Reasoning with graphical models

Slides Set 10 (part a):
Sampling Techniques for Probabilistic and Deterministic Graphical models

Rina Dechter

(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**: 
  \[ f(x^*) = \max_x \prod_{\alpha} f_\alpha(x_\alpha) \]

- **Sum-Inference**: 
  \[ Z = \sum_x \prod_{\alpha} f_\alpha(x_\alpha) \]

- **Mixed-Inference**: 
  \[ f(x^*_M) = \max_{x_M} \sum_{x_S} \prod_{\alpha} f_\alpha(x_\alpha) \]

- **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,
  or asymptotically unbiased, \( \mathbb{E}[U] \to \mathbb{E}[u(x)] \) as \( m \to \infty \)

  – 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:

  \[ \begin{array}{ccc}
  m=1: & m=5: & m=15:
  \end{array} \]

• 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:

\[ \Pr\left[ |U - \mathbb{E}[U]| > \epsilon \right] \leq O\left( \exp(-m\epsilon^2) \right) \]
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 $A\bar{v}_n(f)$ be a sample mean, where the function $f$ has expectation $\mu$ and variance $\sigma^2$. The expectation of the sample mean $A\bar{v}_n(f)$ is $\mu$ and its variance is $\sigma^2/n$.

The estimate $A\bar{v}_n(f)$ 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 $n$. 
Central Limit Theorem

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

Continues to hold if we replace $\sigma^2$ by the sample variance:

$$S^2_n(f) \overset{def}{=} \frac{1}{n-1} \sum_{i=1}^{n} (f(x^i) - \text{Av}_n(f))^2$$

Allows us to compute confidence intervals, even when we do not know the value of variance $\sigma^2$. 
Given a set of variables $X=\{X_1,\ldots,X_n\}$, a sample, denoted by $S^t$ is an instantiation of all variables:

$$S^t = (x_1^t, x_2^t, \ldots, x_n^t)$$
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 $X=1$

So, how to draw a sample from a multi-dimensional distribution?
How to Draw a Sample? Multi-Variate Distribution

• Let $X = \{X_1, \ldots, X_n\}$ be a set of variables

• Express the distribution in product form

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

• 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

\[ p(A, B, C, D) = p(A) \cdot p(B) \cdot p(C \mid A, B) \cdot p(D \mid B, C) \]

Sample:
\[ a \sim p(A) \]
\[ b \sim p(B) \]
\[ c \sim p(C \mid A = a, B = b) \]
\[ d \sim p(D \mid C = c, B = b) \]
Froward Sampling: No Evidence (Henrion 1988)

Input: Bayesian network

\[ X = \{X_1, \ldots, X_N\}, \text{N-#nodes, T - # samples} \]

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. \[ X_i \leftarrow \text{sample } x_i^t \text{ from } P(x_i \mid \text{pa}_i) \]
Forward Sampling (example)

\[ P(X_1, X_2, X_3, X_4) = P(X_1) \times P(X_2 \mid X_1) \times P(X_3 \mid X_1) \times P(X_4 \mid X_2, X_3) \]

No Evidence

// generate sample \( k \)

1. Sample \( x_1 \) from \( P(x_1) \)
2. Sample \( x_2 \) from \( P(x_2 \mid X_1 = x_1) \)
3. Sample \( x_3 \) from \( P(x_3 \mid X_1 = x_1) \)
4. Sample \( x_4 \) from \( P(x_4 \mid X_2 = x_2, X_3 = x_3) \)

No evidence!
Forward Sampling w/ Evidence

Input: Bayesian network
\[ X = \{X_1, \ldots, X_N\}, \text{ N- # nodes} \]
E – evidence, T - # samples

Output: T samples consistent with E

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

\[ P(x_1) \]

\[ P(x_2 | x_1) \]

\[ P(x_3 | x_1) \]

\[ P(x_4 | x_2, x_3) \]

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 $X = \{X_1, X_2, \ldots, X_n\}$, 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:

$$\text{Var}_P[g(x)] = \sum_x \left[ g(x) - E_P[g(x)] \right]^2 P(x)$$

slides10a F2021
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 $S^1, S^2, \ldots S^T$ drawn from $P$, the Monte carlo estimate of $\mathbb{E}_P[g(x)]$ 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

\[
\hat{g} = \frac{40 \times \# \text{samples}(X = 0) + 50 \times \# \text{samples}(X = 1)}{\# \text{samples}}
\]

\[
= \frac{40 \times 4 + 50 \times 6}{10} = 46
\]
Bayes Networks with Evidence

• Estimating posterior probabilities, \( P[A = a | E=e] \)?

