Hidden Markov Models

Course Webpage:
http://www.cs.cornell.edu/Courses/cs4786/2017fa/
Same example:

But you don’t observe location (dark room)

You hear how close the bot is!

What you hear:
Same example:

But you don’t observe location (dark room)

You hear how close the bot is!

What you hear:

Can you catch the Bot?
Hidden Markov Model (HMM)

$X_t$'s are what you hear (observation)
$S_t$'s are the unseen locations (states)

Eg: for $n \times n$ grid we have, $K = n^2$ states

Number of alphabets = 5
(colors you can observe)
**Hidden Markov Model (HMM)**

- $S_1$, $S_2$, $S_3$ are the unseen locations (states)
- $X_1$, $X_2$, $X_3$ are what you hear (observation)

**Example:**

For an $n \times n$ grid we have, $K = n^2$ states

Number of alphabets = 5
(colors you can observe)
Hidden Markov Model (HMM)
What are the parameters?
What are the parameters?

**Transition Probability table:** $T = \text{P}(S_t|S_{t-1})$

**Emission Probabilities:** $E = \text{P}(X_t|S_t)$

**Initial State Probabilities:** $P(S_1)$
Hidden Markov Model (HMM)
• What is probability that bot will be in location $k$ at time $t$ given the entire sequence of observations?
• What is probability that bot will be in location \( k \) at time \( t \) given the entire sequence of observations?

\[
P(S_t = k | X_1, \ldots, X_N)\]
Inference in HMM

\[ P(S_t = k | X_1, \ldots, X_N) \]
Inference in HMM

\[ P(S_t = k | X_1, \ldots, X_N) \]
\[ \propto P(X_{t+1}, \ldots, X_N | S_t = k, X_1, \ldots, X_t) P(S_t = k | X_1, \ldots, X_t) \]
\[ P(S_t = k | X_1, \ldots, X_N) \]
\[ \propto P(X_{t+1}, \ldots, X_N | S_t = k, X_1, \ldots, X_t) P(S_t = k | X_1, \ldots, X_t) \]
\[ \propto P(X_{t+1}, \ldots, X_N | S_t = k, X_1, \ldots, X_t) P(S_t = k, X_1, \ldots, X_t) \]
Inference in HMM

\[ P(S_t = k | X_1, \ldots, X_N) \]
\[ \propto P(X_{t+1}, \ldots, X_N | S_t = k, X_1, \ldots, X_t) P(S_t = k | X_1, \ldots, X_t) \]
\[ \propto P(X_{t+1}, \ldots, X_N | S_t = k, X_1, \ldots, X_t) P(S_t = k, X_1, \ldots, X_t) \]
\[ \propto P(X_{t+1}, \ldots, X_N | S_t = k, X_1, \ldots, X_t) P(X_t | S_t = k, X_1, \ldots, X_{t-1}) P(S_t = k, X_1, \ldots, X_{t-1}) \]
\[ P(S_t = k | X_1, \ldots, X_N) \]
\[ \propto P(X_{t+1}, \ldots, X_N | S_t = k, X_1, \ldots, X_t) P(S_t = k | X_1, \ldots, X_t) \]
\[ \propto P(X_{t+1}, \ldots, X_N | S_t = k, X_1, \ldots, X_t) P(S_t = k, X_1, \ldots, X_t) \]
\[ \propto P(X_{t+1}, \ldots, X_N | S_t = k, X_1, \ldots, X_t) P(X_t | S_t = k, X_1, \ldots, X_{t-1}) P(S_t = k, X_1, \ldots, X_{t-1}) \]
\[ \propto P(X_{t+1}, \ldots, X_N | S_t = k) P(X_t | S_t = k) P(S_t = k, X_1, \ldots, X_{t-1}) \]
Inference in HMM

\[ P(S_t = k|X_1, \ldots, X_N) \]
\[ \propto P(X_{t+1}, \ldots, X_N|S_t = k, X_1, \ldots, X_t)P(S_t = k|X_1, \ldots, X_t) \]
\[ \propto P(X_{t+1}, \ldots, X_N|S_t = k, X_1, \ldots, X_t)P(S_t = k, X_1, \ldots, X_t) \]
\[ \propto P(X_{t+1}, \ldots, X_N|S_t = k, X_1, \ldots, X_t)P(X_t|S_t = k, X_1, \ldots, X_{t-1})P(S_t = k, X_1, \ldots, X_{t-1}) \]
\[ \propto P(X_{t+1}, \ldots, X_N|S_t = k)P(X_t|S_t = k)P(S_t = k, X_1, \ldots, X_{t-1}) \]

