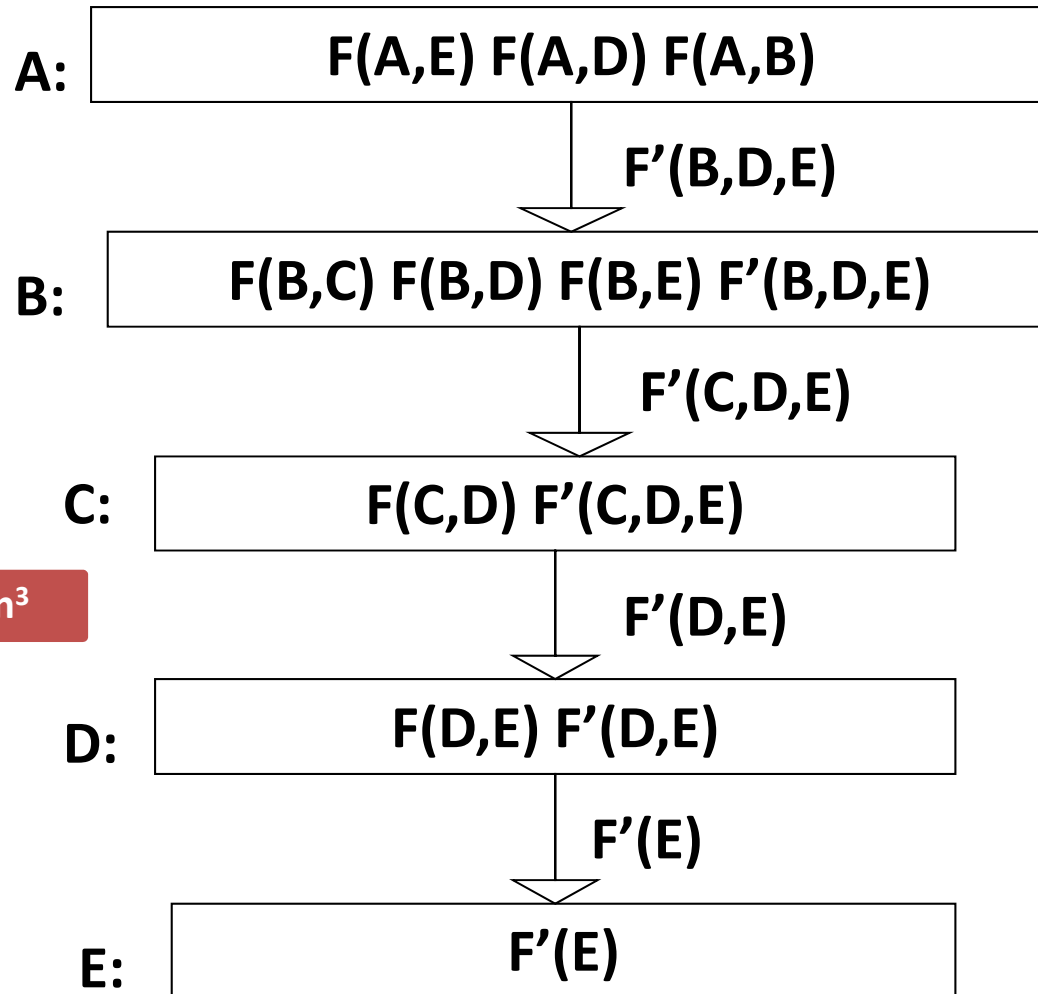
Example

For an event with probability .5 and a sample size of 10000, there is a 95% chance that the **absolute** error is $\approx 4.5\%$. However, for the same confidence level, the relative error increases to $\approx 13.4\%$ if the event has probability .1, and increases again to $\approx 44.5\%$ if the event has probability .01

Bucket Elimination Overview



Complexity: $\text{Exp}(3)$ or n^3



Sampling Direction