67800: Probabilistic Methods in AI **Spring 2017/18**

Recitation 5: Sampling-based Inference

Teaching Assistant: Eitan Richardson

Recommended reading:

- PGM Book, Chapters 12 [1]

5.1 Background

5.1.1 Sampling from a BN

Sampling from a BN is easy – *forward sampling* (aka *ancestral sampling*): Sample each RV from its CPD in topological order.

☞ *The conditional probability (defined by the CPD table) of a discrete RV with k possible values, given its observed parents, is a multinomial distribution with $k-1$ free parameters $p_1 \dots p_k$. There is a "trick" for sampling from such a distribution in $O(\log k)$ – divide the unit interval to sections of length $p_1 \dots p_k$, uniformly sample a value between 0 and 1 and check which section it fell into.*

5.1.2 Sampling-based inference

We saw in previous class that inference (probability query) is a hard problem. In some cases, approximate inference (e.g. Loopy Belief Propagation) is a possible solution (although there are no convergence or error bound guarantees). Sampling (or particle) based approximate inference is another possible solution.

☞ *We generate samples and them use them to answer probability queries (inference). This is different from estimating model parameters from real samples (learning).*

Some (confusing) notations:

- $f(\mathcal{X})$ – a general function $f : \mathcal{X} \mapsto \mathbb{R}$ (defines a new RV)
- $\xi\langle \mathbf{Y} \rangle$ – the assignment in ξ to variables in $\mathbf{Y}$
- $\mathbb{1}\{\xi\langle \mathbf{Y} \rangle = y\}$ – an indicator RV – equals 1 if the assignment in ξ to $\mathbf{Y}$ is y
- $\mathcal{D} = \{\xi[1], \dots, \xi[M]\}$ – A set of M samples
- $y[m]$ – short for $\xi[m]\langle \mathbf{Y} \rangle$ (the assignment in sample $\xi[m]$ to the subset of variables $\mathbf{Y}$)

[1] *Original LaTeX template courtesy of UC Berkeley.*

Approximating the expectation of $f(\mathcal{X})$ by sampling:

$$\mathbb{E}_P f(\mathcal{X}) \approx \hat{\mathbb{E}}_{\mathcal{D}} f(\mathcal{X}) = \frac{1}{M}\sum_{m=1}^{M} f(\xi[m])$$

Specifically, if we choose $f(\mathcal{X}) = \mathbb{1}\{y[m] = y\}$, we get:

$$\mathbb{E}_P[\mathbb{1}\{y[m] = y\}] = P(Y = y) \approx \hat{P}_{\mathcal{D}}(y) = \frac{1}{M}\sum_{m=1}^{M} \mathbb{1}\{y[m] = y\}$$

☞ *This is an approximate estimation of the unconditional marginal probability.*

5.2 Approximation error bounds

How accurate is the sampling-based approximation? How many samples do we need?

$\mathbb{1}\{Y = y\}$ is a Bernoulli RV with $p = P(y)$, so our sample $\mathcal{D}$ defines M independent Bernoulli trials.

Theorem 5.1 *Hoeffding bound*
Let $\{x[1], \ldots x[M]\}$ be M independent Bernoulli trials with success probability p and let $\hat{q} = \frac{1}{M}\sum_{m=1}^{M} x[m]$, then:

$$P(\hat{q} > p + \epsilon) \leq e^{-2M\epsilon^2}\ ,\ \ P(\hat{q} < p - \epsilon) \leq e^{-2M\epsilon^2}$$
$$P(|p - \hat{q}| > \epsilon) \leq 2e^{-2M\epsilon^2}$$

So if we want an estimate with an approximation error not larger than ϵ with probability of at least $1 - \delta$, we need:

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

How many samples do we need if we want to bound the error relative to the event probability (e.g. not more than 1% of the real event probability)?

