

CSE 473: Artificial Intelligence

Bayesian Networks: Inference

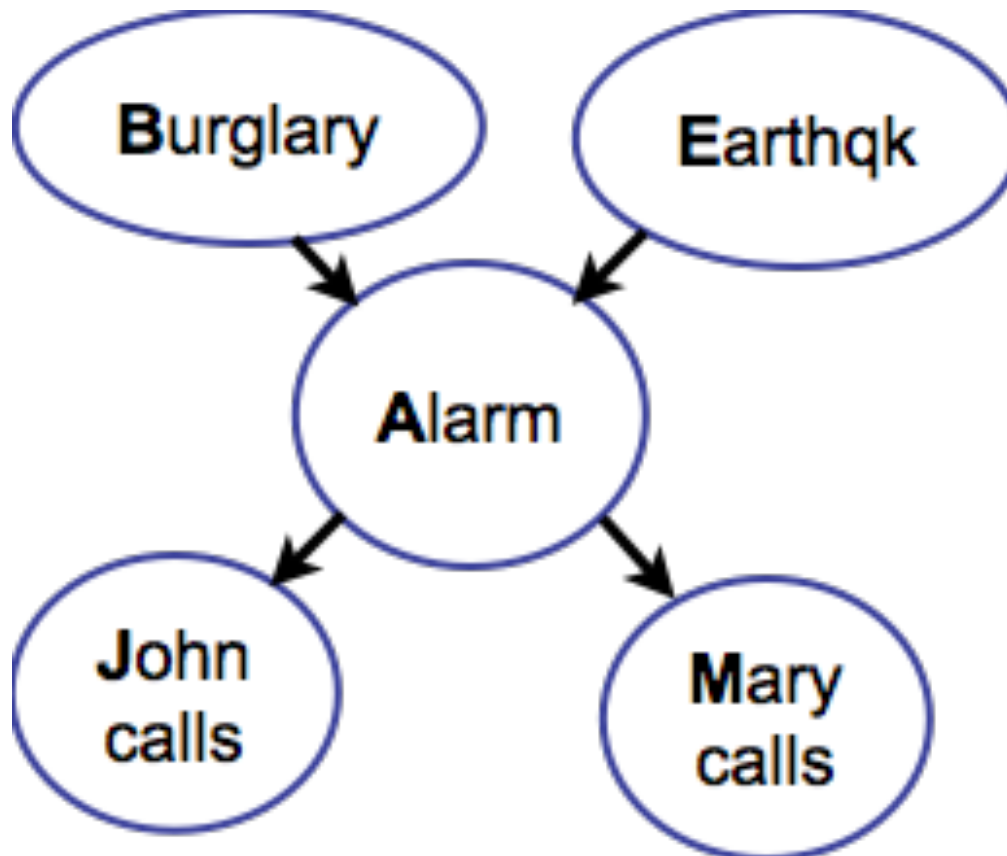
Hanna Hajishirzi

Many slides over the course adapted from either Luke Zettlemoyer, Pieter Abbeel, Dan Klein, Stuart Russell or Andrew Moore

Outline

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

Bayes Net Representation



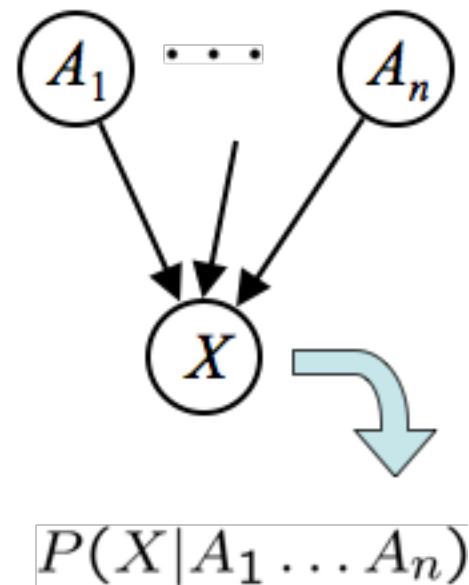
Bayes' Net Semantics

- Let's formalize the semantics of a Bayes' net
- A set of nodes, one per variable X
- A directed, acyclic graph
- A conditional distribution for each node
 - A collection of distributions over X , one for each combination of parents' values

$$P(X|a_1 \dots a_n)$$

- CPT: conditional probability table

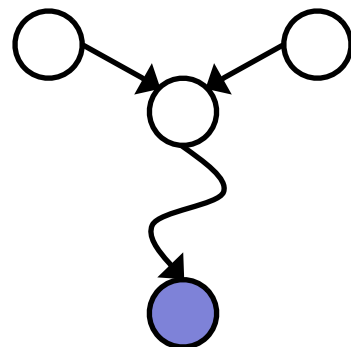
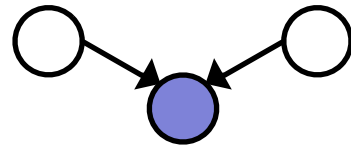
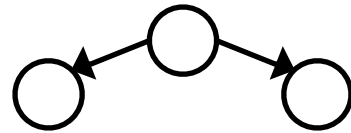
A Bayes net = Topology (graph) + Local Conditional Probabilities



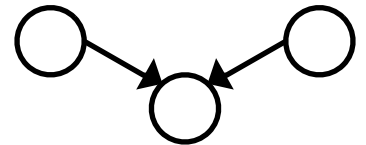
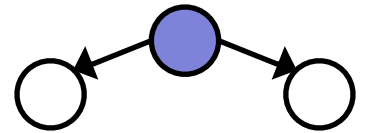
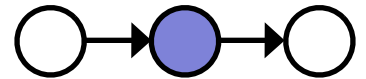
Reachability (D-Separation)

