UC Berkeley – Computer Science
CS188: Introduction to Artificial Intelligence
Josh Hug and Adam Janin

Midterm II, Fall 2016 **Solutions**

This test has **7** questions worth a total of **100** points, to be completed in 110 minutes. The exam is closed book, except that you are allowed to use a single two-sided hand written cheat sheet. No calculators or other electronic devices are permitted. Give your answers and show your work in the space provided.

Write the statement out below in the blank provided and sign. You may do this before the exam begins. Any plagiarism, no matter how minor, will result in an F.

"I have neither given nor received any assistance in the taking of this exam."

_____

_____

Signature: Josh Hug

Name: Josh Hug **Your EdX Login:** _____

SID: _____ **Name of person to left: Josh Hug**

Exam Room: Josh Hug **Name of person to right: Josh Hug**

Primary TA: Adam Janin

- ◯ indicates that only one circle should be filled in.
- ▫ indicates that more than one box may be filled in.
- Be sure to fill in the ◯ and ▫ boxes completely and erase fully if you change your answer.
- There may be partial credit for incomplete answers. Write as much of the solution as you can, but bear in mind that we may deduct points if your answers are much more complicated than necessary.
- There are a lot of problems on this exam. Work through the ones with which you are comfortable first. **Do not get overly captivated by interesting problems or complex corner cases you're not sure about.**
- Not all information provided in a problem may be useful.
- **This exam will be graded during the year 2016.**
- **There are some problems on this exam with a slow brute force approach and a faster, clever approach. Think before you start calculating with lots of numbers!**
- Write the last four digits of your SID on each page in case pages get shuffled during scanning.

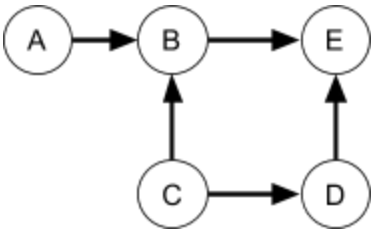
Problem	1	2	3	4	5	6	7
Points	14	9	12	20	14	13	18

Optional. Mark along the line to show your feelings on the spectrum between :(and ☺.

Before exam: [:(**Josh Hug** ☺].

After exam: [:(**Josh Hug** ☺].

1. Bayesics (14 pts) Model-Q1 is a Bayes' Net consisting of the graph and probability tables shown below:



A	P(A)
0	0.6
1	0.4

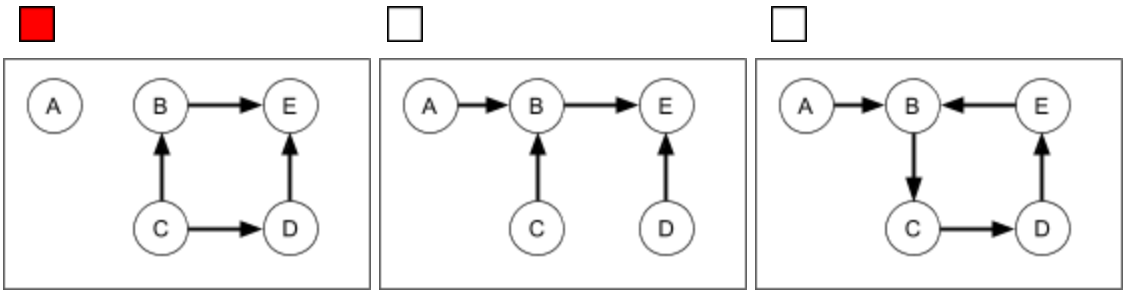
C	D	P(D|C)
0	0	0.2
0	1	0.8
1	0	0.3
1	1	0.7

A	C	B	P(B|A,C)
0	0	0	0.2
0	0	1	0.8
0	1	0	0.4
0	1	1	0.6
1	0	0	0.2
1	0	1	0.8
1	1	0	0.4
1	1	1	0.6

B	D	E	P(E|B,D)
0	0	0	0.75
0	0	1	0.25
0	1	0	0.5
0	1	1	0.5
1	0	0	0.85
1	0	1	0.15
1	1	0	0.4
1	1	1	0.6

C	P(C)
0	0.4
1	0.6

i. (3 pts) Check the boxes above the Bayes' nets below that could also be valid for the above probability tables.



Justification: The first Bayes' net implies B is independent of A given C, which is true: looking at P(B|A,C), we see that $\boldsymbol{P}(\boldsymbol{b}|\boldsymbol{a},\boldsymbol{c}) = \boldsymbol{P}(\boldsymbol{b}|\boldsymbol{c})$ for all $\boldsymbol{a}$.

The second Bayes' net implies C and D are independent, which isn't true: for instance,

$\boldsymbol{P}(\boldsymbol{D}=1) = \boldsymbol{P}(\boldsymbol{D}=1|\boldsymbol{C}=0)\boldsymbol{P}(\boldsymbol{C}=0) + \boldsymbol{P}(\boldsymbol{D}=1|\boldsymbol{C}=1)\boldsymbol{P}(\boldsymbol{C}=1) = (0.8)(0.4) + (0.7)(0.6) = 0.74$ but

$\boldsymbol{P}(\boldsymbol{D}=1|\boldsymbol{C}=0) = 0.8$.

The third Bayes' net is cyclic, and is therefore not a valid Bayes' net.

ii. (2 pts) Caryn wants to compute the distribution P(A,C|E=1) using ***prior sampling*** on Model-Q1 (given at the top of this page). She draws a bunch of samples. The first of these is (0, 0, 1, 1, 0), given in (A, B, C, D, E) format. What's the probability of drawing this sample?

Solution: P(A=0, B=0, C=1, D=1, E=0)
= P(A=0) * P(C=1) * P(B=0 | A=0, C=1) * P(D=1 | C=1) * P(E=0 | B=0, D=1)
= 0.6 * 0.6 * 0.4 * 0.7 * 0.5
= 0.0504

iii. (2 pts) Give an example of an inference query for Model-Q1 with **one query variable and one evidence variable** that could be estimated more efficiently (in terms of runtime) using ***rejection sampling*** than by using ***prior sampling***. If none exist, state "not possible".

Example Solution: P(C|A=0)
Explanation: Rejection sampling provides an efficiency advantage when it allows us to realize that a sample is going to be unusable before it is fully generated. Thus, the query needs to be able to reject a sample (by comparing against the evidence variable) before all variables are sampled. The last variable sampled must be E (since it is the only sink (no outgoing edges) in the Bayes's net), so any solution where the evidence is not E (and the query is not the same as the evidence), then rejection sampling will be more efficient than prior sampling.

Common Mistakes:

- It was important to actually give evidence, e.g. P(A|B=1), and not P(A|B). The latter doesn't actually incorporate evidence for rejection sampling to use.
- While it is true that A and B are independent, Model-Q1 doesn't incorporate this information, and sampling methods applied to this model won't take advantage of this information either. (Key point: sampling methods are applied to a model, and not directly on the joint distribution.)
- Note that the domains of the variables (as visible in the CPTs) are {0, 1}. Specifically, they are not {+a, -a} or anything similar.

iv. (2 pts) Give an example of an inference query for Model-Q1 with **one query variable and one evidence variable** for which ***rejection sampling*** provides no efficiency advantage (in terms of runtime) over using ***prior sampling***. If none exist, state "not possible".

