An Introduction to Probabilistic Graphical Models

Michael I. Jordan
University of California, Berkeley

June 30, 2003

Chapter 8

The exponential family and generalized linear models

In this chapter we extend the scope of our modeling toolbox to accommodate a variety of additional data types, including counts, time intervals and rates. We introduce the exponential family of distributions, a family that includes the Gaussian, binomial, multinomial, Poisson, gamma, Rayleigh and beta distributions, as well as many others. We consider both unconditional and conditional models involving this family.

Much of our discussion is focused on the conditional setting, in which we have a directed model, $X \to Y$, with X and Y observed, and with Y having an exponential family distribution for each value of X. To parameterize this conditional distribution we introduce a class of models known as *generalized linear models (GLIM's)*. GLIM's are a general category of models that include linear regression and linear classification models as special cases. As in those models, GLIM's retain an important role for linearity, while introducing appropriate nonlinearities so as to cope with the idiosyncracies of the particular exponential family distribution at hand. GLIM's have the dual virtue of systematizing the work that we have done thus far and showing how to extend that work to handle a wide range of additional data types.

At first blush this chapter may appear to involve a large dose of mathematical detail, but appearances shouldn't deceive—most of the detail involves working out examples that show how the exponential family and GLIM's relate to more familiar material. The real message of this chapter is the simplicity and elegance of exponential family and GLIM methods. Once the new ideas are mastered, it is often easier to work within the general exponential family and GLIM frameworks than with specific instances.

8.1 The exponential family

A probability density in the exponential family takes the following general form:

$$p(x \mid \eta) = h(x) \exp\{\eta^T T(x) - A(\eta)\} \tag{8.1}$$

for a parameter vector η, often referred to as the *natural parameter*, and for given functions T, a, and h. The function $T(X)$ is referred to as a *sufficient statistic*; the reasons for this nomenclature are discussed below. The form of the function $h(x)$ is not of fundamental importance; it simply reflects the underlying measure with respect to which $p(x \mid \eta)$ is a density. Of rather more importance is the function $A(\eta)$. Integrating Eq. (8.1) with respect to x, we have:

$$A(\eta) = \log \int h(x) \exp\{\eta^T T(x)\} dx \tag{8.2}$$

where we see that $A(\eta)$ can be viewed as the logarithm of a normalization factor. The set of η for which this integral is finite is referred to as the *natural parameter space*.

It is also common to write the exponential family distribution in the following way:

$$p(x \mid \eta) = \frac{1}{Z(\eta)} h(x) \exp\{\eta^T T(x)\}, \tag{8.3}$$

which is equivalent to if we let $A(\eta) = \log Z(\eta)$. Although we focus on Eq. (8.1) throughout this chapter, we will also make use of Eq. (8.72) in later chapters.

8.1.1 Examples

The Bernoulli distribution

The probability mass function of a Bernoulli random variable X is given as follows:

$$\begin{aligned} p(x \mid \pi) &= \pi^x (1-\pi)^{1-x} && (8.4) \\ &= \exp\left\{ \log\left(\frac{\pi}{1-\pi}\right) x + \log(1-\pi) \right\}. && (8.5) \end{aligned}$$

where our trick, here and throughout the chapter, is to take the exponential of the logarithm of the original distribution. Thus we see that the Bernoulli distribution is an exponential family distribution with:

$$\begin{aligned} \eta &= \frac{\pi}{1-\pi} && (8.6) \\ T(x) &= x && (8.7) \\ A(\eta) &= -\log(1-\pi) = \log(1+e^\eta) && (8.8) \\ h(x) &= 1. && (8.9) \end{aligned}$$

Note moreover that the relationship between η and π is invertible. Solving Eq. (8.6) for π, we have:

$$\pi = \frac{1}{1+e^{-\eta}}, \tag{8.10}$$

which is the logistic function.

The Poisson distribution

The probability mass function of a Poisson random variable is given as follows:

$$p(x \mid \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}. \tag{8.11}$$

Rewriting this expression we obtain:

$$p(x \mid \lambda) = \frac{1}{x!} \exp\{x \log \lambda - \lambda\}. \tag{8.12}$$

Thus the Poisson distribution is an exponential family distribution, with:

$$\eta = \log \lambda \tag{8.13}$$
$$T(x) = x \tag{8.14}$$
$$A(\eta) = \lambda = e^{\eta} \tag{8.15}$$
$$h(x) = \frac{1}{x!}. \tag{8.16}$$

Moreover, we can obviously invert the relationship between η and λ:

$$\lambda = e^{\eta}. \tag{8.17}$$

The Gaussian distribution

The (univariate) Gaussian distribution can be written as follows:

$$p(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{8.18}$$
$$= \frac{1}{\sqrt{2\pi}} \exp\left\{\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{1}{2\sigma^2}\mu^2 - \ln\sigma\right\}. \tag{8.19}$$

This is in the exponential family form, with:

$$\eta = \begin{bmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{bmatrix} \tag{8.20}$$
$$T(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix} \tag{8.21}$$
$$A(\eta) = \frac{\mu^2}{2\sigma^2} + \ln\sigma = -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\ln(-2\eta_2) \tag{8.22}$$
$$h(x) = \frac{1}{\sqrt{2\pi}}. \tag{8.23}$$

Note in particular that the univariate Gaussian distribution is a two-parameter distribution and that its sufficient statistic is a vector.