- Question: Are X and Y conditionally independent given evidence vars {Z}?
 - Yes, if X and Y “separated” by Z
 - Look for active paths from X to Y
 - No active paths = independence!
- A path is active if each triple is active:
 - Causal chain $A \rightarrow B \rightarrow C$ where B is unobserved (either direction)
 - Common cause $A \leftarrow B \rightarrow C$ where B is unobserved
 - Common effect (aka v-structure) $A \rightarrow B \leftarrow C$ where B or one of its descendants is observed
- All it takes to block a path is a single inactive segment

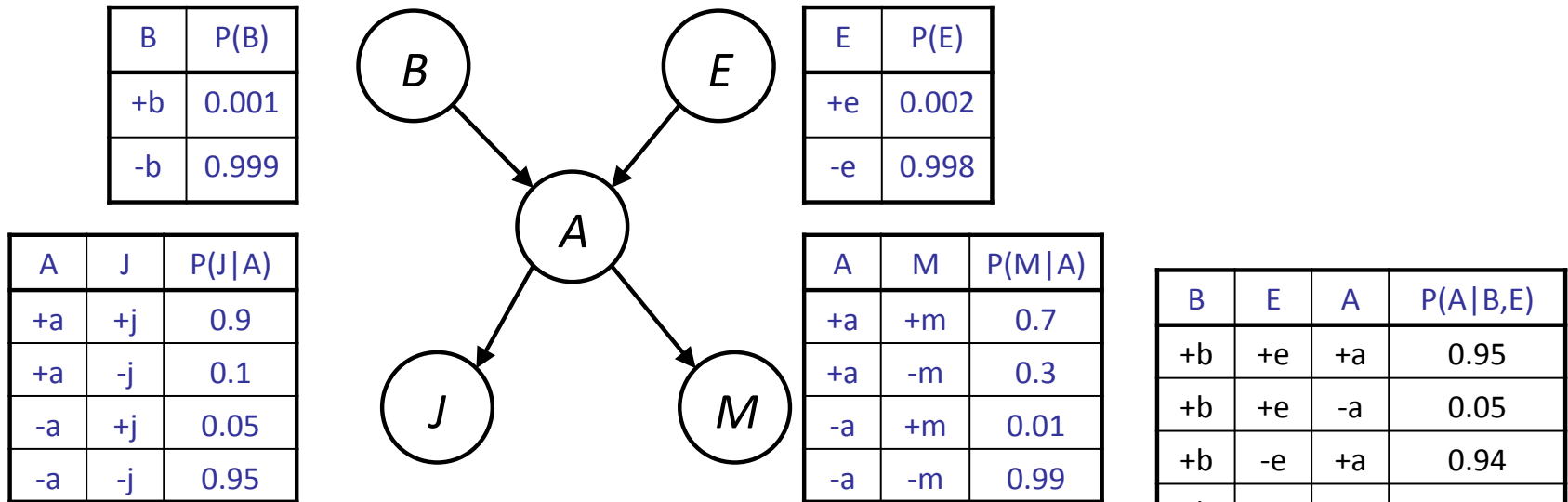
Active Triples
(dependent)



Inactive Triples
(Independent)



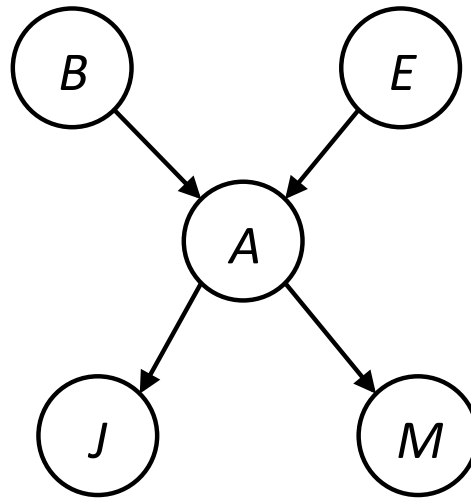
Bayes Net Joint Distribution



$$P(+b, -e, +a, -j, +m) = P(+b)P(-e)P(+a | +b, -e)P(-j | +a)P(+m | +a) =$$

Bayes Net Joint Distribution

B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998

A	J	P(J A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

A	M	P(M A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

B	E	A	P(A B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

$$\begin{aligned}
 P(+b, -e, +a, -j, +m) &= \\
 P(+b)P(-e)P(+a|+b, -e)P(-j|+a)P(+m|+a) &= \\
 0.001 \times 0.998 \times 0.94 \times 0.1 \times 0.7 &
 \end{aligned}$$

Probabilistic Inference

- Probabilistic inference: compute a desired probability from other known probabilities (e.g. conditional from joint)
- We generally compute conditional probabilities
 - $P(\text{on time} \mid \text{no reported accidents}) = 0.90$
 - These represent the agent's beliefs given the evidence
- Probabilities change with new evidence:
 - $P(\text{on time} \mid \text{no accidents, 5 a.m.}) = 0.95$
 - $P(\text{on time} \mid \text{no accidents, 5 a.m., raining}) = 0.80$
 - Observing new evidence causes beliefs to be updated

Inference

- Inference: calculating some useful quantity from a joint probability distribution

- Examples:

- Posterior probability

$$P(Q|E_1 = e_1, \dots, E_k = e_k)$$

- Most likely explanation:

$$\operatorname{argmax}_q P(Q = q|E_1 = e_1 \dots)$$

Inference by Enumeration

- General case:

