Notes on Exponential Families

Jeffrey W. Miller

June 10, 2016

These notes cover some of the basic theoretical properties of exponential families, such as reduction to natural form, smoothness and convexity of the log-partition function, justification of differentiating under the integral, and convexity of the natural parameter space.

1 General form

Let $(\mathcal{Y}, \mathcal{A}, \nu)$ be a measure space, and let $s : \mathcal{Y} \to \mathbb{R}^D$, $h : \mathcal{Y} \to [0, \infty)$, and $\varphi : \mathbb{R}^k \to \mathbb{R}^D$ be measurable functions. Euclidean spaces, such as $\mathbb{R}^D$, are given the Borel sigma-algebra, unless otherwise specified. Define $L : \mathbb{R}^k \to [-\infty, \infty]$ by

$$L(\beta) = \log \int_{\mathcal{Y}} \exp\big(\varphi(\beta)^{\mathtt{T}} s(y)\big) h(y) d\nu(y).$$

By convention, $\log(\infty) = \infty$, $\log(0) = -\infty$, $\exp(\infty) = \infty$, and $\exp(-\infty) = 0$. Define

$$q_\beta(y) = \exp\big(\varphi(\beta)^{\mathtt{T}} s(y) - L(\beta)\big) h(y)$$

for $y \in \mathcal{Y}$, $\beta \in \mathbb{R}^k$, and define

$$Q_\beta(B) = \int_B q_\beta(y) d\nu(y)$$

for $B \in \mathcal{A}$, $\beta \in \mathbb{R}^k$. For any $\beta \in \mathbb{R}^k$ such that $L(\beta) \in (-\infty, \infty)$, it follows that Q_β is a probability measure on $(\mathcal{Y}, \mathcal{A})$, and q_β is the probability density of Q_β with respect to ν.

This is the general form of an exponential family. In order to derive many of the properties of Q_β and $L(\beta)$, it is convenient to first rewrite the distribution in the corresponding "natural form" with density $p_\theta(x) = \exp(\theta^{\mathtt{T}} x - K(\theta))$ with respect to a measure μ on $\mathbb{R}^D$.

2 Reduction to natural form

The basic idea is to absorb $h(y)$ into the measure, make a change of variables to $x = s(y)$, and re-parameterize in terms of $\theta = \varphi(\beta)$. The details are as follows. Let $\mathcal{B}_{\mathbb{R}^D}$ denote the Borel sigma-algebra on $\mathbb{R}^D$, and define $\mu(A) = \int_{\mathcal{Y}} \mathbb{1}_A(s(y)) h(y) d\nu(y)$ for $A \in \mathcal{B}_{\mathbb{R}^D}$. Here, $\mathbb{1}_A$ denotes the indicator function of the set A (that is, $\mathbb{1}_A(x) = 1$ if $x \in A$, and $\mathbb{1}_A(x) = 0$ otherwise). Note that this makes μ a measure on $(\mathbb{R}^D, \mathcal{B}_{\mathbb{R}^D})$. Define $K : \mathbb{R}^D \to [-\infty, \infty]$ by

$$K(\theta) = \log \int_{\mathbb{R}^D} e^{\theta^{\mathtt{T}} x} d\mu(x)$$

and let $\Theta = \{\theta \in \mathbb{R}^D : K(\theta) \in (-\infty, \infty)\}$. Define

$$p_\theta(x) = \exp(\theta^\mathsf{T} x - K(\theta))$$

for $x \in \mathbb{R}^D$, $\theta \in \Theta$, and define

$$P_\theta(A) = \int_A p_\theta(x) d\mu(x)$$

for $A \in \mathcal{B}_{\mathbb{R}^D}$, $\theta \in \Theta$. Then for any $\theta \in \Theta$, P_θ is a probability measure on $(\mathbb{R}^D, \mathcal{B}_{\mathbb{R}^D})$ with density p_θ with respect to μ. The following result shows that μ simultaneously absorbs h into the measure and makes a change of variables to $x = s(y)$.

