6. Representing Markov models as exponential families

4. Exponential families

Definition 1. An exponential family of distributions for a random variable $X \in \mathcal{X}$ is a parametric model with p.d.

$$p(x) = h(x) \exp[\langle \psi(x), \theta \rangle - A(\theta)]$$

where
- $\psi(x) \in \mathbb{R}^n$ is the sufficient statistic
- $\theta \in \mathbb{R}^n$ is the natural parameter
- $h(x) \in \mathbb{R}_+$ is the base measure
- $A(\theta)$ is the log-partition function (aka cumulant function) given by

$$A(\theta) = \log \int h(x) \exp(\langle \psi(x), \theta \rangle) \, dx$$

Example 1

a) Bernoulli distribution $p(x) = p^x (1-p)^{1-x}, \ x = 0, 1$

$$p(x) = \exp[\langle x \log(p) - (1-p) \rangle]$$

with natural parameter $\theta = \log(p)$

b) Normal distribution $p(x) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2} (x-\mu)^2\right]$ is an exponential family with

$$h(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} x^2}, \ \psi(x) = x, \ \theta = \mu, \ A(\mu) = \frac{1}{2} \mu^2$$

Definition 2. An exponential family has minimal representation if $A(\theta) \neq \text{const.} + \langle a, \psi(x) \rangle$, for all $x \in \mathcal{X}$. That is, each distribution of the family is represented by a unique parameter vector $\theta \in \mathbb{R}^n$. A non-minimal representation is called overcomplete.
Proposition 1 (principle of maximum entropy)

Let \( X \in \mathcal{X} \) be a random variable and \( \psi(x) \in \mathbb{R}^n \), \( x \in \mathcal{X} \) a statistic. The probability distribution with highest entropy among distributions \( \mathcal{P}(x) \) with \( \mathbb{E}_\mu \psi(x) = \mu \) is a member of the family

\[
\mathcal{P}(x) = \frac{1}{Z} \exp \left[ \langle \psi(x), \theta \rangle - A(\theta) \right].
\]

Markov models in exponential form

Starting from Definition 16, Sec 1 \Rightarrow

The joint p.d.f. of a Markov chain model with strictly positive probabilities can be written as

\[
p(s) = p(s_1, s_2, \ldots, s_n) = \frac{1}{Z} \prod_{i=1}^{n} \phi_i(s_{i-1}, s_i) = \frac{1}{Z} \exp \sum_{i=2}^{n} u_i (s_{i-1}, s_i)
\]

Remark 1. The partition function \( Z(\mu) \) is defined by

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

and can be computed by an algorithm similar to the one discussed in Sec 3. The potentials \( u_i : \mathbb{R}^2 \to \mathbb{R} \) define the model uniquely. The reverse is not true.

Let us consider the underlying chain of the model as a graph and denote its nodes \( i \in \mathcal{V} \) and its edges \( e \in \mathcal{E} \). A sequence of states \( s = (s_1, s_n) \) labels the nodes \( i \in \mathcal{V} \) by labels \( s \in \mathcal{E} \).

We represent edge labelings \( s_e \), \( e \in \mathcal{E} \) by row vectors \( \mathbf{e} \in \mathbb{R}^n \). The matrices \( \Phi(s) = \Phi(s_e) \), and write the joint p.d.f. as

\[
p(s) = \frac{1}{Z} \exp \sum_{e \in \mathcal{E}} \left\langle \Phi_e(s), \mu_e \right\rangle
\]

Where \( \mu_e \) is a \( K \times K \) matrix of the value of the potential \( \Phi_e : \mathbb{R}^2 \to \mathbb{R} \). If the model is homogeneous, i.e., the potentials \( \mu_e \) are the same for all edges \( e \in \mathcal{E} \), we may write

\[
p(s) = \frac{1}{Z} \exp \left( \sum_{e \in \mathcal{E}} \Phi_e(s) \right) = \frac{1}{Z} \exp \left( \Phi(s) \right).
\]
For the general case, we arrive at a similarly compact notation if we define
\[ H(s) = (P_{s_1}(s), P_{s_2}(s), \ldots, P_{s_n}(s)), \quad U = (u_{e_1}, \ldots, u_{e_{n-1}}). \]

Remark 2. The EF-representation of Markov models is not minimal. The components of the expected statistic \( E[I(s)] \) for a Markov chain model are the pairwise marginal probabilities on the edges \( e \in E \).

