5.1 Review of Markov chains

We fix an alphabet $\Sigma$ with $|\Sigma| = l$ and the length of the chain $n$. The Markov chain model is parameterized by the set of transition matrices

$$\Theta := \{ \theta \in \mathbb{R}^{l \times l} | \text{ row sums} = 1 \}.$$

A Markov chain is defined by the following map:

$$f_{l,n} : \Theta \to \mathbb{R}^{l \times n}, \quad \theta \mapsto (p_{1,1}, \ldots, p_{l,n}),$$

where

$$p_{i_1, i_2, \ldots, i_n} = \theta_{i_1 i_2} \cdot \theta_{i_2 i_3} \cdot \ldots \cdot \theta_{i_{n-1} i_n}.$$

**Remark:** The corresponding laminar family is of the following form:

$$1 \to 2 \to 3 \to \cdots \to n$$

**Example:** For $l = 2$ and $n = 4$ the parameter space $\Theta$ is a square, namely

$$\Theta = \left\{ \begin{pmatrix} \theta_0 & 1 - \theta_0 \\ 1 - \theta_1 & \theta_1 \end{pmatrix} \in \mathbb{R}^{2 \times 2} \right\}.$$

The Markov chain model is the image of $\Theta$ under the map $f_{2,4}$.

5.1.1 Maximum likelihood estimation

We next discuss maximum likelihood estimation for Markov chains. Fix a data vector $u \in \mathbb{N}^n$. A Markov chain model $f_{l,n}$ is defined by an $(l^2 \times l^n)$-matrix $A_{l,n}$. The sufficient statistic $v = A_{l,n} \cdot u \in \mathbb{N}^{l^2}$ is regarded as an $l \times l$ matrix. The entry $v_{ij}$ denotes the number of occurrences of $ij$ as a consecutive pair in the observed sequences.

**Proposition 5.1.** The maximum likelihood estimate for the data $u \in \Sigma^n$ is given by

$$\hat{\theta}_{ij} = \frac{v_{ij}}{\sum_{s \in \Sigma} v_{is}}, \quad i, j \in \Sigma.$$
Proof: Compute the log-likelihood function:
\[
l(\theta) = \sum_{i \in \Sigma} \left( v_{i1} \log(\theta_{i1}) + v_{i2} \log(\theta_{i2}) + \cdots + v_{i,l-1} \log(\theta_{i,l-1}) + v_{il} \log(1 - \sum_{s=1}^{l-1} \theta_{is}) \right).
\]
The result follows by equating the partial derivatives of \(l(\theta)\) to zero. \(\square\)

5.2 Fully observed Markov models

We fix a first alphabet \(\Sigma'\) with \(|\Sigma'| = l'\), a second alphabet \(\Sigma''\) with \(|\Sigma''| = l''\) and the length of the chain \(n\). The model is parameterized by the set
\[
\Theta := \{(\theta', \theta'') \in \mathbb{R}^{l' \times l'} \times \mathbb{R}^{l' \times l''} \mid \text{row sums of } \theta', \theta'' = 1\}.
\]

Definition 5.2. A fully observed Markov model is defined by the following map:
\[
F : \Theta \to \mathbb{R}^{(l')^n \times (l'')^n}, \quad \theta = (\theta', \theta'') \mapsto (p_{\sigma,\tau}), \quad \text{where}
\]
\[
p_{\sigma,\tau} = \theta''_{\sigma_1 \tau_1} \cdot \theta'_{\sigma_1 \sigma_2} \cdot \theta''_{\sigma_2 \sigma_3} \cdots \theta''_{\sigma_n \tau_n}.
\]

Remark: The corresponding laminar family is of the following form:

Example: For \(\Sigma' = \{0, 1\}\) and \(\Sigma'' = \{A,C,G,T\}\) the parameter space \(\Theta\) is
\[
\Theta := \{(\theta', \theta'') \in \mathbb{R}^{2 \times 2} \times \mathbb{R}^{2 \times 4} \mid \text{row sums of } \theta', \theta'' = 1\}.
\]

For example,
\[
p_{0110, AGCA} = \theta''_{01} \cdot \theta'_{01} \cdot \theta''_{10} \cdot \theta'_{10} \cdot \theta''_{0A}.
\]

Remark: The fully observed Markov model is a log-linear model. The maximum likelihood estimate for the data \(u \in \mathbb{N}^{(l')^n \times (l'')^n}\) is given by
\[
\hat{\theta}'_{ij} = \frac{v'_{ij}}{\sum_{s \in \Sigma'} v'_{is}} \quad \text{and} \quad \hat{\theta}''_{ij} = \frac{v''_{ij}}{\sum_{l \in \Sigma''} v''_{il}}.
\]
The proof is analogous to the proof of Proposition 5.1.
5.3 Hidden Markov models

