More on phylogenetic reconstruction

In this homework we will return to the phylogenetic reconstruction problem introduced in homework 4. You might want to have a look at that homework for background and motivation. We copy below the model definition.

In our model, the phylogenetic tree (the ‘tree of life’) is a binary tree, with a single ancestor and $\ell + 1$ generations (see Figure 1 for an example with $\ell = 3$). The generations will be denoted by $\{0, 1, \ldots, \ell\}$. Generation 0 has a single node (a common ancestor to all the species in the tree) that will be called ‘root’ and denoted by $\circ$. Generation $t \in \{1, \ldots, \ell\}$ has $2^t$ nodes. The nodes at the last generation (there is $2^\ell$ of them) will also be called ‘leaves’.

Each node corresponds to a species. Each species (except leaves) has two species that are directed descendants. Each species, except the root, has one species that is its direct ancestor.

In this homework, we will focus on the simplified case whereby each species $i$ is characterized by a length-one DNA sequence, i.e. by a single base $s_i \in \{A, T, C, G\}$. As argued in homework 4, this is the basic building block for understanding sequences of general lengths.

We assume the following model for mutations, with mutation parameter $p \in [0, 3/4]$:

1. The sequence at the root is uniformly random among all $4^n$ possible sequences.

2. If the species $i$ has sequence $s_i$, the sequences of each of the descendants of $i$ is generated independently as follows. Each base in $s_i$ mutates independently with probability $p$. If it mutates, one of the 3 remaining letters replaces the existing one with equal probability.

2’ An equivalent description of the mutation mechanism at the previous point is as follows. Each base is tagged for ‘potential mutation’ independently with probability $4p/3$. Once it is tagged, its new value is given by one of the 4 letters, each with probability $1/4$.

Leaves correspond to species that are alive at present time, and we now know their DNA sequence. Our objective will be to infer the sequence of the root species.
The following is a useful observation that perhaps you discovered while solving homework 4. Consider a certain line of descendants of the root species, say the leftmost branch in the tree. If, at any point along the line of descendants, the basis is tagged for potential mutation, then at all subsequent generations the sequence \( s_i \) is independent of the root value and uniformly random in \( \{A, C, T, G\} \).

\[ X_i = \begin{cases} 1 & \text{if } s_i = s_0, \\ 0 & \text{otherwise.} \end{cases} \]

and let \( Z_\ell \) be the fraction of leaf nodes that are equal to the root in an \( \ell \)-generations tree

\[ Z_\ell = \frac{1}{2^\ell} \sum_{i \in \text{Leaves}(\ell)} X_i. \]

Develop a formula for \( \mathbb{E}(Z_\ell) \) as a function of \( p \) and \( \ell \). Plot the result as a function of \( p \) for \( \ell \in \{5, 10, 15, 20\} \).

[Note: This should be a small variation on something that you did in homework 4.]

(2) Reconsider the program that you wrote for generating random samples from this model, for \( \ell = 10 \) and \( p \in \{0.05, 0.1, 0.15, 0.2, 0.25, 0.3\} \). In each case, generate 1,000 realization of the model and evaluate \( Z_\ell \) for each such realization.

Plot an histogram of \( Z_\ell \) for each of the cases \( p \in \{0.05, 0.1, 0.15, 0.2, 0.25, 0.3\} \). Compare the mean of this histogram with the expectation computed at the previous point.

(3) Develop a formula for \( \text{Var}(Z_\ell) \) as a function of \( p \) and \( \ell \). Compare the result with the empirical variance derived from the histograms at the previous point.

(4) Define the ratio

\[ R(p; \ell) = \frac{\mathbb{E}(Z_\ell) - 1/4}{\sqrt{\text{Var}(Z_\ell)}}. \]

Plot \( R(p; \ell) \) for \( \ell \in \{5, 10, 15\} \) as a function of \( p \). What is the interpretation of this ratio? What does it mean when it is small or large? For which values of \( p \) does it get very small, and for which very large? What happens in particular for \( p \in \{0.05, 0.1, 0.15, 0.2, 0.25, 0.3\} \)?
Reconsider the majority method for reconstructing the ancestral sequence $s_o$. In the present case (length-one sequences) this simply takes a majority vote among the bases at the leaves, in order to decide the letter at the root. Let $\hat{s}_o^{\text{majority}}$ be the corresponding estimate at the root.

Using the above program, estimate the success probability $P(\{\hat{s}_o = s_o\})$ for $p \in \{0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$ and $\ell = 10$.

Show that

$$P(\{\hat{s}_o = s_o\}) \leq P(\{Z_\ell > 1/4\}).$$

Use the simulation results to compare the two quantities in this inequality for $p \in \{0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$ and $\ell = 10$. Compare them with the value of the ratio $R(p; \ell)$.

Can this ratio be used to predict whether the majority method will work or not? Answer on the basis of simulations and your intuition.