GRAPHICAL MARKOV MODELS (WS2017)
4. SEMINAR

Assignment 1. Consider the following probabilistic model for real valued sequences $\boldsymbol{x} = (x_1, \ldots, x_n)$, $x_i \in \mathbb{R}$ of fixed length n. Each sequence is a combination of a leading part $i \leqslant k$ and a trailing part $i > k$. The boundary $k = 1, \ldots, n$ 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 $\mathcal{T}^m = \{\boldsymbol{x}^\ell \in \mathbb{R}^n \mid \ell = 1, \ldots, m\}$, 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 $\alpha_\ell^{(t)}(k) = p(k \mid \boldsymbol{x}^\ell)$ 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

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

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

Assignment 2. Prove equivalence of Definitions 1a and 1b given in the lecture for a Markov model on a tree (see Sec. 10).

Assignment 3. Let $T = (V, E)$ be an undirected tree and s_i, $i \in V$ be a field of K-valued random variables. Suppose that $v_{ij}(k, k')$, $k, k' \in K$ is a system of pairwise probabilities associated with the edges $\{i, j\} \in E$ of the tree. Consider the set $\mathcal{P}(\boldsymbol{v})$ of all joint probability distributions $p(s)$, which have $\boldsymbol{v}$ as pairwise marginals, i.e.

$$\sum_{s \in K^{|V|}} p(s) \delta_{s_i k} \delta_{s_j k'} = v_{ij}(k, k') \quad \forall \{i, j\} \in E,\ \forall k, k' \in K.$$

We want to find the distribution wit highest entropy

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

in $\mathcal{P}(\boldsymbol{v})$. Prove that the unique maximiser is the Markov model on the tree T defined by the edge marginals $\boldsymbol{v}$.

Hint: Formulate and solve the constrained optimisation task by using its Lagrange function.