

# ESTIMATIONS WITH EXPONENTIAL FAMILIES

Doctor SVN Vishwanathan

April 7, 2008

## 1 Defining Exponential Families

First of all, the question is: what *are* exponential families? They are families of parameterized distributions.

The generalised form of distributions is:

$$P(x, \theta) = P_0(x) \exp(\langle \phi(x), \theta \rangle - g(\theta))$$

$P$  is a distribution.  $x$  stands for observations from the domain  $\mathcal{X}$ .  $\theta$  is a parameter.  $P_0(x)$  is the base measure;  $x$  has been drawn from a space where there is a distribution.  $\phi(x)$  is the sufficient statistics.  $g(\theta)$  is the log-partition function; it normalizes the distribution to make it sum to 1. We will study the log-partition function further.

Intuitively,  $\phi(x)$  is the result of adding questions regarding the observations: features. If  $x$  is the class of all documents,  $\phi(x)$  might be questions like “does this word occur in the document”, “what’s the length of the document”, “what language is this document in”, “with what frequency does this word occur”.

Almost every distribution that you can name is a member of the exponential family: Gaussian (which only requires the mean and the variance to graph), Poisson, Laplace, and so on. There is a rich family of distributions to use for estimations in machine learning.

The parametrized family, eg. the Gaussian exponential family, have two levels of selection:

1. Family: choose a member of the family you want to work with, eg. Gaussian, Laplace, lambda, Poisson.
2. Members: choose from the members in the family.

Example:

1. The family chosen is Gaussian; the model will be Gaussian.
2. The data will fit to a mean of 0 and a variance of 2.

Typically, choosing  $\phi(x)$  implies the family has been chosen. Example:

$$\phi(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix}$$

This means that the family of distributions are Gaussians. Next, choose  $\theta$ —the mean and the variance—and you’ve fixed a member of the family.

Choosing a member of the family means choosing  $\phi(x)$  and  $\theta$ , or  $g(\theta)$ . (The log-partition function is connected to the others.)

Choosing the log-partition function can force the choosing of the family. Choosing  $\phi(x)$  is sometimes a sufficient strategy, though for most of the course we will not worry about  $\phi(x)$ . We are not focusing on a family working with some feats and some families.

Most of the work is concentrating on *how* to find  $\theta$  for given data. We decide to fit a Gaussian or other distribution to the data, and then decide which possible Gaussian is the best choice: that is finding  $\theta$ .

*Examples of families.*

A simple example for the *Bernoulli* distribution is the flip of a coin.

$$Pr(\text{head}) = P$$

$$Pr(\text{tail}) = 1 - P$$

$$P(\text{head}) = (x = 1)$$

$$P(\text{tail}) = (x = 0)$$

Equations:

$$P(x) : x \in 0, 1 \in X$$

$$Pr(x) = P(x) = p^x(1-p)^{1-x}$$

$$P(x) = \exp \log P(x)$$

$$P(x) = \exp(x \log p + (1-x) \log(1-p))$$

$$P(x) = \exp\left(\begin{pmatrix} x \\ 1-x \end{pmatrix}^T \begin{pmatrix} \log p \\ \log(1-p) \end{pmatrix}\right)$$

$$\phi(x) = \begin{pmatrix} x \\ 1-x \end{pmatrix}^T$$

$$\theta = \begin{pmatrix} \log p \\ \log(1-p) \end{pmatrix}$$

(Note: there is a restriction of  $p$  and  $p-1$ .)

$$\exp\left(\begin{pmatrix} x \\ 1-x \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}\right) - g(\theta)$$

$$g(\theta) = \log(\exp(\theta_1) + \exp(\theta_2))$$

$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \in \mathbb{R}$$

Another example is the Laplace distribution, using the parameter  $\theta$ . In the time interval  $[x, x+dx]$ , the probability that an atom still exists is proportional to  $\theta x$ . The probability that an atom still exists in this time is  $\theta dx$ .

