Bayesian Networks: Inference

Luke Zettlemoyer

Many slides over the course adapted from either Dan Klein, Stuart Russell or Andrew Moore
Outline

- Bayesian Networks Inference
  - Exact Inference: Variable Elimination
  - Approximate Inference: Sampling
Variable Elimination

- Why is inference by enumeration so slow?
  - You join up the whole joint distribution before you sum out the hidden variables
  - You end up repeating a lot of work!

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

- We’ll need some new notation to define VE
Example: Traffic Domain

- **Random Variables**
  - R: Raining
  - T: Traffic
  - L: Late for class!

- First query: \( P(L) \)

\[
P(l) = \sum_{t} \sum_{r} P(l|t)P(t|r)P(r)
\]
Variable Elimination Outline

- Maintain a set of tables called factors
- Initial factors are local CPTs (one per node)

\[
\begin{array}{c|c}
P(R) & P(T|R) & P(L|T) \\
\hline
+r & +t & 0.8 & +t & +l & 0.3 \\
+r & -t & 0.2 & +t & -l & 0.7 \\
-r & +t & 0.1 & -t & +l & 0.1 \\
-r & -t & 0.9 & -t & -l & 0.9 \\
\end{array}
\]

- Any known values are selected
  - E.g. if we know \( L = +l \), the initial factors are

\[
\begin{array}{c|c}
P(R) & P(T|R) & P(+l|T) \\
\hline
+r & +t & 0.8 & +t & +l & 0.3 \\
+r & -t & 0.2 & -t & +l & 0.1 \\
-r & +t & 0.1 & \text{ - } & \text{ - } & \text{ - } \\
-r & -t & 0.9 & \text{ - } & \text{ - } & \text{ - } \\
\end{array}
\]

- VE: Alternately join factors and eliminate variables
Operation 1: Join Factors

- First basic operation: joining factors
- Combining factors:
  - Just like a database join
  - Get all factors over the joining variable
  - Build a new factor over the union of the variables involved
- Example: Join on R

\[
P(R) \times P(T|R) \rightarrow P(R,T)
\]

| R  | T  | P(R) | P(T|R) | P(R,T) |
|----|----|------|--------|--------|
| +r | +t | 0.1  | +r     | 0.8    |
| +r | -t | 0.2  | +r     | 0.2    |
| -r | +t | 0.1  | -r     | 0.1    |
| -r | -t | 0.9  | -r     | 0.9    |

- Computation for each entry: pointwise products

\[\forall r, t : P(r, t) = P(r) \cdot P(t|r)\]
Example: Multiple Joins

\[ P(R) \]

\[
\begin{array}{c|c}
+r & 0.1 \\ 
-r & 0.9 \\
\end{array}
\]

\[ P(T|R) \]

\[
\begin{array}{c|cc}
+r & +t & 0.8 \\ 
+r & -t & 0.2 \\ 
-r & +t & 0.1 \\ 
-r & -t & 0.9 \\
\end{array}
\]

Join R

\[ P(R,T) \]

\[
\begin{array}{c|cc}
+r & +t & 0.08 \\ 
+r & -t & 0.02 \\ 
-r & +t & 0.09 \\ 
-r & -t & 0.81 \\
\end{array}
\]

\[ P(L|T) \]

\[
\begin{array}{c|cc}
+t & +l & 0.3 \\ 
+t & -l & 0.7 \\ 
-t & +l & 0.1 \\ 
-t & -l & 0.9 \\
\end{array}
\]

\[ P(L|T) \]

\[
\begin{array}{c|cc}
+t & +l & 0.3 \\ 
+t & -l & 0.7 \\ 
-t & +l & 0.1 \\ 
-t & -l & 0.9 \\
\end{array}
\]
Example: Multiple Joins

\[ P(R, T) \]

<table>
<thead>
<tr>
<th></th>
<th>+t</th>
<th>-t</th>
<th>0.08</th>
<th>0.02</th>
<th>0.09</th>
<th>0.81</th>
</tr>
</thead>
<tbody>
<tr>
<td>+r</td>
<td>+t</td>
<td></td>
<td>0.02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+r</td>
<td>-t</td>
<td></td>
<td>0.09</td>
<td>0.08</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-r</td>
<td>+t</td>
<td></td>
<td>0.02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-r</td>
<td>-t</td>
<td></td>
<td>0.81</td>
<td>0.09</td>
<td>0.02</td>
<td>0.08</td>
</tr>
</tbody>
</table>

\[ P(R, T, L) \]

