Assignment 7: Written Exercises II

by the Staff of CSE 415

Due June 2, 2021 via GradeScope

This is an individual-work assignment.

Prepare your answers in a neat, easy-to-read PDF. Our grading rubric will be set up such that when a question is not easily readable or not correctly tagged or with pages repeated or out of order, then points will be be deducted. However, if all answers are clearly presented, in proper order, and tagged correctly when submitted to Gradescope, we will award a 5-point bonus.

If you choose to choose to typeset your answers in Latex using the template file for this document, please put your answers in blue while leaving the original text black.

Paul G. Allen School of Computer Science and Engineering, University of Washington

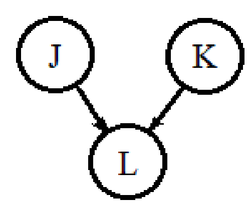
1 Joint Distributions and Factoring

(35 points) Consider the joint probability distribution below.

J	K	L	$P(J, K, L)$
true	true	true	0.024
true	true	false	0.096
true	false	true	0.028
true	false	false	0.252
false	true	true	0.144
false	true	false	0.036
false	false	true	0.294
false	false	false	0.126

(a) (15 points) For each of the three pairs of random variables (J, K), (J, L), and (K, L), provide a computation to prove that either (i) the two variables are independent, or (ii) they are dependent. If you provide any marginal or other derived distributions, make sure they are clearly identified, e.g., "$P(K|L)$:", etc.

(b) (15 points) Suppose that somebody has suggested that the joint distribution can be factored in a way that corresponds to the graph at the right. Prove either that (i) it cannot be factored according to this structure (i.e., some conditional independence assumption would be violated), or (ii) it can be factored according to this structure (i.e., provide the factorization with marginal distributions for J and K, and a CPT (conditional probability table) for $P(L|J, K)$.



(c) (3 points) Whether or not our joint distribution can actually be factored according to this two-parent, one-child structure, explain how many "free parameters" would be involved in the specification.

(d) (2 points) How many free parameters are (or would be) saved by using the factored representation of the joint distribution vs using free parameters of the full joint distribution table at the beginning of this problem?

2 Bayes Nets: D-Separation

(35 points) Consider the Bayes Net graph at the bottom of the page, which represents the topology of a web-server security model. Here the random variables have the following interpretations:

V = Vulnerability exists in web-server code or configs.

C = Complexity to access the server is high. (Passwords, 2-factor auth., etc.)

S = Server accessibility is high. (Firewall settings, and configs are permissive).

A = Attacker is active.

L = Logging infrastructure is state-of-the-art.

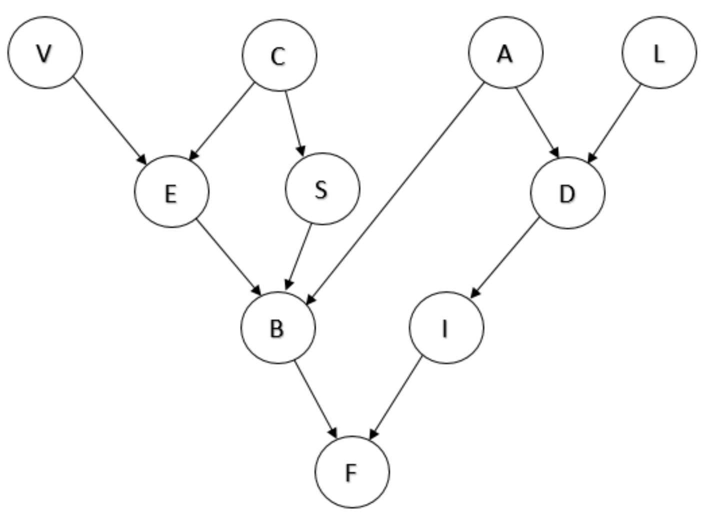
E = Exposure to vulnerability is high.

D = Detection of intrusion attempt.

B = Break-in; the web server is compromised.

I = Incident response is effective.

F = Financial losses are high (due to data loss, customer dissatisfaction, etc.).



For each of the following statements (4 points each), indicate whether (True) or not (False) the topology of the net guarantees that the statement is true. If False, identify an undirected path through which influence propagates between the two random variables being considered. (Be sure that the path follows the D-Separation rules covered in lecture.)

