6. Representing HMMs as exponential families

Def. 16, Sec. 1 $\Rightarrow$ the joint p.d. for a Markov chain model with strictly positive prob's can be written as

$$p(s) = p(s_1, \dots, s_n) = \frac{1}{Z} \prod_{i=2}^{n} g_i(s_{i-1}, s_i) = \frac{1}{Z} \exp\Big[\sum_{i=2}^{n} u_i(s_{i-1}, s_i)\Big]$$

Remark 1 The factors g_i, resp. the potentials u_i define the model uniquely. The reverse is not true.

Remark 2 The normalising factor Z is defined by

$$Z(u) = \sum_{s \in K^n} \exp\Big[\sum_{i=2}^{n} u_i(s_{i-1}, s_i)\Big]$$

and can be computed by an algorithm similar to the one described in Sec. 3.

Denote:

- $\vec{\varphi}(s_i) \in \mathbb{R}^K$ a binary valued indicator vector that denotes the state s_i ("one out of K" encoding), i.e.

$$\vec{\varphi}(s_i = k) = (0, \dots, \underset{\substack{\uparrow \\ \text{pos. } k}}{1}, \dots 0)$$

- U_i denotes the $K \times K$ matrix with values $u_i(s_{i-1}, s_i)$

Then, the joint p.d. can be written as

$$p(s) = \frac{1}{Z(u)} \exp \sum_{i=2}^{n} \langle \vec{\varphi}_{i-1}(s_{i-1}), U_i \cdot \vec{\varphi}_i(s_i) \rangle$$

$$= \frac{1}{Z(u)} \exp \sum_{i=2}^{n} \langle \Phi_i(s_{i-1}, s_i), U_i \rangle ,$$

where

$$\Phi_i(s_{i-1}, s_i) = \vec{\varphi}(s_{i-1}) \otimes \vec{\varphi}(s_i)$$

is a $K \times K$ binary valued indicator matrix and

$$\langle \Phi_i, U \rangle = \mathrm{tr}(\Phi^T U)$$

denotes the Frobenius inner product.

Finally, let us denote $\Phi = (\Phi_2, \Phi_3, \dots, \Phi_n)$ and $U = (U_2, U_3, \dots, U_n)$.
The joint p.d. of a Markov chain model can be written as

$$p(s) = \frac{1}{Z(u)} \exp \langle \Phi(s), U \rangle .$$

Similarly, the joint p.d. of an HMM can be written as

$$p(s) = \frac{1}{Z(u)} \exp \langle \Phi(x, s), U \rangle$$

by using similar notations.

7. Supervised learning, ML-estimator

Given an i.i.d. sample of pairs of sequences

$$\mathcal{T} = \{ (x^j, s^j) \mid x^j \in F^n, s^j \in K^n, j = 1, \dots, \ell \},$$

estimate the model parameters of the HMM by the maximum likelihood estimator

$$U^* \in \arg\max_u \prod_{(x,s) \in \mathcal{T}} p_u(x, s) =$$

$$= \arg\max_u \frac{1}{|\mathcal{T}|} \sum_{(x,s) \in \mathcal{T}} \log p_u(x, s),$$

i.e. find optimal $u_i^*(s_{i-1}, s_i)$, $\tilde{u}_i^*(x_i, s_i)$ or, equivalently, $p(s_{i-1}, s_i)$, $p(x_i, s_i)$

Intuitive answer: u^* is given by

$$p_{u^*}(s_{i-1}, s_i) = \beta(s_{i-1}, s_i)$$
$$p_{u^*}(x_i, s_i) = \beta(x_i, s_i),$$

where β-s denote the frequencies of the corresponding events in T.

Let us prove correctness. The log-likelihood of T is

$$L(u) = \frac{1}{|T|} \sum_{(x,s)\in T} \left[\langle \varphi(x,s), u \rangle - \log Z(u) \right]$$

$$= \langle \psi, u \rangle - \log Z(u),$$

where

$$\psi = \mathbb{E}_T \varphi = \frac{1}{|T|} \sum_{(x,s)\in T} \varphi(x,s).$$

Lemma 1 The log-partition function $\log Z(u)$ of an HMM (with strictly positive p.d.) is convex in u.

Proof

$$\nabla_u \log Z(u) = \frac{1}{Z(u)} \sum_{x,s} \exp\langle \varphi(x,s), u\rangle \, \varphi(x,s) \overset{!}{=} \mathbb{E}_u \varphi$$

The components of $\mathbb{E}_u \varphi$ are the pairwise marginal prob's on the edges of the model.

$$\nabla_u^2 \log Z(u) = \mathbb{E}_u(\varphi \otimes \varphi) - \mathbb{E}_u(\varphi) \otimes \mathbb{E}_u(\varphi)$$
$$= \mathbb{E}_u\left[(\varphi - \mathbb{E}_u\varphi) \otimes (\varphi - \mathbb{E}_u \varphi) \right].$$

The expectation of a positive semidefinite matrix is p.s.d. $\Rightarrow$ $\log Z(u)$ is convex. ∎

The log-likelihood $L(u)$ is concave as a consequence and has global maxima only. They are given by

$$\nabla_u L(u^*) = \frac{1}{|T|} \sum_{(x,s)\in T} \varphi(x,s) - \mathbb{E}_{u^*}\varphi = \mathbb{E}_T \varphi - \mathbb{E}_{u^*}\varphi = 0$$

Recall that the components of $\mathbb{E}_u(\varphi)$ are the pairwise marginal prob's of the model $p_u(x,s)$. Hence, the optimiser u^* defines the model whose pairwise marginal prob's coincide with the empirical marginal frequencies in T.