- Evidence variables: $E_1 \dots E_k = e_1 \dots e_k$
 - Query* variable: Q
 - Hidden variables: $H_1 \dots H_r$
- } X_1, X_2, \dots, X_n
All variables

- We want: $P(Q|e_1 \dots e_k)$

- First, select the entries consistent with the evidence

- Second, sum out H to get joint of Query and evidence:

$$P(Q, e_1 \dots e_k) = \sum_{h_1 \dots h_r} \underbrace{P(Q, h_1 \dots h_r, e_1 \dots e_k)}_{X_1, X_2, \dots, X_n}$$

- Finally, normalize the remaining entries to conditionalize

- Obvious problems:

- Worst-case time complexity $O(d^n)$
- Space complexity $O(d^n)$ to store the joint distribution

Inference in BN by Enumeration

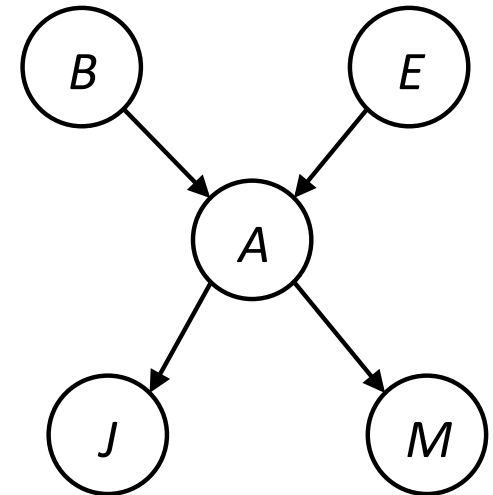
- Given unlimited time, inference in BNs is easy
- Reminder of inference by enumeration by example:

$$P(B \mid +j, +m) \propto_B P(B, +j, +m)$$

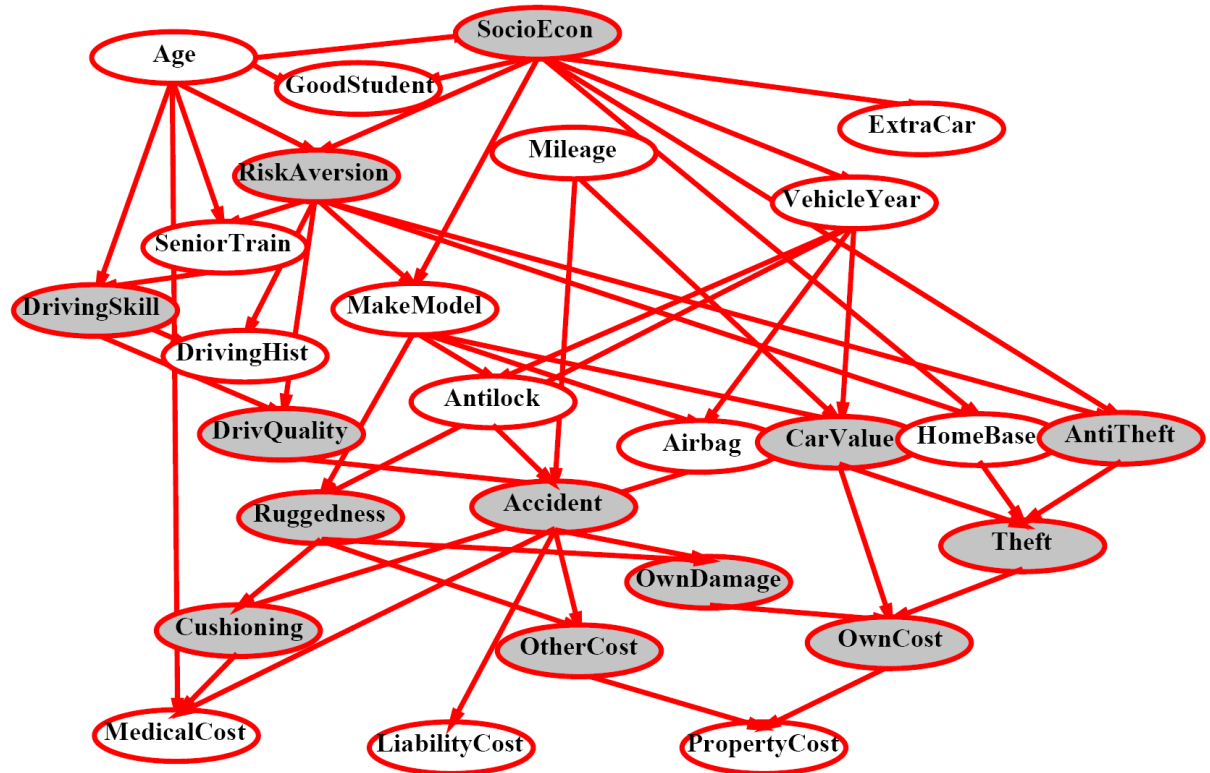
$$= \sum_{e,a} P(B, e, a, +j, +m)$$

$$= \sum_{e,a} P(B)P(e)P(a|B, e)P(+j|a)P(+m|a)$$

$$= P(B)P(+e)P(+a|B, +e)P(+j|+a)P(+m|+a) + P(B)P(+e)P(-a|B, +e)P(+j|-a)P(+m|-a) \\ + P(B)P(-e)P(+a|B, -e)P(+j|+a)P(+m|+a) + P(B)P(-e)P(-a|B, -e)P(+j|-a)P(+m|-a)$$



