## 9. Unsupervised learning: EM algorithm for HMMs

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

Task: $$u_* \in \arg\max_u \frac{1}{|\mathcal{T}|} \sum_{x \in \mathcal{T}} \log \sum_{s \in K^n} p_u(x,s)$$

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 $\alpha(s|x) \geq 0$, $\sum_{s \in K^n} \alpha(s|x) = 1 \quad \forall x \in \mathcal{T}$

$$\geq \frac{1}{|\mathcal{T}|} \sum_{x \in \mathcal{T}} \sum_{s \in K^n} \alpha(s|x) \log p_u(x,s) -$$

$$- \frac{1}{|\mathcal{T}|} \sum_{x \in \mathcal{T}} \sum_{s \in K^n} \alpha(s|x) \log \alpha(s|x)$$

$$= L_B(u, \alpha)$$

This is a lower bound of the log-likelihood (Minka 1998)
Maximise it by block-coordinate ascent w.r.t. $\alpha$-s and $u$-s.
Start with some $u^{(0)}$ and iterate

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

$\boxed{\text{M-step:}}$ set

$$u^{(t+1)} \in \arg\max_u \frac{1}{|\mathcal{T}|} \sum_{x \in \mathcal{T}} \sum_{s \in K^n} \alpha^{(t)}(s|x) \log p_u(x,s)$$

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

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

By denoting

$$\psi = \frac{1}{|T|}\sum_{x\in T}\sum_{s\in K^n} \alpha^{(t)}(s|x)\,\varphi(x,s),$$

we see that the task in the M-step reads

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

~~This task~~ It is equivalent to the supervised learning task analysed in Sec. 7! We already know how to solve it, provided that we know how to compute $\psi$.

- Computing $\psi$:
- For each $x\in T$ we have to compute

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

i.e. we have to compute posterior pairwise marginals $p(s_{i-1},s_i|x)$ $\forall i=2,\dots,n$ and $\forall s_{i-1},s_i\in K$. This can be done by an algorithm similar to the one discussed in Sec. 5 (computing $p(s_i|x)$).

- The components of $\psi$ are obtained by averaging the components of $\psi(x)$ over all $x\in T$, i.e. from

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

**Theorem 1** (w/o proof)

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