Homework 4 - Part 1

Instructions on homework submission. This assignment has a slightly different structure from previous homeworks. There are two separate submissions: First part is Hidden Markov model related questions(this pdf). **You need to submit first part on April 17th.** You are allowed to use only one late day if you need to. This is because the final exam is approaching, and we want to release a solution as soon as possible. The Second part is a programming assignment. This time, however, is quite coding intensive, if you are not familiar with Pytorch. Please check the recitation. To make you familiar with Pytorch, **programming part is due April 24th.** However, **you should start this programming part as early as possible.** For the first part, submit a pdf file. For the second part, submit a .ipynb file.

1. [Hidden Markov Model, 40pts] Consider the following hidden Markov model for annotating the chromatin states given the DNA methylation data. Assume the methylation data have been binarized into either methylated or non-methylated at each locus. The DNA methylation data are given as a sequence of 1's for methylated DNA positions and 0's for non-methylated DNA positions. Assume four chromatin states: promoters (P), enhancers (E), transcribed regions (R), and background (B).

$$\left.\begin{array}{ll} \text{Methylated} & 1 \\ \text{Non- methylated} & 0 \end{array}\right\} \text{Observed}$$

$$\boxed{\begin{array}{lcccc} & \text{P} & \text{E} & \text{R} & \text{B} \\ \pi = & [0.2 & 0.3 & 0.3 & 0.2] \end{array}} \quad \text{Hidden States}$$

$$\mathbf{T} = \begin{bmatrix} & \text{P} & \text{E} & \text{R} & \text{B} \\ P(q_t|q_{t-1} = \text{P}) & 0.3 & 0.1 & 0.3 & 0.3 \\ P(q_t|q_{t-1} = \text{E}) & 0.1 & 0.4 & 0.1 & 0.4 \\ P(q_t|q_{t-1} = \text{R}) & 0.1 & 0.1 & 0.4 & 0.4 \\ P(q_t|q_{t-1} = \text{B}) & 0.1 & 0.1 & 0.1 & 0.7 \end{bmatrix}$$

$$\mathbf{E} = \begin{bmatrix} & \overset{0}{\text{non methylated}} & \overset{1}{\text{methylated}} \\ P(o_t|q_t = \text{P}) & 0.9 & 0.1 \\ P(o_t|q_t = \text{E}) & 0.8 & 0.2 \\ P(o_t|q_t = \text{R}) & 0.9 & 0.1 \\ P(o_t|q_t = \text{B}) & 0.2 & 0.8 \end{bmatrix}$$

Answer the following questions on HMMs.

(a) (5 pts) See the HMM in Figure 1 and answer the questions about conditional independencies.

i. Is q_T conditionally independent of q_3 given o_{T-1}?

No

$P(O_1, \ldots, O_T, q_1, \ldots, q_T) = P(O_T | q_T) P(q_T | q_{T-1}) P(O_1, \ldots, O_{T-1}, q_1, \ldots, q_{T-1})$

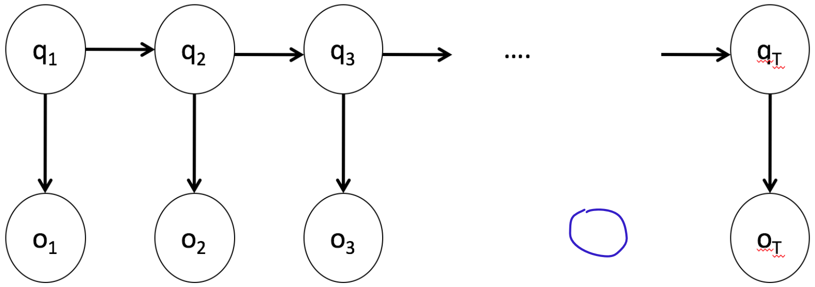


Figure 1: Hidden Markkov model

ii. Is q_T conditionally independent of q_3 given q_{T-1}?

Yes

iii. Is q_T conditionally independent of q_1 given q_3?

Yes

iv. Is q_T conditionally independent of q_1 given o_3?

No

v. Is o_2 conditionally independent of o_1 given q_2?

Yes

(b) (30 pts) Consider the sequence of methylation '010011' of length 6. You may calculate by yourself or use programming for answering the question. However, you do not need to submit the program in the programming submission.

i. (3 pts) Compute the probability $\underbrace{P(\text{PBRRRB}, 010011)}$.

Joint Probability of (hidden States, Observable States)

$P(Y = 010011, X = PBRRRB)$

$= P(X_1 = P) P(Y_1 = 0 | X_1 = P) P(X_2 = B | X_1 = P) P(Y_2 = 1 | X_2 = B) P(X_3 = R | X_2 = B) P(Y_3 = 0 | X_3 = R) P(X_4 = R | X_3 = R)$
$P(Y_4 = 0 | X_4 = R) P(X_5 = R | X_4 = R) P(Y_5 = 1 | X_5 = R) P(X_6 = B | X_5 = R) P(Y_6 = 1 | X_6 = B)$

$= 0.2 \times 0.9 \times 0.3 \times 0.8 \times 0.1 \times 0.9 \times 0.4 \times 0.9 \times 0.4 \times 0.1 \times 0.4 \times 0.8$

$= 0.00001792$

ii. (10 pts) Use the forward-backward algorithm. Compute α's and β's for each position.

