Exponential Families

Neil Girdhar

March 24, 2020

The *exponential families* are an important class of probability distributions that include the normal, gamma, beta, exponential, Poisson, binomial, and Bernoulli distributions. In this section, based on presentations by Nielsen and Garcia (2011) and Shao (2003, p. 66), we describe some of the exponential families' many exciting properties.

1 Definition

For simplicity, we will restrict ourselves to discrete and continuous distributions; the general, measure-theoretic definition (Shao 2003, p. 66) is analogous. A *natural exponential family* is a family of probability distributions parametrized by $\boldsymbol{\eta}$ and whose probability mass function or density function can be decomposed as:

$$f(\mathbf{x} \mid \boldsymbol{\eta}) = \exp\bigl(T(\mathbf{x})^{\mathsf{T}}\boldsymbol{\eta} - g(\boldsymbol{\eta}) + h(\mathbf{x})\bigr), \qquad \mathbf{x} \in \Omega \tag{1}$$

where

- Ω is the *support*,
- $\boldsymbol{\eta}$ are the *natural parameters*,
- $T(\mathbf{x})$ is the *sufficient statistic*,
- $g(\boldsymbol{\eta})$ is the *log-normalizer*, and
- $h(\mathbf{x})$ is the *carrier measure*.

(See Appendix **??** for examples.)

1.1 Natural parameters

The decomposition of an exponential family in equation 1 is not unique. Any transformation

$$\boldsymbol{\eta}' = D\boldsymbol{\eta} \qquad\qquad T' = \left[D^{\mathsf{T}}\right]^{-1} T \tag{2}$$

where D is a nonsingular matrix (a bijective linear map) gives another representation of the same natural exponential family.

If $\boldsymbol{\eta}$ were replaced by an arbitrary function $\boldsymbol{\eta}(\boldsymbol{\theta})$ of parameters $\boldsymbol{\theta}$, then the family of probability distributions is called an *exponential family*. We avoid this general form, preferring the so-called *canonical form*.

1.2 Sufficient statistic

The sufficient statistic is a vector-valued function of only the outcome $\mathbf{x}$. Its name is justified by its connection to the maximum entropy formulation (§1.5) and to maximum likelihood estimation (§3.2).

1.3 Log-normalizer

The log-normalizer is a scalar-valued function of only the natural parameters $\boldsymbol{\eta}$. It is so-named because

$$1 = \int_{\mathbf{x}} f(\mathbf{x} \mid \boldsymbol{\eta})\ \mathrm{d}\mathbf{x} = \int_{\mathbf{x}} \exp\bigl(T(\mathbf{x})^{\mathsf{T}}\boldsymbol{\eta} - g(\boldsymbol{\eta}) + h(\mathbf{x})\bigr)\ \mathrm{d}\mathbf{x} \tag{3}$$

$$\Downarrow$$

$$g(\boldsymbol{\eta}) = \log \int_{\mathbf{x}} \exp\bigl(T(\mathbf{x})^{\mathsf{T}}\boldsymbol{\eta} + h(\mathbf{x})\bigr)\ \mathrm{d}\mathbf{x}\,. \tag{4}$$

The log-normalizer is strictly convex and smooth (infinitely differentiable) (Nielsen and Nock 2011).

1.4 Carrier measure

The carrier measure is a scalar-valued function of only the outcome $\mathbf{x}$. In the measure-theoretic presentation of exponential families (Shao 2003, p. 66), the carrier measure truly is a *measure* on the support. The measure-theoretic intuition is analogous to Shannon's description of continuous entropy (§**??**): the carrier measure is an assumed standard that weights each small volume of the domain by $\exp\{(h(\mathbf{x}))\}$. It represents prior knowledge about the parametrization of the support.

Many formulae are simplified when the carrier measure is zero, in which case it is called a *standard carrier measure*. For continuous distributions, this can always be achieved by a change of variables; for discrete distributions, the carrier measure is rarely a standard carrier measure, and nothing can be done to make it so.

1.5 Maximum entropy formulation

The exponential families are motivated in such situations: Suppose that we make independent realizations of a random variable $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, but we only know (1) the expected sufficient statistic $\mathbb{E}(T(\mathbf{x}))$ for some known function T, (2) the support of the realizations Ω, and (3) optionally some prior carrier measure h on the space. Then, Jaynes (1957) avers the *principle of maximum entropy*: "the maximum-entropy distribution may be asserted for the positive reason that it is uniquely determined as the one which is maximally noncommittal with regard to missing information, instead of the negative one that there was no reason to think otherwise. Thus the concept of entropy supplies the missing criterion of choice…" Gokhale (1975) showed that these constraints uniquely lead to a maximum-entropy exponential family with the given sufficient statistic and support.

