Chapter 9
Generalized linear models and the exponential family

9.1 The exponential family

Before defining the exponential family, we mention several reasons why it is important:

- It can be shown that, under certain regularity conditions, the exponential family is the only family of distributions with finite-sized sufficient statistics, meaning that we can compress the data into a fixed-sized summary without loss of information. This is particularly useful for online learning, as we will see later.
- The exponential family is the only family of distributions for which conjugate priors exist, which simplifies the computation of the posterior (see Section 9.1.5).
- The exponential family can be shown to be the family of distributions that makes the least set of assumptions subject to some user-chosen constraints (see Section 9.1.6).
- The exponential family is at the core of generalized linear models, as discussed in Section 9.2.
- The exponential family is at the core of variational inference, as discussed in Section TODO.

9.1.1 Definition

A pdf or pmf $p(\boldsymbol{x}|\boldsymbol{\theta})$,for $\boldsymbol{x} \in \mathbb{R}^m$ and $\boldsymbol{\theta} \in \mathbb{R}^D$, is said to be in the **exponential family** if it is of the form

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} h(\boldsymbol{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\boldsymbol{x})] \tag{9.1}$$

$$= h(\boldsymbol{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\boldsymbol{x}) - A(\boldsymbol{\theta})] \tag{9.2}$$

where

$$Z(\boldsymbol{\theta}) = \int h(\boldsymbol{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\boldsymbol{x})] \mathrm{d}\boldsymbol{x} \tag{9.3}$$

$$A(\boldsymbol{\theta}) = \log Z(\boldsymbol{\theta}) \tag{9.4}$$

Here $\boldsymbol{\theta}$ are called the **natural parameters** or **canonical parameters**, $\boldsymbol{\phi}(\boldsymbol{x}) \in \mathbb{R}^D$ is called a vector of **sufficient statistics**, $Z(\boldsymbol{\theta})$ is called the **partition function**, $A(\boldsymbol{\theta})$ is called the **log partition function** or **cumulant function**, and $h(\boldsymbol{x})$ is the a scaling constant, often 1. If $\boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{x}$, we say it is a **natural exponential family**.

Equation 9.2 can be generalized by writing

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = h(\boldsymbol{x}) \exp[\boldsymbol{\eta}(\boldsymbol{\theta})^T \boldsymbol{\phi}(\boldsymbol{x}) - A(\boldsymbol{\eta}(\boldsymbol{\theta}))] \tag{9.5}$$

where $\boldsymbol{\eta}$ is a function that maps the parameters $\boldsymbol{\theta}$ to the canonical parameters $\boldsymbol{\eta} = \boldsymbol{\eta}(\boldsymbol{\theta})$.If $\dim(\boldsymbol{\theta}) < \dim(\boldsymbol{\eta}(\boldsymbol{\theta}))$, it is called a **curved exponential family**, which means we have more sufficient statistics than parameters. If $\boldsymbol{\eta}(\boldsymbol{\theta}) = \boldsymbol{\theta}$, the model is said to be in **canonical form**. We will assume models are in canonical form unless we state otherwise.

9.1.2 Examples

9.1.2.1 Bernoulli

The Bernoulli for $x \in \{0,1\}$ can be written in exponential family form as follows:

$$\begin{aligned} \mathrm{Ber}(x|\mu) &= \mu^x (1-\mu)^{1-x} \\ &= \exp[x \log \mu + (1-x) \log(1-\mu)] \end{aligned} \tag{9.6}$$

where $\boldsymbol{\phi}(x) = (\mathbb{I}(x=0), \mathbb{I}(x=1))$ and $\boldsymbol{\theta} = (\log \mu, \log(1-\mu))$.