We know \( P(X_t|S_t = k) \)'s and \( P(S_t|S_{t-1}) \)

Compute \( P(X_{t+1}, \ldots, X_N) \) and \( P(S_t = k, X_1, \ldots, X_{t-1}) \) recursively.
Inference in HMM

\[
message_{S_{t-1} \rightarrow S_t}(k) = P(S_t = k, X_1, \ldots, X_{t-1})
\]

\[
message_{S_{t+1} \rightarrow S_t}(k) = P(X_n, \ldots, X_{t+1} | S_t = k)
\]

\[
P(S_t = k | X_1, \ldots, X_n) \propto message_{S_{t-1} \rightarrow S_t}(k) \times message_{S_{t+1} \rightarrow S_t}(k) \times P(X_t | S_t = k)
\]
Inference in HMM

message_{S_{t-1} \rightarrow S_t}(k) = P(S_t = k, X_1, \ldots, X_{t-1})

message_{S_{t+1} \rightarrow S_t}(k) = P(X_n, \ldots, X_{t+1} | S_t = k)
**Inference in HMM**

\[
\text{message}_{S_{t-1} \rightarrow S_t}(k) = P(S_t = k, X_1, \ldots, X_{t-1})
\]

\[
\text{message}_{S_{t+1} \rightarrow S_t}(k) = P(X_n, \ldots, X_{t+1} | S_t = k)
\]

**Forward:**

\[
P(X_1, \ldots, X_{t-1}, S_t = k) = \sum_{j=1}^{K} P(S_t = k | S_{t-1} = j)P(X_{t-1} | S_{t-1} = j)P(X_1, \ldots, X_{t-2}, S_{t-1} = j)
\]
Inference in HMM

Forward:

\[
P(X_1, \ldots, X_{t-1}, S_t = k) = \sum_{j=1}^{K} P(S_t = k | S_{t-1} = j) P(X_{t-1} | S_{t-1} = j) P(X_1, \ldots, X_{t-2}, S_{t-1} = j)
\]

\[
message_{S_{t-1} \rightarrow S_t}(k) = \sum_{j=1}^{K} P(S_t = k | S_{t-1} = j) P(X_{t-1} | S_{t-1} = j) message_{S_{t-2} \rightarrow S_{t-1}}(j)
\]
Inference in HMM

message_{s_{t-1} \rightarrow s_t}(k) = P(s_t = k, X_1, \ldots, X_{t-1})

message_{s_{t+1} \rightarrow s_t}(k) = P(X_n, \ldots, X_{t+1} | s_t = k)
Inference in HMM

\[ \text{message}_{S_{t-1} \rightarrow S_t}(k) = P(S_t = k, X_1, \ldots, X_{t-1}) \]

\[ \text{message}_{S_{t+1} \rightarrow S_t}(k) = P(X_n, \ldots, X_{t+1}\mid S_t = k) \]

Backward:

\[ P(X_n, \ldots, X_{t+1}\mid S_t = k) = \sum_{j=1}^{K} P(X_n, \ldots, X_{t+2}\mid S_{t+1} = j)P(X_{t+1}\mid S_{t+1} = j)P(S_{t+1} = j\mid S_t = k) \]
Inference in HMM

message $S_{t-1} \mapsto S_t(k) = P(S_t = k, X_1, \ldots, X_{t-1})$

message $S_{t+1} \mapsto S_t(k) = P(X_n, \ldots, X_{t+1} | S_t = k)$

Backward:

$$P(X_n, \ldots, X_{t+1}|S_t = k) = \sum_{j=1}^{K} P(X_n, \ldots, X_{t+2}|S_{t+1} = j)P(X_{t+1}|S_{t+1} = j)P(S_{t+1} = j|S_t = k)$$

message $S_{t+1} \mapsto S_t(k) = \sum_{j=1}^{K} \text{message}_{S_{t+2} \mapsto S_{t+1}}(j)P(X_{t+1}|S_{t+1} = j)P(S_{t+1} = j|S_t = k)$
Now that we have algorithm for inference, what about learning

