The Exponential Family

Scribing Notes for COS513: Foundations of Probabilistic Modeling.

Afonso Bandeira*

October 27, 2010

1 The Exponential Family

In Probability Theory and Statistics it is very usual to consider certain classical probability distributions like Gaussian Distribution, Poisson Distributions, Bernoulli Distributions and many more. Although these families of distributions are very often treated separately one can view them in a unified context, merely as examples of distributions in the exponential family. We say that a parameterized family of distributions (like the Gaussian or the Poisson families) is in the exponential family if, parameterized with a vector η (called the natural parameters), its probability distribution (or density function) $p(X|\eta)$ can be written as

$$p(x|\eta) = e^{\eta^T t(x) - a(\eta)} h(x),$$

for some functions a, t and h[1]. Since $p(X|\eta)$ is a probability distribution it has to integrate to 1 so

$$1 = \int p(x|\eta)dx = \int e^{\eta^T t(x) - a(\eta)} h(x)dx.$$

*Program in Applied and Computational Mathematics, Princeton University, New Jersey, USA (ajsb@math.princeton.edu).

[1] h does not need to be a function, but for now let's not worry about that, we will explain why later, with examples.

Then

$$a(\eta) = \log \int e^{\eta^T t(x)} h(x) dx.$$

For this reason $a(\eta)$ is called the log normalizer.

Before giving a more detailed discussion about the exponential family we'll see examples of distributions that belong to it.

1.1 Gaussian Distribution

The Gaussian distribution, parameterized by its mean μ and variance σ^2, as its density function given by

$$p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

With algebraic manipulation one gets,

$$\begin{array}{rcl} p(x|\mu, \sigma^2) & = & \frac{1}{\sqrt{2\pi}} e^{-\log\sigma + \frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{\mu^2}{2\sigma^2}} \\ & = & \frac{1}{\sqrt{2\pi}} e^{\left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)^T \left(x, x^2\right) - \left(\log\sigma + \frac{\mu^2}{2\sigma^2}\right)}. \end{array}$$

Therefore, setting

$$\begin{array}{rcl} \eta & = & \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right) \\ t(x) & = & (x, x^2) \\ a(\eta) & = & \log\sigma + \frac{\mu^2}{2\sigma^2} \\ & = & -\log(-2\eta_2) - \frac{\eta_1^2}{4\eta_2} \\ h(x) & = & \frac{1}{\sqrt{2\pi}}, \end{array}$$

gives us the Gaussian distribution as a element of the exponential family.

1.2 Poisson Distribution

The Poisson distribution on the non-negative integers, parameterized by its intensity λ is given as

$$p(x|\lambda) = \frac{1}{x!} \lambda^x e^{-\lambda}.$$

Using algebraic manipulation one gets,

$$p(x|\lambda) = \frac{1}{x!} e^{\log(\lambda)x - \lambda}.$$

Therefore, setting

$$\begin{array}{rcl} \eta & = & \log\lambda \\ t(x) & = & x \\ a(\eta) & = & \lambda \\ & = & e^{\eta} \\ h(x) & = & \frac{1}{x!},\ \text{for non-negative } x. \end{array}$$

gives us the Poisson distribution as a element of the exponential family.

Remark 1 *Note that in this example h cannot be considered a function because, as the measure we want it to represent (let's call it ν) gives positive probability to some integers x, it is not absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^n$, so there exists no function such that $d\nu(x) = h(x)d(x)$. However this can be achieved if h is a distribution (a generalization of a function in some sense). Another way of making the statements rigorous would be to adapt the definition of exponential family and instead of considering h, consider the measure itself. For background on measure theory we recommend reading [1].*

1.3 Bernoulli Distribution

The Bernoulli distribution on $0, 1$, parameterized by its mean π is given as

$$p(x|\pi) = \pi^x(1-\pi)^{1-x}.$$

Using algebraic manipulation one gets,

$$\begin{array}{rcl} p(x|\pi) & = & e^{\log(\pi)x+\log(1-\pi)(1-x)} \\ & = & e^{(\log\pi-\log(1-\pi))x-(-\log(1-\pi))}. \end{array}$$

Therefore, setting

$$\begin{array}{rcl} \eta & = & \log\pi-\log(1-\pi) \\ t(x) & = & x \\ a(\eta) & = & -\log(1-\pi) \\ & = & -\log\left(1-\frac{e^{\eta}}{1+e^{\eta}}\right) \\ & = & \log\left(1+e^{\eta}\right) \\ h(x) & = & 1,\ \text{for } x\in\{0,1\}. \end{array}$$

gives us the Bernoulli distribution as a element of the exponential family.