For example: $E \perp\!\!\!\perp S$: False (ECS).

(a) $L \perp\!\!\!\perp I \mid F$

(b) $I \perp\!\!\!\perp E \mid A, S$

(c) $L \perp\!\!\!\perp C \mid S, B, F$

(d) $F \perp\!\!\!\perp C \mid B, L, E$

(e) $L \perp\!\!\!\perp V \mid D, E, F, S$

(f) (15 points) Suppose that the company hired an outside expert to examine the system and she determines that B and E are true: The web server is compromised, and exposure to vulnerability is high. Given this information, your job is to explain to management why getting additional information about A (whether the attacker is active) could have an impact on the probability of V (regarding the existence or non-existence of vulnerabilities). Give your explanation, for the manager of the company, using about between 10 and 20 lines of text, which should be based on what you know about D-separation, applied to this situation. However, your explanation should not use the terminology of D-separation but be **in plain English**. (You can certainly use words like "influence", "probability", "given", but not "active path", "triple", or even "conditionally independent").

3 Markov Models

(30 points) According to an unnamed source, the stock market can be modeled using a Markov model, where there are two states "bull" and "bear." The dynamics of the model are given in the table below:

S_{t-1}	S_t	$P(S_t \mid S_{t-1})$
bull	bull	0.5
bull	bear	0.5
bear	bull	0.3
bear	bear	0.7

(a) (4 points) Draw a visual state-transition diagram to represent the conditional probability table of this Markov model. (Hint: there should be two nodes in your diagram).

(b) (6 points) Compute the stationary probabilities for *bull* and *bear*.

	bull	*bear*
P_∞		

(c) (15 points) Suppose it's given that $S_0 = bull$. Perform 5 rounds of the mini-forward algorithm and calculate $P(S_1), P(S_2), P(S_3), P(S_4), P(S_5)$ for each outcome of *bull* and *bear*. Furthermore, define the mean square error $\mathcal{E}_i$ between the distribution $P(S_i)$ and the stationary distribution P_∞ you've calculated in the previous part as

$$\mathcal{E}_i = \frac{1}{2}\sqrt{(P(S_i = bull) - P_\infty(bull))^2 + (P(S_i = bear) - P_\infty(bear))^2}.$$

Compute the $\mathcal{E}_i$ for the 5 steps. For the error terms, use scientific notation and keep 3 significant digits. Fill in the table below.

Note: You might find it helpful to use software such as Julia[1], NumPy or Matlab to compute these values for you or verify your hand-calculated results. If you have a linear algebra background, you might also find it helpful to express the calculation in terms of matrix vector multiplications.

[1]Our reference Julia solution for the previous part and this part combined takes 12 lines.

	bull	*bear*	$\mathcal{E}_i$
$P(S_1)$			
$P(S_2)$			
$P(S_3)$			
$P(S_4)$			
$P(S_5)$			

(d) (3 points) Plot $\log \mathcal{E}_k$ with k as the horizontal axis. Comment on the convergence behavior of $\mathcal{E}_i$ (i.e., how fast does it converge to 0?).

Comment: Google uses an algorithm based on this idea to compute the PageRank values for each webpage to inform its search ranking. However, as web search evolves, PageRank has become less and less important over time.

(e) (2 points) Suppose a startup uses this model to predict the state of the market in the future. Their analysis software can make a precise observation about the state of the market of previous day S_{t-1} as soon as markets open each day t. One night, a power surge destroyed the server containing all the historic market records. The next day, you are called in to assess the damage.

How has this event affected the accuracy of the market predictions of this startup?