Given observations, how do we estimate parameters for HMM? Three guesses …
EM FOR HMM (BAUM WELCH)

- EM algorithm of course, for HMM its referred to as Baum Welch algorithm
- Initialize Transition and Emission probability tables arbitrarily
- For $i = 1$ to convergence:
  
  **E-step** For every state variable $t \in \{1, \ldots, n\}$, use forward-backward algorithm to compute probabilities of latent variables given observation
  
  **M-step** Optimize weighted log likelihood as usual:

$$\theta^{(i)} = \arg \max_{\theta \in \Theta} \sum_{S_1,\ldots,n} P(S_1,\ldots,n|X_1,\ldots,n, \theta^{(i-1)}) \log P(X_1,\ldots,n, S_1,\ldots,n|\theta)$$
\[ \log P(X_1, \ldots, n, S_1, \ldots, n | \theta) : \]
\[
\log P(X_1,...,n, S_1,...,n | \theta) = \log \left( \prod_{t=1}^{n} P(X_t|S_t, \theta) \prod_{t=1}^{n} P(S_t|S_{t-1}, \theta) \right)
\]


\[ \log P(X_1, \ldots, n, S_1, \ldots, n | \theta) = \log \left( \prod_{t=1}^{n} P(X_t|S_t, \theta) \prod_{t=1}^{n} P(S_t|S_{t-1}, \theta) \right) \]

\[ = \sum_{t=1}^{n} \log P(X_t|S_t, \theta) + \sum_{t=1}^{n} \log P(S_t|S_{t-1}, \theta) \]
\[
\log P(X_1,\ldots,n, S_1,\ldots,n|\theta) = \log \left( \prod_{t=1}^{n} P(X_t|S_t, \theta) \prod_{t=1}^{n} P(S_t|S_{t-1}, \theta) \right) \\
= \sum_{t=1}^{n} \log P(X_t|S_t, \theta) + \sum_{t=1}^{n} \log P(S_t|S_{t-1}, \theta)
\]

Hence,

\[
\sum_{S_1,\ldots,n} P(S_1,\ldots,n|X_1,\ldots,n, \theta^{(i-1)}) \log P(X_1,\ldots,n, S_1,\ldots,n|\theta) \\
= \sum_{t=1}^{n} \sum_{s_t=1}^{K} P(S_t = s_t|X_1,\ldots,n, \theta^{i-1}) \log P(X_t|S_t = s_t, \theta) \\
+ \sum_{t=1}^{n} \sum_{s_t,s_{t-1}=1}^{K} P(S_t = s_t, S_{t-1} = s_{t-1}|X_1,\ldots,n, \theta^{i-1}) \log P(S_t|S_{t-1}, \theta)
\]
E-STEP

Only need to compute \( P(St = st, St - 1 = st - 1, X_1, \ldots, n, i - 1) \) and \( P(S_t = st, St - 1 = st - 1, X_1, \ldots, n, i - 1) \) using forward-backward.

First term is immediate

\[
P(S_t = st, X_1, \ldots, n, i - 1) \propto m_{St - 1} S_{St}(st) m_{St + 1} S_{St}(st) \cdot E_{i - 1}[st, X_t]
\]

For second term,

\[
P(S_t - 1 = st - 1, X_1, \ldots, n, i - 1) \propto m_{St - 1} S_{St}(st - 1) T_{i - 1}[st - 1, st] m_{St - 2} S_{St - 1}(st - 1) \cdot E_{i - 1}[st - 1, X_{t - 1}]
\]

Why?
E-STEP

- Only need to compute \( P(S_t = s_t|X_1,\ldots,n, \theta^{i-1}) \) and \( P(S_t = s_t, S_{t-1} = s_{t-1}|X_1,\ldots,n, \theta^{i-1}) \) using forward-backward
**E-step**

- Only need to compute \( P(S_t = s_t|X_1,...,n, \theta^{i-1}) \) and \( P(S_t = s_t, S_{t-1} = s_{t-1}|X_1,...,n, \theta^{i-1}) \) using forward-backward

- First term is immediate

\[
P(S_t = s_t|X_1,...,n, \theta^{i-1}) \propto m_{S_{t-1} \rightarrow S_t}(s_t) \cdot m_{S_{t+1} \rightarrow S_t}(s_t) \cdot E^{(i-1)}[s_t, X_t]
\]
Only need to compute $P(S_t = s_t|X_1,...,n, \theta^{i-1})$ and $P(S_t = s_t, S_{t-1} = s_{t-1}|X_1,...,n, \theta^{i-1})$ using forward-backward