For example, given a mean μ, and a variance σ^2, the support of the reals, and assuming standard carrier measure, one is spared the maximum-entropy calculation and can arrive directly at the exponential family distribution function:

$$f(x) \propto \exp\left(\begin{bmatrix} x \\ (x-\mu)^2 \end{bmatrix}^{\mathsf{T}} \boldsymbol{\eta}\right) \tag{5}$$

Normalizing this function leads to the normal distribution's density function (§**??**). The parameters $\boldsymbol{\eta}$ are uniquely determined by the given mean and variance.

2 The natural parametrization

The *natural parametrization* of an exponential family is the vector space for combining and scaling evidence from independent sources. The natural parametrization is the one that specifies elements of the exponential family using natural parameters (§1.1).

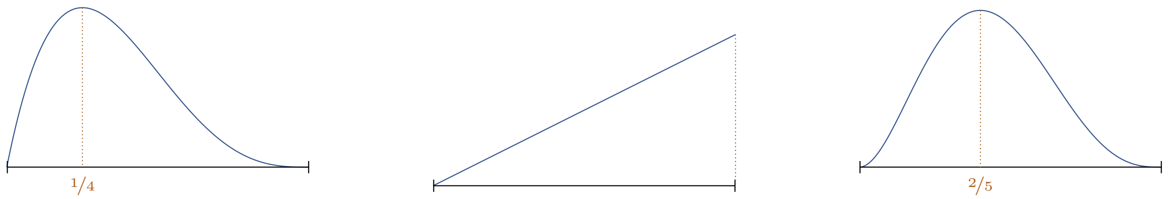
2.1 Bayesian evidence combination

For example, consider that a friend flips a coin four times and secretly records the result. His belief over the coin's bias is distributed X_1 with natural parameters $\boldsymbol{\eta}_{X_1}$. Then, you flip the coin once and record the result yielding a belief distributed X_2 with natural parameters $\boldsymbol{\eta}_{X_2}$. Given the coin, your beliefs are independent. The "Bayesian evidence combination" operation (Figure 1) aggregates such independent information by summing the natural parameters. This is because the combined belief

$$f(\mathbf{x} \mid \boldsymbol{\eta}_{X_1}, \boldsymbol{\eta}_{X_2}) \propto \exp\bigl(T(\mathbf{x})^\mathsf{T}\boldsymbol{\eta}_{X_1} - g(\boldsymbol{\eta}) + h(\mathbf{x})\bigr)\exp\bigl(T(\mathbf{x})^\mathsf{T}\boldsymbol{\eta}_{X_2} - g(\boldsymbol{\eta})\bigr) \qquad \text{(by equation 1).} \tag{6}$$

(The decomposition into a product is by independence, and we take care not to double-count the carrier measure h.)

$$= \exp\bigl(T(\mathbf{x})^\mathsf{T}(\boldsymbol{\eta}_{X_1} + \boldsymbol{\eta}_{X_2}) - g(\boldsymbol{\eta}) + h(\mathbf{x})\bigr). \tag{7}$$



(a) The belief on the bias of a coin that has been flipped four times and landed heads once.

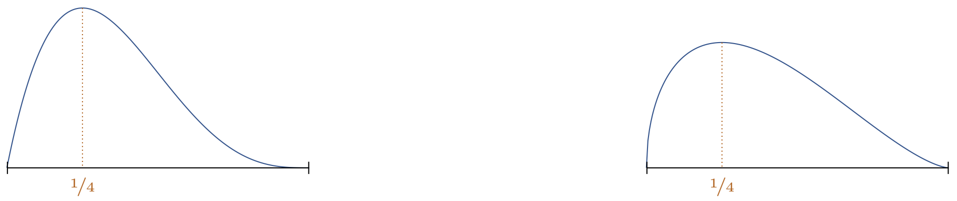
(b) The belief on the bias of a coin that has been flipped once and landed heads.

(c) The combined belief on the bias of a coin that has been flipped five times and landed heads twice.

Figure 1: Bayesian evidence combination with beta-distributed (§**??**) beliefs over the bias of a coin.

2.2 Bayesian evidence scaling

If you value the opinion from a friend more than your own, it is as if n friends provided identical, but independent information. Reasoning from §2.1, "Bayesian evidence scaling" (Figure 2) corresponds to scaling in the space of natural parameters.



(a) The belief on the bias of a coin that has been flipped four times and landed heads once.

(b) The belief on the bias after half of the information due to the coin flips in Figure 2a has been invalidated.

Figure 2: Bayesian evidence scaling with beta-distributed (§**??**) beliefs over the bias of a coin.