The multivariate Gaussian distribution can also be written in the exponential family form; we leave the details to Exercise **??** and Chapter 13.

The multinomial distribution

As a final example, let us consider the multinomial distribution. Let $X = (X_1, X_2, \ldots, X_m)$ be a collection of integer-valued random variables representing event counts, where X_i represents the count of the number of times the ith event occurs in a set of n independent trials. Let π_i represent the probability of the ith event occuring in any given trial. We have:

$$p(x \mid \pi) = \frac{n!}{x_1! x_2! \cdots x_m!} \pi_1^{x_1} \pi_2^{x_2} \cdots \pi_m^{x_m}, \tag{8.24}$$

as the probability mass function for such a collection.

Following the strategy of our previous examples, we rewrite the multinomial distribution as follows:

$$p(x \mid \pi) = \exp\left\{ \sum_{i=1}^{m} x_i \ln \pi_i \right\}. \tag{8.25}$$

While this shows that the multinomial distribution is in the exponential family, there are some troubling aspects to this expression. In particular it appears that the $A(\eta)$ term is equal to zero. As we will be seeing (in Section 8.1.2), one of the principal virtues of the exponential family form is that moments can be calculated by taking derivatives of $A(\eta)$; thus, the disappearance of this term is unsettling.

Our problem is caused by the fact that the parameters satisfy a linear constraint, namely: $\sum_{i=1}^{m} \pi_i = 1$. Let us define an exponential family to be of *full rank* if the parameters satisfy no such constraint—technically we assume that an m-dimensional parameter space contains an open rectangle of dimension m. (In the case of the multinomial the parameters lie on an $m-1$ dimensional simplex, and thus the parameter space does not contain a rectangle of dimension m). To achieve a full rank representation for the multinomial, we parameterize the distribution using the first $m-1$ components of π:

$$\begin{aligned} p(x \mid \pi) &= \exp\left\{ \sum_{i=1}^{m} x_i \ln \pi_i \right\} && (8.26) \\ &= \exp\left\{ \sum_{i=1}^{m-1} x_i \ln \pi_i + \left(1 - \sum_{i=1}^{m-1} x_i\right) \ln\left(1 - \sum_{i=1}^{m-1} \pi_i\right) \right\} && (8.27) \\ &= \exp\left\{ \sum_{i=1}^{m-1} \ln\left(\frac{\pi_i}{1 - \sum_{i=1}^{m-1} \pi_i} \right) x_i + \ln\left(1 - \sum_{i=1}^{m-1} \pi_i\right) \right\}. && (8.28) \end{aligned}$$

where we have used the fact that $\pi_m = 1 - \sum_{i=1}^{m-1} \pi_i$.

From this representation we obtain:

$$\eta_i = \ln\left(\frac{\pi_i}{1 - \sum_{i=1}^{m-1} \pi_i} \right) = \ln\left(\frac{\pi_i}{\pi_m} \right) \tag{8.29}$$

for $i = 1, \ldots, m-1$. For convenience we also can define η_m; Eq. (8.29) implies that if we do so we must take $\eta_m = 0$.

As in the other examples of exponential family distributions, we can invert Eq. (8.29) to obtain a mapping that expresses π_i in terms of η_i. Taking the exponential of Eq. (8.29) and summing we obtain:

$$\pi_i = \frac{e^{\eta_i}}{\sum_{j=1}^{m} e^{\eta_j}}, \tag{8.30}$$

which is the softmax function.

Finally, from Eq. (8.28) we obtain:

$$A(\eta) = -\ln\left(1 - \sum_{i=1}^{m-1} \pi_i\right) = \ln\left(\sum_{i=1}^{m} e^{\eta_i}\right) \tag{8.31}$$

as the log normalization factor for the multinomial.

8.1.2 Moments

An appealing feature of the exponential family representation is that we can obtain moments of the distribution by taking derivatives of the log normalization function $A(\eta)$. Before establishing this fact, let us consider an example.

Recall that in the case of the Bernoulli distribution we have $A(\eta) = \log(1 + e^{\eta})$. Taking a first derivative yields:

$$\begin{aligned} \frac{dA}{d\eta} &= \frac{e^{\eta}}{1+e^{\eta}} && (8.32) \\ &= \frac{1}{1+e^{-\eta}} && (8.33) \\ &= \mu, && (8.34) \end{aligned}$$

which is the mean of a Bernoulli variable.

Taking a second derivative yields:

$$\begin{aligned} \frac{d^2 a}{d\eta^2} &= \frac{d\mu}{d\eta} && (8.35) \\ &= \mu(1-\mu), && (8.36) \end{aligned}$$

which is the variance of a Bernoulli variable.

