CS221 Exam

CS221
November 17, 2015

Name: _____
by writing my name I agree to abide by the honor code

SUNet ID: _____

Read all of the following information before starting the exam:

- This test has 3 problems and is worth 150 points total. It is your responsibility to make sure that you have all of the pages.
- Keep your answers precise and concise. Show all work, clearly and in order, or else points will be deducted, even if your final answer is correct.
- Don't spend too much time on one problem. Read through all the problems carefully and do the easy ones first. Try to understand the problems intuitively; it really helps to draw a picture.
- Good luck!

Problem	Part	Max Score	Score
1	a	10	
	b	10	
	c	10	
	d	10	
	e	10	
2	a	10	
	b	10	
	c	10	
	d	10	
	e	10	
3	a	10	
	b	10	
	c	10	
	d	10	
	e	10	

Total Score: ☐ + ☐ + ☐ = ☐

1. Two views (*50 points*)

Alice and Bob are trying to predict movie ratings. They cast the problem as a standard regression problem,[1] where $x \in \mathcal{X}$ contains information about the movie and $y \in \mathbb{R}$ is the real-valued movie rating.

To split up the effort, Alice will create a feature extractor $A : \mathcal{X} \to \mathbb{R}^d$ using information from the text of the movie reviews and Bob will create a feature extractor $B : \mathcal{X} \to \mathbb{R}^d$ from the movie metadata (genre, cast, number of reviews, etc.) These features will be used to do linear regression as follows: If $\mathbf{u} \in \mathbb{R}^d$ is a weight vector associated with A and $\mathbf{v} \in \mathbb{R}^d$ is a weight vector associated with B, then we can define the resulting predictor as:

$$f_{\mathbf{u},\mathbf{v}}(x) \stackrel{\text{def}}{=} \mathbf{u} \cdot A(x) + \mathbf{v} \cdot B(x). \tag{1}$$

We are interested in the squared loss:

$$\text{Loss}(x, y, \mathbf{u}, \mathbf{v}) \stackrel{\text{def}}{=} (f_{\mathbf{u},\mathbf{v}}(x) - y)^2. \tag{2}$$

Let $\mathcal{D}_{\text{train}}$ be the training set of (x, y) pairs, on which we can define the training loss:

$$\text{TrainLoss}(\mathbf{u}, \mathbf{v}) \stackrel{\text{def}}{=} \sum_{(x,y) \in \mathcal{D}_{\text{train}}} \text{Loss}(x, y, \mathbf{u}, \mathbf{v}). \tag{3}$$

[1] Remember that you should be going beyond this for your project!

a. (*10 points*) Compute the gradient of the training loss with respect to Alice's weight vector $\mathbf{u}$.

$\nabla_{\mathbf{u}} \text{TrainLoss}(\mathbf{u}, \mathbf{v}) =$ _____

Suppose Alice and Bob have found a current pair of weight vectors $(\mathbf{u}, \mathbf{v})$ where the gradient with respect to $\mathbf{u}$ is zero: $\nabla_{\mathbf{u}} \text{TrainLoss}(\mathbf{u}, \mathbf{v}) = 0$. Mark each of the following statements as true or false (no justification is required).

1. Even if Bob updates his weight vector $\mathbf{v}$ to $\mathbf{v}'$, $\mathbf{u}$ is still optimal for the new $\mathbf{v}'$; that is, $\nabla_{\mathbf{u}} \text{TrainLoss}(\mathbf{u}, \mathbf{v}') = 0$ for any $\mathbf{v}'$.
   – True / False

2. The gradient of the loss on each example is also zero: $\nabla_{\mathbf{u}} \text{Loss}(x, y, \mathbf{u}, \mathbf{v}) = 0$ for each $x, y \in \mathcal{D}_{\text{train}}$.
   – True / False

b. (*10 points*)

