1. Show how to compute the forward probabilities for an HMM in log space.

   Solution: (similar to the note posted on Piazza) First we take the log of the base case

   $$\log f_1(k) = \log \pi_k + \log e_k(x_1).$$

   Then in the recursive step, we want to compute

   $$\log f_t(k) = \log e_k(x_t) + \log \sum_j f_{t-1}(j) a_{jk}.$$

   But we haven't actually been computing the forward probabilities directly, we've been computing them in log space. So what we really have is

   $$\log f_t(k) = \log e_k(x_t) + \log \sum_j e^{\log f_{t-1}(j)} \cdot e^{\log a_{jk}}.$$

   In general, we want to compute an expression of the form

   $$\log \sum_i e^{d_i}.$$

   The problem arises when all of the d_i are very large negative numbers, which can happen since d_i is the log of a probability, which could be very small. Then when we take the exponential of a very large negative number, the computer returns 0. If this happens for all d_i, we end up trying to take the log of 0, which returns $-\infty$, and this error propagates into our downstream computation. To fix this, we can subtract off the maximum, $D = \max_i\{d_i\}$, from each d_i. This ensures that for at least one d_k, $e^{d_k - D} = e^0 = 1$, so we won't be taking the log of 0. Overall, our computation becomes

   $$\log \sum_i e^{d_i} = \log \sum_i e^{d_i} \cdot e^{D-D} = D + \log \sum_i e^{d_i - D}.$$

2. Suppose you are given fixed phylogenetic tree branch lengths $t_1, t_2, \cdots, t_B$ and mutation counts $x = x_1, x_2 \cdots, x_B$ for each branch. Assuming mutations occur as a Poisson process with mutation rate μ, find the MLE (maximum likelihood estimator) for μ. Recall that if a random variable X is Poisson distributed with parameter λ, then the pmf (probability mass function) for X is $\mathbb{P}_\lambda(X = x) = \lambda^x e^{-\lambda}/x!$.

   Solution: Let $X = X_1, X_2, \cdots, X_B$. Since each branch is independent, we can rewrite the likelihood of our data as

   $$L(\mu; x) = \mathbb{P}_\mu(X = x) = \prod_{i=1}^{B} \mathbb{P}_\mu(X_i = x_i) = \prod_{i=1}^{B} \frac{(\mu t_i)^{x_i} e^{-\mu t_i}}{x_i!}.$$

   Then we can take the log to find the log likelihood

   $$\ell(\mu; x) = \sum_{i=1}^{B} x_i \log(\mu t_i) - \mu t_i - \log(x_i!).$$

   Then take the derivative with respect to μ and set it equal to 0. Finally solve for μ to get

   $$\frac{\mathrm{d}\ell(\mu; x)}{\mathrm{d}\mu} = \sum_{i=1}^{B} \frac{x_i}{\mu} - t_i = 0 \quad \Rightarrow \quad \hat{\mu} = \frac{\sum_{i=1}^{B} x_i}{\sum_{i=1}^{B} t_i}.$$

3. The dynamic programming table below shows the fill-in step of Nussinov's RNA folding algorithm on the string S. Fill in the two missing entries, $B(3, 10)$ and $B(1, 12)$. Then perform back-tracing on both entries to find the optimal secondary structure(s) for $S[3 \ldots 10]$ and S. Does the number of unique optimal structures equal the number of tracebacks?

Solution:

S		A	U	C	G	G	A	U	C	G	A	A	C
		1	2	3	4	5	6	7	8	9	10	11	12
A	1	0	0	0	0	1	2	3	3	3	3	3	4
U	2		0	0	0	1	2	2	2	2	3	3	3
C	3			0	0	1	1	1	1	2	2	2	2
G	4				0	0	0	0	1	1	1	1	2
G	5					0	0	0	1	1	1	1	2
A	6						0	0	0	0	1	1	1
U	7							0	0	0	1	1	1
C	8								0	0	0	0	1
G	9									0	0	0	1
A	10										0	0	0
A	11											0	0
C	12												0

Let black arrows represent shared paths above. There are 4 traceback paths from $B(3, 10)$ (green, cyan, orange, purple), but only 3 unique secondary structures, shown below. There are 2 traceback paths from $B(1, 12)$ (red, blue), but both of these paths represent the same RNA secondary structure, shown below.

