Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science
6.437 Inference and Information
Spring 2015

9 Exponential Families

Exponential families of probability distributions play a key role in statistical inference and its information theoretic foundations.[1] More generally, exponential families have special properties, are particularly amenable to analysis, and arise in a wealth of applications. Indeed, many well-known parameterized probability distributions take the form of particular exponential families.

The development in this section applies to families of distributions over discrete and continuous alphabets $\mathcal{Y}$. For the most part, we will emphasize the discrete case, to keep our development as simple as possible, but the basic concepts are the same for continuous alphabets. However, there are additional mathematical subtleties in the treatment of the continuous case that our beyond our scope, which will limit our discussion.

9.1 One-parameter Exponential Family

An exponential family is a parameterized family of distributions, where the (typically continuous-valued) parameters specify a particular distribution in the family. To develop our initial insights, we begin with the simplest case, in which there is a single scalar parameter.

Definition 1. *A parameterized family of distributions* $\{p_{\mathbf{y}}(\cdot; x), x \in \mathcal{X}\}$ *over the alphabet* $\mathcal{Y}$ *is a* one-parameter exponential family *if it can be expressed in the form*

$$p_{\mathbf{y}}(\mathbf{y}; x) = \exp\{\lambda(x)\, t(\mathbf{y}) - \alpha(x) + \beta(\mathbf{y})\} \quad \textit{for all } x \in \mathcal{X} \textit{ and } \mathbf{y} \in \mathcal{Y}, \tag{1}$$

for some choice of functions $\lambda(\cdot)\colon \mathcal{X} \mapsto \mathbb{R}$, $t(\cdot)\colon \mathcal{Y} \mapsto \mathbb{R}$, *and* $\beta(\cdot)\colon \mathcal{Y} \mapsto \mathbb{R}$.

In such a family, $\lambda(\cdot)$, $t(\cdot)$, and $\beta(\cdot)$ are referred to as the *natural parameter*, *natural statistic*, and *log base function*, respectively. Together with $\mathcal{X}$ and $\mathcal{Y}$, they specify the family.

Note, however, that the function $\alpha(x)$ in (1) is not part of the specification of the family. For a given choice of $\lambda(\cdot)$, $t(\cdot)$, and $\beta(\cdot)$, the role of $\alpha(x)$ is to normalize each distribution in the family. To see this, observe that (1) can be rewritten in the form

$$p_{\mathbf{y}}(\mathbf{y}; x) = \frac{1}{Z(x)} \exp\{\lambda(x)\, t(\mathbf{y}) + \beta(\mathbf{y})\}, \tag{2}$$

[1] It is important to emphasize that the name *exponential family* is a bit unfortunate since there is potential for confusion with the *exponential distribution*, even though they refer to very different concepts, as our development will make clear.

where the normalization constant[2]

$$Z(x) = e^{\alpha(x)} \tag{3}$$

ensures that $p_{\mathbf{y}}(\cdot; x)$ sums (or integrates) to one. Specifically, we have

$$Z(x) = e^{\alpha(x)} = \sum_{\mathbf{y}} \exp\{\lambda(x)\, t(\mathbf{y}) + \beta(\mathbf{y})\}, \tag{4}$$

where the summation in (4) is replaced with an integral when $\mathbf{y}$ is continuous-valued. Typically the parameter alphabet $\mathcal{X}$ will be some interval of the real line, corresponding to the values of x for which (1) can be normalized.

The normalization constant $Z(x)$ is often referred to as the *partition function*, a term that comes from the use of distributions of this type in statistical physics. Consequently, $\alpha(x)$ is called the *log-partition function*. These quantities play an important role in statistical inference more generally, and have meaningful physical and/or operational interpretations.[3]

We use the notation

$$\mathbf{y} \sim \mathtt{E}(\mathcal{X}, \mathcal{Y}; \lambda(\cdot), t(\cdot), \beta(\cdot)) \tag{5}$$

to indicate that the distribution of $\mathbf{y}$ has a form defined in (1). Note that implicit in this notation is that (4) is finite for all $x \in \mathcal{X}$.

It is important to recognize that the specification of an exponential family is not unique. In particular, for all c_1 and c_2, we have

$$\mathtt{E}(\mathcal{X}, \mathcal{Y}; \lambda(\cdot), t(\cdot), \beta(\cdot)) = \mathtt{E}(\mathcal{X}, \mathcal{Y}; \lambda(\cdot), t(\cdot) - c_1, \beta(\cdot) - c_2). \tag{6}$$

To see this, note that

$$\begin{aligned} \ln p_{\mathbf{y}}(\mathbf{y}; x) &= \lambda(x)\,(t(\mathbf{y}) - c_1) - \alpha(x) + (\beta(\mathbf{y}) - c_2) \\ &= \lambda(x)\, t(\mathbf{y}) - (\alpha(x) + c_1 + c_2) + \beta(\mathbf{y}) \\ &= \lambda(x)\, t(\mathbf{y}) - \tilde{\alpha}(x) + \beta(\mathbf{y}) \end{aligned}$$

where $\tilde{\alpha}(x) = \alpha(x) + c_1 + c_2$ is the renormalization. Thus, when we refer to "the exponential family $\mathtt{E}(\mathcal{X}, \mathcal{Y}; \lambda(\cdot), t(\cdot), \beta(\cdot))$," we mean this family and all its equivalents, i.e., the equivalence class of this exponential family.

One consequence of (6) is that if we set $c_1 = 0$ and choose c_2 such that[4]

$$q(\mathbf{y}) \triangleq e^{\beta(\mathbf{y}) - c_2} \tag{7}$$

[2]We refer to $Z(x)$ as a constant because it does not depend on the value of $\mathbf{y}$.

[3]In many applications, the alphabet $\mathcal{Y}$ is often large enough that evaluating the partition function as (4) prescribes is computationally intractable. As a result, it is frequently necessary to develop approximations, approaches to which will develop in a later installment of the notes.

[4]Note that there will exist such a c_2 precisely when $\lambda(x_0) = 0$ for some $x_0 \in \mathcal{X}$.

is a distribution (i.e., is normalized), then we can rewrite the exponential family (1) in the form

$$p_{\mathbf{y}}(\mathbf{y}; x) \propto q(\mathbf{y}) \exp\{\lambda(x)\, t(\mathbf{y})\}. \tag{8}$$

We refer to $q(\cdot)$ as the *base distribution* of the family.