Theorem 2.1. *For any measurable* $f : \mathbb{R}^D \to [0, \infty]$, *we have*

$$\int_{\mathbb{R}^D} f(x) d\mu(x) = \int_{\mathcal{Y}} f(s(y)) h(y) d\nu(y).$$

Proof. The proof follows a standard argument for establishing the equality of two classes of integrals. If $f = \mathbb{1}_A$ for some $A \in \mathcal{B}_{\mathbb{R}^D}$, then the result holds by the definition of μ. If f is a nonnegative simple function, then it holds by linearity of the integral (Folland, 2013, 2.15). If $f : \mathbb{R}^D \to [0, \infty]$ is measurable, then there exists a sequence of simple functions f_n such that $0 \leq f_1 \leq f_2 \leq \cdots \leq f$ and $f_n \to f$ pointwise (Folland, 2013, 2.10). Thus,

$$\int f\, d\mu = \lim_{n\to\infty} \int f_n\, d\mu = \lim_{n\to\infty} \int f_n(s(y)) h(y) d\nu(y) = \int f(s(y)) h(y) d\nu(y)$$

by two applications of the monotone convergence theorem (Folland, 2013, 2.14), first to $f_n(x)$ and then to $f_n(s(y))h(y)$. $\square$

We use Theorem 2.1 to prove the following theorem, which allows one to obtain results about $L(\beta)$ and Q_β based on results about $K(\theta)$ and P_θ, and vice versa.

Theorem 2.2. *For any* $\beta \in \mathbb{R}^k$, $L(\beta) = K(\varphi(\beta))$. *For any* $\beta \in \mathbb{R}^k$ *such that* $L(\beta) \in (-\infty, \infty)$, *if* $Y \sim Q_\beta$ *then* $s(Y) \sim P_\theta$ *where* $\theta = \varphi(\beta)$.

Proof. By Theorem 2.1 with $f(x) = \exp(\varphi(\beta)^\mathsf{T} x)$,

$$L(\beta) = \log \int_{\mathcal{Y}} \exp\big(\varphi(\beta)^\mathsf{T} s(y)\big) h(y) d\nu(y) = \log \int_{\mathbb{R}^D} \exp\big(\varphi(\beta)^\mathsf{T} x\big) d\mu(x) = K(\varphi(\beta)).$$

Suppose $L(\beta) \in (-\infty, \infty)$ and $\theta = \varphi(\beta)$. Then for any $A \in \mathcal{B}_{\mathbb{R}^D}$,

$$\begin{aligned} \mathbb{P}(s(Y) \in A) &= \int_{\mathcal{Y}} \mathbb{1}_A(s(y)) q_\beta(y) d\nu(y) \\ &= \int_{\mathcal{Y}} \mathbb{1}_A(s(y)) \exp\big(\varphi(\beta)^\mathsf{T} s(y) - L(\beta)\big) h(y) d\nu(y) \\ &= \int_{\mathbb{R}^D} \mathbb{1}_A(x) \exp\big(\varphi(\beta)^\mathsf{T} x - L(\beta)\big) d\mu(x) \\ &= \int_A \exp\big(\theta^\mathsf{T} x - K(\theta)\big) d\mu(x) = P_\theta(A) \end{aligned}$$

where the third step is by Theorem 2.1 with $f(x) = \mathbb{1}_A(x) \exp\big(\varphi(\beta)^\mathsf{T} x - L(\beta)\big)$, and the fourth step is by the first part of this theorem. $\square$

3 Convexity and differentiation properties

Theorem 3.1. *Let μ be a measure on $(\mathbb{R}^d, \mathcal{B}_{\mathbb{R}^D})$. If $K(\theta) = \log \int e^{\theta^\mathsf{T} x} d\mu(x)$ and $\Theta = \{\theta \in \mathbb{R}^D : K(\theta) \in (-\infty, \infty)\}$, then K is convex on $\mathbb{R}^D$, and Θ is a convex set.*

