These are exercises for self-study and exam preparation. All material is examinable unless otherwise mentioned.

Exercise 1. ***Baum-Welch Algorithm***

This question is on the EM algorithm for discrete-valued Hidden Markov Models (HMMs). We assume that the transition and emission distributions of the HMM are parameterised by the matrices $\mathbf{A}$ and $\mathbf{B}$, respectively, with elements

$$p(h_i = k | h_{i-1} = k'; \mathbf{A}) = A_{k,k'}, \qquad p(v_i = m | h_i = k; \mathbf{B}) = B_{m,k}, \tag{1}$$

and the initial state distribution is parameterised by the vector $\mathbf{a}$, with elements

$$p(h_1 = k; \mathbf{a}) = a_k, \tag{2}$$

where $h_i \in \{1, \ldots, K\}$ denotes the hidden (unobserved) and $v_i \in \{1, \ldots, M\}$ the visible (observed) variables.

We assume that we are given n independent sequences (time-series) $\mathcal{D}_1, \ldots, \mathcal{D}_n$, each containing the values of the visibles. The length of sequence $\mathcal{D}_j$ is d_j.

In the course, we have seen that the parameters $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{a}$ can be learned from the data $\mathcal{D}_1, \ldots, \mathcal{D}_n$ by means of the EM (the Baum-Welch) algorithm.

(a) A friend produces the following pseudo-code for one iteration of the EM algorithm, where they use $\boldsymbol{\theta}$ to denote the values of $\mathbf{A}, \mathbf{B}$, and $\mathbf{a}$ from the previous iteration. However, the pseudo-code contains several mistakes. Find and correct them.

Step 1: For each sequence $\mathcal{D}_j$ compute the posteriors $p(h_i, h_{i-1} \mid \mathcal{D}_j; \boldsymbol{\theta})$ and $p(h_i \mid \mathcal{D}_j; \boldsymbol{\theta})$ using the alpha recursion.

Step 2: Update the parameters as follows

$$a_k \leftarrow \frac{1}{n} \sum_{j=1}^{n} p(h_1 = k | \mathcal{D}_j; \boldsymbol{\theta})$$

$$A_{k,k'} \leftarrow \frac{\sum_{j=1}^{n} \sum_{i=2}^{d_j} p(h_i = k, h_{i-1} = k' | \mathcal{D}_j; \boldsymbol{\theta})}{\sum_{j=1}^{n} \sum_{i=2}^{d_j} p(h_i = k, h_{i-1} = k' | \mathcal{D}_j; \boldsymbol{\theta})}$$

$$B_{m,k} \leftarrow \frac{\sum_{j=1}^{n} \sum_{i=1}^{d_j} p(h_i = k | \mathcal{D}_j; \boldsymbol{\theta})}{\sum_{j=1}^{n} \sum_{i=1}^{d_j} p(h_i = k | \mathcal{D}_j; \boldsymbol{\theta})}$$

(b) In the update of the matrices $\mathbf{A}$ and $\mathbf{B}$, we sum over the length of the sequences d_j. What model assumption is the reason for this summation? What potential advantage does the summation have for the estimation of the parameters?