<table>
<thead>
<tr>
<th></th>
<th>+t</th>
<th></th>
<th>+l</th>
<th></th>
<th>0.024</th>
</tr>
</thead>
<tbody>
<tr>
<td>+r</td>
<td>+t</td>
<td></td>
<td>+l</td>
<td></td>
<td>0.024</td>
</tr>
<tr>
<td>+r</td>
<td>-t</td>
<td></td>
<td>-l</td>
<td></td>
<td>0.056</td>
</tr>
<tr>
<td>-r</td>
<td>+t</td>
<td></td>
<td>+l</td>
<td></td>
<td>0.002</td>
</tr>
<tr>
<td>-r</td>
<td>-t</td>
<td></td>
<td>-l</td>
<td></td>
<td>0.018</td>
</tr>
</tbody>
</table>

\[ P(L|T) \]

<table>
<thead>
<tr>
<th></th>
<th>+l</th>
<th>-l</th>
<th>0.3</th>
<th>0.7</th>
<th>0.1</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>+t</td>
<td>+l</td>
<td>-l</td>
<td>0.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+t</td>
<td>-l</td>
<td>+l</td>
<td>0.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-t</td>
<td>+l</td>
<td>-l</td>
<td>0.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-t</td>
<td>-l</td>
<td>+l</td>
<td>0.9</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Join T

\[ R, T, L \]
Operation 2: Eliminate

- Second basic operation: marginalization
- Take a factor and sum out a variable
  - Shrinks a factor to a smaller one
  - A projection operation
- Example:

\[
P(R, T)
\]

<table>
<thead>
<tr>
<th></th>
<th>+t</th>
<th>-t</th>
</tr>
</thead>
<tbody>
<tr>
<td>+r</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td>+r</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>-r</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>-r</td>
<td>0.81</td>
<td></td>
</tr>
</tbody>
</table>

\[
sum R
\]

\[
P(T)
\]

<table>
<thead>
<tr>
<th></th>
<th>+t</th>
<th>-t</th>
</tr>
</thead>
<tbody>
<tr>
<td>+t</td>
<td>0.17</td>
<td></td>
</tr>
<tr>
<td>-t</td>
<td>0.83</td>
<td></td>
</tr>
</tbody>
</table>
Multiple Elimination

\[ R, T, L \]

\[ T, L \]

\[ L \]

\[ P(R, T, L) \]

<table>
<thead>
<tr>
<th></th>
<th>+t</th>
<th>+l</th>
<th></th>
<th>P(R, T, L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>+r</td>
<td>+t</td>
<td>+l</td>
<td></td>
<td>0.024</td>
</tr>
<tr>
<td>+r</td>
<td>+t</td>
<td>-l</td>
<td></td>
<td>0.056</td>
</tr>
<tr>
<td>+r</td>
<td>-t</td>
<td>+l</td>
<td></td>
<td>0.002</td>
</tr>
<tr>
<td>+r</td>
<td>-t</td>
<td>-l</td>
<td></td>
<td>0.018</td>
</tr>
<tr>
<td>-r</td>
<td>+t</td>
<td>+l</td>
<td></td>
<td>0.027</td>
</tr>
<tr>
<td>-r</td>
<td>+t</td>
<td>-l</td>
<td></td>
<td>0.063</td>
</tr>
<tr>
<td>-r</td>
<td>-t</td>
<td>+l</td>
<td></td>
<td>0.081</td>
</tr>
<tr>
<td>-r</td>
<td>-t</td>
<td>-l</td>
<td></td>
<td>0.729</td>
</tr>
</tbody>
</table>

Sum out R

\[ P(T, L) \]

<table>
<thead>
<tr>
<th></th>
<th>+t</th>
<th>+l</th>
<th></th>
<th>P(T, L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>+t</td>
<td>+t</td>
<td>+l</td>
<td></td>
<td>0.051</td>
</tr>
<tr>
<td>+t</td>
<td>-t</td>
<td>+l</td>
<td></td>
<td>0.119</td>
</tr>
<tr>
<td>-t</td>
<td>+t</td>
<td>+l</td>
<td></td>
<td>0.083</td>
</tr>
<tr>
<td>-t</td>
<td>-t</td>
<td>+l</td>
<td></td>
<td>0.747</td>
</tr>
</tbody>
</table>

Sum out T

\[ P(L) \]

<table>
<thead>
<tr>
<th></th>
<th>+l</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>+l</td>
<td>0.134</td>
<td></td>
</tr>
<tr>
<td>-l</td>
<td>0.886</td>
<td></td>
</tr>
</tbody>
</table>
P(L) : Marginalizing Early!