Applying *Chernhoff bound*, we get:

$$P(\hat{q} > p(1 + \epsilon)) \leq e^{-2Mp\epsilon^2/3}\ ,\ \ P(\hat{q} < p(1 - \epsilon)) \leq e^{-2Mp\epsilon^2/3}$$
$$P(\hat{q} \notin p(1 \pm \epsilon)) \leq 2e^{-2Mp\epsilon^2/3}$$

So:

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

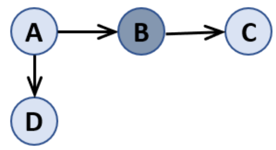
☞ *To estimate the probability of a rare event, we'll need much more data!*

5.3 Conditional Probability Queries

How do we estimate $P(Y = y \mid \mathbf{E} = e)$?

Maybe we can do forward sampling except that we **force** all variables in $\mathbf{E}$ to e?

Example 5.2 *(bad solution) Forward sampling and forcing observed variables:*
Sample A from its prior $P(A)$, set $B = b$, sample C from $P(C|B = b)$ and sample D from $P(D|A = a)$.



☞ The process above will not generate samples from $P(A, C, D \mid B = b)$. The reason is that we're not taking into account that $P(A \mid E = e) \neq P(A)$. This affects both samples of A and of D.

Possible solution: *Rejection Sampling* – sample all variables, reject all samples in which $E \neq e$, calculate as before using remaining samples:

$$P(Y = y \mid \mathbf{E} = e) \approx \frac{\sum_{m=1}^{M} \mathbb{1}\{y[m] = y,\, e[m] = e\}}{\sum_{m=1}^{M} \mathbb{1}\{e[m] = e\}}$$

☞ *Rejection sampling will provide an accurate estimate (with enough samples), but if $P(E = e)$ is small, we'll throw away almost all our samples...*

A better solution is presented below – *Likelihood Weighting*.

5.3.1 Likelihood Weighting

The idea is to perform forward sampling, force observed variables to their evidence value but re-weight the samples according to the likelihood:

$$P(y \mid e) \approx \hat{P}_{\mathcal{D}}(y \mid e) = \frac{\sum_{m=1}^{M} w[m] \mathbb{1}\{y[m] = y\}}{\sum_{m=1}^{M} w[m]}, \quad w[m] = \prod_{E \in \mathbf{E}} P(e \mid Pa_E[m])$$

$w[m]$ is the likelihood of the observed parameters given their parents. Since these are independent events, we take the product of the CPD entries.

Algorithm 1 Likelihood-weighted Sampling (single sample)

procedure LW-SAMPLE($\mathcal{B}, \mathbf{E} = e$)
    $w = 1$
    for $i = 1 \ldots n$ **do** ▷ topological order
        if $X_i \in \mathbf{E}$ **then**
            $x_i = e\langle X_i \rangle$ ▷ Assignment to X_i in the evidence
            $w = w \cdot P(x_i \mid Pa_{X_i})$ ▷ Likelihood of evidence given already sampled parents
        else
            Sample x_i from $P(X_i \mid Pa_{X_i})$
    return $(x_1, \ldots, x_n)$, w

☞ *We didn't prove this (intuitive) method is correct – we'll do it using the more general method of* Importance Sampling.

Example 5.3 *The Stopped Car – Likelihood Weighting*
Estimate $P(M = 1 \mid L = 1, N = 0)$