$$P(x) = \theta \exp(\theta x)$$

$$P(x) = \exp(\langle -x, \theta \rangle - (\log \theta))$$

$\langle -x, \theta \rangle$  is the sufficient statistics ( $-x$ ), and  $\log(\theta)$  is  $g(\theta)$ .  
Or consider the normal distribution.

$$P(x) = \frac{-1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

$$P(x) = \exp\left(\frac{-x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} - \left(\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)\right)\right)$$

$$P(x) = \exp\left(\begin{pmatrix} x \\ x^2 \end{pmatrix}^T \begin{pmatrix} \frac{\mu}{\sigma^2} \\ \frac{-1}{2\sigma^2} \end{pmatrix} - g(\theta)\right)$$

$$\phi(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix}$$

$$g(\theta) = \frac{1}{4}\theta_1^2\theta_2^{-1} + \frac{1}{2}\log(2\theta_2)$$

## 2 Properties of the Exponential Family

$$P(x) = P_0 \exp(\langle \phi(x), \theta \rangle - g(\theta))$$

$$P(x) = P_0(x) \frac{\exp(\langle \phi(x), \theta \rangle)}{Z}$$

$$Z = \exp(g(\theta))$$

$$\int_x P(x, \theta) - 1$$

$$\int_x P(x, \theta) = \frac{\int_x P_0(x) \exp(\langle \phi(x), \theta \rangle) dx}{\exp(g(\theta))}$$

Thus:

$$g(\theta) = \log \int_x P_0(x) \exp(\langle \phi(x), \theta \rangle) dx$$

(This is very expensive to compute if  $x$  is large. We use techniques for computing the integral.)

$g(\theta)$  has many nice properties. Since the function is in the form of a log of a  $\int$  of an exp: we can take derivatives and differential. (It's infinitely differentiable.)

$$g(\theta) = \log \int_x P_0(x) \exp(\langle \phi(x), \theta \rangle) dx$$

$$\delta_\theta g(\theta) = \frac{\int P_0(x) \cdot \phi(x) \cdot \exp(\langle \phi(x), \theta \rangle) dx}{\int P_0(x) \exp(\langle \phi(x), \theta \rangle) dx}$$

$$\delta_\theta g(\theta) = \frac{\int P_0(x) \cdot \phi(x) \cdot \exp(\langle \phi(x), \theta \rangle) dx}{\exp(g(\theta))}$$

$$\delta_\theta g(\theta) = \int P_0(x) \phi(x) \exp(\langle \phi(x), \theta \rangle - g(\theta)) dx$$

$$\delta_\theta g(\theta) = \int \theta(x) P(x, \theta) dx$$

$$\text{Expectation} = \mathbb{E}[\phi(x)]$$

$$\delta_{\theta}^2 g(\theta) = \text{variance}$$

$$\delta_{\theta}^2 g(\theta) = \int_x \phi(x) \delta_{\theta} P(x, \theta) dx$$

$$\delta_{\theta}^2 g(\theta) = \int_x \phi(x) \{\phi(x) - \delta_{\theta} g(\theta)\}^T P(x, \theta) dx$$

$$\delta_{\theta}^2 g(\theta) = \int_x \phi(x)^2 P(x, \theta) dx - \mathbb{E}_{P(x, \theta)}(\phi(x) \delta_{\theta} g(\theta)^T)$$

$$g(\theta) = \log \int_x P_{\theta}(x) \exp(\langle \phi(x), \theta \rangle) dx$$

$$\mathbb{E}\{\phi(x) \phi(x)^T\} - \mathbb{E}[\phi(x)] \mathbb{E}[\phi(x)]^T$$

(The formula for variance.)

The second derivative of a convex function is always positive and semi-definite.

One way to show convexity is the second derivative covariance matrix, to see if it's positive and semi-definite; then  $g(\theta)$  will be convex.

It is very useful when doing estimation to know that there is always a global minimum.

Does every log-partition correspond with a log-partition function? Ask what the inverse relationship is to find out. You can't get an exponential family out of any convex function, so the answer is no. Whether every convex function corresponds to a *family*, though, is unknown.

(Note: in the semi-definite matrix, not all entries have to be positive.)