Example Solution: P(C|E=0)
Explanation: For the same logic as the previous part, if E is the evidence, then rejection and prior sampling will have the same runtime.

v. (2 pts) Now Caryn wants to determine P(A,C|E=1) for Model-Q1 using likelihood weighting. She draws the five samples shown below, which are given in (A, B, C, D, E) format, where the leftmost sample is "Sample 1" and the rightmost is "Sample 5". What are the weights of the samples S1 and S3?

weight(S1): $\boldsymbol{P}(\boldsymbol{E} = 1 \mid \boldsymbol{B} = 0, \boldsymbol{D} = 1) = 0.5$ weight(S3): $\boldsymbol{P}(\boldsymbol{E} = 1 \mid \boldsymbol{B} = 0, \boldsymbol{D} = 1) = 0.5$

S1: (0, 0, 1, 1, 1) S2: (0, 0, 1, 1, 1) S3: (1, 0, 1, 1, 1) S4: (0, 1, 0, 0, 1) S5: (0, 1, 0, 0, 1)

vi. (1 pt) For the same samples as in part v, compute P(A=1,C=1|E=1) for Model-Q1. Express your answer as a simplified fraction (e.g. 2 / 3 instead of 4 / 6).

The weights are (left to right): 0.5, 0.5, 0.5, 0.15, 0.15. This can be found using just two table lookups: the weights of S1, S2, and S3 are all the same as calculated in the previous part, and the weights of S4 and S5 are $\boldsymbol{P}(\boldsymbol{E} = 1 \mid \boldsymbol{B} = 1, \boldsymbol{D} = 0) = 0.15$.
Only S3 matches the query A=1, C=1, so we infer P(A=1,C=1|E=1) to be the following:

weight(S3) / (sum of all weights) = 0.5 / (0.5 + 0.5 + 0.5 + 0.15 + 0.15) = 5 / 18

vii. (2 pts) Select True or False for each of the following:

True	False	
○	●	When there is no evidence, prior sampling is guaranteed to yield the exact same answer as inference by enumeration.
●	○	When collecting a sample during likelihood weighting, evidence variables are not sampled.
○	●	When collecting a sample during rejection sampling, variables can be sampled in any order.
●	○	Gibbs sampling is a technique for performing approximate inference, not exact inference.

2. Pacmanian Language Modeling (9 pts)

Archaeologists uncover ancient Pacmanian ruins and discover a fragment of their rare writing. Your job is to analyze the Pacmanian language based ***only*** on the tiny amount of data uncovered (and your knowledge of material in this class). Specifically, the fragment contains the following 20 word sentence:

```
.ge  7YcY  0go  4mc  'Y`  7YcY  0go  4mc  "gh  7YcY  0go  !ogg  .ge  7YcY  4mc  7YcY  'Y`
7YcY  4mc  7YcY
```

Notes and hints:

- W_i represents the ***i***th word in the fragment. So W_0 = Nom and W_8 = Bop
- The words and counts are: Awoo (1) Bop (1) Gah (2) Nom (2) Pow (3) Tuk (4) Waka (7)
- Your parameters should be computed from the 20 word fragment above!
- For those of you who are up to date on the post-MT2 material, ***do not use Laplacian smoothing***!

i. (1 pt) Given a new four word sentence in Pacmanian, draw a Bayes' Network representing a ***bigram language model***, where a word at position ***i*** can be dependent on the word that precedes it. You do not need to provide the conditional probability tables (CPTs).

Solution: 

ii. (2 pts) Given a bigram language model, write down the formula for the joint probability $P(W_0, W_1, W_2, W_3)$ in terms of the four smaller factors. You do not need to provide the CPTs.

Solution: $P(W_0) * P(W_1 | W_0) * P(W_2 | W_1) * P(W_3 | W_2)$

iii. (2 pts) Given a bigram language model with parameters collected from the 20 word fragment, what is the probability of the Pacmanian sentence "Nom Waka Pow Awoo"? Hint: Only compute the parameters you need.

Solution:

P("Nom") = 2/20

P("Waka" | "Nom") = 2/2

P("Pow" | "Waka") = 3/6 This one is a bit tricky because of edge effects. 3/7 is probably also okay.

P("Awoo" | "Pow") = 1/3

$\frac{2*2*3*1}{20*2*6*3}$ or $\frac{12}{720}$ or $\frac{1}{60}$ or 0.0166

iv. (2 pts) Given a four word sentence in Pacmanian, draw a Bayes' Network representing a ***trigram language model***, where a word at position ***i*** can be dependent on the ***two*** words that precede it. You do not need to provide the CPTs.

Solution: 

v. (2 pts) Given a trigram language model, write down the formula for the joint probability $P(W_0, W_1, W_2, W_3)$ in terms of the smaller factors.

Solution: $P(W_0) * P(W_1 | W_0) * P(W_2 | W_0, W_1) * P(W_3 | W_1, W_2)$

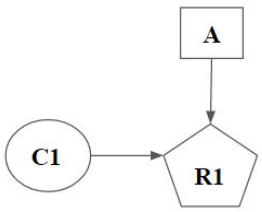
3. Smoke Weed Everyday (12 pts)

Allen is trying to decide whether to vote to legalize marijuana. He wants marijuana to be legalized, but he's pretty lazy and doesn't want to wait in line. He knows that while it's a close decision, he doesn't think his vote will actually matter.

Part A: Allen lives in some County C1, e.g Alameda. Our model has three possible scenarios for C1:

- Scenario —**m**: marijuana is not legalized regardless of Allen's vote (probability 1/2)
- Scenario +**m**: marijuana is legalized regardless of Allen's vote (probability 1/4)
- Scenario **@**: marijuana is legalized but only if Allen votes; otherwise it is not (probability 1/4)

We can model this interaction between Allen's action A (+**v** for vote, —**v** for not vote) and C1 using a new type of node called a variable node. Its value is deterministically given by its inputs. The resulting network is shown below. The table on the right lists the result R1 of the county(C1)'s decision on marijuana, given the possible values for C1 and A. R1 is either (+**m**)arijuana legalized or (—**m**)arijuana not legalized**.** The black squares mean that Allen's vote doesn't matter in those cases.



