# CSE 494/598: Algorithms in Computational Biology - Assignment 4 $$\operatorname{Fall}\xspace 2022$$

Instructor: Heewook Lee	Due Date:	11/28/2022 11:59pm
Hidden Markov Models		

Note: that any program you implement should run within 5 mins (wall time) to produce an output to be satisfactory.

## Hidden Markov Models

Note: Assume initial probabilities are equally likely. Remember to compute probabilities in log space to prevent any underflow errors resulting in probabilities of zero.

1. [Viterbi Algorithm - 60pt] Implement the Viterbi Algorithm for finding the most likely sequence of hidden states  $\pi$ , where  $Pr(x, \pi)$  is maximized.

**Input:** An observed sequence of symbols x, followed by alphabet  $\Sigma$ , followed by a list of states states, followed by transition matrix T, followed by emission matrix E of an HMM  $(\Sigma, states, T, E)$ .

**Output:** The sequence of hidden states that maximizes  $Pr(x, \pi)$ .

## Sample input file content:

```
xyxzzxyxyy
------
x y z
-----
A B
-----
A B
A 0.641 0.359
B 0.729 0.271
-----
x y z
A 0.117 0.691 0.192
B 0.097 0.42 0.483
```

#### Sample output:

AAABBAAAA

2. [Parameter Estimation via Viterbi Learning - 60pt] Implement Viterbi Learning algorithm to estimate the unknown parameters of HMM to maximize  $Pr(x, \pi)$  over all possible parameter sets.

**Input:** A number of iterations i, followed by an observed sequence of symbols x, followed by alphabet  $\Sigma$ , followed by a list of states states, followed by initial transition matrix T, followed by emissions matrix E.

**Output:** Transition matrix followed by emission matrix that maximizes  $Pr(x,\Pi)$  over all  $\Pi$  and parameters (transition and emission matrices).

#### Sample input file content:

100

xxxzyzzxxzxyzxzxyxxzyzyzyyyyzzxxxzzxzyzzzxyxzzzxyzzxxxxzzzxyyxzzzzzyzzzxxzzxxxxx

yzzyxzxxxyxzyxxyzyxz

хуг

A B

-----

Α

A 0.582 0.418

B 0.272 0.728

X  $\mathbf{z}$ A  $0.129 \ 0.35$ 0.52

B 0.422 0.151 0.426

## Sample output:

Α

В

A 0.875 0.125

B 0.011 0.989

X у 0.75A 0.0 0.25

B 0.402 0.174 0.424

3. [Soft Decoding Problem - 60pt] Instead of finding the most likely hidden path, we want to find the conditional probability  $Pr(\Pi_i = k \mid x)$  that the HMM was in hidden state k when emitting ith symbol (at step i).

**Input:** An observed sequence of symbols  $x = x_1 \dots x_n$ , followed by alphabet  $\Sigma$ , followed by a list of states states, followed by transition matrix T, followed by emission matrix E of an HMM  $(\Sigma, states, T, E).$ 

**Output:** All conditional probabilities  $Pr(\Pi_i = k \mid x)$  for each state k and each step i (1 to n)

### Sample input file content:

yzzzyxzxxx

-----

хуг

**BBABABABAB** 

-----

ABC

#### Sample output:

Α В C A 0.01.0 0.0

B 0.8 0.20.0

C = 0.333 = 0.333 = 0.333

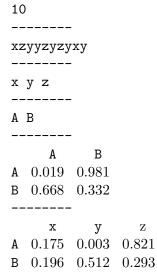
$$\begin{array}{ccccc} & x & y & z \\ \text{A} & 0.25 & 0.25 & 0.5 \\ \text{B} & 0.5 & 0.167 & 0.333 \\ \text{C} & 0.333 & 0.333 & 0.333 \end{array}$$

4. [Baum-Welch Learning for HMM - 60pt] (OPTIONAL for CSE494) Implement Baum-Welch algorithm to learn the unknown parameters of HMM. You will need soft decoding of both nodes and edges in Viterbi graph (node soft decoding done in the previous problem).

**Input:** A number of iterations i, followed by an observed sequence of symbols x, followed by alphabet  $\Sigma$ , followed by a list of states states, followed by initial transition matrix T, followed by emissions matrix E.

**Output:** Transition matrix followed by emission matrix that maximize  $Pr(x,\Pi)$  over all possible  $\Pi$  and parameters (transition and emission matricies).

# Sample input file content:



#### Sample output: