CS 1810 Spring 2025 Section 8: Bayesian Networks and Inference

1 Bayesian Networks

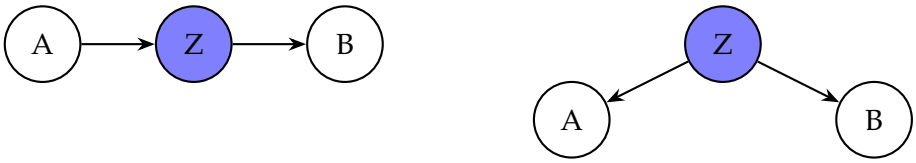
A Bayesian network is a graphical model that represents random variables and their dependencies using a directed acyclic graph. Bayesian networks are useful because they allow us to efficiently model joint distributions over many variables by taking advantage of the local dependencies. With Bayesian networks, we can easily reason about conditional independence and perform inference on large joint distributions.

1.1 D-separation rules

D-separation rules allow us to determine which variables in a network are independent under specific observation assumptions.

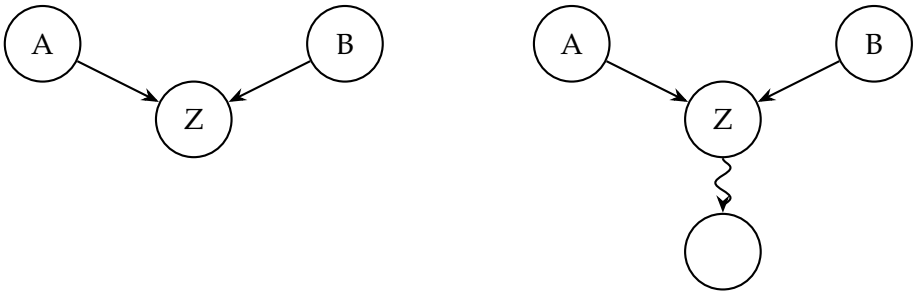
Let X_A and X_B denote sets of variables that we are interested in reasoning about. Recall that an arrow $X_A \to X_B$ indicates that X_B depends on X_A. We say X_A and X_B are *d-separated* by a set of evidence X_E (observed information) if **every** undirected path from X_A to X_B is *blocked*. There are two cases for a path being blocked:

1. There is a node Z with non-converging arrows on the path, and $Z \in X_E$ (i.e., we observe Z). In this example, the shaded node indicates an evidence node.



   Intuitively, observing Z blocks the flow of information from A to B. Given this structure, we can also factor the joint probability distribution for all three variables in the left example as $p(A, Z, B) = p(A)p(Z|A)p(B|Z)$ and $p(A, Z, B) = p(Z)p(A|Z)p(B|Z)$ in the right example.

2. There is a node Z with converging arrows on the path, and neither Z nor its descendants are in X_E (i.e., we don't observe Z or any of its descendants).



   We can factor out the joint distribution here as $p(A, B, Z) = p(A)p(B)p(Z|A, B)$. Intuitively, information flow from A to B is blocked by the unobserved node Z. If we observe Z, then

this might give us information about how much A or B may have contributed to Z adopting a value. This phenomenon where observing Z creates conditional dependence is called *explaining away*.

In each path, only one node Z needs to fall under one of the two cases described above for the whole path to be blocked. Remember though that we need to check **every** undirected path from X_A to X_B to verify d-separation. If there exists an unblocked path, then X_A and X_B are not d-separated.

If X_A and X_B are d-separated by X_E (i.e., all paths are blocked), then X_A and X_B are conditionally independent given X_E, i.e. $X_A \perp\!\!\!\perp X_B \mid X_E$.

