The Exponential Family

David M. Blei
Columbia University

October 27, 2014

Definition

¶ A probability density in the exponential family has this form

$$p(x \mid \eta) = h(x) \exp\{\eta^\top t(x) - a(\eta)\}, \tag{1}$$

where

- η is the *natural parameter*
- $t(x)$ are *sufficient statistics*
- $h(x)$ is the "underlying measure", ensures x is in the right space
- $a(\eta)$ is the log normalizer

Examples of exponential family distributions include Gaussian, gamma, Poisson, Bernoulli, multinomial, Markov models.

Examples of distributions that are not in this family include student-t, mixtures, and hidden Markov models. (We are considering these families as distributions of data. The latent variables are implicitly marginalized out.)

¶ The statistic $t(x)$ is called *sufficient* because the probability of x under η only depends on x through $t(x)$.

¶ The exponential family has fundamental connections to the world of graphical models. For our purposes, we'll use exponential families as components in directed graphical models, e.g., in the mixtures of Gaussians.

¶ The log normalizer ensures that the density integrates to 1,

$$a(\eta) = \log \int h(x) \exp\{\eta^\top t(x)\} dx \tag{2}$$

This is the negative logarithm of the normalizing constant.

¶ **Example: Bernoulli.** As an example, let's put the Bernoulli (in its usual form) into its exponential family form. The Bernoulli you are used to seeing is

$$p(x \mid \pi) = \pi^x (1-\pi)^{1-x} \quad x \in \{0, 1\} \tag{3}$$

In exponential family form

$$\begin{aligned} p(x \mid \pi) &= \exp\left\{\log\left(\pi^x (1-\pi)^{1-x}\right)\right\} && (4)\\ &= \exp\{x \log \pi + (1-x)\log(1-\pi)\} && (5)\\ &= \exp\{x \log \pi - x \log(1-\pi) + \log(1-\pi)\} && (6)\\ &= \exp\{x \log(\pi/(1-\pi)) + \log(1-\pi)\} && (7) \end{aligned}$$

This reveals the exponential family where

$$\begin{aligned} \eta &= \log(\pi/(1-\pi)) && (8)\\ t(x) &= x && (9)\\ a(\eta) &= -\log(1-\pi) = \log(1+e^\eta) && (10)\\ h(x) &= 1 && (11) \end{aligned}$$

Note the relationship between π and η is invertible

$$\pi = 1/(1+e^{-\eta}) \tag{12}$$

This is the *logistic function*.

¶ **Example: Gaussian.** The familiar form of the univariate Gaussian is

$$p(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \tag{13}$$

We put it in exponential family form by expanding the square

$$p(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}} \exp\left\{\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{1}{2\sigma^2}\mu^2 - \log\sigma\right\} \tag{14}$$

We see that

$$\begin{aligned} \eta &= \langle \mu/\sigma^2, -1/2\sigma^2 \rangle && (15)\\ t(x) &= \langle x, x^2 \rangle && (16)\\ a(\eta) &= \mu^2/2\sigma^2 + \log\sigma && (17)\\ &= -\eta_1^2/4\eta_2 - (1/2)\log(-2\eta_2) && (18)\\ h(x) &= 1/\sqrt{2\pi} && (19) \end{aligned}$$

Moments of an exponential family

¶ Let's go back to the general family. We are going to take derivatives of the log normalizer. This gives us moments of the sufficient statistics,

$$\begin{aligned}
\nabla_\eta a(\eta) &= \nabla_\eta \{\log \int \exp\{\eta^\top t(x)\} h(x) dx\} && (20)\\
&= \frac{\nabla_\eta \int \exp\{\eta^\top t(x)\} h(x) dx}{\int \exp\{\eta^\top t(x)\} h(x) dx} && (21)\\
&= \int t(x) \frac{\exp\{\eta^\top t(x)\} h(x)}{\int \exp\{\eta^\top t(x)\} h(x) dx} dx && (22)\\
&= \mathrm{E}_\eta[t(X)] && (23)
\end{aligned}$$

Check for yourself: Higher order derivatives give higher order moments.

¶ There is a 1-1 relationship between the mean of the sufficient statistics $\mathbb{E}[t(X)]$ and natural parameter η. In other words, the mean is an alternative parameterization of the distribution. (Many of the forms of exponential families that you know are the mean parameterization.)

¶ Consider again the Bernoulli. We saw this with the logistic function, where note that $\pi = \mathbb{E}[X]$ (because X is an indicator).

¶ Now consider the Gaussian.

