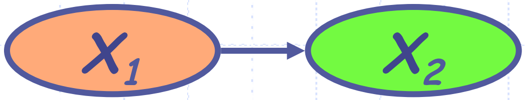
# VARIABLE ELIMINATION

# Approaches to inference

- Exact inference algorithms
  - The variable elimination algorithm
  - The junction tree algorithms (not covered)
- Approximate inference techniques
  - Stochastic simulation / sampling methods
  - Markov chain Monte Carlo methods
  - Variational algorithms (not covered)

# Inference in Simple Chains

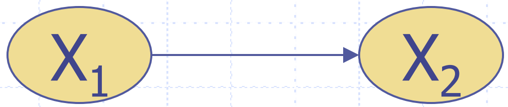



How do we compute $P(X_2)$?

$$P(x_2) = \sum_{x_1} P(x_1, x_2) = \sum_{x_1} P(x_1) P(x_2 \mid x_1)$$

# BN Inference

- Simplest Case:



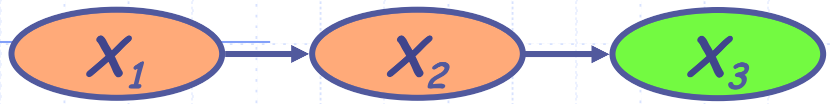

Compute $P(X_2)$:

$$P(x_2) = P(x_1)P(x_2|x_1) + P(\sim x_1)P(x_2|\sim x_1)$$

$$P(\sim x_2) = P(x_1)P(\sim x_2|x_1) + P(\sim x_1)P(\sim x_2|\sim x_1)$$

$$P(x_2) = \sum_{x_1} P(x_1)P(x_2 \mid x_1)$$

# Inference in Simple Chains (cont.)



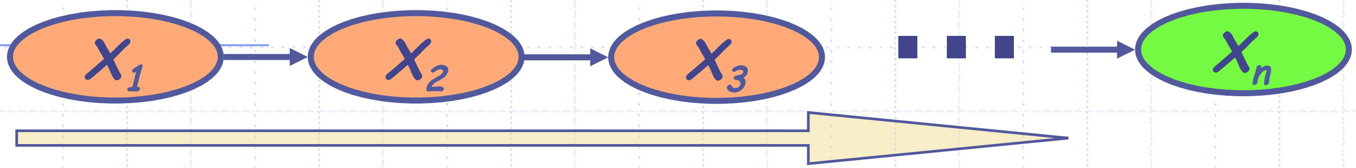

How do we compute *$P(X_3)$*?

$$P(x_3) = \sum_{x_2} P(x_2, x_3) = \sum_{x_2} P(x_2) P(x_3 \mid x_2)$$

- we already know how to compute *$P(X_2)$*...

$$P(x_2) = \sum_{x_1} P(x_1, x_2) = \sum_{x_1} P(x_1) P(x_2 \mid x_1)$$

# Inference in Simple Chains (cont.)




How do we compute $P(X_n)$?

- Compute $P(X_1)$, $P(X_2)$, $P(X_3)$, …
- We compute each term by using the previous one

$$P(x_{i+1}) = \sum_{x_i} P(x_i) P(x_{i+1} \mid x_i)$$

Complexity:

- Each step costs $O(|Val(X_i)| * |Val(X_{i+1})|)$ operations
- Compare to naïve evaluation, that requires summing over joint values of $n-1$ variables

# Elimination in Chains

- We now try to understand the simple chain example using first principles

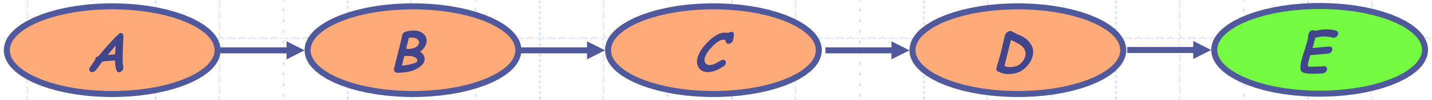
$$A \to B \to C \to D \to E$$

- Using definition of probability, we have

$$P(e) = \sum_{d}\sum_{c}\sum_{b}\sum_{a} P(a,b,c,d,e)$$

a naïve summation needs to enumerate over an exponential number of terms

# Elimination in Chains




- By chain decomposition, we get

$$P(e) = \sum_d \sum_c \sum_b \sum_a P(a,b,c,d,e)$$

$$= \sum_d \sum_c \sum_b \sum_a P(a)P(b \mid a)P(c \mid b)P(d \mid c)P(e \mid d)$$

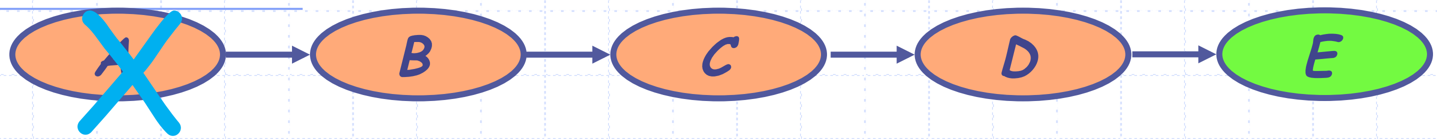
# Elimination in Chains

$$A \rightarrow B \rightarrow C \rightarrow D \rightarrow E$$

- Rearranging terms ...

$$
\begin{aligned}
P(e) &= \sum_d \sum_c \sum_b \sum_a P(a)P(b \mid a)P(c \mid b)P(d \mid c)P(e \mid d) \\
&= \sum_d \sum_c \sum_b P(c \mid b)P(d \mid c)P(e \mid d) \sum_a P(a)P(b \mid a)
\end{aligned}
$$

# Elimination in Chains



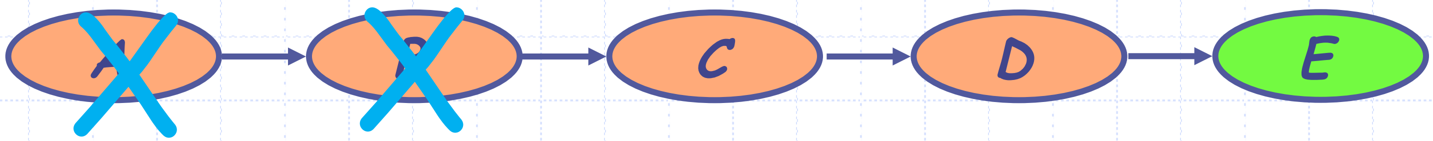

- Now we can perform innermost summation

$$P(e) = \sum_d \sum_c \sum_b P(c \mid b) P(d \mid c) P(e \mid d) \underbrace{\sum_a P(a) P(b \mid a)}$$

$$= \sum_d \sum_c \sum_b P(c \mid b) P(d \mid c) P(e \mid d) p(b)$$

- This summation, is exactly the first step in the forward iteration we describe before