\[
P(R)
\]

<table>
<thead>
<tr>
<th></th>
<th>+r</th>
<th>-r</th>
</tr>
</thead>
<tbody>
<tr>
<td>+r</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>-r</td>
<td>0.9</td>
<td></td>
</tr>
</tbody>
</table>

Join R

\[
P(T|R)
\]

<table>
<thead>
<tr>
<th></th>
<th>+r</th>
<th>+t</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>+r</td>
<td></td>
<td>+t</td>
<td>0.8</td>
</tr>
<tr>
<td>+r</td>
<td></td>
<td>-t</td>
<td>0.2</td>
</tr>
<tr>
<td>-r</td>
<td>+t</td>
<td></td>
<td>0.1</td>
</tr>
<tr>
<td>-r</td>
<td>-t</td>
<td></td>
<td>0.9</td>
</tr>
</tbody>
</table>

Sum out R

\[
P(R,T)
\]

<table>
<thead>
<tr>
<th></th>
<th>+r</th>
<th>+t</th>
<th>0.08</th>
</tr>
</thead>
<tbody>
<tr>
<td>+r</td>
<td></td>
<td>+t</td>
<td>0.08</td>
</tr>
<tr>
<td>+r</td>
<td></td>
<td>-t</td>
<td>0.02</td>
</tr>
<tr>
<td>-r</td>
<td>+t</td>
<td></td>
<td>0.09</td>
</tr>
<tr>
<td>-r</td>
<td>-t</td>
<td></td>
<td>0.81</td>
</tr>
</tbody>
</table>

\[
P(T)
\]

<table>
<thead>
<tr>
<th></th>
<th>+t</th>
<th>-t</th>
</tr>
</thead>
<tbody>
<tr>
<td>+t</td>
<td>0.17</td>
<td></td>
</tr>
<tr>
<td>-t</td>
<td>0.83</td>
<td></td>
</tr>
</tbody>
</table>

\[
P(L|T)
\]

<table>
<thead>
<tr>
<th></th>
<th>+t</th>
<th>+l</th>
<th>0.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>+t</td>
<td></td>
<td>+l</td>
<td>0.3</td>
</tr>
<tr>
<td>+t</td>
<td></td>
<td>-l</td>
<td>0.7</td>
</tr>
<tr>
<td>-t</td>
<td>+l</td>
<td></td>
<td>0.1</td>
</tr>
<tr>
<td>-t</td>
<td>-l</td>
<td></td>
<td>0.9</td>
</tr>
</tbody>
</table>

\[
P(L|R)
\]

\[
P(L|R,T)
\]

<table>
<thead>
<tr>
<th></th>
<th>+t</th>
<th>+l</th>
<th>0.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>+t</td>
<td></td>
<td>+l</td>
<td>0.3</td>
</tr>
<tr>
<td>+t</td>
<td></td>
<td>-l</td>
<td>0.7</td>
</tr>
<tr>
<td>-t</td>
<td>+l</td>
<td></td>
<td>0.1</td>
</tr>
<tr>
<td>-t</td>
<td>-l</td>
<td></td>
<td>0.9</td>
</tr>
</tbody>
</table>

\[
P(L)
\]

<table>
<thead>
<tr>
<th></th>
<th>+t</th>
<th>+l</th>
<th>0.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>+t</td>
<td></td>
<td>+l</td>
<td>0.3</td>
</tr>
<tr>
<td>+t</td>
<td></td>
<td>-l</td>
<td>0.7</td>
</tr>
<tr>
<td>-t</td>
<td>+l</td>
<td></td>
<td>0.1</td>
</tr>
<tr>
<td>-t</td>
<td>-l</td>
<td></td>
<td>0.9</td>
</tr>
</tbody>
</table>

\[
P(L)
\]

<table>
<thead>
<tr>
<th></th>
<th>+t</th>
<th>+l</th>
<th>0.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>+t</td>
<td></td>
<td>+l</td>
<td>0.3</td>
</tr>
<tr>
<td>+t</td>
<td></td>
<td>-l</td>
<td>0.7</td>
</tr>
<tr>
<td>-t</td>
<td>+l</td>
<td></td>
<td>0.1</td>
</tr>
<tr>
<td>-t</td>
<td>-l</td>
<td></td>
<td>0.9</td>
</tr>
</tbody>
</table>
Marginalizing Early (aka VE*)

\[ P(T) \]

<table>
<thead>
<tr>
<th></th>
<th>+t</th>
<th>-t</th>
</tr>
</thead>
<tbody>
<tr>
<td>+t</td>
<td>0.17</td>
<td></td>
</tr>
<tr>
<td>-t</td>
<td>0.83</td>
<td></td>
</tr>
</tbody>
</table>

\[ P(L|T) \]