In general, the support $\operatorname{supp}(p_{\mathbf{y}}(\cdot; x))$ of a member distribution $p_{\mathbf{y}}(\cdot; x)$— i.e., the values of $\mathbf{y} \in \mathcal{Y}$ for which $p_{\mathbf{y}}(\mathbf{y}; x) > 0$— can depend on x as well. A simple example where such a case arises is as follows.

Example 1. Let the continuous random variable y be uniformly distributed over the interval $[0, x)$ for $x > 0$, i.e., $\mathsf{y} \sim \mathtt{U}([0, x))$. Then $\mathcal{Y} = [0, \infty)$, and we have

$$p_{\mathsf{y}}(y; x) = \frac{1}{x} = e^{-\ln x}, \quad 0 \le y < x.$$

This defines an exponential family with

$$\lambda(x) = 0, \quad t(y) = y, \quad \beta(y) = 0, \quad \text{and} \quad \alpha(x) = \ln x,$$

for all $x \in \mathcal{X} = (0, \infty)$. Clearly in this case $\operatorname{supp}(p_{\mathsf{y}}(\cdot; x)) = [0, x)$, which depends on $x \in \mathcal{X}$.

When $\operatorname{supp}(p_{\mathsf{y}}(\cdot; x))$ depends on x, exponential family analysis can be more complicated. Accordingly, in the sequel, we restrict our attention to exponential families $\mathtt{E}(\mathcal{X}, \mathcal{Y}; \lambda(\cdot), t(\cdot), \beta(\cdot))$ where for every x the support of $p_{\mathbf{y}}(\cdot; x)$ does *not* depend on x, i.e., every distribution in the family has the same support. Such exponential families are termed *regular*. Consider the following examples.

Example 2. Let y be a Bernoulli random variable, taking on values 0 or 1, with $x = \mathbb{P}(\mathsf{y} = 1)$, i.e., $\mathsf{y} \sim \mathtt{B}(x)$. So $\mathcal{Y} = \{0, 1\}$, and we can write the distribution for y conveniently as

$$p_{\mathsf{y}}(y; x) = x^{\mathbb{1}_{y=1}} (1 - x)^{\mathbb{1}_{y=0}} = x^y (1 - x)^{1-y}, \quad y \in \{0, 1\},$$

whence

$$\ln p_{\mathsf{y}}(y; x) = y \ln\left(\frac{x}{1 - x}\right) + \ln(1 - x).$$

This defines a exponential family with, for example,

$$\lambda(x) = \ln\left(\frac{x}{1 - x}\right), \quad t(y) = y, \quad \beta(y) = 0, \quad \text{and} \quad \alpha(x) = -\ln(1 - x).$$

If $\mathcal{X} = (0, 1)$ then $\operatorname{supp}(p_{\mathsf{y}}(\cdot; x)) = \mathcal{Y}$ for all $x \in \mathcal{X}$, and the exponential family is regular. However, if, e.g., $\mathcal{X} = [0, 1]$ then the family will no longer be regular; indeed, $\operatorname{supp}(p_{\mathsf{y}}(\cdot; 0)) = \{0\} \neq \{1\} = \operatorname{supp}(p_{\mathsf{y}}(\cdot; 1))$.

Example 3. Let y be a scalar Gaussian random variable with mean x and unit variance, i.e., $\mathsf{y} \sim \mathtt{N}(x, 1)$. Then $\mathcal{Y} = \mathbb{R}$ and $\mathcal{X} = \mathbb{R}$, and

$$\ln p_{\mathsf{y}}(y; x) = -\frac{1}{2}\ln(2\pi) - \frac{(y-x)^2}{2} = xy - \frac{x^2}{2} - \frac{1}{2}\ln(2\pi) - \frac{y^2}{2}, \quad x \in \mathcal{X},\ y \in \mathcal{Y}$$

from which we see

$$\lambda(x) = x, \quad t(y) = y, \quad \beta(y) = -\frac{y^2}{2}, \quad \text{and} \quad \alpha(x) = \frac{x^2}{2} + \frac{1}{2}\ln(2\pi).$$

Moreover, the base distribution that corresponds to this exponential family is the normal distribution

$$q(y) = (2\pi)^{-1/2} e^{-y^2/2}.$$

Evidently, this exponential family is regular.

Example 4. Let y be an exponential random variable with mean $1/x$, so $\mathcal{Y} = [0, \infty)$ and

$$p_{\mathsf{y}}(y; x) = x e^{-xy}, \quad y \in \mathcal{Y},\ x \in \mathcal{X}$$

where $\mathcal{X} = (0, \infty)$. Then since

$$\ln p_{\mathsf{y}}(y; x) = -xy + \ln x,$$

this is a regular exponential family with

$$\lambda(x) = x, \quad t(y) = -y, \quad \beta(y) = 0, \quad \text{and} \quad \alpha(x) = -\ln x.$$

Hence, there is a particular exponential family whose members are each exponential distributions. Note that there is no base distribution for this family, since $e^{\beta(y)} = 1$ cannot be normalized over $\mathcal{Y}$ via (7).

9.2 Linear and Canonical Exponential Families

For some exponential families, the natural parameter is the parameter itself, i.e., $\lambda(x) = x$. We have already encountered an exponential family in this form in Examples 3 and 4. Because the arise frequently, we give a special name to such families:

Definition 2. *A* linear *(one-parameter) exponential family is an exponential family whose natural parameter is equal to the underlying parameter, i.e., $\lambda(x) = x$, so*

$$\ln p_{\mathbf{y}}(\mathbf{y}; x) = x\, t(\mathbf{y}) - \alpha(x) + \beta(\mathbf{y}), \quad \textit{for all } x \in \mathcal{X} \textit{ and } y \in \mathcal{Y}. \tag{9}$$

Furthermore, in both Examples 3 and 4 the natural statistic is also equal to the data, i.e., $t(y) = y$. We give a special name to linear exponential families with this additional property as well.

