Exponential families and learning in (undirected) graphical models

Guillaume Obozinski

Swiss Data Science Center

SDSC

African Masters of Machine Intelligence, 2018-2019, AIMS, Kigali

Exponential family

Definition

An exponential family is a family of distributions of the form

$$p(x;\theta)\,d\nu(x) = h(x)\exp\left\{\langle b(\theta),\phi(x)\rangle - \tilde{A}(\theta)\right\}d\nu(x),$$

where

- $h(x)$ the ancillary statistic,
- $d\nu(x)$ the reference measure (or base measure),
- $\phi(x)$ the sufficient statistic (also called feature vector),
- θ the parameter,
- $\eta = b(\theta)$ the canonical parameter,
- $\tilde{A}(\theta) = A(\eta) = \log Z(\eta)$ the log-partition function.

Canonical exponential family

A canonical exponential family is an exponential family with

$$b(\theta) = \theta = \eta$$

so that

$$p(x; \eta) = h(x) \exp(\langle \eta, \phi(x) \rangle - A(\eta))$$

Partition function and log-partition functions

Note that, in the discrete case since $\sum_{x \in \mathcal{X}} p(x; \eta) = 1$, we necessarily have $A(\eta) = \log Z(\eta)$ with

$$Z(\eta) = \sum_{x \in \mathcal{X}} h(x) e^{\langle \eta, \phi(x) \rangle}.$$

Similarly, in the continuous case

$$A(\eta) = \log Z(\eta) = \log \int_{x \in \mathcal{X}} h(x) e^{\langle \eta, \phi(x) \rangle} d\nu(x)$$

Multinomial distribution in exponential family form

Let X be a random variable on $\mathcal{X} = \{0,1\}^K$. X follows a multinomial distribution of parameter $\pi \in [0,1]^K$.

$$p(x;\pi) = \prod_{k=1}^{K} \pi_k^{x_k} = \exp\Big(\sum_{k=1}^{K} x_k \log \pi_k\Big) = \exp\Big(\sum_{k=1}^{K} x_k \eta_k\Big) = \exp(\langle x, \eta\rangle)$$

that we need to identify with

$$h(x) \exp(\langle \eta, \phi(x)\rangle - A(\eta)).$$

So, we easily recognize:

- $\eta = (\log \pi_1, \log \pi_2, \dots, \log \pi_K)^\top$;
- $\phi(x) = x$;
- $h(x) = 1$ the constant function equal to one;

But we don't recognize $A(\eta)$...

Multinomial distribution in exponential family form

Let us compute explicitly $A(\eta)$:

$$A(\eta) = \log\Big(\sum_{x \in \mathcal{X}} \exp(\eta^T x)\Big) = \log\Big(\sum_{k=1}^{K} \exp(\eta_k)\Big)$$

If $\eta = (\log \pi_1, \log \pi_2, \cdots, \log \pi_K)^T$ then

$$A(\eta) = \log \sum_{k'=1}^{K} \exp \eta_k = 0.$$

The canonical parameter is however not constrained in general to satisfy this contraint.

Many exponential families

Many of the families of distributions that are classical actually are actually exponential families:

- Bernoulli, Binomial, Multinomial distribution
- Gaussian distributions
- Poisson distributions
- Geometric distributions
- Exponential distributions
- Gamma distributions
- Wishart distributions
- Beta distributions
- Dirichlet distributions
- and more

Ising model: binary variables with pairwise interactions

$$p_{\eta_0}(x) = \frac{1}{Z(\eta_0)} \exp \sum_{(i,j)\in E} \psi_{ij}(x_i, x_j; \eta_0)$$

$$\psi_{ij}(x_i, x_j; \eta_0) = V_{ij}^{11} x_i x_j + V_{ij}^{10} x_i (1-x_j) + V_{ij}^{01}(1-x_i) x_j + V_{ij}^{00}(1-x_i)(1-x_j)$$

$$\eta_0 = (V_{ij}^{kk'})_{\substack{(i,j)\in E\\ k,k'\in\{0,1\}}} \qquad \text{and} \qquad \phi(x) = \begin{pmatrix} x_i x_j \\ (1-x_i)x_j \\ \vdots \end{pmatrix}_{(i,j)\in E}$$

This first expression is overparametrized. We can rewrite the expression with just one parameter per pair (x_i, x_j):

$$\boxed{p_\eta(x) = \frac{1}{Z(\eta)} \prod_{(i,j)\in E} \exp\left(\eta_{ij}\, x_i x_j\right) \prod_{i\in V} \exp\left(\eta_i\, x_i\right)}$$

Potts' model: multinomial variables with pairwise interactions

We associate to node i a multinomial variable (with one-hot encoding)

$$X_i = (X_{i1}, \dots, X_{iK}),$$

encoding K possible states.