Proof. If $\mu(\mathbb{R}^D) = 0$, then $K(\theta) = -\infty$ for all $\theta \in \mathbb{R}^D$, and $\Theta = \varnothing$, so the result is trivial. Suppose $\mu(\mathbb{R}^D) \in (0, \infty]$. Then $\int e^{\theta^\mathsf{T} x} d\mu(x) \in (0, \infty]$ and $K(\theta) > -\infty$ for all $\theta \in \mathbb{R}^D$ (Folland, 2013, 2.23). Let $\theta, \eta \in \mathbb{R}^D$, and let $a, b > 0$ such that $a + b = 1$. Let $p = 1/a$ and $q = 1/b$. Then by Hölder's inequality (Folland, 2013, 6.2),

$$\begin{aligned}
\int e^{(a\theta+b\eta)^\mathsf{T} x} d\mu(x) &= \int (e^{a\theta^\mathsf{T} x})(e^{b\eta^\mathsf{T} x}) d\mu(x) \\
&\leq \Big(\int (e^{a\theta^\mathsf{T} x})^p d\mu(x) \Big)^{1/p} \Big(\int (e^{b\eta^\mathsf{T} x})^q d\mu(x) \Big)^{1/q} \\
&= \Big(\int e^{\theta^\mathsf{T} x} d\mu(x) \Big)^a \Big(\int e^{\eta^\mathsf{T} x} d\mu(x) \Big)^b.
\end{aligned}$$

Taking logs, we have $K(a\theta + b\eta) \leq aK(\theta) + bK(\eta)$. Therefore, K is convex. In particular, if $\theta, \eta \in \Theta$, then $-\infty < K(a\theta + b\eta) \leq aK(\theta) + bK(\eta) < \infty$, so $a\theta + b\eta \in \Theta$. Hence, Θ is convex. $\square$

We use S° to denote the interior of a set S.

Theorem 3.2. *Let μ be a measure on $(\mathbb{R}^d, \mathcal{B}_{\mathbb{R}^D})$. Define $G : \mathbb{R}^D \to [0, \infty]$ by $G(\theta) = \int e^{\theta^\mathsf{T} x} d\mu(x)$, and let $S = \{\theta \in \mathbb{R}^D : G(\theta) < \infty\}$. Then G is C^∞ on S°, and for any $k \in \{0, 1, 2, \ldots\}$, $i_1, \ldots, i_k \in \{1, \ldots, D\}$, $\theta \in S^\circ$, we have $\int |x_{i_1} \cdots x_{i_k}| e^{\theta^\mathsf{T} x} d\mu(x) < \infty$ and*

$$\frac{\partial}{\partial \theta_{i_1}} \cdots \frac{\partial}{\partial \theta_{i_k}} G(\theta) = \int_{\mathbb{R}^D} x_{i_1} \cdots x_{i_k} e^{\theta^\mathsf{T} x} d\mu(x). \tag{1}$$

Proof. We proceed by induction. By assumption, $G(\theta) = \int e^{\theta^\mathsf{T} x} d\mu(x) < \infty$ for all $\theta \in S^\circ$. Suppose that for some $k \in \{0, 1, 2, \ldots\}$ and some $i_1, \ldots, i_k \in \{1, \ldots, D\}$, we have that for all $\theta \in S^\circ$, Equation 1 holds and

$$\int |x_{i_1} \cdots x_{i_k}| e^{\theta^\mathsf{T} x} d\mu(x) < \infty. \tag{2}$$

Let $j \in \{1, \ldots, D\}$ and let $u = (0, \ldots, 1, \ldots, 0)$ denote the unit vector with 1 in the jth position. Let $\theta_0 \in S^\circ$. Since S° is open, there exists $\varepsilon > 0$ such that $\theta_0 + tu \in S^\circ$ for all $t \in [-2\varepsilon, 2\varepsilon]$. Define

$$f(x, t) = x_{i_1} \cdots x_{i_k} e^{(\theta_0 + tu)^\mathsf{T} x}$$

for $x \in \mathbb{R}^D$, $t \in [-2\varepsilon, 2\varepsilon]$, and note that $\int |f(x, t)| d\mu(x) < \infty$ for all $t \in [-2\varepsilon, 2\varepsilon]$, since Equation 2 holds for all $\theta \in S^\circ$ by assumption. Define

$$g(x) = \frac{1}{\varepsilon} |f(x, 2\varepsilon)| + \frac{1}{\varepsilon} |f(x, -2\varepsilon)|$$

for $x \in \mathbb{R}^D$, and note that $\int g(x)d\mu(x) < \infty$. It can be shown that $|\frac{\partial f}{\partial t}(x,t)| \leq g(x)$ for all $x \in \mathbb{R}^D$, $t \in [-\varepsilon, \varepsilon]$, but to avoid getting mired in details, let's assume this for now and return to it later.

