STAT 538 Lecture 3

Exponential Family Models

©Marina Meilă
mmp@stat.washington.edu

Reading: Gill (Chapters 2–4)

1 Build your own exponential family in four easy steps

Sample space

Let Ω be a *sample space*, e.g $\{0,1\}, \mathbb{N}, \mathbb{R}$ with a dominating measure. Denote by x and element of Ω. The set Ω will be the support for all densities in the exponential family $\mathcal{P}$ that we are about to define.

Sufficient statistics

Define the function(s) $t(x)$, $t : \Omega \longrightarrow \mathbb{R}^k$, with components $t_{1:k}$. We call k the **order** of the exponential family $\mathcal{P}$. If $t_{1:k}$ are *affinely independent* on Ω

$$\text{i.e. } c^T t(x) = c_0 \text{ for } x \in \Omega \quad \Rightarrow \quad c_0, c = 0 \tag{1}$$

then the family $\mathcal{P}$ (or its parametrization) is called **minimal**.

For example, the functions

$$t(x) = (x\ x^2) \quad \text{on} \quad \mathbb{R} \quad \text{are affinely independent} \tag{2}$$
$$t(x) = (x\ x^2) \quad \text{on} \quad \{0,1\} \quad \text{are NOT affinely independent} \tag{3}$$
$$t(x) = (x\ 1-x) \quad \text{on} \quad \{0,1\} \quad \text{are NOT affinely independent} \tag{4}$$

Natural parameter space Θ

Define

$$\Theta = \{\theta \,|\, Z(\theta) = \int_\Omega e^{\theta^T t(x)} dx < \infty\} \subseteq \mathbb{R}^k. \tag{5}$$

Note that even though Ω can be any sample space, the natural parameter space is a subset of $\mathbb{R}^k$, so that any $\theta \in \Theta$ is a k-dimensional *natural parameters vector*. The set Θ is convex [*Exercise: verify this.*] If Θ is also open then $\mathcal{P}$ is **linear**.

Exponential family model in the natural parametrization.

Finally we are ready to define the elements of our model class $\mathcal{P}$.

$$p_\theta(x) \;=\; \frac{1}{Z(\theta)} e^{\theta^T t(x)} \tag{6}$$

The sufficient statistics are also called **natural coordinates** of the exponential family, and Z is the **normalization constant**.

It is easy to see that $Z(\theta)$ is convex (in θ) as sum (or integral) of the convex functions $e^{x^T\theta}$.

We will find it useful to work with $\ln Z(\theta)$ which is the **partition function** or the **cumulant function**.

$$\psi(\theta) \;=\; \ln Z(\theta) \;=\; \ln \sum_x e^{\theta^T t(x)} \tag{7}$$

This function is always convex in θ as the composition of the convex increasing function $\log \sum_i e^{y_i}$ with the linear functions $y_i = x_i^T\theta$.

Using (7) we can express the distribution $p(x)$ as

$$p_\theta(x) \;=\; e^{\theta^T t(x) - \psi(\theta)} \tag{8}$$

Remarks 1. The general form of an exponential family model is $\log p(x) = \frac{\theta^T t(x) - \log Z_(\theta)}{a(\gamma)} + \log c(x)$, with γ another parameter called *nuisance parameter* and $a > 0$ a scaling function, and $c(x)$ a given probability measure. Here we will ignore a and c, as they do not have any influence on the estimation of the parameter θ, nor on any of the properties of the exponential family that we study here. Often times, $t(x) = x$ and we will make this substitution automatically when there is no ambiguity.

2. Exponential families are defined for $x \in \Omega \subseteq \mathbb{R}^n$, which can be a discrete or a continuous sample space. Therefore we will alternate between $\sum_x$ and $\int dx$, where the sum and the integral are over Ω.

Exponential sub-families

Define a function $\theta(\eta)$, $\eta \in Q\mathbb{R}^{k'} \rightarrow \theta(\eta) \in \Theta$, with $k' \leq k$. This mapping defines a subfamily of $\mathcal{P}$ by $\mathcal{P}_Q = \{p_\eta = p_\theta(\eta) \in \mathcal{P}, \eta \in Q\}$. If the range of $\theta(\eta)$ is a vector subspace of dimension $k' < k$ of Θ, then we call $\mathcal{P}_Q$ a **linear subfamily**. Otherwise we call it a **curved exponential (sub)-family**. The original $\mathcal{P}$ is called the **full** exponential family w.r.t $\mathcal{P}_Q$.