2 Exercise: Network Basics

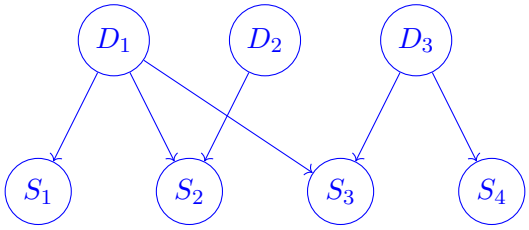
A patient goes to the doctor for a medical condition, and the doctor suspects 3 diseases as the cause of the condition. The 3 diseases are D_1, D_2, and D_3, and they are independent from each other (given no other observations). There are 4 symptoms S_1, S_2, S_3, and S_4, and the doctor wants to check for their presence in order to find the most probable cause. Suppose we know that

- S_1 can be caused by D_1
- S_2 can be caused by D_1 and D_2
- S_3 can be caused by D_1 and D_3
- S_4 can be caused by D_3.

Assume all random variables are Bernoulli, i.e. the patient has the disease/symptom or not.

Q1: Draw a Bayesian network to represent this problem with the following variable ordering: $D_1, D_2, D_3, S_1, S_2, S_3, S_4$.

Solution: Note that there are many valid networks (depending on the chosen variable ordering), some more efficient (i.e. requiring fewer parameters) than others. Here is a compact representation that comes from variable ordering $D_1, D_2, D_3, S_1, S_2, S_3, S_4$. (Recall that all dependencies to earlier variables need to be indicated with edges).



Q2: Write down the expression for the joint probability distribution given this network.

Solution:

$$p(D_1, D_2, D_3, S_1, S_2, S_3, S_4)$$
$$= p(D_1)p(D_2)p(D_3)p(S_1|D_1)p(S_2|D_1, D_2)p(S_3|D_1, D_3)p(S_4|D_3)$$

Q3: How many parameters are required to describe this joint distribution?

Solution: Recall that the unconditional distribution of a Bernoulli random variable has only one parameter. We'll add more parameters for the conditional distributions depending on the number of random variables we condition on and the number of possible values they take on.

Conditional Probability Table	Number of Parameters
$p(D_1)$	1
$p(D_2)$	1
$p(D_3)$	1
$p(S_1\|D_1)$	2
$p(S_2\|D_1, D_2)$	4
$p(S_3\|D_1, D_3)$	4
$p(S_4\|D_3)$	2
Total Number of Parameters	15

Q4: How many parameters would be required to represent the CPTs in a Bayesian network if there were no conditional independences between variables?

Solution: The network would be structured as a clique, and considering order $D_1, D_2, D_3, S_1, S_2, S_3, S_4$, the number of parameters for the CPTs would be $1 + 2 + 4 + 8 + 16 + 32 + 64 = 127$.

Conditional Probability Table	Number of Parameters
$p(D_1)$	1
$p(D_2\|D_1)$	2
$p(D_3\|D_1, D_2)$	4
$p(S_1\|D_1, D_2, D_3)$	8
$p(S_2\|D_1, D_2, D_3, S_1)$	16
$p(S_3\|D_1, D_2, D_3, S_1, S_2)$	32
$p(S_4\|D_1, D_2, D_3, S_1, S_2, S_3)$	64
Total Number of Parameters	127

(We can see there is no saving relative to specifying the joint probability distribution directly, which would require $2^7 - 1 = 127$ numbers.)

Q5: What diseases do we gain information about when observing the fourth symptom ($S_4 = true$)?

Solution: We have independence relations $I(D_1, S_4)$ (since the path is blocked without observing S_3 and $I(D_2, S_4)$ (since the path is blocked at both S_2 and S_3). What is left is dependence between D_3 and S_4. Thus, we only learn information about D_3.

Q6: Suppose we know that the third symptom is present ($S_3 = true$). What does observing the fourth symptom ($S_4 = true$) tell us now?