The derivative with respect to η_1 is

$$\begin{aligned}
\frac{da(\eta)}{d\eta_1} &= -\frac{\eta_1}{2\eta_2} && (24)\\
&= \mu && (25)\\
&= \mathbb{E}[X] && (26)
\end{aligned}$$

The derivative with respect to η_2 is

$$\begin{aligned}
\frac{da(\eta)}{d\eta_2} &= \frac{\eta_1^2}{4\eta_2^2} - \frac{1}{2\eta_2} && (27)\\
&= \sigma^2 + \mu^2 && (28)\\
&= \mathbb{E}\left[X^2\right] && (29)
\end{aligned}$$

This means that the variance is

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 && (30)\\
&= -\frac{1}{2\eta_2} && (31)\\
&= \sigma^2 && (32)
\end{aligned}$$

¶ To see the $1-1$ relationship between $\mathbb{E}[t(X)]$ and η, note that

- $\text{Var}(t(X)) = \nabla^2 a_\eta$ is positive.
- $\rightarrow a(\eta)$ is convex.
- $\rightarrow$ 1-1 relationship between its argument and first derivative

Here is some notation for later (when we discuss generalized linear models). Denote the mean parameter as $\mu = \mathbb{E}[t(X)]$; denote the inverse map as $\psi(\mu)$, which gives the η such that $\mathbb{E}[t(X)] = \mu$.

Maximum likelihood estimation of an exponential family.

¶ The data are $x_{1:n}$. We seek the value of η that maximizes the likelihood.

¶ The log likelihood is

$$\begin{aligned} \mathcal{L} &= \sum_{n=1}^{N} \log p(x_n \mid \eta) && (33) \\ &= \sum_{n=1}^{N} (\log h(x_n) + \eta^\top t(x_n) - a(\eta)) && (34) \\ &= \textstyle\sum_{n=1}^{N} \log h(x_n) + \eta^\top \sum_{n=1}^{N} t(x_n) - N \cdot a(\eta) && (35) \end{aligned}$$

As a function of η, the log likelihood only depends on $\sum_{n=1}^{N} t(x_n)$. Note that it has fixed dimension; there is no need to store the data. (And note that it is *sufficient* for estimating η.)

¶ Take the gradient of the likelihood and set it to zero,

$$\nabla_\eta \mathcal{L} = \sum_{n=1}^{N} t(x_n) - N \nabla_\eta a(\eta). \tag{36}$$

It's now easy to solve for the mean parameter:

$$\mu_{\text{ML}} = \frac{\sum_{n=1}^{N} t(x_n)}{N}. \tag{37}$$

This is, as you might guess, the empirical mean of the sufficient statistics. The inverse map gives us the natural parameter,

$$\eta_{\text{ML}} = \psi(\mu_{ML}). \tag{38}$$

¶ Consider the Bernoulli. The MLE of the mean parameter μ_{ML} is the sample mean; the MLE of the natural parameter is the corresponding log odds.

¶ Consider the Gaussian. The MLE of the mean parameters are the sample mean and the sample variance.

Conjugacy

¶ Consider the following set up:

$$\eta \sim F(\cdot \mid \lambda) \tag{39}$$
$$x_i \sim G(\cdot \mid \eta) \quad \text{for } i \in \{1, \ldots, n\}. \tag{40}$$

This is a classical Bayesian data analysis setting. And, this is used as a component in more complicated models, e.g., in hierarchical models.

¶ The posterior distribution of η given the data $x_{1:n}$ is

$$p(\eta \mid x_{1:n}, \lambda) \propto F(\eta \mid \lambda) \prod_{i=1}^{n} G(x_i \mid \eta). \tag{41}$$

Suppose this distribution is in the same family as F, i.e., its parameters are in the space indexed by λ. Then F and G are a **conjugate pair**.

¶ For example,

- Gaussian likelihood with fixed variance; Gaussian prior on the mean
- Multinomial likelihood; Dirichlet prior on the probabilities
- Bernoulli likelihood; beta prior on the bias
- Poisson likelihood; gamma prior on the rate

In all these settings, the conditional distribution of the parameter given the data is in the same family as the prior.

¶ Suppose the data come from an exponential family. Every exponential family has a conjugate prior,

$$p(x_i \mid \eta) = h_\ell(x) \exp\{\eta^\top t(x_i) - a_\ell(\eta)\} \tag{42}$$
$$p(\eta \mid \lambda) = h_c(\eta) \exp\{\lambda_1^\top \eta + \lambda_2(-a_\ell(\eta)) - a_c(\lambda)\}. \tag{43}$$

The natural parameter $\lambda = \langle \lambda_1, \lambda_2 \rangle$ has dimension $\dim(\eta) + 1$. The sufficient statistics are $\langle \eta, -a(\eta) \rangle$.

The other terms $h_c(\cdot)$ and $a_c(\cdot)$ depend on the form of the exponential family. For example, when η are multinomial parameters then the other terms help define a Dirichlet.

