COMP 571: Homework #2

Vi Nguyen

March 2, 2017

1 HMMs and the Viterbi Algorithm [20 pts]

1.

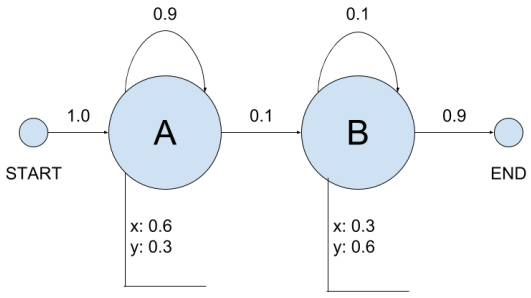
$$P(X, \Pi) = P(\Pi|X)P(X)$$

$$P(X) = \sum_{\pi} P(x, \pi)$$

$P(X)$ is constant for all paths Π. Therefore, $P(X, \Pi)$ is directly proportional with $P(\Pi|X)$, since when computing argmax on Π we are only varying Π, not X. Thus, argmax on Π for both $P(X, \Pi)$ and $P(\Pi|X)$ result in the same Π^*.

2. Consider the HMM below (fig. 1) and $X = xyyy$.

Figure 1: Example HMM for question 1-2.



This $\Pi^* = argmax_{\Pi} P(X|\Pi)$ is the path that, if given, has the highest probability of producing X. Therefore, in this case, Π^* would be $ABBB$,

since it is a valid path through the states and it maximizes the probability of producing X. It maximizes the probability of producing X because state A has the highest probability of emitting x and B has the highest probability of emitting y, so $ABBB$ has the highest probability of producing $X = xyyy$. However, the problem with this Π^* is that it doesn't take into account the probability of the path itself. Note here that $ABBB$ is a relatively improbable path through the states of the HMM.

Consider instead $\Pi^* = argmax_\Pi P(X, \Pi)$.

$$P(X, \Pi) = a_{0\Pi_1} \cdot e_{\Pi_1}(x) \cdot a_{\Pi_1\Pi_2} \cdot e_{\Pi_2}(x) \cdot a_{\Pi_2\Pi_3} \cdot e_{\Pi_3}(x) \cdot a_{\Pi_1\Pi_2}$$

This also takes into account the probability of transitioning from state to state. Thus, $\Pi^* = argmax_\Pi P(X, \Pi)$ is also dependent on how probable Π^* is itself. Thus, we do not want to consider $\Pi^* = argmax_\Pi P(X|\Pi)$.

2 HMMs and Length Distribution [25 pts]

1. We want to compute the probability $P(L_x = t)$. To create a sequence x of length t, we first need to enter state X, self-loop back to state X t times, and then exit state X.

   Since these are all independent events, we can compute

   $$P(L_x = t) = (1-\eta)(\prod_{i=1}^{t}(1-\eta))(\eta) = (\eta - \eta^2)\prod_{i=1}^{t}(1-\eta)$$
   $$P(L_x = t) = (\eta - \eta^2)(1-\eta)^t$$

   where $t \leq n$. If $t > n$, then $P(L_x = t) = 0$.

2. Let random variable X be the length of sequence x and random variable Y be the length of sequence y. Let us first consider $E[X]$.

   The expected value of a random variable is $E[X] = \sum x_i p_i$. In this case,

   $$E[X] = \sum_{i=0}^{n} i \cdot P(L_x = i) = \sum_{i=0}^{n} i \cdot (\eta - \eta^2)(1-\eta)^i$$

   $$= (\eta - \eta^2)\sum_{i=0}^{n} i(1-\eta)^i$$

   $$= \frac{(\eta - \eta^2)(1-\eta)(1-(n+1)(1-\eta)^n + n(1-\eta)^{n+1})}{\eta^2}$$

   $$= \frac{(1-\eta)(1-\eta)(1-(n+1)(1-\eta)^n + n(1-\eta)^{n+1})}{\eta}$$

$$= \frac{(1-\eta)^2(1-(n+1)(1-\eta)^n + n(1-\eta)^{n+1})}{\eta}$$

$$= \frac{(1-\eta)^2(1-(n+1)(1-\eta)^n + n(1-\eta)^{n+1})}{\eta}$$

We assume that L_x is not bounded, so we take the limit as $n \to \infty$. Note that $\eta < 1$, so $(1-\eta)^n$ and $(1-\eta)^{n+1}$ go to 0.

