CS281A/Stat241A: Statistical Learning Theory

Exponential Family

Lecturer: Michael I. Jordan *Scribes: Barbara Engelhardt and Xuanlong Nguyen*

1 Exponential Family

The general distribution describing the exponential family is as follows:

$$P(x|\eta) = h(x)\exp\{\eta^T T(x) - A(\eta)\},$$

where

- $x = (x^1, ..., x^m)$.
- $T(x) = [T_1(x), ..., T_k(x)]^T$ is the *sufficient statistic* vector. $T(x)$ is also known as a set of *features* on x.
- η is the *canonical parameter*, or the natural parameter.
- $A(\eta)$ is the *cumulant generating function*, or log normalization.
- $h(x) \geq 0$ is the *reference density*, or base density.

This equation defines a generic distribution on a graphical model. In particular, one can use a factor graph to represent an exponential probability distribution (see Figure 1), where each term $\exp\{\eta_k T_k(x)\}$ corresponds to a factor in the graph. A nice fact about exponential families is that the product of densities in an exponential family is also in the exponential family.

Notice that the exponent contains an affine function of $T(x)$. The choice of $T(x)$, or the sufficient statistic, defines a specific family of distributions.

Since $A(\eta)$ is a normalizing factor, it is constrained by a choice of $T(x), \eta$ and $h(x)$, given that the integral of the distribution over x must be 1 since the distribution is a probability measure:

$$1 = \int h(x)\exp\{\eta^T T(x) - A(\eta)\}\mathrm{d}x$$

$$= e^{-A(\eta)} \int h(x)\exp\{\eta^T T(x)\}\mathrm{d}x,$$

from which we can conclude that

$$A(\eta) = \log\left[\int h(x)\exp\{\eta^T T(x)\}\mathrm{d}x\right],$$

in other words, $A(\eta)$ acts as a normalization term for the distribution.

There are two properties of the exponential family that we can use in subsequent proofs.

Property 1: $\{\eta : A(\eta) < \infty\}$ is convex.

Property 2: $A(\eta)$ is a convex function of η.

These will be described in more detail and property 2 will be proved after the examples.

2 Examples

The exponential family is pervasive in statistics. We will go through a few examples of how it can be used in some distributions we are already familiar with.

2.1 Bernoulli Distribution

$$\begin{aligned} p(x|\pi) &= \pi^x (1-\pi)^{1-x} \\ &= \exp\{\ln(\pi^x (1-\pi)^{1-x})\} \\ &= \exp\{x \ln \pi + (1-x)\ln(1-\pi)\} \\ &= \exp\left\{\ln\left(\frac{\pi}{1-\pi}\right) x + \ln(1-\pi)\right\} \end{aligned}$$

Writing this out in terms of exponential parameters:

- $h(x) = 1$
- $\eta = \ln \frac{\pi}{1-\pi}$
- $A(\eta) = -\ln(1-\pi)$
- $T(x) = x$

Note that $\eta = \ln \frac{\pi}{1-\pi}$ is some function of the underlying parameterization of the distribution. There is a one-to-one mapping between π and η and this is generally true in all distributions in the exponential family. We will go over this one-to-one mapping in more detail in later sections.

In the case of the Bernoulli distribution, $\eta = \ln \frac{\pi}{1-\pi} \Leftrightarrow \pi = \frac{1}{1+e^{-\eta}}$. Here, η is also known as the *log-odds ratio*. Note that the inverse of η, which is π, is the logistic function. Notice that we used the logisitic function to model the conditional (posterior) distribution for a variable with a Bernoulli distribution, which is a special case of the GLIM model to be discussed in future lectures.

$A(\eta) = -\ln(1-\pi) = \log(1+e^{\eta})$ is the cumulant generating function for the Bernoulli distribution.

2.2 Poisson Distribution

$$\begin{aligned} p(x|\lambda) &= \frac{\lambda^x e^{-\lambda}}{x!} \\ &= \frac{1}{x!}\exp\{x \ln \lambda - \lambda\} \end{aligned}$$

where $h(x) = \frac{1}{x!}$, $\eta = \ln \lambda$, $T(x) = x$, and $A(\eta) = \lambda = e^{\eta}$.

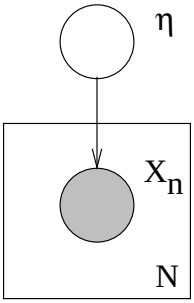

Figure 1: Graphical model describing the generative model of the exponential family and its parameter η.

