# Approximate Inference 1

1

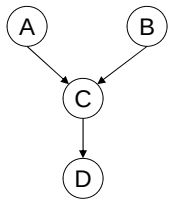
## Forward Sampling

- This section on approximate inference relies on samples / particles
- Full particles: complete assignment to all network variables eg. ($X_1 = x_1$, $X_2 = x_2$, ..., $X_N = x_N$)

2

# Forward Sampling

- Topological sort or order: An ordering of the nodes in the DAG where X comes before Y in the ordering if there is a directed path from X to Y in the graph.
- A topological order is equivalent to a partial order on the nodes of the graph
- There may be several topological orderings




Examples of Topological orders:

A,B,C,D

B,A,C,D

3

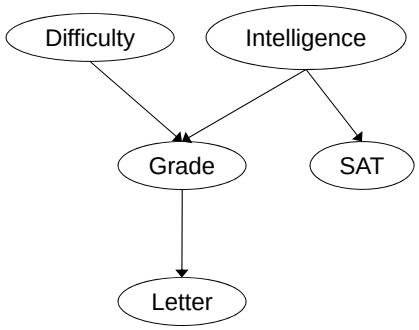
# Forward Sampling

Student Example

| D | P(D) |
|---|---|
| low | 0.6 |
| high | 0.4 |

| I | P(I) |
|---|---|
| low | 0.7 |
| high | 0.3 |

| D | I | G | P(G\|D,I) |
|---|---|---|---|
| low | low | C | 0.3 |
| low | low | B | 0.4 |
| low | low | A | 0.3 |
| low | high | C | 0.02 |
| low | high | B | 0.08 |
| low | high | A | 0.9 |
| high | low | C | 0.7 |
| high | low | B | 0.25 |
| high | low | A | 0.05 |
| high | high | C | 0.2 |
| high | high | B | 0.3 |
| high | high | A | 0.5 |



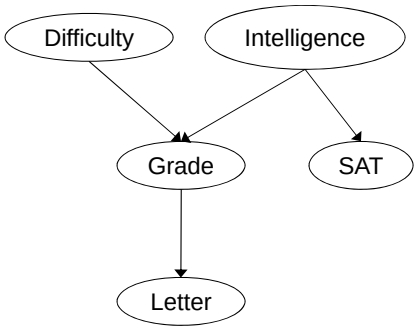

| I | S | P(S\|I) |
|---|---|---|
| low | low | 0.95 |
| low | high | 0.05 |
| high | low | 0.2 |
| high | high | 0.8 |

| G | L | P(L\|G) |
|---|---|---|
| C | weak | 0.99 |
| C | strong | 0.01 |
| B | weak | 0.4 |
| B | strong | 0.6 |
| A | weak | 0.1 |
| A | strong | 0.9 |

4

# Forward Sampling

Topological ordering: D, I, G, S, L




1. Sample D from P(D) (Say you get D=high)
2. Sample I from P(I) (Say you get I=low)
3. Sample G from P(G|I=low,D=high) (Say you get G=C)
4. Sample S from P(S|I=low) (Say you get S=low)
5. Sample L from P(L|G=C) (Say you get L=weak)

You now have a sample (D=high, I=low, G=C, S=low, L=weak)

5

# Forward Sampling

Suppose you want to calculate $P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$ using forward sampling on a Bayesian network. The algorithm:

1. Do a topological sort of the nodes in the Bayesian network.

2. For *j = 1* to *NUM_SAMPLES*
   For each node *i* in the ordering (starting from the top of the Bayesian network down)
   Sample the value $\hat{x}_i$ from the distribution $P(X_i \mid Parents(X_i))$
   Add $\{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n\}$ to your collection of samples

3. Let $P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \approx \frac{\#\text{ of samples with } X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n}{NUM\_SAMPLES}$

6

# Forward Sampling

- How do you sample from $P(X_i \mid Parents(X_i))$?
- Note: $P(X_i \mid Parents(X_i))$ is a multinomial distribution $P(x_i^1, ..., x_i^k \mid \theta_1, ..., \theta_k)$?

7

# Forward Sampling

How do you sample from a multinomial distribution $P(x_i^1, ..., x_i^k \mid \theta_1, ..., \theta_k)$?

- Generate a sample *s* uniformly from [0,1]
- Partition interval into *k* subintervals: $[0, \theta_1)$, $[\theta_1, \theta_1+\theta_2)$, ...
- More generally, the *i*th interval is
$$\left[\sum_{j=1}^{i-1}\theta_j, \sum_{j=1}^{i}\theta_j\right)$$
- If *s* is in the ith interval, the sample value is $x_i$.
- Use binary search to find the interval for *s* in time $O(\log k)$

8

# Forward Sampling

Suppose your list of samples looks like the following table:

| D | I | G | S | L |
|---|---|---|---|---|
| low | low | B | low | weak |
| low | high | A | high | strong |
| low | high | A | high | weak |
| high | high | A | high | strong |
| high | low | C | low | weak |

P(I=high) = 3/5 = 0.6

Note that this value becomes a lot more accurate as the number of samples heads to infinity.