Definition 3. *A* canonical *(one-parameter) exponential family is a linear exponential family with $\mathcal{Y} \subset \mathbb{R}$ whose natural statistic is equal to the underlying variable, i.e., $t(y) = y$, so*

$$\ln p_{\mathcal{Y}}(y; x) = xy - \alpha(x) + \beta(y). \tag{10}$$

Linear (and canonical) families have many special properties. For example, the log-partition function serves as a cumulant-generating function for linear families:[5]

$$\begin{aligned}
\dot{\alpha}(x) &= \frac{\mathrm{d}}{\mathrm{d}x} \ln Z(x) \\
&= \frac{1}{Z(x)} \frac{\mathrm{d}Z(x)}{\mathrm{d}x} \\
&= e^{-\alpha(x)} \frac{\mathrm{d}}{\mathrm{d}x} \left(\sum_{\mathbf{y}} \exp\{xt(\mathbf{y}) + \beta(\mathbf{y})\} \right) && (11) \\
&= \sum_{\mathbf{y}} e^{-\alpha(x)} \frac{\mathrm{d}}{\mathrm{d}x} \exp\{xt(\mathbf{y}) + \beta(\mathbf{y})\} \\
&= \sum_{\mathbf{y}} t(\mathbf{y}) \exp\{xt(\mathbf{y}) + \beta(\mathbf{y}) - \alpha(x)\} && (12) \\
&= \sum_{\mathbf{y}} t(\mathbf{y}) p_{\mathbf{y}}(\mathbf{y}; x) && (13) \\
&= \mathbb{E}\left[t(\mathbf{y})\right], && (14)
\end{aligned}$$

where to obtain (11) we have used (4), and to obtain (13) we have used (9).

A similar computation establishes that the second derivative yields the variance of $t(\mathbf{y})$. Specifically, we have

$$\begin{aligned}
\ddot{\alpha}(x) = \frac{\mathrm{d}}{\mathrm{d}x}\dot{\alpha}(x) &= \frac{\mathrm{d}}{\mathrm{d}x} \sum_{\mathbf{y}} t(\mathbf{y}) \exp\{xt(\mathbf{y}) + \beta(\mathbf{y}) - \alpha(x)\} && (15) \\
&= \sum_{\mathbf{y}} t(\mathbf{y}) \frac{\mathrm{d}}{\mathrm{d}x} \exp\{xt(\mathbf{y}) + \beta(\mathbf{y}) - \alpha(x)\} \\
&= \sum_{\mathbf{y}} t(\mathbf{y}) \left[t(\mathbf{y}) - \dot{\alpha}(x)\right] \exp\{xt(\mathbf{y}) + \beta(\mathbf{y}) - \alpha(x)\} \\
&= \sum_{\mathbf{y}} t(\mathbf{y}) \left[t(\mathbf{y}) - \mathbb{E}\left[t(\mathbf{y})\right]\right] p_{\mathbf{y}}(\mathbf{y}; x) && (16) \\
&= \sum_{\mathbf{y}} \left[t(\mathbf{y}) - \mathbb{E}\left[t(\mathbf{y})\right]\right]^2 p_{\mathbf{y}}(\mathbf{y}; x) + \underbrace{\sum_{\mathbf{y}} \mathbb{E}\left[t(\mathbf{y})\right] \left[t(\mathbf{y}) - \mathbb{E}\left[t(\mathbf{y})\right]\right] p_{\mathbf{y}}(\mathbf{y}; x)}_{0} \\
&= \mathrm{var}\left[t(\mathbf{y})\right], && (17)
\end{aligned}$$

[5]We use the notation $\dot{}$ and $\ddot{}$ to denote first and second derivatives, respectively.

where to obtain (15) we have used (12), where to obtain (16) we have used (14) and (9).

Moreover, since

$$\frac{\partial}{\partial x} \ln p_{\mathsf{y}}(\mathsf{y}; x) = \frac{\partial}{\partial x}(x\, t(\mathsf{y}) - \alpha(x) + \beta(\mathsf{y})) = t(\mathsf{y}) - \dot{\alpha}(x) = t(\mathsf{y}) - \mathbb{E}\left[t(\mathsf{y})\right],$$

it follows that the Fisher information in $\mathbf{y}$ about x is

$$J_{\mathbf{y}}(x) = \mathbb{E}\left[\left(\frac{\partial}{\partial x} \ln p_{\mathbf{y}}(\mathbf{y}; x)\right)^2\right] = \mathrm{var}\left[t(\mathbf{y})\right] = \ddot{\alpha}(x). \tag{18}$$

For canonical families, these derivatives obviously yield the mean and variance of y. Moreover, when the canonical family can be written in the form (8), as is always the case when $\mathcal{Y}$ is finite, the partition function can be interpreted as the *moment-generating function* corresponding to the base distribution $q(\cdot)$, i.e.,

$$Z(x) = e^{\alpha(x)} = \sum_{\mathbf{y}} \exp\{xy + \ln q(y)\} = \mathbb{E}_q\left[e^{x\mathsf{y}}\right], \tag{19}$$

where the notation $\mathbb{E}_q[\cdot]$ expresses that the expectation is with respect to the base distribution $q(\cdot)$. In turn, the log-partition function $\alpha(x)$ can be interpreted as the corresponding *cumulant-generating function*.

9.3 Geometric Means and Tiltings of Distributions

As we have seen, many familiar parameterized distributions can be expressed as exponential families. In this section, we describe two ways in which exponential families naturally arise when we study the space of probability distributions. As we will see, for the case of finite alphabets $\mathcal{Y}$, these lead to useful alternative characterizations of exponential families.