α values

	0	1	0	0	1	1
P	0.1800	0.0109	0.0278	0.0165	0.0010	0.0003
E	0.24	0.029	0.0299	0.0174	0.0025	0.0007
R	0.27	0.0190	0.0329	0.0254	0.0018	0.00037
B	0.0400	0.2288	0.0365	0.01182	0.02425	0.0152

β- values

	0	1	0	0	1	1
P	0.019	0.0947	0.145	0.173	0.32	1
E	0.024	0.078	0.137	0.232	0.42	1
R	0.022	0.083	0.141	0.219	0.39	1
B	0.032	0.051	0.103	0.351	0.6	1

iii. (3 pts) Compute the probability $P(010011)$.

$$P(O) = P(O_1, \ldots, O_T) = P(O_1 = 0, O_2 = 1, O_3 = 0, O_4 = 0, O_5 = 1, O_6 = 1)$$

$$= \sum_{i=1}^{4} P(O_1, \ldots, O_T, q_T = S_i) = \sum_{i=1}^{4} \alpha_T(i)$$

$$= \sum_{i=1}^{4} \alpha(i,6) = 0.000316 + 0.000739 + 0.00037 + 0.015199$$

$$= 0.016624$$

iv. (3 pts) Compute the probability $P(\text{PBRRRB}|010011)$.

$$P(Q = \text{PBRRRB} | O = 010011) = \frac{P(Q = \text{PBRRRB}, O = 010011)}{P(O = 010011)}$$

$$= \frac{0.00001792}{0.016624}$$

$$= 0.00107796$$

v. (3 pts) Compute the probability $P(q_3 = R|010011)$.

$$P(q_3 = R | 010011) = \frac{\alpha_t(R)\beta_t(R)}{\sum_{i=1}^{4} \alpha(i)\beta(i)} = \frac{(0.032985)(0.141281)}{\alpha_t(1)\beta_t(1) + \alpha_t(2)\beta_t(2) + \alpha_t(3)\beta_t(3) + \alpha_t(4)\beta_t(4)}$$

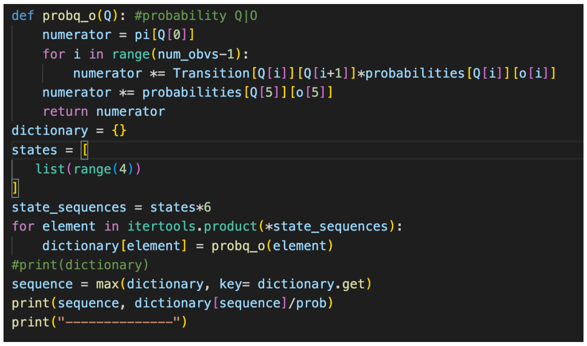
$$= \frac{(0.032985)(0.141281)}{(0.02785)(0.145789) + (0.0173816)(0.137945) + (0.032985)(0.141281) + (0.22880)(0.103187)}$$

$$= 0.280$$

vi. (8 pts) Compute the most probable sequence of annotation given this methylation

$$P(Q|O) = \frac{P(Q,O)}{P(O)} \rightarrow P(Q_1)\, P(O_1|Q_1)\, P(Q_2|Q_1)\, P(Q_3|Q_2)\, P(O_2|Q_2)\, P(Q_4|Q_3)\, P(O_3|Q_3)$$
$$P(O_4|Q_4)\, P(Q_5|Q_4)\, P(O_5|Q_5)\, P(Q_6|Q_5)\, P(O_6|Q_6)$$

data.



```
def probq_o(Q): #probability Q|O
    numerator = pi[Q[0]]
    for i in range(num_obvs-1):
        numerator *= Transition[Q[i]][Q[i+1]]*probabilities[Q[i]][o[i]]
    numerator *= probabilities[Q[5]][o[5]]
    return numerator
dictionary = {}
states = [
   list(range(4))
]
state_sequences = states*6
for element in itertools.product(*state_sequences):
    dictionary[element] = probq_o(element)
#print(dictionary)
sequence = max(dictionary, key= dictionary.get)
print(sequence, dictionary[sequence]/prob)
print("-------------")
```

$P(Q^*|O) = \max_Q P(Q|O)$

$Q^* = \text{argmax}_Q P(Q|O)$

$= \text{path defined by } \text{argmax}_j\, \delta_T(i)$

Used code to get the sequence [2,3,3,3,3,3]

Most probable path RBBBBB

(c) (5 pts) Calculate the parameters of this model given the following data. Assume the annotation is given. [Hint: There should be 4 initial parameters, 16 transition probabilities, 8 emission probabilities].