2.3 Bayesian evidence combination is better than product of experts

Hinton (2002) calls a similar operation—the pointwise product of probability measures—a *product of experts*. For an exponential family, this operation is equivalent to "Bayesian evidence combination" except when the carrier measure (§1.4) is nonzero. In that case, the carrier measure, which represents prior knowledge about the parametrization of the support, is double-counted. In other words, Bayesian evidence combination is invariant under reparametrization unlike *product of experts*.

3 The expectation parametrization

Suppose we have a random variable X distributed according to a distribution in family $\mathcal{F}$ (which is a natural exponential family). Then, X has an *expectation parametrization*, which is the one whose parameters are the expected sufficient statistic

$$\boldsymbol{\chi} \triangleq \mathbb{E}(T(X)). \tag{8}$$

This parametrization is convenient for *parametric density estimation*: the problem of estimating a distribution's parameters given its realizations.

Like the natural parametrization (equation 2), the expectation parametrization is unique up to a bijective linear map. Unlike the natural parametrization, the expectation parametrization does not have meaningful vector space operations (constant scaling and summation); instead, only weighted averages are meaningful.

3.1 Conjugate prior distribution

If the random variable X is distributed according to an exponential family distribution with unknown natural parameters $\boldsymbol{\eta}$, then independent realizations $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ of X induce a likelihood over $\boldsymbol{\eta}$: $\mathcal{L}(\boldsymbol{\eta} \mid \mathbf{x}_1, \ldots, \mathbf{x}_n)$. This distribution must belong to a family $\mathcal{F}'$ (called the conjugate prior of $\mathcal{F}$) that is also an exponential family since

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\eta} \mid \mathbf{x}_1, \ldots, \mathbf{x}_n) &= f(\mathbf{x}_1, \ldots, \mathbf{x}_n \mid \boldsymbol{\eta}) && (9)\\
&\propto \prod_i f(\mathbf{x}_i \mid \boldsymbol{\eta}) \qquad \text{(by independence)} && (10)\\
&= \textstyle\prod_i \exp\Big(T\left(\mathbf{x}_i\right)^\mathsf{T} \boldsymbol{\eta} - g(\boldsymbol{\eta}) + h(\mathbf{x}_i)\Big) && (11)
\end{aligned}$$

(Since $\mathbf{x}_i$ are fixed, $\prod_i \exp\{(h(\mathbf{x}_i))\}$ is constant)

$$\begin{aligned}
&\propto \textstyle\prod_i \exp\Big(T\left(\mathbf{x}_i\right)^\mathsf{T} \boldsymbol{\eta} - g(\boldsymbol{\eta})\Big) && (12)\\
&= \exp\Big(\big(\textstyle\sum_i T\left(\mathbf{x}_i\right)\big)^\mathsf{T} \boldsymbol{\eta} - n g(\boldsymbol{\eta})\Big) && (13)\\
&= \exp\big(T'(\boldsymbol{\eta})^\mathsf{T} \boldsymbol{\eta}'\big) && (14)
\end{aligned}$$

where

$$T'(\boldsymbol{\eta}) = \begin{bmatrix} \boldsymbol{\eta} \\ -g(\boldsymbol{\eta}) \end{bmatrix} \qquad\qquad \boldsymbol{\eta}' = \begin{bmatrix} \sum_i T(\mathbf{x}_i) \\ n \end{bmatrix}. \tag{15}$$

In equation 14, we can see that the vector of *hyperparameters* $\boldsymbol{\eta}'$ are natural parameters of the distribution. Thus, "Bayesian evidence combination" and "Bayesian evidence scaling" correspond to vector addition and scaling of these induced parameters, one of which n is a real-valued pseudo-observation count.

3.2 Maximum likelihood distribution

Continuing the parametric density estimation problem from §3.1, suppose we have an induced likelihood over $\boldsymbol{\eta}$:

$$\mathcal{L}(\boldsymbol{\eta} \mid \mathbf{x}_1, \ldots, \mathbf{x}_n) \propto \exp\big(T'(\boldsymbol{\eta})^\mathsf{T} \boldsymbol{\eta}'\big). \tag{14 revisited}$$

Then, the maximum likelihood distribution of X given the realizations $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is the mode of equation 14. So,

$$
\begin{aligned}
0 &= \frac{\partial \exp\bigl(T'(\boldsymbol{\eta})^\intercal \boldsymbol{\eta}'\bigr)}{\partial \boldsymbol{\eta}} && (16)\\
&= \exp\bigl(T'(\boldsymbol{\eta})^\intercal \boldsymbol{\eta}'\bigr) \frac{\partial\left(\begin{bmatrix} \eta_1 \\ \vdots \\ \eta_k \\ -g(\boldsymbol{\eta}) \end{bmatrix}^\intercal \begin{bmatrix} \sum_i T(\mathbf{x}_i)_1 \\ \vdots \\ \sum_i T(\mathbf{x}_i)_k \\ n \end{bmatrix}\right)}{\partial \boldsymbol{\eta}} && (17)\\
&\Downarrow \\
\frac{\partial g(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} &= \frac{\sum_i T(\mathbf{x}_i)}{n} && (18)\\
&\Downarrow \\
\boldsymbol{\chi} &= \frac{\sum_i T(\mathbf{x}_i)}{n} \qquad \text{(by equation 29).} && (19)
\end{aligned}
$$

Thus, the maximum likelihood distribution has expectation parameters equal to the expected sufficient statistics of the samples. This motivates the expectation parametrization, and the term *sufficient statistic*.

3.3 Aggregating maximum likelihood distributions

Suppose that instead of n independent realizations $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ of our exponential family distribution X (as in §3.1), we collect realizations m times. The ith collection yields n_i realizations from which we calculate a maximum likelihood distribution X_i having expectation parameters $\boldsymbol{\chi}_i$ (as per §3.2). After the collection, we discard the realization, so that all we have are the n_is and X_is. How can we combine these into one maximum likelihood distribution over all $\sum_{i=1}^m n_i$ realizations had been collected.

From equations 15 and 19, we can conclude that the natural parameters of the conjugate prior distribution for each i is

$$\boldsymbol{\eta}'_i = \begin{bmatrix} n_i \boldsymbol{\chi}_i \\ n_i \end{bmatrix}. \tag{20}$$

From §2.1, we know that we can combine these into one conjugate prior distribution with parameters:

$$\boldsymbol{\eta}_i = \begin{bmatrix} \sum_{i=1}^m n_i \boldsymbol{\chi}_i \\ \sum_{i=1}^m n_i \end{bmatrix}. \tag{21}$$

From equation 19, we can conclude the maximum likelihood distribution given all of the realizations has expectation parameters:

$$\frac{\sum_{i=1}^m n_i \boldsymbol{\chi}_i}{\sum_{i=1}^m n_i}. \tag{22}$$

Therefore, weighted average in the space of expectation parameters represents combining maximum likelihood distributions as if the realizations they were based on had been aggregated.

3.4 Duality of parametrizations

If the random variable X has known natural parameters $\boldsymbol{\eta}$, then Nielsen and Nock (2011) show that the function that converts natural parameters to expectation parameters is the gradient of the log-normalizer:

$$
\begin{aligned}
\frac{\partial g(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} &= \frac{\partial \log \int_{\mathbf{x}} \exp\bigl(T(\mathbf{x})^{\mathsf{T}}\boldsymbol{\eta} + h(\mathbf{x})\bigr)\ \mathrm{d}\mathbf{x}}{\partial \boldsymbol{\eta}} && \text{(by equation 4)} && (23)\\
&= \frac{\int_{\mathbf{x}} T(\mathbf{x}) \exp\bigl(T(\mathbf{x})^{\mathsf{T}}\boldsymbol{\eta} + h(\mathbf{x})\bigr)\ \mathrm{d}\mathbf{x}}{\int_{\mathbf{x}} \exp\bigl(T(\mathbf{x})^{\mathsf{T}}\boldsymbol{\eta} + h(\mathbf{x})\bigr)\ \mathrm{d}\mathbf{x}} && && (24)\\
&= \frac{\int_{\mathbf{x}} T(\mathbf{x}) \exp\bigl(T(\mathbf{x})^{\mathsf{T}}\boldsymbol{\eta} + h(\mathbf{x})\bigr)\ \mathrm{d}\mathbf{x}}{\exp\bigl(g(\boldsymbol{\eta})\bigr)} && \text{(by equation 4)} && (25)\\
&= \int_{\mathbf{x}} T(\mathbf{x}) \exp\bigl(T(\mathbf{x})^{\mathsf{T}}\boldsymbol{\eta} - g(\boldsymbol{\eta}) + h(\mathbf{x})\bigr)\ \mathrm{d}\mathbf{x} && && (26)\\
&= \int_{\mathbf{x}} T(\mathbf{x}) f(\mathbf{x} \mid \boldsymbol{\eta})\ \mathrm{d}\mathbf{x} && \text{(by equation 1)} && (27)\\
&= \mathbb{E}(T(X)) && && (28)\\
&= \boldsymbol{\chi} && \text{(by equation 8).} && (29)
\end{aligned}
$$

