Statistics 210A Lecture 3 Notes

Daniel Raban

September 2, 2021

1 Exponential Families and Differential Identities

1.1 Examples of exponential families

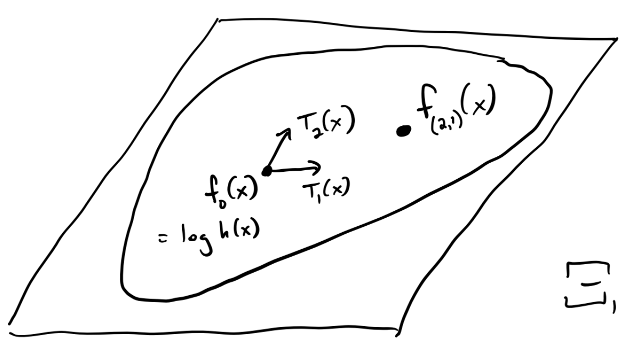
Recall from last time that an **s-parameter exponential family** is a family $\mathcal{P} = \{P_\eta : \eta \in \Xi\}$ with densities

$$p_\eta(x) = e^{\eta^\top T(x) - A(\eta)} h(x)$$

with respect to a base measure μ on $\mathcal{X}$. Here,

- $T : \mathcal{X} \to \mathbb{R}^s$ is called the **sufficient statistic**,
- $h : \mathcal{X} \to [0, \infty)$ is called the **carrier/base density**,
- $\eta \in \Xi \subseteq \mathbb{R}^s$ is called the **natural parameter**,
- $A : \mathbb{R}^s \to \mathbb{R}$ is called the **cumulant generating function** (or the **normalizing constant**).

Last time, we mentioned that we can think of an s-parameter exponential family as an s dimensional hyperplane in the space of log densities.



An important thing to note about this picture is that it shows us that the h and T are not unique. Only the span really matters.

Sometimes it is more convenient to use a different parameterization than the natural parameter:

$$p_\theta(x) = e^{\eta(\theta)^\top T(x) - B(\theta)} h(x), \qquad B(\theta) = A(\eta(\theta)).$$

Example 1.1. Consider the family of Gaussian distributions, $X \sim N(\mu, \sigma^2)$ with $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. Here, $\theta = (\mu, \sigma^2)$. To describe this as an exponential family, we have

$$\begin{aligned} p_\theta(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)} \\ &= \exp\left(\frac{\mu}{\sigma^2} x - \frac{1}{2\sigma^2} x^2 - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) \right). \end{aligned}$$

So we have

$$\eta(\theta) = \begin{bmatrix} \mu/\sigma^2 \\ -1/(2\sigma^2) \end{bmatrix}, \qquad T(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}, \qquad h(x) = 1, \qquad B(\theta) = \frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2).$$

In terms of η, we can say

$$p_\eta(x) = \exp\left(\eta^\top \begin{bmatrix} x \\ x^2 \end{bmatrix} - A(\eta) \right), \qquad A(\eta) = -\frac{\eta_1^2}{4\eta_2} + \frac{1}{2}\log(-\pi/\eta_2).$$

Example 1.2. Now suppose $X_1, \dots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$. Then

$$\begin{aligned} p_\theta(x) &= \prod_{i=1}^n p_\theta^{(i)}(x_i) \\ &= \exp\left(\sum_{i=1}^n \left[\frac{\mu}{\sigma^2} x_i - \frac{1}{2\sigma^2} x_i^2 - \left(\frac{\mu}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2) \right) \right] \right) \\ &= \exp\left(\frac{\mu}{\sigma^2} \sum_{i=1}^n x_i - \frac{1}{2\sigma^2} \sum_{i=1}^n x_i^2 - n\left(\frac{\mu}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2) \right) \right). \end{aligned}$$

So we have

$$\eta(\theta) = \begin{bmatrix} \mu/\sigma^2 \\ -1/(2\sigma^2) \end{bmatrix}, \qquad T(x) = \begin{bmatrix} \sum_i x_i \\ \sum_i x_i^2 \end{bmatrix}, \qquad h(x) = 1, \qquad B(\theta) = nB^{(1)}(\theta).$$

Proposition 1.1. *Suppose $X_1, \dots, X_n \overset{\text{iid}}{\sim} p_\eta^{(i)}(x) = e^{\eta^\top T(x) - A(\eta)} h(x)$. Then the distribution of $X = (X_1, \dots, X_n)$ follows an exponential family with sufficient statistic $\sum_{i=1}^n T(x_i)$ and cumulant generating function $nA(\eta)$.*

Proof.

$$\begin{aligned}
X &\sim \prod_{i=1}^n p_\eta^{(i)}(x_i) \\
&= \prod_{i=1}^n e^{\eta^\top T(x_i) - A(\eta)} h(x_i) \\
&= \exp\left(\eta^\top \sum_i T(x_i) - nA(\eta)\right) \prod_{i=1}^n h(x_i).
\end{aligned}$$

□

$T(X)$ also follows a closely related exponential family.

Proposition 1.2. *Suppose $X \in \mathcal{X}$ and $T(X) \in \mathcal{T} \subseteq \mathbb{R}^s$ with $h(x) = 1$ and $X \sim p_\eta(x) = e^{\eta^\top T(x) - A(\eta)}$ with respect to μ. For a set $B \subseteq \mathcal{T}$, define $\nu(B) = \mu(T^{-1}(B))$. Then*

$$T(X) \sim q_\eta(t) = e^{\eta^\top t - A(\eta)}$$

with respect to ν.