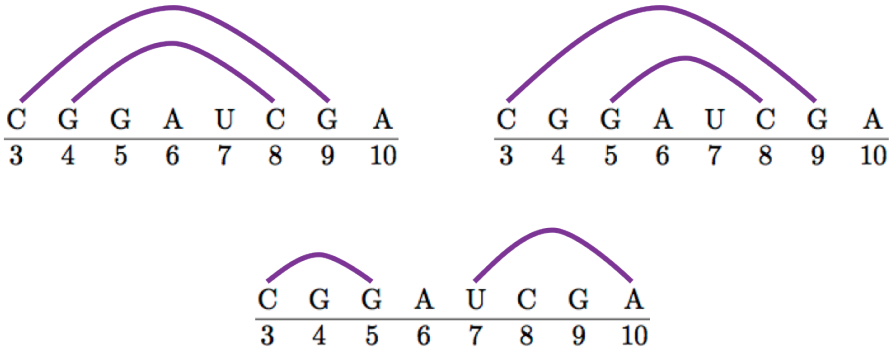

Figure 1: There are three optimal secondary structures for $S[3 \ldots 10]$.

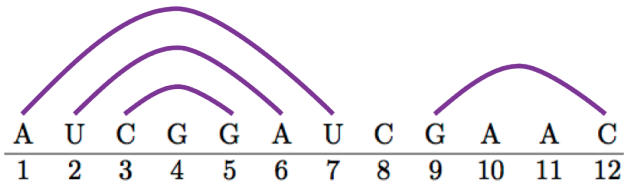

Figure 2: There is one optimal secondary structure for S.

4. Suppose a weighted coin, with probability θ of turning up heads is flipped until the first head occurs. Let X represent the total number of flips prior to getting heads. X is said to be geometrically distributed. Compute the pmf for X, $f(x) = \mathbb{P}(X = x)$. Show that $f(x)$ is a proper probability distribution. Also compute the cmf (cumulative mass function) for X, $F(x) = \mathbb{P}(X \leq x)$.

   Solution: The probability of getting x tails is $(1-\theta)^x$. Therefore the probability of getting x tails, then getting a head is

   $$f(x) = (1-\theta)^x \theta.$$

   To show this is a valid probability distribution, we need to make sure the sum of $f(x)$ over all possible x is 1. Using the formula for an infinite geometric series,

   $$\sum_{x=0}^{\infty} f(x) = \theta \sum_{x=0}^{\infty} (1-\theta)^x = \theta \cdot \frac{1}{1-(1-\theta)} = 1.$$

   Finally, we can use the formula for a finite geometric series to find $F(x)$,

   $$F(x) = \theta \sum_{y=0}^{x} (1-\theta)^y = \theta \cdot \frac{1-(1-\theta)^x}{1-(1-\theta)} = 1-(1-\theta)^x.$$

5. The following are two key properties of expectation.

   (a) Let X and Y independent discrete random variables. Prove that

   $$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y].$$

   Does this property imply independence?

   Solution: If X and Y are independent, then $\mathbb{P}(X, Y) = \mathbb{P}(X)\mathbb{P}(Y)$. So writing out the expectation of XY, we get

   $$\begin{aligned}\mathbb{E}[XY] &= \sum_{x,y} xy\mathbb{P}(x,y) = \sum_x \sum_y xy\mathbb{P}(x)\mathbb{P}(y) \\ &= \left(\sum_x x\mathbb{P}(x)\right) \cdot \left(\sum_y y\mathbb{P}(y)\right) = \mathbb{E}[X]\mathbb{E}[Y].\end{aligned}$$

   This property does not necessarily imply independence; you can construct random variables that satisfy this equation, but are not independent.

   (b) Now let X and Y be arbitrary discrete random variables, and $c, d \in \mathbb{R}$ arbitrary constants. Prove that

   $$\mathbb{E}[cX + dY] = c\mathbb{E}[X] + d\mathbb{E}[Y].$$

   Does this result require independence of X and Y? Show that if $\mathbb{P}(X \leq Y) = 1$ then $\mathbb{E}[X] \leq \mathbb{E}[Y]$.

Solution: Beginning with the LHS:

$$\begin{aligned}
\mathbb{E}[cX + dY] &= \sum_{x,y} (cx + dy)\mathbb{P}(x, y) \\
&= c\sum_{x,y} x\mathbb{P}(x, y) + d\sum_{x,y} y\mathbb{P}(x, y) \\
&= c\sum_{x} x\sum_{y} \mathbb{P}(x, y) + d\sum_{y} y\sum_{x} \mathbb{P}(x, y) \\
&= c\sum_{x} x\mathbb{P}(x) + d\sum_{y} y\mathbb{P}(y) \\
&= c\mathbb{E}[X] + d\mathbb{E}[Y].
\end{aligned}$$

No where did we require that X and Y be independent. For the next part of this problem, we can use the result above with $c = -1$ and $d = 1$:

$$\mathbb{E}[Y] - \mathbb{E}[X] = \mathbb{E}[Y - X] = \sum_{x,y} (y - x)\mathbb{P}(x, y) \geq 0$$

$$\Rightarrow \quad \mathbb{E}[Y] \geq \mathbb{E}[X].$$