CS 369 Assignment 3 2017

Due Monday May 8 5:00 pm

Write your submission in a Jupyter notebook and write any code in Python 3. You should submit the .ipynb file and a .html file with all output displayed. The primary document the markers will look at is the .html file.

In your report, include explanations of what you are doing and comments in the code. Where you are making mathematical derivations, show your working.

Submit your notebook as an .ipynb file and as an .html file with all output showing to `https://adb.auckland.ac.nz/` by 6:00 pm on the due date.

1. *[5 marks total]* In this problem you model the length of insertions in a genetic sequence using the geometric distribution and find the maximum likelihood estimate. Show your working in the below.

   (a) Explain why the correct choice of geometric distribution for modelling the length of insertions is the one that takes values in $\{1, 2, 3, \dots\}$ and has probability mass function $\Pr(X = x) = (1-p)^{x-1}p$.

   (b) Suppose we have sampled n independent insertions and found the ith insertion to have length k_i. Let the data vector be $D = (k_1, k_2, \dots, k_n)$. Write down an expression for the likelihood of data D, $\mathcal{L}(p; D) = \Pr(D|p)$. Make sure you simplify the expression so that any sums or products are moved as far to the right as possible.

   (c) By finding the root of the derivative of the likelihood $\mathcal{L}(p; D)$, show that the maximum likelihood estimate of p is $\frac{1}{m}$, where $m = \frac{1}{n}\sum_{i=1}^{n} k_i$ is the sample mean.

2. *[4 marks total]* In this short question, you will write methods for generating samples from the exponential and Poisson distributions.

   (a) Using the inversion method, write function `rand_exp` that draws samples from the exponential distribution. It should take two parameters: `rate` which is the rate paramter of the exponential distribution (called λ in the notes) and `size` which is a positive integer number of samples to return. It should return an ndarray of length `size`.

   (b) Using your `rand_exp` function from part (a), write a function `rand_poiss` that draws samples from the Poisson distribution. It should take two parameters: `rate` which is the rate parameter of the Poisson distribution (called λ in the notes) and `size` which is an integer number of samples to return. It should return an ndarray of length `size`.

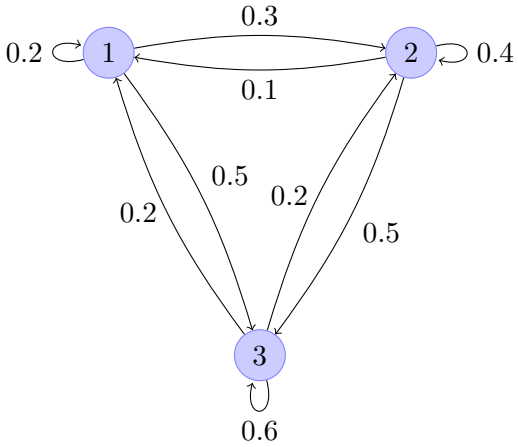
   (c) Provide evidence that your implementations of `rand_exp` and `rand_poiss` are correct by making a histogram of 10000 samples from each function with the rate parameter set at $\lambda = 2$ and comparing them to histograms of samples drawn using analogous functions in the numpy.random library.

3. *[6 marks total]* In this problem you will write a method that simulates a Markov chain on a finite state space with a given transition matrix.

   (a) Write a method `check_matrix` that takes as input a matrix P and returns a boolean output which is 1 if P is a valid transtion matrix and 0 if is not.

   (b) Write a method `markov_chain` that takes as input a matrix P, an integer $n > 0$ and a start state s and returns a sampled path of length n starting in state s from the Markov chain with transition matrix P. The method should check that all inputs are valid. Use the `numpy.random.choice` method but no other random number generation methods for this problem.

   (c) Write down the transition matrix for the Markov model shown here and simulate a chain of length 10 starting in state 1.



   (d) Simulate a single long chain starting at 1 and use it to estimate the probabilities $p_{1i}(\infty)$ for $i \in \{1, 2, 3\}$. It is not necessary to print the chain to output. Check your answer by comparison with the n-step transition matrix, P^n, for some large value of n.

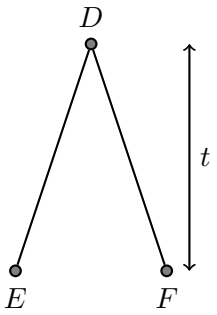
4. *[9 marks total]* The Jukes-Cantor model of DNA sequence evolution is simple: each site mutates at rate μ and when a mutation occurs, a new base is chosen uniformly at random from the four possible bases, $\{A, C, G, T\}$. If we ignore mutations from base X to base X, the mutation rate is $\frac{3}{4}\mu$. All sites mutate independently of each other.

   Thus we observe mutations at a site after an exponentially distributed waiting time with rate $\frac{3}{4}\mu$. At a mutation, choose from the 3 possible bases to mutate to with equal probability.

   A sequence that has evolved over time according to the Jukes-Cantor model has each base equally likely to occur at each site.

   Your programs should write sequences consisting of $\{A, C, G, T\}$, though it may be easier internally to translate the bases to integers, $\{1, 2, 3, 4\}$ for example.

   (a) Write a method that simulates pairs of sequences that have diverged from a recent common ancestor t time units ago. Assume that evolution has occurred according to the Jukes Cantor model. The distribution for the sequence of the most recent common ancestor is uniform over the four possible bases at each site. The method should take sequence length, time t and mutation rate μ as inputs. It should return the ancestral sequence (E in the figure) and



Sequence D is the most recent common ancestor of sequences E and F.
The time since E split from F is t time units.

the descendant sequences (E and F in the figure). You may use methods `choice, exponential` and `poisson` from the `numpy.random` library.

Simulate a pair of sequences of length 50 with $\mu = 0.01$ and $t = 10$. Print the resulting sequences along with the ancestral sequence. Report the number of sites at which each sequence differs from the ancestral sequence and from its sibling sequence (i.e., the number of sites difference between D and E, D and F and E and F.

(b) Explain why you would expect the number of mutations that occur on a tree to be Poisson distributed with parameter $2tL\frac{3}{4}\mu$, where L is the sequence length. Simulate 1000 pairs of sibling sequences of length 1000 with $\mu = 0.01$ and $t = 25$. For each simulated pair, count the number of sites at which they differ from each other. Report the mean and variance of the number of differing sites. Is this number Poisson distributed with parameter $2tL\frac{3}{4}\mu$? Explain why or why not.

(c) Simulate a pair of sibling sequences, E and F, of length 10000 with $t = 10$ and $\mu = 0.03$ and use them to calculate the probabilities that a G is aligned to base X for $X \in \{A, C, G, T\}$, p_{GX}. Assume that $p_{ab} = p_{ba}$.

Your estimates for p_{GX} should be close to the theoretical values

$$p_{GX} = \begin{cases} \frac{1}{4} + \frac{3}{4}\exp(-2t\mu) & \text{if } X = G \\ \frac{1}{4} - \frac{1}{4}\exp(-2t\mu) & \text{if } X \neq G. \end{cases}$$

(d) Consider now a model of sequence mutation that also includes insertions and deletions. Insertions and deletions occur at exponentially distributed intervals with per site rate $\mu/10$ (so that a sequence of length k would experience insertions at rate $k\mu/10$ and deletions at rate $k\mu/10$). At an insertion, 3 randomly chosen bases are inserted immediately after the chosen site. At a deletion, the chosen site and the two following (if they exist) are deleted. If this model were run for time t on a sequence of initial length L, explain why the number of insertion and deletion events would not be Poisson distributed with parameter $2Lt\mu/10$.