Notes and Solutions for "Theoretical Statistics" by Robert W. Keener

Arya Pourzanjani

May 19, 2016

1 Motivation

Many of the popular distributions for random variables belong to the exponential family of distributions, including the normal, the binomial, and the Poisson. We list some applications, and the exponential family distribution they can be modeled with:

- **The Normal Distribution and Population Parameters:** The distribution of certain traits of a population such as height, weight, or IQ can often be modeled as normally distributed. If we are studying injuries we can model the speed at which blood clots across a population. Perhaps we can assign two different normal distributions for a subpopulation of patients with brain injuries and subpopulation of patients without.
- **Waiting Times and the Exponential Distribution:** In certain cloud computing settings we could be concerned with the possibility that a remote instance dies on us. The time until an instance dies can be modeled as an exponential random variable.
- **Classification and Bernoulli Distribution:** Any sort of classification task can be modeled using a Bernoulli random variable. If there are multiple classes a Multinoulli distribution can be used instead.
- **Counts and the Poisson Distribution:** Documents such as doctor's notes or newspaper articles are often modeled by how many times certain words show up. A doctor's note that describes a patient who was in a motor vehicle crash will probably contain multiple occurrences of the word car and driver. These counts can be modeled by the Poisson distribution. Another domain where we see counts is in chemical and biological reactions. The number of times a yeast cell will mate in an hour can be modeled as a Poisson random random variable.

Exponential families all come in a single form, which means whatever we learn about them applies to many useful and common distributions. Likewise software

for using exponential family distributions can be written generally but applied to a wide variety of distributions and use cases. In fact, we will later on see that all exponential family distributions are easy to fit because they all have convex likelihoods.

Despite the wide-use of simplicity of exponential families they are still an active area of research. Neural networks can be thought of as modeling a Bernoulli random variable whose mean is the linear combination of other Bernoulli random variables who themselves have means that are modeled as more Bernoulli random variables! Ranganath and et al point this out in their paper "Deep Exponential Families", and give examples of where creating a hierarchy of exponential family random variables is useful, but thankfully still tractable!

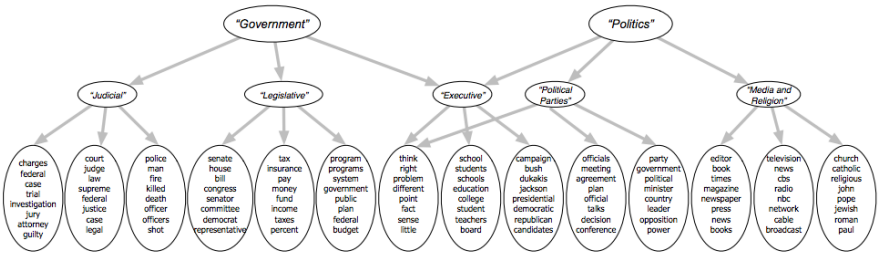

Figure 1: Poisson deep exponential network depicting learned concepts from New York Times Articles.

2 Definition

Let X be a random variable. We don't know what value it will take on, but we might know certain values are more probable than others and we might know the probability of it falling in a certain range, i.e. we might know the density of the random variable. For a given value of X and given values of parameters the density gives us a real number (that is incidentally not a probability).

If X is a member of the exponential family it has a density that can be written as follows:

$$
\begin{aligned}
p_\eta(x) &= \exp\left[\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta)\right] h(x) && (1)\\
&= \exp\left[\sum_{i=1}^{s} \eta_i T_i(x)\right] \exp\left[-A(\eta)\right] h(x) \\
&= \frac{\exp\left[\sum_{i=1}^{s} \eta_i T_i(x)\right] h(x)}{\exp\left[A(\eta)\right]} && (2)
\end{aligned}
$$

Our specific choice of the functions $T_i(x)$, the number of parameters s, and $h(x)$ function define a set of allowable distributions $\{p_\eta(x)\}$ indexed by η. For example, a certain choice of $T_i(x)$ will lead to the set of all the one-dimensional normal distributions.

2.1 The Cumulant Generating Function

The function $A(\eta)$ is a normalization term called the **cumulant generating function** in statistics or the **free energy function** in physics. It is defined as follows:

$$A(\eta) := \log \int \exp\left[\sum_{i=1}^{s} \eta_i T_i(x)\right] h(x)\, d\mu(x). \tag{3}$$

We only define this function for values of η that make the integral converge. That is, we can't just consider any value of η, but only the subset of values of η where we have finite free energy. This leads us to two very special properties of the cumulant generating function that we will discuss in detail:

- The set of valid values for η is a convex set in parameter space and the cumulant generating function itself is a convex function over these values, which will make fitting of exponential family models easy.
- The derivatives of the cumulant generating function are the expected values of the functions $T_i(x)$, which are called sufficient statistics.