<table>
<thead>
<tr>
<th></th>
<th>+t</th>
<th>-t</th>
<th>+l</th>
<th>-l</th>
</tr>
</thead>
<tbody>
<tr>
<td>+t</td>
<td>0.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+t</td>
<td></td>
<td>0.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-t</td>
<td></td>
<td></td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>-t</td>
<td></td>
<td></td>
<td>0.9</td>
<td></td>
</tr>
</tbody>
</table>

\[ P(T,L) \]

<table>
<thead>
<tr>
<th></th>
<th>+t</th>
<th>-t</th>
<th>+l</th>
<th>-l</th>
</tr>
</thead>
<tbody>
<tr>
<td>+t</td>
<td></td>
<td>0.17</td>
<td>+l</td>
<td></td>
</tr>
<tr>
<td>+t</td>
<td>0.83</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-t</td>
<td></td>
<td></td>
<td>+l</td>
<td>0.3</td>
</tr>
<tr>
<td>-t</td>
<td></td>
<td></td>
<td>-l</td>
<td>0.7</td>
</tr>
</tbody>
</table>

\[ P(L) \]

<table>
<thead>
<tr>
<th></th>
<th>+l</th>
<th>-l</th>
</tr>
</thead>
<tbody>
<tr>
<td>+l</td>
<td>0.134</td>
<td></td>
</tr>
<tr>
<td>-l</td>
<td>0.886</td>
<td></td>
</tr>
</tbody>
</table>

* VE is variable elimination
Evidence

- If evidence, start with factors that select that evidence
  - No evidence uses these initial factors:

  \[
  \begin{align*}
  P(R) & & P(T|R) & & P(L|T) \\
  +r & 0.1 & +r & +t & 0.8 & +t & +l & 0.3 \\
  -r & 0.9 & +r & -t & 0.2 & +t & -l & 0.7 \\
  -r & +t & 0.1 & -t & +l & 0.1 \\
  -r & -t & 0.9 & -t & -l & 0.9 \\
  \end{align*}
  \]

- Computing \( P(L|+r) \), the initial factors become:

  \[
  \begin{align*}
  P(+r) & & P(T|+r) & & P(L|T) \\
  +r & 0.1 & +r & +t & 0.8 & +t & +l & 0.3 \\
  +r & -t & 0.2 & +t & -l & 0.7 \\
  -r & +t & 0.1 & -t & +l & 0.1 \\
  -r & -t & 0.9 & -t & -l & 0.9 \\
  \end{align*}
  \]

- We eliminate all vars other than query + evidence
Evidence II

- Result will be a selected joint of query and evidence
  - E.g. for $P(L \mid +r)$, we’d end up with:

<table>
<thead>
<tr>
<th></th>
<th>$+l$</th>
<th>$-l$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$+r$</td>
<td>0.026</td>
<td>0.074</td>
</tr>
<tr>
<td>$+l$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$-l$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Normalize

<table>
<thead>
<tr>
<th></th>
<th>$+l$</th>
<th>$-l$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$+l$</td>
<td>0.26</td>
<td></td>
</tr>
<tr>
<td>$-l$</td>
<td>0.74</td>
<td></td>
</tr>
</tbody>
</table>

- To get our answer, just normalize this!
- That’s it!
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 Bayes Rule

Start / Select

\[ P(B) \]

<table>
<thead>
<tr>
<th>B</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>+b</td>
<td>0.1</td>
</tr>
<tr>
<td>-b</td>
<td>0.9</td>
</tr>
</tbody>
</table>

Add / Remove

\[ P(A|B) \rightarrow P(a|B) \]

<table>
<thead>
<tr>
<th>B</th>
<th>A</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>+b</td>
<td>+a</td>
<td>0.8</td>
</tr>
<tr>
<td>-b</td>
<td>-a</td>
<td>0.2</td>
</tr>
<tr>
<td>+a</td>
<td>-b</td>
<td>0.1</td>
</tr>
<tr>
<td>b</td>
<td>a</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Join on B

\[ a, B \]

\[ P(a, B) \]

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>+a</td>
<td>+b</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td>+a</td>
<td>-b</td>
<td>0.09</td>
<td></td>
</tr>
</tbody>
</table>

Normalize

\[ P(B|a) \]

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>+a</td>
<td>+b</td>
<td>8/17</td>
<td></td>
</tr>
<tr>
<td>+a</td>
<td>-b</td>
<td>9/17</td>
<td></td>
</tr>
</tbody>
</table>
Example