A log-partition function: must be convex; it generates movements of distributions, and is differentiable.

$$P(x, \theta) = P_0(x) \exp \langle \phi(x), \theta \rangle - g(\theta)$$

How do we use the nice properties?

For example, we can observe some random variable:

$$\{x_0, x_1, \dots, x_m\} \in X$$

(The  $x$ es can be documents, or anything.)

We can choose to fit a Gaussian to the data: we have to find Gaussian parameters which explain how the data is generated. We make the standard

assumptions (the data is identical and independently distributed), which may or may not be correct. What is the probability of set  $X$  being so distributed?

$$\prod_{i=1}^M P(x_i, \theta) = P(x_i, \theta)$$

How do we estimate the best  $\theta$ ? We use the maximum likelihood estimation—we insure that the data was generated by the most likely distribution. We can maximise:

$$\max_{\theta} P(x_i, \theta)$$

$$\theta = \arg \max_{\theta} P(x_i, \theta)$$

Or else, the log-likelihood of the data can be minimised.

$$\min_{\theta} -Z \log P(x_i, \theta)$$

(assuming  $P(x_i, \theta)$  was generated from the family distribution.)

Remember that  $\phi(x)$  is the features. Once that has been given, the family is known. It only remains to estimate the parameters.

$$\sum_{i=1}^m - \langle \phi(x_i), \theta \rangle + mg(\theta)$$

$$\min_{\theta} mg(\theta) \cdot m \sum_{i=1}^m \langle \phi(x), \theta \rangle$$

$$\min_{\theta} g(\theta) \cdot \sum_{i=1}^m \langle \phi(x), \theta \rangle$$

$$\min_{\theta} g(\theta) - \langle \varphi, \theta \rangle$$

( $g(\theta)$  and  $\langle \varphi, \theta \rangle$  are both convex or convex-linear functions.)

$$\varphi = \frac{1}{m} \sum \phi(x_i)$$

A convex function reaches the minimum where the gradient of the curve is zero.

$$\delta_{\theta} g(\theta) = \varphi$$

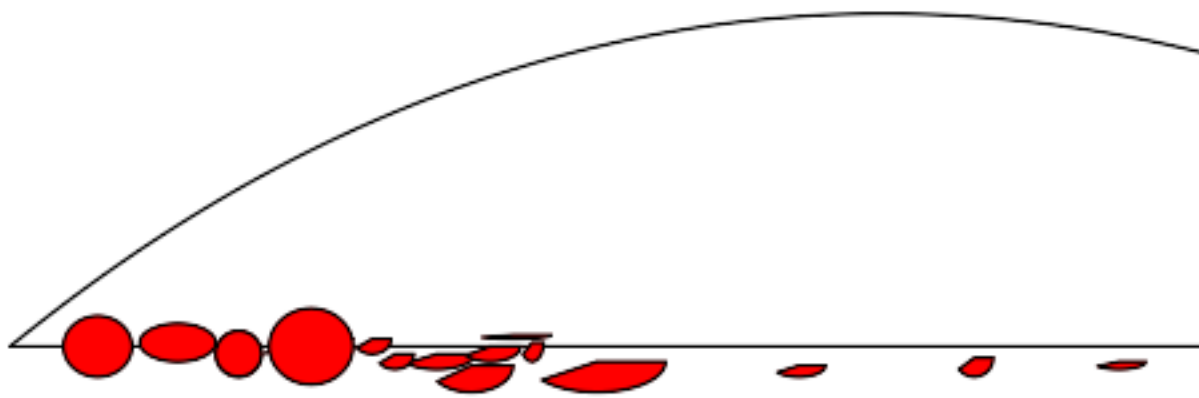
(The gradient of the log-partition function is the expectation;  $\varphi$  is the sample mean.)

$$\delta_{\theta} g(\theta) = \mathbb{E}[\phi(x)]$$

You can then apply the reverse solver; computing the gradient and taking expectations, using the maximum likelihood estimate.

MLE, however, is not robust to outliers, unless there is an infinite data limit.