(4) (4x4)

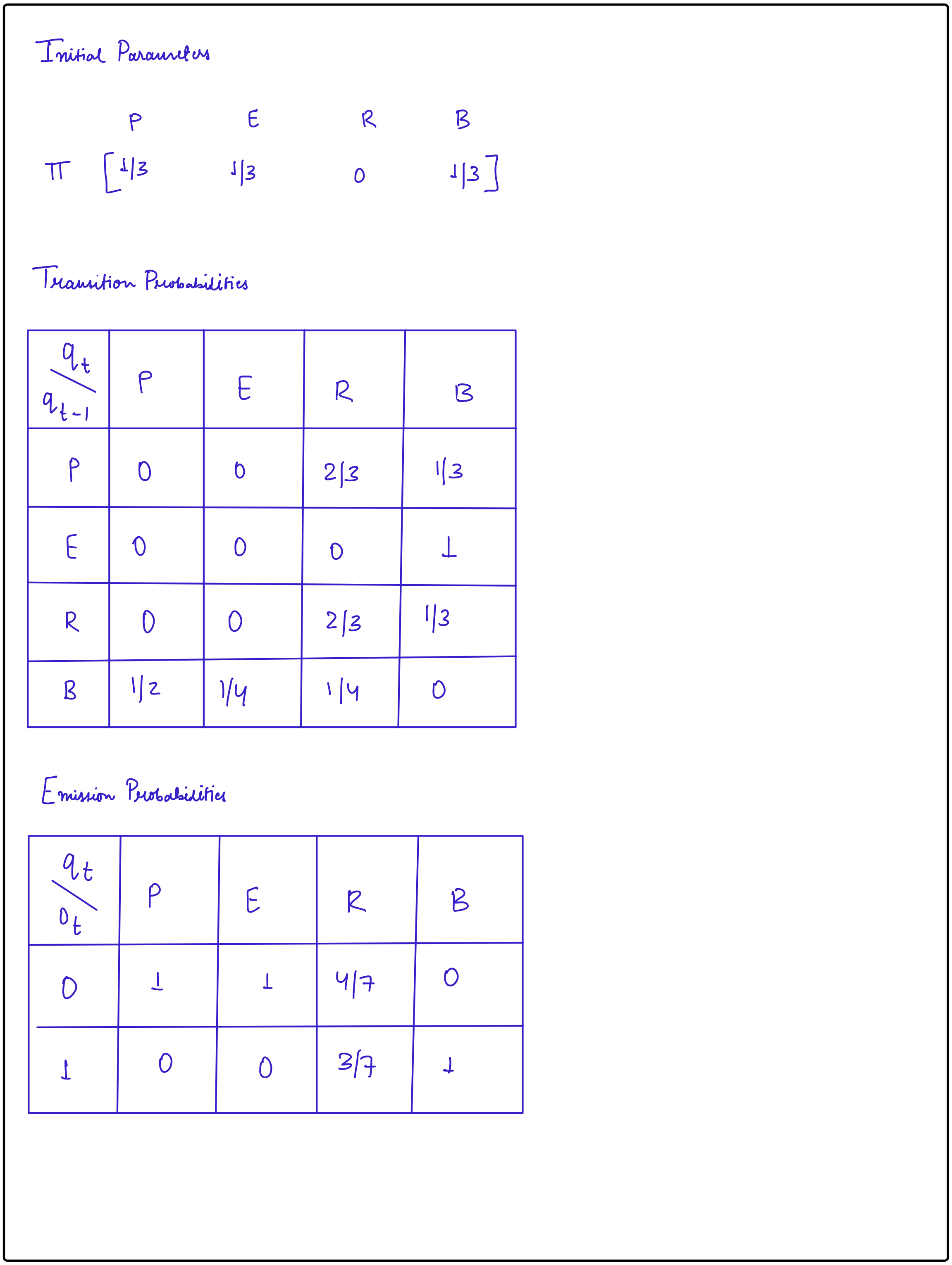
(4x2)

PBRRRB	010011
BEBPRR	101001
EBPRRB	010101

$P(Q_i)$

$P(O_i|Q_i)$

$P(Q_{i+1}|Q_i)$



Initial Parameters

$$\begin{array}{cc} & \begin{array}{cccc} P & E & R & B \end{array} \\ \pi & \begin{bmatrix} 1/3 & 1/3 & 0 & 1/3 \end{bmatrix} \end{array}$$

Transition Probabilities

q_t / q_{t-1}	P	E	R	B
P	0	0	2/3	1/3
E	0	0	0	1
R	0	0	2/3	1/3
B	1/2	1/4	1/4	0

Emission Probabilities

q_t / o_t	P	E	R	B
0	1	1	4/7	0
1	0	0	3/7	1

$(O_1, \dots, O_6) = 010011$

	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$	$t=6$
P						
E						
R						
B						

$$\alpha_{it} = P(O_1, \dots, O_t, q_t = i)$$

$$P(O_1, O_2, O_3, q_3 = P) = P(0, 1, 0, q_3 = P)$$

$$P(O_1, O_2, O_3, O_4, O_5, O_6, q_6 = P) = P(0, 1, 0, 0, 1, 1\ q_6 = P)$$