Alice and Bob got into an argument and now aren't on speaking terms. Each of them therefore trains his/her weight vector separately: Alice computes

$$\mathbf{u}_A = \arg\min_{\mathbf{u}\in\mathbb{R}^d} \sum_{(x,y)\in\mathcal{D}_{\text{train}}} (\mathbf{u}\cdot A(x) - y)^2$$

and Bob computes

$$\mathbf{v}_B = \arg\min_{\mathbf{v}\in\mathbb{R}^d} \sum_{(x,y)\in\mathcal{D}_{\text{train}}} (\mathbf{v}\cdot B(x) - y)^2.$$

Just for reference, if they had worked together, then they would have gotten:

$$(\mathbf{u}_C, \mathbf{v}_C) = \arg\min_{\mathbf{u},\mathbf{v}} \text{TrainLoss}(\mathbf{u}, \mathbf{v}).$$

Let L_A, L_B, L_C be the training losses of the following predictors:

- (L_A) Alice works alone: $x \mapsto \mathbf{u}_A \cdot A(x)$
- (L_B) Bob works alone: $x \mapsto \mathbf{v}_B \cdot B(x)$
- (L_C) They work together from the beginning: $x \mapsto \mathbf{u}_C \cdot A(x) + \mathbf{v}_C \cdot B(x)$

For each statement, mark it as necessarily true, necessarily false, or neither:

1. $L_A \le L_B$ — True / False / Neither
2. $L_C \le L_A$ — True / False / Neither
3. $L_C \le L_B$ — True / False / Neither

Now, let L'_A, L'_B, L'_C be the corresponding losses on the *test set*. For each statement, mark it as necessarily true, necessarily false, or neither:

1. $L'_A \le L'_B$ — True / False / Neither
2. $L'_C \le L'_A$ — True / False / Neither
3. $L'_C \le L'_B$ — True / False / Neither

c. (*10 points*)
Alice and Bob have reconciled, and together they collected a small training set:[2]

	Review text	# reviews	Genre	Rating
Example 1	great movie	100	action	5
Example 2	great quality	10	action	4
Example 3	poor quality	10	action	2
Example 4	poor movie	1	action	1

Alice and Bob wrote down the following features:

- (Alice) review text contains "great"
- (Alice) review text contains "poor"
- (Alice) review text contains "quality"
- (Alice) review text contains "movie"
- (Bob) # reviews
- (Bob) $\log_{10}$(# reviews)
- (Bob) genre is "comedy"
- (Bob) 1 (bias feature)

For example, if you set all feature weights to 1, then the prediction on example 1 is $1 + 0 + 0 + 1 + 100 + \log_{10}(100) + 0 + 1 = 105$, and the squared loss on that example would be $(105 - 5)^2 = 10000$ (not so good!). For the following three questions, try to use your intuition about linear models; don't brute force all possibilities.

[2]For your final project, you hopefully have more data than they do!

(i) Alice decides to choose exactly *one* of her features in a linear model. Which feature should she choose and what is the associated feature weight in order to obtain the smallest total squared loss across all training examples? You don't need to give the value of the squared loss.

Alice's Feature: _____

Feature weight: _____

(ii) Similarly, Bob decides to choose exactly *one* of his features. Which feature should he choose and what is the associated feature weight in order to obtain the smallest total squared loss across all training examples? You don't need to give the value of the squared loss.

Bob's Feature: _____

Feature weight: _____

(iii) What is the smallest set of features from both Alice and Bob's set that would allow us to get 0 prediction error? What are the weights of the chosen features?

