CBMF 4671 Homework 5

Alex Ying

I discussed the problems with Jason Mohabir and Harry Lee.

Problem 1

a. For events A and B which have probabilities of $P(A)$ and $P(B)$ of occurring, their joint probability is $P(A \cap B) = P(A)P(B)$. By using log probabilities, this multiplication operation is changed to addition, as $log(P(A \cap B)) = log(P(A)P(B)) = log(P(A)) + log(P(B))$. Since addition accumulates changes in orders of magnitude far more slowly than multiplication, the risk of underflow (and overflow) error is greatly reduced. This greatly increase the number of probabilities that can be worked with, as the log of the float limit in Python, $log(2.2 \times 10^{-308})$, is only -307.7, which means that the probabilities can get far closer to 0 before exceeding a decimal limit.

b. Using log probabilities is problematic for the forward-backward algorithm because the foward algorithm requires summing probabilities, in that the probability $f_k(t+1) = P(y(1), ..., y(t+1), p(t+1) = k)$ is calculated from:

$$f_k(t+1) = \sum_{i=1}^{K} (f_i(t) T_{i,k} E_{k,y(t+1)})$$

d. The derivation of $P(p(n) = k, p(n+1) = k', Y) = f_k(n) T_{k,k'} E_{k',y(n+1)} b_{k'}(n+1)$ is given below:

Let events $p(n) = k$ and $p(n+1) = k'$ be denoted as $\pi_{n,k}$ and $\pi_{n+1,k'}$, respectively. The probability of observing sequence Y can be viewed as the probability of observing each term in Y from 1 to N, where N is the length of the sequence of observations.

$$P(\pi_{n,k}, \pi_{n+1,k'}, Y) = P(\pi_{n,k}, \pi_{n+1,k'}, y(1), ..., y(N))$$

Rearranging gives:

$$P(y(1), ..., y(n), \pi_{n,k}, \pi_{n+1,k'}, y(n+1), y(n+2), ..., y(N))$$
$$= P(y(n+2), ..., y(N) | \pi_{n,k}, \pi_{n+1,k'}, y(1), ..., y(n), y(n+1)) P(\pi_{n,k}, \pi_{n+1,k'}, y(1), ..., y(n), y(n+1))$$

Note that the probability of an observation $y(t)$ is conditionally independent of any past observation, so:

$$P(y(n+2), ..., y(N) | \pi_{n,k}, \pi_{n+1,k'}, y(1), ..., y(n), y(n+1)) = P(y(n+2), ..., y(N) | \pi_{n,k}, \pi_{n+1,k'})$$

Observations $y(n+2), ..., y(N)$ are also conditionally independent of the state at n, $p(n)$. This is because the only implication of $p(n)$ is how it can transition to $p(n+1)$, which is already given, and how it can emit, which is irrelevant. $y(n+2), ..., y(N)$ is dependent on $p(n+1)$ because $p(n+1)$ dictates the distribution of states $p(n+2)$, which dictates the probaility of all future observations. Therefore:

$$P(y(n+2), ..., y(N) | \pi_{n,k}, \pi_{n+1,k'}) = P(y(n+2), ..., y(N) | \pi_{n+1,k'}) = b_{k'}(n+1)$$

The probability $P(\pi_{n,k}, \pi_{n+1,k'}, y(1), ..., y(n), y(n+1))$ can be further simplified as:

$$P(\pi_{n,k}, \pi_{n+1,k'}, y(1), ..., y(n), y(n+1)) = P(y(n+1) | \pi_{n,k}, \pi_{n+1,k'}, y(1), ..., y(n)) P(\pi_{n,k}, \pi_{n+1,k'}, y(1), ..., y(n))$$

$$= P(y(n+1)|\pi_{n,k}, \pi_{n+1,k'}, y(1), ..., y(n))P(\pi_{n,k}, \pi_{n+1,k'}, y(1), ..., y(n))$$

$y(n+1)$ is conditionally independent of $\pi_{n,k}$ and $y(1), ..., y(n)$, simplifying to:

$$P(y(n+1)|\pi_{n,k}, \pi_{n+1,k'}, y(1), ..., y(n)) = P(y(n+1)|\pi_{n+1,k'}) = E_{k',y(n+1)}$$

The probability $P(\pi_{n,k}, \pi_{n+1,k'}, y(1), ..., y(n))$ can be further simplified as:

$$P(\pi_{n,k}, \pi_{n+1,k'}, y(1), ..., y(n)) = P(\pi_{n+1,k'}|\pi_{n,k}, y(1), ..., y(n))P(\pi_{n,k}, y(1), ..., y(n))$$

$$= P(\pi_{n+1,k'}|\pi_{n,k}, y(1), ..., y(n))f_k(n)$$

Since the hidden states of an HMM are only dependent on the previous hidden state:

$$P(\pi_{n+1,k'}|\pi_{n,k}, y(1), ..., y(n)) = P(\pi_{n+1,k'}|\pi_{n,k}) = T_{k,k'}$$

This gives the desired result:

$$P(p(n) = k, p(n+1) = k', Y) = f_k(n)T_{k,k'}E_{k',y(n+1)}b_{k'}(n+1)$$

e. The derivation of $P(p(n+1) = k'|p(n) = k, Y) = P(p(n+1) = k', p(n) = k, Y)/(f_k(n)b_k(n))$ is known simply from:

$$P(A, B, C) = P(A|B, C)P(B, C)$$

and that:

$$P(p(n) = k, Y) = f_k(n)b_k(n)$$

f. Part (c) requires adding probabilities, which means that underflow is a risk as mentioned in part (b), since the addition of probabilities cannot be done using their logs. Probabilities are not added in part (e), but multiplied, which means that their log can be used, avoiding underflow.