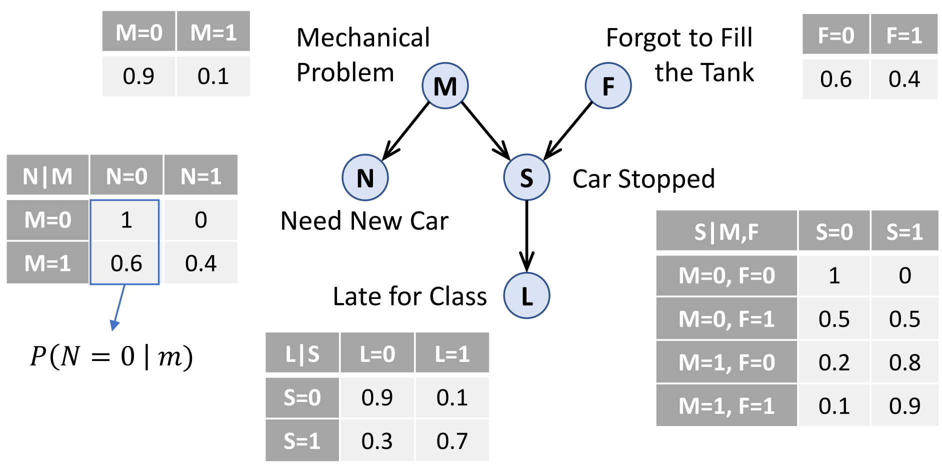


Figure 5.1: Bayesian Network example – *The Stopped Car*

We sample m from $P(M)$, f from $P(F)$, set n to 0, sample s from $P(S \mid m, f)$ and set l to 1. The weight of the sample is $P(N = 0 \mid m) \cdot P(L = 1 \mid s)$.

iteration	m	f	s	$P(N=0 \mid m)$	$P(L=1 \mid s)$	w	$\hat{P}_{\mathcal{D}}(M=1 \mid L=1, N=0)$
0	0	1	1	1.0	0.7	0.7	0.0
1	0	0	0	1.0	0.1	0.1	0.0
2	0	0	0	1.0	0.1	0.1	0.0
3	1	0	1	0.6	0.7	0.42	0.31818181818181823
4	0	1	0	1.0	0.1	0.1	0.29577464788732394
5	0	0	0	1.0	0.1	0.1	0.2763157894736842
6	0	0	0	1.0	0.1	0.1	0.25925925925925924
7	0	0	0	1.0	0.1	0.1	0.24418604651162787
8	0	1	0	1.0	0.1	0.1	0.23076923076923073
9	0	1	0	1.0	0.1	0.1	0.21874999999999994
10	0	0	0	1.0	0.1	0.1	0.20792079207920786
...							
995	0	1	0	1.0	0.1	0.1	0.15340604326837382
996	0	1	1	1.0	0.7	0.7	0.15292754656447996
997	0	0	0	1.0	0.1	0.1	0.15285943345804645
998	0	1	0	1.0	0.1	0.1	0.1527913809990232
999	0	0	0	1.0	0.1	0.1	0.1527233891064462

5.4 Importance Sampling

Normalized and un-normalized distributions: $P(X) = \frac{1}{Z}\tilde{P}(X)$
(relevant for MNs and for BNs with evidence ($Z = P(e)$))

Calculating an expectation over distribution P using a second distribution Q:

$$\begin{aligned}\mathbb{E}_P[f(x)] &= \int P(x)f(x)dx = \int Q(x)f(x)\frac{P(x)}{Q(x)}dx = \frac{1}{Z}\int Q(x)f(x)\frac{\tilde{P}(x)}{Q(x)}dx \\ &= \frac{1}{Z}\mathbb{E}_Q[f(x)\frac{\tilde{P}(x)}{Q(x)}] = \frac{1}{Z}\mathbb{E}_Q[f(x)w(x)]\end{aligned}$$

The Un-normalized Importance Sampling estimator (for $Z = 1$):

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

The Normalized Importance Sampling estimator (for the general case):

$$\mathbb{E}_Q[w(x)] = \mathbb{E}_Q[\frac{\tilde{P}(x)}{Q(x)}] = \int \tilde{P}(x)dx = Z$$

$$\mathbb{E}_P[f(x)] = \frac{1}{Z}\mathbb{E}_Q[f(x)w(x)] = \frac{\mathbb{E}_Q[f(x)w(x)]}{\mathbb{E}_Q[w(x)]}$$

So the normalized estimator is:

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

5.4.1 Likelihood Weighting as Importance Sampling

☞ *The Likelihood Weighting method has a similar form to NIS... What Q did we use?*

Definition 5.4 *The Mutilated BN*
Let $\mathcal{B}$ be a BN with evidence $\boldsymbol{E} = e$. We define the Mutilated BN $\mathcal{B}_{\boldsymbol{E}=e}$ *as follows:*

- *Incoming edges to each node $X_i \in \boldsymbol{E}$ are removed (i.e. no parents) and its CPD is set to $P(X_i = e\langle X_i\rangle) = 1$*