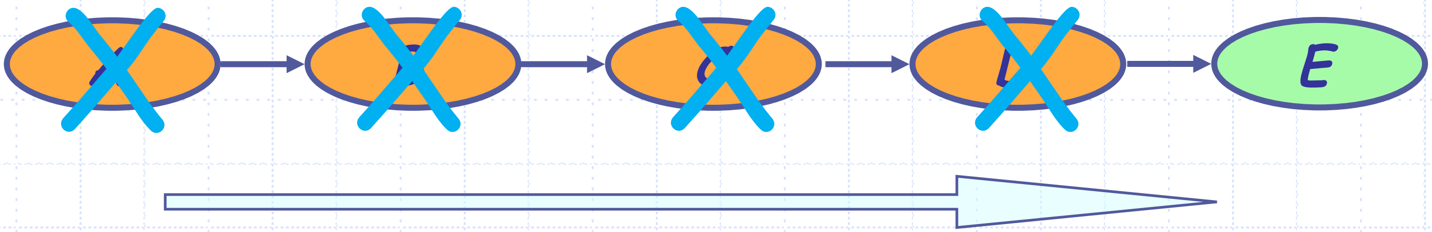
# Elimination in Chains




- Rearranging and then summing again, we get

$$
\begin{aligned}
P(e) &= \sum_d \sum_c \sum_b P(c \mid b) P(d \mid c) P(e \mid d) p(b) \\
&= \sum_d \sum_c P(d \mid c) P(e \mid d) \sum_b P(c \mid b) p(b) \\
&= \sum_d \sum_c P(d \mid c) P(e \mid d) p(c)
\end{aligned}
$$

# Elimination in Chains




- Eliminate nodes one by one all the way to the end, we get

$$P(e) = \sum_d P(e \mid d) p(d)$$

# General Inference w/ Variable Elimination

## General idea:

- Write query in the form

$$P(X_1, \boldsymbol{e}) = \sum_{x_n} \cdots \sum_{x_3} \sum_{x_2} \prod_i P(x_i \mid pa_i)$$

  - this suggests an "elimination order" of variables to be marginalized

- Iteratively
  - Move all irrelevant terms outside of innermost sum
  - Perform innermost sum, getting a new term
  - Insert the new term into the product

- wrap-up

$$P(X_1 \mid \boldsymbol{e}) = \frac{P(X_1, \boldsymbol{e})}{P(\boldsymbol{e})}$$

# Variable Elimination

General idea:

- Write query in the form

$$P(X_n, \mathbf{e}) = \sum_{x_k} \cdots \sum_{x_3} \sum_{x_2} \prod_i P(x_i \mid pa_i)$$

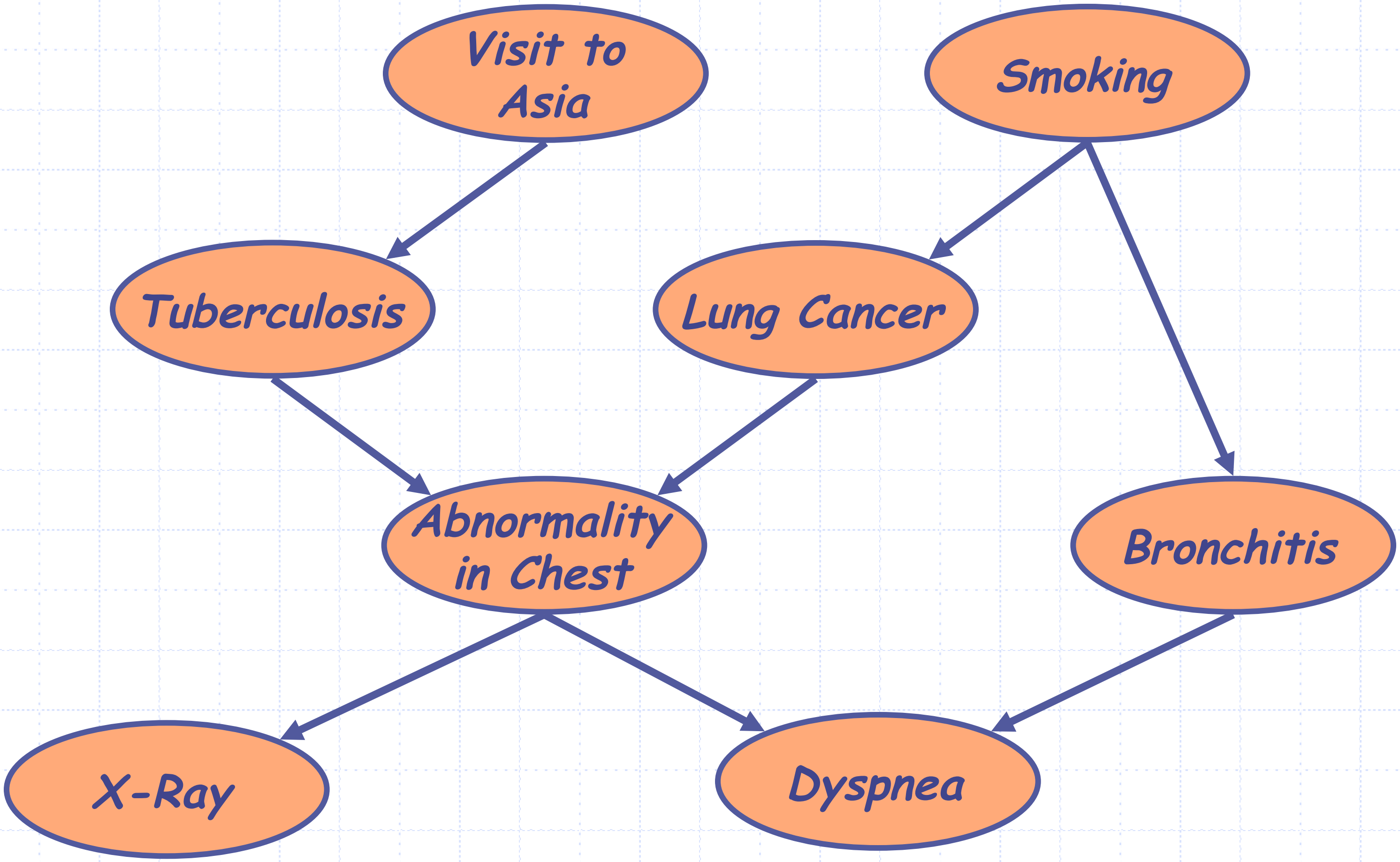
- Iteratively
  - Move all irrelevant terms outside of innermost sum
  - Perform innermost sum, getting a new term
  - Insert the new term into the product

# Variable Elimination Algorithm

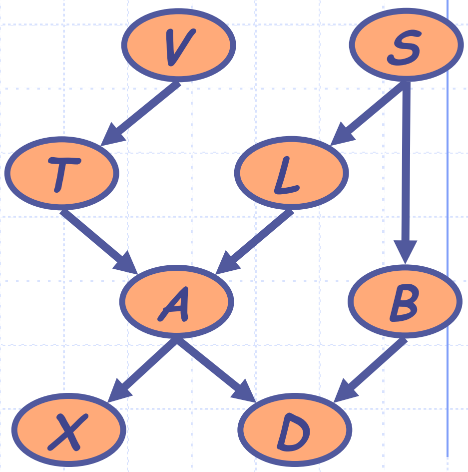
- Let X1,…, Xm be an ordering on the non-query variables $$\sum_{X_1}\sum_{X_2}\cdots\sum_{X_m}\prod_j P(X_j \mid \text{Parents}(X_j))$$
- For I = m, …, 1
  - Leave in the summation for Xi only factors mentioning Xi
  - Multiply the factors, getting a factor that contains a number for each value of the variables mentioned, including Xi
  - Sum out Xi, getting a factor f that contains a number for each value of the variables mentioned, not including Xi
  - Replace the multiplied factor in the summation

# A More Complex Example

- "Asia" network:



