1: DNA Sequencing and Gene Finding (50 + 15 Pts)

DNA is a substance that carries genetic information in the cells of the body. It made from a long sequence of four bases called A, C, G and T. DNA sequence consists of two region types: gene and non-gene (Fig. 1). The gene regions play an important role in production of proteins in the body so that many genetic diseases are related to these regions.

One of the most important problems in genetics is to detect the gene regions. Generally, the statistical pattern of bases differs in gene and non-gene regions, for example, bases C and G are more frequent in gene regions. Thus we can detect the regions by exploiting these statistical patterns. HMM is the most popular method for this purpose. In this Method, the DNA bases are considered as observation sequence of the HMM and there are two possible values (i.e. gene and non-gene) for each hidden state.

Sometimes there is no direct access to DNA sequence for detecting gene regions, thus we should obtain the sequence at first and then by further processing detect the gene regions. There are several technologies for DNA sequencing. One of the modern methods for DNA sequencing is Nanopore. In this technology for each base we receive an electrical signal that determines type of that base. Following figure shows a sample of received signals for a short DNA sequence. Since this signal is continuous, we can use a GMM with four components as its distribution.

In this problem we want, by observing the Nanopore signals, to detect gene and non-gene regions of a DNA and also obtain the most probable base sequence. For this purpose we will use a model called tied-Mixture HMM.

Tied-mixture HMM is used for modeling continuous observations with discrete hidden states. In this model, observations are modeled by a GMM as follow:

\[ p(x_t \mid z_t = i, \theta) = \sum_{k=1}^{K} w_{ik} N(x_t \mid \mu_k, \Sigma_k), \quad i = 1, \ldots, L; \quad t = 1, \ldots, T \]  

(1)

where \( x_t \) and \( z_t \) are observed and hidden variables at time \( t \), respectively. \( \theta \) consists all the model
parameters that are listed below:

\[ \Pi = [\pi_1, \ldots, \pi_L] \]
\[ \pi_i = p(z_1 = i), \quad i = 1, \ldots, L \]
\[ W = \{w_{ik}\}, \quad i = 1, \ldots, L; \quad k = 1, \ldots, K \]
\[ a_{ij} = p(z_t = j \mid z_{t-1} = i), \quad i = 1, \ldots, L; \quad j = 1, \ldots, L \]
\[ B = \{\mu_k, \Sigma_k\}, \quad k = 1, \ldots, K \]

Note that in this model we use \( K \) Gaussian components for all observations and only the weights of these components, \( w_{ik} \)'s, are different for each hidden state value.

In our sequencing problem, \( T \) is the DNA length, \( z_t \) determines region type (gene or non-gene) of \( t \)-th base, \( x_t \) is the received electrical signal for \( t \)-th base, \( K = 4 \) and \( L = 2 \). In this model we denote the value of \( t \)-th base by \( y_t \) which is called the mixture variable of the model.

1. (5 Pts) Draw the graphical model.

2. (15 Pts) Let we are given \( S \) sequences \( \{x_1^s, \ldots, x_T^s\}_{s=1}^S \). Derive E and M steps of EM-algorithm for estimating parameters of the model.

(a) E-Step: Write the total log-likelihood \( \log P(z, y, x \mid \theta^{(l)}) \) and derive its expectation

\[ Q(\theta \mid \theta^{(l)}) = \mathbb{E}\{\log P(z, y, x \mid \theta) \mid x, \theta^{(l)}\} \]

Show that the expected sufficient statistics are:

\[ \mathbb{E}\{\mathbb{I}(y_t^s = k) \mid x^s, \theta^{(l)}\} \]
\[ \mathbb{E}\{\mathbb{I}(y_t^s = k)\mathbb{I}(z_t^s = i) \mid x^s, \theta^{(l)}\} \]
\[ \mathbb{E}\{\mathbb{I}(z_t^s = i) \mid x^s, \theta^{(l)}\} \]
\[ \mathbb{E}\{\mathbb{I}(z_{t-1}^s = i)\mathbb{I}(z_t^s = j) \mid x^s, \theta^{(l)}\} \]

where \( \mathbb{I}(.) \) is a binary indicator function.

(b) M-Step: Take derivative from \( Q(\theta \mid \theta^{(l)}) \) with respect to all the parameters and show that the update equations will be:
We have learned about factor analysis in the class. One of the successful paradigms in speaker verification is based on this model. Speaker verification is the process to accept or reject an identity claim by comparing two speech samples: one that is used as a reference of the identity (speaker model) and the other that is collected during the test from the person who makes the claim. However, the performance of speaker verification suffers from the mismatch between the reference speech and the test speech. This mismatch can occur due to e.g., environmental noise, different microphones, channel-effects and within-speaker variabilities such as ageing. In order to establish a robust speaker verification system we need to compensate these variabilities which are
usually called channel variability. In the speaker verification based on factor analysis the aim is to reduce the dimensionality of the feature vector (speaker model) as well as compensating the channel variability. In this model, instead of one latent factor in traditional factor analysis we have two latent factors corresponding to speaker and channel subspaces. Therefore, after using the model we have two low dimensional vectors that contain information about speaker and channel and in this way, we can separate channel variability from speaker information.