Example of an overdawn curve due to outliers:



MAP estimates can help when there are outliers. Overall, MLE is not generally good, especially on outliers or small amounts of data.

To be able to compute the expectation  $g(\theta)$ , you must be able to compute  $\phi(x)$ , and do integrations over the entire domain.

$$\int \phi(x)P(x_i, \theta)dx$$

We can do two things:

1. Given data, we can estimate density.
2. We can predict labels of data (a more common request).

Looking at the joint density, a lot of people say: can we arrive at a joint estimate? One problem with joint density is that things may be marginalised; this can lead to a wasted effort in modelling the data itself.

Example: a task to identify digits and discover whether they are integers or not. One possible way to build the machine is to model all possible ways to generate an integer and find a label. Now, given an instance of a digit, we have to get the probability that its label is “integer”. This involves learning a lot of extra information, which should not be wasteful (for example, we should not have to learn how to write an integer). Instead, we could say: “Given that I have seen the observations, I can tell you the probability it is an integer.”

We can then immediately see that it makes sense to model, not the joint distribution, but  $P(y|x)$ ,  $P(x, y, \theta)$ . When modelling exponential families, we can still model for conditional distribution.

$$P(y|x_i, \theta) = \exp(\langle \phi(x, y), \theta \rangle - g(\theta|x))$$

$$g(\theta|x) = \log \int_x \exp(\langle \phi(x, y), \theta \rangle) dy$$

Conditional models usually make log partition computation easier.

Example: given some labelled data, the task is to build a classifier. (This will be easier to do if the set of labels is small.) We have  $X$  and  $Y$  where  $X$  represents the set of data and  $Y$  the set of labels ( $x_{1..m} \in X$ ,  $y_{1..m} \in Y$ ). We are to discover  $P(Y|X, \theta)$ .

$$\min_{\theta} -\log P(y|x_i, \theta)$$

$$\min_{\theta} \sum_{i=1}^m -\langle \phi(x_i, y), \theta \rangle - g(\theta|x_i)$$

The gradient is set to zero.

$$\sum_{i=1}^m \delta_{\theta} g(\theta|x) = \sum_{i=1}^m \phi(x_i, y_i)$$

$$\mathbb{E}_{P(y|x, \theta)}[\phi(x, y)] = \sum_{i=1}^m \phi(x, y_i)$$

We can take the Bayesian approach. We should have some knowledge as to where the parameters are going to lie.

Setting the prior:  $p(\theta)$ ; we should know some  $\theta$ , and some distribution over the parameters. Then the data arrives, and changes the prior to the posterior. We can convert the prior to the posterior using Bayes' rule:

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$

$$P(\theta|x, y) = \frac{P(y|x; \theta)P(\theta)}{P(y, x)}$$

$$\propto P(y|x_i, \theta)P(\theta)$$

If the distribution is a Gaussian,  $P(\theta)$  is normally distributed and we could choose that prior.

Using that prior and the exponential family is a *gaussian process*.

Question: Assume we have two variables,  $x$  and  $y$ .  $x$  is normally distributed with a mean and variance;  $y$  is also normally distributed. Covariance matrix:

$$\begin{pmatrix} A & C \\ C^T & B \end{pmatrix}$$

What is the distribution of  $P(x|y)$ ?

$$x \sim N(a, A)$$

$$y \sim N(b, B)$$

$$z = \begin{pmatrix} x \\ y \end{pmatrix}$$

$$z \sim N\left(\begin{pmatrix} a \\ b \end{pmatrix}, \begin{pmatrix} A & C \\ C^T & B \end{pmatrix}\right)$$

Chances are that it is normally distributed.

$$X|Y \sim N(\text{mean}, \text{variance})$$

Two random variables are normal: conditioning on one should be joint distributed as well.

What if you extend the concept to the stochastic process? Example:

$$t(x) \Rightarrow \mathbb{R}$$

This is a stochastic process.

If we have a stochastic process and integrate it at  $m$  number of points, the result is a collection of  $m$  random variables. If  $m$  variables are normally distributed, the stochastic process is a Gaussian; only the mean and variance are needed to specify the Gaussian.