3.5 Higher moments of the sufficient statistic

The higher moments of the sufficient statistic are the higher-order gradients of the log-normalizer:

$$
\begin{aligned}
\boldsymbol{\nabla}^{n}_{\boldsymbol{\eta}} g(\boldsymbol{\eta}) &= \boldsymbol{\nabla}^{n-1}_{\boldsymbol{\eta}} \int_{\mathbf{x}} T(\mathbf{x}) \exp\bigl(T(\mathbf{x})^{\mathsf{T}}\boldsymbol{\eta} - g(\boldsymbol{\eta}) + h(\mathbf{x})\bigr)\ \mathrm{d}\mathbf{x} && \text{(by equation 26)} && (30)\\
&= \int_{\mathbf{x}} T(\mathbf{x}) \otimes \left[\otimes_{i=1}^{n-1}(T(\mathbf{x}) - \frac{\partial g(\boldsymbol{\eta}))}{\partial \boldsymbol{\eta}}\right] \exp\bigl(T(\mathbf{x})^{\mathsf{T}}\boldsymbol{\eta} - g(\boldsymbol{\eta}) + h(\mathbf{x})\bigr)\ \mathrm{d}\mathbf{x} && && (31)\\
&= \int_{\mathbf{x}} T(\mathbf{x}) \otimes \left[\otimes_{i=1}^{n-1}(T(\mathbf{x}) - \frac{\partial g(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}})\right] f(\mathbf{x} \mid \boldsymbol{\eta})\ \mathrm{d}\mathbf{x} && \text{(by equation 1)} && (32)\\
&= \mathbb{E}\bigl(T(X) \otimes \left[\otimes_{i=1}^{n-1}(T(X) - \frac{\partial g(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}})\right]\bigr) && && (33)\\
&= \mathbb{E}\bigl(T(X) \otimes \bigl[\otimes_{i=1}^{n-1}(T(X) - \mathbb{E}(T(X)))\bigr]\bigr) && \text{(by equation 28).} && (34)
\end{aligned}
$$

So, for example, the covariance matrix of the sufficient statistic is the Hessian of the log-normalizer:

$$
\begin{aligned}
\frac{\partial^2 g(\boldsymbol{\eta})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}} &= \mathbb{E}\bigl(T(X) \otimes (T(X) - \mathbb{E}(T(X)))\bigr) && (35)\\
&= \operatorname{Var}(T(X)). && (36)
\end{aligned}
$$

In particular, as described by Efron (1978),

$$
\begin{aligned}
\frac{\partial \boldsymbol{\chi}}{\partial \boldsymbol{\eta}} &= \frac{\partial^2 g(\boldsymbol{\eta})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}} && \text{(by equation 29)} && (37)\\
&= \operatorname{Var}(T(X)) && \text{(by equation 36).} && (38)
\end{aligned}
$$

4 Statistics of exponential families

4.1 Information theoretic statistics

Consider a data-generating distribution X, and an approximating distribution Y in the same exponential family, having natural parameters $\boldsymbol{\eta}_X$ and $\boldsymbol{\eta}_Y$, and expectation parameters $\boldsymbol{\chi}_X$ and $\boldsymbol{\chi}_Y$. Their cross entropy

(§**??**) is

$$\begin{aligned}
\mathcal{H}^{\times}(X;Y) &= -\int_{\mathbf{x}} f_X(\mathbf{x}) \log f_Y(\mathbf{x}) \, \mathrm{d}\mathbf{x} && \text{(by equation ??)} && (39)\\
&= -\int_{\mathbf{x}} f_X(\mathbf{x}) \big(T(\mathbf{x})^{\mathsf{T}} \boldsymbol{\eta}_Y - g(\boldsymbol{\eta}_Y) + h(\mathbf{x}) \big) \, \mathrm{d}\mathbf{x} && \text{(by equation 1)} && (40)\\
&= -\boldsymbol{\chi}_X^{\mathsf{T}} \boldsymbol{\eta}_Y + g(\boldsymbol{\eta}_Y) - \mathop{\mathbb{E}}_{\mathbf{x}\sim X}(h(\mathbf{x})) && \text{(by equation 29).} && (41)
\end{aligned}$$

The entropy (§**??**) of X is

$$\begin{aligned}
\mathcal{H}(X) &= \mathcal{H}^{\times}(X;X) && \text{(by equation ??)} && (42)\\
&= -\boldsymbol{\chi}_X^{\mathsf{T}} \boldsymbol{\eta}_X + g(\boldsymbol{\eta}_X) - \mathop{\mathbb{E}}_{\mathbf{x}\sim X}(h(\mathbf{x})) && && (43)
\end{aligned}$$

and their relative entropy (§**??**) is

$$\begin{aligned}
\mathcal{H}^{\mathrm{KL}}(X;Y) &= \mathcal{H}^{\times}(X;Y) - \mathcal{H}(X) && \text{(by equation ??)} && (44)\\
&= g(\boldsymbol{\eta}_Y) - g(\boldsymbol{\eta}_X) - (\boldsymbol{\eta}_Y - \boldsymbol{\eta}_X)^{\mathsf{T}} \boldsymbol{\chi}_X . && && (45)
\end{aligned}$$

So, for exponential families, the information theoretic statistics are easily calculated from the natural and expectation parameters (Figure 3).