Inference by Enumeration



$$P(\textit{Antilock} | \textit{observed variables}) = ?$$

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

Review

- Joint distribution: $P(X, Y)$
 - Entries $P(x, y)$ for all x, y
 - Sums to 1

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$P(\text{cold}, W)$

- Selected joint: $P(x, Y)$
 - A slice of the joint distribution
 - Entries $P(x, y)$ for fixed x , all y
 - Sums to $P(x)$

T	W	P
cold	sun	0.2
cold	rain	0.3

Review

- Family of conditionals:

$P(X | Y)$

- Multiple conditionals
- Entries $P(x | y)$ for all x, y
- Sums to $|Y|$

$P(W|T)$

T	W	P	
hot	sun	0.8	} $P(W hot)$
hot	rain	0.2	
cold	sun	0.4	} $P(W cold)$
cold	rain	0.6	

- Single conditional: $P(Y | x)$
 - Entries $P(y | x)$ for fixed x , all y
 - Sums to 1

T	W	P
cold	sun	0.4
cold	rain	0.6

Review

$$P(\text{rain}|T)$$

- Specified family: $P(y | X)$
 - Entries $P(y | x)$ for fixed y , but for all x
 - Sums to ... who knows!

T	W	P
hot	rain	0.2
cold	rain	0.6

$$\left. \begin{array}{l} \\ \end{array} \right\} P(\text{rain}|\text{hot})$$
$$\left. \begin{array}{l} \\ \end{array} \right\} P(\text{rain}|\text{cold})$$

- In general, when we write $P(Y_1 \dots Y_N | X_1 \dots X_M)$
 - It is a “factor,” a multi-dimensional array
 - Its values are all $P(y_1 \dots y_N | x_1 \dots x_M)$
 - Any assigned X or Y is a dimension missing (selected) from the array

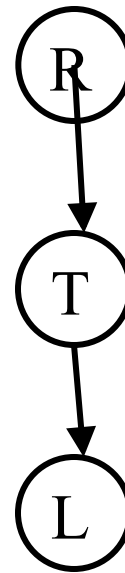
Inference

- Inference is expensive with enumeration
- Variable elimination:
 - Interleave joining and marginalization: Store initial results and then join with the rest

Example: Traffic Domain

- Random Variables

- R: Raining
- T: Traffic
- L: Late for class!



$P(R)$

+r	0.1
-r	0.9

$P(T|R)$

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

$P(L|T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

$$P(l) = \sum_t \sum_r P(l|t)P(t|r)P(r)$$

Variable Elimination Outline

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

+r	0.1
-r	0.9

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

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

+r	0.1
-r	0.9

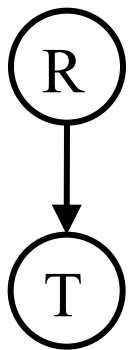
+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

