2.3 Exponential Families

Distributions from the exponential family are some of the most versatile tools for statistical inference. Gaussians, Poisson, Gamma and Wishart distributions all form part of the exponential family. They play a key role in dealing with graphical models, classification, regression and conditional random fields which we will encounter in later parts of this book. Some of the reasons for their popularity are that they lead to convex optimization problems and that they allow us to describe probability distributions by linear models.

2.3.1 Basics

Densities from the exponential family are defined by

\[ p(x; \theta) := \exp \left( \langle \phi(x), \theta \rangle - g(\theta) \right). \]  

(2.39)
Here \( \phi(x) \) is a map from \( x \) to the sufficient statistics \( \phi(x) \). \( \theta \) is commonly referred to as the natural parameter. It lives in the space dual to \( \phi(x) \). Moreover, \( g(\theta) \) is a normalization constant which ensures that \( p(x) \) is properly normalized. \( g \) is often referred to as the log-partition function. The name stems from physics where \( Z = e^{\theta(\theta)} \) denotes the number of states of a physical ensemble. \( g \) can be computed as follows:

\[
g(\theta) = \log \int_X \exp (\langle \phi(x), \theta \rangle) \, dx.
\] (2.40)

**Example 2.5 (Binary Model)** Assume that \( X = \{0; 1\} \) and that \( \phi(x) = x \). In this case we have \( g(\theta) = \log [e^\theta + e^\theta] = \log [1 + e^\theta] \). It follows that \( p(x = 0; \theta) = \frac{1}{1 + e^\theta} \) and \( p(x = 1; \theta) = \frac{e^\theta}{1 + e^\theta} \). In other words, by choosing different values of \( \theta \) one can recover different Bernoulli distributions.

One of the convenient properties of exponential families is that the log-partition function \( g \) can be used to generate moments of the distribution itself simply by taking derivatives.

**Theorem 2.14 (Log partition function)** The function \( g(\theta) \) is convex. Moreover, the distribution \( p(x; \theta) \) satisfies

\[
\partial_\theta g(\theta) = E_x [\phi(x)] \quad \text{and} \quad \partial^2_\theta g(\theta) = \text{Var}_x [\phi(x)].
\] (2.41)

**Proof** Note that \( \partial^2_\theta g(\theta) = \text{Var}_x [\phi(x)] \) implies that \( g \) is convex, since the covariance matrix is positive semidefinite. To show (2.41) we expand

\[
\partial_\theta g(\theta) = \frac{\int_X \phi(x) \exp (\langle \phi(x), \theta \rangle) \, dx}{\int_X \exp (\langle \phi(x), \theta \rangle)} = \int \phi(x) p(x; \theta) \, dx = E_x [\phi(x)].
\] (2.42)

Next we take the second derivative to obtain

\[
\partial^2_\theta g(\theta) = \int_X \phi(x) [\phi(x) - \partial_\theta g(\theta)]^\top p(x; \theta) \, dx
\]

\[
= E_x [\phi(x) \phi(x)] - E_x [\phi(x)] E_x [\phi(x)]^\top
\] (2.43)

which proves the claim. For the first equality we used (2.42). For the second line we used the definition of the variance.

One may show that higher derivatives \( \partial^n g(\theta) \) generate higher order cumulants of \( \phi(x) \) under \( p(x; \theta) \). This is why \( g \) is often also referred as the cumulant-generating function. Note that in general, computation of \( g(\theta) \) is nontrivial since it involves solving a high-dimensional integral. For many
2.3 Exponential Families

In fact, the computation is NP hard, for instance when the domain of permutations \( FJ95 \). Throughout the book we will discuss a number of approximation techniques which can be applied in such a case.

Let us briefly illustrate (2.42) using the binary model of Example 2.5. We have that \( \partial \theta = \frac{e^\theta}{1 + e^\theta} \) and \( \partial^2 \theta = \frac{e^\theta}{(1 + e^\theta)^2} \). This is exactly what we would have obtained from direct computation of the mean \( p(x = 1; \theta) \) and variance \( p(x = 1; \theta) - p(x = 1; \theta)^2 \) subject to the distribution \( p(x; \theta) \).

2.3.2 Examples

A large number of densities are members of the exponential family. Note, however, that in statistics it is not common to express them in the dot product formulation for historic reasons and for reasons of notational compactness. We discuss a number of common densities below and show why they can be written in terms of an exponential family. A detailed description of the most commonly occurring types are given in a table.