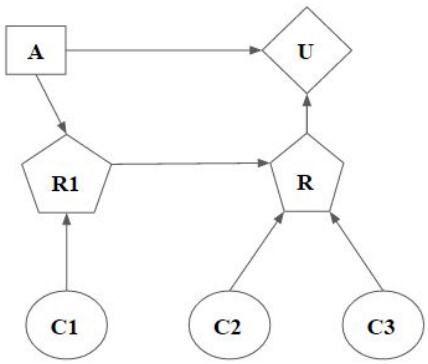
C1	P(C1)
+m	1/4
–m	1/2
@	1/4

C1	A	R1
+m		+m
–m		–m
@	+v	+m
@	–v	–m

i. (0.5 pt) What is the chance of marijuana legalization assuming Allen votes (i.e. P(R1 = +m | A = +v))?
Solution: If Allen votes, marijuana is legalized iff $C1 = +m$ or $C1 = @$, which has probability 1/2.

ii. (0.5 pt) What is the chance of marijuana legalization assuming Allen doesn't vote (i.e. P(R1 = +m | A = −v))?
Solution: If Allen doesn't vote, marijuana is legalized iff $C1 = +m$, which has probability 1/4.

Part B: Allen lives in a small state with only three counties, C1, C2, and C3. In counties C2 and C3, (+**m**)arijuana legalized or (−**m**)arijuana not legalized have a 50% chance each of occurring and Allen's vote does not affect those counties. Let R represent the outcome of the vote for the state, which can either be **R = +m** for legalization, or **R = −m** for not legalizing marijuana. The majority result of the counties determines the result of the state, e.g. if R1 = + m, C2 = +m, C3 = −m, then R = +m. The resulting model (including Allen's utility function) is given in the figure at the top of the next page. For space reasons, no table is shown for R in the figure. Larger version on next page.



Make sure to show your work ***clearly*** on this page, as it will be used for determining partial credit!

iii. (6 pts) **What is** $MEU(\oslash)$, i.e. the maximum expected utility if Allen doesn't have any evidence? To maximize his utility, **should Allen vote**? Note that Allen's utility is a function of the result AND his action.

Explanation: We start by calculating the CPT $P(R = +m|A)$.

We note that if $C2 = C3$ (1/2 probability), then $R1$ does not matter. If they disagree, then the election falls to the result of $R1$, which we calculated the distribution of in the previous part. Thus, we have:

$P(R = +m|A = +v) = P(C2 = C3 = +v) + P(C2 \neq C3) * P(R1 = +m|A = +v) = 1/4 + 1/2 * 1/2 = 1/2$

$P(R = +m|A = -v) = P(C2 = C3 = +m) + P(C2 \neq C3) * P(R1 = +m|A = -v) = 1/4 + 1/2 * 1/4 = 3/8$

Now, we can calculated the expected utilities for each possible action:

$EU(\{+v\}) = U(+m, +v) * P(R = +m|A = +v) + U(-m, +v) * P(R = -m|A = +v) = 8 * 1/2 + 0 * 1/2 = 4$

$EU(\{-v\}) = U(+m, -v) * P(R = +m|A = -v) + U(-m, -v) * P(R = -m|A = -v) = 16 * 3/8 + 4 * 5/8 = 8.5$

Thus, $MEU(\oslash) = max(4, 8.6) = 8.5$

$MEU(\oslash)$**: 8.5** Should Allen vote? **no**

iv. (3 pts) Suppose Allen could somehow know that C1 = +m. How much would he value additional information about the results of C2? In other words, what is VPI(C2 | C1 = +m)?

Solution: 0. If $C1 = +m$, then Allen's vote cannot change the result of the election, so it will be optimal for Allen to note vote. From there, knowing any other variable will not change Allen's optimal action, so the VPI of any other variable is 0.

v. (2 pts) Is VPI(C2|C1= @) greater than, less than, or equal to VPI(C2|C1= +m)?

● greater than ○ less than ○ equal

Explanation: As noted above, $VPI(C2|C1 = +m) = 0$. We can also reason that $VPI(C2|C1 = @) > 0$, without calculating the value exactly, because knowing that $C1 = @$ (i.e. Allen's vote can matter) means that Allen does want more information to judge whether or not he should bother voting. Thus,

$VPI(C2|C1 = @) > VPI(C2|C1 = +m)$.

4. Probability Potpourri Pain Problem (20 pts)

i. (6 pts) Suppose we know that $A \perp B$ and $B \perp C \mid D$. Which of the following statements (if any) are ***definitely true***? Which are ***sometimes*** true? Which are ***never*** true? Here $\perp$ represents the independence symbol.

DEFINITELY	SOMETIMES	NEVER	
○	●	○	$P(A, B, C, D) = P(A)\,P(B)\,P(C)\,P(D)$

Explanation: This can be true (e.g. if the four variables are mutually independent), but is not always true.

○	●	○	$P(A, B, C, D) = P(A)\,P(B)\,P(C\mid D)$

Explanation: This can be true (e.g. if the four variables are mutually independent and P(D) = 1), but is not always true. (This was, by far, the hardest of these question parts.)

●	○	○	$P(A, B, C, D) = P(A)\,P(B\mid A)\,P(C\mid A, B)\,P(D\mid A, B, C)$

Explanation: This is always true as a direct invocation of the chain rule.

○	●	○	$P(B, C\mid D) = P(B)\,P(C\mid D)$

Explanation: This can be true (e.g. if B and D are independent and B and C are independent conditioned on D), but is not always true.

○	●	○	$P(A, C\mid D) = P(C)\,P(A\mid D)$

Explanation: This can be true (e.g. if C and D are independent and A and C are independent conditioned on D), but is not always true.

○	●	○	$P(X, Y\mid Z) = P(X\mid Y)\,P(Y\mid Z)$

Explanation: This is true iff X and Z are independent conditioned on Y.

Comment: Notice that none of the expressions are never true. In general, it's often possible to find weird cases where the probabilities just happen to be correct. (Looking for cases when one or more of the probability terms are equal to 0 or 1 is a useful way of doing this.)

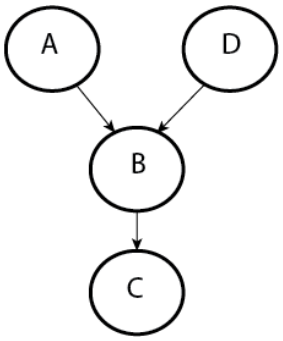
ii. (3 pt) Draw a Bayes' Net that is consistent with the assumptions that $A \perp B$ and $B \perp C \mid D$. Your Bayes' Net may include additional independence assumptions other than these.

Solution: There were many possible solutions for this problem.
The simplest one (which unfortunately only 8% of students got) was simply a Bayes's net with no edges (i.e. where all variables are mutually independent). This was allowed because the question allowed the Bayes's net to make additional independence assumptions.
The most common answer was one where (B, D, C) was an inactive triple (e.g. $B \rightarrow D \rightarrow C$ or $B \leftarrow D \rightarrow C$) and A was either unconnected or connected in a way that kept A and B as independent. (It seems like many students assumed the answer needed to have B and C as potentially dependent when not conditioned on D.)
The most common ***incorrect*** answer included the triple $B \rightarrow D \leftarrow C$, which would make B and C potentially dependent when D is observed.

iii. (1 pt) ○ True ● False: If X is independent of Y, then X is independent of Y given Z, i.e. $X \perp Y \rightarrow X \perp Y \mid Z$

Solution: If the structure of the Bayes's net was $X \rightarrow Z \leftarrow Y$, then this would be false. (See below in part v.)

iv. (4 pts) Given the Bayes' Net below, prove algebraically that A is independent of D. Briefly (less than five words per step) explain or justify each step of your proof. Use the lines given for the steps of your proof. You may not need all lines.



