Bayes Net Representation

- A directed, acyclic graph, one node per random variable
- A conditional probability table (CPT) for each node
  - A collection of distributions over $X$, one for each combination of parents’ values

\[
P(X|a_1 \ldots a_n)
\]

- Bayes nets implicitly encode joint distributions
  - As a product of local conditional distributions
  - To see what probability a BN gives to a full assignment, multiply all the relevant conditionals together:

\[
P(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} P(x_i|\text{parents}(X_i))
\]
Inference by Enumeration vs. Variable Elimination

- Why is inference by enumeration so slow?
  - You join up the whole joint distribution before you sum out the hidden variables

- Idea: interleave joining and marginalizing!
  - Called “Variable Elimination”
  - Still NP-hard, but usually much faster than inference by enumeration
Traffic Domain

Inference by Enumeration

\[ P(L) = ? \]

\[ \sum_t \sum_r P(L|t)P(r)P(t|r) \]

- Join on \( r \)
- Join on \( t \)
- Eliminate \( r \)
- Eliminate \( t \)

Variable Elimination

\[ = \sum_t P(L|t) \sum_r P(r)P(t|r) \]

- Join on \( r \)
- Eliminate \( r \)
- Join on \( t \)
- Eliminate \( t \)
General Variable Elimination

- Query: \[ P(Q|E_1 = e_1, \ldots E_k = e_k) \]

- Start with initial factors:
  - Local CPTs (but instantiated by evidence)

- While there are still hidden variables (not Q or evidence):
  - Pick a hidden variable \( H \)
  - Join all factors mentioning \( H \)
  - Eliminate (sum out) \( H \)

- Join all remaining factors and normalize
### Variable Elimination Efficiency

- Interleave joining and marginalizing, instead of fully joining all at once (i.e. enumeration)

- $d^k$ entries computed for a factor over $k$ variables with domain sizes $d$

- Ordering of elimination of hidden variables can affect size of factors generated

- Worst case: running time exponential in the size of the Bayes net (NP-hard)
Approximate Inference: Sampling
Sampling

- **Basic idea**
  - Draw N samples from a sampling distribution S
  - Compute an approximate posterior probability
  - Show this converges to the true probability $P$

- **Why sample?**
  - Learning: get samples from a distribution you don’t know
  - Inference: getting samples can be faster than computing the right answer (e.g. with variable elimination)
Sampling

- Sampling from given distribution
  - Step 1: Get sample $u$ from uniform distribution over $[0, 1)$
    - E.g. random() in python
  - Step 2: Convert this sample $u$ into an outcome for the given distribution by having each outcome associated with a sub-interval of $[0,1)$ with sub-interval size equal to probability of the outcome

- Example

<table>
<thead>
<tr>
<th>C</th>
<th>P(C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>red</td>
<td>0.6</td>
</tr>
<tr>
<td>green</td>
<td>0.1</td>
</tr>
<tr>
<td>blue</td>
<td>0.3</td>
</tr>
</tbody>
</table>

- If random() returns $u = 0.83$, then our sample is $C = \text{blue}$
- E.g., after sampling 8 times:

  $0 \leq u < 0.6$, \( \rightarrow C = \text{red} \)
  $0.6 \leq u < 0.7$, \( \rightarrow C = \text{green} \)
  $0.7 \leq u < 1$, \( \rightarrow C = \text{blue} \)
Sampling in Bayes Nets

- Prior Sampling
- Rejection Sampling
- Likelihood Weighting
- Gibbs Sampling
Prior Sampling
Prior Sampling

$P(C)$
- $+c$: 0.5
- $-c$: 0.5

Samples:
- $+c$, $-s$, $+r$, $+w$
- $-c$, $+s$, $-r$, $+w$

...
Prior Sampling

- For $i=1, 2, \ldots, n$
  - Sample $x_i$ from $P(X_i | \text{Parents}(X_i))$
- Return $(x_1, x_2, \ldots, x_n)$
Prior Sampling

- This process generates samples with probability:

\[ S_{PS}(x_1 \ldots x_n) = \prod_{i=1}^{n} P(x_i|\text{Parents}(X_i)) = P(x_1 \ldots x_n) \]

...i.e. the BN’s joint probability

- Let the number of samples of a particular event be \( N_{PS}(x_1 \ldots x_n) \) and the total number of samples of all events be \( N \).