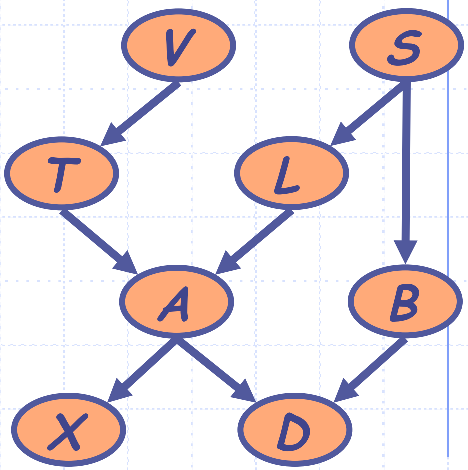
- We want to compute *P(d)*
- Need to eliminate: *v,s,x,t,l,a,b*




Initial factors

$$P(v)P(s)P(t \mid v)P(l \mid s)P(b \mid s)P(a \mid t,l)P(x \mid a)P(d \mid a,b)$$

- We want to compute *P(d)*
- Need to eliminate: *v,s,x,t,l,a,b*




## Initial factors

$$P(v)P(s)P(t \mid v)P(l \mid s)P(b \mid s)P(a \mid t,l)P(x \mid a)P(d \mid a,b)$$

Eliminate: $v$

Compute:

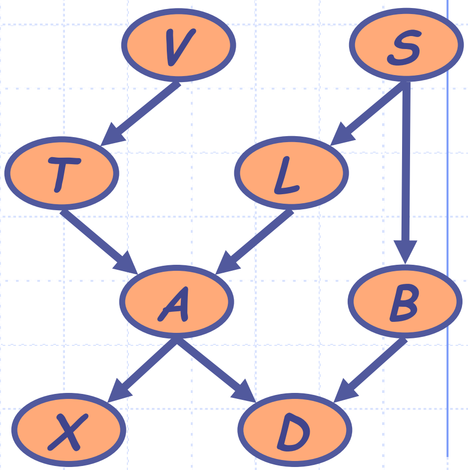
$$f_v(t) = \sum_v P(v)P(t \mid v)$$

$$\Rightarrow f_v(t)P(s)P(l \mid s)P(b \mid s)P(a \mid t,l)P(x \mid a)P(d \mid a,b)$$

Note: $f_v(t) = P(t)$

In general, result of elimination is not necessarily a probability term

- We want to compute *P(d)*
- Need to eliminate: *s,x,t,l,a,b*




- Initial factors

$$P(v)P(s)P(t \mid v)P(l \mid s)P(b \mid s)P(a \mid t,l)P(x \mid a)P(d \mid a,b)$$

$$\Rightarrow f_v(t)\underline{P(s)}\,\underline{P(l \mid s)}\,\underline{P(b \mid s)}P(a \mid t,l)P(x \mid a)P(d \mid a,b)$$

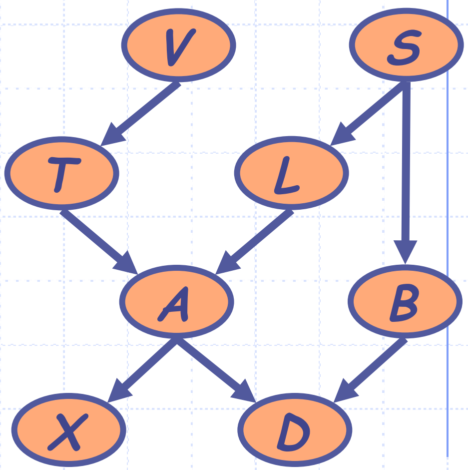
Eliminate: $s$

Compute: $$f_s(b,l) = \sum_s P(s)P(b \mid s)P(l \mid s)$$

$$\Rightarrow f_v(t)\underline{f_s(b,l)}P(a \mid t,l)P(x \mid a)P(d \mid a,b)$$

Summing on $s$ results in a factor with two arguments $f_s(b,l)$
In general, result of elimination may be a function of several variables

- We want to compute *P(d)*
- Need to eliminate: *x,t,l,a,b*




- Initial factors

$$P(v)P(s)P(t\mid v)P(l\mid s)P(b\mid s)P(a\mid t,l)P(x\mid a)P(d\mid a,b)$$

$$\Rightarrow f_v(t)P(s)P(l\mid s)P(b\mid s)P(a\mid t,l)P(x\mid a)P(d\mid a,b)$$

$$\Rightarrow f_v(t)f_s(b,l)P(a\mid t,l)\underline{P(x\mid a)}P(d\mid a,b)$$

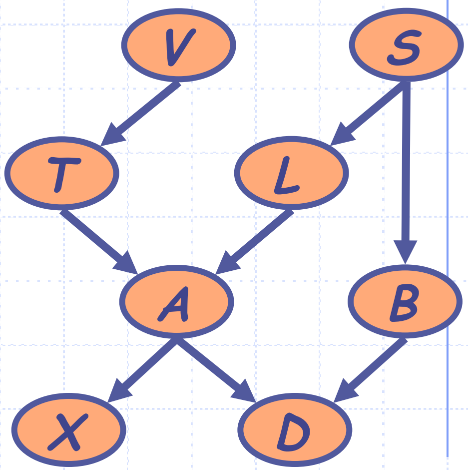
Eliminate: *x*

Compute: $$f_x(a) = \sum_x P(x\mid a)$$

$$\Rightarrow f_v(t)f_s(b,l)\underline{f_x(a)}P(a\mid t,l)P(d\mid a,b)$$

Note: $f_x(a) = 1$ for all values of *a* !!

- We want to compute *P(d)*
- Need to eliminate: *t,l,a,b*




- Initial factors

$$P(v)P(s)P(t \mid v)P(l \mid s)P(b \mid s)P(a \mid t,l)P(x \mid a)P(d \mid a,b)$$

$$\Rightarrow f_v(t)P(s)P(l \mid s)P(b \mid s)P(a \mid t,l)P(x \mid a)P(d \mid a,b)$$

$$\Rightarrow f_v(t)f_s(b,l)P(a \mid t,l)P(x \mid a)P(d \mid a,b)$$

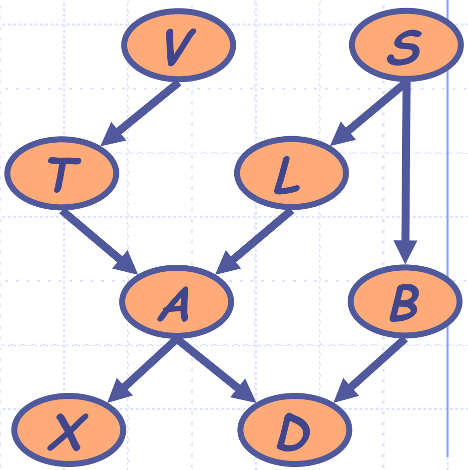
$$\Rightarrow \underline{f_v(t)}f_s(b,l)f_x(a)\underline{P(a \mid t,l)}P(d \mid a,b)$$