**Gaussian** Let \( x, \mu \in \mathbb{R}^d \) and let \( \Sigma \in \mathbb{R}^{d \times d} \) where \( \Sigma > 0 \), that is, \( \Sigma \) is a positive definite matrix. In this case the normal distribution can be expressed via

\[
p(x) = (2\pi)^{-\frac{d}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left( -\frac{1}{2} (x - \mu)^\top \Sigma^{-1} (x - \mu) \right)
\]

\[
= \exp\left( x^\top [\Sigma^{-1} \mu] + \text{tr} \left( -\frac{1}{2} xx^\top \right) \left[ \Sigma^{-1} \right] \right) - c(\mu, \Sigma)
\]

where \( c(\mu, \Sigma) = \frac{1}{2} \mu^\top \Sigma^{-1} \mu + \frac{d}{2} \log 2\pi + \frac{1}{2} \log |\Sigma| \). By combining the terms in \( x \) into \( \phi(x) := (x, -\frac{1}{2} xx^\top) \) we obtain the sufficient statistics of \( x \). The corresponding linear coefficients \( (\Sigma^{-1} \mu, \Sigma^{-1}) \) constitute the natural parameter \( \theta \). All that remains to be done to express \( p(x) \) in terms of (2.39) is to rewrite \( g(\theta) \) in terms of \( c(\mu, \Sigma) \). The summary table on the following page contains details.

**Multinomial** Another popular distribution is one over \( k \) discrete events. In this case \( \mathcal{X} = \{1, \ldots, k\} \) and we have in completely generic terms \( p(x) = \pi_x \) where \( \pi_x \geq 0 \) and \( \sum_x \pi_x = 1 \). Now denote by \( e_x \in \mathbb{R}^k \) the \( x \)-th unit vector of the canonical basis, that is \( \langle e_x, e_{x'} \rangle = \delta_{x,x'} \). In this case we may rewrite \( p(x) \) via

\[
p(x) = \pi_x = \exp \left( \langle e_x, \log \pi \rangle \right)
\]

where \( \log \pi = (\log \pi_1, \ldots, \log \pi_k) \). In other words, we have succeeded in rewriting the distribution as a member of the exponential family.
where $\phi(x) = e^x$ and where $\theta = \log \pi$. Note that in this definition $\theta$ is restricted to a $k - 1$ dimensional manifold. If we relax those constraints we need to ensure that $p(x)$ remains normalized. Details are given in the summary table.

**Poisson** This distribution is often used to model distributions over discrete events. For instance, the number of raindrops which fall on a given surface area in a given amount of time, the number of stars in a given volume of space, or the number of Prussian soldiers killed by horse-kicks in the Prussian cavalry all follow this distribution. It is given by

$$p(x) = \frac{e^{-\lambda} \lambda^x}{x!} = \frac{1}{x!} \exp(x \log \lambda - \lambda) \text{ where } x \in \mathbb{N}_0.$$  \hfill (2.47)

By defining $\phi(x) = x$ we obtain an exponential families model. Note that things are a bit less trivial here since $\frac{1}{x!}$ is the nonuniform counting measure on $\mathbb{N}_0$. The case of the uniform measure which leads to the exponential distribution is discussed in Problem 2.15.

The reason why many discrete processes follow the Poisson distribution is that it can be seen as the limit over the average of a large number of Bernoulli draws: denote by $z \in \{0, 1\}$ a random variable with $p(z = 1) = \alpha$. Moreover, denote by $z_n$ the sum over $n$ draws from this random variable. In this case $z_n$ follows the multinomial distribution with $p(z_n = k) = \binom{n}{k} \alpha^k (1 - \alpha)^{n-k}$. Now assume that we let $n \to \infty$ such that the expected value of $z_n$ remains constant. That is, we rescale $\alpha = \frac{\lambda}{n}$. In this case we have

$$p(z_n = k) = \frac{n! \lambda^k}{(n-k)! n^k} \left(1 - \frac{\lambda}{n}\right)^{n-k} \right)$$  \hfill (2.48)

For $n \to \infty$ the second term converges to $e^{-\lambda}$. The third term converges to 1, since we have a product of only $2k$ terms, each of which converge to 1. Using the exponential families notation we may check that $E[x] = \lambda$ and that moreover also $\text{Var}[x] = \lambda$.

**Beta** This is a distribution on the unit interval $\mathcal{X} = [0, 1]$ which is very versatile when it comes to modelling unimodal and bimodal distributions. It is given by

$$p(x) = x^{a-1} (1-x)^{b-1} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)},$$  \hfill (2.49)
Taking logarithms we see that this, too, is an exponential families distribution, since \( p(x) = \exp((a - 1) \log x + (b - 1) \log(1 - x) + \log \Gamma(a + b) - \log \Gamma(a) - \log \Gamma(b)) \).

Figure 2.10 has a graphical description of the Poisson distribution and the Beta distribution. For a more comprehensive list of exponential family distributions see the table below and [Fel71, FT94, MN83]. In principle any map \( \phi(x) \), domain \( X \) with underlying measure \( \mu \) are suitable, as long as the log-partition function \( g(\theta) \) can be computed efficiently.