9.3.1 Geometric Mean of Two Distributions

Consider two strictly positive probability distributions, $p_1(\cdot)$ and $p_2(\cdot)$ and define the (normalized) *weighted geometric mean* of these distributions via

$$p_{\mathbf{y}}(\mathbf{y}; x) = \frac{p_1(\mathbf{y})^x p_2(\mathbf{y})^{1-x}}{Z(x)}, \quad x \in \mathcal{X} = [0, 1]. \tag{20}$$

Since

$$\ln p_{\mathbf{y}}(\mathbf{y}; x) = x \ln \frac{p_1(\mathbf{y})}{p_2(\mathbf{y})} + \ln p_2(\mathbf{y}) - \ln Z(x), \tag{21}$$

we see that (20) describes an exponential family with

$$\lambda(x) = x, \quad t(\mathbf{y}) = \ln \frac{p_1(\mathbf{y})}{p_2(\mathbf{y})}, \quad \beta(\mathbf{y}) = \ln p_2(\mathbf{y}), \quad \alpha(x) = \ln Z(x). \tag{22}$$

Example 5. Consider the case $\mathcal{Y} = \{0, 1\}$ and let $p_1 = \mathtt{B}(1/(1+e^{-1}))$ and $p_2 = \mathtt{B}(1/2)$. Then the normalized weighted geometric mean is $p_y(\cdot; x) = \mathtt{B}(1/(1+e^{-x}))$, which can be verified via

$$\frac{p_y(1;x)}{p_y(0;x)} = \frac{\left(\frac{e}{1+e}\right)^x \left(\frac{1}{2}\right)^{1-x}}{\left(\frac{1}{1+e}\right)^x \left(\frac{1}{2}\right)^{1-x}} = e^x.$$

Example 6. Consider the case $\mathcal{Y} = \mathbb{R}$ and let $p_1 = \mathtt{N}(1, 1)$ and $p_2 = \mathtt{N}(0, 1)$. Then their weighted geometric mean satisfies

$$\begin{aligned} p_y(y;x) &\propto \exp\left\{-\frac{x}{2}(y-1)^2\right\} \exp\left\{-\frac{(1-x)}{2}y^2\right\} \\ &= \exp\left\{-\frac{x}{2}\left[(y-1)^2 - y^2\right] - \frac{1}{2}y^2\right\} \\ &\propto \exp\left\{-\frac{1}{2}(y-x)^2\right\}, \end{aligned}$$

i.e., $p_y(\cdot; x) = \mathtt{N}(x, 1)$.

In fact, it follows readily that *any* linear exponential family of the form (9) over a finite alphabet $\mathcal{Y}$ can be expressed as the geometric mean of two suitably chosen distributions. To see this, suppose we are given a family specified by $t(\cdot)$ and $\beta(\cdot)$. Then choose c_1 and $p_2(\cdot)$ such that

$$p_2(\mathbf{y}) = c_1\, e^{\beta(\mathbf{y})},$$

where we see $c_1 > 0$ is a normalizing constant.[6] Next, we choose c_2 and $p_1(\cdot)$ such that

$$p_1(\mathbf{y}) = c_2\, p_2(\mathbf{y})\, e^{t(\mathbf{y})},$$

where $c_2 > 0$ is also a normalizing constant. Then the normalized geometric mean of p_1 and p_2 is

$$\begin{aligned} p(\mathbf{y};x) &\propto \exp\left\{x \ln \frac{p_1(\mathbf{y})}{p_2(\mathbf{y})} + \ln p_2(\mathbf{y})\right\} && (23) \\ &= \exp\left\{x(t(\mathbf{y}) + \ln c_2) + (\beta(\mathbf{y}) + \ln c_1)\right\} && (24) \\ &\propto \exp\left\{xt(\mathbf{y}) + \beta(\mathbf{y})\right\} && (25) \end{aligned}$$

so

$$p_y(\mathbf{y};x) = \exp\left\{xt(\mathbf{y}) - \alpha(x) + \beta(\mathbf{y})\right\},$$

with, as usual, $\alpha(x)$ chosen to normalize the distribution.

In addition, the preceding development implies, in turn, that given any two distributions $p_1(\cdot)$ and $p_2(\cdot)$, not only is there a linear exponential family of the form (9) that includes both as members, but that linear exponential family is *unique* (within the equivalence class discussed earlier). In particular, to within additive constants $t(\cdot)$ and $\beta(\cdot)$ must be of the form given in (22).

[6]Note that for a infinite alphabet $\mathcal{Y}$, such a normalizing constant may not exist.

9.3.2 Tilting a Distribution

Let's consider distribution $q(\cdot)$ defined for a scalar random variable and define

$$p_{\mathbf{y}}(\mathbf{y}; x) = \frac{q(\mathbf{y})e^{x\mathbf{y}}}{Z(x)}. \tag{26}$$

$p_{\mathbf{y}}(\cdot; x)$ is called a *tilted distribution* because it weighs larger values of $\mathbf{y}$ exponentially more than the base distribution $q(\cdot)$. It is a member of the one-parameter exponential family defined by

$$\lambda(x) = x, \quad t(\mathbf{y}) = \mathbf{y}, \quad \beta(\mathbf{y}) = \ln q(\mathbf{y}), \quad \alpha(x) = \ln Z(x).$$

Example 7. Let $\mathcal{Y} = \{0, 1\}$ and $q \sim \mathtt{B}(1/2)$. Then the tilted distribution is $p_y(\cdot; x) \sim \mathtt{B}(1/(1 + e^{-x}))$ since

$$\frac{p_y(1; x)}{p_y(0; x)} = \frac{e^x/2}{1/2} = e^x.$$

Evidently, tilting skews the Bernoulli distribution to make $y = 1$ more probable.

Example 8. For $\mathcal{Y} = \mathbb{R}$, the distribution $p_y(\cdot; x) \sim \mathtt{N}(x, 1)$ is an example of a tilted distribution whose base density is $q \sim \mathtt{N}(0, 1)$. In this case, tilting shifts the mean of the distribution by the value of the parameter x but does not change the shape of the distribution.

Evidently, *any* canonical exponential family of the form (10) over a finite alphabet $\mathcal{Y}$ can be expresed as a tilted distribution. In particular, the base distribution q is derived via (7).[7]

9.4 Efficient Estimators and Exponential Families

There is a close connection between exponential families and the existence of efficient estimators, i.e., unbiased estimators that achieve the Cramér-Rao Bound, as we now describe for the case of a scalar parameter.

We begin with a necessary condition for efficiency.

Claim 1. *If an efficient estimator exists for estimating a nonrandom parameter x from observations $\mathbf{y}$, then the model $p_{\mathbf{y}}(\mathbf{y}; x)$ is a member of an exponential family with*

$$\lambda(x) = \int^x J_{\mathbf{y}}(u)\,\mathrm{d}u \quad \textit{and} \quad t(\mathbf{y}) = \hat{x}_{\mathrm{ML}}(\mathbf{y}), \tag{27}$$

where $\hat{x}_{\mathrm{ML}}(\mathbf{y})$ is the maximum likelihood estimate of x based on $\mathbf{y}$.

[7] We emphasize that for infinite alphabets $\mathcal{Y}$, such a base distribution may not exist.