We now show that in general the first derivative of $A(\eta)$ is equal to the mean of $T(X)$. We treat the case of scalar η for simplicity; the (straightforward) extension to vector η is considered in Exercise **??**. Calculating the first derivative of $A(\eta)$ yields:

$$\begin{aligned} \frac{dA}{d\eta} &= \frac{d}{d\eta}\left\{\log \int \exp\{\eta T(x)\} h(x) dx\right\} && (8.37) \\ &= \frac{\int T(x)\exp\{\eta T(x)\} h(x) dx}{\int \exp\{\eta T(x)\} h(x) dx} && (8.38) \\ &= \int T(x)\exp\{\eta^T T(x) - A(\eta)\} h(x) dx && (8.39) \\ &= ET(X). && (8.40) \end{aligned}$$

Thus we see that the first derivative of $A(\eta)$ is equal to the mean of the sufficient statistic.

Let us now take a second derivative:

$$\begin{aligned}
\frac{d^2a}{d\eta^2} &= \int T(x)\exp\{\eta T(x) - A(\eta)\}(T(x) - a'(\eta))h(x)dx && (8.41)\\
&= \int T(x)\exp\{\eta T(x) - A(\eta)\}(T(x) - ET(X))h(x)dx && (8.42)\\
&= \int T^2(x)\exp\{\eta T(x) - A(\eta)\}h(x)dx - ET(X)\int T(x)\exp\{\eta T(x) - A(\eta)\}h(x)dx\\
&= ET^2(x) - (ET(X))^2 && (8.43)\\
&= \mathrm{Var}[T(x)], && (8.44)
\end{aligned}$$

and thus we see that the second derivative of $A(\eta)$ is equal to the variance of the sufficient statistic.

Example

Let us calculate the moments of the univariate Gaussian distribution. Recall the form taken by $A(\eta)$:

$$A(\eta) = -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\ln(-2\eta_2), \tag{8.45}$$

where $\eta_1 = \mu/\sigma^2$ and $\eta_2 = -1/2\sigma^2$.

Taking the derivative with respect to η_1 yields:

$$\begin{aligned}
\frac{\partial A}{\partial \eta_1} &= \frac{\eta_1}{2\eta_2} && (8.46)\\
&= \frac{\mu/\sigma^2}{1/\sigma^2} && (8.47)\\
&= \mu, && (8.48)
\end{aligned}$$

which is the mean of X, the first component of the sufficient statistic.

Taking a second derivative with respect to η_1 yields:

$$\begin{aligned}
\frac{\partial^2 A}{\partial \eta_1^2} &= -\frac{1}{2\eta_2} && (8.49)\\
&= \sigma^2, && (8.50)
\end{aligned}$$

which is the variance of X.

Given that X^2 is the second component of the sufficient statistic, we can also compute the variance by calculating the partial of a with respect to η_2. Moreover, we can calculate third moments by computing the mixed partial, and fourth moments by taking the second partial with respect to η_2 (see Exercise **??**).

8.1.3 The moment parameterization

In the previous section we have seen that it is possible to obtain the mean, $\mu \triangleq ET(X)$, as a function of the canonical parameter η:

$$\mu = \frac{dA}{d\eta}. \tag{8.51}$$

It turns out that this relationship is invertible.

To see this, note from Eq. (8.44) that the second derivative of $A(\eta)$ is a variance and hence positive. This implies that $A(\eta)$ is a convex function. For a convex function there is necessarily a one-to-one relationship between the argument to the function and the first derivative of the function. Hence the mapping from η to μ is invertible.

We will represent the inverse mapping as $\eta = \psi(\mu)$ in the remainder of the chapter.

This argument implies that a distribution in the exponential family can be parameterized not only by η—the canonical parameterization—but also by μ—the *moment parameterization*. Many distributions are traditionally parameterized using the moment parameterization; indeed, in Section 8.1.1 our starting point was the moment parameterization for each of the examples. We subsequently reparameterized these distribution using the canonical parameterization. We also computed the mapping from η to μ in each case, recovering some familiar functions, including the logistic function and the softmax function. We will return to this topic in Section 8.2 when we discuss generalized linear models.

8.1.4 Sufficiency

In this section we discuss the important concept of *sufficiency*. Sufficiency characterizes what is essential in a data set, or, alternatively, what is inessential and can therefore be thrown away. While the notion of sufficiency is broader than the exponential family, the ties to the exponential family are close, and it is natural to introduce the concept here.

A *statistic* is a function of a random variable. In particular, let X be a random variable and let $T(X)$ be a statistic. Suppose that the distribution of X depends on a parameter θ. The intuitive notion of sufficiency is that $T(X)$ is sufficient for θ if there is no information in X regarding θ beyond that in $T(X)$. That is, having observed $T(X)$, we can throw away X for the purposes of inference with respect to θ. Let us make this notion more precise.