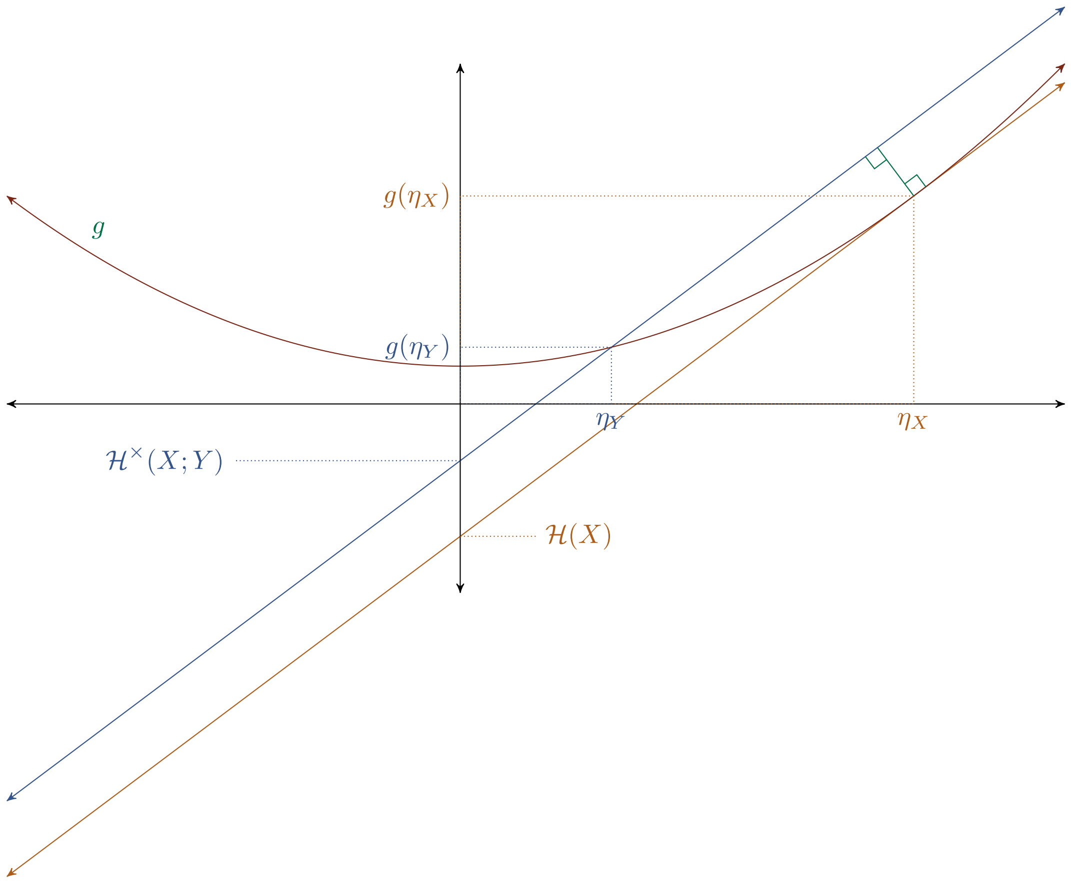


Figure 3: Graphical illustration of the entropy, cross entropy, and relative entropy of exponential families with standard carrier measure adapted from Nielsen and Nock (2011).

4.2 Parameter estimation statistics

The statistical score (§**??**) of $\boldsymbol{\eta}_Y$ given data $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is

$$\begin{aligned}
\mathcal{V}(\boldsymbol{\eta}_Y \mid \mathbf{x}_1, \ldots, \mathbf{x}_n) &= \frac{\partial \log \mathcal{L}(\boldsymbol{\eta}_Y \mid \mathbf{x}_1, \ldots, \mathbf{x}_n)}{\partial \boldsymbol{\eta}_Y} && \text{(46)} \\
&= \frac{\partial \sum_{i=1}^{n} \log \mathcal{L}(\boldsymbol{\eta}_Y \mid \mathbf{x}_i)}{\partial \boldsymbol{\eta}_Y} && \text{(47)} \\
&= \sum_{i=1}^{n} \frac{\partial\big(T(\mathbf{x}_i)^{\mathsf{T}} \boldsymbol{\eta}_Y - g(\boldsymbol{\eta}_Y) + h(\mathbf{x}_i)\big)}{\partial \boldsymbol{\eta}_Y} && \text{(by equation 1)} \quad \text{(48)} \\
&= \sum_{i=1}^{n} \left(T(\mathbf{x}_i) - \frac{\partial g(\boldsymbol{\eta}_Y)}{\partial \boldsymbol{\eta}_Y} \right) && \text{(49)} \\
&= \sum_{i=1}^{n} \left(T(\mathbf{x}_i) - \boldsymbol{\chi}_Y \right) && \text{(by equation 29).} \quad \text{(50)}
\end{aligned}$$