Proof. Recall that if an efficient estimator exists, then it must satisfy

$$\hat{x}_{\text{eff}}(\mathbf{y}) = x + J_{\mathbf{y}}^{-1}(x)\frac{\partial}{\partial x}\ln p_{\mathbf{y}}(\mathbf{y};x).$$

Rearranging terms, we obtain

$$\frac{\partial}{\partial x}\ln p_{\mathbf{y}}(\mathbf{y},x) = \hat{x}_{\text{eff}}(\mathbf{y})J_{\mathbf{y}}(x) - xJ_{\mathbf{y}}(x),$$

afterwhich integrating with respect to x yields

$$\ln p_{\mathbf{y}}(\mathbf{y},x) = \left(\int^x J_{\mathbf{y}}(u)\,\mathrm{d}u\right)\hat{x}_{\text{eff}}(\mathbf{y}) - \int^x uJ_{\mathbf{y}}(u)\,\mathrm{d}u + \beta(\mathbf{y}),$$

where $\beta(\mathbf{y})$ is an arbitrary function of $\mathbf{y}$ that does not depend on x. But this is exactly the form of an element of an exponential family with

$$\lambda(x) = \int^x J_{\mathbf{y}}(u)\,\mathrm{d}u, \quad t(\mathbf{y}) = \hat{x}_{\text{eff}}(\mathbf{y}), \quad \alpha(x) = \int^x uJ_{\mathbf{y}}(u)\,\mathrm{d}u.$$

Finally, as we developed earlier, if an efficient estimator exists, it is the maximum likelihood estimator, whence (27). □

Hence, we see that for an efficient estimator to exist, it is necessary for the model to be in an exponential family. However, it should be emphasized that the model being in an exponential family is not sufficient for an efficient estimator to exist.

A sufficient condition for efficiency is as follows.

Claim 2. *If the model is from a linear exponential family, and the Fisher information $J_{\mathbf{y}}(x)$ does not depend on the parameter x, then an efficient estimator exists.*

Proof. Let $J_{\mathbf{y}}(x) \triangleq J$. From (18), we then have that $\ddot{\alpha}(x) = J$. Hence, $\dot{\alpha}(x) = Jx - c$ for some constant c. Now an efficient estimator exists since

$$\begin{aligned}
\hat{x}_{\text{eff}}(\mathbf{y}) &= x + J_{\mathbf{y}}^{-1}(x)\frac{\partial}{\partial x}\ln p_{\mathbf{y}}(\mathbf{y};x) && (28)\\
&= x + \frac{1}{J}\left[t(\mathbf{y}) - \dot{\alpha}(x)\right] && (29)\\
&= \frac{c}{J} + \frac{1}{J}t(\mathbf{y}) && (30)\\
&= \frac{c}{J} + \frac{1}{J}\left[J\hat{x}_{\text{ML}}(\mathbf{y}) - c\right] && (31)\\
&= \hat{x}_{\text{ML}}(\mathbf{y}) && (32)
\end{aligned}$$

does not depend on x, where we have used that for the linear exponential family

$$\frac{\partial}{\partial x}\ln p_{\mathbf{y}}(\mathbf{y};x) = t(\mathbf{y}) - \dot{\alpha}(x), \tag{33}$$

and, hence, that for this family

$$t(\mathbf{y}) = \dot{\alpha}(\hat{x}_{\mathrm{ML}}(\mathbf{y}). \tag{34}$$

□

At least within the class of linear exponential families, it follows from the above proof that the Fisher information being constant is also necessary for efficiency.

Corollary 1. *If the model is from a linear exponential family and an efficient estimator exists, then the Fisher information $J_{\mathsf{y}}(x)$ does not depend on x.*

9.5 Exponential Families with Multiple Parameters

The ideas we discussed so far in this section extend to distributions with multiple parameters.

Definition 4. *A parameterized family of distributions $p(\cdot;\mathbf{x})$ is a K-parameter exponential family with natural parameter $\boldsymbol{\lambda}(\cdot) = [\lambda_1(\cdot),\ldots,\lambda_K(\cdot)]^{\mathrm{T}}\colon \mathcal{X} \mapsto \mathbb{R}^K$, natural statistic $\mathbf{t}(\cdot) = [t_1(\cdot),\ldots,t_K(\cdot)]^{\mathrm{T}}\colon \mathcal{Y} \mapsto \mathbb{R}^K$, and log base function $\beta(\cdot)\colon \mathcal{Y} \mapsto \mathbb{R}$ if each member of the family is of the form*

$$\begin{aligned} p_{\mathsf{y}}(\mathbf{y};\mathbf{x}) &= \exp\left\{\sum_{i=1}^{K} \lambda_i(\mathbf{x})\, t_i(\mathbf{y}) - \alpha(\mathbf{x}) + \beta(\mathbf{y})\right\} \\ &= \exp\left\{\boldsymbol{\lambda}^{\mathrm{T}}(\mathbf{x})\, \mathbf{t}(\mathbf{y}) - \alpha(\mathbf{x}) + \beta(\mathbf{y})\right\} \end{aligned} \tag{35}$$

We begin with a simple example.

Example 9. Let y be a scalar Gaussian random variable with mean x_1 and variance x_2. Then

$$\ln p_{\mathsf{y}}(y;\mathbf{x}) = -\frac{1}{2}\ln(2\pi x_2) - \frac{(y-x_1)^2}{2x_2} = \frac{x_1}{x_2}y - \frac{1}{2x_2}y^2 - \frac{x_1^2}{2x_2} - \frac{1}{2}\ln(2\pi x_2),$$

which corresponds to a two-parameter exponential family with

$$\boldsymbol{\lambda}(\mathbf{x}) = \begin{bmatrix} x_1/2x_2 \\ -1/2x_2 \end{bmatrix}, \quad \mathbf{t}(y) = \begin{bmatrix} y \\ y^2 \end{bmatrix}, \quad \beta(y) = 0, \quad \text{and} \quad \alpha(\mathbf{x}) = \frac{x_1^2}{2x_2} + \frac{1}{2}\ln 2\pi x_2.$$