Sufficiency is defined in somewhat different ways in the Bayesian and frequentist frameworks. Let us begin with the Bayesian approach, which is arguably more natural. In the Bayesian approach, we treat θ as a random variable, and are therefore licensed to consider conditional independence relationships involving θ. We say that $T(X)$ is sufficient for θ if the following conditional independence statement holds:

$$\theta \perp\!\!\!\perp X \mid T(X). \tag{8.52}$$

We can also write this in terms of probability distributions:

$$p(\theta \mid T(x), x) = p(\theta \mid T(x)). \tag{8.53}$$

Thus, as shown graphically in Figure 8.1(a), sufficiency means that θ is independent of X, when

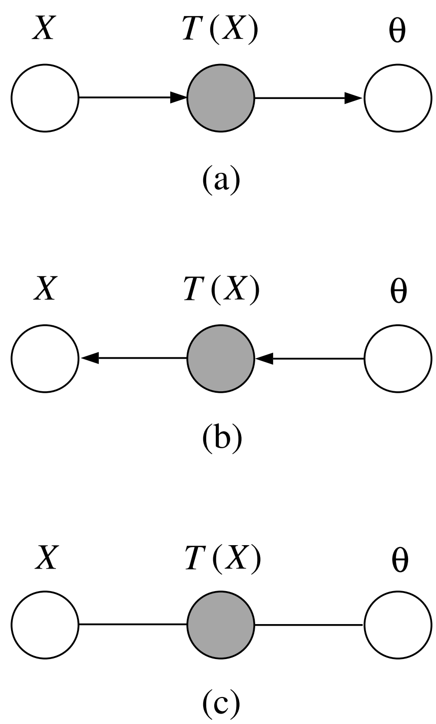

Figure 8.1: Graphical models whose conditional independence properties capture the notion of sufficiency in three equivalent ways.

we condition on $T(X)$. This captures the intuitive notion that $T(X)$ contains all of the essential information in X regarding θ.

To obtain a frequentist definition of sufficiency, let us consider the graphical model in Figure 8.1(b). This model expresses the same conditional independence semantics as Figure 8.1(a), asserting that θ is independent of X conditional on $T(X)$, but the model is parameterized in a different way. From the factorized form of the joint probability we obtain:

$$p(x \mid T(x), \theta) = p(x \mid T(x)). \tag{8.54}$$

This expression suggests a frequentist definition of sufficiency. In particular, treating θ as a label rather than a random variable, we define $T(X)$ to be sufficient for θ if the conditional distribution of X given $T(X)$ is not a function of θ.

Both the Bayesian and frequentist definitions of sufficiency imply a factorization of $p(x \mid \theta)$, and it is this factorization which is generally easiest to work with in practice. To obtain the factorization we use the undirected graphical model formalism. Note in particular that Figure 8.1(c) expresses the same conditional independence semantics as Figure 8.1(a) and Figure 8.1(b). Moreover, from Figure 8.1(c), we know that we can express the joint probability as a product of potential functions ψ_1 and ψ_2:

$$p(x, T(x), \theta) = \psi_1(T(x), \theta)\psi_2(x, T(x)), \tag{8.55}$$

where we have absorbed the constant of proportionality Z in one of the potential functions. Now $T(x)$ is a deterministic function of x, which implies that we can drop $T(x)$ on the left-hand side of the equation. Dividing by $p(\theta)$ we therefore obtain:

$$p(x \mid \theta) = g(T(x), \theta)h(x, T(x)), \tag{8.56}$$

for given functions g and h. Although we have motivated this result by using a Bayesian calculation, the result can be utilized within either the Bayesian or frequentist framework. Its equivalence to the frequentist definition of sufficiency is known as the Neyman factorization theorem.

8.1.5 Sufficiency and the exponential family

An important feature of the exponential family is that it one can obtain sufficient statistics by inspection, once the distribution is expressed in the standard form. Recall the definition:

$$p(x \mid \eta) = h(x) \exp\{\eta^T T(x) - A(\eta)\}. \tag{8.57}$$

From Eq. (8.56) we see immediately that $T(X)$ is a sufficient statistic for η.

8.1.6 IID sampling

The reduction obtainable by using a sufficient statistic is particularly notable in the case of IID sampling. Suppose that we have a collection of N independent random variables, $X = (X_1, X_2, \ldots, X_N)$, characterized by the same exponential family density. Taking the product, we obtain the joint density:

$$p(x \mid \eta) = \prod_{n=1}^{N} h(x_n) \exp\{\eta^T T(x_n) - A(\eta)\} = \left(\prod_{n=1}^{N} h(x_n)\right) \exp\left\{\eta^T \sum_{n=1}^{N} T(x_n) - NA(\eta)\right\}. \tag{8.58}$$

From this result we see that X is itself an exponential distribution, with sufficient statistic $\sum_{n=1}^{N} T(X_n)$.

For several of the examples we discussed earlier (in Section 8.1.1), including the Bernouilli, the Poisson, and the multinomial distributions, the sufficient statistic $T(X)$ is equal to the random variable X. For a set of N IID observations from such distributions, the sufficient statistic is equal to $\sum_{n=1}^{N} x_n$. Thus in this case, it suffices to maintain a single value, the sum of the observations. The individual data points can be thrown away.

For the univariate Gaussian the sufficient statistic is the pair $T(X) = (X, X^2)$, and thus for N IID Gaussians it suffices to maintain the sum $\sum_{n=1}^{N} x_n$, and the sum of squares $\sum_{n=1}^{N} x_n^2$.