+t	+l	0.3
-t	+l	0.1

- 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



$$P(R) \times$$

+r	0.1
-r	0.9

$$P(T|R)$$

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9



$$P(R, T)$$

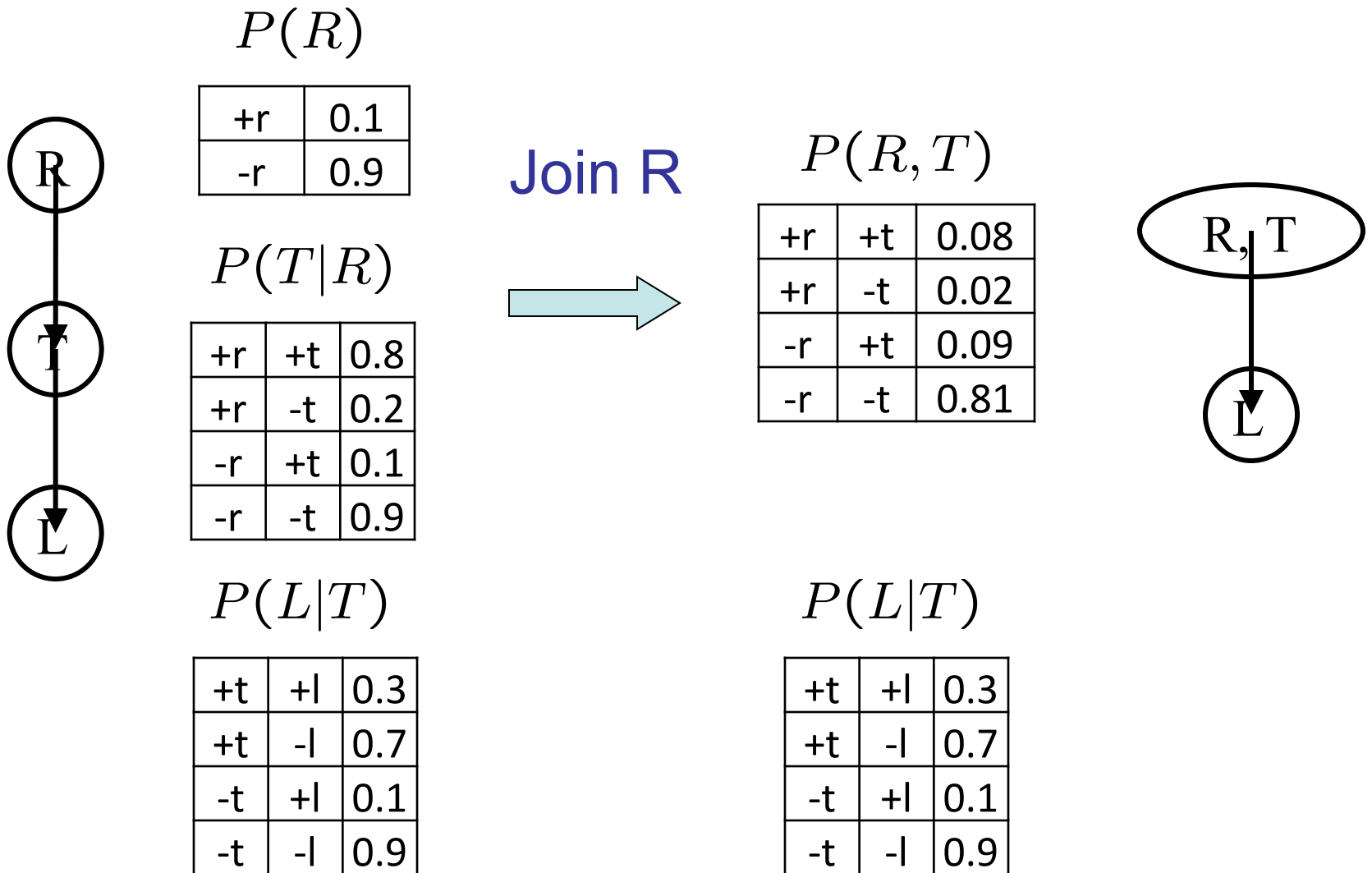


+r	+t	0.08
+r	-t	0.02
-r	+t	0.09
-r	-t	0.81

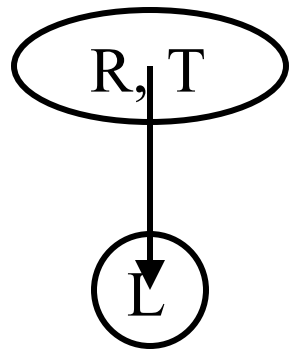
- Computation for each entry: pointwise products

$$\forall r, t : P(r, t) = P(r) \cdot P(t|r)$$

Example: Multiple Joins



Example: Multiple Joins



$P(R, T)$

+r	+t	0.08
+r	-t	0.02
-r	+t	0.09
-r	-t	0.81

$P(L|T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

Join T




R, T, L

$P(R, T, L)$

+r	+t	+l	0.024
+r	+t	-l	0.056
+r	-t	+l	0.002
+r	-t	-l	0.018
-r	+t	+l	0.027
-r	+t	-l	0.063
-r	-t	+l	0.081
-r	-t	-l	0.729

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:

$P(R, T)$			sum R	$P(T)$	
+r	+t	0.08		+t	0.17
+r	-t	0.02		-t	0.83
-r	+t	0.09			
-r	-t	0.81			

Multiple Elimination

R, T, L

$P(R, T, L)$

+r	+t	+l	0.024
+r	+t	-l	0.056
+r	-t	+l	0.002
+r	-t	-l	0.018
-r	+t	+l	0.027
-r	+t	-l	0.063
-r	-t	+l	0.081
-r	-t	-l	0.729

Sum
out R



T, L

$P(T, L)$

+t	+l	0.051
+t	-l	0.119
-t	+l	0.083
-t	-l	0.747

Sum
out T



L

$P(L)$

+l	0.134
-l	0.886

P(L) : Marginalizing Early!

$P(R)$

+r	0.1
-r	0.9

Join R

$P(T|R)$

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9



$P(R, T)$

+r	+t	0.08
+r	-t	0.02
-r	+t	0.09
-r	-t	0.81

Sum out R

$P(T)$

+t	0.17
-t	0.83

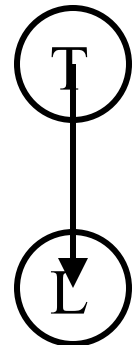
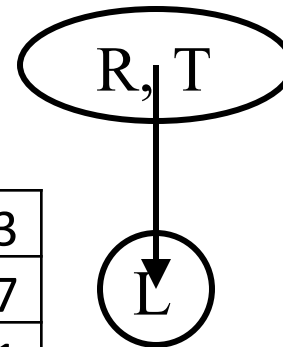
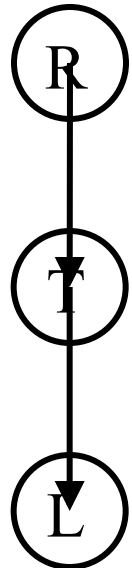
R, T