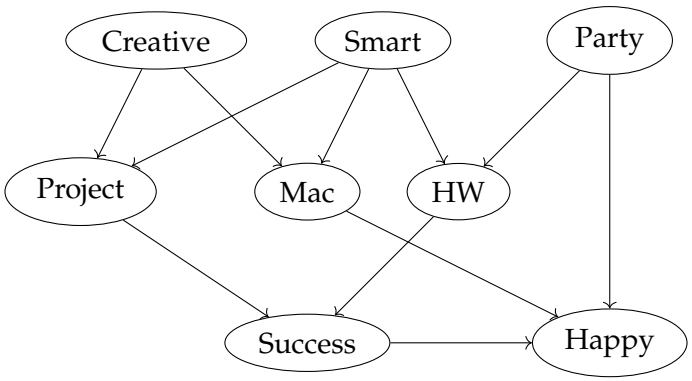
Solution: With $S_3 = true$, observing $S_4 = true$ now also gives us information about D_1 (via 'explaining away', or using d-separation, because the D_1 to S_4 path is no longer blocked at S_3). We still don't learn any information about D_2 because the D_2 to S_4 path remains blocked at S_2.

3 Exercise: D-Separation

As part of a comprehensive study of the role of CS 181 on people's happiness, we have been collecting important data from students. In an entirely optional survey that all students are required to complete, we ask the following highly objective questions:

Question	Variable name	Possible values
Do you party frequently?	Party	Yes / No
Are you smart?	Smart	Yes / No
Are you creative?	Creative	Yes / No
Did you do well on all your homework assignments?	HW	Yes / No
Do you use a Mac?	Mac	Yes / No
Did your last major project succeed?	Project	Yes / No
Did you succeed in your most important class?	Success	Yes / No
Are you currently happy?	Happy	Yes / No

After consulting behavioral psychologists we build the following model:



Q1: True or False: $Party$ is independent of $Success$ given HW.

Solution: False; there is a path that is not blocked: $Party - HW - Smart - Project - Success$ has neither a converging arrows not in the set of evidence or a non-converging arrows in the set.

Q2: True or False: $Creative$ is independent of $Happy$ given Mac.

Solution: False; there is a path that is not blocked: $Creative - Project - Success - Happy$

Q3: True or False: $Party$ is independent of $Smart$ given $Success$.

Solution: False; there is a path that is not blocked between $Party$ and $Smart$: the path $Party - HW - Success$ is not blocked because the converging arrows node at HW has a descendant ($Success$) in the evidence.

Q4: True or False: $Party$ is independent of $Creative$ given $Happy$.

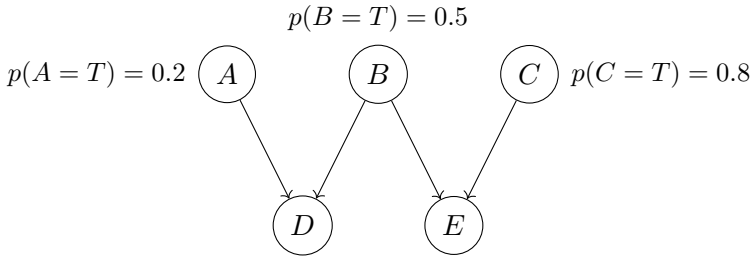
Solution: False; there is a path that is not blocked between $Party$ and $Creative$ through the converging arrows at $Happy$. There are actually multiple not-blocked paths – can you find them?

Q5: True or False: $Party$ is independent of $Creative$ given $Success$, $Project$ and $Smart$.

Solution: True! All paths between $Party$ and $Creative$ are blocked. Working from $Party$, the paths that come through $Happy$ are blocked there (converging arrows, no evidence). Those that come through HW and $Smart$ are blocked at $Smart$. Those that come through $HW, Success, Project$ are blocked at $Project$.

4 Exercise: Inference

Consider the following Bayesian network, where all variables are Bernoulli.



A	B	$p(D = T\|A, B)$
F	F	0.9
F	T	0.6
T	F	0.5
T	T	0.1

B	C	$p(E = T\|B, C)$
F	F	0.2
F	T	0.4
T	F	0.8
T	T	0.3

Q1: What is the probability that all five variables are simultaneously false (F)?