Example 1.3. In the discrete case, this is

$$\begin{aligned}
\mathbb{P}_\eta(T(X) \in B) &= \sum_{x : T(x) \in B} e^{\eta^\top T(x) - A(\eta)} \mu(\{x\}) \\
&= \sum_{t \in B} \sum_{x : T(x) = t} e^{\eta^\top t - A(\eta)} \mu(\{x\}) \\
&= \sum_{t \in B} e^{\eta^\top t - A(\eta)} \underbrace{\mu(T^{-1}(\{x\}))}_{\nu(\{x\})}.
\end{aligned}$$

So $T \sim e^{\eta^\top t - A(\eta)}$ with respect to ν.

Example 1.4. Let $X \sim \text{Binomial}(n, \theta)$. We can turn this into an exponential family as follows: For $\theta \in (0, 1)$,

$$\begin{aligned}
p_\theta(x) &= \theta^x (1-\theta)^{n-x} \binom{n}{x} \\
&= \left(\frac{\theta}{1-\theta}\right)^x (1-\theta)^n \binom{n}{x} \\
&= \exp\left(x \log \frac{\theta}{1-\theta} + n \log(1-\theta)\right) \binom{n}{x}
\end{aligned}$$

The natural parameter is $\eta(\theta) = \log \frac{\theta}{1-\theta}$.

Example 1.5. Let $X \sim \text{Pois}(\theta)$ with density $p_\lambda(x) = \frac{\lambda^x e^{-\lambda}}{x!}$ with respect to counting measure on $\mathbb{N}$. This is an exponential family

$$p_\lambda(x) = \exp\left((\log \lambda)x - \lambda\right) \frac{1}{x!}$$

with natural parameter $\eta(\lambda) = \log \lambda$.

Most of the families of distributions you can find on, say, Wikipedia, will be exponential families.

1.2 Differential identities for the cumulant generating function

Begin with the equation

$$e^{A(\eta)} = \int e^{\eta^\top T(x)} h(x)\, d\mu(x)$$

and then differentiate. Here is a criterion which lets us differentiate under the integral:

Theorem 1.1 (Theorem 2.4 in Keener)**.** *For $f : \mathcal{X} \to \mathbb{R}$, let $\Xi_f = \{\eta \in \mathbb{R}^s : \int |f| e^{\eta^\top T} h\, d\mu < \infty\}$. Then $g(\eta) = \int f e^{\eta^\top T} h\, d\mu$ has continuous partial derivatives of all orders for interior points $\eta \in \Xi_f^0$, and we can find them by differentiating under the integral.*

In particular, letting $f = 1$, we get that $A(\eta)$ has infinitely many partial derivatives in Ξ_1^0. So we can calculate

$$\frac{\partial}{\partial \eta_j} e^{A(\eta)} = \int \frac{\partial}{\partial \eta_j} e^{\eta^\top T(x)} h(x)\, d\mu(x),$$

which gives

$$\begin{aligned} \frac{\partial A}{\partial \eta_j}(\eta) &= \int T_j(x) e^{\eta^\top T(x) - A(\eta)} h(x)\, d\mu(x) \\ &= \mathbb{E}_\eta[T_j(X)]. \end{aligned}$$

This shows that

Proposition 1.3.

$$\nabla A(\eta) = \mathbb{E}_\eta[T(X)].$$

Taking second derivatives, we have

$$\frac{\partial^2 A}{\partial \eta_j \partial \eta_k} e^{A(\eta)} = \int \frac{\partial^2}{\partial \eta_j \partial \eta_k} e^{\eta^\top T(x)} h(x)\, d\mu(x),$$

which gives us

$$\left(\frac{\partial^2 A}{\partial \eta_j} - \frac{\partial A}{\partial \eta_j}\frac{\partial A}{\partial \eta_k}\right) = \int T_j T_k e^{\eta^\top T - A(\eta)} h \, d\mu.$$

So we get

$$\frac{\partial^2 A}{\partial \eta_j \partial \eta_k}(\eta) = \mathbb{E}_\eta[T_j T_k] - \mathbb{E}_\eta[T_j]\,\mathbb{E}_\eta[T_k] = \mathrm{Cov}(T_j, T_k).$$

In total, we get

Proposition 1.4.

$$\nabla^2 A(\eta) = \mathit{Var}_\eta(T(X)).$$

Differentiating repeatedly, we get

$$e^{-A(\eta)} \frac{\partial^{k_1 + \cdots + k_s}}{\partial_{\eta_1}^{k_1} \cdots \partial_{\eta_s}^{k_s}} (e^{A(\eta)}) = \mathbb{E}_\eta[T_1^{k_1} \cdots T_s^{k_s}].$$

This is because $M_\eta^T(u) = e^{A(\eta+u) - A(\eta)}$ is the **moment generating function (MGF)** of $T(X)$ when $X \sim p_\eta$:

$$\begin{aligned} M_\eta^{T(X)}(u) &= \mathbb{E}_\eta[e^{u^\top T(X)}] \\ &= \int e^{u^\top T} e^{\eta^\top T - A(\eta)} h \, d\mu \\ &= e^{A(\eta+u) - A(\eta)} \underbrace{\int e^{(\eta+u)^\top T - A(\eta+u)} h \, d\mu}_{=1}. \end{aligned}$$