This implies that $t \mapsto \int f(x,t)d\mu(x)$ is differentiable on $(-\varepsilon, \varepsilon)$ and $\frac{\partial}{\partial t}\int f(x,t)d\mu(x) = \int \frac{\partial f}{\partial t}(x,t)d\mu(x)$ for $t \in (-\varepsilon, \varepsilon)$ (Folland, 2013, 2.27b). Therefore,

$$\begin{aligned}
\frac{\partial}{\partial \theta_j}\Big|_{\theta=\theta_0} \frac{\partial}{\partial \theta_{i_1}} \cdots \frac{\partial}{\partial \theta_{i_k}} G(\theta) &= \frac{\partial}{\partial \theta_j}\Big|_{\theta=\theta_0} \int x_{i_1} \cdots x_{i_k} e^{\theta^\intercal x} d\mu(x) \\
&= \frac{\partial}{\partial t}\Big|_{t=0} \int x_{i_1} \cdots x_{i_k} e^{(\theta_0 + tu)^\intercal x} d\mu(x) \\
&= \frac{\partial}{\partial t}\Big|_{t=0} \int f(x,t) d\mu(x) \\
&= \int \frac{\partial f}{\partial t}(x,0) d\mu(x) \\
&= \int x_{i_1} \cdots x_{i_k} x_j e^{\theta_0^\intercal x} d\mu(x),
\end{aligned}$$

where the first equality is by the induction hypothesis. To complete the induction, note also that

$$\int |x_{i_1} \cdots x_{i_k} x_j| e^{\theta_0^\intercal x} d\mu(x) = \int \Big|\frac{\partial f}{\partial t}(x,0)\Big| d\mu(x) \leq \int g(x) d\mu(x) < \infty.$$

Now, to finish the proof, we have to justify the claim that $|\frac{\partial f}{\partial t}(x,t)| \leq g(x)$ for all $x \in \mathbb{R}^D$, $t \in [-\varepsilon, \varepsilon]$. First, suppose $x_j \leq 0$. Then $|x_j| = x_j \leq e^{\varepsilon x_j}/\varepsilon$ (since $a \leq e^a$ for any $a \in \mathbb{R}$). Also, $|f(x,t)| = |x_{i_1} \cdots x_{i_k}| e^{\theta_0^\intercal x} e^{t x_j} \leq |x_{i_1} \cdots x_{i_k}| e^{\theta_0^\intercal x} e^{\varepsilon x_j} = |f(x,\varepsilon)|$. Therefore,

$$\Big|\frac{\partial f}{\partial t}(x,t)\Big| = |x_j||f(x,t)| \leq \frac{1}{\varepsilon} e^{\varepsilon x_j} |f(x,\varepsilon)| = \frac{1}{\varepsilon}|f(x, 2\varepsilon)|.$$

Meanwhile, if $x_j \leq 0$, then by a completely symmetrical argument, $|\frac{\partial f}{\partial t}(x,t)| \leq \frac{1}{\varepsilon}|f(x,-2\varepsilon)|$. Therefore, in either case, $|\frac{\partial f}{\partial t}(x,t)| \leq \frac{1}{\varepsilon}|f(x,2\varepsilon)| + \frac{1}{\varepsilon}|f(x,-2\varepsilon)| = g(x)$. □

Theorem 3.3. *Let μ be a measure on $(\mathbb{R}^D, \mathcal{B}_{\mathbb{R}^D})$. Define $K(\theta) = \log \int e^{\theta^\intercal x} d\mu(x)$ for $\theta \in \mathbb{R}^D$, let $\Theta = \{\theta \in \mathbb{R}^D : K(\theta) \in (-\infty, \infty)\}$, and define $P_\theta(A) = \int_A \exp(\theta^\intercal x - K(\theta)) d\mu(x)$ for $A \in \mathcal{B}_{\mathbb{R}^D}$, $\theta \in \Theta$. For any $\theta \in \Theta^\circ$, if $X \sim P_\theta$ then*

1. $\dfrac{\partial}{\partial \theta_i} K(\theta) = \mathbb{E} X_i$,
2. $\dfrac{\partial^2}{\partial \theta_i \partial \theta_j} K(\theta) = \mathbb{E}\big((X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j)\big) = \mathrm{Cov}(X_i, X_j)$, *and*
3. $\dfrac{\partial^3}{\partial \theta_i \partial \theta_j \partial \theta_k} K(\theta) = \mathbb{E}\big((X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j)(X_k - \mathbb{E}X_k)\big)$

for any $i, j, k \in \{1, \ldots, D\}$.