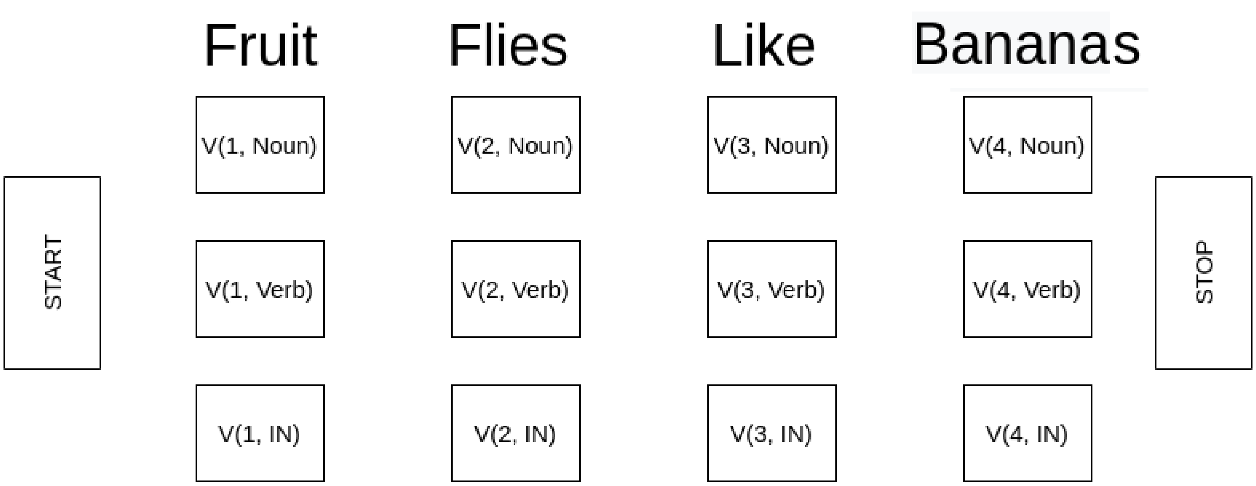
Give an explanation for how you arrived at your assessment.

4 The Viterbi Algorithm

(35 Points) One of the applications where the Viterbi algorithm can help us is in POS (Part-Of-Speech) tagging in Natural Language Processing. That is, given a sequence of words (e.g., a sentence), assign the correct grammatical tag (Noun, Verb, etc.) to each word based on its definition and context.

- Read more about POS-tagging here
- Read more about the Viterbi algorithm here
- Watch a video to help you understand how to approach this problem here

In this problem we perform POS tagging using the Viterbi algorithm for a small sentence and only considering Nouns (N), Prepositions or subordinating conjunctions (IN) and Verbs (V). While we do not use any other POS tags in this question, if you are curious you can find a list of them here. For the setup below we will be using the Viterbi algorithm to first determine the score for each POS tag for each word and then determine the most probable POS-tag sequence for the entire sentence based on the scores we have calculated.



Let Y_i be random variables representing the emissions (at position i) which are English words and X_i be random variables representing the POS tags. $P(y_i|x_i)$ represents the emission probability of word y_i given tag x_i and $P(x_i|x_{i-1})$ represents the probability of transitioning to tag x_i from tag x_{i-1}. Therefore, for computing score V_{i,x_i} for each state, we have:

$$V_{i,x_i} = \max_{x_1 \dots x_{i-1}} P(x_1 \dots x_i, y_1 \dots y_i) = \max_{x_{i-1}} P(y_i|x_i)P(x_i|x_{i-1})V_{i-1,x_{i-1}}$$

(a) (25 points) In this part we want to compute the maximum score of the tags for each word. Below we have provided the emission and transition probabilities. You will have to use them and the formula above to calculate the score for each state. Show all calculations. (You can use a separate page if you want, but be sure to include it in the Gradescope submission.) Please put your maximum score above the corresponding tag box (e.g., put a number above the $V(1,N)$ box). DO NOT round any numbers/final answers and show all calculations.

Transition model:

$$\begin{array}{rrrr}
P(N|\text{START}) = 0.6 & P(N|N) = 0.4 & P(N|V) = 0.5 & P(N|IN) = 0.7 \\
P(V|\text{START}) = 0.3 & P(V|N) = 0.3 & P(V|V) = 0.1 & P(V|IN) = 0.1 \\
P(IN|\text{START}) = 0.1 & P(IN|N) = 0.1 & P(IN|V) = 0.2 & P(IN|IN) = 0.1 \\
P(\text{STOP}|\text{START}) = 0.0 & P(\text{STOP}|N) = 0.2 & P(\text{STOP}|V) = 0.2 & P(\text{STOP}|IN) = 0.1
\end{array}$$