Eliminate: *t*

Compute: $$f_t(a,l) = \sum_t f_v(t)P(a \mid t,l)$$

$$\Rightarrow f_s(b,l)f_x(a)\underline{f_t(a,l)}P(d \mid a,b)$$

- We want to compute *P(d)*
- Need to eliminate: *l,a,b*




- Initial factors

$$P(v)P(s)P(t|v)P(l|s)P(b|s)P(a|t,l)P(x|a)P(d|a,b)$$

$$\Rightarrow f_v(t)P(s)P(l|s)P(b|s)P(a|t,l)P(x|a)P(d|a,b)$$

$$\Rightarrow f_v(t)f_s(b,l)P(a|t,l)P(x|a)P(d|a,b)$$

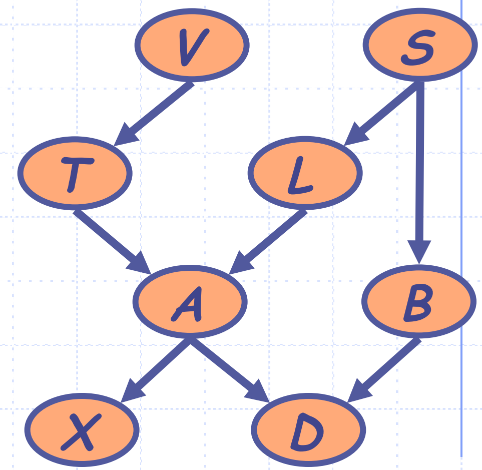
$$\Rightarrow f_v(t)f_s(b,l)f_x(a)P(a|t,l)P(d|a,b)$$

$$\Rightarrow \underline{f_s(b,l)}f_x(a)\underline{f_t(a,l)}P(d|a,b)$$

Eliminate: *l*

Compute: $$f_l(a,b) = \sum_l f_s(b,l)f_t(a,l)$$

$$\Rightarrow \underline{f_l(a,b)}f_x(a)P(d|a,b)$$




- We want to compute *P(d)*
- Need to eliminate: *b*

- Initial factors

$$P(v)P(s)P(t\,|\,v)P(l\,|\,s)P(b\,|\,s)P(a\,|\,t,l)P(x\,|\,a)P(d\,|\,a,b)$$

$$\Rightarrow f_v(t)P(s)P(l\,|\,s)P(b\,|\,s)P(a\,|\,t,l)P(x\,|\,a)P(d\,|\,a,b)$$

$$\Rightarrow f_v(t)f_s(b,l)P(a\,|\,t,l)P(x\,|\,a)P(d\,|\,a,b)$$

$$\Rightarrow f_v(t)f_s(b,l)f_x(a)P(a\,|\,t,l)P(d\,|\,a,b)$$

$$\Rightarrow f_s(b,l)f_x(a)f_t(a,l)P(d\,|\,a,b)$$

$$\Rightarrow \underline{f_l(a,b)}\,\underline{f_x(a)}\,\underline{P(d\,|\,a,b)} \Rightarrow \underline{f_a(b,d)} \Rightarrow \underline{f_b(d)}$$

Eliminate: *a,b*

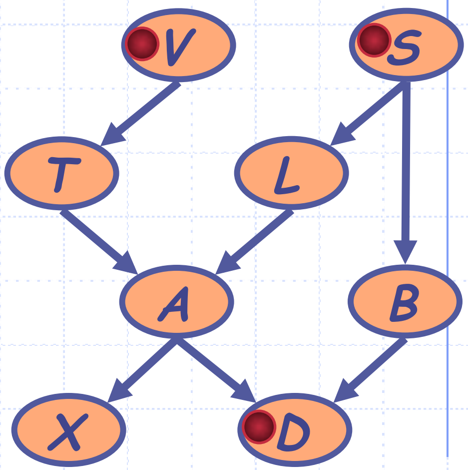
Compute:

$$f_a(b,d) = \sum_a f_l(a,b)f_x(a)p(d\,|\,a,b) \qquad f_b(d) = \sum_b f_a(b,d)$$

# Variable Elimination

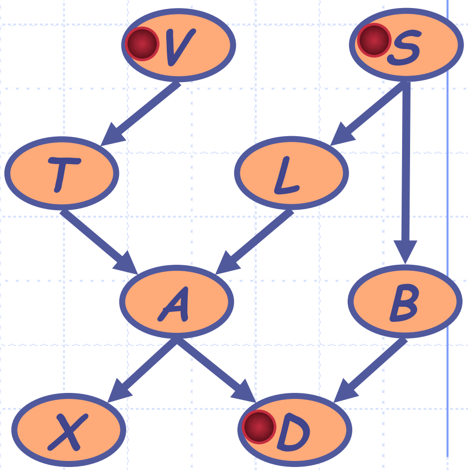
- We now understand variable elimination as a sequence of **rewriting** operations
- Actual computation is done in elimination step
- Computation depends on order of elimination

# Dealing with evidence




- How do we deal with evidence?
- Suppose get evidence $V = t, S = f, D = t$
- We want to compute $P(L | V = t, S = f, D = t)$

# Dealing with Evidence




- We start by writing the factors:

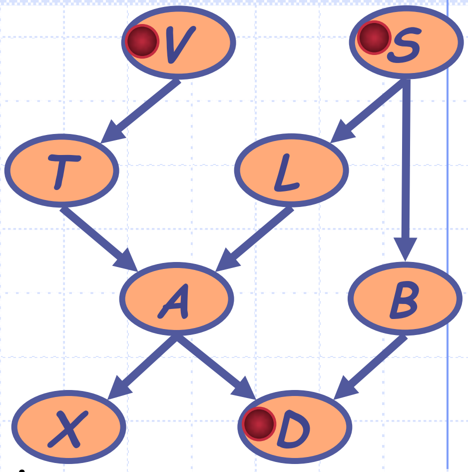
$$P(v)P(s)P(t \mid v)P(l \mid s)P(b \mid s)P(a \mid t,l)P(x \mid a)P(d \mid a,b)$$

- Since we know that $V = t$, we don't need to eliminate $V$
- Instead, we can replace the factors $P(V)$ and $P(T|V)$ with

$$f_{P(V)} = P(V = t) \qquad f_{p(T|V)}(T) = P(T \mid V = t)$$

- These "select" the appropriate parts of the original factors given the evidence
- Note that $f_{p(V)}$ is a constant, and thus does not appear in elimination of other variables

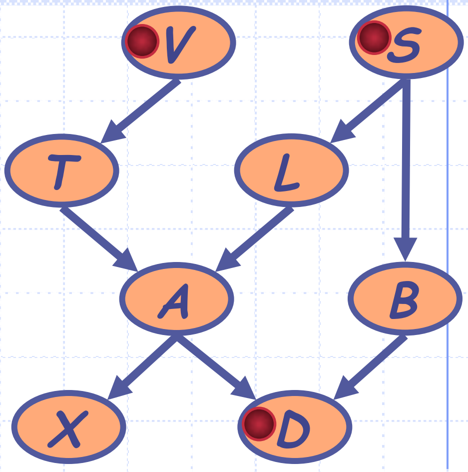
# Dealing with Evidence




- Given evidence $V = t, S = f, D = t$
- Compute $P(L, V = t, S = f, D = t)$
- Initial factors, after setting evidence:

$$f_{P(v)} f_{P(s)} f_{P(t|v)}(t) f_{P(l|s)}(l) f_{P(b|s)}(b) P(a \mid t,l) P(x \mid a) f_{P(d|a,b)}(a,b)$$

# Dealing with Evidence



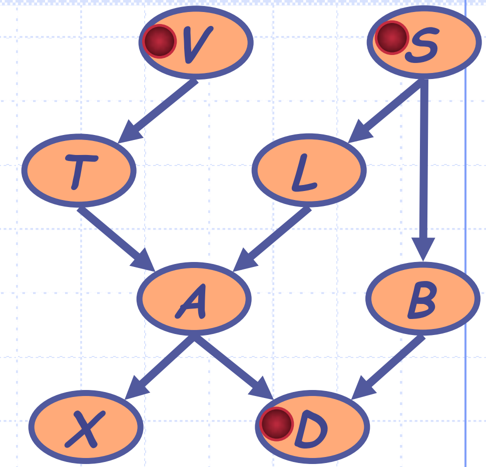

- Given evidence $V = t, S = f, D = t$
- Compute $P(L, V = t, S = f, D = t\,)$
- Initial factors, after setting evidence:

$$f_{P(v)} f_{P(s)} f_{P(t|v)}(t) f_{P(l|s)}(l) f_{P(b|s)}(b) P(a\,|\,t,l) P(x\,|\,a) f_{P(d|a,b)}(a,b)$$