- Then

\[
\lim_{N \to \infty} \hat{P}(x_1, \ldots, x_n) = \lim_{N \to \infty} \frac{N_{PS}(x_1, \ldots, x_n)}{N} = S_{PS}(x_1, \ldots, x_n) = P(x_1 \ldots x_n)
\]

- I.e., the sampling procedure is consistent
We’ll get a bunch of samples from the BN:

+\(c\), -\(s\), +\(r\), +\(w\)
+\(c\), +\(s\), +\(r\), +\(w\)
-\(c\), +\(s\), +\(r\), -\(w\)
-\(c\), -\(s\), +\(r\), +\(w\)
+\(c\), -\(s\), -\(r\), +\(w\)

If we want to know \(P(W)\)

- We have counts <+\(w\):4, -\(w\):1>
- Normalize to get \(P(W) = <+\(w\):0.8, -\(w\):0.2>\)
- This will get closer to the true distribution with more samples
- Can estimate anything else, too
- What about \(P(C \mid +w)\)? \(P(C \mid +r, +w)\)? \(P(C \mid -r, -w)\)?
- Fast: can use fewer samples if less time (what’s the drawback?)
Rejection Sampling
Rejection Sampling

- Let’s say we want \( P(C) \)
  - No point keeping all samples around
  - Just tally counts of \( C \) as we go

- Let’s say we want \( P(C \mid +s) \)
  - Same thing: tally \( C \) outcomes, but ignore (reject) samples which don’t have \( S=+s \)
  - This is called rejection sampling
  - It is also consistent for conditional probabilities (i.e., correct in the limit)
Rejection Sampling

- **IN: evidence instantiation**
- **For** $i=1, 2, ..., n$
  - Sample $x_i$ from $P(X_i \mid \text{Parents}(X_i))$
  - If $x_i$ not consistent with evidence
    - Reject: Return, and no sample is generated in this cycle
  - Return $(x_1, x_2, ..., x_n)$
Likelihood Weighting
\textbf{Likelihood Weighting}

- **Problem with rejection sampling:**
  - If evidence is unlikely, rejects lots of samples
  - Evidence not exploited as you sample
  - Consider $P(\text{Shape} \mid \text{blue})$

- **Idea:** fix evidence variables and sample the rest
  - Problem: sample distribution not consistent!
  - Solution: weight by probability of evidence given parents

<table>
<thead>
<tr>
<th>Shape</th>
<th>Color</th>
</tr>
</thead>
<tbody>
<tr>
<td>pyramid, green</td>
<td></td>
</tr>
<tr>
<td>pyramid, red</td>
<td></td>
</tr>
<tr>
<td>sphere, blue</td>
<td></td>
</tr>
<tr>
<td>cube, red</td>
<td></td>
</tr>
<tr>
<td>sphere, green</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Shape</th>
<th>Color</th>
</tr>
</thead>
<tbody>
<tr>
<td>pyramid, blue</td>
<td></td>
</tr>
<tr>
<td>pyramid, blue</td>
<td></td>
</tr>
<tr>
<td>sphere, blue</td>
<td></td>
</tr>
<tr>
<td>cube, blue</td>
<td></td>
</tr>
<tr>
<td>sphere, blue</td>
<td></td>
</tr>
</tbody>
</table>
Likelihood Weighting

$P(C)$

<table>
<thead>
<tr>
<th></th>
<th>+c</th>
<th>-c</th>
</tr>
</thead>
<tbody>
<tr>
<td>+c</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>-c</td>
<td>0.5</td>
<td></td>
</tr>
</tbody>
</table>

$P(S|C)$

<table>
<thead>
<tr>
<th></th>
<th>+s 0.1</th>
<th>+s 0.5</th>
<th>-s 0.9</th>
<th>-s 0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>+c</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-c</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$P(R|C)$

<table>
<thead>
<tr>
<th></th>
<th>+r 0.8</th>
<th>+r 0.2</th>
<th>-r 0.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>+c</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-c</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$P(W|S, R)$