Query: $P(B|j, m)$

Choose A

$P(A|B, E) \times P(j|A) \times P(m|A) \quad \sum \quad P(j, m|B, E)$
Example

Choose E

\[
P(E) \times P(j, m | B, E) \sum P(j, m | B)
\]

Finish with B

\[
P(B) \times P(j, m, B) \text{ Normalize} P(B | j, m)
\]
Exact Inference: Variable Elimination

- Remaining Issues:
  - Complexity: exponential in tree width (size of the largest factor created)
  - Best elimination ordering? NP-hard problem

- What you need to know:
  - Should be able to run it on small examples, understand the factor creation / reduction flow
  - Better than enumeration: saves time by marginalizing variables as soon as possible rather than at the end

- We have seen a special case of VE already
  - HMM Forward Inference
Approximate Inference

- Simulation has a name: sampling
- Sampling is a hot topic in machine learning, and it’s really simple
- 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 a sample is faster than computing the right answer (e.g. with variable elimination)
Prior Sampling

\[ P(C) \]

\[
\begin{array}{c|c}
+c & 0.5 \\
-c & 0.5 \\
\end{array}
\]

\[ P(S|C) \]

\[
\begin{array}{c|c|c}
+c & +s & 0.1 \\
+s & 0.9 \\
-c & +s & 0.5 \\
-s & 0.5 \\
\end{array}
\]

\[ P(R|C) \]

\[
\begin{array}{c|c|c}
+c & +r & 0.8 \\
+r & 0.2 \\
-c & +r & 0.2 \\
-r & 0.8 \\
\end{array}
\]

\[ P(W|S, R) \]

\[
\begin{array}{c|c|c|c}
+s & +r & +w & 0.99 \\
+r & -w & 0.01 \\
+s & -r & +w & 0.90 \\
-r & -w & 0.10 \\
-s & +r & +w & 0.90 \\
+r & -w & 0.10 \\
-s & -r & +w & 0.01 \\
-r & -w & 0.99 \\
\end{array}
\]

Samples:

+\( c \), -S, +r, +w
-\( c \), +S, -r, +w
...

---

Cloudy

Sprinkler

Rain

WetGrass
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 an event be \( N_{PS}(x_1 \ldots x_n) \)

- Then

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

- I.e., the sampling procedure is consistent
Example

- 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| +w)? P(C| +r, +w)? P(C| -r, -w)?
  - Fast: can use fewer samples if less time (what’s the drawback?)
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 $P(C|+s)$
  - 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)
Likelihood Weighting

- Problem with rejection sampling:
  - If evidence is unlikely, you reject a lot of samples
  - You don’t exploit your evidence as you sample
  - Consider $P(B|+a)$

- Idea: fix evidence variables and sample the rest

- Problem: sample distribution not consistent!
- Solution: weight by probability of evidence given parents
Likelihood Weighting

\[ P(C) \]
\[ \begin{array}{c|c}
  +c & 0.5 \\
  -c & 0.5 \\
\end{array} \]

\[ P(S|C) \]
\[ \begin{array}{c|c}
  +c & +s \ 0.1 \\
    & -s \ 0.9 \\
-\!c & +s \ 0.5 \\
    & -s \ 0.5 \\
\end{array} \]

\[ P(R|C) \]
\[ \begin{array}{c|c}
  +c & +r \ 0.8 \\
    & -r \ 0.2 \\
-\!c & +r \ 0.2 \\
    & -r \ 0.8 \\
\end{array} \]

\[ P(W|S, R) \]
\[ \begin{array}{c|c|c|c}
  & +r & +w & 0.99 \\
  & -r & -w & 0.01 \\
+\!s & +w & 0.90 \\
  & -w & 0.10 \\
-\!s & +w & 0.90 \\
  & -w & 0.10 \\
\end{array} \]

Samples:
\[ +c, +s, +r, +w \]
\[ \ldots \]
\[ w = 1.0 \times 0.1 \times 0.99 \]
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
  - E.g. here, W’s value will get picked based on the evidence values of S, R
  - More of our samples will reflect the state of the world suggested by the evidence

- **Likelihood weighting doesn’t solve all our problems**
  - Evidence influences the choice of downstream variables, but not upstream ones (C isn’t more likely to get a value matching the evidence)

- We would like to consider evidence when we sample every variable
Markov Chain Monte Carlo*

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

- **Gibbs Sampling**: resample one variable at a time, conditioned on the rest, but keep evidence fixed.

- **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.