Solution: Based on this network, we can factor out the joint distribution as

$$p(A, B, C, D, E) = p(A)p(B)p(C)p(D|A, B)p(E|B, C)$$

Now let X be the event that all five variables are false. Then we have

$$\begin{aligned} P(X) &= p(A = F, B = F, C = F, D = F, E = F) \\ &= p(A = F)p(B = F)p(C = F)p(D = F|A = F, B = F)p(E = F|B = F, C = F) \\ &= (0.8)(0.5)(0.2)(0.1)(0.8) \\ &= 0.0064 \end{aligned}$$

Q2: What is the probability that A is false given that the remaining variables are all known to be true (T)?

Solution: For this part, we need to calculate $p(A = F|B = T, C = T, D = T, E = T)$. Let Y be the event that B, C, D, E are all true. Then we want to find $P(A = F|Y)$.

By the definition of conditional probability,

$$p(A = F|Y) = \frac{p(A = F, Y)}{P(Y)} = \frac{p(A = F, Y)}{P(A = F, Y) + P(A = T, Y)}$$

The joint probabilities $p(A = F, Y)$ and $p(A = T, Y)$ can be computed as:

$$\begin{aligned} p(A = F, Y) &= p(A = F)p(B = T)p(C = T)p(D = T|A = F, B = T)p(E = T|B = T, C = T) \\ &= (0.8)(0.5)(0.8)(0.6)(0.3) \\ &= (0.05760) \end{aligned}$$

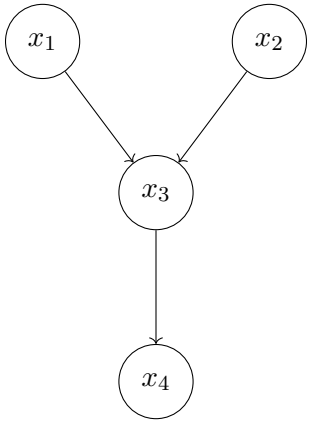
$$\begin{aligned} p(A = T, Y) &= p(A = T)p(B = T)p(C = T)p(D = T|A = T, B = T)p(E = T|B = T, C = T) \\ &= (0.2)(0.5)(0.8)(0.1)(0.3) \\ &= (0.00240) \end{aligned}$$

Finally, we can plug this in to get:

$$p(A = F|B = T, C = T, D = T, E = T) = p(A = F|Y) = \frac{.05760}{.05760 + .00240} = .96$$

5 Variable Elimination in Bayesian Networks

We apply an inference algorithm called variable elimination to the following Bayesian network:



Assume that all of the random variables are Categorical, meaning they are discrete random variables that can each take on k possible values. In this network, we can factor the joint distribution as

$$p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2)p(x_3|x_1, x_2)p(x_4|x_3)$$

If we wanted to calculate the marginal distribution of x_4 – the distribution of x_4 without conditioning on any other information – we could naively marginalize out all other variables (also known as LOTP):

$$\begin{aligned} p(x_4) &= \sum_{x_1}\sum_{x_2}\sum_{x_3} p(x_1, x_2, x_3, x_4) \\ &= \sum_{x_1}\sum_{x_2}\sum_{x_3} p(x_1)p(x_2)p(x_3|x_1, x_2)p(x_4|x_3) \end{aligned}$$

To obtain the full distribution, we need to compute this sum for k possible values of x_4 (we say k for simplicity, but we could actually just compute the sum $k-1$ times and find the last $p(x_4)$ via complementary probability). For each of these values of x_4, we then have to consider the k^3 possible values that (x_1, x_2, x_3) could take on. Hence, this naive calculation takes $O(k^4)$ computations. We thus see that if we had n variables, the number of computations grows exponentially, i.e. $O(k^n)$.