¶ Let's compute the posterior in the general case,

$$\begin{aligned} p(\eta \mid x_{1:n}, \lambda) &\propto p(\eta \mid \lambda) \prod_{i=1}^{n} p(x_i \mid \eta) && (44)\\ &= h(\eta) \exp\{\lambda_1^\top \eta + \lambda_2(-a(\eta)) - a_c(\lambda)\} && (45)\\ &\quad \cdot \left(\textstyle\prod_{i=1}^n h(x_i)\right) \exp\{\eta^\top \textstyle\sum_{i=1}^n t(x_i) - n a_\ell(\eta)\} && (46)\\ &\propto h(\eta) \exp\{(\lambda_1 + \textstyle\sum_{i=1}^n t(x_i))^\top \eta + (\lambda_2 + n)(-a(\eta))\} && (47) \end{aligned}$$

This is the same exponential family as the prior, with parameters

$$\begin{aligned} \hat{\lambda}_1 &= \lambda_1 + \sum_{i=1}^{n} t(x_i) && (48)\\ \hat{\lambda}_2 &= \lambda_2 + n. && (49) \end{aligned}$$

Posterior predictive distribution

¶ Let's compute the posterior predictive distribution. This comes up frequently in applications of models (i.e., when doing prediction) and in inference (i.e., when doing collapsed Gibbs sampling).

¶ Consider a new data point x_{new}. Its posterior predictive is

$$\begin{aligned} p(x_{\text{new}} \mid x_{1:n}) &= \int p(x_{\text{new}} \mid \eta) p(\eta \mid x_{1:n}) d\eta \\ &\propto \int \exp\{\eta^\top x_{\text{new}} - a(\eta)\} \exp\{\hat{\lambda}_1^\top \eta + \hat{\lambda}_2(-a(\eta)) - a(\hat{\lambda})\} d\eta \\ &= \frac{\int \exp\{(\hat{\lambda}_1 + x_{\text{new}})^\top \eta + (\hat{\lambda}_2 + 1)(-a(\eta)) d\eta}{\exp\{a(\hat{\lambda}_1, \hat{\lambda}_2)\}} \\ &= \exp\{a(\hat{\lambda}_1 + x_{\text{new}}, \hat{\lambda}_2 + 1) - a(\hat{\lambda}_1, \hat{\lambda}_2)\} \end{aligned}$$

¶ In other words, the posterior predictive density is a ratio of normalizing constants. In the numerator is the posterior with the new data point added; in the denominator is the posterior without the new data point.

¶ This is how we can derive collapsed Gibbs samplers.

Canonical prior

¶ There is a variation on the form of the simple conjugate prior that is useful for understanding its properties. We set $\lambda_1 = x_0 n_0$ and $\lambda_2 = n_0$. (And, for convenience, we drop the sufficient statistic so that $t(x) = x$.)

¶ In this form the conjugate prior is

$$p(\eta \mid x_0, n_0) \propto \exp\{n_0 x_0^\top \eta - n_0 a(\eta)\}. \tag{50}$$

Here we can interpret x_0 as our prior idea of the expectation of x, and n_0 as the number of "prior data points".

Consider $\mathbb{E}[\mathbb{E}[X \mid \eta]]$, where the first expectation is with respect to the prior and the second is respect to the data distribution. As an exercise, show that

$$\mathbb{E}[\mathbb{E}[X \mid \eta]] = x_0. \tag{51}$$

¶ Assume data $x_{1:n}$. The mean is

$$\bar{x} = (1/n) \sum_{i=1}^{n} x_i \tag{52}$$

Given the data and a prior, the posterior parameters are

$$\hat{x}_0 = \frac{n_0 x_0 + n\bar{x}}{n + n_0} \tag{53}$$

$$\hat{n} = n_0 + n \tag{54}$$

This is easy to see from our previous derivation of $\hat{\lambda}_1, \hat{\lambda}_2$. Further, we see that the posterior expectation of X is a convex combination of x_0 (our prior idea) and $\bar{x}$ (the data mean).

Example: Data from a unit variance Gaussian

¶ Suppose the data x_i come from a unit variance Gaussian

$$p(x \mid \mu) = \frac{1}{\sqrt{2\pi}} \exp\{-(x-\mu)^2/2\}. \tag{55}$$

This is a simpler exponential family than the previous Gaussian

$$p(x \mid \mu) = \frac{\exp\{-x^2/2\}}{\sqrt{2\pi}} \exp\{\mu x - \mu^2/2\}. \tag{56}$$

In this case

$$\begin{aligned} \eta &= \mu && (57)\\ t(x) &= x && (58)\\ h(x) &= \frac{\exp\{-x^2/2\}}{\sqrt{2\pi}} && (59)\\ a(\eta) &= \mu^2/2 = \eta^2/2. && (60) \end{aligned}$$

¶ What is the conjugate prior? It is

$$p(\eta \mid \lambda) = h(\eta) \exp\{\lambda_1 \eta + \lambda_2(-\eta^2/2) - a_c(\lambda)\} \tag{61}$$

This has sufficient statistics $\langle \eta, -\eta^2/2 \rangle$, which means it's a *Gaussian distribution*.