First term is immediate

$$P(S_t = s_t|X_1,...,n, \theta^{i-1}) \propto m_{S_{t-1} \rightarrow s_t}(s_t) \cdot m_{S_{t+1} \rightarrow s_t}(s_t) \cdot E(i-1)[s_t, X_t]$$

For second term,

$$P(S_t = s_i, S_{t-1} = s_{t-1}|X_1,...,n, \theta^{i-1})$$

$$\propto m_{S_{t-1} \rightarrow s_t}(s_t)T^{(i-1)}[s_{t-1}, s_t]P(S_{t-1} = s_{t-1}|X_1,...,n, \theta^{i-1})$$

$$\propto m_{S_{t-1} \rightarrow s_t}(s_t)T^{(i-1)}[s_{t-1}, s_t]m_{S_{t-2} \rightarrow s_{t-1}}(s_{t-1})m_{S_t \rightarrow s_{t-1}}(s_{t-1})E(i-1)[s_{t-1}, X_{t-1}]$$

Why?
\[ P(S_t = s_t, S_{t-1} = s_{t-1} | X_1, \ldots, n, \theta^{i-1}) \]
\[
P(S_t = s_t, S_{t-1} = s_{t-1}|X_1,\ldots,n, \theta^{i-1}) \\
= P(S_t = s_t, |S_{t-1} = s_{t-1}, X_1,\ldots,n, \theta^{t-1})P(S_{t-1} = s_{t-1}|X_1,\ldots,n, \theta^{i-1})
\]
\[ P(S_t = s_t, S_{t-1} = s_{t-1}|X_1,\ldots,n, \theta^{i-1}) \]
\[ = P(S_t = s_t, |S_{t-1} = s_{t-1}, X_1,\ldots,n, \theta^{t-1})P(S_{t-1} = s_{t-1}|X_1,\ldots,n, \theta^{i-1}) \]
\[ = P(S_t = s_t, |S_{t-1} = s_{t-1}, X_t,\ldots,n, \theta^{i-1})P(S_{t-1} = s_{t-1}|X_1,\ldots,n, \theta^{i-1}) \]
\[
P(S_t = s_t, S_{t-1} = s_{t-1}| X_1, ..., n, \theta^{i-1}) \\
= P(S_t = s_t, | S_{t-1} = s_{t-1}, X_1, ..., n, \theta^{t-1}) P(S_{t-1} = s_{t-1}| X_1, ..., n, \theta^{i-1}) \\
= P(S_t = s_t, | S_{t-1} = s_{t-1}, X_t, ..., n, \theta^{i-1}) P(S_{t-1} = s_{t-1}| X_1, ..., n, \theta^{i-1}) \\
\propto P(X_t, ..., n| S_t = s_t, S_{t-1} = s_{t-1}, \theta^{i-1}) \\
P(S_t = s_t| S_{t-1} = s_{t-1}, \theta^{i-1}) P(S_{t-1} = s_{t-1}| X_1, ..., n, \theta^{i-1})
\]
E-STEP

\[ P(S_t = s_t, S_{t-1} = s_{t-1} | X_1, \ldots, n, \theta^{i-1}) \]

\[ = P(S_t = s_t, | S_{t-1} = s_{t-1}, X_1, \ldots, n, \theta^{t-1}) P(S_{t-1} = s_{t-1} | X_1, \ldots, n, \theta^{i-1}) \]

\[ = P(S_t = s_t, | S_{t-1} = s_{t-1}, X_t, \ldots, n, \theta^{i-1}) P(S_{t-1} = s_{t-1} | X_1, \ldots, n, \theta^{i-1}) \]

\[ \propto P(X_t, \ldots, n | S_t = s_t, S_{t-1} = s_{t-1}, \theta^{i-1}) \]

\[ P(S_t = s_t | S_{t-1} = s_{t-1}, \theta^{i-1}) P(S_{t-1} = s_{t-1} | X_1, \ldots, n, \theta^{i-1}) \]

\[ \propto P(X_t, \ldots, n | S_t = s_t, \theta^{i-1}) \]