Problem 2

a. The nucleotide sequence as described can be modeled by an HMM with hidden states representing the 4 bases, (A, C, G, and T), and the termination state (S), and emits the current base with 100% probability. The given occurence frequencies of particular pairs define the transition matrix. It is first necessary to check that $\sum_X P(XpY) = P(Y)$, as the frequency of a base, Y, should not depend on whether it is part of a pair or not.

Since length of the chain is geometrically distributed around a mean length of 1 Mb, the probability of the chain terminating is 10^{-8} at every base. For simplicity, let $\alpha = 1 - 10^{-8}$.

	A	C	G	T	S
A	$\alpha/4$	$\alpha/4$	$9\alpha/40$	$11\alpha/40$	10^{-8}
C	$7\alpha/18$	$\alpha/4$	$\alpha/18$	$11\alpha/36$	10^{-8}
G	$\alpha/4$	$9\alpha/40$	$\alpha/4$	$11\alpha/40$	10^{-8}
T	$5\alpha/22$	$9\alpha/44$	$7\alpha/22$	$\alpha/4$	10^{-8}
S	0	0	0	0	1

The emission matrix is simply:

	A	C	G	T
A	1	0	0	0
C	0	1	0	0
G	0	0	1	0
T	0	0	0	1
S	0	0	0	0

b. The nucleotide sequence with CpG islands can be modeled by an HMM with 6 hidden states, representing the 4 bases in CpG-depleted distribution (A_n, C_n, G_n, and T_n), a single state representing being in a CpG island (I), and the termination state (S). Let it be that the probability of going from an X_n state to I is subject to CpG-depleted distribution and the probability of going from I to a Y_n state is subject to CpG island distribution. The probability of going from any X_n state to I should sum to $10^{-5}(1-10^{-8})$, as the length of CpG-depleted regions are 100 kb in length on average, geometrically distributed, and do not terminate. Likewise, the probability of going from any X_i state to I should sum to $10^{-3}(1-10^{-8})$, as the length of CpG islands is given to be 1 kb in length, geometrically distributed, and do not terminate. The probability of transitioning to the termination state from any other state should be 10^{-8}, as described in (a). Let $\alpha = 1-10^{-8}$, $\beta = 1-10^{-5}$, and $\gamma = 1-10^{-3}$.

This gives the transition matrix:

	A_n	C_n	G_n	T_n	A_i	C_i	G_i	T_i	S
A_n	$\alpha\beta/4$	$\alpha\beta/4$	$9\alpha\beta/40$	$11\alpha\beta/40$	$(\alpha 10^{-5})/4$	$(\alpha 10^{-5})/4$	$9(\alpha 10^{-5})/40$	$11(\alpha 10^{-5})/40$	10^{-8}
C_n	$7\alpha\beta/18$	$\alpha\beta/4$	$\alpha\beta/18$	$11\alpha\beta/36$	$7(\alpha 10^{-5})/18$	$(\alpha 10^{-5})/4$	$(\alpha 10^{-5})/18$	$11(\alpha 10^{-5})/36$	10^{-8}
G_n	$\alpha\beta/4$	$9\alpha\beta/40$	$\alpha\beta/4$	$11\alpha\beta/40$	$(\alpha 10^{-5})/4$	$9(\alpha 10^{-5})/40$	$(\alpha 10^{-5})/4$	$11(\alpha 10^{-5})/40$	10^{-8}
T_n	$5\alpha\beta/22$	$9\alpha\beta/44$	$7\alpha\beta/22$	$\alpha\beta/4$	$5(\alpha 10^{-5})/22$	$9(\alpha 10^{-5})/44$	$7(\alpha 10^{-5})/22$	$(\alpha 10^{-5})/4$	10^{-8}
A_i	$\alpha\gamma/4$	$\alpha\gamma/4$	$\alpha\gamma/4$	$\alpha\gamma/4$	$11(\alpha 10^{-3})/40$	$9(\alpha 10^{-3})/40$	$9(\alpha 10^{-3})/40$	$11(\alpha 10^{-3})/40$	10^{-8}
C_i	$\alpha\gamma/4$	$\alpha\gamma/4$	$\alpha\gamma/4$	$\alpha\gamma/4$	$11(\alpha 10^{-3})/40$	$9(\alpha 10^{-3})/40$	$9(\alpha 10^{-3})/40$	$11(\alpha 10^{-3})/40$	10^{-8}
G_i	$\alpha\gamma/4$	$\alpha\gamma/4$	$\alpha\gamma/4$	$\alpha\gamma/4$	$11(\alpha 10^{-3})/40$	$9(\alpha 10^{-3})/40$	$9(\alpha 10^{-3})/40$	$11(\alpha 10^{-3})/40$	10^{-8}
T_i	$\alpha\gamma/4$	$\alpha\gamma/4$	$\alpha\gamma/4$	$\alpha\gamma/4$	$11(\alpha 10^{-3})/40$	$9(\alpha 10^{-3})/40$	$9(\alpha 10^{-3})/40$	$11(\alpha 10^{-3})/40$	10^{-8}
S	0	0	0	0	0	0	0	0	1

The emission matrix is:

	A	C	G	T
A_n	1	0	0	0
C_n	0	1	0	0
G_n	0	0	1	0
T_n	0	0	0	1
A_i	1	0	0	0
C_i	0	1	0	0
G_i	0	0	1	0
T_i	0	0	0	1
S	0	0	0	0