Step	Explanation
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____

$P(A, B, C, D) = P(A)P(D)P(B|A, D)P(C|B)$ [from the bayes net]

$P(A, D) = \sum_{b,c} P(A, b, c, D) = \sum_{b,c} P(A)P(D)P(b|A, D)P(c|b)$ [introduce B and C]

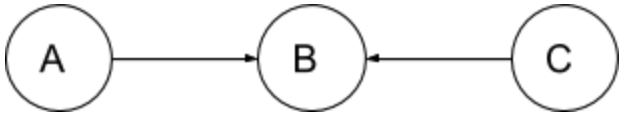
$P(A, D) = P(A)P(D)$ [definition of independence]

OR

$P(A, D) = P(A)P(D)$ [from the bayes net]

A independent of D if $P(A, D) = P(A)P(D)$ [definition of independence]

v. (2 pts) Suppose we have the Bayes' Net model below:

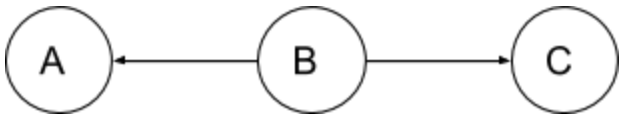


Which of the following statements regarding independence **must be true** given the model?

☐ $A \perp B$ ☐ $A \perp C \mid B$
■ $A \perp C$ ☐ $A \perp B \mid C$
☐ $B \perp C$ ☐ $B \perp C \mid A$

Explanation: The direct edge between A and B rules out any potential guarantees that the two are independent, regardless of conditioning; likewise for B and C. Thus, we can immediately rule out four potential answers. Looking at the relationship between A and C, we see that B is a common child, so A and C are guaranteed to be independent only when B is not observed.

vi. (2 pts) Consider the Bayes' Net model below.

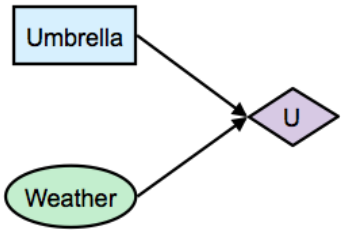


Which of the following independence assumptions **must be true** given the model?

☐ $A \perp B$ ■ $A \perp C \mid B$
☐ $A \perp C$ ☐ $A \perp B \mid C$
☐ $B \perp C$ ☐ $B \perp C \mid A$

Explanation: With the same reasoning as the previous part, we can immediately rule out the same four potential answers. This time, however, when looking at the relationship between A and C, we see that B is a common ***parent***, so A and C are guaranteed to be independent only when B ***is*** observed.

vii. (2 pts) Consider the simple model shown below:



W	P(W)
rain	0.1
sun	0.9

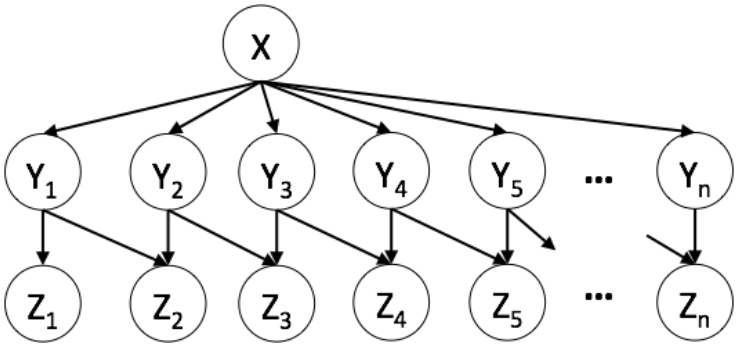
Umbrella	W	U(A, W)
take	rain	0
take	sun	100
leave	rain	0
leave	sun	100

Trivially, we can calculate that the maximum expected utility in the absence of evidence is $\boldsymbol{MEU}(\oslash) = 90$. Suppose we want to calculate the value of knowing that it is raining VPI(w = rain). Calculating, we find that $\boldsymbol{MEU}(\boldsymbol{rain}) = 0$. This implies that the value of knowing that it is raining is -90. In class, we said that VPI is always non-negative. Explain where this argument fails.

Solution: The value of perfect information about a random variable is always nonnegative. However, "VPI(W = rain)" doesn't make sense: we don't actually know that W = rain, as there is only a 10% chance of this happening. In the other 90% probability, w = sun, and $\boldsymbol{MEU}(\boldsymbol{W} = \boldsymbol{sun}) = 100$. Thus, the expected $\boldsymbol{MEU}$, knowing the weather, is still 90, and so $\boldsymbol{VPI}(\boldsymbol{W}) = 0$, which is nonnegative.

5. Bayes' Nets Inference (14 pts)

Consider the following Bayes' Net where all variables are binary.



i. (2 pts) Suppose we somehow calculate the distribution $\boldsymbol{P}(\boldsymbol{Y_n}|\boldsymbol{Z}_1 = \boldsymbol{z}_1, \boldsymbol{Z}_2 = \boldsymbol{z}_2, \cdots, \boldsymbol{Z_n} = \boldsymbol{z_n})$. What is the size of this factor (number of rows)? Note that the Z variables are all observed evidence!

Number of rows in this factor: 2
Explanation: The only variable in the probability distribution that is not observed is $\boldsymbol{Y_n}$, so the factor needs to include both possible values $\boldsymbol{Y_n}$ can take on.

ii. (3 pts) Suppose we try to calculate $\boldsymbol{P}(\boldsymbol{Y_n}|\boldsymbol{Z}_1 = \boldsymbol{z}_1, \boldsymbol{Z}_2 = \boldsymbol{z}_2, \cdots, \boldsymbol{Z_n} = \boldsymbol{z_n})$ using variable elimination, and start by eliminating the variable Y_1. What new factor is generated? ***Give your answer in standard probability distribution notation***, as opposed to using notation that involves f, e.g. use P(A, B | C), not f_1(A, B, C).

What is the size of the probability table for this new factor (number of rows)? As before, assume all the Z variable are observed evidence.

New factor generated: $\boldsymbol{P}(\boldsymbol{z}_1, \boldsymbol{z}_2|\boldsymbol{Y}_2, \boldsymbol{X})$ Size of this new factor: 4

Explanation: We join together all of the CPTs that include $\boldsymbol{Y}_1$, and then sum over $\boldsymbol{Y}_1$. We can thus calculate this factor as $\boldsymbol{P}(\boldsymbol{z}_1, \boldsymbol{z}_2|\boldsymbol{Y}_2, \boldsymbol{X}) = \sum_{\boldsymbol{y}_1} \boldsymbol{P}(\boldsymbol{y}_1|\boldsymbol{X})\boldsymbol{P}(\boldsymbol{z}_1|\boldsymbol{y}_1)\boldsymbol{P}(\boldsymbol{z}_2|\boldsymbol{y}_1, \boldsymbol{Y}_2)$

A Space For Anything.

iii. (2 pts) Suppose we decide to perform variable elimination to calculate $\boldsymbol{P}(\boldsymbol{Y_n}|\boldsymbol{Z}_1 = \boldsymbol{z}_1, \boldsymbol{Z}_2 = \boldsymbol{z}_2, \cdots, \boldsymbol{Z_n} = \boldsymbol{z_n})$. Suppose that we eliminate the variables in the order Y_1, Y_2, … Y_{n-1}, X. What factor is generated when we finally eliminate X? ***Give your answer in standard probability distribution notation.*** Note Y_n is not eliminated!