Remark 3. We can extend this to EF-representations of HMMs by introducing statistics for all edges of the model:

\[ x_1 \rightarrow x_2 \rightarrow \cdots \rightarrow x_n \]

\[ s_1 \rightarrow s_2 \rightarrow \cdots \rightarrow s_n \]
Given an i.i.d. sample of sequences $\mathcal{T} = \{ s_j \}_{j \in \mathcal{K}^n}$, estimate the model parameters of the Markov model by the maximum likelihood estimator
\[
P^* \in \arg \max_P \prod_{s \in \mathcal{T}} P(s) = \arg \max_P \frac{1}{|\mathcal{T}|} \sum_{s \in \mathcal{T}} \log P(s).
\]

Intuitive answer: $P^*$ is given by $P^*(s_{-1}, s_1) = \hat{P}(s_{-1}, s_1)$, where $\hat{P}$ denotes the frequencies of the corresponding events in $\mathcal{T}$. Let us prove correctness.

The log-likelihood of $\mathcal{T}$ is
\[
L(u) = \frac{1}{|\mathcal{T}|} \sum_{s \in \mathcal{T}} \left[ \langle \Phi(s), u \rangle - \log Z(u) \right]
= \mathbb{E}_T \left[ \langle \Phi(s), u \rangle \right] - \log Z(u)
= \langle \Phi^*, u \rangle - \log Z(u)
\]
where $\Phi^* = \mathbb{E}_T \left[ \Phi(s) \right]$.

Remark: Observe that all we need to know from the training data $\mathcal{T}$ is $\Phi^* = \mathbb{E}_T \left[ \Phi(s) \right]$.

Lemma: The log-partition function $\log Z(u)$ of a Markov model is convex in $u$.

Proof
\[
\nabla_u \log Z(u) = \frac{1}{Z(u)} \sum_{s \in \mathcal{K}^n} \exp \langle \Phi(s), u \rangle \nabla_u \Phi(s) = \mathbb{E}_{\Phi_u} \left[ \nabla_u \Phi(s) \right]
\]

Recall that the components of $\mathbb{E}_{\Phi_u} \left[ \Phi(s) \right]$ are the pairwise marginal probs on the model edges.
\[
\nabla_u^2 \log Z(u) = \mathbb{E}_{\Phi_u} \left[ \Phi(s) \otimes \Phi(s) \right] - \mathbb{E}_{\Phi_u} \left[ \Phi(s) \right] \otimes \mathbb{E}_{\Phi_u} \left[ \Phi(s) \right]
= \mathbb{E}_{\Phi_u} \left[ (\Phi - \mathbb{E}_{\Phi_u} \Phi) \otimes (\Phi - \mathbb{E}_{\Phi_u} \Phi) \right]
\]
The expectation of a positive semidefinite matrix is p.s.d. $\Rightarrow \log Z(u)$ is convex.
The log-likelihood is concave, and as a consequence, has only global maxima. They are given by:

\[ \nabla_u \log L(u) = E_T [q(s)] - E_{p_u} [q(s)] = 0 \]

Hence, the optimiser \( u^* \) defines the model whose pairwise marginal probs coincide with the empirical marginal frequencies in \( T \).

This is easily generalised to learning of HMMs on i.i.d. training data \( T \) which consist of pairs of sequences \((x,s)\). Recall that an HMM is defined as

\[ p(x,s) = \prod_{i=1}^n p(x_i|x_{i-1}, s_i) \prod_{i=2}^n p(s_i; s_{i-1}) \]

\[ \frac{p(x,s)}{p(x) p(s)} \]

Both model parts are learned independently. The log-likelihood of \( p(x,s) \) further splits into the sum

\[ \log p(x,s) = \sum_{i=1}^n \log p(x_i|x_{i-1}, s_i) \]

so that each \( \log p(x_i|x_{i-1}, s_i) \) can be learned independently.