We assume that the training data contains I utterances from J speakers and denote the $i^{th}$ utterance of the $j^{th}$ speaker by $x_{ij}$. The following Equation shows the data generation process using this model:

$$x_{ij} = \mu + Fs_i + Gc_{ij} + \epsilon_{ij}$$

where $\mu$ represents the overall mean of the training dataset and $F$ and $G$ are matrices containing a basis for speaker and channel subspaces respectively. $s_i$ is the latent factor for speaker that is identical for all the utterances from a speaker. $c_{ij}$ is the channel factor corresponding to the $i^{th}$ utterance of the $j^{th}$ speaker. Remaining unexplained data variation is explained by the residual noise term $\epsilon_{ij}$ which is defined to be Gaussian with diagonal covariance $\Sigma$. The speaker and channel latent factors are also assumed to have Gaussian distribution with zero mean and identity covariance matrix.

1. (8 Pts) Draw the probabilistic graphical model corresponding this method.

To train this model we need to take a set of data points $x_{ij}$ and find the parameter set $\Theta = \{\mu, \Sigma, F, G\}$. This would be easy if we knew the values of the latent variables $s_i$ and $c_{ij}$. Likewise it would be easy to estimate $s_i$ and $c_{ij}$ given $\Theta$. Unfortunately, none of these variables are known. The Expectation Maximization (EM) algorithm alternately estimates the two sets of parameters in such a way that the likelihood is guaranteed to increase at each iteration. More specifically, in the E-Step, we calculate the posterior distribution over the latent variables $s_i$ and $c_{ij}$ for fixed parameter values. In the M-Step, we optimize point estimates of the parameters $\Theta = \{\mu, \Sigma, F, G\}$. To simplify this step we can define the concatenation of the two subspace matrices $F$ and $G$ as $B = [F \ G]$ and the concatenation of the two factor loading vectors as $z_{ij} = \left[ \begin{array}{c} s_i \\ c_{ij} \end{array} \right]$. So the above equation is written as:

$$x_{ij} = \mu + [F \ G] \left[ \begin{array}{c} s_i \\ c_{ij} \end{array} \right] + \epsilon_{ij}$$

2. (12 Pts) Derive the log of the complete likelihood and show that the parameter update rules for the M-step will be:

$$\mu = \frac{1}{IJ} \sum_{i,j} x_{ij}$$

$$B = \left( \sum_{i,j} (x_{ij} - \mu) E[z_i^T] \right) \left( \sum_{i,j} E[z_i z_i^T] \right)^{-1}$$

$$\Sigma = \frac{1}{IJ} \sum_{i,j} \text{Diag} \left[ (x_{ij} - \mu) (x_{ij} - \mu)^T - BE[z_i] (x_{ij} - \mu)^T \right]$$
3: Conditional Random Fields (15 Pts)

In this problem we aim to use CRF for the text tagging problem. We want to extract the title, authors, year, and conference names from a researchers Web page. Suppose we use the graphical model in Fig. 2. \( x_{1:T} \) are observations and \( y_{1:T} \) are hidden labels and the scalar \( \lambda_i \) is the weight for feature \( f_i(.) \). Suppose \( D = [(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), ..., (x^{(N)}, y^{(N)})] \) are i.i.d. training samples.

1. (4 Pts) What is the conditional log-Likelihood of the model?
2. (4 Pts) Define two feature functions for this model.
3. (7 Pts) Find the update equation for the parameter \( \lambda_i \) according to gradient descent algorithm.

4: Kalman Filters (15 Pts)

Consider the below Kalman filter

\[
Y_t = HX_t + V_t, V_t \sim N(0, R)
\]
\[
X_{t+1} = FX_t + GU_t, U_t \sim N(0, Q)
\]

and

\[
X_0 \sim N(0, \Sigma_0)
\]

in which \( U_t \) and \( V_t \) are correlated for the same \( t \) with covariance \( E[U_t V_t^T] = C \), and \( U_t \) and \( V_t \) are mutually independent for all different values of \( t \).

Let define

\[
\hat{X}_{t|t} = E[X_t|Y_{0:t}]
\]
\[
\Sigma_{t|t} = Cov[X_t|Y_{0:t}] = E[(X_t - \hat{X}_{t|t})(X_t - \hat{X}_{t|t})^T|Y_{0:t}]
\]

1. (7.5 Pts) Find \( \hat{X}_{t+1|t} \) in terms of \( \hat{X}_{t|t} \), \( \hat{X}_{t|t-1} \), \( \Sigma_{t|t} \).
2. (7.5 Pts) Find \( \Sigma_{t+1|t} \) in terms of \( \Sigma_{t|t} \), \( \Sigma_{t|t-1} \).