This chapter deals with the class of distributions known as the exponential family, which includes the Gaussian, Bernoulli, uniform, and gamma distributions, as well as many others. Knowing the properties of this family of distributions allows us to derive theorems and algorithms that can be applied more broadly.

The Bernoulli distribution for $x \in \{0, 1\}$ is over-complete when written in exponential family form since there is a linear dependence between the features, so $\theta$ is not uniquely identifiable. Because of this it is common to require that the representation be minimal, meaning that there is a unique $\theta$ associated with the distribution.

The Pitman-Koopman-Darmois theorem states that under certain regularity conditions, the exponential family is the only family of distributions with finite sufficient statistics, where finite means of a size independent of the size of the data set. A condition required by this theorem is that the support of the distribution cannot be dependent on the parameter. This can require the use of moment matching, where when the gradient is set to zero, at the MLE, the empirical average of the sufficient statistics must equal the model's theoretical expected sufficient statistics.

Linear and logistic regression models are examples of generalized linear models, also called GLMs, which are defined as models where the output density is in the exponential family and the mean parameters are a linear combination of the inputs passed through a function which can be nonlinear (for example, the logistic function). In a GLM, there will be a dispersion parameter often set to 1, a natural parameter, a partition function, and a normalization constant. When converting from the mean parameter to the natural parameter, we can use a function that is uniquely determined by the form of the exponential family distribution. Afterwards, the mean of the distribution should be made some invertible monotonic function of the linear combination, which is known by convention as the mean function, represented by $g^{-1}()$. The inverse of the mean function is called the link function and is represented by $g()$, and pretty much any function can be chosen for the link function as long as it is invertible and $g^{-1}$ has the appropriate range. One particularly simple form of the link function that is used is $g = \psi$, which is known as the canonical link function.

One of the useful things about GLMs is that they can be fit using the same methods used for logistic regression. If it turns out that the expected Hessian is the same as the Hessian given when the canonical link is used, then using the expected Hessian instead of the actual Hessian is known as the Fisher scoring method.