8.1.7 Maximum likelihood estimates

In this section we show how to obtain maximum likelihood estimates in exponential family distributions. We obtain a generic formula which generalizes our earlier work on density estimation in Chapter 5.

Consider an IID data set, $\mathcal{D} = (x_1, x_2, \ldots, x_N)$. From Eq. (8.58) we obtain the following log likelihood:

$$l(\eta \,|\, \mathcal{D}) = \log\left(\prod_{n=1}^{N} h(x_n)\right) + \eta^T \left(\sum_{n=1}^{N} T(x_n)\right) - NA(\eta). \tag{8.59}$$

Taking the gradient with respect to η yields:

$$\nabla_\eta l = \sum_{n=1}^{N} T(x_n) - N\nabla_\eta A(\eta), \tag{8.60}$$

and setting to zero gives:

$$\nabla_\eta A(\hat{\eta}) = \frac{1}{N}\sum_{n=1}^{N} T(x_n). \tag{8.61}$$

Finally, defining $\mu \triangleq E[T(x)]$, and recalling Eq. (8.40), we obtain:

$$\hat{\mu}_{ML} = \frac{1}{N}\sum_{n=1}^{N} T(x_n) \tag{8.62}$$

as the general formula for maximum likelihood estimation in the exponential family.

It should not be surprising that our formula involves the data only via the sufficient statistic $\sum_{n=1}^{N} T(X_n)$. This gives operational meaning to sufficiency—for the purpose of estimating parameters we retain only the sufficient statistic.

For distributions in which $T(X) = X$, which include the the Bernoulli distribution, the Poisson distribution, and the the multinomial distribution, our result shows that the sample mean is the maximum likelihood estimate of the mean.

For the univariate Gaussian distribution, we see that the sample mean is the maximum likelihood estimate of the mean and the sample variance is the maximum likelihood estimate of the variance. For the multivariate Gaussian we obtain the same result, where by "variance" we mean the covariance matrix.

8.1.8 Maximum likelihood and the Kullback-Leibler divergence

In this section we point out a simple relationship between the maximum likelihood problem and the Kullback-Leibler (KL) divergence. This relationship is general; it has nothing to do specifically with the exponential family. We discuss it in the current chapter, however, because we have a hidden agenda. Our agenda, to be gradually revealed in Chapters 9, 11 and 19, involves building a number of very interesting and important relationships between the exponential family and the Kullback-Leibler (KL) divergence. By introducing a statistical interpretation of the KL divergence in the current chapter, we hope to hint subliminally at deeper connections to come.

To link the KL divergence and maximum likelihood, let us first define the *empirical distribution*, $\tilde{p}(x)$. This is a distribution which places a point mass at each data point x_n in our data set $\mathcal{D}$. We have:

$$\tilde{p}(x) \triangleq \frac{1}{N}\sum_{n=1}^{N}\delta(x, x_n), \tag{8.63}$$

where $\delta(x, x_n)$ is a Kronecker delta function in the continuous case. In the discrete case, $\delta(x, x_n)$ is simply a function that is equal to one if its arguments agree and equal to zero otherwise.

If we integrate (in the continuous case) or sum (in the discrete case) $\tilde{p}(x)$ against a function of x, we evaluate that function at each point x_n. In particular, the log likelihood can be written this way. In the discrete case we have:

$$\begin{aligned}
\sum_x \tilde{p}(x)\log p(x\,|\,\theta) &= \sum_x \frac{1}{N}\sum_{n=1}^{N}\delta(x, x_n)\log p(x\,|\,\theta) && (8.64)\\
&= \frac{1}{N}\sum_{n=1}^{N}\sum_x \delta(x, x_n)\log p(x\,|\,\theta) && (8.65)\\
&= \frac{1}{N}\sum_{n=1}^{N}\log p(x_n\,|\,\theta) && (8.66)\\
&= \frac{1}{N}l(\theta\,|\,\mathcal{D}). && (8.67)
\end{aligned}$$

Thus by computing a cross-entropy between the empirical distribution and the model, we obtain the log likelihood, scaled by the constant $1/N$. We obtain an identical result in the continuous case by integrating.

Let us now calculate the KL divergence between the empirical distribution and the model $p(x\,|\,\theta)$. We have:

$$\begin{aligned}
D(\tilde{p}(x) \parallel p(x\,|\,\theta)) &= \sum_x \tilde{p}(x)\log\frac{\tilde{p}(x)}{p(x\,|\,\theta)} && (8.68)\\
&= \sum_x \tilde{p}(x)\log\tilde{p}(x) - \sum_x \tilde{p}(x)\log p(x\,|\,\theta) && (8.69)\\
&= +\sum_x \tilde{p}(x)\log\tilde{p}(x) - \frac{1}{N}l(\theta\,|\,\mathcal{D}). && (8.70)
\end{aligned}$$

The first term, $\sum_x \tilde{p}(x)\log\tilde{p}(x)$, is independent of θ. Thus, the minimizing value of θ on the left-hand side is equal to the maximizing value of θ on the right-hand side.