$P(L|T)$

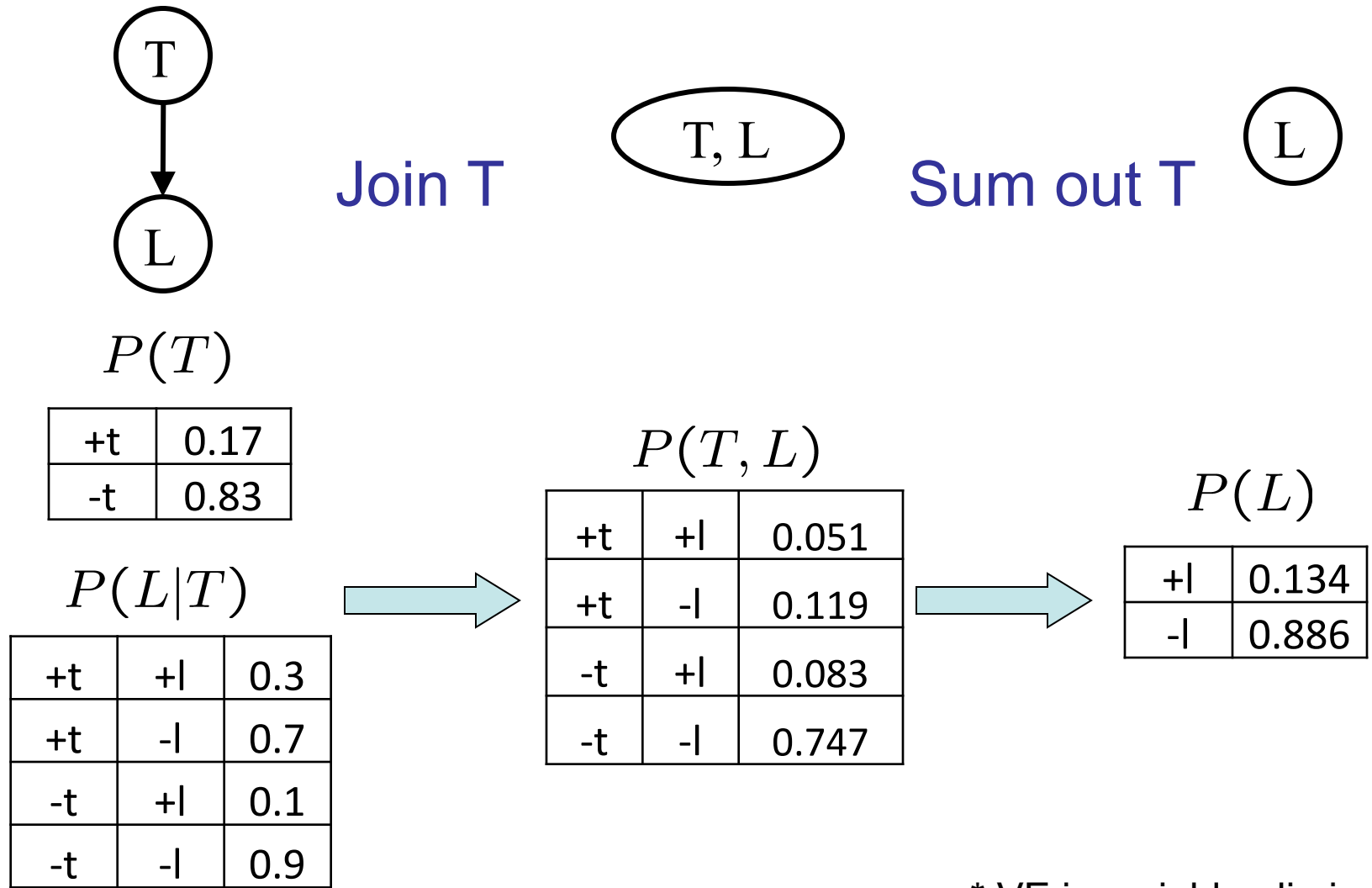
+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

$P(L|T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

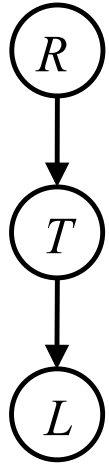


Marginalizing Early (aka VE*)



* VE is variable elimination

Traffic Domain



$$P(L) = ?$$

■ Inference by Enumeration

$$= \sum_t \sum_r P(L|t) \underbrace{P(r)P(t|r)}_{\text{Join on } r}$$

$$\underbrace{\hspace{10em}}_{\text{Join on } t}$$

$$\underbrace{\hspace{10em}}_{\text{Eliminate } r}$$

$$\underbrace{\hspace{10em}}_{\text{Eliminate } t}$$

■ Variable Elimination

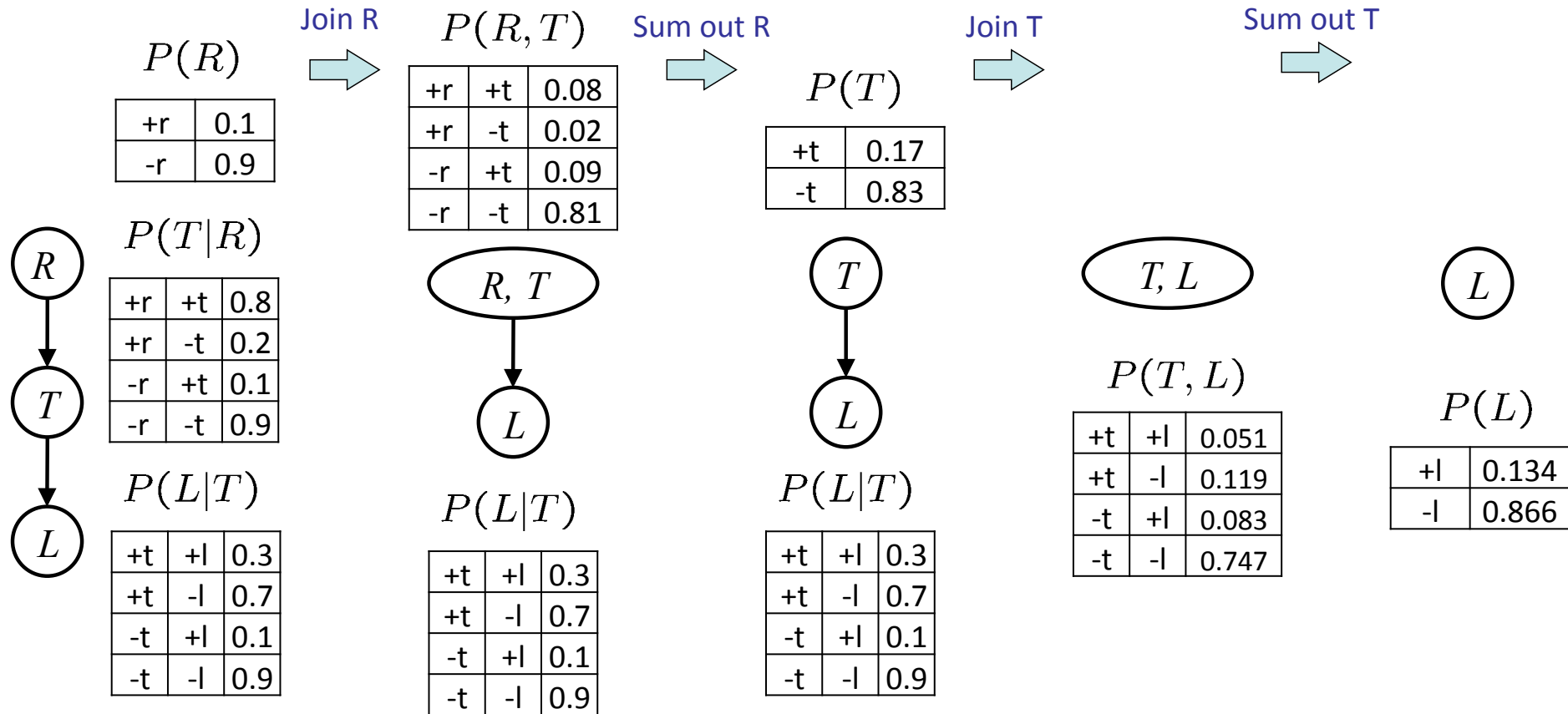
$$= \sum_t P(L|t) \underbrace{\sum_r P(r)P(t|r)}_{\text{Join on } r}$$