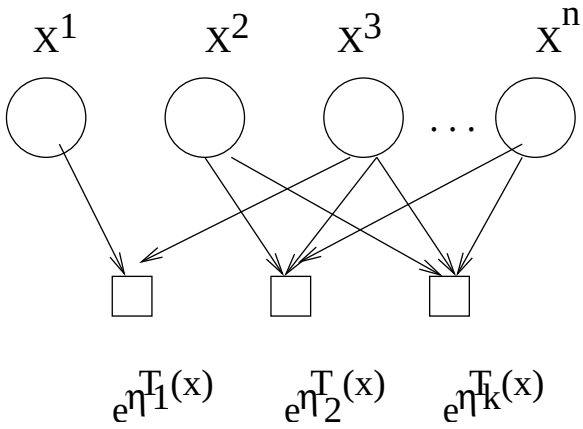

Figure 2: Factor graph describing features in the exponential family, which can be functions of arbitrary combinations of the variables X.

2.3 Gaussian Distribution

Note that this is the single dimensional case. As an exercise, extend this to the multidimensional case.

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

where, using the exponential family parameterization,

- $h(x) = \frac{1}{\sqrt{2\pi}}$.
- $\eta = [\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}]^T$.
- $T(x) = [x, x^2]^T$.
- $A(\eta) = \frac{-\mu^2}{2\sigma^2} + \ln\sigma$

It is a straightforward exercise to invert η to establish the one-to-one mapping with the parameters θ for the univariate Gaussian distribution.

3 Exponential Family Properties

We will prove the following very using fact in full generality.

Fact 1: $\bigtriangledown_\eta A(\eta) = E[T(x)]$

Example: For Bernoulli distribution we have:

$$E[T(x)] = EX = \bigtriangledown_\eta \log(1+e^\eta)$$

$$= \frac{e^\eta}{1+e^\eta}$$

$$= \frac{1}{1+e^{-\eta}} = \pi.$$

Fact 2: $\bigtriangledown^2_\eta A(\eta) = Var[T(x)]$

Example: For Bernoulli distribution,

$$Var[T(x)] = \frac{d^2A}{d\eta^2}$$

$$= \frac{d}{d\eta}\frac{1}{1+e^{-\eta}}$$

$$= \pi(1-\pi)$$

In fact, repeatedly taking the derivative of A with respect to η will yield the cumulants, or central moments, of $T(x)$. Specifically, taking the derivative four times give the mean, variance, and kurtosis of $T(x)$, which are not to be confused with the first, second, and fourth moments of $T(x)$.

Proof of Fact 1: Using the fact that the integral over x of the density $p(x|\eta)$ must equal 1, we can rewrite $A(\eta) = \int h(x)e^{\eta^T T(x)}\mathrm{d}x$. Taking the derivative wrt η gives

$$\begin{aligned}\bigtriangledown_\eta A(\eta) &= \bigtriangledown_\eta \log\{\int h(x)e^{\eta^T T(x)}\mathrm{d}x\} \\ &= \frac{\int h(x)e^{\eta^T T(x)}T(x)\mathrm{d}x}{\int h(x)e^{\eta^T T(x)}\mathrm{d}x} \\ &= \int h(x)e^{\eta^T T(x)-A(\eta)}T(x)\mathrm{d}x \\ &= \int P(x|\eta)T(x)\mathrm{d}x \equiv E[T(x)]\end{aligned}$$

which completes the proof.

Proof of Fact 2: Taking second derivative wrt η in the same manner give:

$$\bigtriangledown_\eta^2 A(\eta) = \int T(x)h(x)e^{\eta^T T(x)-A(\eta)}(T(x) - \bigtriangledown_\eta A)^T\mathrm{d}x$$

but notice that $\bigtriangledown_\eta A = E[T(x)]$ from Fact 1,

$$\begin{aligned}&\int T(x)T^T(x)p(x|\eta)\mathrm{d}x - \int T(x)p(x|\eta)E[T^T(x)]\mathrm{d}x \\ &= E[T(x)T^T(x)] - E[T(x)]E[T^T(x)] \equiv Var[T(x)]\end{aligned}$$

This completes the proof.

Returning to the first section, we can now easily show Property 2.

Proof of Property 2:

$\bigtriangledown_\eta^2 A(\eta) = Var[T(x)] > 0$ is positive semidefinite, and is (strictly) positive definite when $T(x)$ is not constant. This implies the convexity of $A(\eta)$.

Fact 1 implies that $\mu = E[T(x)]$ is a function of the canonical parameter η through the relation: $\mu = \bigtriangledown_\eta A(\eta)$. Moreover, we can easily prove:

Theorem: The relationship between μ, η is invertable.

To be precise, we need the condition that the function $T(x)$ is not a constant. It follows from the proof of Property 2 that $A(\eta)$ is *strictly* convex, which implies that the first derivative of $A(\eta)$ is strictly monotone. In other words, μ is a strictly monotonic function of η. This implies that this function is invertible.

Example: Poisson distribution

$A(\eta) = e^\eta$, then $E[T(x)] = E(x) = e^\eta = \lambda$.