$$E[X] = \frac{(1-\eta)^2(1-0+0)}{\eta} = \frac{(1-\eta)^2}{\eta}$$

$E[Y]$ is computed in a similar manner to $E[X]$ since all of the transition probabilities in and out of state Y are the same as those for state X. Thus, the expected lengths of sequences produced by this HMM are

$$E[X] = E[Y] = \frac{(1-\eta)^2}{\eta}$$

Now we will compute what value η should be set to so that $E[X] = E[Y] = L^*$.

$$L^* = \frac{(1-\eta)^2}{\eta} \Rightarrow L^*\eta = (1-\eta)^2 \Rightarrow L^*\eta = 1 - 2\eta + \eta^2$$

$$\Rightarrow 0 = \eta^2 - 2\eta - L^*\eta + 1 \Rightarrow 0 = \eta^2 + (-2 - L^*)\eta + 1$$

Apply the quadratic formula to solve for η.

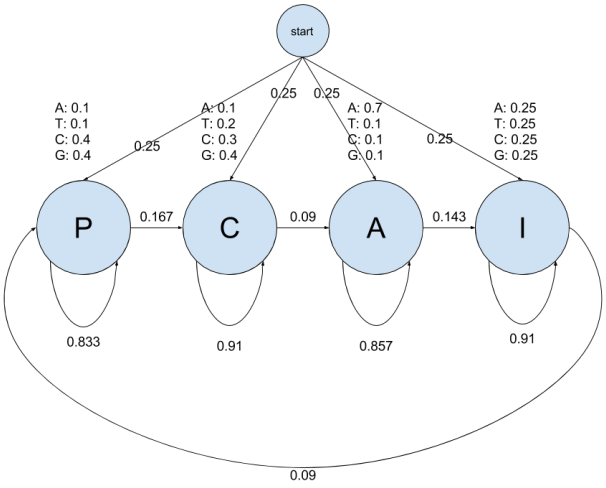
$$\eta = \frac{2 + L^* \pm \sqrt{(-2-L)^2 - 4}}{2} = \frac{2 + L^* \pm \sqrt{4 + 2L^* + (L^*)^2 - 4}}{2}$$

$$\eta = \frac{2 + L^* \pm \sqrt{2L^* + (L^*)^2}}{2}, \qquad 0 \leq \eta \leq 1$$

3 Finding Gene Structure using HMMs [30 pts]

1. See the HMM below (fig 2). P refers to promoter, C to coding region, A to poly-A tail, and I to intergenic region. Transition probabilities were computed in a similar manner to the one described in problem 2.

Figure 2: Constructed HMM for question 3-1.



2. I implemented my HMM using the HMM package for Matlab specified in the problem set pdf.

   Let the states be P (promoter), C (coding region), A (poly-A tail), and I (intergenic region).

   The starting probabilities matrix is as follows:

$$a_0 = \begin{bmatrix} 0.25 & 0.25 & 0.25 & 0.25 \end{bmatrix}$$

   The transition matrix is as follows:

$$a = \begin{bmatrix} 0.8333 & 0.0000 & 0.0000 & 0.0909 \\ 0.1667 & 0.9091 & 0.0000 & 0.0000 \\ 0.0000 & 0.0909 & 0.8571 & 0.0000 \\ 0.0000 & 0.0000 & 0.1429 & 0.9091 \end{bmatrix}$$

   The emission probabilities matrix is as follows:

$$e = \begin{bmatrix} 0.10 & 0.10 & 0.40 & 0.40 \\ 0.10 & 0.20 & 0.30 & 0.40 \\ 0.70 & 0.10 & 0.10 & 0.10 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{bmatrix}$$

   Generated DNA sequence (X): G T C C C G G T C G A A A A A A A A G C T T T T A A A T A C G G G C G G A G A C A A T A C G T C G C G A C C C G C G C G T T G T T T G C A A A A A A C C C T G G G T G A C C C T C G A A G T T G G G G G C A C G C C T