Feature set: _____

Feature weights: _____

d. (*10 points*)

Alice and Bob were so excited about their dataset and feature extractors, that they immediately started running feature extraction. For each original training example (x_i, y_i) $(i = 1, \ldots, n)$, they computed Alice's feature vector $A(x_i)$ and Bob's feature vector $B(x_i)$. But they had a bug in their code that caused them to switch the feature vectors once in a while. Specifically, their code overwrote the original example (x_i, y_i) with $(A(x_i), B(x_i), y_i)$ some of the time and $(B(x_i), A(x_i), y_i)$ some of the time. Whoops![3]

They think for a bit and develop an ingenious scheme to take into account this shuffling. The intuition is that they will try treating each example (c_i, d_i, y_i) as either $(A(x_i), B(x_i), y_i)$ or $(B(x_i), A(x_i), y_i)$ depending on which one yields lower loss. Here's the objective function they wrote down:

$$\text{ShuffledTrainLoss}(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^{n} \min\{(\mathbf{u} \cdot c_i + \mathbf{v} \cdot d_i - y_i)^2, (\mathbf{u} \cdot d_i + \mathbf{v} \cdot c_i - y_i)^2\}. \quad (4)$$

Help Alice and Bob derive an alternating minimization algorithm. Let $z_i \in \{0, 1\}$ denote whether the feature vectors are swapped or not for each $i = 1, \ldots, n$. Specifically, $z_i = 0$ corresponds to the case where $c_i = A(x_i)$ and $d_i = B(x_i)$; $z_i = 1$ corresponds to the case where $d_i = A(x_i)$ and $c_i = B(x_i)$.

(i) Write the closed form update for each z_i given $\mathbf{u}, \mathbf{v}$ (don't worry about ties):

(ii) Write the optimization problem for $(\mathbf{u}, \mathbf{v})$ given $z_1, \ldots, z_n$:

[3] For your final project, always save and backup your raw data.

e. (*10 points*)

Alice and Bob realize that it wasn't a bug at all, but it was Competitive Carol from another group up to no good! Here's what Carol was doing: for each example $i = 1, \ldots, n$, she would adversarily choose to swap the features or not so that Alice and Bob would get the worst (highest) loss. Worse, Carol had found out a way to do the same adversarial swapping at test time too.

After suffering ten minutes of standing with their mouths agape in horror, Alice and Bob eventually wrote down the following objective function to model Carol's machinations:

$$\text{AdversarialTrainLoss}(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^{n} \max\{(\mathbf{u} \cdot c_i + \mathbf{v} \cdot d_i - y)^2, (\mathbf{u} \cdot d_i + \mathbf{v} \cdot c_i - y)^2\}. \tag{5}$$

Compute the gradient $\nabla_{\mathbf{u}}\text{AdversarialTrainLoss}(\mathbf{u}, \mathbf{v})$ with respect to Alice's weight vector $\mathbf{u}$ (Bob's would be similar, so you don't need to do it). For your convenience, to keep your equations compact, define:

$$r_i = \mathbf{u} \cdot c_i + \mathbf{v} \cdot d_i - y \tag{6}$$

$$s_i = \mathbf{u} \cdot d_i + \mathbf{v} \cdot c_i - y. \tag{7}$$

$\nabla_{\mathbf{u}}\text{AdversarialTrainLoss}(\mathbf{u}, \mathbf{v}) =$

2. Rafting (*50 points*)

You are going on a rafting trip! The river is modeled as a grid of positions (x, y), where $x \in \{-m, -(m-1), \ldots, 0, \ldots, (m-1), m\}$ represents the horizontal offset from the middle of the river and $y \in \{0, \ldots, n\}$ is how far down the river you are. To make things more challenging, there are a number of rocks in the river: For each position (x, y), let $R(x, y) = 1$ if there is a rock and 0 otherwise. You can assume that the start and end positions do not have rocks.

Here's how you can control the raft. From any position (x, y), you can:

- go straight down to $(x, y+1)$ (which takes 1 second),
- veer left to $(x-1, y+1)$ (which takes 2 seconds), or
- veer right to $(x+1, y+1)$ (which takes 2 seconds).

If the raft enters a position with a rock, there is an extra 5 second delay. The raft starts in $(0, 0)$ and you want to end up in any position (x, y) where $y = n$.