In other words: *minimizing the KL divergence to the empirical distribution is equivalent to maximizing the likelihood.* This simple result will prove to be very useful in our later work.

8.1.9 Conjugacy and Bayesian estimates

[Section not yet written].

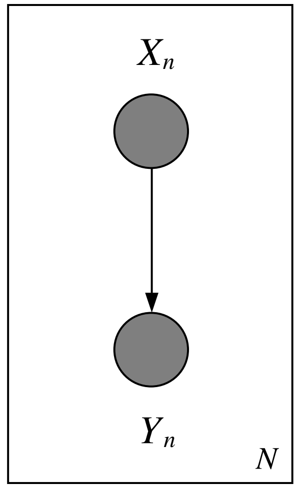

Figure 8.2: The graphical model representation of a generalized linear model.

8.2 Generalized linear models

We now turn to problems involving a pair of variables, X and Y, where both X and Y are assumed to be observed (see Figure 8.2). As in the linear regression and (discriminative) linear classification models that we discussed in Chapters 6 and 7, we focus on the conditional relationship between X and Y.

A common feature of both the linear regression and discriminative linear classification models is a particular choice of representation for the conditional expectation of Y. Letting μ denote the modeled value of the conditional expectation, we can summarize the structural component of both types of models by writing:

$$\mu = f(\theta^T x). \tag{8.71}$$

In the case of linear regression the function $f(\cdot)$ is the identity function. For the linear classification models, we studied a variety of possible choices for $f(\cdot)$, including the logistic function (for logistic regression) and the cumulative Gaussian (for probit regression).

To complete the model specification, we endow Y with a particular conditional probability distribution, having μ as a parameter. For linear regression, this distribution is Gaussian, whereas for the linear classification models, this distribution is Bernoulli (for the binary case) or multinomial (for the multiway case).

The *generalized linear model (GLIM)* framework extends these ideas beyond the Gaussian, Bernoulli and multinomial settings to the more general exponential family. A GLIM makes three assumptions regarding the form of the conditional probability distribution $p(y \mid x)$:

- The observed input x is assumed to enter into the model via a linear combination $\xi = \theta^T x$,
- The conditional mean μ is represented as a function $f(\xi)$ of the linear combination ξ, where f is known as the *response function*,

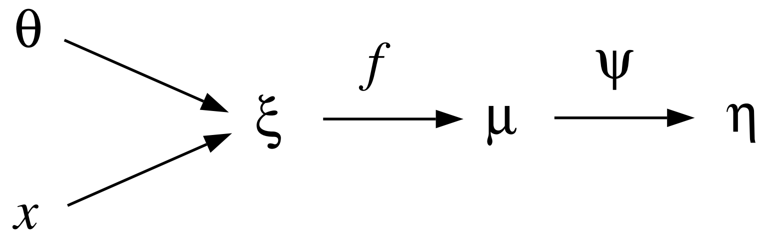

Figure 8.3: A diagram summarizing the relationships between the variables in a GLIM model.

- The observed output y is assumed to be characterized by an exponential family distribution with conditional mean μ.

These assumptions are summarized diagrammatically in Figure 8.3. Note that the diagram includes the mapping from μ to η, which we denote as $\eta = \psi(\mu)$. This mapping allows us to use the canonical parameterization to represent the exponential family distribution for Y.

Within the GLIM framework, it is convenient to work with a slight variation on the exponential family theme. In particular, in the GLIM framework we assume that the conditional distribution of Y takes the following form:

$$p(x \mid \eta, \phi) = h(x, \phi) \exp\left\{ \frac{\eta^T x - A(\eta)}{\phi} \right\}, \tag{8.72}$$

where we have augmented the representation in Eq. (8.1) to include an explicit *scale parameter* ϕ. Many exponential family distributions, including the Gaussian and the gamma, are naturally expressed in this form. In particular, we can write the Gaussian distribution as follows:

$$\begin{aligned} p(x \mid \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\} \qquad (8.73) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{ -\frac{x^2}{2\sigma^2} \right\} \right) \exp\left\{ \frac{\mu x - \mu^2/2}{\sigma^2} \right\}. \qquad (8.74) \end{aligned}$$

As we have seen, we can equivalently bundle the parameters μ and σ^2 into a single parameter vector, and express the Gaussian as a two-parameter exponential family distribution. The scale-parameter form is, however, often more natural. Note, moreover, that although the Gaussian yields a two-parameter exponential family when we bundle the canonical parameter and the scale parameter, in general we do not require that $p(x \mid \eta, \phi)$ is expressible as a two-parameter exponential family. This gives some useful flexibility.

Note also that for the purposes of GLIM modeling we drop the $T(\cdot)$ function in the exponential family representation. Thus we focus on distributions for which the observable Y is itself a sufficient statistic.

There are two principal choice points in the specification of a GLIM: (1) the choice of exponential family distribution, and (2) the choice of the response function $f(\cdot)$.

The choice of exponential family distribution is generally rather strongly constrained by the nature of the data Y. Thus, class labels are naturally represented by Bernoulli or multinomial distributions, counts by the Poisson distribution, intervals by the exponential or gamma distributions, etc.