Exponential family models comprise (multivariate) normal distributions, Markov random fields (with positive distributions), binomial and multinomial models, etc. They have many convenient properties, some of which are evident from the definition above. For example, exponential family models are essentially the only parametric models that have fixed dimensional sufficient statistics[1]; they have **conjugate priors**; from the differential geometry p.o.v, exponential families repreent **flat manifolds**, i.e affine function spaces spanned by the vectors θ_i. We will show some of these properties in section **??**.

2 Examples

Normal univariate distribution with unit variance

$$p_\mu(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}} = exp\{\underbrace{-\frac{x^2}{2}}_{c(x)} + \mu x - \underbrace{\mu^2/2 + \frac{1}{2}\ln(2\pi)}_{\psi(\mu)}\} \tag{9}$$

The natural parameter is $\mu \in \mathbb{R}$, the vector of sufficient statistics is one dimensional, equal to x, and there is a non-trivial $c(x)$ component of the model, that influences what $Z(\theta)$ is, but not the natural parameter μ.

Normal univariate distribution

$$p_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{10}$$

$$= e^{\frac{-1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{1}{2}[\frac{\mu^2}{\sigma^2} + \ln(2\pi\sigma^2)]} \tag{11}$$

This univariate distribution has 2 natural parameters $\theta_1 = \frac{-1}{2\sigma^2}$, $\theta_2 = \frac{\mu}{\sigma^2}$ and a vector of sufficient statistics $t(x) = [x^2 \; x] \in \mathbb{R}^2$. Note that in this case

[1]Distributions that are piecewise uniform may also have finite sufficient statistics. In their case, the sufficient statistics are intervals in which the data lie.

the natural coordinates/sufficient statistics have a different dimension than the original variable x. The log-partition function ψ expressed in natural parameters is

$$\psi(\theta) \;=\; \frac{-\theta_2^2}{\theta_1} + \ln(-\pi/\theta_1) \tag{12}$$

which is strictly convex (verify by taking the Hessian) when $\theta_1 < 0$. The domain Θ of the natural parameters is $(-\infty, 0) \times \mathbb{R}$.

Note that the first example presents a linear subfamily of the second. To obtain a curved one-parameter subfamily of $\mathcal{P} = \{p_{\mu,\sigma^2}\}$, take the family of all distribution where $\mu = \sigma > 0$.

Bernoulli distribution Here $\Omega = \{0, 1\}$ and let $p = Pr[X = 1]$ the probability of success in the Bernoulli trial.

$$P_p(x) \;=\; p^x(1-p)^{1-x} \;=\; e^{(\ln p)x + \ln(1-p)(1-x)} \tag{13}$$

and the natural parameters are $\theta = [\ln p \;\; \ln(1-p)]$ with $\psi(\theta) = 0$, $Z(\theta) = 1$. This is not a minimal model. We will return to this model in the next section.

3 Expectations, moments and covexity

1. $\boxed{E_\theta[X] \;\equiv\; \mu(\theta) \;=\; \nabla\psi(\theta)}$
Proof

$$\begin{aligned}
\nabla\psi(\theta) \;&=\; \frac{\nabla_\theta \left(\sum_x e^{\theta^T x}\right)}{Z(\theta)} && (14)\\
&=\; \frac{\sum_x x e^{\theta^T x}}{Z(\theta)} && (15)\\
&=\; \sum_x x \frac{e^{\theta^T x}}{Z(\theta)} && (16)\\
&=\; \sum_x x p(x) \;=\; E_\theta[X] && (17)
\end{aligned}$$

2. $\boxed{Var_\theta[X] \;=\; \nabla^2\psi(\theta)}$
Proof

$$\begin{aligned}
\nabla^2\psi(\theta) &= \nabla_\theta^T\left[\frac{\sum_x xe^{\theta^T x}}{Z(\theta)}\right] && (18)\\
&= \sum_x\left\{xx^T\frac{e^{\theta^T x}}{Z(\theta)} + xe^{\theta^T x}\left[-\frac{\nabla^T Z(\theta)}{Z^2(\theta)}\right]\right\} && (19)\\
&= \left\{\sum_x xx^T p(x) - xe^{\theta^T x}\left[-\frac{\sum_{x'} x'e^{\theta^T x'}}{Z^2(\theta)}\right]^T\right\} && (20)\\
&= E_\theta[xx^T] - E_\theta[x](E_\theta[x])^T \;=\; Var_\theta X && (21)
\end{aligned}$$

3. From Property 2, because the variance is always positive definite, we obtain an alternative proof that $\psi(\theta)$ is convex.

