

## Lecture 6 — 09/13

Lecturer: Martin Wainwright

Scribe: Caroline Uhler



This is the danger environment.

**Warning:** These scribe notes have only been mildly proofread.

## Exponential families

**Definition** Given a set of functions  $T_1, \dots, T_d$  ('sufficient statistics'), exponential families are families of density functions [mass functions] with common support  $\mathcal{X}$  of the form

$$p(x; \theta) = h(x) \exp \left\{ \sum_{i=1}^d \eta_i(\theta) T_i(x) - A(\theta) \right\},$$

where  $h(x) \geq 0$  is some fixed function,  $\eta_i : \Omega \rightarrow \mathbb{R}$ ,  $\theta \mapsto \eta_i(\theta)$ ,  $i \in \{1, \dots, d\}$ , specify the parameterization and  $A(\theta)$  is the log normalization constant

$$A(\theta) = \begin{cases} \log \left( \int h(x) \exp \{ \langle \eta(\theta), T(x) \rangle \} \right) \\ \log \left( \sum_{x \in \mathcal{X}} h(x) \exp \{ \langle \eta(\theta), T(x) \rangle \} \right) \end{cases} .$$

**Definition** An exponential family is normal or of full rank, if  $\eta(\Omega)$  has a non-empty interior, and  $\nexists c \in \mathbb{R}^d$  such that  $\sum_{i=1}^d c_i T_i(x)$  is constant for almost all  $x \in \mathcal{X}$ .

### Examples

a)  $T_1(x) = x$ ,  $T_2(x) = 1 - x$ ,  $x \in \{0, 1\}$ .

As  $T_1(x) + T_2(x) = 1 \quad \forall x \in \{0, 1\}$ , this exponential family is not of full rank.

b)  $\eta_1(\theta) = \theta$ ,  $\eta_2(\theta) = \theta^2$ ,  $\theta \in \Omega \equiv \mathbb{R}$

As the interior of  $\eta(\Omega)$  is empty, this exponential family is not of full rank.

**Proposition 6.1.** In a full rank exponential family  $(T_1, \dots, T_d)$  is minimal sufficient.

**Proof:** By the Neyman factorization criterion  $(T_1, \dots, T_d)$  is sufficient for the family of distributions  $\mathcal{P} = \{p(x; \theta) \mid \theta \in \Omega\}$ . Choose a set of  $(d+1)$  vectors  $\theta^0, \dots, \theta^d \in \Omega \subset \mathbb{R}^d$  such that the vectors  $(\eta(\theta^1) - \eta(\theta^0), \eta(\theta^2) - \eta(\theta^0), \dots, \eta(\theta^d) - \eta(\theta^0))$  are linearly independent. Look at the subfamily  $\mathcal{P}_0 = \{p(x, \theta^i) \mid i = 0, \dots, d\}$ . From problem set 2 we know that

$$\left\{ \log \left( \frac{p(x; \theta^1)}{p(x; \theta^0)} \right), \dots, \log \left( \frac{p(x; \theta^d)}{p(x; \theta^0)} \right) \right\} = \{ (\eta(\theta^1) - \eta(\theta^0)) T(x), \dots, (\eta(\theta^d) - \eta(\theta^0)) T(x) \}$$

is minimal sufficient for  $\mathcal{P}_0$ .

By linear independence  $T(x)$  is minimal sufficient for  $\mathcal{P}_0$  and so we get from problem set 2 that  $T(x)$  is also minimal sufficient for  $\mathcal{P}$ . □

**Remark** In a full rank exponential family  $(T_1, \dots, T_d)$  is not only minimal sufficient but also complete. However, this proof is more technical than the previous proof and we will therefore omit it.

**Definition** An exponential family is in canonical form if  $\eta_i(\theta) = \theta_i$ ,  $i = 1, \dots, d$ . Then  $p(x; \theta) = h(x) \exp \left\{ \sum_{i=1}^d \theta_i T_i(x) - A(\theta) \right\}$ .

**Remark** For an exponential family in canonical form,  $A(\theta) = \log \left( \int \exp \left\{ \sum_{i=1}^d \theta_i T_i(x) \right\} h(x) dx \right)$  and  $\Omega = \{ \theta \in \mathbb{R}^d \mid A(\theta) < \infty \}$ .

**Example** (Gaussian  $\mathcal{N}(\mu, \sigma^2)$  in canonical form)

$$p(x; \theta) \propto \exp \{ \theta_1 x + \theta_2 x^2 \}$$

and

$$A(\theta) = \log \left( \int_{\mathbb{R}} \exp \{ \theta_1 x + \theta_2 x^2 \} dx \right) = \begin{cases} < \infty & \theta_2 < 0 \\ \infty & \theta_2 \geq 0 \end{cases} .$$

So  $\Omega = \{ (\theta_1, \theta_2) \mid \theta_2 < 0 \}$ .