\[ T^{(i-1)}[s_{t-1}, s_t] P(S_{t-1} = s_{t-1} | X_1, \ldots, n, \theta^{i-1}) \]
\[ P(S_t = s_t, S_{t-1} = s_{t-1}|X_1,...,n, \theta^{i-1}) \]
\[ = P(S_t = s_t, S_{t-1} = s_{t-1}, X_1,...,n, \theta^t) P(S_{t-1} = s_{t-1}|X_1,...,n, \theta^{i-1}) \]
\[ = P(S_t = s_t, S_{t-1} = s_{t-1}, X_t,...,n, \theta^{i-1}) P(S_{t-1} = s_{t-1}|X_1,...,n, \theta^{i-1}) \]
\[ \propto P(X_t,...,n|S_t = s_t, S_{t-1} = s_{t-1}, \theta^{i-1}) \]
\[ P(S_t = s_t|S_{t-1} = s_{t-1}, \theta^{i-1}) P(S_{t-1} = s_{t-1}|X_1,...,n, \theta^{i-1}) \]
\[ \propto P(X_t,...,n|S_t = s_t, \theta^{i-1}) \]
\[ T^{(i-1)}[s_{t-1}, s_t] P(S_{t-1} = s_{t-1}|X_1,...,n, \theta^{i-1}) \]
\[ \propto m_{S_{t-1} \rightarrow s_t}(s_t) \cdot T^{(i-1)}[s_{t-1}, s_t] \cdot P(S_{t-1} = s_{t-1}|X_1,...,n, \theta^{i-1}) \]
E-step

\[
P(S_t = s_t, S_{t-1} = s_{t-1}|X_1,\ldots,n, \theta^{i-1}) \\
= P(S_t = s_t, |S_{t-1} = s_{t-1}, X_1,\ldots,n, \theta^{t-1})P(S_{t-1} = s_{t-1}|X_1,\ldots,n, \theta^{i-1}) \\
= P(S_t = s_t, |S_{t-1} = s_{t-1}, X_t,\ldots,n, \theta^{i-1})P(S_{t-1} = s_{t-1}|X_1,\ldots,n, \theta^{i-1}) \\
\propto P(X_t,\ldots,n|S_t = s_t, S_{t-1} = s_{t-1}, \theta^{i-1}) \\
\propto P(X_t,\ldots,n|S_t = s_t, \theta^{i-1}) \\
T^{(i-1)}[s_{t-1}, s_t]P(S_{t-1} = s_{t-1}|X_1,\ldots,n, \theta^{i-1}) \\
\propto m_{S_{t-1}\mapsto s_t}(s_t) \cdot T^{(i-1)}[s_{t-1}, s_t] \cdot P(S_{t-1} = s_{t-1}|X_1,\ldots,n, \theta^{i-1}) \\
\propto m_{S_{t-1}\mapsto s_t}(s_t) \cdot T^{(i-1)}[s_{t-1}, s_t] \\
m_{S_{t-2}\mapsto s_{t-1}}(s_{t-1}) \cdot m_{S_t\mapsto s_{t-1}}(s_{t-1}) \cdot E^{(i-1)}[s_{t-1}, X_{t-1}]}
\]
Baum Welch Algorithm

Initialize $T_0$, $E_0$ probability tables

For $i = 1$ to convergence

E-step: Run Forward-Backward algorithm and compute messages

For every $t$ compute $P(S_t = s_t, S_{t-1} = s_{t-1} \mid X_1, ..., n, ✓_{i-1})$ and $P(S_t = s_t \mid X_1, ..., n, ✓_{i-1})$ as in previous slides

M-step: $\forall u, v T(i) [u, v] = \sum_{n \geq 2} P(S_t = v, S_{t-1} = u \mid X_1, ..., n, ✓_{i-1})$ $\sum_{n \geq 2} P(S_{t-1} = u \mid X_1, ..., n, ✓_{i-1})$

$\forall v, e E(i) [v, e] = \sum_{n \geq 1} P(S_t = v \mid X_1, ..., n, ✓_{i-1}) \cdot 1_{X_t = e}$ $\sum_{n \geq 1} P(S_t = v \mid X_1, ..., n, ✓_{i-1})$
Baum Welch Algorithm