<table>
<thead>
<tr>
<th></th>
<th>+w 0.99</th>
<th>-w 0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>+s</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>+r</td>
<td>-r</td>
</tr>
<tr>
<td>+w</td>
<td>0.90</td>
<td>0.10</td>
</tr>
<tr>
<td>-w</td>
<td>0.90</td>
<td>0.10</td>
</tr>
<tr>
<td>+w</td>
<td>0.90</td>
<td>0.10</td>
</tr>
<tr>
<td>-w</td>
<td>0.90</td>
<td>0.10</td>
</tr>
<tr>
<td>+w</td>
<td>0.99</td>
<td>0.01</td>
</tr>
<tr>
<td>-w</td>
<td>0.99</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Samples:

+ +c, +s, +r, +w

$w = 1.0 \times 0.1 \times 0.99$
Likelihood Weighting

- **IN:** evidence instantiation
- \( w = 1.0 \)
- for \( i = 1, 2, \ldots, n \)
  - if \( X_i \) is an evidence variable
    - \( X_i = \) observation \( x_i \) for \( X_i \)
    - Set \( w = w \times P(x_i | \text{Parents}(X_i)) \)
  - else
    - Sample \( x_i \) from \( P(X_i | \text{Parents}(X_i)) \)
- return \((x_1, x_2, \ldots, x_n), w\)
Likelihood Weighting

- Sampling distribution if \( z \) sampled and \( e \) fixed evidence

\[
S_{WS}(z, e) = \prod_{i=1}^{l} P(z_i|\text{Parents}(Z_i))
\]

- Now, samples have weights

\[
w(z, e) = \prod_{i=1}^{m} P(e_i|\text{Parents}(E_i))
\]

- Together, weighted sampling distribution is consistent

\[
S_{WS}(z, e) \cdot w(z, e) = \prod_{i=1}^{l} P(z_i|\text{Parents}(z_i)) \prod_{i=1}^{m} P(e_i|\text{Parents}(e_i))
\]

\[= P(z, e)\]
Likelihood Weighting

- Likelihood weighting is good
  - We have taken evidence into account as we generate the sample
  - Our samples will reflect the state of the world suggested by the evidence
  - No need for rejection!

- Likelihood weighting doesn’t solve all our problems
  - Evidence influences the choice of downstream variables, but not upstream ones (not more likely to get a value matching the evidence)
  - Can cause many very small weights → inefficient!

- We would like to consider evidence when we sample every variable
  → Gibbs sampling
Gibbs Sampling
Gibbs Sampling Example: $P(S \mid +r)$

- **Step 1:** Fix evidence
  - $R = +r$

- **Step 2:** Initialize other variables
  - Randomly

- **Step 3:** Repeat the following:
  - Choose a non-evidence variable $X$
  - Resample $X$ from $P(X \mid \text{all other variables})$

Sample from $P(S \mid c, -w, +r)$
Sample from $P(C \mid s, -w, +r)$
Sample from $P(W \mid s, +c, +r)$
Gibbs Sampling

How is this better than sampling from the full joint?

- In a Bayes Net, sampling a variable given all the other variables (e.g. $P(R|S,C,W)$) is usually much easier than sampling from the full joint distribution.
- Only requires one join on the variable to be sampled (in this case, a join on $R$).
Further Reading on Gibbs Sampling*

- Gibbs sampling produces sample from the query distribution $P(Q | e)$ in limit of re-sampling infinitely often.

- Gibbs sampling is a special case of more general methods called Markov chain Monte Carlo (MCMC) methods.
  - Metropolis-Hastings is one of the more famous MCMC methods (in fact, Gibbs sampling is a special case of Metropolis-Hastings).

- You may read about Monte Carlo methods – they’re just sampling.
Markov Chain Monte Carlo*

- **Idea:** instead of sampling from scratch, create samples that are each like the last one.

- **Procedure:** resample one variable at a time, conditioned on all the rest, but keep evidence fixed. E.g., for \( P(b|c) \):

  ![Diagram](image)

- **Properties:** Now samples are not independent (in fact they’re nearly identical), but sample averages are still consistent estimators!

- **What’s the point:** both upstream and downstream variables condition on evidence.
Bayes Net Sampling Summary

- Prior Sampling  $P(Q)$
- Rejection Sampling  $P(Q | e)$
- Likelihood Weighting  $P(Q | e)$
- Gibbs Sampling  $P(Q | e)$