The hidden Markov model $f$ is derived from the fully observed Markov model $F$ by summing out the first indices $\sigma \in (\Sigma')^n$. Consider the marginalization map

$$\rho : \mathbb{R}^{(l')^n \times (l'')^n} \rightarrow \mathbb{R}^{(l'')^n}$$

obtained by taking the column sums of an $(l')^n \times (l'')^n$-matrix. The hidden Markov model is defined by composing the fully observed Markov model $F$ with the marginalization map $\rho$:

$$f = \rho \circ F : \Theta \rightarrow \mathbb{R}^{(l'')^n}, \quad \theta = (\theta', \theta'') \mapsto (p_\tau), \quad \text{where} \quad p_\tau = \sum_{\sigma \in \Sigma')^n} p_{\sigma, \tau}.$$

Remark: The corresponding laminar family is of the following form:

Example: For $\Sigma' = \{0, 1\}$ and $\Sigma'' = \{A, C, G, T\}$ we get for example

$$p_{AGCA} = \sum_{\sigma \in \{0, 1\}^4} p_{\sigma, AGCA}.$$

Remark: The hidden Markov model is not a log-linear model.

5.3.1 Defining the ML problem

Data: Our data consist of a sequence of observations $i_1, i_2 \ldots i_t$ where each $i_j$ represents a sequence of the observed states of the HMM model, each of length $n$. We define $u \in \mathbb{N}^{\Sigma'|^N}$ where $u_k$ is the number of indices $j \in [t]$ such that $i_j = k$. The vector $u$ is a sufficient statistic of the HMM model.

For $\theta = (\theta', \theta'')$, $i \in |\Sigma''|^n = N$ and $j \in |\Sigma'|^n = M$, $f_{ij}(\theta)$ is the probability of an observed sequence $i$ and a hidden sequence $j$ given $\theta$.

We define

$$f_i(\theta) := \sum_{j=1}^M f_{ij}(\theta).$$

5-3
Given a vector of (observed) data $u \in \mathbb{N}^N$, the ML problem for the HMM model consists of finding $\theta$ that maximizes the likelihood function:

$$L_f(\theta) = f_1(\theta)^{u_1} \cdot f_2(\theta)^{u_2} \cdots f_N(\theta)^{u_N}. \quad (5.1)$$

As the different $f_i(\theta)$'s involve summations, finding the maximizing $\theta$ does not become simpler when taking the log of this equation.

**Definition 5.3.** Given a vector $u \in \mathbb{N}^N$ and an HMM model $\theta \in \Theta$, define the expected hidden data matrix $U = (u_{ij}) \in \mathbb{R}^{N \times M}$ by

$$u_{ij} = u_i \cdot \frac{f_{ij}(\theta)}{\sum_{j=1}^{M} f_{ij}(\theta)} = \frac{u_i}{f_i(\theta)} \cdot f_{ij}(\theta).$$

Given a hidden data matrix, we assume knowledge of all the states of the HMM, both observed and hidden. Given a matrix $U$ we can define the full ML problem for the HMM model as maximizing

$$L_F(\theta) = f_{11}(\theta)^{u_{11}} \cdot f_{12}(\theta)^{u_{12}} \cdots f_{n1}(\theta)^{u_{n1}} \cdots f_{nm}(\theta)^{u_{nm}}. \quad (5.2)$$

For this equation finding the maximizing solution is easy.

We define $l_f(\theta)$ and $l_F(\theta)$ to be the log-likelihood functions of $L_f(\theta)$ and $L_F(\theta)$ respectively.

As we want to estimate $\theta$ with data consisting of the observed states only, we use the EM algorithm.

### 5.4 The EM Algorithm

We make use of the EM (Expectation Maximization) algorithm to find a solution for the ML problem of HMMs, i.e. maximizing $l_f(\theta)$. However, the solution achieved by using this method is promised to be a critical point of $l_f(\theta)$ but is not necessarily the global maximum. The algorithm uses the fact that solving the ML problem for $L_F(\theta)$ is easy.

*The EM Algorithm:*

Input: An $N \times M$ matrix of polynomials $f_{ij}(\theta)$ representing the hidden model $F$ and a vector of observed data $u \in \mathbb{N}^N$

Output: A proposed maximum $\hat{\theta} \in \Theta$ of the log-likelihood function $l_f(\theta)$

0. Pick a threshold $\epsilon > 0$.

1. Initialize $\theta$ by choosing a random $\theta \in \Theta$.

2. **E-Step:** Compute the expected hidden data matrix, $U$, using $u$ and $\theta$. 
3. M-Step: Compute the solution $\theta^* \in \Theta$ of the full ML problem for the model $F = (f_{ij})$ by maximizing $l_F(\theta)$.

4. If $l_f(\theta) - l_f(\theta^*) > \epsilon$ let $\theta = \theta^*$ and go to step (2). Otherwise, output $\hat{\theta} := \theta^*$.

**Theorem 5.4.** The values of the likelihood function $l_f(\theta)$ weakly increase during each step of the EM algorithm, establishing $l_f(\theta) \leq l_f(\theta^*)$. Furthermore, if $l_f(\theta) = l_f(\theta^*)$ then $\theta^*$ is a critical point of the likelihood function $l_f(\theta)$.