New factor generated: $\boldsymbol{P}(\boldsymbol{Z}_1 = \boldsymbol{z}_1, ..., \boldsymbol{Z_n} = \boldsymbol{z_n}, \boldsymbol{Y_n})$

Solution 1: We can manually run variable elimination, noting that after eliminating $\boldsymbol{Y_i}$ we generate the factor $\boldsymbol{P}(\boldsymbol{Z}_1 = \boldsymbol{z}_1, ..., \boldsymbol{Z_{i+1}} = \boldsymbol{z_{i+1}}|\boldsymbol{Y_{i+1}}, \boldsymbol{X})$, such that when we eliminate $\boldsymbol{X}$, we join $\boldsymbol{P}(\boldsymbol{Z}_1 = \boldsymbol{z}_1, ..., \boldsymbol{Z_n} = \boldsymbol{z_n}|\boldsymbol{Y_n}, \boldsymbol{X})$, $\boldsymbol{P}(\boldsymbol{Y_n}|\boldsymbol{X})$, and $\boldsymbol{P}(\boldsymbol{X})$, and sum over $\boldsymbol{X}$ to get $\boldsymbol{P}(\boldsymbol{Z}_1 = \boldsymbol{z}_1, ..., \boldsymbol{Z_n} = \boldsymbol{z_n}, \boldsymbol{Y_n})$.

Solution 2: After we eliminate all of the variables, since the only unconditioned probability table is one for an eliminated variable, the resulting factor must just be the joint distribution over all remaining variables (with the evidence variables incorporating the evidence).

Common Mistakes:

- Note that $\boldsymbol{P}(\boldsymbol{Z}_1 = \boldsymbol{z}_1, ..., \boldsymbol{Z_n} = \boldsymbol{z_n}, \boldsymbol{Y_n})$ and $\boldsymbol{P}(\boldsymbol{Z}_1, ..., \boldsymbol{Z_n}, \boldsymbol{Y_n})$ are not the same term. The former has two rows (reflecting that $\boldsymbol{Y_n}$ is the only unspecified variable), while the latter has all 2^{n+1} rows of every combination of the variables' values. We ended up giving full credit for this answer only because we couldn't tell the difference between hand-written upper-case ("Z") and lower-case ("z") letters.
- If a variable is eliminated, then it naturally could not be a part of any resulting factor.
- The question specified to give an answer in standard probability distribution notation. Thus, $\boldsymbol{f}(\boldsymbol{Z}_1 = \boldsymbol{z}_1, ..., \boldsymbol{Z_n} = \boldsymbol{z_n}, \boldsymbol{Y_n})$ was incorrect.

iv. (4 pts) Find the best and worst variable elimination orderings for calculating $\boldsymbol{P}(\boldsymbol{Y_n}|\boldsymbol{Z}_1 = \boldsymbol{z}_1, \boldsymbol{Z}_2 = \boldsymbol{z}_2, \cdots, \boldsymbol{Z_n} = \boldsymbol{z_n})$. An ordering is considered better than another if the sum of the sizes of the factors that are generated is smaller. You do not need to calculate the precise value of this sum to answer this question. If there are orderings that are tied, give just one ordering.

Best ordering: $\boldsymbol{Y}_1, ..., \boldsymbol{Y_{n-1}}, \boldsymbol{X}$ or $\boldsymbol{Y}_1, ..., \boldsymbol{Y_{n-2}}, \boldsymbol{X}, \boldsymbol{Y_{n-1}}$

Worst ordering: $\boldsymbol{X}$, ($\boldsymbol{Y}_1, \boldsymbol{Y}_2, ..., \boldsymbol{Y_{n-1}}$ in any order)

Explanation: The best ordering is to eliminate first the $\boldsymbol{Y_i}$'s for $\boldsymbol{i} = 1, \cdots, \boldsymbol{n-2}$, and then to eliminate $\boldsymbol{X}$ and $\boldsymbol{Y_{n-2}}$ in either order. The sizes of the factors generated are 4 for each of the first $\boldsymbol{n-1}$ factors, and 2 for the last factor, for a total size of $4\boldsymbol{n}-2$. We can argue that this is the best by noting that any factor generated for this query (other than the final factor) must have a size of at least 4, so this is minimal.

The worst ordering is to eliminate $\boldsymbol{X}$ first, which creates the largest possible factor of size $2^{\boldsymbol{n}}$; afterwards, in each iteration after eliminating any $\boldsymbol{k}$ of the $\boldsymbol{Y}$s, we will generate a factor of size $2^{\boldsymbol{n-k}}$. Our total sum is thus $\sum_{\boldsymbol{k}=0}^{\boldsymbol{n-1}} 2^{\boldsymbol{n-k}} = \sum_{\boldsymbol{i}=1}^{\boldsymbol{n}} 2^{\boldsymbol{i}} = 2^{\boldsymbol{n}+1} - 2$.

Common Mistakes:

- The most common mistake was eliminating $\boldsymbol{Y_n}$, the query variable.
- Similarly, other common mistakes were to eliminate the wrong set of variables (i.e. eliminate any $\boldsymbol{Z_i}$, or to ***not*** eliminate $\boldsymbol{X}$ or any other $\boldsymbol{Y_i}$).
- For the best ordering, note that reversing the order of the $\boldsymbol{Y}$s is a worse ordering. Eliminating $\boldsymbol{Y_{n-1}}$ creates the factor $\boldsymbol{P}(\boldsymbol{Z_{n-1}} = \boldsymbol{z_{n-1}}, \boldsymbol{Z_n} = \boldsymbol{z_n}|\boldsymbol{X}, \boldsymbol{Y_{n-2}}, \boldsymbol{Y_n})$ which has size 8, and likewise for each successive factor generated.

v. (2 pts) Assume now we want to use variable elimination to calculate **a new query** $P(Z_n|Y_1, Y_2, \cdots, Y_{n-1})$. Mark all of the following variables that produce a ***constant*** factor after being eliminated for all possible elimination orderings. Note that for this problem, the set of evidence variables is different than in previous parts.