The expression for the Ising model generalizes to

$$\boxed{p_\eta(x) = \exp\Big(\sum_{i \in V} \sum_{k=1}^{K} \eta_{ik}\, x_{ik} \;+ \sum_{(i,j) \in E} \sum_{k,\ell=1}^{K} \eta_{ijk\ell}\, x_{ik}\, x_{j\ell} \;- A(\eta)\Big)}$$

Gibbs model in exponential family form

In the general case of a discrete graphical model **such that** $p(x) > 0$ for all $x \in \mathcal{X}$, we have:

$$\begin{aligned} p(x) &= \frac{1}{Z} \prod_{c \in \mathcal{C}} \Psi_c(x_c) \\ &= \frac{1}{Z} \exp\left\{ \sum_{c \in \mathcal{C}} \log \Psi_c(x_c) \right\} \\ &= \frac{1}{Z} \exp\left\{ \sum_{c \in \mathcal{C}} \sum_{y_c \in \mathcal{X}_c} \delta_{\{y_c = x_c\}} \log \Psi_c(y_c) \right\} \end{aligned}$$

where $\mathcal{X}_c = \{$ set of all possible values of the r.v. on the clique $c\}$.
We recognize:

$$\phi(x) = \big(\delta_{\{x_c = y_c\}}\big)_{y_c \in \mathcal{X}_c,\, c \in \mathcal{C}}$$

and

$$\eta = \big(\log \Psi_c(y_c)\big)_{y_c \in \mathcal{X}_c,\, c \in \mathcal{C}}$$

Maximum likelihood in a canonical exponential family

Assume an i.i.d. sample $x^{(1)}, \ldots, x^{(n)}$ For a model which is an exponential family, the likelihood of the parameter η

$$\mathcal{L}(\eta) = \prod_{i=1}^{n} p_\eta(x^{(i)}) = \prod_{i=1}^{n} h(x^{(i)}) \exp\big(\langle \eta, \phi(x^{(i)}) \rangle - A(\eta)\big)$$

So that the log-likelihood is

$$\ell(\eta) = \sum_{i=1}^{n} \log h(x^{(i)}) + \sum_{i=1}^{n} \langle \eta, \phi(x^{(i)}) \rangle - nA(\eta).$$

Equivalently,

$$\boxed{\frac{1}{n}\ell(\eta) = \langle \eta, \bar{\phi} \rangle - A(\eta) + c}$$

with $\bar{\phi} = \dfrac{1}{n}\sum_{i=1}^{n} \phi(x^{(i)})$ and $c = \dfrac{1}{n}\sum_{i=1}^{n} \log h(x^{(i)})$.

Qualities of canonical exponential family likelihoods

$$\boxed{\frac{1}{n}\ell(\eta) = \langle \eta, \bar{\phi} \rangle - A(\eta) + c}$$

Proposition

For an exponential family, A is a $\mathcal{C}^\infty$ **convex** function.

Corollary

In a canonical exponential family, ℓ is a **concave** function.

⚠ Does not hold in a *curved* (i.e. non-canonical) exponential family

Proposition

The maximum likelihood parameter $\widehat{\eta}_{\mathrm{ML}}$ satisfies

$$\boxed{\nabla A(\widehat{\eta}_{\mathrm{ML}}) = \bar{\phi}.}$$

Proof: The maxima of a concave differentiable function are exactly its stationary points, i.e. points such that $\nabla \ell(\eta) = 0$.

Who is ∇A?

Consider the discrete case

$$\nabla A(\eta) = \nabla\big(\log Z(\eta)\big) = \frac{1}{Z(\eta)} \nabla Z(\eta).$$

But

$$\begin{aligned}
\nabla Z(\eta) &= \sum_{x \in \mathcal{X}} \nabla\big(e^{\langle \eta, \phi(x) \rangle}\big) \\
&= \sum_{x \in \mathcal{X}} \phi(x)\, e^{\langle \eta, \phi(x) \rangle} \\
&= Z(\eta) \sum_{x \in \mathcal{X}} \phi(x)\, e^{\langle \eta, \phi(x) \rangle - A(\eta)} \\
&= Z(\eta)\, \mathbb{E}_\eta[\phi(X)]
\end{aligned}$$

So that

$$\boxed{\mu(\eta) := \nabla A(\eta) = \mathbb{E}_\eta[\phi(X)]}$$

$\mu(\eta)$ is called the *moment parameter* of the exponential family.