Consistent with (5), we use $\mathtt{E}(\mathcal{X},\mathcal{Y};\boldsymbol{\lambda}(\cdot),\mathbf{t}(\cdot),\beta(\cdot))$ to denote a family of the form (35) and its equivalents. In the case of multiple parameters, however, the equivalence

class is larger still. To see this, consider the case of a finite alphabet $\mathcal{Y} = \{1, 2, \ldots, M\}$, with $M \geq m$, in which case we can write (35) in vector-matrix form as

$$\begin{bmatrix} \ln p_y(1;\mathbf{x}) \\ \vdots \\ \ln p_y(M;\mathbf{x}) \end{bmatrix} = \underbrace{\begin{bmatrix} t_1(1) & \cdots & t_K(1) \\ \vdots & \ddots & \vdots \\ t_1(M) & \cdots & t_K(M) \end{bmatrix}}_{\triangleq \mathbf{T}} \begin{bmatrix} \lambda_1(\mathbf{x}) \\ \vdots \\ \lambda_K(\mathbf{x}) \end{bmatrix} - \alpha(\mathbf{x}) \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} + \begin{bmatrix} \beta(1) \\ \vdots \\ \beta(M) \end{bmatrix}. \quad (36)$$

If the columns of $\mathbf{T}$ are linearly dependent,[8] then the representation is redundant, and (36) can be equivalently expressed using fewer parameters a linearly independent subset of the columns of $\mathbf{T}$. When $\mathbf{T}$ has full rank (i.e., $\text{rank}(\mathbf{T}) = K$), there is no such redundancy. In this case we refer to the representation as *minimal*. In the case of continuous alphabets $\mathcal{Y}$, there is a corresponding notion of equivalence and minimality based on the properties of the Jacobian matrix[9] $\partial \mathbf{t}(\mathbf{y})/\partial \mathbf{y}$, but the analysis is more complicated.

The possible relationships between the actual and natural parameter vectors is also richer than in the one-parameter case. For example, for K-parameter exponential families, the dimensionality of the parameter $\mathbf{x}$ need not be equal to the number of natural parameters K, i.e., $\mathcal{X}$ need not be a subset of $\mathbb{R}^K$. More generally, it is useful to at least coarsely distinguish different classes of behavior.

Case 1: $\text{rank}(\partial \boldsymbol{\lambda}(\mathbf{x})/\partial \mathbf{x}) < L$ **for** $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^L$**.** In this case the family is *over-parameterized*, and the mapping $\boldsymbol{\lambda}(\cdot)$ is not invertible, so the parameters $\mathbf{x}$ are not identifiable. Such cases do not arise in well-formulated inference problems. Simple examples with $\mathbf{0} \notin \mathcal{X}$ include

$$\lambda(\mathbf{x}) = x_1 + x_2, \quad \boldsymbol{\lambda}(\mathbf{x}) = [x_1 + x_2^2, x_1 + x_2^2]^{\mathrm{T}}, \quad \text{and} \quad \boldsymbol{\lambda}(\mathbf{x}) = [x_1 + x_2, (x_1 + x_2)^2]^{\mathrm{T}}.$$

Since $\text{rank}(\partial \boldsymbol{\lambda}(\mathbf{x})/\partial \mathbf{x} \leq K$, this case always occurs when $K < L$.

Case 2: $\text{rank}(\partial \boldsymbol{\lambda}(\mathbf{x})/\partial \mathbf{x}) = L = K$ **for** $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^L$**.** In this case the family is referred to as *full*. Simple examples with $\mathbf{0} \notin \mathcal{X}$ include

$$\boldsymbol{\lambda}(\mathbf{x}) = \mathbf{x}, \quad \boldsymbol{\lambda}(\mathbf{x}) =, \quad \text{and} \quad \boldsymbol{\lambda}(\mathbf{x}) = [x_1 + x_2^2, x_1^2 + x_2]^{\mathrm{T}}.$$

[8] Recall that linear dependence means that one column can be expressed as a linear combination of the others.

[9] Recall that an infinite differentiable vector-valued function $\mathbf{g}(\dot{})$ of a vector argument can be approximated in a neighborhood of the argument $\mathbf{u}_0$ via

$$\mathbf{g}(\mathbf{u}) \cong \mathbf{g}(\mathbf{u}_0) + \frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}}(\mathbf{u} - \mathbf{u}_0).$$

so the Jacobian describes the locally linear behavior of the function.

Case 3: $L \leq \operatorname{rank}(\partial \boldsymbol{\lambda}(\mathbf{x})/\partial \mathbf{x}) = L < K$ **for** $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^L$**.** In this case the family is referred to as *curved*. A simple example with $\mathbf{0} \notin \mathcal{X}$ is

$$\boldsymbol{\lambda}(x) = [x, x^2]^{\mathrm{T}}.$$

Example 10. Let $\mathcal{Y} = \mathbb{R}$ and consider the family of distributions $p_y(\cdot; x) = \mathtt{N}(x, x^2)$ for $x \in \mathcal{X} = (0, \infty)$. Then

$$\ln p_y(y; x) \propto -\frac{1}{2x^2}(y - x)^2 \propto \frac{-1}{2x^2}\, y^2 + \frac{1}{x}\, y, \tag{37}$$

from which we see that this is a curved 2-parameter exponential family with

$$\boldsymbol{\lambda}(x) = \begin{bmatrix} -1/(2x^2) & 1/x \end{bmatrix}^{\mathrm{T}}, \quad \mathbf{t}(y) = \begin{bmatrix} y^2 & y \end{bmatrix}^{\mathrm{T}}, \quad \text{and} \quad \beta(y) \equiv 1.$$