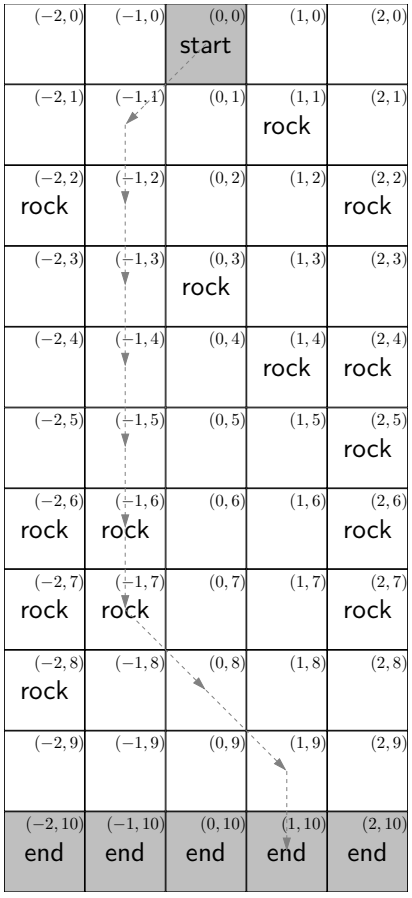


Figure 1: The goal is to raft down the river from 'start' to one of the 'end' positions in the least amount of time. For this example, $m = 2, n = 10$.

a. (*10 points*)

Let $C(x, y)$ denote the minimum time that it takes to get from position (x, y) to an end position. For convenience, let $C(x, y) = \infty$ for invalid positions (x, y) (those that are off the grid).

Write down the dynamic programming recurrence (you should have a base case and a recursive case):

$C(x, y) =$ _____

b. (*10 points*)

During the rafting trip, you are planning to raft down not just one $m \times n$ section of the river, but K of them. You are given K maps $R_1, \ldots, R_K$, where each $R_j(x, y) = 1$ if there is a rock in position (x, y) in the j-th map. Assume that each map has at most p rocks (where p is much smalller than mn). You could find the minimum cost path for each of the K maps separately, but that's a lot of work. Upon inspection, you find that the maps are quite similar, so you suspect there's a way to save some work.

Indeed there is! Precisely define a *single* heuristic $h(x, y)$ that can be used to do A* search on each of the K maps. Your heuristic must be

- consistent in the search problem defined by *each* of the K maps,
- non-trivial, which means you must use the information from the rocks somehow (not just define something like the Manhattan distance), and
- computable in $O(kp + mn)$ time.

Write one sentence justifying why your heuristic is consistent, and write one sentence explaining the running time for computing the heuristic for each position (x, y).

c. (*10 points*)

You find out that the maps are actually all wrong (and you can't get your money back either!), so you're going to have to start rafting without full knowledge about the location of the rocks. Fortunately, you have eyes, so at position (x, y) you can see whether there's a rock at $(x', y+1)$ for all x'. Furthermore, you assume that each position has an independent probability α of having a rock. To simplify, assume that there are no rocks in the $y = 0$ and $y = 1$ rows. You want to minimize the *expected* time to reach an end goal.

Show that you can solve this problem by defining an MDP whose maximum expected utility policy minimizes the expected the time to reach an end goal. Reduce the number of states as much as you can while still maintaining optimality; for full credit, you should be able to store only $O(mn)$ states. Define precisely what your states, transition probabilities are, etc., but don't worry too much about the corner cases. (Hint: think carefully about what you actually need to remember in the state, and explicitly define the state tuple $s =$ _____.)

- $s_{\text{start}} =$
- $\text{Actions}(s) = \{a \in \{-1, 0, +1\} : \text{ moving horizontally in direction } a \text{ keeps you on the grid}\}$
- $T(s, a, s') =$
- $\text{Reward}(s, a, s') =$
- $\text{IsEnd}(s) =$
- $\gamma =$

d. (*10 points*)

As before, assume that you don't have the map describing the position of the rocks, but you know that the probability distribution over the map is $R(x, y) = 1$ independently with probability α.