Moment matching property of the MLE

Combining the fact that $\nabla A(\eta) = \mathbb{E}_\eta[\phi(X)]$ for any η and that for the MLE we have $\nabla A(\widehat{\eta}_{\mathrm{ML}}) = \bar{\phi}$, we get

Theorem

The maximum likelihood estimator(s) is(/are) characterized by the *moment matching condition*:

$$\boxed{\boxed{\mu(\widehat{\eta}_{\mathrm{ML}}) := \mathbb{E}_{\widehat{\eta}_{\mathrm{ML}}}[\phi(X)] = \bar{\phi}}}$$

Interpretation: the MLE is the set of parameters such that the expected value of the vector of sufficient statistics under the chosen parameters $\mathbb{E}_{\widehat{\eta}_{\mathrm{ML}}}[\phi(X)]$ matches the empirical average value $\bar{\phi}$ of the vector of *sufficient statistics* in the data.

Computing the MLE

- The moment matching condition gives immediately the MLE for the moment parameter since

$$\widehat{\mu}_{\mathrm{ML}} = \mu(\widehat{\eta}_{\mathrm{ML}}) = \bar{\phi}.$$

- Solving for $\widehat{\eta}_{\mathrm{ML}}$ can most of the time not be done in closed form

$\Rightarrow$ Need to use numerical methods, e.g. gradient based methods.

$\Rightarrow$ Need to compute the gradient of ℓ...

Gradient of the log-likelihood

$$\nabla \ell(\eta) = \bar{\phi} - \mathbb{E}_\eta[\phi(X)]$$

- How to compute $\mathbb{E}_\eta[\phi(X)]$?

Example 1: Ising model

Reminder: $X = (X_i)_{i \in V}$ is a vector of random variables, taking value in $\{0,1\}^{|V|}$, whose distribution has the following exponential form:

$$p(x) = e^{-A(\eta)} \prod_{i \in V} e^{\eta_i x_i} \prod_{(i,j) \in E} e^{\eta_{i,j} x_i x_j}$$

The associated log-likelihood is this:

$$\ell(\eta) = \sum_{i \in V} \eta_i x_i + \sum_{(i,j) \in E} \eta_{i,j} x_i x_j - A(\eta)$$

with sufficient statistics

$$\phi(x) = \begin{pmatrix} (x_i)_{i \in V} \\ (x_i x_j)_{(i,j) \in E} \end{pmatrix}$$

Example 1: Ising model

So with

$$\ell(\eta) = \phi(x)^T \eta - A(\eta)$$
$$\nabla_\eta \, \ell(\eta) = \phi(x) - \underbrace{\nabla_\eta \, A(\eta)}_{\mathbb{E}_\eta[\phi(X)]}$$

We therefore need to compute $\mathbb{E}_\eta[\phi(X)]$.
In the case of the Ising model, we get:

$$\mathbb{E}_\eta[X_i] = \mathbb{P}_\eta[X_i = 1]$$
$$\mathbb{E}_\eta[X_i X_j] = \mathbb{P}_\eta[X_i = 1, X_j = 1]$$

Example 2: Potts model

Reminder: C_i are random variables, taking value in $\{1, \ldots, K_i\}$. We note X_{ik} the random variable such that $X_{ik} = 1$ if and only if $C_i = k$. Then,

$$p(x) = \exp\left[\sum_{i \in V} \sum_{k=1}^{K_i} \eta_{i,k} x_{ik} + \sum_{(i,j) \in E} \sum_{k=1}^{K_i} \sum_{k'=1}^{K_j} \eta_{i,j,k,k'} x_{ik} x_{jk'} - A(\eta)\right]$$

and

$$\phi(x) = \begin{pmatrix} (x_{ik})_{i,k} \\ (x_{ik} x_{jk'})_{i,j,k,k'} \end{pmatrix}$$

So that the expected value of the vector of sufficient statistics has components:

$$\mathbb{E}_\eta[X_{ik}] = \mathbb{P}_\eta[C_i = k]$$
$$\mathbb{E}_\eta[X_{ik} X_{jk'}] = \mathbb{P}_\eta[C_i = k, C_j = k']$$

On ties between learning and inference

In an exponential family

- *learning with the maximum likelihood principle* is the problem of computing η given a fixed value of $\mu(\eta) = \bar{\phi}$
- *performing probabilistic inference* is the problem of computing $\mu(\eta)$ given η.

So we can think of these problems as inverse of each other.

Learning η numerically using a gradient method requires to solve an inference problem at each iteration.
Some recent methods exploiting convex duality avoid to have to solve a whole inference problem at each iteration (Meshi et al., 2010; Pletscher et al., 2010; Meshi et al., 2015), but the connection and potential hardness related to inference is inescapable.

References I

Meshi, O., Sontag, D., Globerson, A., and Jaakkola, T. S. (2010). Learning efficiently with approximate inference via dual losses. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 783–790.

Meshi, O., Srebro, N., and Hazan, T. (2015). Efficient training of structured SVMs via soft constraints. In *AISTATS*.

Pletscher, P., Ong, C. S., and Buhmann, J. M. (2010). Entropy and margin maximization for structured output learning. In *ECML*.