10. Unsupervised learning, EM algorithm

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

Task: $\vec{u}^* \in \operatorname{argmax}_{\vec{u}} \frac{1}{|T|} \sum_{x \in T} \log \sum_{s \in K^n} p_{\vec{u}}(x,s)$

Substitute $p_{\vec{u}} = \frac{1}{Z(\vec{u})} \exp\langle \vec{\varphi}(x,s), \vec{u} \rangle$:

$$L(\vec{u}) = \underbrace{\log Z(\vec{u})}_{g(\vec{u})} - \underbrace{\frac{1}{|T|} \sum_{x \in T} \log \sum_{s \in K^n} \exp\langle \vec{\varphi}(x,s), \vec{u} \rangle}_{h(\vec{u})} \to \min_{\vec{u}}$$

- $L(\vec{u})$ is a difference of convex functions
- $\nabla L(\vec{u}) = \mathbb{E}_{\vec{u}}(\vec{\varphi}) - \frac{1}{|T|} \sum_{x \in T} \mathbb{E}_{\vec{u}}(\vec{\varphi} \mid x)$

  i.e. for an optimal $\vec{u}^*$ its pairwise marginal statistics (on edges) coincide with its averaged conditional pairwise marginal statistics on the training data.

The optimisation task can be solved by a DC-algorithm, which turns out to be a „re-incarnation" of the EM-algorithm (aka Baum-Welch algorithm for HMMs)

Start with an arbitrary $\vec{u}^{(0)}$ and iterate

a) E-step: compute

$$\vec{v}^{(t)} = \nabla h(\vec{u}^{(t)}) = \frac{1}{|T|} \sum_{x \in T} \mathbb{E}_{\vec{u}^{(t)}}(\vec{\varphi} \mid x)$$

The posterior marginal probabilities $p(s_{i-1}, s_i \mid x)$ for each example $x \in T$ can be computed by an algorithm similar to the one given in sec. 5

b) M-step: compute $\vec{u}^{(t+1)} \in \partial g^*(\vec{v}^{(t)})$

We know from sec. 9:

$$\vec{u} \in \partial g^*(\vec{v}) \iff \vec{v} \in \partial g(\vec{u})$$

Hence

$$\vec{u}^{(t+1)} \in \underbrace{\operatorname{argmax}_{\vec{u}} \left[\langle \vec{v}^{(t)}, \vec{u} \rangle - \log Z(\vec{u}) \right]}_{\text{MLE}}$$

and we know how to solve this task for HMMs (see sec. 7)

Theorem 1 (w/o proof)

- The sequences $g(\vec{u}^{(t)}) - h(\vec{u}^{(t)})$ and $h^*(\vec{v}^{(t)}) - g^*(\vec{v}^{(t)})$ are non increasing
- The sequence $\vec{v}^{(t)}$ is convergent. Its fixpoint is a local minimum of $h^*(\vec{v}) - g^*(\vec{v})$.