4. $\boxed{\ln p_\theta(x) = \theta^T x - \psi(\theta) \text{ is concave in } \theta}$ and linear in x. Hence p is **log-concave** in θ, and is a **log-linear model** in x.

5. From 4 we also expect that, (under mild regularity conditions) the Maximum Likelihood estimate (when it exists) to be unique, and computationally easy to find, as the unique local maximum of the log-likelihood.

   Remark: Let us assume that the sufficient statistics $x_1, \ldots x_n$ are *affinely independent* random variables. Then, $VarX$ is non-singular, and consequently, $\nabla^2 \ln p_\theta(x) = -\nabla^2\psi(\theta) \prec 0$, implying that the log-likelihood is strictly concave, and has at most one global maximum.

   Example: Univariate Normal By taking the gradient of $\psi(\theta)$, we obtain

$$\nabla\psi \;=\; \begin{bmatrix} \frac{-1}{2\theta_1} + \frac{\theta_2^2}{4\theta_1^2} \\ \frac{-\theta_2}{2\theta_1} \end{bmatrix} \;=\; \begin{bmatrix} \sigma^2 + \mu^2 \\ \mu \end{bmatrix} \;=\; \begin{bmatrix} E[X] \\ E[X^2] \end{bmatrix} \qquad (22)$$

   Furthermore (Exercise) the Hessian of ψ will give us (on its diagonal)the variance of X^2, respectively the variance of X, which is of course σ^2.

   Example: Bernoulli In the above, we obtained $\phi(\theta) \equiv 0$ for the Bernoulli. Hence, its gradient cannot give us the expectation of X. What is wrong? The problem is that the sufficient statistic $t(x) =$

$[x\ 1-x]$ is not a vector of affinely independent functions of x. (This happens generally for distributions over discrete sample spaces if we are not careful.). It is said that the model (13) is not in *standard form.*

We reparametrize P_p using a single sufficient statistic and a single parameter θ.

$$\begin{aligned} x\ln p + (1-x)\ln(1-p) &= x[\ln p - \ln(1-p)] + \ln(1-p) \quad (23)\\ \theta &= \ln\frac{p}{1-p} \quad (24)\\ \psi(\theta) &= \ln(1-p) = \ln\frac{1}{1+e^{\theta}} \quad (25)\end{aligned}$$

Now, $\psi'(\theta) = \frac{e^{\theta}}{1+e^{\theta}} = p = E[X]$ (by replacing θ with its value in (24)).

Let us examine ML estimation closer. Assume we have an i.i.d sample $x^1, x^2, \ldots x^n$. The likelihood of the sample is

$$\begin{aligned} p_\theta(x^{1:n}) &= \prod_{i=1}^{n} e^{\theta^T x^i - \psi(\theta)} \quad (26)\\ &= e^{\theta^T \sum_{i=1}^n x^i - n\psi(\theta)} \quad (27)\\ &= e^{n[\theta^T \bar{x} - \psi(\theta)]} \quad (28)\end{aligned}$$

and the ML estimation equation is

$$\max_\theta g(\theta, \bar{x}) = \bar{x}^T\theta - \psi(\theta) \quad (29)$$

Comparing the above equation with (**??**) we find that $\boxed{\theta^{ML} \text{ is Legendre conjugate with } \bar{x} = (\sum_{i=1}^n x^i)/n}$ and that the max log-likelihood (= log-likelihood at θ^{ML}) $\boxed{\phi(\bar{x}) \text{ is the Legendre conjugate function of } \psi(\theta)}$. Moreover, maximizing the likelihood is equivalent to solving the equations

$$\bar{x} = \nabla\psi(\theta); \quad (30)$$

but from Property 1 we know that $\nabla\psi(\theta) = E_\theta[X]$. Hence, the ML equations for an exponential family model amount to solving for θ in

$$E_\theta[X] = \frac{\sum_i x^i}{n} \quad (31)$$

In other words, θ^{ML} is the parameter value for which the model expectation equals the sample mean of the data (=the expectation under the empirical distribution). (Exercise: Normal distribution).