$$\underbrace{\hspace{10em}}_{\text{Eliminate } r}$$

$$\underbrace{\hspace{10em}}_{\text{Join on } t}$$

$$\underbrace{\hspace{10em}}_{\text{Eliminate } t}$$

Marginalizing Early



Evidence

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

$$P(R)$$

+r	0.1
-r	0.9

$$P(T|R)$$

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

$$P(L|T)$$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

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

$$P(+r)$$

+r	0.1
----	-----

$$P(T|+r)$$

+r	+t	0.8
+r	-t	0.2

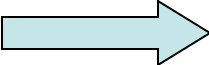
$$P(L|T)$$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

- 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:

$P(+r, L)$			Normalize	$P(L \mid +r)$	
+r	+l	0.026		+l	0.26
+r	-l	0.074		-l	0.74

- To get our answer, just normalize this!
- That's it!

General Variable Elimination

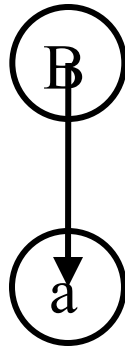
- Query: $P(Q|E_1 = e_1, \dots, E_k = e_k)$
- Start with initial factors:
 - Local CPTs (but instantiated by evidence)
- While there are still hidden variables (not Q or evidence):
 - Pick a hidden variable H
 - Join all factors mentioning H
 - Eliminate (sum out) H
- Join all remaining factors and normalize

Variable Elimination Bayes Rule

Start / Select

$P(B)$

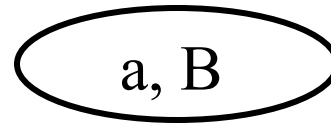
B	P
+b	0.1
-b	0.9



$P(A|B) \rightarrow P(a|B)$

B	A	P
+b	+a	0.8
b	-a	0.2
-b	+a	0.1
-b	-a	0.9

Join on B



$P(a, B)$

A	B	P
+a	+b	0.08
+a	-b	0.09

Normalize

$P(B|a)$

A	B	P
+a	+b	8/17
+a	-b	9/17

Example

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

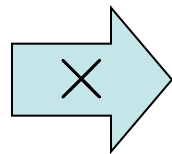
$P(B)$	$P(E)$	$P(A B, E)$	$P(j A)$	$P(m A)$
--------	--------	-------------	----------	----------

Choose A

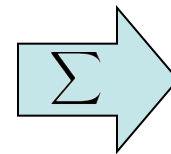
$P(A|B, E)$

$P(j|A)$

$P(m|A)$



$P(j, m, A|B, E)$



$P(j, m|B, E)$

$P(B)$	$P(E)$	$P(j, m B, E)$
--------	--------	----------------

Example

$P(B)$	$P(E)$	$P(j, m B, E)$
--------	--------	----------------

Choose E

$$\begin{array}{l} P(E) \\ P(j, m|B, E) \end{array} \xrightarrow{\times} P(j, m, E|B) \xrightarrow{\Sigma} P(j, m|B)$$

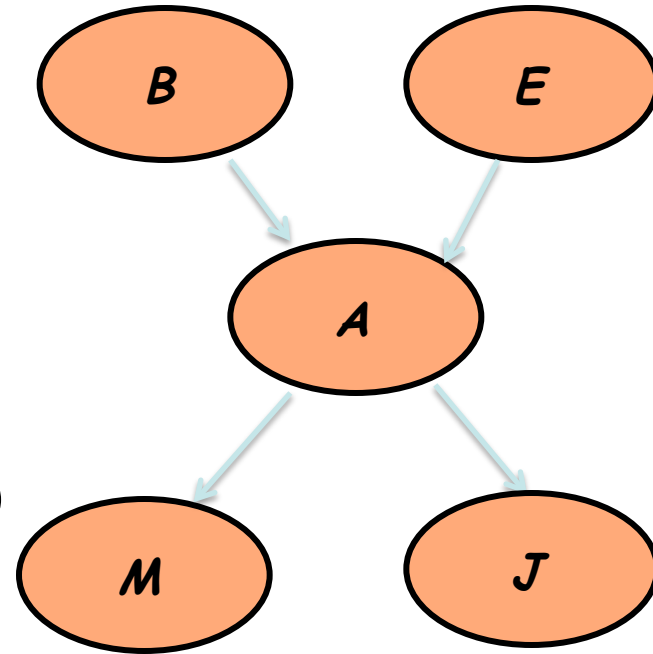
$P(B)$	$P(j, m B)$
--------	-------------