Therefore, the expected value of the score is

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{x}\sim X} \left(\mathcal{V}(\boldsymbol{\eta}_Y \mid \mathbf{x})\right) &= \boldsymbol{\chi}_X - \boldsymbol{\chi}_Y && \text{(by equation 8)} \quad \text{(51)} \\
&= -\frac{\partial \mathcal{H}^{\times}(X; Y)}{\partial \boldsymbol{\eta}_Y} && \text{(by equation ??).} \quad \text{(52)}
\end{aligned}$$

The Fisher information (§**??**) is

$$\begin{aligned}
\mathcal{I}(\boldsymbol{\eta}_Y) &= -\mathop{\mathbb{E}}_{\mathbf{x}\sim X} \left(\frac{\partial^2 \log f(\mathbf{x} \mid \boldsymbol{\eta}_Y)}{\partial \boldsymbol{\eta}_Y \partial \boldsymbol{\eta}_Y} \right) && \text{(53)} \\
&= -\mathop{\mathbb{E}}_{\mathbf{x}\sim X} \left(\frac{\partial^2 \big(T(\mathbf{x})^{\mathsf{T}} \boldsymbol{\eta}_Y - g(\boldsymbol{\eta}_Y) + h(\mathbf{x})\big)}{\partial \boldsymbol{\eta}_Y \partial \boldsymbol{\eta}_Y} \right) && \text{(by equation 1)} \quad \text{(54)} \\
&= \mathop{\mathbb{E}}_{\mathbf{x}\sim X} \left(\frac{\partial^2 g(\boldsymbol{\eta}_Y)}{\partial \boldsymbol{\eta}_Y \partial \boldsymbol{\eta}_Y} \right) && \text{(55)} \\
&= \frac{\partial^2 g(\boldsymbol{\eta}_Y)}{\partial \boldsymbol{\eta}_Y \partial \boldsymbol{\eta}_Y}. && \text{(56)}
\end{aligned}$$

The Jeffreys prior (§**??**) for a natural exponential family is thus

$$f(\boldsymbol{\eta}) \propto \sqrt{\det \mathcal{I}(\boldsymbol{\eta})} = \sqrt{\det \frac{\partial^2 g(\boldsymbol{\eta})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}}}. \tag{57}$$

5 Altering exponential families

§**??** defines *generalized linear models*, which are a kind of *regression* (§**??**) that makes an exponential family distributional assumption about the targets and uses cross entropy loss. This section explores what happens in the case of three alterations of the assumed exponential family.

Learning in generalized linear models depends only on the gradient of the cross entropy, which is the difference of the expectation parameters of the target values and those of predictions (by equation 51). If we find that the expectation parameters are affected by an alteration, then learning is affected, which means that the model is different. Otherwise, the alteration has no effect on the model.

5.1 Transformation of an exponential family

The beta distribution (§**??**) is the *conjugate prior* (§3.1) of a Bernoulli distribution (§**??**) parametrized by a probability $p \in [0, 1]$. If instead we had parametrized the Bernoulli using *odds* $o = \frac{p}{p+1}$, the conjugate prior is *beta-prime*. Therefore, for any beta-distributed X, there is a beta-prime-distributed $\frac{X}{X+1}$. Is regression with a beta distributional assumption the same as regression with a beta-prime assumption?

In general, suppose that we have an exponential family $\mathcal{F}$ with sufficient statistics $T_{\mathcal{F}}$ and carrier measure $h_{\mathcal{F}}$ over support $\mathcal{S}$. For any distribution $D \in \mathcal{F}$, let $X \sim D$ be a random variable with density f_X and distribution function F_X.

Let $a : \mathcal{S} \to \mathcal{T}$ be a smooth, invertible function that is independent of the parameters of X and let $Y = a(X)$ be a random variable with density f_Y and distribution function F_Y. The distribution function of Y is

$$F_Y(y) = F_X(a^{-1}(y)) \tag{58}$$

$$\Downarrow$$

$$f_Y(y) \triangleq \frac{\mathrm{d}F_Y(y)}{\mathrm{d}y} = \frac{\mathrm{d}F_X(a^{-1}(y))}{\mathrm{d}y} \tag{59}$$

$$= f_X(a^{-1}(y))\frac{\mathrm{d}a^{-1}(y)}{\mathrm{d}y}. \tag{60}$$

Therefore, Y's distribution belongs to an exponential family $\mathcal{G}$ with sufficient statistics

$$T_{\mathcal{G}}(y) = T_{\mathcal{F}}(a^{-1}(y)), \tag{61}$$

carrier measure

$$h_{\mathcal{G}}(y) = h_{\mathcal{F}}(y) + \log\left(\frac{\mathrm{d}a^{-1}(y)}{\mathrm{d}y}\right), \tag{62}$$

support $\mathcal{T}$, and the same log-normalizer as $\mathcal{F}$.