This leaves us with the choice of the response function as the principal degree of freedom in the specification of a GLIM. There are constraints that we generally want to impose on this function, reflecting constraints on the conditional expectation. For example, in the case of the Bernoulli and multinomial distributions, the conditional expectation must lie between 0 and 1, and this suggests that we should choose a response function whose range is $(0, 1)$. Similarly, for a gamma distribution, the random variable is nonnegative, and we should presumably choose a response function whose range is $(0, \infty)$. Such constraints only give rough guidance, however, and in general for any given distribution there are many possible choices of response function. There is, however, a particular response function—the *canonical response function*—that is uniquely associated with a given exponential family distribution and has some appealing mathematical properties. In particular, if we assume that $\xi = \eta$, or equivalently that $f(\cdot) = \psi^{-1}(\cdot)$, we obtain the canonical response function. Note that the function $\psi(\cdot)$ is determined once we have chosen a particular exponential family density. Thus if we decide to use the canonical response function the choice of the exponential family density completely determines the GLIM.

We will explore some of the properties of the canonical response function in the remainder of this chapter. It should be emphasized, however, that the canonical response function is by no means the universally best choice for all problems. Indeed, as we have already seen in the case of the classification models, different choices of response function can be appropriate in different situations, reflecting different underlying assumptions about the way that the data are generated. We might view the canonical response function as a reasonable default.

The first point to note about the canonical response function is that it automatically passes a sanity check with regards to the constraints on its range. That is, the modeled values $\mu = f(\eta)$ are guaranteed to be possible values of the conditional expectation. To see this, note that:

$$f(\eta) = \psi^{-1}(\eta) = a'(\eta) = E[Y \mid \eta]. \tag{8.75}$$

Thus, for any value η such that $a'(\eta)$ exists, we see that $f(\eta)$ is equal to the conditional mean of an exponential family distribution in which η is the canonical parameter.

Some examples of canonical response functions are provided in the following table; these have been collected from the examples in Section 8.1 and from the exercises.

8.2.1 Maximum likelihood estimation

In this section we write down the likelihood for generalized linear models and present on-line and batch methods for maximizing the likelihood. We restrict ourselves to scalar Y in order to simplify the presentation; the results go through for vector Y with straightforward notational alterations (see Exercise **??**).

Consider an IID data set, $\mathcal{D} = \{(x_n, y_n); n = 1, \ldots, N\}$. Taking the logarithm of a product of N copies of the exponential family distribution in Eq. (8.72), we obtain the following log likelihood tb

Model	Canonical response function
Gaussian	$\mu = \eta$
Bernoulli	$\mu = 1/(1+e^{-\eta})$
multinomial	$\mu_i = \eta_i / \sum_j e^{\eta_j}$
Poisson	$\mu = e^{\eta}$
gamma	$\mu = -\eta^{-1}$

Figure 8.4: The canonical response functions for several exponential family distributions.

for GLIM models:

$$
\begin{aligned}
l(\theta \,|\, \mathcal{D}) &= \log\left(\prod_{n=1}^{N} h(y_n)\exp\{\eta_n y_n - A(\eta_n)\}\right) && (8.76)\\
&= \sum_{n=1}^{N}\log h(y_n) + \sum_{n=1}^{N}\left(\eta_n y_n - A(\eta_n)\right), && (8.77)
\end{aligned}
$$

where $\eta_n = \psi(\mu_n)$, $\mu_n = f(\xi_n)$ and $\xi_n = \theta^T x_n$.

In the case of the canonical response function, for which $\eta_n = \theta^T x_n$, the log likelihood simplifies:

$$
\begin{aligned}
l(\theta \,|\, \mathcal{D}) &= \sum_{n=1}^{N}\log h(y_n) + \sum_{n=1}^{N}\left(\theta^T x_n y_n - A(\eta_n)\right) && (8.78)\\
&= \sum_{n=1}^{N}\log h(y_n) + \theta^T\sum_{n=1}^{N} x_n y_n - \sum_{n=1}^{N} A(\eta_n). && (8.79)
\end{aligned}
$$

From this expression we can draw an important conclusion—*the sum* $\sum_{n=1}^{N} x_n y_n$ *is a sufficient statistic for* θ. This sum has a fixed, finite dimension (the dimension of the vector x_n), for any value of N. This has the very practical consequence that we can allocate a fixed amount of storage for collecting the information needed to estimate θ, whatever the sample size N. This is an important motivation for considering canonical response functions.

Let us now calculate the gradient of the log likelihood:

$$
\begin{aligned}
\nabla_\theta l &= \sum_{n=1}^{N}\frac{dl}{d\eta_n}\nabla_\theta \eta_n && (8.80)\\
&= \sum_{n=1}^{N}(y_n - a'(\eta_n))\nabla_\theta \eta_n && (8.81)\\
&= \sum_{n=1}^{N}(y_n - \mu_n)\frac{d\eta_n}{d\mu_n}\frac{d\mu_n}{d\xi_n}x_n. && (8.82)
\end{aligned}
$$

For the canonical response function, we have $\eta_n = \xi_n$, and thus the derivatives cancel, leaving us with the following simple expression for the log likelihood:

$$\nabla_\theta l = \sum_{n=1}^{N} (y_n - \mu_n) x_n. \tag{8.83}$$

This expression has the appealing feature that the parameter vector θ and the "error" $(y_n - \mu_n)$ are on the same scale.