Finish with B

$$\begin{array}{l} P(B) \\ P(j, m|B) \end{array} \xrightarrow{\times} P(j, m, B) \xrightarrow{\text{Normalize}} P(B|j, m)$$

Variable Elimination

$$\begin{aligned} P(B, j, m) &= \sum_{A, E} P(b, j, m, A, E) = \\ &\sum_{A, E} P(B)P(E)P(A|B, E)P(m|A)P(j|A) \\ &\sum_E P(B)P(E) \sum_A \underline{P(A|B, E)P(m|A)P(j|A)} \\ &= \sum_E P(B)P(E) \sum_A \underline{P(m, j, A|B, E)} \\ &= \sum_E \underline{P(B)P(E)P(m, j|B, E)} = P(B) \sum_E \underline{P(m, j, E|B)} \\ &= P(B)P(m, j|B) \end{aligned}$$



Another Example

Query: $P(X_3|Y_1 = y_1, Y_2 = y_2, Y_3 = y_3)$

Start by inserting evidence, which gives the following initial factors:

$$p(Z)p(X_1|Z)p(X_2|Z)p(X_3|Z)p(y_1|X_1)p(y_2|X_2)p(y_3|X_3)$$

Eliminate X_1 , this introduces the factor $f_1(Z, y_1) = \sum_{x_1} p(x_1|Z)p(y_1|x_1)$, and we are left with:

$$p(Z)f_1(Z, y_1)p(X_2|Z)p(X_3|Z)p(y_2|X_2)p(y_3|X_3)$$

Eliminate X_2 , this introduces the factor $f_2(Z, y_2) = \sum_{x_2} p(x_2|Z)p(y_2|x_2)$, and we are left with:

$$p(Z)f_1(Z, y_1)f_2(Z, y_2)p(X_3|Z)p(y_3|X_3)$$

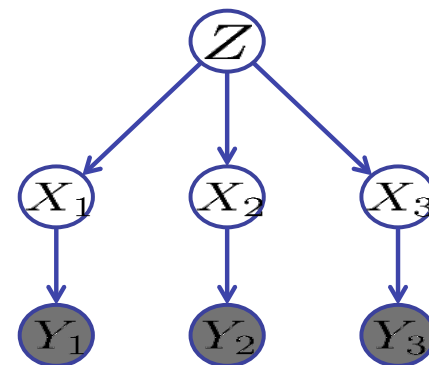
Eliminate Z , this introduces the factor $f_3(y_1, y_2, X_3) = \sum_z p(z)f_1(z, y_1)f_2(z, y_2)p(X_3|z)$, and we are left:

$$p(y_3|X_3), f_3(y_1, y_2, X_3)$$

No hidden variables left. Join the remaining factors to get:

$$f_4(y_1, y_2, y_3, X_3) = P(y_3|X_3)f_3(y_1, y_2, X_3).$$

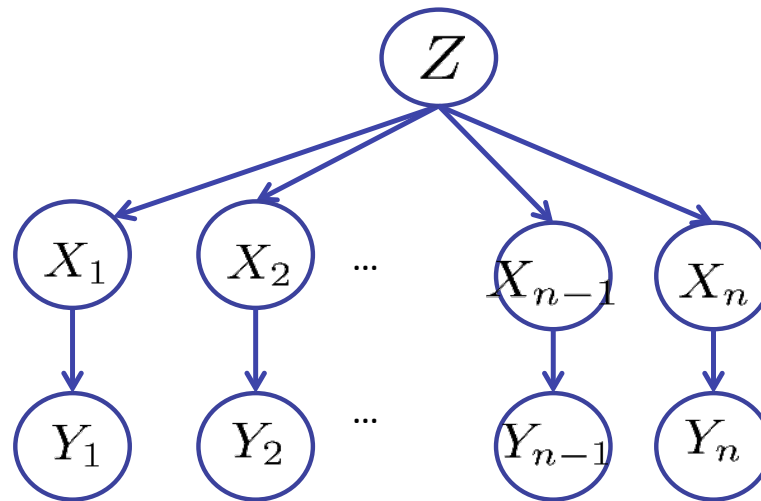
Normalizing over X_3 gives $P(X_3|y_1, y_2, y_3)$.



Computational complexity critically depends on the largest factor being generated in this process. Size of factor = number of entries in table. In example above (assuming binary) all factors generated are of size 2 --- as they all only have one variable (Z , Z , and X_3 respectively).

Variable Elimination Ordering

- For the query $P(X_n | y_1, \dots, y_n)$ work through the following two different orderings as done in previous slide: Z, X_1, \dots, X_{n-1} and X_1, \dots, X_{n-1}, Z . What is the size of the maximum factor generated for each of the orderings?



- Answer: 2^{n+1} versus 2^2 (assuming binary)
- In general: the ordering can greatly affect efficiency.

VE: Computational and Space Complexity

- The computational and space complexity of variable elimination is determined by the largest factor
- The elimination ordering can greatly affect the size of the largest factor.
 - E.g., previous slide's example 2^n vs. 2
- Does there always exist an ordering that only results in small factors?
 - No!

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

Variable Elimination

- Interleave joining and marginalizing
- 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