Emission model:

$$\begin{array}{rrr}
P(\text{fruit}|N) = 0.4 & P(\text{fruit}|V) = 0.1 & P(\text{fruit}|IN) = 0.0 \\
P(\text{flies}|N) = 0.2 & P(\text{flies}|V) = 0.4 & P(\text{flies}|IN) = 0.0 \\
P(\text{like}|N) = 0.1 & P(\text{like}|V) = 0.4 & P(\text{like}|IN) = 0.3 \\
P(\text{bananas}|N) = 0.1 & P(\text{bananas}|V) = 0.1 & P(\text{bananas}|IN) = 0.7
\end{array}$$

(b) (10 points) Now that we have computed the score, determine the maximum-probability sequence of states by working backwards from the STOP state. Consider the STOP state as the first value of a variable CURRENT-STATE. At each stage, move to the left, to the state that was selected during the arg-maxing for CURRENT-STATE. Then make that state be CURRENT-STATE, and iterate until all the way left at START. Reading the sequence along this path from left to right, you'll have the maximum-probability sequence of states that could give rise to the sentence "Fruit flies like bananas." Clearly draw the path, and also write down the final maximum-probability tag sequence.

5 Disambiguating Syntax with PCFGs

(35 points) Consider the sentence "Mary analyzed the algorithm with an equation." This might mean that, for example, Mary used an equation to analyze the equation, or rather differently it might mean that Mary analyzed the algorithm, which happened to contain an equation.

A grammar relevant to this example is given below. Consider the number at the right of a production to be the conditional probability of applying that production given that the symbol to be expanded, during a derivation, is the symbol on the left-hand side of the production.

In this problem, you'll convert probabilities of productions into scores. Then, with the given probabilistic context-free grammar, you will find two legal parses for the sentence, and compute a score for each parse. You'll then convert the overall parse scores back to probabilities of each parse. Then you'll identify the more probable parse using the parse probabilities.

```
S   ::=  NP   VP         1.000
NP  ::=  NN              0.250
NP  ::=  NP   PP         0.250
NP  ::=  NNP             0.125
NP  ::=  DT   NN         0.250
NN  ::=  NNP             0.125
VP  ::=  VBD  NP  PP     0.250
VP  ::=  VBD  NP         0.125
PP  ::=  IN   NP         0.250
```

(a) (8 points) Scores for each production rule: Convert each probability into a score by taking score $= -\log_2(p)$. Write the scores next to the probabilities above.

(b) (6 points) Find parse number 1. The parse will assume that the terminal symbols have been converted to non-terminals as shown. Make this parse correspond to the interpretation that Mary used an equation.

```
NNP      VBD      DT       NN      IN  DT    NN
Mary   analyzed  the algorithm with an equation.
```

(c) (6 points) Parse number 2. Make this parse according to an interpretation in which the algorithm contains an equation.

```
NNP      VBD      DT      NN       IN  DT    NN
Mary   analyzed  the algorithm with an equation.
```

(d) (6 points) Total score ___ and overall probability ___ for parse number 1.

(e) (6 points) Total score___ and overall probability ___ for parse number 2.

(f) (3 point) Which parse is more probable? _____ ___

You are welcome but not required to use the online parser at https://parser.kitaev.io/ in this problem. Note that it doesn't return all parses of a sentence.

6 The Laws of Robotics and Ethics in AI

(30 points) Read the short story *Little Lost Robot* by Isaac Asimov. Answer the questions listed below the links (5 points each).

Short Story: https://canvas.uw.edu/courses/1448875/files?preview=77766644

(a) List three facts about Isaac Asimov, including one that you think would be of interest to computer scientists.

(b) What are Asimov's 3 Laws of Robotics?

(c) Is the ordering of the laws important? If so, explain why (e.g. give an example of what might happen if the ordering of the laws changed). If not, explain why not. (You may find referring to this resource helpful: https://xkcd.com/1613/))

(d) Can you describe an example of when a robot didn't obey Asimov's Laws? Do you think the example you described is consistent with the reported modification to the laws?

(e) Did you find the end of the story satisfying – i.e. did it seem reasonable that the robot could be tricked to reveal that it could differentiate between different kinds of radiation? Explain.

(f) In the story, Bogert and Calvin disagree on how potentially dangerous the modified robots are. Do you agree with Calvin's opinion that the robots are dangerous enough that all 63 robots should be destroyed if the lost robot can't be found? Explain.