Remark 2 *If we adopt a different perspective and consider a probability distribution X (not a family of parameterized distributions) with a density function $d(x)$, then we can always write it in the exponential family simply by choosing $h(x) = d(x)$, $t(x) = 0$ and $a(\eta)$=0. However, this is cheating in some sense because we are interested in estimating parameters of distributions and in the case described there are no parameters to be estimated. To be more precise we should only consider the setting where we are interested in a conditional distribution over parameters $p(X|\eta)$, and in the particular case (and always the case in practical applications) when different values of the parameters η correspond to different conditional probability distributions. When this is the case then the definition of being in the exponential family is far more restrictive since it very accurately describes the way the probability distribution varies with the parameters η.*

2 Moments of the Exponential Family

It turns out that there exists a remarkable relation between the log normalization function a and the moments of the sufficient statistics $t(x)$, we will now derive that relation.

$$\begin{aligned}
a'(\eta) &= \tfrac{d}{d\eta}\left(\log \textstyle\int e^{\eta^T t(x)} h(x)dx\right) \\
&= \frac{\frac{d}{d\eta}\int e^{\eta^T t(x)} h(x)dx}{\int e^{\eta^T t(x)} h(x)dx} \\
&= \frac{\int t(x) e^{\eta^T t(x)} h(x)dx}{\int e^{\eta^T t(x)} h(x)dx} \\
&= \frac{\int t(x) e^{\eta^T t(x)} h(x)dx}{e^{a(\eta)}} \\
&= \textstyle\int t(x) e^{\eta^T t(x) - a(\eta)} h(x)dx \\
&= \mathbb{E}[t(X)].
\end{aligned}$$

One can also prove that,

$$a''(\eta) = \mathbb{E}[t(X)^2] - \mathbb{E}[t(X)]^2 = \mathrm{Var}[t(X)].$$

Remark 3 *There is a "high level" reason for this relation between the derivatives of a and the moments of $t(X)$, we will try to make a brief exposition of it:*

Let f be the moment generating function[2] *of $t(X)$ (it is given as $f(\alpha) = \mathbb{E}[e^{\alpha t(X)}]$). Then*

$$\begin{aligned} f(\alpha) &= \mathbb{E}(e^{\alpha t(X)}) \\ &= \int e^{\alpha^t(x)} e^{\eta^t(x) - a(\eta)} h(x)dx \\ &= \int e^{(\alpha+\eta)^t(x) - a(\eta)} h(x)dx \\ &= e^{-a(\eta)} \int e^{(\alpha+\eta)^t(x)} h(x)dx \\ &= e^{-a(\eta)} e^{a(\alpha+\eta)} \\ &= e^{a(\alpha+\eta) - a(\eta)} \end{aligned}$$

Also, the cumulants[3] *generating function g is the* log *of the moments generating function, therefore*

$$g(\alpha) = \log f(\alpha) = a(\alpha + \eta) - a(\eta)$$

The cumulants can be recovered from g by differentiation at zero, the k^{th} cumulant is equal to $g^{(k)}(0)$. Also, the first cumulant is the mean and the second is the variance. Thus,

$$\left\{ \begin{array}{rcl} \mathbb{E}[t(X)] &=& \frac{d}{d\alpha} g(\alpha)_{|\alpha=0} = \frac{d}{d\alpha} a(\alpha+\eta) - a(\eta)_{|\alpha=0} = a'(\eta) \\ \mathrm{Var}[t(X)] &=& \frac{d^2}{d\alpha^2} g(\alpha)_{|\alpha=0} = \frac{d^2}{d\alpha^2} a(\alpha+\eta) - a(\eta)_{|\alpha=0} = a''(\eta) \end{array} \right.$$

Example 1 *Let's compute the mean of the Bernoulli distribution using this method. In the Bernoulli setting we have $a(\eta) = \log(1 + e^{\eta})$ so,*

$$a'(\eta) = \frac{e^{\eta}}{1+e^{\eta}} = \frac{1}{1+e^{-\eta}} = \pi = \mathbb{E}(X) = 1p(X=1) + 0p(X=0) = p(X=1).$$

Generally, we can also parameterize distributions in the exponential family with the moment parameters μ given by $\mu = \mathbb{E}[t(X)]$. We then have an application

$$\mu = \frac{da(\eta)}{d\eta},$$

[2] In one dimension one can think of a generating function of a sequence a_n as the series $\sum a_n x^n$, if $t(X)$ is high dimensional then one thinks about componentwise.