**Example** (Multinomial distribution)

For  $x \in \{0, 1, \dots, d-1\}$

$$p(x; \theta) \propto \exp \{ \theta_1 x + \theta_2 x^2 + \dots + \theta_{d-1} x^{d-1} \}$$

and

$$A(\theta) = \log \left( \sum_{x \in \{0, 1, \dots, d-1\}} \exp \left\{ \sum_{i=1}^{d-1} \theta_i x^i \right\} \right) < \infty \quad \forall \theta \in \mathbb{R}^d$$

So  $\Omega = \mathbb{R}^d$ .

Now let's look at  $A(\theta)$  more generally:

**Proposition 6.2.** *In an exponential family,  $A$  is convex on its domain  $\Omega$  (which is a convex set).*

**Proof:** Take  $\theta^1, \theta^2 \in \Omega$  and  $\alpha \in [0, 1]$ . We want to show that  $A(\alpha\theta^1 + (1 - \alpha)\theta^2) \leq \alpha A(\theta^1) + (1 - \alpha)A(\theta^2)$ .

By Hölder's inequality:

$$\begin{aligned} A(\alpha\theta^1 + (1 - \alpha)\theta^2) &= \log \left( \int \exp \{ (\alpha\theta^1 + (1 - \alpha)\theta^2) T(x) \} h(x) dx \right) \\ &= \log \left( \int (\exp \{ \theta^1 T(x) \})^\alpha (\exp \{ \theta^2 T(x) \})^{1-\alpha} h(x) dx \right) \\ &\leq \log \left( \left( \int \exp \{ \theta^1 T(x) \} h(x) dx \right)^\alpha \left( \int \exp \{ \theta^2 T(x) \} h(x) dx \right)^{1-\alpha} \right) \\ &= \alpha A(\theta^1) + (1 - \alpha)A(\theta^2) \end{aligned}$$

□

**Proposition 6.3.** For an exponential family in canonical form,  $A$  is  $C^\infty$  on its domain  $\Omega$  and moreover

a)

$$\frac{\partial A}{\partial \theta_i} = \mathbb{E}_\theta[T_i(X)], \quad \nabla A(\theta) = \mathbb{E}_\theta[T(X)],$$

b)

$$\frac{\partial^2 A}{\partial \theta_i \partial \theta_j} = \text{cov}_\theta\{T_i(X), T_j(X)\}.$$

**Sketch of the proof**  $A(\theta) = \log \left( \int \exp \left\{ \sum_{i=1}^d \theta_i T_i(x) \right\} h(x) dx \right)$

a) We assume that differentiation and integration can be exchanged. This follows from dominated convergence (see Keener for details). Then

$$\begin{aligned} \frac{\partial A}{\partial \theta_k} &= \frac{1}{e^{A\theta}} \int \frac{\partial}{\partial \theta_k} \exp \left\{ \sum_{i=1}^d \theta_i T_i(x) \right\} h(x) dx \\ &= \frac{1}{e^{A\theta}} \int T_k(x) \exp \left\{ \sum_{i=1}^d \theta_i T_i(x) \right\} h(x) dx \\ &= \mathbb{E}_\theta[T_k(X)]. \end{aligned}$$

b) The proof follows from a similar argument.

**Note:** Hence  $\nabla^2 A(\theta) = \text{cov}\{T_1(X), \dots, T_d(X)\}$ , which is positive semidefinite. This provides an alternative proof of the convexity of  $A$ .

**Examples**

a)  $\mathcal{N}(\theta, 1)$  distribution:

$$\begin{aligned} p(x; \theta) &= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x - \theta)^2\right\} \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \exp\left\{\theta x - \frac{1}{2}\theta^2\right\} \\ \Rightarrow &\begin{cases} A(\theta) = \frac{1}{2}\theta^2 \\ A'(\theta) = e^\theta = \mathbb{E}_\theta[X] \\ A''(\theta) = 1 = \text{var}_\theta(X) \end{cases} . \end{aligned}$$

b) Poisson distribution:

$$\begin{aligned} p(x; \theta) &= \frac{1}{x!} \exp\{\theta x - e^\theta\} \\ \Rightarrow &\begin{cases} A(\theta) = e^\theta \\ A'(\theta) = e^\theta = \mathbb{E}_\theta[X] \\ A''(\theta) = e^\theta = \text{var}_\theta(X) \end{cases} . \end{aligned}$$

**Exercise**  $X \sim \mathcal{N}(\mu, \Omega)$ ,  $\mu \in \mathbb{R}^d$ ,  $\Omega \in \mathbb{R}^{d \times d}$  positive semidefinite and symmetric. Write this family of distributions as exponential family and compute  $A$ , derivatives, etc.

**Remark** (Mean parameters)

We have seen that quantities like

$$\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_d \end{pmatrix} = \begin{pmatrix} \mathbb{E}[T_1(X)] \\ \vdots \\ \mathbb{E}[T_d(X)] \end{pmatrix} = \nabla A(\theta) \in \mathbb{R}^d$$

often arise. We will see later that there is a one-to-one transformation from  $\theta$  to  $\mu$ .