G C C G A A T C C G T T T T C T A A G G A A T G G A A T G T C G C T C A G T T G T T A C A A C T A C G G G G C G G T G G G G T A A A C A A G A T C G T G G A G C C C G C A C C C A G A C G

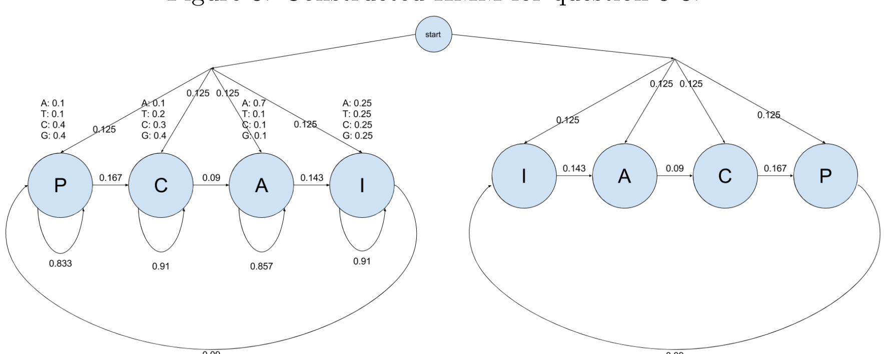
Generated labeling (P): P P P P P P P C C C A A A A A A A I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I P P P P P P P P P P P P P P C C C C C C A A A A A A A I I I I I I I I I I I I I I I I I I I I I I P C C C C C C C C C C C A I I I I I I I I P P P C C C C C C A A A A A A A I I I I I I I I I I I I I I I I I I I I I P P P P P P P P P C C C C C A A A A A A A A A A I I I I I I I I I P P P P C C C C C C C C C

Computed labeling (E): C C C C C C C C C C A A A A A A A A I I I I I I I I I I I I I I I I I I I I I I I I I I I I I P P P P P P P P P P P P C C C C C C C C C A A A A A A I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I P C C C C C C C C C C C C C A A A A A A I I I I I I I I I I I I I I I I I I I I I I I I

I mainly see errors in differences between generated and computed labeling. This may be due to an error in computing the B matrix due an incorrect distribution. Particularly, there is a very long string of I-states in the computed labeling that doesn't exist in the generated labeling.

3. The following HMM will read both positive strands or negative strands (not strands that have both positive and components). Note that on both "halves" of the HMM, the emission probabilities remain the same for the states and the self-loop probabilities remain the same.

Figure 3: Constructed HMM for question 3-3.



4 Profile HMMs [25 pts]

1. First, here is a table that expresses the counts that I computed from the multiple alignment (using Laplace's rule). The columns represent the

subscripts of the states (i.e. M_1, M_2, I_0, etc.) For transitions, column i refers to a transition from state-i to state-(i+1).

	0	1	2	3
A	-	3	6	1
C	-	1	1	4
T	-	1	2	1
G	-	7	2	6
A	1	1	1	1
C	1	3	1	1
T	1	1	1	1
G	1	1	1	1
M-M	9	6	8	9
M-D	1	2	1	1
M-I	1	3	1	1
I-M	1	3	1	1
I-D	1	1	1	1
I-I	1	1	1	1
D-M	1	1	2	1
D-D	1	1	1	1
D-I	1	1	1	1

From the above counts table follows the following emission and transition probabilities.

	0	1	2	3
A	-	0.250	0.545	0.083
C	-	0.083	0.091	0.333
T	-	0.083	0.182	0.083
G	-	0.583	0.182	0.500
A	0.250	0.167	0.250	0.250
C	0.250	0.500	0.250	0.250
T	0.250	0.167	0.250	0.250
G	0.250	0.167	0.250	0.250
M-M	0.818	0.545	0.800	0.818
M-D	0.091	0.273	0.100	0.091
M-I	0.091	0.182	0.100	0.091
I-M	0.333	0.600	0.333	0.333
I-D	0.333	0.200	0.333	0.333
I-I	0.333	0.200	0.333	0.333
D-M	-	0.333	0.500	0.333
D-D	-	0.333	0.250	0.333
D-I	-	0.333	0.250	0.333