The definitions of linear and canonical exponential families naturally extend to the vector case. A linear K-parameter exponential family is characterized by $\boldsymbol{\lambda}(\mathbf{x}) = \mathbf{x}$. A canonical K-parameter exponential family must further satisfy $\mathbf{t}(\mathbf{y}) = \mathbf{y}$. For a linear family, the log-partition function plays the role of a cumulant-generating function. In particular, it is easy to show that

$$\frac{\partial \alpha(\mathbf{x})}{\partial x_i} = \mathbb{E}\left[t_i(\mathbf{y})\right], \quad i = 1, \ldots, K, \tag{38}$$

$$\frac{\partial^2 \alpha(\mathbf{x})}{\partial x_i \partial x_j} = \operatorname{cov}(t_i(\mathbf{y}), t_j(\mathbf{y})), \quad i, j = 1, \ldots, K, \tag{39}$$

or equivalently in vector form,

$$\frac{\partial \alpha(\mathbf{x})}{\partial \mathbf{x}} = \mathbb{E}\left[\mathbf{t}(\mathbf{y})\right]^{\mathrm{T}} = \boldsymbol{\mu}_{\mathbf{t}}^{\mathrm{T}},$$

$$\frac{\partial^2 \alpha(\mathbf{x})}{\partial \mathbf{x}^2} = \mathbb{E}\left[(\mathbf{t}(\mathbf{y}) - \boldsymbol{\mu}_{\mathbf{t}})(\mathbf{t}(\mathbf{y}) - \boldsymbol{\mu}_{\mathbf{t}})^{\mathrm{T}}\right] = \boldsymbol{\Lambda}_{\mathbf{t}}.$$

There is also an analogous relationship between efficiency, maximum likelihood estimation, and exponential families for the vector parameter case. We leave the derivation as an exercise for the reader.

Example 11. Let $p_y(y; \mathbf{x})$ with $\mathbf{x} = (x_1, \ldots, x_K)$ denote the weighted and normalized geometric mean of $K + 1$ strictly positive distributions $p_1(\cdot), \ldots, p_{K+1}(\cdot)$, i.e.,

$$p_y(y; \mathbf{x}) = \frac{1}{Z(\mathbf{x})} p_1(y)^{x_1} p_2(y)^{x_2} \cdots p_K(y)^{x_K} p_{K+1}(y)^{1 - x_1 - x_2 - \cdots - x_K}.$$

Then

$$\begin{aligned} p_y(y) &= x_1 \ln p_1(y) + x_2 \ln p_2(y) + \cdots + x_K \ln p_K(y) \\ &\qquad + (1 - x_1 - x_2 - \cdots - x_K) \ln p_{K+1}(y) - \alpha(\mathbf{x}) \end{aligned} \tag{40}$$

$$= \ln p_{K+1}(y) - \alpha(\mathbf{x}) + \sum_{i=1}^{K} x_i \ln \frac{p_i(y)}{p_K(y)}, \tag{41}$$

so this is a K-parameter linear exponential family with

$$t_i(y) = \ln \frac{p_i(y)}{p_K(y)}, \quad i = 1, \ldots, K \qquad \text{and} \qquad \beta(y) = \ln p_{K+1}(y). \tag{42}$$

Conversely, as in the one-parameter case, any (regular) K-parameter linear exponential family can be expressed as the weighted and normalized geometric mean of $K+1$ suitably chosen distributions. Moreover, given any (strictly positive) distributions $p_1(\cdot), \ldots, p_{K+1}(\cdot)$ over a finite alphabet $\mathcal{Y}$, there is, to within the usual equivalence class, a *unique* minimal K-parameter exponential family that includes all these $K+1$ distributions, and this is the exponential family specified in (42) where we have freely chosen $\lambda(\cdot)$ to be the identity.

9.6 Additional Perspectives on Exponential Families

We conclude with some useful further insights into exponential families.

9.6.1 Reparameterization

When we specify an exponential family, we are specifying not only the collection of distributions it comprises, but also a particular parameterization of this collection. The collection itself is primarily specified through the choice of $\mathbf{t}(\cdot)$ and $\beta(\cdot)$, while the parameterization is specified through the choice of $\boldsymbol{\lambda}(\cdot)$.[10]

As such, the parameterization represents how we index the members of the family. In some cases the underlying parameter $\mathbf{x}$ has a physically meaningful interpretation in the problem at hand, and is thus of direct interest. Moreover, the choice of mapping $\boldsymbol{\lambda}(\cdot)$ it can resolve some member distributions of the collection more than others.

In some applications, however, only the collection itself is of interest, and the parameterization itself is immaterial. In such cases, we often choose a parameterization that is convenient for the application of interest. From this perspective, the linear exponential families, whereby $\boldsymbol{\lambda}(\mathbf{x}) = \mathbf{x}$, represent one simple parameterization that is convenient for many computations. Ultimately then, in cases when only the collection matters and not its parameterization, the class of exponential families that are equivalent to a given one is even larger than discussed earlier and we can typically take the corresponding linear exponential family as its representative, so that $\boldsymbol{\lambda}(\cdot)$ is no longer part of the specification of the family, only $\mathcal{X}$, $\mathbf{t}(\cdot)$, and $\beta(\cdot)$.

To see this more concretely, note that if we reparameterize an exponential family, it remains an exponential family with the same natural statistic and log base function. In particular, if $\mathbf{g}(\cdot)$ is an invertible mapping and we choose the reparameterization

[10] Obviously, the choices of $\boldsymbol{\lambda}(\cdot)$ and $\mathcal{X}$ together can also affect the extent of the collection, but this need not be the case.

$\mathbf{x} = \mathbf{g}(\tilde{\mathbf{x}})$, then

$$\begin{aligned}\ln p_{\mathbf{y}}(\mathbf{y};\mathbf{x}) &= \boldsymbol{\lambda}(\mathbf{x})^{\mathrm{T}}\,\mathbf{t}(\mathbf{y}) - \alpha(\mathbf{x}) + \beta(\mathbf{y}), \quad \mathbf{x} \in \mathcal{X} \\ &= \boldsymbol{\lambda}(\mathbf{g}(\tilde{\mathbf{x}}))^{\mathrm{T}}\,\mathbf{t}(\mathbf{y}) - \alpha(\mathbf{g}(\tilde{\mathbf{x}})) + \beta(\mathbf{y}), \quad \tilde{\mathbf{x}} \in \mathbf{g}^{-1}(\mathcal{X}) \\ &= \tilde{\boldsymbol{\lambda}}(\tilde{\mathbf{x}})^{\mathrm{T}}\,\mathbf{t}(\mathbf{y}) - \tilde{\alpha}(\tilde{\mathbf{x}}) + \beta(\mathbf{y}) \triangleq \ln \tilde{p}_{\mathbf{y}}(\mathbf{y};\tilde{\mathbf{x}}), \quad \tilde{\mathbf{x}} \in \tilde{\mathcal{X}},\end{aligned}$$

where $\tilde{\boldsymbol{\lambda}}(\cdot) = \boldsymbol{\lambda}(\mathbf{g}(\cdot))$, $\tilde{\alpha}(\cdot) = \alpha(\mathbf{g}(\cdot))$, and $\tilde{\mathcal{X}} = \mathbf{g}^{-1}(\mathcal{X})$. In other words, in this case, we are describing exactly the same family of distributions, but indexing them via a different parameterization. From this perspective, $\mathtt{E}(\mathcal{X}, \mathcal{Y}; \boldsymbol{\lambda}(\cdot), \mathbf{t}(\cdot), \beta(\cdot))$ and $\mathtt{E}(\tilde{\mathcal{X}}, \mathcal{Y}; \tilde{\boldsymbol{\lambda}}(\cdot), \mathbf{t}(\cdot), \beta(\cdot))$ refer to the same exponential family.