- Eliminating $x$, we get

$$f_{P(v)} f_{P(s)} f_{P(t|v)}(t) f_{P(l|s)}(l) f_{P(b|s)}(b) P(a\,|\,t,l) f_x(a) f_{P(d|a,b)}(a,b)$$

# Dealing with Evidence




- Given evidence $V = t, S = f, D = t$
- Compute $P(L, V = t, S = f, D = t)$
- Initial factors, after setting evidence:

$$f_{P(v)} f_{P(s)} f_{P(t|v)}(t) f_{P(l|s)}(l) f_{P(b|s)}(b) P(a \mid t,l) \underline{P(x \mid a)} f_{P(d|a,b)}(a,b)$$

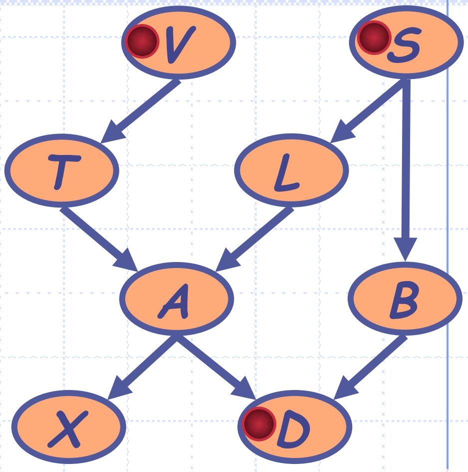
- Eliminating $x$, we get

$$f_{P(v)} f_{P(s)} \underline{f_{P(t|v)}(t)} f_{P(l|s)}(l) f_{P(b|s)}(b) \underline{P(a \mid t,l)} f_x(a) f_{P(d|a,b)}(a,b)$$

- Eliminating $t$, we get

$$f_{P(v)} f_{P(s)} f_{P(l|s)}(l) f_{P(b|s)}(b) f_t(a,l) f_x(a) f_{P(d|a,b)}(a,b)$$

# Dealing with Evidence




- Given evidence $V = t, S = f, D = t$
- Compute $P(L, V = t, S = f, D = t)$
- Initial factors, after setting evidence:

$$f_{P(v)} f_{P(s)} f_{P(t|v)}(t) f_{P(l|s)}(l) f_{P(b|s)}(b) P(a \mid t, l) \underline{P(x \mid a)} f_{P(d|a,b)}(a,b)$$

- Eliminating $x$, we get

$$f_{P(v)} f_{P(s)} \underline{f_{P(t|v)}(t)} f_{P(l|s)}(l) f_{P(b|s)}(b) \underline{P(a \mid t, l)} f_x(a) f_{P(d|a,b)}(a,b)$$

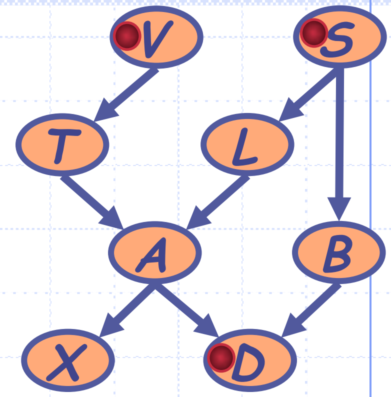
- Eliminating $t$, we get

$$f_{P(v)} f_{P(s)} f_{P(l|s)}(l) f_{P(b|s)}(b) \underline{f_t(a,l) f_x(a) f_{P(d|a,b)}(a,b)}$$

- Eliminating $a$, we get

$$f_{P(v)} f_{P(s)} f_{P(l|s)}(l) f_{P(b|s)}(b) f_a(b,l)$$

# Dealing with Evidence




- Given evidence *V = t, S = f, D = t*
- Compute *P(L, V = t, S = f, D = t )*
- Initial factors, after setting evidence:

$$f_{P(v)} f_{P(s)} f_{P(t|v)}(t) f_{P(l|s)}(l) f_{P(b|s)}(b) P(a \mid t,l) P(x \mid a) f_{P(d|a,b)}(a,b)$$

- Eliminating *x*, we get

$$f_{P(v)} f_{P(s)} f_{P(t|v)}(t) f_{P(l|s)}(l) f_{P(b|s)}(b) P(a \mid t,l) f_x(a) f_{P(d|a,b)}(a,b)$$

- Eliminating *t*, we get

$$f_{P(v)} f_{P(s)} f_{P(l|s)}(l) f_{P(b|s)}(b) \cancel{f_t(a,l) f_x(a) f_{P(d|a,b)}(a,b)}$$

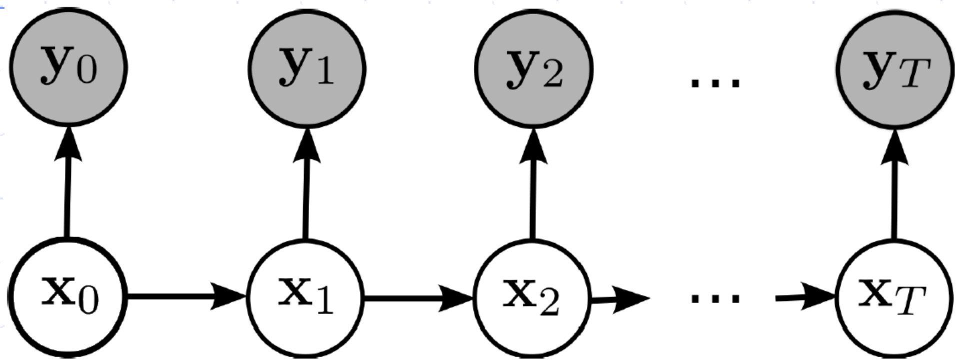
- Eliminating *a*, we get

$$f_{P(v)} f_{P(s)} f_{P(l|s)}(l) f_{P(b|s)}(b) f_a(b,l)$$

- Eliminating *b*, we get

$$f_{P(v)} f_{P(s)} f_{P(l|s)}(l) f_b(l)$$

# Hidden Markov Models



# Sequences of Observations

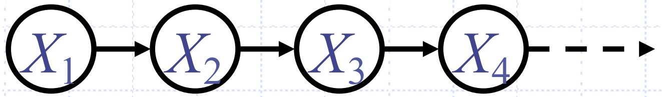
- Often, we want to reason about a sequence of observations

  - Speech recognition
  - Autonomous car localization
  - User attention modeling
  - Robot state estimation
  - Medical monitoring

- Need to introduce time into our models

# Markov Models

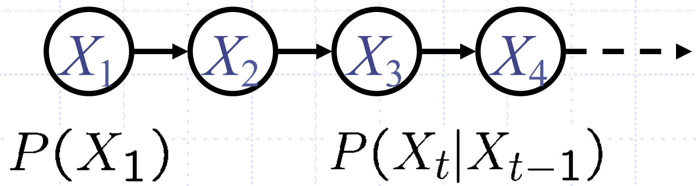
- Value of *X* at a given time is called the state



$P(X_1)$ $P(X_t|X_{t-1})$

- Parameters: called transition probabilities or dynamics, specify how the state evolves over time (also, initial state probabilities)
- Stationarity assumption: transition probabilities the same at all times
- Same as MDP transition model, but no choice of action
  - e.g., an agent acting in an MDP with a known, fixed policy

# Joint Distribution of a Markov Model




- Joint distribution:

$$P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2|X_1)P(X_3|X_2)P(X_4|X_3)$$