The expectation parameters are unchanged since

$$\mathbb{E}(T_{\mathcal{G}}(Y)) = \mathbb{E}\left(T_{\mathcal{F}}(a^{-1}(Y))\right) \qquad \text{(by equation 61)} \tag{63}$$

$$= \mathbb{E}\left(T_{\mathcal{F}}(X)\right). \tag{64}$$

This shows that changing the distributional assumption of a generalized linear model from $\mathcal{F}$ to $\mathcal{G}$ by smoothly transforming its values has no effect on the model.

5.2 Truncation of an exponential family

Linear regression is equivalent to an assumption of normality. However, if the target values are known to be from a subset of the reals, then is the corresponding *truncated normality* assumption equivalent to the original model?

As in the previous section, suppose that we have an exponential family $\mathcal{F}$ with log-normalizer $g_{\mathcal{F}}$ over a support $\mathcal{S}$. For any distribution $D \in \mathcal{F}$, let $X \sim D$ be a random variable with density f_X and distribution function F_X.

Define another exponential family $\mathcal{G}$ with the same sufficient statistics T and carrier measure h as $\mathcal{F}$, but over support $\mathcal{T} \subseteq \mathcal{S}$. Let Y be a random variable corresponding to X such that they have the same natural parameters $\boldsymbol{\eta}$. Let its density be f_Y and its distribution function be F_Y. We have:

$$f_Y(y) = \frac{f_X(y)}{\mathcal{P}(X \in \mathcal{T})}. \tag{65}$$

The divisor $\mathcal{P}(X \in \mathcal{T})$ depends on the parameters $\boldsymbol{\eta}$, which means that $\mathcal{G}$ has a different log-normalizer than $\mathcal{F}$:

$$g_{\mathcal{G}}(\boldsymbol{\eta}) = g_{\mathcal{F}}(\boldsymbol{\eta}) + \log\left(\mathcal{P}(X \in \mathcal{T})\right). \tag{66}$$

X and Y having different log-normalizers means that their expectation parameters $\boldsymbol{\chi}_X$ and $\boldsymbol{\chi}_Y$ are different even though their natural parameters are the same:

$$\boldsymbol{\chi}_Y \triangleq \mathbb{E}(T(Y)) \qquad \text{(by equation 8)} \tag{67}$$

$$= \int_{\mathcal{T}} f_Y(y)T(y)\,\mathrm{d}y\,. \tag{68}$$

This shows that clipping the distributional assumption of a generalized linear model changes the model.

5.3 Altering the carrier measure

Truncation (§5.2) of the sample space of an exponential family is equivalent to setting the carrier measure h to $-\infty$ over the truncated region. In the previous section, this would mean that we could have left $\mathcal{G}$'s support as $\mathcal{S}$, but set its carrier measure to

$$h_{\mathcal{G}}(x) = \begin{cases} h_{\mathcal{F}}(x) & \text{if } x \in \mathcal{T} \\ -\infty & \text{otherwise.} \end{cases} \tag{69}$$

If one desires a softer version of truncation, then one can softly decrease the carrier measure. This affects the expectation parameters—which are the expected value of the sufficient statistics—because it focuses that expectation where the carrier measure is larger. Therefore, altering the carrier measure changes the model.

References

[1] B. Efron, "The Geometry of Exponential Families," *The Annals of Statistics*, vol. 6, no. 2, pp. 362–376, 1978.

[2] D. B. Gokhale, "Maximum Entropy Characterizations of Some Distributions," in *A Modern Course on Statistical Distributions in Scientific Work: Models and structures*, ser. NATO Advanced Study Institutes series: Mathematical and physical sciences, G. P. Patil, S. Kotz, and J. K. Ord, Eds., Volume 3, Dordecht, Netherlands: D. Reidel Publishing Company, 1975, pp. 299–304.

[3] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, Aug. 2002.

[4] E. T. Jaynes, "Information Theory and Statistical Mechanics," *Phys. Rev.*, vol. 106, no. 4, pp. 620–630, May 1957.

[5] F. Nielsen and V. Garcia, "Statistical exponential families: A digest with flash cards," *CoRR*, vol. abs/0911.4, 2011.

[6] F. Nielsen and R. Nock, "Entropies and Cross-entropies of Exponential Families," *Proceedings of the International Conference on Image Processing, ICIP 2010, September 26-29, Hong Kong, China*, pp. 3621–3624, 2011.

[7] J. Shao, *Mathematical Statistics*, ser. Springer Texts in Statistics. Springer, 2003.