6. Representing an HMM as an exponential family

According to Def. 1b in sec. 1, the joint p.d. for a Markov model on a chain can be written as

$$p(s) = p(s_1, \dots, s_n) = \prod_{i=2}^{n} g_i(s_{i-1}, s_i)$$

To allow arbitrary non-negative g-s, we introduce a normalising factor Z. If, in addition, all g-s are strictly positive, we may write

$$p(s) = \frac{1}{Z(u)} \exp\left[\sum_{i=2}^{n} u_i(s_{i-1}, s_i)\right]$$

Remark 1 Notice that the normalising factor Z depends on the model parameters u, i.e. Z is a function of u. It is defined by

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

and can be computed by the algorithm described in sec. 3

Remark 2 The formula

$$p(s) = p(s_1, \dots, s_n) = \frac{p(s_1, s_2) \cdot p(s_2, s_3) \cdot \dots \cdot p(s_{n-1}, s_n)}{p(s_2) \cdot p(s_3) \cdot \dots \cdot p(s_{n-1})}$$

for a Markov chain (see sec. 1) provides an easy way to compute the u-s given the pairwise marginal prob's.

Remark 3 The factors g_i resp. the potentials u_i define a Markov model uniquely but are themselves not unique

$$u_2(s_1, s_2) + \dots + u_i(s_{i-1}, s_i) + u_{i+1}(s_i, s_{i+1}) + \dots u_n(s_{n-1}, s_n)$$

$$\underbrace{+ V_i(s_i)}_{\tilde{u}_i} \qquad \underbrace{- V_i(s_i)}_{\tilde{u}_{i+1}}$$

To emphasize the structure in which the model depends on its parameters u, we may write

$$p(s) = \frac{1}{Z(u)} \exp\left[\sum_{i=2}^{n} u_i(s_{i-1}, s_i)\right] = \frac{1}{Z(\vec{u})} \exp\langle \vec{\varphi}(s), \vec{u} \rangle$$

The vectors $\vec{\varphi}(s), \vec{u}$ are elements of $\mathbb{R}^{(n-1)K^2}$

$\vec{\varphi}(s)$ is a binary valued indicator vector with components defined as follows

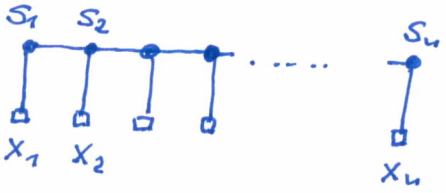
$$\varphi_{ikk'}(s) = \varphi_{ikk'}(s_{i-1}, s_i) = \delta(s_{i-1}, k) \cdot \delta(s_i, k')$$

$$i = 2, \dots, n, \quad k, k' \in K$$

where δ denotes the Kronecker delta.

$\vec{u}$ is the parameter vector with components $u_{ikk'} = u_i(k, k')$

A hidden Markov model (on a chain) is defined on the graph



$s_i \in K$

$x_i \in \bar{F}$

Its joint p.d. can be written as

$$p(x,s) = \prod_{i=2}^{n} g_i(s_{i-1}, s_i) \prod_{i=1}^{n} \tilde{g}_i(x_i, s_i)$$

$$= \frac{1}{Z(u, \tilde{u})} \exp\left[\sum_{i=2}^{n} u_i(s_{i-1}, s_i) + \sum_{i=1}^{n} \tilde{u}_i(x_i, s_i)\right]$$

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

7. Supervised learning. ML-estimator

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

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

Maximum likelihood estimator:

$$\vec{u}^* = \operatorname*{argmax}_{\vec{u}} \prod_{(x,s) \in T} p_{\vec{u}}(x,s)$$

$$= \operatorname*{argmax}_{\vec{u}} \frac{1}{|T|} \sum_{(x,s) \in T} \log p_{\vec{u}}(x,s)$$

i.e. find optimal

$\vec{u}$: $u_i(s_{i-1}, s_i)$, $\tilde{u}_i(x_i, s_i)$ or, equivalently

$p(s_{i-1}, s_i)$, $p(x_i, s_i)$

Intuitive answer: $\vec{u}^*$ is given by

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

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

Let us prove that this is correct. The objective function for the learning task is

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

$$= \langle \vec{\Psi}, \vec{u} \rangle - \log Z(\vec{u})$$

where $\vec{\Psi}$ denotes the empirical mean of the random vector $\vec{\Phi}$:

$$\vec{\Psi} = \frac{1}{|T|} \sum_{(x,s) \in T} \vec{\varphi}(x,s)$$

The first term in $L(\vec{u})$ is linear and thus also concave. Let us prove that $\log Z(\vec{u})$ is convex.

$$\log Z(\vec{u}) = \log \sum_{x,s} \exp \langle \vec{\varphi}(x,s), \vec{u} \rangle$$

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

i.e. $\nabla \log Z(\vec{u}) \hat{=}$ expectation of the random vector $\vec{\Phi}$

components: pairwise marginal prob's on edges

$$\nabla^2 \log Z(\vec{u}) = \mathbb{E}_{\vec{u}}(\vec{\Phi} \otimes \vec{\Phi}) - \mathbb{E}_{u}(\vec{\Phi}) \otimes \mathbb{E}_{u}(\vec{\Phi})$$

$$= \mathbb{E}_{\vec{u}}\left[\left(\vec{\Phi} - \mathbb{E}_{\vec{u}}(\vec{\Phi})\right) \otimes \left(\vec{\Phi} - \mathbb{E}_{\vec{u}}(\vec{\Phi})\right) \right]$$

i.e. $\nabla^2 \log Z(\vec{u}) \hat{=}$ covariance matrix of the random vector $\vec{\Phi}$ $\Rightarrow$ symmetric and positive semidefinite

<u>Lemma 1</u> The partition function $\log Z(\vec{u})$ of an HMM (with strictly positive p.d.) is <u>convex</u> in $\vec{u}$. □

Hence, the objective function $L(\vec{u})$ is concave and has global maxima only. They are given by

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

But, the components of $\mathbb{E}_{\vec{u}}(\vec{\Phi})$ are the pairwise marginals of the model $p_{\vec{u}}(x,s)$. The optimiser $\vec{u}^*$ defines the model which has precisely the same pairwise marginals as the empirical marginal frequencies of T.

Concavity of log-Likelihood $L(\vec{u})$ also ensures consistency of the estimator.

Theorem 1 (w/o proof) The maximum likelihood estimator for HMM-s is consistent, i.e.

$$P_{\vec{u}}\left(\| \vec{u}_*(T_\ell) - \vec{u} \| > \varepsilon \right) \xrightarrow{\ell \to \infty} 0$$

for every $\varepsilon > 0$