▣ Z_1 ▣ Z_2 ▣ Z_{n-1} ☐ X

Explanation: Eliminating the leaf nodes $Z_1, Z_2, \cdots, Z_{n-1}$ can produce a constant factor. The elimination of Z_i $(i = 2, \cdots, n-1)$ will generate the factor $\sum_{z_i} P(z_i|Y_{i-1}, Y_i) = 1$, and eliminating Z_1 will also generate a constant factor $\sum_{z_1} P(z_1|Y_1) = 1$.

vi. (1 pt) ● True ○ False: We can simply delete any factor that will produce a constant factor from our feature set before computing any distribution.

Explanation: Remember in the last step of variable elimination, we re-normalize the last factor to get the query distribution, so it won't affect the query result to delete all constant factors generated along the way of variable elimination.

6. The Homebody Ladybug (13 pts)

Consider a ladybug who lives on a number line where, at each step, the ladybug changes her position x by flying one step to the right (setting x = x + 1) or one step to the left (setting x = x − 1) with probabilities:

$$P_{left} = \frac{1}{2} + \frac{1}{2}\frac{x}{c+|x|}$$, where c is a constant greater than 0.

$$P_{right} = 1 - P_{left}$$

For example, if the constant c equals 1, the probabilities of a move to the left at positions $x = -2,-1,0,1,2$ are given by ⅛, ¼, ½, ¾, and ⅝, respectively.

i. (1.5 pts) Draw the Bayes' Net for the ladybug's position, using X_i to represent the ladybug's position at time step i. Do not assume the ladybug starts at the origin.

Solution: $(X_0) \rightarrow (X_1) \rightarrow (X_2) \rightarrow \ldots$

ii. (1.5 pts) What is the domain of the variable X_i? Do not assume the ladybug starts at the origin.
Solution: There are a variety of possible answers. One answer is the integers ($\{0, \pm 1, \pm 2, \ldots\}$), or reals, or complex numbers, et cetera, depending on what the starting position X_0 is.

iii. (3 pts) Assume that at time 0 the ladybug starts at the origin ($X_0 = 0$). What is $P(X_{10} = 5 \mid X_0 = 0)$ if c = 1?
Solution: 0. If the ladybug starts at the origin, then after an even number of time steps, the ladybug's location must be even.

iv. (2 pts) You'd like to estimate the probability that the ladybug is at the origin at time t=4, given that you observe that the ladybug is at +1 in timestep t = 1, i.e. $P(X_4 = 0 \mid X_1 = 1)$. Consider the following possible samples for $\{X_0, \ldots, X_4\}$. Put a check ***in the box to the left the samples that are valid for computing this query and could have come from the ladybug model***.

☒ áùÄ úÄ ûÄ üÄ ýâ ☐ áùÄ ÈúÄ ùÄ úÄ ùâ ☐ áùÄ ÈúÄ ùÄ ûÄ Èûâ ☒ áùÄ úÄ ùÄ ÈúÄ ùâ

☒ áùÄ úÄ ûÄ úÄ ùâ ☐ áùÄ ÈúÄ ÈûÄ ÈúÄ ùâ ☐ áùÄ ÈúÄ ÈûÄ ÈúÄ ùâ ☐ áùÄ úÄ úÄ ûÄ ùâ

v. (1 pt) From the valid samples in part iv, what is $P(X_4 = 0 \mid X_1 = 1)$?
Solution: 2/3, since out of the three valid samples, $X_4 = 0$ in two of them.

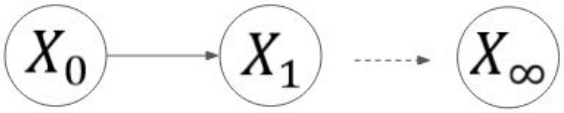
vi. (2 pts) How would you expect $P(X_4 = 0 \mid X_0 = 0)$ to change as c increases?
Solution: As c increases, the ladybug is less likely to move back to the origin (regardless of where it is), so this probability ***decreases***. (Note that this applies regardless of whether the ladybug is at a positive or negative location.)

vii. (2 pts) Suppose you are trying to use Gibbs Sampling to estimate $P(X_4 = 0 \mid X_0 = 0)$. If we start by initializing X_1, X_2, X_3, and X_4 to random 32 bit integers, will the Gibbs Sampling process ***yield a reasonable estimate***? Will the sampling procedure complete in a ***reasonable amount of time***? Explain your answer.
Solution: No, it will not yield a reasonable estimate. If the X_i s are assigned to values that are not compatible (with each other or with $X_0 = 0$), then it can be impossible to assign values to other X_i s. For example, if $X_2 = 3$ initially, then there is no valid distribution $P(X_1 | X_0 = 0, X_2 = 3)$ to sample from. We gave everybody 2 points on this because there was too much variety to grade fairly.

7. Ms. Pac-Man's Spicy Psychic Stake Out (18 pts)

The parts of this problem are independent. <u>You can do any of them without doing the rest.</u>

i) (4 pts) Ms. Pac-Man is training herself to understand the habits of ghosts. She has captured a ghost named Clyde, and forced it to walk down a hallway to train her model. She builds the model below, and calculates the conditional probability tables (CPTs) for this model, i.e. $P(X_0)$, $P(X_1 \mid X_0)$, $P(X_2 \mid X_1)$, etc. where X_i is the location of the ghost at time i.



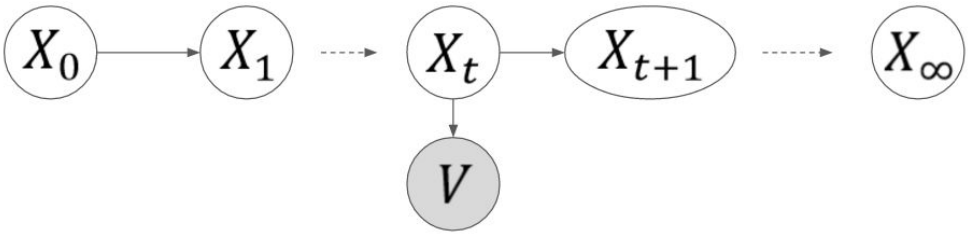
Give an expression for computing **$P(X_i)$** in terms of these CPTs.

Solution:

$$P(X_i) = \Sigma_{X_{i-1}} P(X_i \mid X_{i-1}) \ldots \Sigma_{X_1} P(X_2 \mid X_1) \Sigma_{X_0} P(X_1 \mid X_0) P(X_0)$$

This is repeated application of the forward algorithm, starting with the initial distribution at timestep 0.

ii) (4 pts) Ms. Pac-Man has been experimenting with SPICE to gain psychic abilities. When she takes SPICE, she sees a psychic vision v that provides her with a belief about the ghost's future position at some specified time t, shown below. In other words, she now knows v, and thus also knows the distribution $P(v \mid X_t)$.



Suppose Ms. Pac-Man wants to figure out where the ghost will go next in the timestep after t. Give an expression for computing **$P(X_{t+1} \mid v)$** in terms of distributions either given as part of her model, or calculated in part i. In total this includes the CPTs, $P(v \mid X_t)$, and $P(X_i)$. You may not need all of them.