Consider the following two scenarios:

- A genie reveals the entire map to you right as you get on your raft and at that instant, you run your blazing fast code to find the minimum time path (using part (a)). Let T_1 be this expected minimum time of getting to an end goal.
- There is sadly no genie, and you have to use your own eyes to look at the position of the next row as you're rafting. But you solve the MDP from part (c) to find the best policy. Let T_2 be this minimum expected time of getting to an end goal.

Prove that $T_1 \leq T_2$. (Intuitively this should be true; you must argue it mathematically.)

e. (*10 points*)

Suppose we don't actually know how long it takes to go over rocks or what the distribution of rocks is, so we will use Q-learning to learn a good policy.

- Suppose the state s includes the position (x, y) and the map that's revealed so far, i.e. $R(x', y')$ for all x' and $y' \leq y + 1$.
- The actions are $a \in \{-1, 0, +1\}$, corresponding to going left, straight, or right.
- The reward is the negative time it takes to travel from state s to the new state s'.
- Assume the discount $\gamma = 1$.

For each state s and action a, let $H(s, a) = 1$ if a causes the raft to hit a rock and 0 otherwise. Now define the approximate Q-function to be:

$$Q(s, a; \alpha, \beta) = \alpha H(s, a) + \beta,$$

where α and β are parameters to be learned. Suppose we sample once from an exploration policy, which led to the trajectory shown in Figure **??**.

(i) Write down the Q-learning updates on α on experience (s, a, r, s') using a step size η. Your formula should be in terms of $H(s, a), \alpha, \beta$ and should not contain generic RL notation.

$\alpha \leftarrow \alpha - \eta$_____

$\beta \leftarrow \beta - \eta$_____

(ii) On how many of the $n = 10$ updates could α change its value? _____

3. Voting (*50 points*)

You decide to visit Bayesland for Thanksgiving. It happens to be prime voting season there. Having just learned about Bayesian networks, you see that the voting process (on any given issue) can be modeled by following Bayesian network (see Figure **??**):

1. The head of state makes a statement either in favor of the issue ($H = 1$) with probability $\frac{1}{2}$ or against the issue ($H = 0$) with probability $\frac{1}{2}$.

2. Each of the n citizens casts an independent vote $C_i \in \{0, 1\}$ given the head's statement, deviating from H with probability α; that is: $p(c_i \mid h) = \alpha$ if $c_i \neq h$ and $1 - \alpha$ if $c_i = h$.

3. Because of the backwards technology in Bayesland, each vote C_i gets flipped with probability β, and this noisy version D_i is sent to the central agency.

4. Finally, the central agency reports $R = [\sum_{i=1}^{n} D_i > n/2]$ as the final result, which is whether over half the of votes received were in favor of the issue.