• Rejection sampling
  – Draw \( x \sim p(x) \), but discard if \( E \neq e \)
  – Resulting samples are from \( p(x | E=e) \); 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 \textit{sample bounds} require low \textit{relative} error!
  – Again, performs poorly when evidence probability is small
  – \textbf{What bounds can we get?}
Bayes Networks With Evidence

- Estimating the probability of evidence, $P[E=e]$ (absolute error):

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

  - Finite sample bounds: $u(x) \in [0,1]$ [e.g., Hoeffding]
    $$\Pr\left[|U - \mathbb{E}[U]| > \epsilon\right] \leq 2 \exp(-2m\epsilon^2)$$

  - 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 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

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

\[ \int p(x)u(x) = \int q(x)\frac{p(x)}{q(x)}u(x) \approx \frac{1}{m} \sum_{i} \frac{p(\tilde{x}^{(i)})}{q(\tilde{x}^{(i)})} u(\tilde{x}^{(i)}) \quad \tilde{x}^{(i)} \sim q(x) \]
Importance Sampling

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

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

“importance weights”

\[ w^{(i)} = \frac{p(\tilde{x}^{(i)})}{q(\tilde{x}^{(i)})} \]
Estimating $P(E)$ and $P(X|e)$
Importance Sampling For $P(e)$
(for discrete variables)

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

- **Unbiased.**
  \[
  E_Q[\hat{P}(e)] = P(e)
  \]

- **Variance:**
  \[
  \text{Var}_Q[\hat{P}(e)] = \text{Var}_Q\left[\frac{1}{T} \sum_{i=1}^{N} w(z^i)\right] = \frac{\text{Var}_Q[w(z)]}{T}
  \]
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 $P(X_i \mid e)$

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

\[
P(x_i \mid e) = \frac{P(x_i, e)}{P(e)} = \frac{\sum z \delta_{x_i}(z)P(z, e)}{\sum z P(z, e)} = E_Q \left[ \frac{\delta_{x_i}(z)P(z, e)}{Q(z)} \right]
\]

Idea: Estimate numerator and denominator by IS.

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

Estimate is biased: $E[\overline{P}(x_i \mid e)] \neq P(x_i \mid e)$
Properties of the IS estimator for $P(X_i | e)$

- Convergence: By Weak law of large numbers
  \[
  \bar{P}(x_i | e) \to P(x_i | e) \quad \text{as } T \to \infty
  \]

- Asymptotically unbiased
  \[
  \lim_{T \to \infty} E_P[\bar{P}(x_i | e)] = P(x_i | e)
  \]

- Variance
  - Harder to analyze
  - Liu suggests a measure called “Effective sample size”
End of class
Generating Samples From Q

• No restrictions on “how to”
• Typically, express Q in product form:
  – Q(Z)=Q(Z₁)×Q(Z₂|Z₁)×...×Q(Zₙ|Z₁,...,Zₙ₋₁)
• Sample along the order Z₁,...,Zₙ
• Example:
  – Z₁ ← Q(Z₁)=(0.2,0.8)
  – Z₂ ← Q(Z₂|Z₁)=(0.1,0.9,0.2,0.8)
  – Z₃ ← Q(Z₃|Z₁,Z₂)=Q(Z₃)=(0.5,0.5)
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 \( q(.) \) is easy to sample from

- IS can have lower variance if \( q(.) \) 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 \( q(x) << u(x) p(x) \): 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?
Likelihood Weighting
(Fung and Chang, 1990; Shachter and Peot, 1990)

Is an instance of importance sampling!