Solution:

We came up with two basic approaches. In both, we start with an expression and try to manipulate it so that it is in terms of the given variables. The first approach starts from:

And the second starts from:

$$P(X_{t+1}|v) = \sum_{X_t} P(X_{t+1}, X_t|v)$$

Approach 2 is a harder 'start' to come up with, but is ultimately an easier approach.

Technique ii-1 [hard mode]:

Given the starting point shown below this line. The goal is to compute P(v) and $P(X_{t+1}, v)$ in terms of the CPTs and what we calculated in part i.

$$P(X_{t+1}|v) = \frac{\sum_{X_t} P(X_{t+1}, X_t, v)}{P(v)}$$ (by introducing X_t)

$$P(X_{t+1}|v) = \frac{\sum_{X_t} P(v|X_{t+1}, X_t)P(X_{t+1}, X_t)}{P(v)}$$ (by product rule)

$$P(X_{t+1}|v) = \frac{\sum_{X_t} P(v|X_t)P(X_{t+1}, X_t)}{P(v)}$$ (v is conditionally independent of X_{t+1} given X_t)

$$P(X_{t+1}|v) = \frac{\sum_{X_t} P(v|X_t)P(X_{t+1}|X_t)P(X_t)}{P(v)}$$ (by product rule)

$$P(X_{t+1}|v) = \frac{\sum_{X_t} P(v|X_t)P(X_{t+1}|X_t)P(X_t)}{\sum_{X_t} P(v|X_t)P(X_t)}$$ (by introducing X_t to denominator)

At this point, everything is in terms of either CPTs given in model or part i of the problem.

Technique ii-2 [easier mode]:

$$P(X_{t+1}|v) = \sum_{X_t} P(X_{t+1}, X_t|v)$$

$$P(X_{t+1}|v) = \sum_{X_t} P(X_{t+1}|X_t)P(X_t|v)$$

(by product rule, now just need $P(X_t \mid v)$)

$$P(X_{t+1}|v) = \frac{\sum_{X_t} P(X_{t+1}|X_t)P(v|X_t)P(X_t)}{P(v)}$$

(Bayes rule, now just need P(v))

$$P(X_{t+1}|v) = \frac{\sum_{X_t} P(X_{t+1}|X_t)P(v|X_t)P(X_t)}{\sum_{X_t} P(v|X_t)P(X_t)}$$

(introduce X_t to denominator)

At this point, everything is in terms of either CPTs given in model or part i of the problem. Note that our solution is exactly the same as in technique ii-1.

Alternate presentation of technique ii-2 (bottom up -- just in case it's easier to understand):
Note: In this problem, we will require use of the following conditional probability, $P(X_t \mid v)$, which we derive here:

$$P(X_t \mid v) = P(v \mid X_t) \cdot P(X_t)/P(v)$$

This is an application of Bayes's Rule. We are given from the previous sub-problem (and in the current problem statement) that $P(X_t)$ can be readily calculated. Now, we must calculate the denominator, P(v).

$$P(v) = \Sigma_{X_t} P(v, X_t)$$

In the above equation, we are marginalizing out X_t. This will give us P(v), which is what we want.

$$P(v) = \Sigma_{X_t} P(v \mid X_t) P(X_t)$$

Now, we have split up the joint using chain rule. Note that we have access to both of these probabilities. The first comes from the CPT for random variable V. The second can be calculated as in the previous sub-problem.

Putting it all together, we are left with the following:

$$P(X_t \mid v) = P(v \mid X_t) \cdot P(X_t)/(\Sigma_{X_t} P(v \mid X_t) P(X_t))$$

We will refer to $P(X_t|v)$ henceforth, since we have derived it.

Now, to derive $P(X_{t+1} \mid v)$.

$$P(X_{t+1} \mid v) = \Sigma_{X_t} P(X_{t+1}, X_t \mid v)$$

In the above step we marginalize out X_t. We can then apply the chain rule:

$$P(X_{t+1} \mid v) = \Sigma_{X_t} P(X_{t+1} \mid X_t, v) \cdot P(X_t \mid v)$$

And by applying conditional independence assumptions:

$$P(X_{t+1} \mid v) = \Sigma_{X_t} P(X_{t+1} \mid X_t) \cdot P(X_t \mid v)$$

Whew! We're done!

iii) (6 pts) Now suppose Ms. Pac-Man wants to figure out where the ghost was the timestep **before** t given her SPICE vision. Give an expression for computing $\mathbf{P(X_{t-1} \mid v)}$ in terms of distributions either given as part of her model, or calculated in part i or ii of Question 7. You may not need all of them.

Throughout this solution we assume that $P(X_t \mid v)$ and $P(v)$ are available, as they were required to solve the earlier parts of this problem.

Technique iii-1: Simple Bayes rule (tricky part is computing $P(v \mid X_{t-1})$)

$$P(X_{t-1} \mid v) = \frac{P(v|X_{t-1})P(X_{t-1})}{P(v)} \text{ by Bayes Rule}$$

$$P(X_{t-1} \mid v) = \frac{\sum_{X_t} P(v|X_t)P(X_t|X_{t-1})P(X_{t-1})}{P(v)} \text{ by introducing } X_t$$

Problem iv note: Reusing technique iii-1 to calculate $P(X_0 \mid v)$ requires that we compute $P(v \mid X_i)$ for all X_i somehow. It doesn't work for part iv without this extra step.

Technique iii-2: marginalize out X_t from conditional distribution (tricky part is computing $P(X_{t-1} \mid X_t)$)

$$P(X_{t-1} \mid v) = \sum_{X_t} P(X_t, X_{t-1} \mid v) \text{ by introducing } X_t$$

$$P(X_{t-1} \mid v) = \sum_{X_t} P(X_{t-1}|X_t)P(X_t|v) \text{ by product rule}$$

$$P(X_{t-1} \mid v) = \sum_{X_t} \frac{P(X_t|X_{t-1})P(X_{t-1})}{P(X_t)} P(X_t|v) \text{ by Bayes Rule}$$

This approach is equivalent to reversing the graph. (You may wonder: conceptually, why does it make sense to reverse the graph? After all, in a typical Bayes' net, the direction of the arrows do matter, and reversing the graph may break independence assumptions. In this case, however, the relevant part of the graph is actually a straight line: $\boldsymbol{X_0 \rightarrow X_1 \rightarrow \ldots \rightarrow X_t \rightarrow v}$. Reversing the edges of ***this*** particular Bayes' net leads to a new Bayes' net that encodes the exact same set of independence assumptions.)

Problem iv note: If you used this technique (and only this technique!) for problem iii, you can simply repeat the technique without modification for part iv.