The following example illustrates the first point. The second point we will touch on later.

Example The Exponential Distribution The exponential distribution of the exponential family is used to model failure times. Its density is derived by picking $h(x) = 1_{(0,\infty)}(x)$, $s = 1$, and $T_1(x) = x$. This gives us the unnormalized density

$$\widetilde{p}_\eta(x) = e^{\eta x}\, 1_{(0,\infty)}(x) \tag{4}$$

To normalize we calculate the cumulant generating function

$$\begin{aligned} A(\eta) &= \log \int_0^\infty e^{\eta x}\, 1_{(0,\infty)}(x)\, dx \\ &= \begin{cases} \log(-1/\eta), & \eta < 0 \\ \infty, & \eta \geq 0 \end{cases} \end{aligned} \tag{5}$$

Thus our normalized density is only defined when η is less than zero and it is

$$p_\eta(x) = \exp[\eta x - \log(-1/\eta)]1_{(0,\infty)}(x) \tag{6}$$

Note that the cumulant generating function is a convex function, and the possible η values that are valid form a convex set.

2.2 Reparameterizations (Non-canonical forms)

Often times we will be interested in parameters that are not nicely linearly related to the sufficient statistics. For example, in the normal distribution our two parameters are the mean, μ and the variance σ^2. Any normal distribution is thus indexed or parameterized by $\theta := (\mu, \sigma^2)$, that is its density is

$$\begin{aligned} p_\theta(x) &= \frac{1}{2\pi\sigma^2}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \\ &= \frac{1}{\sqrt{2\pi}}\exp\left[\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \left(\frac{\mu^2}{2\sigma^2} + \log\sigma\right)\right] \end{aligned} \tag{7}$$

Although we don't we have separated out functions of x, from the parameters, namely $T_1(x) = x$ and $T_2(x) = x^2$ the coefficients they are multiplied by are not simply our parameters μ and σ^2, but some other function of them $\eta_1(\theta)$ and $\eta_2(\theta)$. We call $\eta(\theta)$ a **reparameterization** of θ.

The transformation from θ to η is one-to-one, and thus we can use either one. If we use η however then the result is an exponential family distribution that is in **canonical form**.

$$\begin{aligned} p_\eta(x) &= \frac{1}{\sqrt{2\pi}}\exp\left[\eta_1(\theta)x + \eta_2(\theta)x^2 - B(\theta)\right] \\ &= \frac{1}{\sqrt{2\pi}}\exp\left[\eta_1(\theta)T_1(x) + \eta_2(\theta)T_2(x) - B(\theta)\right] \end{aligned} \tag{8}$$

The reason we don't always use canonical parameterizations is because sometimes a certain parameterization like the $\theta = (\mu, \sigma^2)$ parameterization has an intuitive interpretation, like μ being a mean, whereas it's harder to interpret what $\eta_1(\theta) = \mu/\sigma^2$ means.

Lastly, it's worth noting that if we plotted the valid values of $\eta_1(\theta) = \mu/\sigma^2$ and $\eta_2(\theta) = 1/(2\sigma^2)$ on the xy-plane the only valid values would be in the upper right quadrant, which is a convex set.

2.3 IID Samples of Exponential Families

If $X_1, \cdots, X_n$ is a random sample from an exponential family density

$$\exp\left[\sum_{i=1}^{s} \eta_i(\theta) T_i(x) - B(\theta)\right] h(x), \tag{9}$$

then the joint distribution of these random variables is also parameterized by η and is just the product of each random variables marginal density by itself

$$\begin{aligned} p_\eta(x_1, \cdots, x_n) &= \prod_{j=1}^{n} \exp\left[\sum_{i=1}^{s} \eta_i(\theta) T_i(x_j) - B(\theta)\right] h(x_j) \\ &= \exp\left[\sum_{i=1}^{s} \eta_i(\theta) \left(\sum_{j=1}^{n} T_i(x_j)\right) - nB(\theta)\right] \prod_{j=1}^{n} h(x_i) \end{aligned} \tag{10}$$

This result is itself an exponential family though! The random variable in consideration here is actually the random vector $x = (x_1 \cdots x_n)$ and the sufficient statistics $\widetilde{T}_i(x)$ are just

$$\widetilde{T}_i(x) = \sum_{j=1}^{n} T_i(x_j). \tag{11}$$

3 Differential Identities of the Cumulant Generating Function