6. Returning to the general expression of the log-likelihood, for any θ, the Legendre conjugate parameter μ is given by (**??**) $\mu = \nabla_\theta \psi = E_\theta[X]$. In other words, the conjugate pairs θ, μ represent the (parameter, mean value) pairs. The dual parametrization of the model in terms of $\mu, \phi(\mu)$ is called the **Mean value parametrization**.

   The domain of $\phi(\mu)$, i.e the set $\{E_\theta[X]\}_\theta$ is called the **marginal polytope** of the exponential family. Exercise: is the normal distribution's (μ, σ^2) parametrization a mean value parametrization? For the Bernoulli, since $p = E[X]$ (Exercise: check this via $\nabla\psi(\theta)$), the usual parametrization is the mean value parametrization.

7. The gradient of the log-likelihood w.r.t the parameters has the simple formula

$$\nabla_\theta \frac{1}{N} \ln p_\theta(x^{1:N}) \;=\; \bar{x} - \nabla_\theta \psi(\theta) \;=\; \bar{x} - E_\theta[x] \tag{32}$$

   Thus, when we fit the models by e.g gradient ascent, the direction of ascent is the difference between the data expectations and the model expectations.

   Example: **Generalized Linear Models (GLM)**
   A GLM is a regression where the "noise" distribution is in the exponential family.

   - $y \in \mathbb{R},\ y \sim P_\theta$ with

$$P_\theta(y) \;=\; e^{\theta y - \ln \psi(\theta)} \tag{33}$$

   - the parameter θ is a linear function of $x \in \mathbb{R}^d$

$$\theta \;=\; \beta^T x \tag{34}$$

   - We denote $E_\theta[y] = \mu$. The function $g(\mu) = \theta$ that relates the mean parameter to the natural parameter is called the **link function**. The link function is given by $g(\mu) = (\nabla\psi)^{-1}(\mu)$.

The log-likelihood (w.r.t. β) is

$$l(\beta) \;=\; \ln P_\theta(y|x) \;=\; \theta y - \psi(\theta) \quad \text{where } \theta = \beta^T x \tag{35}$$

and the gradient w.r.t. β is therefore

$$\nabla_\beta l \;=\; \nabla_\theta l \nabla_\beta(\beta^T x) \;=\; (y-\mu)x \tag{36}$$

This simple expression for the gradient is the generalization of the gradient expression you obtained for the two layer neural network in STAT 535. [Exercise: This means that the sigmoid function is the *inverse link function* defined above. Find what is the link function that corresponds to the neural network.]

8. $\boxed{H(p_\theta) \equiv H(\theta) = \psi(\theta) - \theta^T E[X]}$

$$\begin{aligned} \textit{Proof} \; -H(\theta) &= \sum_x p_\theta(x) \ln p_\theta(x) && (37)\\ &= \sum_x p_\theta(x)[\theta^T x - \psi(\theta)] && (38)\\ &= \theta^T \sum_x p_\theta(x) x - \psi(\theta) && (39)\\ &= \theta^T \mu(\theta) - \psi(\theta) \;=\; \phi(\mu) && (40) \end{aligned}$$

It follows also that $\boxed{H(\theta) = -\psi^*(\mu) \equiv -\phi(\mu)}$. The conjugate of ψ is the negative entropy.

9. $\boxed{KL(\theta_1, \theta_2) = d_\psi(\theta_2, \theta_1) = d_\phi(\mu_1, \mu_2)}$

Proof We need to prove only one of the equalities, because the other follows from Property **??** of the Bregman divergence.

$$\begin{aligned} KL(\theta_1, \theta_2) &= \sum_x p_{\theta_1}(x) \ln \frac{p_{\theta_1}(x)}{p_{\theta_2}(x)} && (41)\\ &= \sum_x p_{\theta_1}(x)[\theta_1^T x - \psi(\theta_1) - \theta_2^T x + \psi(\theta_2)] && (42)\\ &= (\theta_1 - \theta_2)^T [\underbrace{\sum_x p_{\theta_1}(x) x}_{\mu(\theta_1) = \nabla\psi(\theta_1)}] - \psi(\theta_1) + \psi(\theta_2) && (43)\\ &= \psi(\theta_2) - \psi(\theta_1) + (\theta_1 - \theta_2)^T \nabla\psi(\theta_1) && (44)\\ &= d_\psi(\theta_2, \theta_1) && (45) \end{aligned}$$

10. Likelihood and KL divergence (see Lecture 8)