GRAPHICAL MARKOV MODELS (WS2018)
4. SEMINAR

Assignment 1. Consider the following probabilistic model for real valued sequences $\boldsymbol{x} = (x_1, \dots, x_n)$, $x_i \in \mathbb{R}$ of fixed length n. Each sequence is a combination of a leading part $i \leqslant k$ and a trailing part $i > k$. The boundary $k = 1, \dots, n$ is random with some categorical distribution $\boldsymbol{\pi} \in \mathbb{R}^n_+$, $\sum_k \pi_k = 1$. The values x_i, in the leading and trailing part are statistically independent and distributed with some probability density function $p_1(x)$ and $p_2(x)$ respectively. Altogether the distribution for pairs $(\boldsymbol{x}, k)$ reads

$$p(\boldsymbol{x}, k) = \pi_k \prod_{i=1}^{k} p_1(x_i) \prod_{j=k+1}^{n} p_2(x_j). \tag{1}$$

The densities p_1 and p_2 are known. Given an i.i.d. sample of sequences $\mathcal{T}^m = \{\boldsymbol{x}^\ell \in \mathbb{R}^n \mid \ell = 1, \dots, m\}$, the task is to estimate the unknown boundary distribution $\boldsymbol{\pi}$ by the EM-algorithm.

a) The E-step of the algorithm requires to compute the values of auxiliary variables $\alpha^{(t)}_\ell(k) = p(k \mid \boldsymbol{x}^\ell)$ for each example $\boldsymbol{x}^\ell$ given the current estimate $\boldsymbol{\pi}^{(t)}$ of the boundary distribution. Give a formula for computing these values from model (1).

b) The M-step requires to solve the optimisation problem

$$\frac{1}{m} \sum_{\ell=1}^{m} \sum_{k=1}^{n} \alpha^{(t)}_\ell(k) \log p(\boldsymbol{x}^\ell, k) \to \max_{\boldsymbol{\pi}}.$$

Substitute the model (1) and solve the optimisation task.

Assignment 2. (breakpoint detection) Consider the following probabilistic model for real valued sequences $\boldsymbol{x} = (x_1, \dots, x_n)$, $x_i \in \mathbb{R}$ of fixed length n. Each sequence is a combination of a leading part $i \leqslant k$ and a trailing part $i > k$. The boundary $k = 1, \dots, n$ is random with some categorical distribution $\boldsymbol{\pi} \in \mathbb{R}^n_+$, $\sum_k \pi_k = 1$. The p.d.s for the leading and trailing parts of the sequence arise from two homogeneous HMM models:

$$p(x_{1:k}) = \sum_{s_{1:k}} p_1(x_{1:k}, s_{1:k}) \;\text{ and }\; p(x_{k+1:n}) = \sum_{s_{k+1:n}} p_2(x_{k+1:n}, s_{k+1:n})$$

The HMMs p_1 and p_2 and the distribution $\boldsymbol{\pi}$ are known. Find an algorithm for inferring the boundary k for a given sequence $\boldsymbol{x}$, assuming that the loss function is $\ell(k, k') = (k - k')^2$.

Assignment 3. Let $s = (s_1, \dots, s_n)$, be a sequence of K-valued random variables. Suppose that $v_i(k, k')$, $i = 2, \dots, n$, $k, k' \in K$ is a system of pairwise probabilities associated with consecutive pairs s_{i-1}, s_i. Consider the set $\mathcal{P}(\boldsymbol{v})$ of all joint probability distributions $p(s)$, which have $\boldsymbol{v}$ as pairwise marginals, i.e.

$$\sum_{s \in K^n} p(s) \delta_{s_{i-1}k} \delta_{s_i k'} = v_i(k, k') \quad \forall i = 2, \dots, n, \; \forall k, k' \in K.$$

We want to find the distribution with highest entropy

$$H(p) = -\sum_{s \in K^n} p(s) \log p(s)$$

in $\mathcal{P}(\boldsymbol{v})$. Prove that the unique maximiser is the Markov chain model defined by the pairwise marginals $\boldsymbol{v}$.
Hint: Formulate and solve the constrained optimisation task by using its Lagrange function.