9. Unsupervised learning: EM algorithm for HMMs

Given: i.i.d. training data $\mathcal{T} = \{x^j \in F^n \,|\, j = 1, \dots, m\}$

Task: $$u^* \in \operatorname*{argmax}_u \frac{1}{|\mathcal{T}|} \sum_{x \in \mathcal{T}}' \log \sum_{s \in K^n}' p_u(x,s)$$

Recall EM-algorithm

$$L(u) = \frac{1}{|\mathcal{T}|} \sum_{x \in \mathcal{T}}' \log \sum_{s \in K^n}' p_u(x,s) =$$

$$= \frac{1}{|\mathcal{T}|} \sum_{x \in \mathcal{T}}' \log \sum_{s \in K^n}' \frac{\alpha(s|x)}{\alpha(s|x)} p_u(x,s)$$

where $\alpha(s|x) \geq 0$, $\sum_{s \in K^n}' \alpha(s|x) = 1 \quad \forall x \in \mathcal{T}$

Using concavity of log, we get a lower bound

$$L(u) \geq L_B(u) = \frac{1}{|\mathcal{T}|} \sum_{x \in \mathcal{T}}' \sum_{s \in K^n}' \alpha(s|x) \log p_u(x,s) -$$

$$- \frac{1}{|\mathcal{T}|} \sum_{x \in \mathcal{T}}' \sum_{s \in K^n}' \alpha(s|x) \log \alpha(s|x)$$

Maximising it by block-coordinate ascent w.r.t. to α-s and u-s gives the EM-alogithm:

Start with some $u^{(0)}$ and iterate

E-step: set $\alpha^{(t)}(s|x) = p_{u^{(t)}}(s|x) \quad \forall s \in K^n, \forall x \in \mathcal{T}$

M-step: set

$$u^{(t+1)} \in \operatorname*{argmax}_u \frac{1}{|\mathcal{T}|} \sum_{x \in \mathcal{T}}' \sum_{s \in K^n}' \alpha^{(t)}(s|x) \log p_u(x,s)$$

Let us analyse the M-step for exponential families and for HMMs in particular. The objective is

$$\frac{1}{|\mathcal{T}|}\sum_{x\in\mathcal{T}}\sum_{s\in K^n}\alpha^{(t)}(s\mid x)\langle\varphi(x,s),u\rangle - \log Z(u) \to \max_u$$

Denoting $\psi = \frac{1}{|\mathcal{T}|}\sum_{x\in\mathcal{T}}\sum_{s\in K^n}\alpha^{(t)}(s\mid x)\,\varphi(x,s)$, we get the task

$$\langle\psi,u\rangle - \log Z(u) \to \max_u .$$

It is equivalent to the supervised learning task analysed in Sec. 7. We know how to solve it, provided we know how to compute ψ.

Computing ψ:

- For each $x\in\mathcal{T}$ we have to compute

$$\psi(x) = \sum_{s\in K^n}\alpha^{(t)}(s\mid x)\,\varphi(x,s) = \sum_{s\in K^n} p_{u^{(t)}}(s\mid x)\,\varphi(x,s)$$

i.e. we have to compute posterior pairwise marginals $p(s_{i-1}, s_i \mid x)$ $\forall\, i=2,\dots,n$ and $\forall\, s_{i-1}, s_i\in K$. This can be done by an alogithm similar to the one discussed in Sec. 5 (see also Assignment 1 from Seminar 3)

- The components of ψ are then obtained by averaging the components of $\psi(x)$ over all $x\in\mathcal{T}$, i.e.

$$\psi = \frac{1}{|\mathcal{T}|}\sum_{x\in\mathcal{T}}\psi(x)$$

Theorem 1 (w/o proof)

The sequence $L(u^{(t)})$ is monotonously increasing and the sequence $\alpha^{(t)}$ is convergent.