An on-line algorithm

A general on-line estimation algorithm can be obtained by following the stochastic gradient of the log likelihood function. Consider first the case of the canonical response function. Given an estimate $\theta^{(t)}$ at the tth iteration of the algorithm, we obtain:

$$\theta^{(t+1)} = \theta^{(t)} + \rho(y_n - \mu_n^{(t)}) x_n, \tag{8.84}$$

where $\mu_n^{(t)} = f(\theta^{(t)T} x_n)$ and where ρ is a step size.

We have obtained an algorithm that is formally identical to the LMS algorithm. Moreover, the geometry of the LMS algorithm that we discussed in Chapter 6 carries over to this more general setting. That is, as in Chapter 6, the on-line algorithm steps in the direction of the input vector x_n, weighted by the prediction error $(y_n - \mu_n^{(t)})$. The specific GLIM model makes its appearance only through the definition of the conditional expectation μ_n.

If we do not use the canonical response function, then the gradient also includes the derivatives of $f(\cdot)$ and $\psi(\cdot)$. These can be viewed as scaling coefficients that alter the step size ρ, but otherwise leave the general LMS form intact. Thus we have obtained a result worth remembering—the LMS-like algorithm in Eq. (8.84) is the generic stochastic gradient algorithm for models throughout the GLIM family.

A batch algorithm

In our discussion of logistic regression (Section 7.3.1) we introduced the *iteratively reweighted least squares (IRLS) algorithm*—a Newton-Raphson algorithm for batch estimation of parameters. The algorithm goes through with essentially no change to the general GLIM setting. For completeness we present the algorithm and its derivation here.

We will assume the canonical response function, and indicate the changes that are needed to accommodate noncanonical response functions at the end of the section.

We begin by writing the gradient in vector notation:

$$\nabla_\theta l = \frac{1}{\phi} \sum_n (y_n - \mu_n) x_n = \frac{1}{\phi} X^T (y - \mu), \tag{8.85}$$

where X is the design matrix whose rows are the vector x_n^T, y is defined as the $N \times 1$ vector whose components are the values y_n, and similarly μ is the $N \times 1$ vector whose components are the values μ_n.

Taking a second derivative, we calculate the Hessian matrix:

$$H = -\frac{1}{\phi}\sum_n \frac{d\mu_n}{d\eta_n} x_n x_n^T \tag{8.86}$$

$$= -\frac{1}{\phi} X^T W X, \tag{8.87}$$

where we have defined the diagonal weight matrix:

$$W \triangleq \mathrm{diag}\left\{\frac{d\mu_1}{d\eta_1}, \frac{d\mu_1}{d\eta_2}, \ldots, \frac{d\mu_1}{d\eta_N}\right\}, \tag{8.88}$$

where $d\mu_n/d\eta_n$ can be computed by calculating the second derivative of $A(\eta_n)$.

Note that the weights depend on the parameter vector θ, and thus the weight matrix W depends on θ. We use the notation $W^{(t)}$ to denote the weight matrix at the tth iteration of the algorithm. Similarly, we use the notation $\mu^{(t)}$ to denote the value of μ at the tth iteration.

The Newton-Raphson algorithm is obtained by multiplying the gradient by the inverse Hessian matrix and subtracting the result from the current parameter vector:

$$\theta^{(t+1)} = \theta^{(t)} + (X^T W^{(t)} X)^{-1} X^T (y - \mu^{(t)}) \tag{8.89}$$

$$= (X^T W^{(t)} X)^{-1} \left[X^T W^{(t)} X \theta^{(t)} + X^T (y - \mu^{(t)})\right] \tag{8.90}$$

$$= (X^T W^{(t)} X)^{-1} X^T W^{(t)} z^{(t)}, \tag{8.91}$$

where we define:

$$z^{(t)} = \eta + [W^{(t)}]^{-1}(y - \mu^{(t)}). \tag{8.92}$$

The algorithm in Eq. (8.91) is the IRLS algorithm.

If we extend the derivation to handle noncanonical response functions, we find that the Hessian matrix has another term (see Exercise **??**). Including this term yields the Newton-Raphson algorithm for noncanonical response functions. There is, however, an alternative approach. If we use the *expected Hessian* in place of the Hessian in the Newton-Raphson update formula, we obtain an alternative algorithm known as the *Fisher scoring method*.[1] This algorithm is a simplification in the case of noncanonical response functions—the extra term that appears in the Hessian contains the factor $(y - \mu)$, and this term therefore vanishes when we take expectations (see Exercise **??**). Thus, the Fisher scoring method takes the form shown in Eq. (8.91) in all cases. It is generally the preferred way to implement the IRLS algorithm.

8.3 Historical remarks and bibliography

[1]Note that the expectation of the Hessian matrix is the Fisher information matrix.