Initialize $T^0, E^0$ probability tables
Initialize $T^0, E^0$ probability tables
For $i = 1$ to convergence
Initialize $T^0, E^0$ probability tables
For $i = 1$ to convergence
  • E-step:
    • Run Forward-Backward algorithm and compute messages
Initialize $T^0$, $E^0$ probability tables
For $i = 1$ to convergence
  - E-step:
    - Run Forward-Backward algorithm and compute messages
    - For every $t$ compute $P(S_t = s_t, S_{t-1} = s_{t-1}|X_1, ..., n, \theta^{i-1})$ and
      $P(S_t = s_t|X_1, ..., n, \theta^{i-1})$ as in previous slides
Initialize $T^0, E^0$ probability tables
For $i = 1$ to convergence
  - **E-step:**
    - Run Forward-Backward algorithm and compute messages
    - For every $t$ compute $P(S_t = s_t, S_{t-1} = s_{t-1}|X_1,...,n, \theta^{i-1})$ and
      $P(S_t = s_t|X_1,...,n, \theta^{i-1})$ as in previous slides
  - **M-step:**

\[
\forall u, v \quad T^{(i)}[u, v] = \frac{\sum_{t=2}^{n} P(S_t = v, S_{t-1} = u|X_1,...,n, \theta^{i-1})}{\sum_{t=2}^{n} P(S_{t-1} = u|X_1,...,n, \theta^{i-1})}
\]

\[
\forall v, e \quad E^{(i)}[v, e] = \frac{\sum_{t=1}^{n} P(S_t = v|X_1,...,n, \theta^{i-1}) \cdot \mathbf{1}_{X_t = e}}{\sum_{t=1}^{n} P(S_t = v|X_1,...,n, \theta^{i-1})}
\]
Inference for general BN
Bayesian Networks

- Directed acyclic graph (DAG): $G = (V, E)$
- Joint distribution $P_\theta$ over $X_1, \ldots, X_n$ that factorizes over $G$:
  $$P_\theta(X_1, \ldots, X_n) = \prod_{i=1}^{N} P_\theta(X_i|\text{Parent}(X_i))$$
- Hence Bayesian Networks are specified by $G$ along with CPD’s over the variables (given their parents)
Variable Elimination: Examples

- Marginals are enough:

\[ P(X_j = x_j, X_k = x_k | X_i = x_i, X_h = x_h) = \frac{P(X_j = x_j, X_k = x_k, X_i = x_i, X_h = x_h)}{P(X_i = x_i, X_h = x_h)} \]
VARIABLE ELIMINATION: EXAMPLES

\[
\begin{align*}
X_1 &\rightarrow X_2 & X_3 \\
&\rightarrow X_4
\end{align*}
\]
Variable Elimination: Examples

\[ P(X_4) = \sum_{x_1} \sum_{x_2} \sum_{x_3} P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4) \]
Variable Elimination: Examples

\[ P(X_4) = \sum_{x_1} \sum_{x_2} \sum_{x_3} P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4) \]

\[ = \sum_{x_1} \left( P(X_1 = x_1) \sum_{x_2} \left( P(X_2 = x_2 | X_1 = x_1) P(X_4 | X_2 = x_2) \left( \sum_{x_3} P(X_3 = x_3 | X_2 = x_2) \right) \right) \right) \]
**Variable Elimination: Examples**

\[ P(X_4) = \sum_{x_1} \sum_{x_2} \sum_{x_3} P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4) \]

\[ = \sum_{x_1} \left( P(X_1 = x_1) \sum_{x_2} \left( P(X_2 = x_2|X_1 = x_1)P(X_4|X_2 = x_2) \left( \sum_{x_3} P(X_3 = x_3|X_2 = x_2) \right) \right) \right) \]

\[ = \sum_{x_1} \left( P(X_1 = x_1) \left( \sum_{x_2} P(X_2 = x_2|X_1 = x_1)P(X_4|X_2 = x_2) \right) \right) \]
Initialize \texttt{List} with conditional probability distributions

Pick an order of elimination $I$ for remaining variables

\textbf{For} each $X_i \in I$

\hspace{1em} Find distributions in \texttt{List} containing variable $X_i$ and remove them

\hspace{1em} Define new distribution as the sum (over values of $X_i$) of the product of these distributions

\hspace{1em} Place the new distribution on \texttt{List}

\textbf{End}

Return \texttt{List}
Variable Elimination: Order Matters

Right order: $O(n)$

Wrong order: $O(2^n)$