Proof. Define $G(\theta) = \exp(K(\theta)) = \int e^{\theta^\intercal x} d\mu(x)$, and note that $G(\theta) \in (0, \infty)$ for any $\theta \in \Theta$. Thus, by Theorem 3.2, G is C^∞ on Θ° and the partial derivatives of G are as given in Equation 1. To de-clutter the notation, let us denote $G_i(\theta) = \frac{\partial}{\partial\theta_i} G(\theta)$, $G_{ij}(\theta) = \frac{\partial^2}{\partial\theta_i\partial\theta_j} G(\theta)$, and so on. From here on, we will suppose $\theta \in \Theta^\circ$. Then by Equation 1,

$$\frac{\partial}{\partial\theta_i} K(\theta) = \frac{\partial}{\partial\theta_i} \log G(\theta) = \frac{G_i(\theta)}{G(\theta)} = \frac{1}{G(\theta)} \int x_i e^{\theta^\intercal x} d\mu(x) = \int x_i e^{\theta^\intercal x - K(\theta)} d\mu(x) = \mathbb{E} X_i.$$

Using this, we have

$$\begin{aligned} \frac{\partial^2}{\partial\theta_i\partial\theta_j} K(\theta) &= \frac{\partial}{\partial\theta_j} \frac{G_i(\theta)}{G(\theta)} = \frac{G_{ij}(\theta)}{G(\theta)} - \frac{G_i(\theta)}{G(\theta)} \frac{G_j(\theta)}{G(\theta)} = \frac{1}{G(\theta)} \int x_i x_j e^{\theta^\intercal x} d\mu(x) - \mathbb{E} X_i \mathbb{E} X_j \\ &= \mathbb{E} X_i X_j - \mathbb{E} X_i \mathbb{E} X_j = \mathbb{E}\big((X_i - \mathbb{E} X_i)(X_j - \mathbb{E} X_j)\big) = \operatorname{Cov}(X_i, X_j). \end{aligned}$$

Finally,

$$\begin{aligned} \frac{\partial^3}{\partial\theta_i\partial\theta_j\partial\theta_k} K(\theta) &= \frac{\partial}{\partial\theta_k} \left(\frac{\partial^2}{\partial\theta_i\partial\theta_j} K(\theta) \right) = \frac{\partial}{\partial\theta_k} \left(\frac{G_{ij}(\theta)}{G(\theta)} - \frac{G_i(\theta)}{G(\theta)} \frac{G_j(\theta)}{G(\theta)} \right) \\ &= \left(\frac{G_{ijk}(\theta)}{G(\theta)} - \frac{G_{ij}(\theta)}{G(\theta)} \frac{G_k(\theta)}{G(\theta)} \right) - \Big(\mathbb{E} X_i \operatorname{Cov}(X_j, X_k) + \mathbb{E} X_j \operatorname{Cov}(X_i, X_k) \Big) \\ &= \big(\mathbb{E} X_i X_j X_k - \mathbb{E} X_i X_j \mathbb{E} X_k \big) - \big(\mathbb{E} X_i \mathbb{E} X_j X_k - \mathbb{E} X_i \mathbb{E} X_j \mathbb{E} X_k \big) \\ &\quad - \big(\mathbb{E} X_j \mathbb{E} X_i X_k - \mathbb{E} X_i \mathbb{E} X_j \mathbb{E} X_k \big) \\ &= \mathbb{E}\big((X_i - \mathbb{E} X_i)(X_j - \mathbb{E} X_j)(X_k - \mathbb{E} X_k)\big). \end{aligned}$$

$\square$

References

G. B. Folland. *Real Analysis: Modern Techniques and Their Applications.* John Wiley & Sons, 2013.