¶ Put it in canonical form,

$$p(\eta \mid n_0, x_0) = h(\eta) \exp\{n_0 x_0 \eta - n_0(\eta^2/2) - a_c(n_0, x_0)\}. \tag{62}$$

¶ The posterior parameters are

$$\begin{aligned} \hat{x}_0 &= \frac{n_0 x_0 + n\bar{x}}{n_0 + n} && (63)\\ \hat{n}_0 &= n + n_0. && (64) \end{aligned}$$

This implies that the posterior mean and variance are

$$\begin{aligned} \hat{\mu} &= \hat{x}_0 && (65)\\ \hat{\sigma}^2 &= 1/(n_0 + n). && (66) \end{aligned}$$

These are results that we asserted earlier.

¶ Intuitively, when we haven't seen any data then our estimate of the mean is the *prior mean*. As we see more data, our estimate of the mean moves towards the *sample mean*.

Before seeing data, our "confidence" about the estimate is the prior variance. As we see more data, the confidence decreases.

Example: Data from a Bernoulli

¶ The Bernoulli likelihood is

$$p(x \mid \pi) = \pi^x (1-\pi)^{1-x}. \tag{67}$$

Above we have seen its form as a minimal exponential family. Let's rewrite that, but keeping the mean parameter in the picture,

$$p(x \mid \pi) = \exp\{x \log(\pi/(1-\pi)) - \log(1-\pi)\} \tag{68}$$

¶ The canonical conjugate prior therefore looks like this,

$$p(\pi \mid x_0, n_0) = \exp\{n_0 x_0 \log(\pi/(1-\pi)) + n_0 \log(1-\pi) - a(n_0, x_0)\} \tag{69}$$

This simplifies to

$$p(\pi \mid x_0, n_0) = \exp\{n_0 x_0 \log(\pi) + n_0(1-x_0) \log(1-\pi)) - a(n_0, x_0)\} \tag{70}$$

Putting this in non-exponential family form,

$$p(\pi \mid x_0, n_0) \propto \pi^{n_0 x_0} (1-\pi)^{n_0(1-x_0)} \tag{71}$$

which *nearly* looks like the familiar Beta distribution.

¶ To get the beta, we set $\alpha \triangleq n_0 x_0 + 1$ and $\beta \triangleq n_0(1-x_0) + 1$. Bringing in the resulting normalizer, we have

$$p(\pi \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi^{\alpha-1} (1-\pi)^{\beta-1}. \tag{72}$$

¶ The posterior distribution follows the usual recipe, from which we can derive that the corresponding updates are

$$\hat{\alpha} = \alpha + \sum_{i=1}^{n} x_i \tag{73}$$

$$\hat{\beta} = \beta + \sum_{i=1}^{n} (1-x_i) \tag{74}$$

To see this, update $\hat{x}_0$ and $\hat{n}_0$ as above and compute $\hat{\alpha}$ and $\hat{\beta}$ from the definitions.

The big picture

¶ Exponential families and conjugate priors can be used in many graphical models. Gibbs sampling is straightforward when each complete conditional involves a conjugate "prior" and likelihood pair.

¶ For example, we can now define an exponential family mixture model. The mixture components are drawn from a conjugate prior; the data are drawn from the corresponding exponential family.

¶ Imagine a generic model $p(x_{1:n}, z_{1:n} \mid \alpha)$. Suppose each complete conditional is in the exponential family,

$$p(z_i \mid z_{-i}, x, \alpha) = h(z_i) \exp\{\eta_i(z_{-i}, x)^\top z_i - a(\eta_i(z_{-i,x}))\}. \tag{75}$$

Notice that the natural parameter is a function of the variables we condition on. This is called a *conditionally conjugate model*. Provided we can compute the natural parameter, Gibbs sampling is immediate.

¶ Many models from the machine learning research literature are conditionally conjugate. Examples include HMMs, mixture models, hierarchical linear regression, probit regression, factorial models, and others.

¶ Consider the HMM, for example, but with its parameters unknown,

[graphical model]

Notice that this is like a mixture model, but where the mixture components are embedded in a Markov chain.

Each row of the transition matrix is a point on the $(K-1)$-simplex; each set of emission probabilities is a point on the $(V-1)$-simplex. We can place Dirichlet priors on these components. (We will talk about the Dirichlet later. It is a multivariate generalization of the beta distribution, and is the conjugate prior to the categorical/multinomial distribution.)

We can easily obtain the complete conditional of z_i. The complete conditionals of the transitions & emissions follow from our discussion of conjugacy.

Collapsed Gibbs sampling is possible, but is quite expensive. One needs to run forward-backward at each sampling of a z_i.