Technique iii-3: marginalize out X_t from joint distribution (tricky part is computing $P(X_t, X_{t-1}, v)$)

$$P(X_{t-1} \mid v) = \frac{P(X_{t-1}, v)}{P(v)} \text{ by the definition of conditional probability}$$

$$P(X_{t-1} \mid v) = \frac{\sum_{X_t} P(X_t, X_{t-1}, v)}{P(v)} \text{ by introducing } X_t$$

$$P(X_{t-1} \mid v) = \frac{\Sigma_{X_t} P(v \mid X_t, X_{t-1}) \cdot P(X_t \mid X_{t-1}) \cdot P(X_{t-1})}{P(v)}$$ by the chain rule

$$P(X_{t-1} \mid v) = \frac{\Sigma_{X_t} P(v \mid X_t) \cdot P(X_t \mid X_{t-1}) \cdot P(X_{t-1})}{P(v)}$$ as $\boldsymbol{v}$ and $\boldsymbol{X_{t-1}}$ are independent conditioned on $\boldsymbol{X_t}$.

The only term here we aren't given is $\boldsymbol{P(v)}$, but which we can derive as in the previous subproblem, or which we can calculate by normalizing the numerator $\boldsymbol{P(X_{t-1}, v)}$.

Problem iv note: Reusing technique iii-3 to calculate $P(X_0 \mid v)$ requires that we compute $P(v \mid X_i)$ for all X_i somehow. It doesn't work for part iv without this extra step.

Common Mistakes:

- Keep in mind that the chain rule says that $\boldsymbol{P(A|B)P(B|C) = P(A, B|C) \neq P(A|C)}$. If you want $\boldsymbol{P(A|C)}$, you need to sum out $\boldsymbol{B}$.

iv) (2 pts) Suppose Ms. Pac-Man takes SPICE at time 0, which provides her a vision v corresponding to time t = 188, i.e. she knows v and $P(v \mid X_{188})$. Explain how she would calculate Clyde's ***most likely current position (i.e. at time 0)*** in terms of the distributions given in parts i, ii, and iii. You may not need all of them.

Technique iv-1,3: Apply either technique iii-1 or technique iii-3 repeatedly, interspersed with usages of Bayes Rule to calculate $P(v \mid X_i)$

In this approach, we can start by first obtaining $P(X_{187} \mid v)$ by using technique iii-1 or technique iii-3. We then try to work our way backwards by next calculating $P(X_{186} \mid v)$. The tricky thing is that these techniques require that we know $P(v \mid X_{187})$ to compute $P(X_{186} \mid v)$. We can do this, for example, by using Bayes rule, as shown below:

$P(v \mid X_{187}) = P(X_{187} \mid v) * P(v) / P(X_{187})$

By alternating between usages of technique iii-1 (or technique iii-3) and usages of Bayes Rule, we can eventually arrive back at $P(X_0 \mid v)$, at which point, we can get the most likely position of Clyde using:

$$\arg\max_{X0} P(X_0|v)$$

Minor note: We could also have started working our way backwards from $P(X_{188} \mid v)$ instead of $P(X_{187} \mid v)$.

Technique iv-2: Repeatedly apply technique iii-2

If you came up with technique iii-2 for part iii, you can simply repeat the process backwards until you get to $P(X_0 \mid v)$. In this approach, we can start by first obtaining $P(X_{188} \mid v)$ as in part ii..

We then "go backwards" by simply applying technique iii-2 one step.

$$P(X_{187} \mid v) = \Sigma_{X_{188}} P(X_{187} \mid X_{188}) \cdot P(X_{188} \mid v)$$

We showed how to calculate $P(X_{187} \mid X_{188})$ using Bayes Rule in the previous sub-problem.

Note that applying this step recursively lets us "step back" once more in time:

$$P(X_{186} \mid v) = \Sigma_{X_{187}} P(X_{186} \mid X_{187}) \cdot P(X_{187} \mid v)$$

In general, we can recursively apply this "backwards algorithm" all the way back to $P(X_0 \mid v)$. To find the most likely position, we simply pick the X_0 that maximizes the conditional distribution.

$$\arg\max_{X0} P(X_0|v)$$

Minor note: We could also have started working our way backwards from $P(X_{187} \mid v)$ instead of $P(X_{188} \mid v)$.

Technique iv-4: Just use normal inference

An alternate approach is to recognize that we are performing inference on a Bayes's net. Thus, we can just apply our usual techniques for inference. For example, we could use inference by joint enumeration to directly arrive at:

$$P(X_0|v) \propto P(X_0, v) = P(X_0) \sum_{X_1 \ldots X_{188}} \prod_{i=1}^{188} P(X_i|X_{i-1}) P(v|X_{188}))$$

Note that this is just technique iv-2, except with the recursion expanded out.

Alternatively, we could have taken an iterative variable elimination approach, either forward (elimination order $\boldsymbol{X}_1, ..., \boldsymbol{X}_{188}$, as a small adaptation of the forward algorithm) or backward (elimination order $\boldsymbol{X}_{188}, ..., \boldsymbol{X}_1$, which is different from the first solution in that it works with $\boldsymbol{P}(\boldsymbol{v}|\boldsymbol{X}_i)$ instead of $\boldsymbol{P}(\boldsymbol{X}_i|\boldsymbol{v})$).

Common Mistakes:

- Viterbi does not work for this problem, even if the formula was adapted to take into account the lack of regular evidence / not using a standard HMM model, because it finds the most probable ***path*** to the evidence, and not the most likely $\boldsymbol{X}_0$. (Finding this most likely path and then looking at the $\boldsymbol{X}_0$ it uses also doesn't work, because that might not be the most likely $\boldsymbol{X}_0$ across all paths.)
- Vague, hand-wavy descriptions of "working backward" or "do the previous parts repeatedly" were not given credit.
- **Very tricky point** about this problem: Saying "recursively apply the solution from part iii" is not necessarily correct. Part (iii) asked you to simply compute the distribution of X at the timestep before the vision, i.e. $\boldsymbol{P}(\boldsymbol{X}_{187}|\boldsymbol{v})$ and not at any other time. For example, if I asked "how do I get from California to Nevada" and you give me precise street-by-street instructions, and then I ask "now how would I get to Utah from there", then you can't just repeat the instructions. However, consider if you instead answered "how do I get from California to Nevada" by explaining how to use Google Maps and then said "use what I told you about Google Maps to go one state east". In that case, if I said "now how would I get to Utah from there", you would just say "use Google Maps like I told you to go one state east again". Technique iii-2 is the equivalent of explaining how to go one state east, and techniques iii-1 and iii-3 are the equivalent of giving precise driving directions for the state (since they both rely on P(v | Xt), or roughly equivalently P(roads_to_use | California)). It's not a perfect analogy, but I hope it helps elucidate this incredibly subtle and not terribly important mathematical detail.

v) (1 pt) ○ True ● False: If we assume our model is stationary (i.e. $P(X_n | X_{n-1}) = P(X_{n-1} | X_{n-2})$), then the limit without evidence as time goes to infinity must be a uniform distribution., i.e. $P(X_\infty)$ is uniform.

vi) (1 pt) ○ True ● False: If we assume our model is stationary (i.e. $P(X_n | X_{n-1}) = P(X_{n-1} | X_{n-2})$), then the limit with evidence as time goes to infinity must be a uniform distribution., i.e. $P(X_\infty | v)$ is uniform.

Explanation: While the limit without evidence should converge, there is no reason to expect it to converge to the uniform distribution.