However, this representation is **over-complete** since $\mathbf{1}^T \boldsymbol{\phi}(x) = \mathbb{I}(x=0) + \mathbb{I}(x=1) = 1$. Consequently $\boldsymbol{\theta}$ is not uniquely identifiable. It is common to require that the representation be **minimal**, which means there is a unique $\boldsymbol{\theta}$ associated with the distribution. In this case, we can just define

$$\mathrm{Ber}(x|\mu) = (1-\mu) \exp\left(x \log \frac{\mu}{1-\mu}\right) \tag{9.7}$$

$$\text{where } \phi(x) = x, \theta = \log \frac{\mu}{1-\mu}, Z = \frac{1}{1-\mu}$$

We can recover the mean parameter μ from the canonical parameter using

$$\mu = \mathrm{sigm}(\theta) = \frac{1}{1+e^{-\theta}} \tag{9.8}$$

9.1.2.2 Multinoulli

We can represent the multinoulli as a minimal exponential family as follows:

$$\begin{aligned}
\mathrm{Cat}(\boldsymbol{x}|\boldsymbol{\mu}) &= \prod_{k=1}^{K} = \exp\left(\sum_{k=1}^{K} x_k \log \mu_k\right) \\
&= \exp\left[\sum_{k=1}^{K-1} x_k \log \mu_k + (1 - \sum_{k=1}^{K-1} x_k) \log(1 - \sum_{k=1}^{K-1} \mu_k)\right] \\
&= \exp\left[\sum_{k=1}^{K-1} x_k \log \frac{\mu_k}{1 - \sum_{k=1}^{K-1} \mu_k} + \log(1 - \sum_{k=1}^{K-1} \mu_k)\right] \\
&= \exp\left[\sum_{k=1}^{K-1} x_k \log \frac{\mu_k}{\mu_K} + \log \mu_K\right], \text{ where } \mu_K \triangleq 1 - \sum_{k=1}^{K-1} \mu_k
\end{aligned}$$

We can write this in exponential family form as follows:

$$\mathrm{Cat}(\boldsymbol{x}|\boldsymbol{\mu}) = \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\boldsymbol{x}) - A(\boldsymbol{\theta})] \tag{9.9}$$

$$\boldsymbol{\theta} \triangleq (\log \frac{\mu_1}{\mu_K}, \cdots, \log \frac{\mu_{K-1}}{\mu_K}) \tag{9.10}$$

$$\boldsymbol{\phi}(\boldsymbol{x}) \triangleq (x_1, \cdots, x_{K-1}) \tag{9.11}$$

We can recover the mean parameters from the canonical parameters using

$$\mu_k = \frac{e^{\theta_k}}{1 + \sum_{j=1}^{K-1} e^{\theta_j}} \tag{9.12}$$

$$\mu_K = 1 - \frac{\sum_{j=1}^{K-1} e^{\theta_j}}{1 + \sum_{j=1}^{K-1} e^{\theta_j}} = \frac{1}{1 + \sum_{j=1}^{K-1} e^{\theta_j}} \tag{9.13}$$

and hence

$$A(\boldsymbol{\theta}] = -\log \mu_K = \log(1 + \sum_{j=1}^{K-1} e^{\theta_j}) \tag{9.14}$$

9.1.2.3 Univariate Gaussian

The univariate Gaussian can be written in exponential family form as follows:

$$\begin{aligned}
\mathcal{N}(x|\mu, \sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right] \\
&= \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}\mu^2\right] \\
&= \frac{1}{Z(\boldsymbol{\theta})} \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(x)]
\end{aligned} \tag{9.15}$$

where

$$\boldsymbol{\theta} = (\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}) \tag{9.16}$$

$$\boldsymbol{\phi}(x) = (x, x^2) \tag{9.17}$$

$$Z(\boldsymbol{\theta}) = \sqrt{2\pi}\sigma \exp(\frac{\mu^2}{2\sigma^2}) \tag{9.18}$$

9.1.2.4 Non-examples

Not all distributions of interest belong to the exponential family. For example, the uniform distribution,$X \sim U(a,b)$, does not, since the support of the distribution depends on the parameters. Also, the Student T distribution (Section TODO) does not belong, since it does not have the required form.

