1. **EM algorithm**

We will define a mixture model for genetic data and propose an EM algorithm for inference in this model.

We observe data at $m$ SNPs across $n$ individuals. There are $K$ ancestral populations. Each individual $i$ is associated with a single population denoted by the random variable $Z_i$ where $Z_i$ is a binary vector with a single one in the entry corresponding to the population that individual $i$ belongs to. The population membership $Z_i$ is chosen according to a vector of mixture proportions $\pi, \pi_k \geq 0, \sum_k \pi_k = 1$. For each SNP $j$ in individual $i$, $X_{i,j}$ denotes the genotype (0 or 1) at SNP $j$ in individual $i$ (we simplify this model so that each individual has a genotype that takes two values instead of three values in most datasets). If individual $i$ belongs to population $k$, $X_{i,j}$ are drawn according to a Bernoulli distribution with parameter $f_{j,k}$.

We can write our model as:

$$Z_i|\pi \sim \text{iid Mult}(\pi)$$

$$X_{i,j}|Z_i = k, f_{j,k} \sim \text{Ber}(f_{j,k})$$

Our goal is to estimate the parameters $\theta = (\pi, f_j)$ and to infer the population labels of each individual $Z_i$.

a) The complete data log likelihood for this model can be written as

$$\sum_{i=1}^{n} \log P(x_i,z_i|\theta) = \sum_{i=1}^{n} \log \left( \prod_{k=1}^{K} (\pi_k) \prod_{j=1}^{m} P(x_{i,j}|f_{j,k}) \right)^{z_{i,k}}$$

Write the expected complete data log-likelihood: $Q(\theta; \theta^{(t)})$. Express your answer in terms of $r_{i,k} = P(Z_i = k|x_i,1:m, \theta^{(t)})$.

b) We would like to obtain the MLE of $\theta$. Derive the M-step updates for $\pi$ and $f_{j,k}$ in terms of $r_{i,k}$.

2. **Coding**: You are given two datasets each of which contains genotype data from 1000 individuals at 5000 SNPs. Dataset 1 contains the values of the hidden values and parameters that you can use to debug.

For this question, you will implement an EM algorithm and obtain MLE of the parameters of the mixture model for each dataset as well as infer the hidden variables. The first dataset...
(dataset 1) includes the true values of the $\theta$ and $Z$ and can be used to test your implementation. The second dataset (dataset 2) does not include the true parameter values.

Use the following specifications for the EM.

- **Initialization**: To initialize, draw each $f_{j,k}$ from a uniform random variable. Draw each $\pi_k$ from an exponential random variable and normalize (remember that $\pi_k$ needs to sum to 1). Alternately, draw $\pi$ from a Dirichlet distribution.
- **Stopping criterion**: Run for atmost 100 iterations or till the change in log likelihood $< 10^{-8}$.
- **Number of random restarts**: Since EM only reaches a local optimum, we often run the algorithm with multiple random initializations and pick the solution with the highest log likelihood. We use 3 restarts.
- **The number of mixture components or populations $K$**: Set $K = 2$.

a) For dataset 1, plot the log likelihood as a function of iteration for a single run of EM (i.e. for a single initialization of EM).

b) For dataset 1, what is the MLE of $\pi$?

c) For dataset 1, fixing $\pi$ and $f$ at its MLE from the previous question, we can compute the posterior probability of $Z_i$ for each individual. Note that these posterior probabilities are computed in the last E-step of the EM algorithm.

We would like to compute the accuracy of the inferred $Z_i$ by comparing the posterior probability of $Z_i$ to the true $Z_i$ available in mixture1.ganc file (each row of this file indicates which population the corresponding individual belongs to). The inferred population label for individual $i$ is set to the MAP (maximum a posteriori) estimate, i.e., choose the population with the maximum value of the posterior probability. Our measure of accuracy is simply the fraction of individuals for which the MAP estimate matches the true population label.

On dataset 1, What is the accuracy of the inferred population label?

d) How similar are the solutions on dataset 1 across the different initializations. Report the log likelihoods of the optimum found as well as the accuracy of the population label.

e) To see how accuracy of predicting population label 1 changes as we get more SNPs, redo the EM using only the first 10, 100, 1000 SNPs and 5000 SNPs. Plot accuracy as a function of number of SNPs.

f) For dataset 2, run EM and report the MLE of $\pi$.

g) For dataset 2, plot how the log likelihood varies with $K$ (try $K = 1, \ldots, 4$). Which value of $K$ would you choose based on this plot?

**Implementation details**: The fixed point to which EM algorithm converges is not guaranteed to be the global optimum. To obtain a better solution, it is recommended to start the algorithm at multiple random points in the parameter space and pick the theta with the maximum likelihood across the random restarts.

The log likelihood is guaranteed to be non-decreasing in each iteration and is a useful check of your implementation.

You will need to compute the posterior probabilities $r_{i,k}$ by applying Bayes theorem. One difficulty is that the likelihood, $P(x_{i,1:m}|z_i = k, \theta)$ can become very small when the number of
SNPs is large leading to underflow errors. One solution is to work with the log likelihood and to notice that the posterior probability is invariant if we re-scale each likelihood by the same amount. Let \( l_i^* = \max_{k'} l_{i, k'} \). We can then write

\[
    r_{i, k} = \frac{\exp(l_{i, k})}{\sum_{k'} \exp(l_{i, k'})} = \frac{\exp(l_{i, k} - l_i^*)}{\sum_{k'} \exp(l_{i, k'} - l_i^*)}
\]

The second line is better behaved in practice.