We must know the  $t(x)$  expectation and covariance evaluation.

$$\mathbb{E}[t(x)] = \varphi(x)$$

$$\text{Covariance}(t(x), t(x')) = k(x, x')$$

$$t(x) = \langle \phi(x), \theta \rangle$$

Mapping function:

$$t(x) \Rightarrow \mathbb{R}$$

We have a whole parameterized family of functions to work with.  $\theta$  is drawn from a distribution. To get the value of  $t(x)$ , can first take  $x$ , pass it through  $\phi(x)$ , and then get the dot product.

$$\theta \sim N[0, \sigma]$$

Given  $x$ : do  $\phi(x)$ , choose  $\theta$ , and get a number of random variables because  $\theta$  is not fixed. If you are placing a prior over  $\theta$ , note that all the parameterized family of linear functions have a prior. You can place a prior over an entire family of functions.

Why is it a Gaussian?

$$\mathbb{E}_\theta[t(x)] = \langle \phi(x), \mathbb{E}[\theta] \rangle$$

$$\mathbb{E}[\langle \phi(x), \theta \rangle \langle \phi(x), \theta \rangle]$$

In the family of functions,  $\mathbb{E}$  with regard to all the family members is 0.

$$= \sigma^2 \phi(x)' \phi(x) - \sigma^2 \langle \phi(x), \phi(x') \rangle$$

$$= k(x, x')$$

$$-\log(x, \theta) + g(\theta) = \langle \phi(x), \theta \rangle$$

$$p(\theta) \propto \exp\left(\frac{-1}{2\sigma^2} \|\theta\|^2\right)$$

$$\prod_{i=1}^m \exp(\langle \phi(x_i, y), \theta \rangle - g(\theta|x_i)) \exp\left(\frac{-1}{2\sigma^2} \|\theta\|^2\right)$$

$$t(y|x) = \langle \phi(x, y) | \theta \rangle$$

For Gaussians, you can place priors over your function class to use for estimation. You can relate it back to the exponential family and estimate by maximising the posterior.

Conditional models:

$$P(y|x_i, \theta) = \exp \langle \phi(x, y), \theta \rangle - g(\theta|x)$$

We are trying to make an estimate.

Having performed inference, we have found a fixed value of  $\theta$ . In ML estimation: we have lots of data and labels given, then we use test data to predict labels. How are we going to do that? For each  $y$ , see the expression:

$$P(y|x, \theta)$$

Each label has a probability of occurring—maybe the labels range from 0-9 for the digit identification problem. Given some input, we have a distribution over the digits. How do we produce a digit? We can look at the label  $y$  with the maximum probability.

$$\arg \max_y \exp(\langle \phi(x, y), \theta \rangle)$$

or:

$$\arg \max_y \exp(\langle \phi(x, y), \theta \rangle)$$

The task, of course, is when given a fixed  $x$ , to predict the label  $y$ . The estimation function is  $g(\theta|x)$ . We find  $\theta$  by using Bayes' rule. The posterior distribution of  $\theta$  given  $x$  and  $y$  is proportional to the prior  $xP(y|x, \theta)$ :

$$P(\theta|x, y) \propto P(\theta) \cdot P(y|x_i, \theta)$$

Can you find the maximum value of that distribution?

$$\max_{\theta} P(\theta|x_i, y)$$

$$\min_{\theta} -\log P(y|x, \theta) - \log P(\theta)$$

$$\theta \sim N(0, \sigma^2, \mathbb{I})$$

$$\min_{\theta} \frac{1}{2\sigma^2} \|\theta\|^2 + \sum_{i=1}^m g(\theta|x_i) - \langle \phi(x_i, y_i), \theta \rangle$$

*normal prior* = *Gaussian*.

### 3 Summary

1. What is an exponential family of distributions?
2. Examples of exponential family of distributions. (See also: Novi's notes.)
3. Properties of the log-partition function.
  - Complex
  - Moment generating
1. Exponential families in unconditional estimation.
2. Exponential families in conditional estimation.
3. Placing a normal prior over your parameters implies that you are working with a Gaussian.