**Remark:** Notice that the theorem does not guarantee any rate of convergence.

**Proof:** First, notice that for any positive number $x$

$$\log(x) \leq x - 1 \quad \text{with equality iff} \quad x = 1. \quad (5.3)$$

Observing the difference between the values of the likelihood function $l_f(\theta)$ for $\theta$ and $\theta^*$ we obtain

$$l_f(\theta^*) - l_f(\theta) = \sum_{i=1}^{N} u_i [\log(f_i(\theta^*)) - \log(f_i(\theta))] \quad (5.4)$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{M} u_{ij} [\log(f_{ij}(\theta^*)) - \log(f_{ij}(\theta))] \quad (5.5)$$

$$= \sum_{i=1}^{N} u_i \cdot \left[ \log \left( \frac{f_i(\theta^*)}{f_i(\theta)} \right) - \sum_{j=1}^{M} \frac{u_{ij}}{u_i} \cdot \log \left( \frac{f_{ij}(\theta^*)}{f_{ij}(\theta)} \right) \right].$$

Equation (5.5) is achieved by adding and subtracting $\sum_{j=1}^{M} u_{ij} \cdot \log \left( \frac{f_{ij}(\theta^*)}{f_{ij}(\theta)} \right)$. We need to show that this difference is non-negative. First, notice that (1) = $l_F(\theta^*) - l_F(\theta) \geq 0$. This is true since the algorithm chooses $\theta^*$ to be the global maximum of $l_F(\theta)$ for a given matrix $U$. To prove that the whole expression is non-negative it suffices to show that (2) is non-negative.

$$\left(2\right) = \log \left( \frac{f_i(\theta^*)}{f_i(\theta)} \right) - \sum_{j=1}^{M} \frac{u_{ij}}{u_i} \cdot \log \left( \frac{f_{ij}(\theta^*)}{f_{ij}(\theta)} \right) \quad (5.6)$$

$$= \log \left( \frac{f_i(\theta^*)}{f_i(\theta)} \right) + \sum_{j=1}^{M} \frac{f_{ij}(\theta)}{f_i(\theta)} \cdot \log \left( \frac{f_{ij}(\theta)}{f_{ij}(\theta^*)} \right) \quad (5.7)$$

$$= \sum_{j=1}^{M} \frac{f_{ij}(\theta)}{f_i(\theta)} \cdot \log \left( \frac{f_i(\theta^*)}{f_i(\theta)} \right) + \sum_{j=1}^{M} \frac{f_{ij}(\theta)}{f_{ij}(\theta^*)} \cdot \log \left( \frac{f_{ij}(\theta)}{f_{ij}(\theta^*)} \right) \quad (5.8)$$

$$= \sum_{j=1}^{M} \frac{f_{ij}(\theta)}{f_i(\theta)} \cdot \log \left( \frac{f_i(\theta^*)}{f_{ij}(\theta^*)} \cdot \frac{f_{ij}(\theta)}{f_i(\theta)} \right) \quad (5.9)$$
The expression (5.7) is achieved as 
\[
\frac{u_{ij}}{u_i} := \frac{u_i f_{ij}(\theta)}{u_i f_i(\theta)},
\]
(5.8) is achieved as \( \sum_{j=1}^{M} \frac{f_{ij}(\theta)}{f_i(\theta)} = 1 \) and the argument by which it is multiplied has no instances of \( j \). To show (5.9) is non-negative, we define \( \pi_j := \frac{f_{ij}(\theta)}{f_i(\theta)} \) and \( \sigma_j := \frac{f_{ij}(\theta^*)}{f_i(\theta^*)} \) and notice that these are non-negative quantities and moreover, \( \pi_1 + \ldots + \pi_M = \sigma_1 + \ldots + \sigma_M = 1 \). The vectors \( \pi \in \mathbb{R}^M \) and \( \sigma \in \mathbb{R}^M \) can thus be regarded as probability distributions, and equation (5.9) equals the Kullback-Leibler distance, defined for probability distributions \( p \) and \( q \) as 
\[
H(p||q) := - \sum_{i=1}^{k} p_i \cdot \log \left( \frac{q_i}{p_i} \right).
\]
Therefore we have
\[
(5.9) = H(\pi||\sigma) = - \sum_{j=1}^{M} \pi_j \cdot \log \left( \frac{\sigma_j}{\pi_j} \right) \geq \sum_{j=1}^{M} (-\pi_j) \cdot \left( 1 - \frac{\sigma_j}{\pi_j} \right) = 0. 
\]
(5.10)
The inequality follows from (5.3).

To complete the proof it remains to show that if \( l_f(\theta) = l_f(\theta^*) \), then \( \theta^* \) is a critical point of \( l_f(\theta) \). This was left as a homework assignment.

\[\square\]

### 5.5 Homework

Complete the proof of Theorem 5.4 - Show that if \( l_f(\theta) = l_f(\theta^*) \) then \( \theta^* \) is a critical point of \( l_f(\theta) \).