- More generally:

$$\begin{aligned} P(X_1, X_2, \ldots, X_T) &= P(X_1)P(X_2|X_1)P(X_3|X_2)\ldots P(X_T|X_{T-1}) \\ &= P(X_1)\prod_{t=2}^{T} P(X_t|X_{t-1}) \end{aligned}$$

# Chain Rule and Markov Models

$$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4 \dashrightarrow$$

- From the chain rule, every joint distribution over $X_1, X_2, X_3, X_4$ can be written as:

$$P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)P(X_4|X_1, X_2, X_3)$$

- Assuming that

$$X_3 \perp\!\!\!\perp X_1 \mid X_2 \quad \text{and} \quad X_4 \perp\!\!\!\perp X_1, X_2 \mid X_3$$

results in the expression posited on the previous slide:

$$P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2|X_1)P(X_3|X_2)P(X_4|X_3)$$

# Implied Conditional Independencies

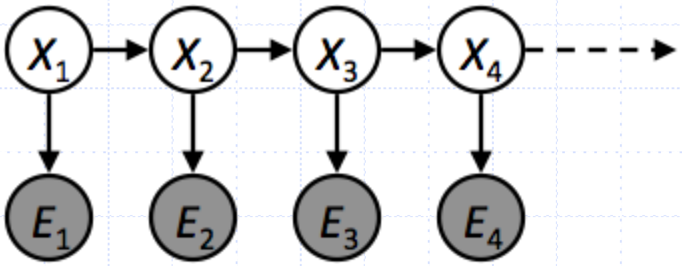
$$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4 \dashrightarrow$$

- We assumed: $X_3 \perp\!\!\!\perp X_1 \mid X_2$ and $X_4 \perp\!\!\!\perp X_1, X_2 \mid X_3$
- Do we also have $X_1 \perp\!\!\!\perp X_3, X_4 \mid X_2$ ?
  - Yes!
  - Proof:

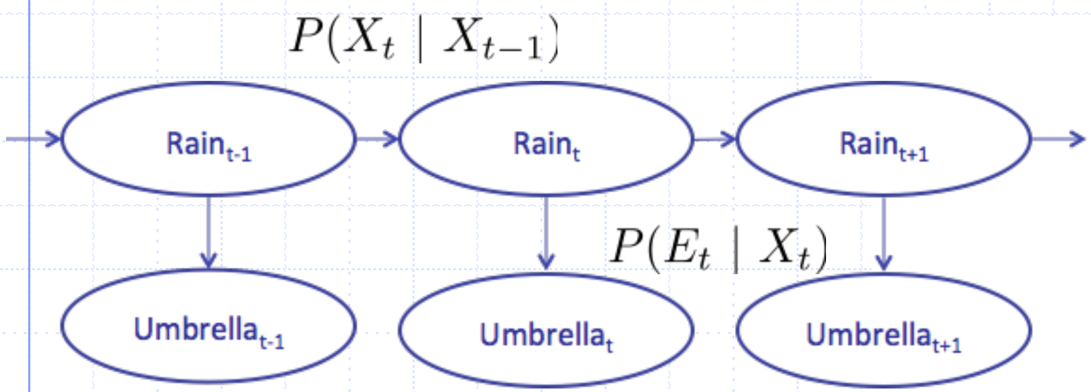
$$
\begin{aligned}
P(X_1 \mid X_2, X_3, X_4) &= \frac{P(X_1, X_2, X_3, X_4)}{P(X_2, X_3, X_4)} \\
&= \frac{P(X_1)P(X_2 \mid X_1)\cancel{P(X_3 \mid X_2)P(X_4 \mid X_3)}}{\sum_{x_1} P(x_1)P(X_2 \mid x_1)\cancel{P(X_3 \mid X_2)P(X_4 \mid X_3)}} \\
&= \frac{P(X_1, X_2)}{P(X_2)} \\
&= P(X_1 \mid X_2)
\end{aligned}
$$

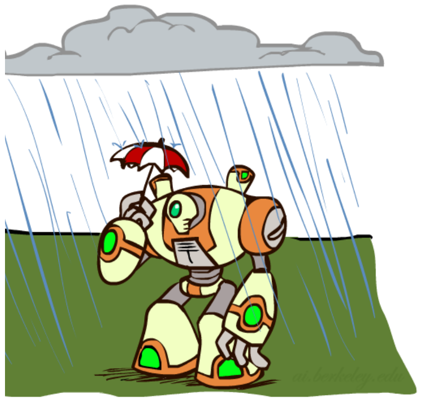
# Hidden Markov Models

- Markov chains not so useful for most agents
  - Need observations to update your beliefs
- Hidden Markov models (HMMs)
  - Underlying Markov chain over states *X*
  - You observe outputs (effects) at each time step




# Example: Weather HMM



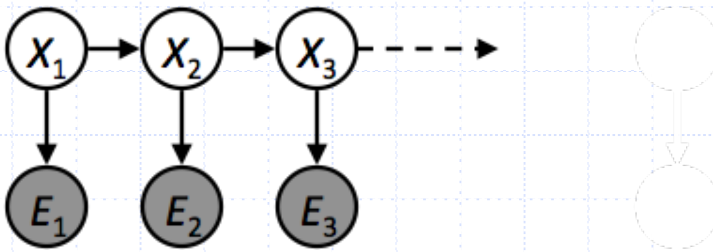



- An HMM is defined by:
  - Initial distribution: $P(X_1)$
  - Transitions: $P(X_t \mid X_{t-1})$
  - Emissions: $P(E_t \mid X_t)$

| $R_t$ | $R_{t+1}$ | $P(R_{t+1} \mid R_t)$ |
|---|---|---|
| +r | +r | 0.7 |
| +r | -r | 0.3 |
| -r | +r | 0.3 |
| -r | -r | 0.7 |

| $R_t$ | $E_t$ | $P(E_t \mid R_t)$ |
|---|---|---|
| +r | +u | 0.9 |
| +r | -u | 0.1 |
| -r | +u | 0.2 |
| -r | -u | 0.8 |

# Joint Distribution of an HMM



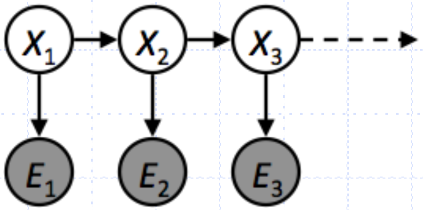

- Joint distribution:

$$P(X_1, E_1, X_2, E_2, X_3, E_3) = P(X_1)P(E_1|X_1)P(X_2|X_1)P(E_2|X_2)P(X_3|X_2)P(E_3|X_3)$$

- More generally (and compactly):

$$P(X_1, E_1, \ldots, X_T, E_T) = P(X_1)P(E_1|X_1)\prod_{t=2}^{T} P(X_t|X_{t-1})P(E_t|X_t)$$

# Chain Rule and HMMs




- From the chain rule, ***every*** joint distribution over $X_1, E_1, X_2, E_2, X_3, E_3$ can be written as:

$$P(X_1, E_1, X_2, E_2, X_3, E_3) = P(X_1)P(E_1|X_1)P(X_2|X_1, E_1)P(E_2|X_1, E_1, X_2) \\ P(X_3|X_1, E_1, X_2, E_2)P(E_3|X_1, E_1, X_2, E_2, X_3)$$