Note that Bayesian nets encode dependencies between variables, which we can use to calculate the marginal distribution more efficiently. By reordering the sums and eliminating one variable at

a time, we derive the variable elimination procedure:

$$\begin{aligned}
p(x_4) &= \sum_{x_1}\sum_{x_2}\sum_{x_3} p(x_1)p(x_2)p(x_3|x_1,x_2)p(x_4|x_3) \\
&= \sum_{x_3} p(x_4|x_3) \sum_{x_2} p(x_2) \sum_{x_1} p(x_3|x_1,x_2)p(x_1) \\
&= \sum_{x_3} p(x_4|x_3) \sum_{x_2} p(x_2)p(x_3|x_2) \\
&= \sum_{x_3} p(x_4|x_3)p(x_3) \\
&= p(x_4)
\end{aligned}$$

The elimination procedure goes from the innermost sum to the outermost sum:

1. First, we eliminate x_1 to obtain a k by k matrix which stores the value of $p(x_3|x_2)$ for each possible pair (x_2, x_3). These values are computed by using the identity

   $$p(x_3|x_2) = \sum_{x_1} p(x_3|x_1,x_2)p(x_1)$$

   Note that since we sum over the k possible values of x_1 for each of the k^2 pairs (x_2, x_3), this step takes $O(k^3)$ computations.

2. Now in the next step we eliminate x_2 and compute the k-dimensional vector $p(x_3)$. Here we use the identity

   $$p(x_3) = \sum_{x_2} p(x_2)p(x_3|x_2),$$

   so we have to sum over the k values of x_2 for each of the k values of x_3. This results in $O(k^2)$ computations.

3. Finally, we eliminate x_3 and compute the k-dimensional vector $p(x_4)$ through using the final identity

   $$p(x_4) = \sum_{x_3} p(x_4|x_3)p(x_3)$$

   Here we sum over k possible values of x_3 for k values of x_4 to get a total of $O(k^2)$ computations.

It follows that we have performed $O(k^3) + O(k^2) + O(k^2) = O(k^3)$ total computations here, rather than the $O(k^4)$ computations we needed in the naive method.

Alternatively, we could have eliminated variables in a different order:

$$\begin{aligned}
p(x_4) &= \sum_{x_1}\sum_{x_2}\sum_{x_3} p(x_1)p(x_2)p(x_3|x_1,x_2)p(x_4|x_3) \\
&= \sum_{x_1} p(x_1) \sum_{x_2} p(x_2) \sum_{x_3} p(x_3|x_1,x_2)p(x_4|x_3) \\
&= \sum_{x_1} p(x_1) \sum_{x_2} p(x_2)p(x_4|x_1,x_2) \\
&= \sum_{x_1} p(x_1)p(x_4|x_1) \\
&= p(x_4)
\end{aligned}$$

However, the ordering matters. Here when we eliminate x_3 first, we compute a $k \times k \times k$ dimensional object $p(x_4|x_1, x_2)$ where each entry requires summing over the k values of x_3. Hence, this is already $O(k^4)$ computations, which is the same complexity as the naive method.

In general, the computational cost of variable elimination depends on the number of variables in these intermediate factors, in particular the largest object computed ('tree-width'). The problem of choosing an optimal ordering is actually NP-hard, if we don't make any assumptions on the graph structure. However, there exists an optimal strategy if the graph is a polytree, i.e. a directed, acyclic graph that would be a tree if we made it undirected:

1. Prune any variables that are not ancestors of the query or evidence. Note that in an expression like $p(x|y)$, x is the query and y is the evidence.

2. Find the variables which are furthest from the query and work backwards to perform variable elimination.

We see that this aligns with how our first elimination procedure in the example above performed better than the second elimination procedure, since x_3 is closer to the query x_4 than x_1.

5.1 Exercise: Variable Elimination

Consider the Bayesian network described in above, and assume the following Conditional Probability Table (CPT). Let $x_i \in \{0, 1\}$ denote the values that variable X_i can take. Our goal is to find $p(x_4)$.

x_1	$p(x_1)$
0	0.3
1	0.7

x_2	$p(x_2)$
0	0.6
1	0.4

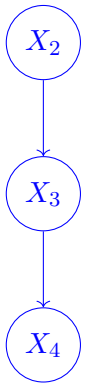
x_3	x_1	x_2	$p(x_3\|x_1, x_2)$
0	0	0	0.5
0	0	1	0.2
0	1	0	0.9
0	1	1	0.5
1	0	0	0.5
1	0	1	0.8
1	1	0	0.1
1	1	1	0.5

x_4	x_3	$p(x_4\|x_3)$
0	0	0.7
0	1	0.1
1	0	0.3
1	1	0.9

1. Eliminate X_1 first. Draw the resulting Bayesian network and compute the missing CPT.
2. Eliminate X_3 first. Draw the resulting Bayesian network and compute the missing CPT.
3. How many sum-product calculations do each of these variable elimination orders require? Which one is preferable?