9.6.2 Local Families of Distributions

In this section, we show that sufficiently well-behaved families of distributions are effectively "locally" exponential.

To see this, suppose we have a family of strictly positive distributions $\{p_y(y;x), x \in \mathcal{X} \subset \mathbb{R}\}$ whose members are infinitely differentiable. Without loss of generality, assume $x = 0$ is in the interior of $\mathcal{X}$. Then we can expand $\ln p_y(y;x)$ in the following Taylor series about $x = 0$:

$$\ln p_y(y;x) = \ln p_y(y;0) + x\left[\frac{\partial}{\partial x}\ln p_y(y;x)\right]\Bigg|_{x=0} + \mathcal{O}(x^2), \quad \text{as } x \to 0, \tag{43}$$

where $\mathcal{O}(x^2)$ denotes the higher order terms—the terms whose contribution is negligible compared to the linear term. Hence, we see that $p_y(y;x)$ behaves like a linear exponential family with

$$t(y) = \frac{\partial}{\partial x}\ln p_y(y;x)\Bigg|_{x=0} \qquad \text{and} \qquad \beta(y) = \ln p_y(y;0)$$

in a neighborhood of $x = 0$.

An analogous result is possible for corresponding families distributions with vector parameters, i.e., $\{p_y(y;\mathbf{x}), \mathbf{x} \in \mathcal{X} \subset \mathbb{R}^K\}$. In this case, we obtain

$$\begin{aligned}\ln p_y(y;\mathbf{x}) &\cong \ln p_y(y;\mathbf{0}) + \mathbf{x}^{\mathrm{T}}\left[\frac{\partial}{\partial x}\ln p_y(y;\mathbf{x})\right]\Bigg|_{\mathbf{x}=\mathbf{0}} \\ &= \ln p_y(y;\mathbf{0}) + \sum_{i=1}^{K} x_i\left[\frac{\partial}{\partial x_i}\ln p_y(y;\mathbf{x})\right]\Bigg|_{\mathbf{x}=\mathbf{0}}\end{aligned}$$

via the vector Taylor series expansion, from which we conclude that $p_y(y;\mathbf{x})$ behaves like a K-parameter linear exponential family with

$$t_i(y) = \frac{\partial}{\partial x_i}\ln p_y(y;\mathbf{x})\Bigg|_{\mathbf{x}=\mathbf{0}} \qquad \text{and} \qquad \beta(y) = \ln p_y(y;\mathbf{0})$$

in a neighborhood of $\mathbf{x} = \mathbf{0}$.

9.6.3 The Global Family of Distributions

In this section, returning to a global view of exponential distributions, we consider the question: given how many common classes of distributions are in exponential families, to what degree can an exponential family "cover" the collection of all possible probability distributions. In the case of distributions over a finite alphabet $\mathcal{Y}$ the answer is straightforward: it is possible to construct a linear exponential family that includes *all* such distributions.

To see this, with, e.g., $\mathcal{Y} \triangleq \{1, \ldots, M\}$, let the M-dimensional linear family be

$$p_{\mathsf{y}}(y; \mathbf{x}) = \exp\left[\sum_{i=1}^{M} x_i \, t_i(y) - \alpha(\mathbf{x})\right]$$

with the natural statistics

$$t_i(y) = \mathbb{1}_{y=i}, \qquad i = 1, \ldots, M.$$

Then

$$p_{\mathsf{y}}(j; \mathbf{x}) \propto e^{x_j},$$

so an arbitrary distribution $q(\cdot)$ over $\mathcal{Y}$ is a member of the family corresponding to the following choice of $\mathbf{x}$:

$$x_i = \ln q(i), \qquad i = 1, 2, \ldots, M.$$

In effect, we have generated this exponential family as the weighted and normalized geometric mean of the M elementary distributions $p_i(y) = \mathbb{1}_{y=i}$, $i = 1, \ldots, M$.

Note that for large alphabets, the associated exponential family is high-dimensional. Obviously, in practice, it would be desirable to have relatively low-dimensional exponential families that "cover" the space of all distributions as much as possible even when the alphabet is large.

For the case of infinite alphabets corresponding to continuous-valued random variables, it is possible to show that there exists a sequence of exponential families that are "dense" in the space of all suitably well-behaved distributions over continuous alphabets $\mathcal{Y} \subset \mathbb{R}$. To show this requires one additional concept we don't yet have: the right measure of "closeness" of two distributions with respect to problems of inference. We therefere briefly postpone our development of this exponential family approximation result.

9.6.4 A Computational View of Inference with Exponential Families

As a final note, observe that when computing the probability distribution $p_{\mathbf{y}}(\mathbf{y}; \mathbf{x})$ of an exponential family (35) the data $\mathbf{y}$ participates in only through the natural statistics $\mathbf{t}(\mathbf{y}) = [t_1(\mathbf{y}), \ldots, t_K(\mathbf{y})]^{\mathrm{T}}$ and the log base function $\beta(\mathbf{y})$. Once we computed

these $K+1$ numbers, we can throw away the data and still perform perfect inference based on the saved statistics. In fact, we will show that for exponential families, we don't even need to keep $\beta(\mathbf{y})$. The next installment of the notes develops the notion of *sufficient statistics*, which are functions that summarize data for the purpose of inference without any loss.

9.7 Further reading

The text by Bernardo and Smith provides a precise if dense introduction to the topic of exponential families. Other references on our list use exponential families and briefly discuss them, but do not cover the topic in depth. Some of them also discuss exponential families in conjunction with information theoretic measures, which we will develop in an upcoming installment of the notes.