9.1.3 Log partition function

An important property of the exponential family is that derivatives of the log partition function can be used to generate **cumulants** of the sufficient statistics.[20] For this reason, $A(\boldsymbol{\theta})$ is sometimes called a **cumulant function**. We will prove this for a 1-parameter distribution; this can be generalized to a K-parameter distribution in a straightforward way. For the first derivative we have

For the second derivative we have

$$\begin{aligned}
\frac{\mathrm{d}A}{\mathrm{d}\theta} &= \frac{\mathrm{d}}{\mathrm{d}\theta}\left\{\log \int \exp[\theta\phi(x)]\, h(x)\mathrm{d}x\right\} \\
&= \frac{\frac{\mathrm{d}}{\mathrm{d}\theta}\int \exp[\theta\phi(x)]\, h(x)\mathrm{d}x}{\int \exp[\theta\phi(x)]\, h(x)\mathrm{d}x} \\
&= \frac{\int \phi(x) exp[\theta\phi(x)]\, h(x)\mathrm{d}x}{\exp(A(\theta))} \\
&= \int \phi(x) \exp[\theta\phi(x) - A(\theta)]\, h(x)\mathrm{d}x \\
&= \int \phi(x) p(x)\mathrm{d}x = \mathbb{E}[\phi(x)]
\end{aligned} \tag{9.19}$$

For the second derivative we have

$$\begin{aligned}
\frac{\mathrm{d}^2 A}{\mathrm{d}\theta^2} &= \int \phi(x) \exp[\theta\phi(x) - A(\theta)]\, h(x) \left[\phi(x) - A'(\theta)\right] \mathrm{d}x \\
&= \int \phi(x) p(x) \left[\phi(x) - A'(\theta)\right] \mathrm{d}x \\
&= \int \phi^2(x) p(x)\mathrm{d}x - A'(\theta) \int \phi(x) p(x)\mathrm{d}x \\
&= \mathbb{E}[\phi^2(x)] - \mathbb{E}[\phi(x)]^2 = \mathrm{var}[\phi(x)]
\end{aligned} \tag{9.20}$$

In the multivariate case, we have that

[20] The first and second cumulants of a distribution are its mean $\mathbb{E}[X]$ and variance $\mathrm{var}[X]$, whereas the first and second moments are its mean $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$.

$$\frac{\partial^2 A}{\partial \theta_i \partial \theta_j} = \mathbb{E}[\phi_i(x)\phi_j(x)] - \mathbb{E}[\phi_i(x)]\mathbb{E}[\phi_j(x)] \quad (9.21)$$

and hence

$$\nabla^2 A(\boldsymbol{\theta}) = \text{cov}[\boldsymbol{\phi}(\boldsymbol{x})] \quad (9.22)$$

Since the covariance is positive definite, we see that $A(\boldsymbol{\theta})$ is a convex function (see Section A.1).

9.1.4 MLE for the exponential family

The likelihood of an exponential family model has the form

$$p(\mathcal{D}|\boldsymbol{\theta}) = \left[\prod_{i=1}^{N} h(\boldsymbol{x}_i)\right] g(\boldsymbol{\theta})^N \exp\left[\boldsymbol{\theta}^T \left(\sum_{i=1}^{N} \boldsymbol{\phi}(\boldsymbol{x}_i)\right)\right] \quad (9.23)$$

We see that the sufficient statistics are N and

$$\boldsymbol{\phi}(\mathcal{D}) = \sum_{i=1}^{N} \boldsymbol{\phi}(\boldsymbol{x}_i) = (\sum_{i=1}^{N} \phi_1(\boldsymbol{x}_i), \cdots, \sum_{i=1}^{N} \phi_K(\boldsymbol{x}_i)) \quad (9.24)$$

The **Pitman-Koopman-Darmois theorem** states that, under certain regularity conditions, the exponential family is the only family of distributions with finite sufficient statistics. (Here, finite means of a size independent of the size of the data set.)

One of the conditions required in this theorem is that the support of the distribution not be dependent on the parameter.

9.1.5 Bayes for the exponential family

TODO

9.1.5.1 Likelihood

9.1.6 Maximum entropy derivation of the exponential family *

9.2 Generalized linear models (GLMs)

Linear and logistic regression are examples of **generalized linear models**, or **GLM**s (McCullagh and Nelder 1989). These are models in which the output density is in the exponential family (Section 9.1), and in which the mean parameters are a linear combination of the inputs, passed through a possibly nonlinear function, such as the logistic function. We describe GLMs in more detail below. We focus on scalar outputs for notational simplicity. (This excludes multinomial logistic regression, but this is just to simplify the presentation.)

9.2.1 Basics

9.3 Probit regression

9.4 Multi-task learning