Solution:

1. The resulting network is:



The variable elimination process eliminates X_1 by marginalizing out X_1

$$p(x_3|x_2) = \sum_{x_1} p(x_3|x_1, x_2)p(x_1).$$

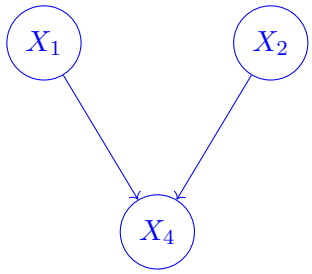
For example:

$$\begin{aligned} p(X_3 = 0|X_2 = 0) &= \sum_{x_1 \in \{0,1\}} p(X_3 = 0|X_1 = x_1, X_2 = 0)p(X_1 = x_1) \\ &= 0.5 \cdot 0.3 + 0.9 \cdot 0.7 \\ &= 0.78 \end{aligned}$$

This is a sum-product calculation, and we need to do one for each value of X_2 and X_3. Thus, there are four sum-product calculations in total. The resulting CPT is:

x_3	x_2	$p(x_3\|x_2)$
0	0	0.78
0	1	0.41
1	0	0.22
1	1	0.59

2. The resulting network is



The variable elimination process eliminates X_3 by marginalizing out X_3: $p(x_4|x_1, x_2) = \sum_{x_3} p(x_4|x_3)p(x_3 \mid x_1, x_2)$. This would be the first intermediate term. For example:

$$\begin{aligned} p(X_4 = 0|X_1 = 0, X_2 = 0) &= \sum_{x_3 \in \{0,1\}} p(X_4 = 0|X_3 = x_3)p(X_3 = x_3|X_1 = 0, X_2 = 0) \\ &= 0.7 \cdot 0.5 + 0.1 \cdot 0.5 \\ &= 0.40 \end{aligned}$$

We need to do this for each combination of values for X_1, X_2 and X_4. Thus, there are eight sum-product calculations in total. The resulting CPT is:

x_4	x_1	x_2	$p(x_4\|x_1, x_2)$
0	0	0	0.40
0	0	1	0.22
0	1	0	0.64
0	1	1	0.40
1	0	0	0.60
1	0	1	0.78
1	1	0	0.36
1	1	1	0.60

3. In these variable elimination operations, we need to compute intermediate terms. The cost of computing these depends on the number of variables that they mention, since each variable increases the number of required sum-product calculations by a factor of $k = 2$.

   For the first ordering, the intermediate terms are:

   - $p(x_3 \mid x_2)$: mentions x_2 and x_3, and thus requires four sum-product calculations (for each row in the original CPT)
   - $p(x_3)$: mentions x_3 and thus requires two sum-product calculations
   - $p(x_4)$: mentions x_4 and thus requires two sum-product calculations

   We have a total of $4 + 2 + 2 = 8$ sum-product calculations.

   For the second ordering, the intermediate terms are:

   - $p(x_4 \mid x_1, x_2)$: mentions x_1, x_2 and x_4, and thus requires eight sum-product calculations (for each row in the original CPT)
   - $p(x_4 \mid x_1)$: mentions x_1 and x_4, and thus requires four sum-product calculations
   - $p(x_4)$: mentions x_4 and thus requires two sum-product calculations

   We have a total of $8 + 4 + 2 = 14$ sum-product calculations.

   Thus, we see that the first ordering is preferable since it requires fewer computational steps.