- *All other edges and CPDs are unchanged*

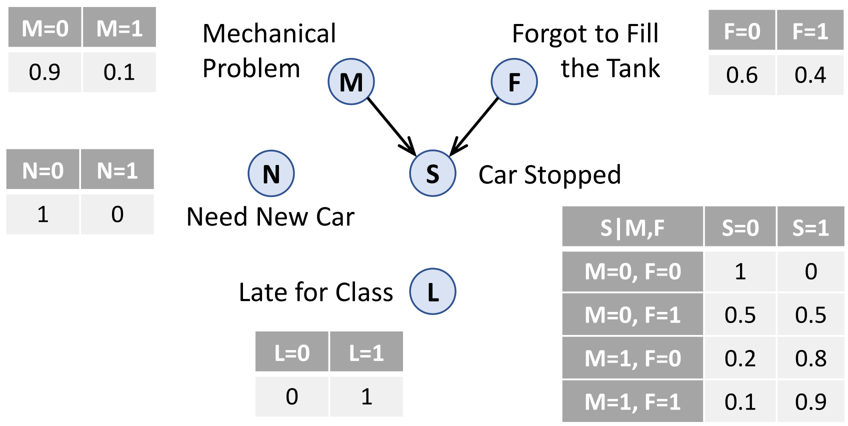


Figure 5.2: Mutilated Bayesian Network example for $\mathbf{E} = \{N = 0, L = 1\}$

Proposition 5.5 *LW is equivalent to NIS with* $Q(X) = P_{\mathcal{B}_{\mathbf{E}=e}}(X)$

Proof: Proof sketch
We need to show that:

1. $x[m] \sim P_{\mathcal{B}_{\mathbf{E}=e}}(X)$ – The LW samples are drawn from the mutilated BN distribution
2. $w[m] = \frac{P_{\mathcal{B}}(x[m])}{P_{\mathcal{B}_{\mathbf{E}=e}}(x[m])}$

Proof for 1:
For $X_i \notin \mathbf{E} \cup Desc_{\mathbf{E}}$, we sample from $P_{\mathcal{B}}$, which is identical to $\mathcal{B}_{\mathbf{E}=e}$ above the first evidence.
For $X_i \in \mathbf{E}$, we force the evidence, which is consistent with the deterministic CPDs.
For the remaining $X_i \in Desc_{\mathbf{E}}$ we can show (by induction, from $E \in \mathbf{E}$ downwards) that $P_{\mathcal{B}_{\mathbf{E}=e}}(X_i | Par_{X_i}) = P_{\mathcal{B}}(X_i \mid Par_{X_i}, \mathbf{E} = e)$

Proof for 2:
Let's start with our example: (the $[m]$ index was removed to keep the expression short)

$$\begin{aligned}\frac{P_{\mathcal{B}}(x)}{P_{\mathcal{B}_{\mathbf{E}=e}}(x)} &= \frac{P_{\mathcal{B}}(m)P_{\mathcal{B}}(f)P_{\mathcal{B}}(N=0 \mid m)P_{\mathcal{B}}(s \mid m,f)P_{\mathcal{B}}(L=1 \mid s)}{P_{\mathcal{B}_{\mathbf{E}=e}}(m)P_{\mathcal{B}_{\mathbf{E}=e}}(f)P_{\mathcal{B}_{\mathbf{E}=e}}(N=0)P_{\mathcal{B}_{\mathbf{E}=e}}(s \mid m,f)P_{\mathcal{B}_{\mathbf{E}=e}}(L=1)} \\ &= \frac{P_{\mathcal{B}}(m)P_{\mathcal{B}}(f)P_{\mathcal{B}}(N=0 \mid m)P_{\mathcal{B}}(s \mid m,f)P_{\mathcal{B}}(L=1 \mid s)}{P_{\mathcal{B}}(m)P_{\mathcal{B}}(f) \cdot 1 \cdot P_{\mathcal{B}}(s \mid m,f) \cdot 1} \\ &= P_{\mathcal{B}}(N=0 \mid m) \cdot P_{\mathcal{B}}(L=1 \mid s) = w\end{aligned}$$

It is easy to show that this is true in the general case. ■

References

[1] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.