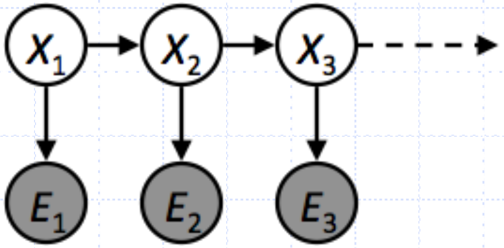
- ***Assuming*** that

$$X_2 \perp\!\!\!\perp E_1 \mid X_1, \quad E_2 \perp\!\!\!\perp X_1, E_1 \mid X_2, \quad X_3 \perp\!\!\!\perp X_1, E_1, E_2 \mid X_2, \quad E_3 \perp\!\!\!\perp X_1, E_1, X_2, E_2 \mid X_3$$

gives us the expression posited on the previous slide:

$$P(X_1, E_1, X_2, E_2, X_3, E_3) = P(X_1)P(E_1|X_1)P(X_2|X_1)P(E_2|X_2)P(X_3|X_2)P(E_3|X_3)$$

# Implied Conditional Independencies




- Many implied conditional independencies, e.g.,

$$E_1 \perp\!\!\!\perp X_2, E_2, X_3, E_3 \mid X_1$$

- To prove them
  - *Approach 1*: follow a similar algebraic approach to what we did with Markov chains *(try this!)*
  - *Approach 2*: directly from the graph structure
    - Intuition: If path between *U* and *V* goes through *W*, then

$$U \perp\!\!\!\perp V \mid W$$

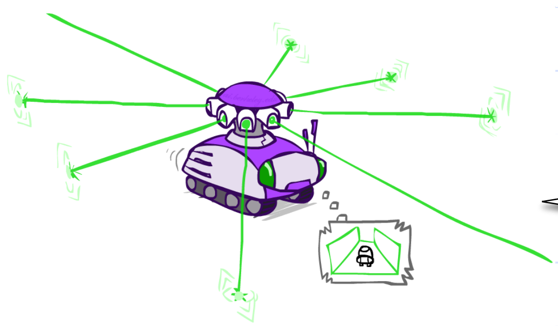
# Filtering

- Filtering is the task of tracking a ***belief distribution***

  $B_t(X) = P_t(X_t \mid e_1, \ldots, e_t)$ over time

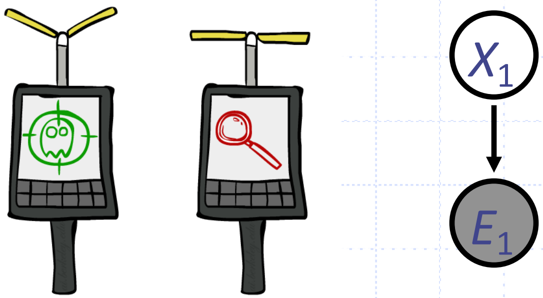
- We start with $B_1(X)$ in an initial setting, e.g., uniform

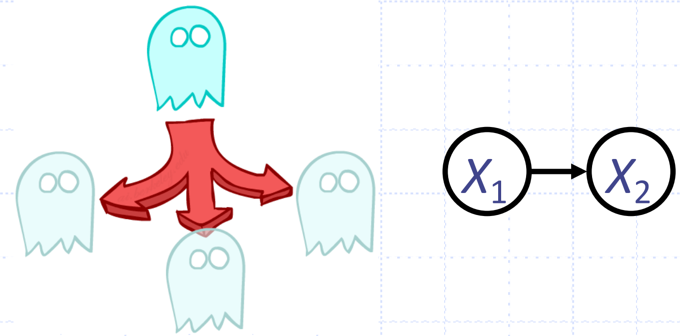
- As time passes, or we get observations, we update $B(X)$



Where am I?

# Inference: Base Cases




$$P(X_1|e_1)$$

$$P(x_1|e_1) = P(x_1, e_1)/P(e_1)$$

$$= \frac{P(e_1|x_1)P(x_1)}{P(e_1)}$$

$$P(X_2)$$

$$P(x_2) = \sum_{x_1} P(x_1, x_2)$$

$$= \sum_{x_1} P(x_1)P(x_2|x_1)$$

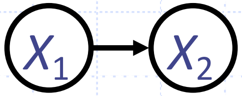
# Passage of Time

- Assume we have current belief *P(X | evidence to date)*

$$B(X_t) = P(X_t|e_{1:t})$$

- Then, after one time step passes:

$$P(X_{t+1}|e_{1:t}) = \sum_{x_t} P(X_{t+1}, x_t|e_{1:t})$$
$$= \sum_{x_t} P(X_{t+1}|x_t, e_{1:t})P(x_t|e_{1:t})$$
$$= \sum_{x_t} P(X_{t+1}|x_t)P(x_t|e_{1:t})$$




**Or compactly:**

$$B'(X_{t+1}) = \sum_{x_t} P(X'|x_t)B(x_t)$$

- Basic idea: beliefs get "pushed" through the transitions
  - With the "B" notation, we have to be careful about what time step *t* the belief is about, and what evidence it includes

# Example: Passage of Time

- As time passes, uncertainty “accumulates”

| | | | | | |
|---|---|---|---|---|---|
| <0.01 | <0.01 | <0.01 | <0.01 | <0.01 | <0.01 |
| <0.01 | <0.01 | <0.01 | <0.01 | <0.01 | <0.01 |
| <0.01 | <0.01 | 1.00 | <0.01 | <0.01 | <0.01 |
| <0.01 | <0.01 | <0.01 | <0.01 | <0.01 | <0.01 |

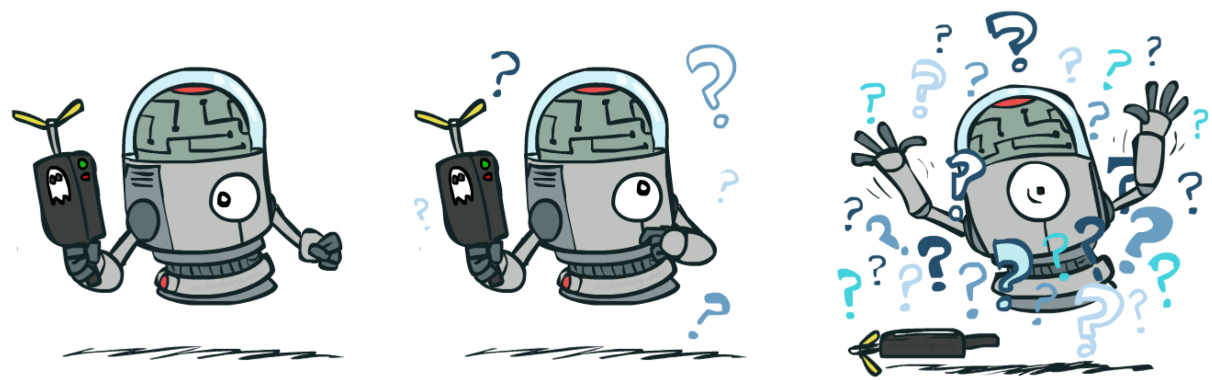
T = 1

| | | | | | |
|---|---|---|---|---|---|
| <0.01 | <0.01 | <0.01 | <0.01 | <0.01 | <0.01 |
| <0.01 | <0.01 | 0.06 | <0.01 | <0.01 | <0.01 |
| <0.01 | 0.76 | 0.06 | 0.06 | <0.01 | <0.01 |
| <0.01 | <0.01 | 0.06 | <0.01 | <0.01 | <0.01 |