“Clamping” evidence+
logic sampling (Forward 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

\[ Q(X \setminus E) = \prod_{X_i \in X \setminus E} P(X_i | pa_i, e) \]

Notice: \(Q\) is another Bayesian network

Example:
Given a Bayesian network: \(P(X_1, X_2, X_3) = P(X_1) \times P(X_2 | X_1) \times P(X_3 | X_1, X_2)\) and evidence \(X_2 = x_2\).

\[ Q(X_1, X_3) = P(X_1) \times P(X_3 | X_1, X_2 = x_2) \]

Weights:

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

\[ w = \frac{P(x, e)}{Q(x)} = \frac{\prod_{X_i \in X \setminus E} P(x_i | pa_i, e) \times \prod_{E_j \in E} P(e_j | pa_j)}{\prod_{X_i \in X \setminus E} P(x_i | pa_i, e)} \]

\[ = \prod_{E_j \in E} P(e_j | pa_j) \]
Likelihood Weighting: Estimates

Estimate $P(e)$:

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

Estimate Posterior Marginals:

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

$$g_{x_i}(x^{(t)}) = 1 \text{ if } x_i = x_i^t \text{ and equals zero otherwise}$$
Properties of Likelihood Weighting

• Converges to exact posterior marginals
• Generates Samples Fast
• Sampling distribution is close to prior (especially if E ⊆ Leaf Nodes)
• Increasing sampling variance
  ⇒ Convergence may be slow
  ⇒ Many samples with $P(x^{(t)})=0$ rejected (because weight is zero)
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.

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

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

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

\[
\therefore Q(z) = P(z | e)
\]
Perfect sampling using Bucket Elimination

• Algorithm:
  – Run Bucket elimination on the problem along an ordering \( d=(X_N,\ldots,X_1) \).
  – Sample along the reverse ordering: \( (X_1,\ldots,X_N) \).
  – At each variable \( X_i \), recover the probability \( P(X_i | x_1,\ldots,x_{i-1}) \) by referring to the bucket.
How to sample from a Markov network?

Exact sampling via inference

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

\[
\begin{align*}
\tilde{b} &\sim f(\tilde{a}, b) \cdot f(b, \tilde{c}) \cdot f(b, \tilde{d}) \cdot f(b, \tilde{e}) / \lambda_{B \rightarrow C} \\
\tilde{c} &\sim f(c, \tilde{a}) \cdot f(c, \tilde{e}) \cdot \lambda_{B \rightarrow C}(\tilde{a}, c, \tilde{d}, \tilde{e}) / \lambda_{C \rightarrow D} \\
\tilde{d} &\sim f(\tilde{a}, d) \cdot \lambda_{B \rightarrow D}(d, \tilde{e}) / \lambda_{D \rightarrow E}(\tilde{a}, \tilde{e}) \\
\tilde{e} &\sim \lambda_{D \rightarrow E}(\tilde{a}, e) / \lambda_{E \rightarrow A}(\tilde{a}) \\
\tilde{a} &\sim p(A) = f(a) \cdot \lambda_{E \rightarrow A}(a) / Z
\end{align*}
\]

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 (BE)

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

bucket C: \( P(C|A) \), \( h_b^B(A,D,C,e) \)

bucket D: \( h_c^C(A,D,e) \)

bucket E: \( h_d^D(A,e) \)

bucket A: \( P(a) \), \( h_e^E(a) \)

\( P(e) \)

\[ \sum \prod_b \]

Elimination operator
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 $C$: $P(C \mid A)$  $h^B(A, D, C, e)$  

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

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

Evidence bucket: ignore

bucket $A$:  $P(A)$  $h^E(A)$  

$Q(A) \propto P(A) \times h^E(A)$

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.

Approximation of $P(e)$
Sampling from the output of MBE

- **bucket B:** 
  - $P(e|B,C)$
  - $P(B|A) P(D|B,A)$
- **bucket C:** 
  - $P(C|A)$
  - $h^B(C,e)$
- **bucket D:** 
  - $h^B(A,D)$
- **bucket E:** 
  - $h^C(A,e)$
- **bucket A:** 
  - $h^E(A)$
  - $h^D(A)$

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!
Choosing a proposal (wmb-IS)

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

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

  **Weighted mixture:**
  - use minibucket 1 with probability $w_1$
  - or, minibucket 2 with probability $w_2 = 1 - w_1$

  where

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

  $$\vdots$$

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

  **Key insight:** provides bounded importance weights!

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

[slides10a F2021]
WMB-IS Bounds

- Finite sample bounds on the average
  \[ \Pr\left[|\hat{Z} - Z| > \epsilon\right] \leq 1 - \delta \]

- Compare to forward sampling

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

“Empirical Bernstein” bounds
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)
  - Ex: [Cheng & Druzdzel 2000] [Lapeyre & Boyd 2010]
    - Lose “iid”-ness of samples
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