Notice that $Var(x) = \lambda$, and in fact since every subsequent derivative of e^η is also λ, all higher order cumulants are λ for the Poisson distribution.

Example: Gaussian distribution

It is easy to see that:

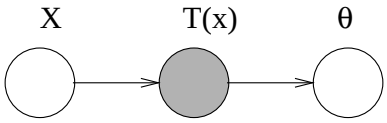

Figure 3: Graphical model representing the Bayesian perspective on the conditional independence relationships for sufficient statistics.

$$\begin{aligned} E[T(x)] &= E([x, x^2]^T) = [\mu, \mu^2 + \sigma^2]^T \\ &= \bigtriangledown_\eta (\frac{-\mu^2}{2\sigma^2} + \ln \sigma) \\ &= \bigtriangledown_\eta A(\eta) \end{aligned}$$

The canonical parameter η can be written as $\eta(\theta)$, which sets up the relationship between η and θ. The power of the exponential family is that a difficult integral (e.g., $E_\theta(x)$) can be found by taking a (possibly much simpler) derivative of function $A(\eta)$. Alternatively, expectations may be better calculated as integrals. Recall the example in the homework of the Dirichlet distribution: $E(\theta_i)$ was simple to calculate using the integral; $E(\log \theta_i)$ is easier to calculate from the derivative of $A(\eta)$, using the exponential family.

4 Sufficient Statistics

In this section we formalize the notion of *sufficient statistic*.

Let X be a random variable. A function $T(X)$ of X is called a *statistic*, if $T(X)$ does not depend on a parameter of the distribution from which X was generated (e.g., θ).

A statistic $T(X)$ is *sufficient* for θ if there is no information in X regarding θ beyond that in $T(X)$. In other words, if one can throw away data X and keep $T(X)$ instead without losing any information about θ (i.e. the distribution $P(X|\theta)$), then $T(X)$ is a sufficient statistic for X. Often $T(X)$ is very compact, and this results in a significant reduction in the size of the data collected, while no information about the value of θ is lost.

Example: Univariate Gaussian distribution. The running sum and N, the number of samples, is one sufficient statistic for the Gaussian distribution. Specifically $\hat{\mu} = \frac{1}{N} \sum_{n=1}^{N} x_n$, or the running mean, can be calculated using these two numbers, instead of storing data of size N: $x_1, x_2, ..., x_N$.

Now we will be more precise about our definition of sufficient statistic following both Bayesian and frequentist approaches. The Bayesian approach provides a particularly natural definition and interpretation.

4.1 Bayesian Sufficiency

Parameter θ can be viewed from the Bayesian perspective as a random variable. Then $T(x)$ is sufficient for θ if and only if $\theta \amalg X|T(x)$. This conditional independence relationship can be represented by the graphical model in figure (3). For this chain, the parameter θ is independent of the data X given the sufficient statistic $T(x)$.

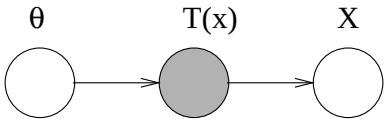

Figure 4: Graphical model representing the frequentist perspective on the conditional independence relationships for sufficient statistics.

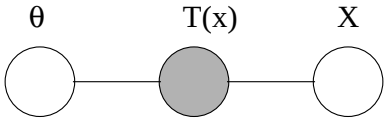

Figure 5: Graphical model representing an undirected graph, which is equivalent to the Bayesian and frequentist graphs describing the conditional independencies involved in sufficient statistics.

4.1.1 Frequentist Sufficiency

Frequentist sufficiency can be summed up by the following quote: "$T(x)$ is sufficient for θ if and only if the conditional distribution of X given $T(x)$ is not a function of θ." In other words, since θ is not considered a random variable, it cannot have a possibly unobserved random variable as a parent, as in the Bayesian sufficency model. Instead, if we require the fixed parameter θ to have no parents, we can draw a graphical model with equivalent conditional independence assumptions to figure (3), shown here in figure (4).

For the model in figure (4), consider θ fixed and condition $T(x)$ on θ, so the sufficient statistics conditioned on θ generate the data X.

Note that both of these chains are equivalent in terms of their conditional independencies to the undirected chain in figure (5).

Neyman Factorization Theorem: $T(x)$ is a sufficient statistic for X if and only if $p(x|\theta) = g(T(x), \theta)h(x, T(x))$.

In other words, we know that $T(x)$ is a sufficient statistic for X when the distribution describing X can be factorized into two parts, one of which is a function of only the sufficient statistic and θ, and the other of which is a function of only x and the sufficient statistic (and vice versa). Note that letting $T(x)$ be the sufficient statistics in the generic exponential family density function, $p(x|\theta)$ factorizes into the two functions g, h trivially. This proves that $T(x)$ is the sufficient statistic for the exponential family distribution.