T = 2

| | | | | | |
|---|---|---|---|---|---|
| 0.05 | 0.01 | 0.05 | <0.01 | <0.01 | <0.01 |
| 0.02 | 0.14 | 0.11 | 0.35 | <0.01 | <0.01 |
| 0.07 | 0.03 | 0.05 | <0.01 | 0.03 | <0.01 |
| 0.03 | 0.03 | <0.01 | <0.01 | <0.01 | <0.01 |

T = 5



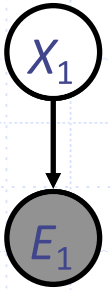
(Transition model: ghosts usually go clockwise)

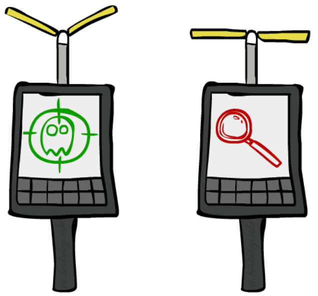
# Observation

- Assume we have current belief *P(X | previous evidence)*:

$$B'(X_{t+1}) = P(X_{t+1}|e_{1:t})$$

- Then, after evidence comes in:





$$\begin{aligned} P(X_{t+1}|e_{1:t+1}) &= P(X_{t+1}, e_{t+1}|e_{1:t})/P(e_{t+1}|e_{1:t}) \\ &\propto_{X_{t+1}} P(X_{t+1}, e_{t+1}|e_{1:t}) \\ &= P(e_{t+1}|e_{1:t}, X_{t+1})P(X_{t+1}|e_{1:t}) \\ &= P(e_{t+1}|X_{t+1})P(X_{t+1}|e_{1:t}) \end{aligned}$$

- Or, compactly:

$$B(X_{t+1}) \propto_{X_{t+1}} P(e_{t+1}|X_{t+1})B'(X_{t+1})$$

- Basic idea: beliefs "reweighted" by likelihood of evidence
- Unlike passage of time, we have to renormalize

# Example: Observation

- As we get sensor observations, beliefs get reweighted
- Generally our uncertainty "decreases"

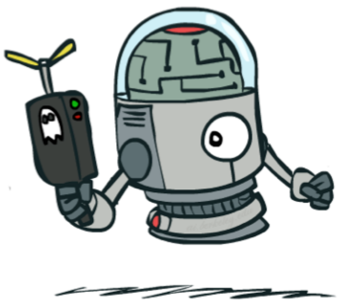
| | | | | | |
|---|---|---|---|---|---|
| 0.05 | 0.01 | 0.05 | <0.01 | <0.01 | <0.01 |
| 0.02 | 0.14 | 0.11 | 0.35 | <0.01 | <0.01 |
| 0.07 | 0.03 | 0.05 | <0.01 | 0.03 | <0.01 |
| 0.03 | 0.03 | <0.01 | <0.01 | <0.01 | <0.01 |

Before observation

| | | | | | |
|---|---|---|---|---|---|
| <0.01 | <0.01 | <0.01 | <0.01 | 0.02 | <0.01 |
| <0.01 | <0.01 | <0.01 | 0.83 | 0.02 | <0.01 |
| <0.01 | <0.01 | 0.11 | <0.01 | <0.01 | <0.01 |
| <0.01 | <0.01 | <0.01 | <0.01 | <0.01 | <0.01 |

After observation

$$B(X) \propto P(e|X)B'(X)$$



# Example: Weather HMM

$B(+r) = 0.5$
$B(-r) = 0.5$

$B'(+r) = 0.5$
$B'(-r) = 0.5$

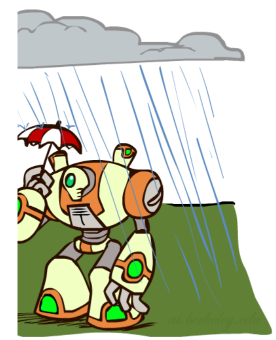
$U=1$

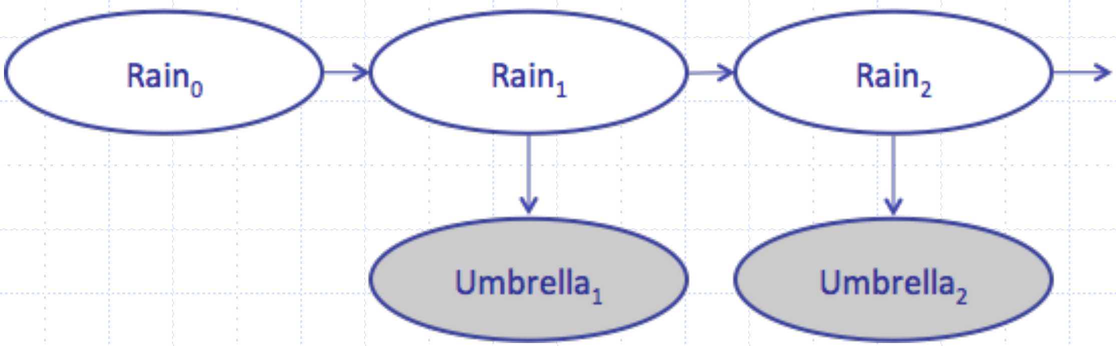
$B(+r) = 0.818$
$B(-r) = 0.182$

$B'(+r) = 0.627$
$B'(-r) = 0.373$

$U=1$

$B(+r) = 0.883$
$B(-r) = 0.117$






| $R_t$ | $R_{t+1}$ | $P(R_{t+1}\|R_t)$ |
|---|---|---|
| +r | +r | 0.7 |
| +r | -r | 0.3 |
| -r | +r | 0.3 |
| -r | -r | 0.7 |

| $R_t$ | $U_t$ | $P(U_t\|R_t)$ |
|---|---|---|
| +r | +u | 0.9 |
| +r | -u | 0.1 |
| -r | +u | 0.2 |
| -r | -u | 0.8 |

# The Forward Algorithm

- We are given evidence at each time, 1…t, and want to know

$$B_t(X) = P(X_t|e_{1:t})$$

- We can derive the following update rule for a DP algorithm

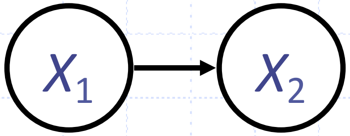
$$
\begin{aligned}
P(x_t|e_{1:t}) &\propto_X P(x_t, e_{1:t}) \\
&= \sum_{x_{t-1}} P(x_{t-1}, x_t, e_{1:t}) \\
&= \sum_{x_{t-1}} P(x_{t-1}, e_{1:t-1}) P(x_t|x_{t-1}) P(e_t|x_t) \\
&= P(e_t|x_t) \sum_{x_{t-1}} P(x_t|x_{t-1}) P(x_{t-1}, e_{1:t-1})
\end{aligned}
$$

We can normalize as we go if we want to have *P(x|e)* at each time step, or just once at the end…

# Online Belief Updates

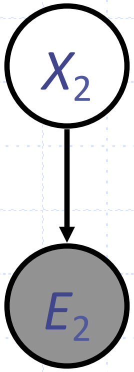
- Every time step, we start with current *P(X | evidence)*
- We update for time:

$$P(x_t|e_{1:t-1}) = \sum_{x_{t-1}} P(x_{t-1}|e_{1:t-1}) \cdot P(x_t|x_{t-1})$$




- We update for evidence:

$$P(x_t|e_{1:t}) \propto_X P(x_t|e_{1:t-1}) \cdot P(e_t|x_t)$$




- Space and time complexity?

# Summary

- Several lectures on ***probabilistic reasoning***
- Probabilistic models encode joint distributions over sets of random variables
- Important example of a probabilistic model: Hidden Markov Models (HMMs)
  - Useful for reasoning about sequences over time and space
  - Robot state estimation, audio processing, etc.