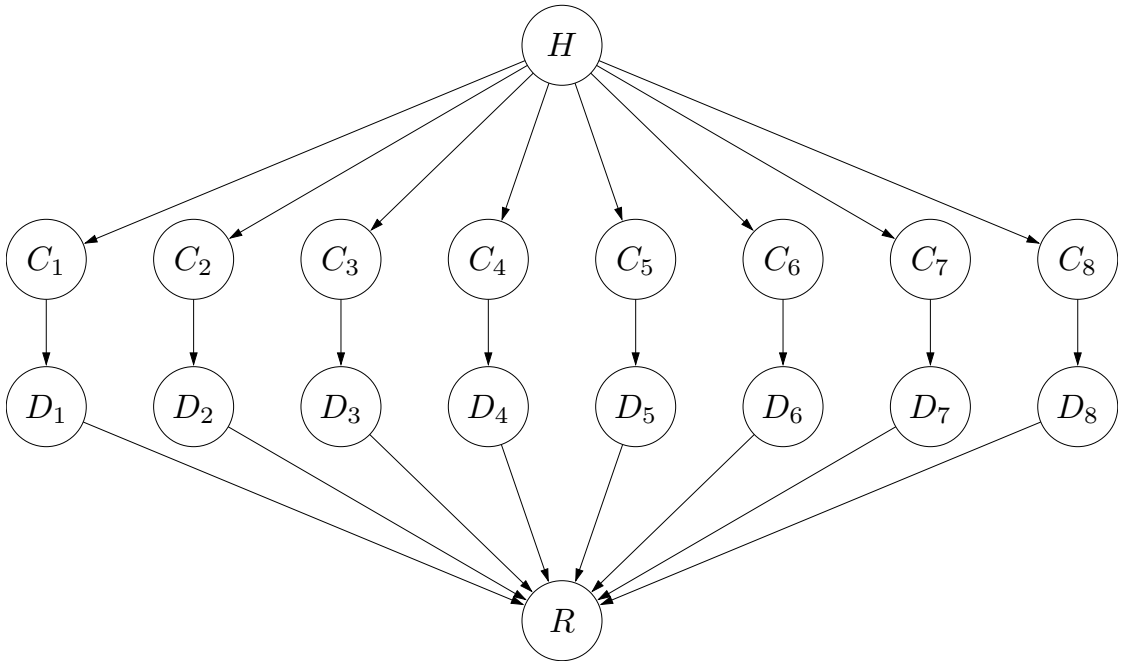


Figure 2: Bayesian network 4 25 corresponding to the voting process in Bayesland.

a. (*10 points*)

You suspect that the voting process is actually much simpler than it appears if you could only expose some of the conditional independence assumptions. Mark each of the following statements as either true or false.

i	$C_2 \perp\!\!\!\perp C_4$	True / False
ii	$C_2 \perp\!\!\!\perp C_4 \mid H$	True / False
iii	$C_2 \perp\!\!\!\perp C_4 \mid R$	True / False
iv	$C_2 \perp\!\!\!\perp C_4 \mid H, R$	True / False
v	$C_2 \perp\!\!\!\perp C_4 \mid D_2, D_4$	True / False
vi	$C_2 \perp\!\!\!\perp C_4 \mid D_2, D_4, R$	True / False
vii	$C_2 \perp\!\!\!\perp C_4 \mid H, D_2$	True / False
viii	$C_2 \perp\!\!\!\perp C_4 \mid H, D_2, R$	True / False
ix	$C_2 \perp\!\!\!\perp C_4 \mid H, D_2, D_4$	True / False
x	$C_2 \perp\!\!\!\perp C_4 \mid H, D_2, D_4, R$	True / False

b. (*10 points*)

After having discovered the qualitative structure of the voting process, you are now interested in digging a bit deeper into how various people influence each other. Compute the following (write each expression in terms of α and β). Hint: for each query, try to marginalize away all the irrelevant quantities before you start writing down formulas. If things get hairy, you might be doing something wrong!

i What did the head say given the citizens' votes?

$\mathbb{P}(H = 1 \mid C_1 = 1, C_2 = 1, C_3 = 0) =$

---

ii How does knowing one citizen's vote influence our belief of another citizen's vote?

$\mathbb{P}(C_5 = 1 \mid C_8 = 1) =$

---

iii You're getting tired of computing all these queries by hand, so you decide to use Gibbs sampling. Describe the Gibbs sampling update for D_i as a function of β. You must simplify as much as possible and explain what the resulting conditional distribution of D_i is.