9

# Forward Sampling

- From a set of particles ***D*** = *{ξ[1], ..., ξ[M]}*, we can estimate the expectation of any function *f* as:

$$\hat{E}_D(f) = \frac{1}{M}\sum_{m=1}^{M} f(\xi[m])$$

- To estimate *P(**y**)*

$$\hat{P}_D(\boldsymbol{y}) = \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{I}\{\boldsymbol{y}[m] = \boldsymbol{y}\}$$

This is the values of the variables in ***Y*** in the particle ξ[m]

10

# Forward Sampling

Define

- $M$ = total # of particles generated
- $n = |\mathcal{X}|$
- $p = \max_i |Pa_{Xi}|$
- $d = \max_i |Val(X_i)|$

Overall cost of sampling is $O(M\, n\, p\, \log(d))$

- To get the CPD entry for X given $Pa_X$, it costs $O(p)$
- Sampling process for $P(X|Pa_X)$ costs $O(\log(d))$

11

# Forward Sampling

How accurate is this estimate? Using the Hoeffding bound:

$$P_D(\hat{P}_D(\boldsymbol{y}) \notin [P(\boldsymbol{y}) - \varepsilon, P(\boldsymbol{y}) + \varepsilon]) \leq 2e^{-2M\varepsilon^2}$$

How many samples are required to achieve an estimate whose error is bounded by $\varepsilon$, with probability at least $(1-\delta)$? Setting

$$2e^{-2M\varepsilon^2} \leq \delta \text{ we get } M \geq \frac{\ln(2/\delta)}{2\varepsilon^2}$$

12

# Forward Sampling

How accurate is this estimate? Using the Chernoff bound:

$$P_D(\hat{P}_D(\boldsymbol{y}) \notin P(\boldsymbol{y})(1 \pm \varepsilon)) \leq 2e^{-MP(\boldsymbol{y})\varepsilon^2/3}$$

Note: This requires us to know P(**y**)

How many samples are required to achieve an estimate whose error is bounded by ε, with probability at least (1-δ)?

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

13

# Rejection Sampling

14

# Rejection Sampling

What if we want to estimate $P(\mathbf{y}|\mathbf{E}=\mathbf{e})$?

- Rejection sampling: do forward sampling but throw out samples where $\mathbf{E} \neq \mathbf{e}$

Example:

P(I=high|L=weak) = 1/3

| D | I | G | S | L |
|---|---|---|---|---|
| low | low | B | low | weak |
| low | high | A | high | strong |
| low | high | A | high | weak |
| high | high | A | high | strong |
| high | low | C | low | weak |

15

# Rejection Sampling

What if the evidence $\mathbf{E}=\mathbf{e}$ is very very rare?

- For example, if $P(\mathbf{e}) = 0.001$, then for 10,000 samples, we get 10 unrejected samples
- To obtain at least $M^*$ unrejected samples, we need to generate on average $M = M^*/P(\mathbf{e})$ samples
- If evidence is rare, we end up generating a lot of samples which wastes time

16

## Rejection Sampling

Bad news:

- Rare evidence is the norm!
- As # of evidence variables $k = |\boldsymbol{E}|$ grows, the probability of the evidence decreases exponentially with $k$

Need something better than rejection sampling!

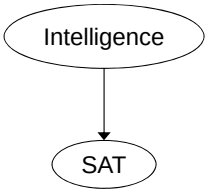
17

## Likelihood Weighting

18

# Likelihood Weighting

Intuition: Weight samples according to probability of the evidence




| I | P(I) |
|---|---|
| low | 0.7 |
| high | 0.3 |

| I | S | P(S\|I) |
|---|---|---|
| low | low | 0.95 |
| low | high | 0.05 |
| high | low | 0.2 |
| high | high | 0.8 |

Drawing I = high and S = high should be 80% of a sample

Drawing I = low and S = high should be 5% of a sample

19

# Likelihood Weighting

Weighted particles:

$$D = \langle \xi[1], w[1] \rangle, \ldots, \langle \xi[M], w[M] \rangle$$

Estimate:

$$\hat{P}_D(\boldsymbol{y} \mid \boldsymbol{e}) = \frac{\sum_{m=1}^{M} w[m] \boldsymbol{I}\{\boldsymbol{y}[m] = \boldsymbol{y}\}}{\sum_{m=1}^{M} w[m]}$$

20

# Likelihood Weighting

**Procedure** LW-Sample(
  $\beta$,  // Bayesian network over $\mathcal{X}$
  **Z**=**z**  // Event in the network
)

1. Let $X_1, \ldots, X_n$ be a topological ordering of $\mathcal{X}$
2. $w \leftarrow 1$
3. **for** i = 1, …, n
4.  $\mathbf{u_i} \leftarrow \mathbf{x}\langle Pa_{Xi}\rangle$  // Assignment to $Pa_{Xi}$ in $x_1, \ldots, x_{i-1}$
5.  **if** $X_i \notin \mathbf{Z}$ **then**
6.  Sample $x_i$ from $P(X_i \mid \mathbf{u_i})$
7.  else
8.  $x_i \leftarrow \mathbf{z}\langle X_i\rangle$ // Assignment to $X_i$ in **z**
9.  $w \leftarrow w \cdot P(x_i \mid \mathbf{u_i})$ // Multiply weight by probability of desired value
10.  **return** $(x_1, \ldots, x_n)$, w