1. **Graphical models**

a) In class, we studied HMMs as an example of a graphical model where the graph is structured as a chain. In this problem, we will study graphical models structured as trees. Tree-structured graphical models are a natural model to represent how populations evolve. Let \( X_1 \in [0, 1] \) denote the frequency of an allele at a single SNP in population 1. Individuals from population 1 split into 2 populations, 2 and 3. The allele frequency in 2 and 3 are independent given \( X_1 \). Population 2 then splits into 4 and 5. We can represent the joint distribution of allele frequencies as a graphical model (see below) where each \( X_i \in [0, 1] \).

In practice, we can only observe data from the leaves (\( X_3, X_4 \) and \( X_5 \)) which correspond to present-day populations but not from the internal nodes which correspond to ancestral populations. One common problem in evolutionary biology is that we would like to test if the tree specified below could have generated the data (\( X_3, X_4 \) and \( X_5 \)). Show that in the model below, \( \mathbb{E}[(X_5 - X_4)X_3] = 0 \).

i. To show this, first consider the distribution over (\( X_3, X_4, X_5 \)) conditioned on \( X_1 \) and \( X_2 \). Using the conditional independence properties of this model, first show that \( \mathbb{E}[(X_5 - X_4)X_3|X_1, X_2] = 0 \). For this result, you will also need to use a property of the conditional distributions \( P(x_i|x_{\text{parent}(i)}) \) common in evolutionary biology, i.e., \( \mathbb{E}[X_i|X_{\text{parent}(i)}] = X_{\text{parent}(i)} \) (this assumption captures the idea that the average allele frequencies do not change over time).

ii. Having shown the above result, \( \mathbb{E}[(X_5 - X_4)X_3|X_1, X_2] = 0 \), we can now use the property \( \mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X] \) (known as the tower property of expectation) to obtain the final result.
2. We will apply this theory to real data from a European, an African and from an extinct population (Neanderthals). In this setting, $X_5$, $X_4$ and $X_3$ represent European, African and Neanderthal population. This data consists of 1614 SNPs (one per line) from the three populations (one per column) where for each SNP $j,j \in \{1 : 1614\}$, we observe an independent sample from the joint distribution $(X_3^j, X_4^j, X_5^j)$. We can estimate $E[(X_5 - X_4)X_3]$ by the statistic: $d = \frac{\sum_{j=1}^{1614} (x_5^j - x_4^j)x_3^j}{m}$. What is $d$ from this data?

3. We would like to test the null hypothesis $H_0 : E[(X_5 - X_4)X_3] = 0$. Under $H_0 : \hat{d} \sim N(0, \sigma^2)$ where $\sigma^2 = 4 \times 10^{-3}$ (we won’t get into the details of how we compute $\sigma$ here). Compute the p-value for $H_0$.

4. Based on the p-value, can you accept the tree that relates Europeans, Africans and Neanderthals?

5. Hidden Markov Models

   a) Consider a HMM with hidden variables $Z_{1:m}$ and observed variables $X_{1:m}$. Assume that each $Z_t \in \{1 : K\}$. In class, we discussed how the forward-backward algorithms allow us to efficiently compute the probabilities $\gamma_t(j) = P(Z_t = j|X_{1:m})$ for all $t \in \{1 : m\}$. It is also useful to be able to compute the probabilities $\psi_t(i, j) = P(Z_t = i, Z_{t+1} = j|X_{1:m})$. Show how $\psi_t$ can be computed assuming we are given all the $\alpha_t$ and $\beta_t$, $t \in \{1 : m\}$.

   b) In class, we discussed the forward algorithm. Show how we can write $\beta_{t-1}(j) = \sum_{i=1}^{K} \beta_t(i)P(Z_t = i|Z_{t-1} = j)P(X_t|Z_t = i)$.

   c) The forward-backward algorithm has computational complexity of $O(K^2)$. This does not use any further assumptions about the joint probability except the conditional independece assumptions associated with the HMM. We can reduce the computational cost if we are given additional information about the probability distribution.

   Consider a HMM with the following transition probability between time $t - 1$ and $t$. We first draw $R_t$ - a binary random variable. If $R_t = 1$, then $Z_t = Z_{t-1}$. If $R_t = 1$, then $Z_t$ is chosen independently of $Z_{t-1}$. Formally:

   $$R_t \sim \text{Ber}(p)$$
   $$Z_t|Z_{t-1}, R_t = 1 \sim \text{Unif}\{\{1, \ldots, K\}\}$$
   $$Z_t|Z_{t-1}, R_t = 0 = Z_{t-1}$$

   Write the computationally efficient recursion for the forward computation. What is the new computational complexity?

   (This HMM is used often in genetics problems such as genotype imputation or phasing. Consider the problem of genotype imputation. We have a reference panel of $K$ individuals and a test individual that we would like to impute (“fill in missing SNPs”). We will use the idea that the genome of the test individual can be obtained by copying segments of the genomes from the reference panel. However, different positions along the genome might copy from different reference individuals because of recombination.

   We formalize this problem as a HMM with $Z_t$ indicating which individual is copied into the test individual at SNP $t$. While modeling genetic sequences, we can use the idea that a recombination between SNPs at position $t - 1$ and $t$ leads to the alleles at the adjacent SNPs choose the individual to copy from independently. If there is no recombination, then
they choose the same individual. This property is referred to as context-specific independence because the independence of $Z_t$ and $Z_{t-1}$ depends on whether or not recombination occurred. Here $R_t = 1$ indicates that recombination occurred and the probability of recombination is $p$.