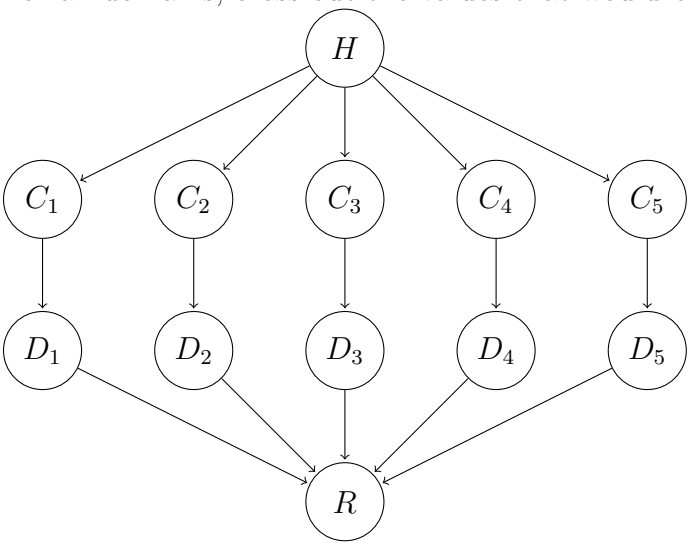
c. (*10 points*)

You're still trying to make your Gibbs sampling go fast. You remember that a Bayesian network is an instance of a factor graph, so you decide to apply AC-3.

Specifically, suppose there are $n = 5$ citizens and $\beta = 0$ (no noise in transferring votes). Consider the factor graph corresponding to

$$\mathbb{P}(H, C_1, C_2, C_3, C_4, C_5, D_3, D_4, D_5 \mid D_1 = 0, D_2 = 0, R = 1).$$

For all domains, cross out the values that would be pruned by running AC-3:



1. $H \in \{0, 1\}$
2. $C_1 \in \{0, 1\}$
3. $C_2 \in \{0, 1\}$
4. $C_3 \in \{0, 1\}$
5. $C_4 \in \{0, 1\}$
6. $C_5 \in \{0, 1\}$
7. $D_3 \in \{0, 1\}$
8. $D_4 \in \{0, 1\}$
9. $D_5 \in \{0, 1\}$

d. (*10 points*)

Again assume $\beta = 0$ and suppose you're interested in how citizens voted given that the overall outcome was positive:

$$\mathbb{P}(C_1, \ldots, C_n \mid R = 1).$$

Suppose you want to use particle filtering for this task. In class, we defined particle filtering for an HMM with transitions and emissions, so we'll need to generalize slightly.

(i) Define a set of new variables $E_1, \ldots, E_n$, where $E_i = 1$ if it is possible for $R = 1$ based on $C_1, \ldots, C_i$ and $E_i = 0$ otherwise. The upshot is that conditioning on $E_1, \ldots, E_n$ is the same as conditioning on $R = 1$, but the evidence can be partially evaluated without seeing all of $C_1, \ldots, C_n$. Define the local conditional distribution formally:

$p(e_i \mid c_1, \ldots, c_i) =$ _____

(ii) In an HMM, we proposed extensions to a particle $(c_1, \ldots, c_{i-1})$ by sampling C_i from $p(c_i \mid c_{i-1})$. For an non-HMM, we need to sample C_i conditioned on the entire history. Compute this quantity (Your answer does not have to be in terms of α and β, but it must be in terms of known probabilities): $\mathbb{P}(C_i = c_i \mid C_1 = c_1, \ldots, C_{i-1} = c_{i-1}) \propto$

_____

e. (*10 points*)

Having done all these calculations, you realize that you don't actually know what α and β actually are! So you decide to record all the communications in Bayesland in order to estimate these values.

Suppose we have $n = 5$ citizens, and they voted on two issues. For the first issue, the head of state chose $H = 1$, and the following results were received (and the result was $R = 1$):

i	1	2	3	4	5
C_i	1	0	1	1	0
D_i	0	1	1	1	0

For the second issue, the head of state chose $H = 0$, and the following results were received (and the result was $R = 0$):

i	1	2	3	4	5
C_i	0	0	0	1	1
D_i	0	0	0	1	1

Taking all this data into account, what is the maximum likelihood estimate of α and β with add λ smoothing / laplace smoothing? Your answer should be in terms of λ.

$\alpha =$ _____

$\beta =$ _____