[3] Cumulants are a set of quantities analogous to moments, they are related to moments but are not the same.

that maps the natural parameters to the moment parameters. Since $a(\eta)$ is a convex function (we will not prove this here) then its derivative is injective, so we can define a inverse of this application, an application ψ mapping the moment parameters to the natural parameters,

$$\eta = \psi(\mu).$$

3 Sufficiency

In this section we will try to give a definition of sufficient statistic and show that $t(x)$ above is actually a sufficient statistic in the sense of this definition.

First of all, a statistic of a certain random variable X is a function $t(X)$ of the random variable, more precisely it is a new random variable Y that can be written as a composition of X with some other function. Let's now present a definition of sufficient statistic for a parameterized distribution.

Definition 1 (Sufficient statistics) *Let X be a probability distribution parameterized by η. We say that a statistic $t(X)$ is sufficient if $\mathbb{E}(t(X))$ gives all the information in X regarding η, i.e.*

$$p(X|\mathbb{E}(t(X)), \eta) = p(X|\mathbb{E}(t(X))).$$

In other words, X and η are independent conditioned on $\mathbb{E}(t(X))$, i.e.

$$\eta \perp X | \mathbb{E}(t(X)).$$

A remarkable property of the sufficient statistics is that estimating its expected value is enough to estimate the parameters. This means that, if one was data points $X_1, ..., X_n$, then, in order to estimate the parameter η, one only needs the information in $\mathbb{E}(t(X))$ (because X is independent of η conditioned on it), so one only needs to estimate $\mathbb{E}(t(X))$ and this is achieved by $\frac{1}{N}\sum_{n=1}^{N} t(X_n)$. This fact motivates a, data focused perspective, definition of sufficient statistics simply as the statistics $t(X)$ such that when you have data X_n, then the empirical mean of $t(X)$, $\frac{1}{N}\sum_{n=1}^{N} t(X_n)$, gives you all the information in $(X_1, ..., X_N)$ regarding the parameter η. This perspective is also usual (see e.g. [2]). We will see in the next section that only $\frac{1}{N}\sum_{n=1}^{N} t(X_n)$ will be used in the maximum likelihood estimation for the exponencial family.

Remark 4 *One can prove that the sufficient statistic $t(X)$ in the definition of the Exponential Family are actually sufficient in terms of the above definition. This can be easily done by noticing that, since $\mathbb{E}(t(X)) = a'(\eta)$ and $a(\eta)$ is a convex function, then the application $\psi(\mathbb{E}(t(X))) = \eta$ is well defined.*

4 Maximum Likelihood Estimation for the Exponential Family

Let's now consider the Maximum Likelihood Estimation for a parameterized family of distribution in the Exponential Family.

We begin by calculating the likelihood function,

$$\begin{aligned} p(x_{1:N}|\eta) &= \Pi_{n=1}^{N} \left(h(x_n) e^{\eta^T t(x_n) - a(\eta)} \right) \\ &= \left(\Pi_{n=1}^{N} h(x_n) \right) e^{\eta^T \sum_{n=1}^{N} t(x_n) - N a(\eta)}. \end{aligned}$$

Remark 5 *Notice that, as the discussion above predicted, all we need, to determine the MLE of η, is the sum of the sufficient statistics. That means that, e. g., if we were estimating a Gaussian distribution that we would only need $\sum X_n$ and $\sum X_n^2$ and if we wanted to estimate a Bernoulli distribution then we would only need $\sum X_n$.*

The log-likelihood function is then given by

$$l = \log p(x_{1:N}|\eta) = \sum_{n=1}^{N} \log h(x_n) + \eta^T \left(\sum_{n=1}^{N} t(x_n) \right) - N a(\eta)$$

This function is convex (as $a(\eta)$ is convex), then its maximum is achieved at its unique stationary point. That point ($\hat{\eta}^{MLE}$) is given by setting $\frac{\partial l}{\partial \eta} = 0$.

So $\hat{\eta}^{MLE}$ satisfies

$$\sum_{n=1}^{N} t(x_n) = N \frac{da}{d\eta} (\hat{\eta}^{MLE}).$$

However, as $a'(\eta) = \mathbb{E}(t(X))$, we have that $\hat{\eta}^{MLE}$ satisfies, the far more practical equality,

$$\frac{1}{N} \sum_{n=1}^{N} t(x_n) = \mathbb{E}(t(X)).$$

References

[1] Shakarchi, R. and Stein, E. *Real Analysis: Measure Theory, Integration, and Hilbert Spaces*, Princeton Lectures in Analysis III, 2005.

[2] Wasserman, L. *All of Statistics*, Springer, 2004.