6. Representing Markov models as exponential families

A. Exponential families

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

$$p_\theta(x) = h(x)\exp\left[\langle \varphi(x), \theta\rangle - A(\theta)\right]$$

where

- $\varphi(x) \in \mathbb{R}^n$ 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

$$A(\theta) = \log \int h(x)\exp\langle \varphi(x), \theta\rangle \, d\nu(x)$$

□

Example 1

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

$$p(x) = \exp\left[x \log\frac{\beta}{1-\beta} - \log(1-\beta)\right]$$

with natural parameter $\theta = \log\frac{\beta}{1-\beta}$

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

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

Definition 2 An exponential family has minimal representation if $\nexists\, a \in \mathbb{R}^n$ s.t. $\langle a, \varphi(x)\rangle = \text{const.}\ \forall x \in \mathcal{X}$. I.e. 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 $\varphi(x) \in \mathbb{R}^n$, $x \in \mathcal{X}$ a statistic. The probability distr. with highest entropy among distributions $p(x)$ with $\mathbb{E}_p[\varphi(x)] = \mu$ is a member of the family

$$p(x) = \exp\left[\langle \varphi(x), \theta \rangle - A(\theta)\right].$$

□

B. Markov models in exponential form

Starting from Definition 16, Sec. 1 $\Rightarrow$
The joint p.d. of a Markov chain model with strictly positive probabilities can be written as

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

Remark 1 The partition function $Z(u)$ is defined by

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

and can be computed by an algorithm similar to the one discussed in Sec. 3. The potentials $u_i : K^2 \to \mathbb{R}$ 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 $i \in V$ and its edges $e \in E$. A sequence of states $s = (s_1, \dots, s_n)$ labels the nodes $i \in V$ by labels $s_i \in K$. We represent edge labellings s_e, $e \in E$ by one-hot $K \times K$ matrices $\Phi_e(s) = \Phi_e(s_e)$, and write the joint p.d. as

$$p(s) = \frac{1}{Z} \exp \sum_{e \in E} \langle \Phi_e(s), U_e \rangle$$

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

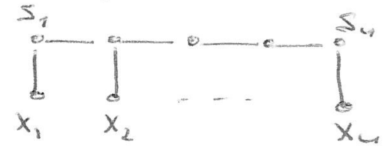
$$p(s) = \frac{1}{Z} \exp \left\langle \sum_{e \in E} \Phi_e(s), U \right\rangle = \frac{1}{Z} \exp \langle \Phi(s), U \rangle$$

For the general case, we arrive at a similarly compact notation if we define

$$\varphi(s) = \left(\varphi_{e_1}(s), \varphi_{e_2}(s) \dots \varphi_{e_{n-1}}(s)\right), \qquad u = \left(u_{e_1}, \dots, u_{e_{n-1}}\right).$$

Remark 2 The EF-representation of Markov models is not minimal. The components of the expected statistic $\mathbb{E}_p[\varphi(s)]$ 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



□

7. Supervised learning for Markov models and HMMS

Given an i.i.d. sample of sequences $\mathcal{T} = \{s^j \mid s^j \in K^n,\ j = 1, \dots, m\}$ estimate the model parameters of the Markov model by the maximum likelihood estimator

$$p_* \in \operatorname*{argmax}_p \prod_{s \in \mathcal{T}} p(s) = \operatorname*{argmax}_p \frac{1}{|\mathcal{T}|} \sum_{s \in \mathcal{T}} \log p(s).$$

<u>Intuitive answer:</u> p_* is given by $p_*(s_{i-1}, s_i) = \beta_i(s_{i-1}, s_i)$, where β-s denote the frequencies of the corresponding events in $\mathcal{T}$. Let us prove correctness.

The log-likelihood of $\mathcal{T}$ is

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

$$= \mathbb{E}_{\mathcal{T}} \left[\langle \varphi(s), u \rangle \right] - \log Z(u)$$

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

where $\psi = \mathbb{E}_{\mathcal{T}} [\varphi(s)]$.

<u>Remark 1</u> Observe that all we need to know from the training data $\mathcal{T}$ is $\psi = \mathbb{E}_{\mathcal{T}}[\varphi(s)]$. □

<u>Lemma 1</u> The log-partition function $\log Z(u)$ of a Markov model is convex in u.

<u>Proof</u>

$$\nabla_u \log Z(u) = \frac{1}{Z(u)} \sum_{s \in K^n} \exp\langle \varphi(s), u \rangle \, \varphi(s) \stackrel{!}{=} \mathbb{E}_{p_u} [\varphi(s)]$$

Recall that the components of $\mathbb{E}_{p_u}[\varphi(s)]$ are the pairwise marginal prob's on the model edges.

$$\nabla_u^2 \log Z(u) = \mathbb{E}_{p_u} [\varphi(s) \otimes \varphi(s)] - \mathbb{E}_{p_u}[\varphi(s)] \otimes \mathbb{E}_{p_u}[\varphi(s)]$$

$$= \mathbb{E}_{p_u} \left[(\varphi - \mathbb{E}_{p_u} \varphi) \otimes (\varphi - \mathbb{E}_{p_u} \varphi) \right]$$

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

$$\nabla_u L(u) = \mathbb{E}_T[\varphi(s)] - \mathbb{E}_{p_u}[\varphi(s)] = 0$$

Hence, the optimiser u_* defines the model whose pairwise marginal prob's 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

$$p(x, s) = \underbrace{\prod_{i=1}^{n} p(x_i \mid s_i)}_{p(x \mid s)} \; \underbrace{p(s_1) \prod_{i=2}^{n} p(s_i \mid s_{i-1})}_{p(s)}$$

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

$$\log p(x|s) = \sum_{i=1}^{n} \log p(x_i \mid s_i) \; ,$$

so that each $\log p(x_i \mid s_i)$ can be learned independently.