6. Representing HMMs as exponential families

**Definition 1**

An exponential family of distributions for a random variable \( X \in \mathbb{X} \) is a parametric model with densities

\[
P_\theta(x) = h(x) \exp \left[ \langle \psi(x), \theta \rangle - a(\theta) \right]
\]

where

- \( \theta \in \mathbb{R}^n \) is the natural parameter
- \( \psi(x) \in \mathbb{R}^n \) is the sufficient statistics
- \( h(x) \geq 0 \) is the base measure
- \( a(\theta) \) is the log partition function (cumulant function) given by

\[
a(\theta) = \log \int h(x) \exp \langle \psi(x), \theta \rangle \, d\nu(x).
\]

**Example 1**

a) Bernoulli distribution \( P_\theta(x) = \theta^x (1-\theta)^{1-x}, \ x = 0, 1 \)

\[
P_\theta(x) = \exp \left[ x \log \frac{\theta}{1-\theta} + \log(1-\theta) \right]
\]

b) Univariate normal distribution \( P_\mu(x) = \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{1}{2} (x-\mu)^2 \right] \)

\[
h(x) = \frac{1}{\sqrt{2\pi}}, \quad e^{-\frac{1}{2} x^2}
\]

\[
\psi(x) = x
\]

\[
a(\mu) = \frac{1}{2} \mu^2
\]
Minimal representation:

\[ \forall \theta \in \mathbb{R}^n : \quad \langle \theta, \psi(x) \rangle = \text{const} + t \in \chi \]

\[ \forall b \in \mathbb{R}^n : \quad \langle b, \theta \rangle = \text{const} + t \in \Theta \]

Def 16, Sec. 4: the joint p.d. of a Markov chain model with strictly positive prob's can be written as

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

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

Remark 2: The partition function \( Z(u) \) is defined by

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

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

Denote:

1. \( \psi(s_i) \in \mathbb{R}^K \) the binary valued indicator vector that denotes the state \( s_i \in \mathcal{X} \) in "one out of \( K \)" encoding, i.e.

\[ \psi(s_i = k) = (0, \ldots, 1, \ldots, 0) \]

2. \( U_i \), the \( K \times K \) matrix with values \( u_i(s_{i-1}, s_i) \)
The joint p.d.f. of a strictly positive Markov chain model can be written as

\[ p(s) = \frac{1}{Z(u)} \exp \sum_{i=2}^{n} \langle \psi(s_{-i}), U_i \psi(s_i) \rangle \]

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

where

\[ \Phi(s_{i-1}, s_i) = \psi(s_{i-1}) \otimes \psi(s_i) \]

is a \( K \times K \) binary valued indicator matrix and

\[ \langle \Phi, U \rangle = Tr(\Phi^T U) \]

denotes the Frobenius inner product.

Finally, denote \( \Phi = (\Phi_2, \ldots, \Phi_n) \) and \( U = (U_2, \ldots, U_n) \) and write

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

The joint p.d.f. of an HMM can be written as

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

by using similar notations.

**Remark 3** The EEF-representations of Markov models / HMMs are not minimal.

**Remark 4** The components of the expectation

\[ \mathbb{E}_{\sup_n \pi(s)} \left[ \Phi(s) \right] \]

for a Markov chain model are the pairwise marginal probabilities for pairs of **consecutive** states.
Given an i.i.d. sample of pairs of sequences

\[ T = \{ (x^j, s^j) \mid x^j \in \mathbb{F}^n, s^j \in \mathbb{K}^n, j = 1, \ldots, L \} \]

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

\[ u^* \in \arg \max_u \prod \limits_{(x,s) \in T} \rho_u (x,s) = \arg \max_u \frac{1}{|T|} \sum \limits_{(x,s) \in T} \log \rho_u (x,s), \]

i.e. find optimal \( u^* (x; s_i), u^* (s_{i-1}; s_i) \) or, equivalently, \( p(x_i; s_i), p(s_{i-1}; s_i) \).

Intuitive answer: \( u^* \) is given by

\[ \rho_{u^*} (s_{i-1}; s_i) = \beta (s_{i-1}; s_i) \]

\[ \rho_{u^*} (x_i; s_i) = \beta (x_i; s_i) \]

where \( \beta \)-s denote the frequencies of the corresponding events in \( T \).

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

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

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

where

\[ \psi = \mathbb{E}_T \Phi = \frac{1}{|T|} \sum \limits_{(x,s) \in T} \Phi(x,s) \]

Remark 1: Observe that all we need to know from the sample \( T \) is \( \psi = \mathbb{E}_T \Phi \).
Lemma 1: The log-partition function $\log Z(u)$ of an HMM is convex in $u$.

Proof:

\[
\nabla_u \log Z(u) = \frac{1}{Z(u)} \sum_{x,s} \exp<\phi(x,s), u> \phi(x,s) \nabla \phi = E_u \phi
\]

Recall that the components of $E_u \phi$ are the pairwise marginal probs on the edges of the model.

\[
\nabla_u^2 \log Z(u) = E_u [\phi \phi] - E_u [\phi] \otimes E_u [\phi]
\]

\[
= E_u [(\phi - E_u \phi) \otimes (\phi - E_u \phi)]
\]

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

The log-likelihood is concave and has global maximun only as a consequence. They are given by

\[
\nabla_u L(u^*) = \frac{1}{T} \sum_{(x,s) \in T} \phi(x,s) - E_{u^*}[\phi] = E_T[\phi] - E_{u^*}[\phi] = 0
\]

Recall that the components of $E_u [\phi]$ are the pairwise marginal probs of the model $p_u(x,s)$. Hence, the optimiser $u^*$ defines the model whose pairwise marginal probs coincide with the empirical marginal frequencies in $T$. 