**Theorem 2.15 (Convex feasible domain)** The domain of definition \( \Theta \) of \( g(\theta) \) is convex.

**Proof** By construction \( g \) is convex and differentiable everywhere. Hence the below-sets for all values \( c \) with \( \{x|g(x) \leq c\} \) exist. Consequently the domain of definition is convex.

Having a convex function is very valuable when it comes to parameter inference since convex minimization problems have unique minimum values and global minima. We will discuss this notion in more detail when designing maximum likelihood estimators.

### 2.4 Estimation

In many statistical problems the challenge is to estimate parameters of interest. For instance, in the context of exponential families, we may want to estimate a parameter \( \theta \) such that it is close to the “true” parameter \( \theta^* \) in the distribution. While the problem is fully general, we will describe the
<table>
<thead>
<tr>
<th>Name</th>
<th>Domain</th>
<th>Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bernoulli</td>
<td>{0,1}</td>
<td>Counting</td>
</tr>
<tr>
<td>Multinomial</td>
<td>{1..N}</td>
<td>Counting</td>
</tr>
<tr>
<td>Exponential</td>
<td>(\mathbb{N}^+)</td>
<td>Counting</td>
</tr>
<tr>
<td>Poisson</td>
<td>(\mathbb{N}^+)</td>
<td>Counting</td>
</tr>
<tr>
<td>Laplace</td>
<td>([0,\infty))</td>
<td>Lebesgue</td>
</tr>
<tr>
<td>Gaussian</td>
<td>(\mathbb{R})</td>
<td>Lebesgue</td>
</tr>
<tr>
<td>Inverse (\chi^2)</td>
<td>(\mathbb{R}^n)</td>
<td>((-\infty, 0))</td>
</tr>
<tr>
<td>Logarithmic</td>
<td>(\mathbb{R})</td>
<td>((-\infty, 0))</td>
</tr>
<tr>
<td>Conjugate</td>
<td>(\Theta)</td>
<td>Conjugate</td>
</tr>
</tbody>
</table>

\(\Theta_n\) denotes the probability simplex in \(n\) dimensions. \(\mathbb{C}_n\) is the cone of positive semidefinite matrices in \(\mathbb{R}^{n \times n}\).
relevant steps in obtaining estimates for the special case of the exponential family. This is done for two reasons — firstly, exponential families are an important special case and we will encounter slightly more complex variants on the reasoning in later chapters of the book. Secondly, they are of a sufficiently simple form that we are able to show a range of different techniques. In more advanced applications only a small subset of those methods may be practically feasible. Hence exponential families provide us with a working example based on which we can compare the consequences of a number of different techniques.

2.4.1 Maximum Likelihood Estimation

Whenever we have a distribution $p(x; \theta)$ parametrized by some parameter $\theta$ we may use data to find a value of $\theta$ which maximizes the likelihood that the data would have been generated by a distribution with this choice of parameter.

For instance, assume that we observe a set of temperature measurements $X = \{x_1, \ldots, x_m\}$. In this case, we could try finding a normal distribution such that the likelihood $p(X; \theta)$ of the data under the assumption of a normal distribution is maximized. Note that this does not imply in any way that the temperature measurements are actually drawn from a normal distribution. Instead, it means that we are attempting to find the Gaussian which fits the data in the best fashion.

While this distinction may appear subtle, it is critical: we do not assume that our model accurately reflects reality. Instead, we simply try doing the best possible job at modeling the data given a specified model class. Later we will encounter alternative approaches at estimation, namely Bayesian methods, which make the assumption that our model ought to be able to describe the data accurately.

Definition 2.16 (Maximum Likelihood Estimator) For a model $p(\cdot; \theta)$ parametrized by $\theta$ and observations $X$ the maximum likelihood estimator (MLE) is

$$\hat{\theta}_{\text{ML}}[X] := \arg\max_{\theta} p(X; \theta). \quad (2.50)$$

In the context of exponential families this leads to the following procedure: given $m$ observations drawn iid from some distribution, we can express the
joint likelihood as
\[ p(X; \theta) = \prod_{i=1}^{m} p(x_i; \theta) = \prod_{i=1}^{m} \exp \left( \langle \phi(x_i), \theta \rangle - g(\theta) \right) \] (2.51)
\[ = \exp \left( m \langle \mu[X], \theta \rangle - g(\theta) \right) \] (2.52)
where \( \mu[X] := \frac{1}{m} \sum_{i=1}^{m} \phi(x_i) \). (2.53)

Here \( \mu[X] \) is the empirical average of the map \( \phi(x) \). Maximization of \( p(X; \theta) \) is equivalent to minimizing the negative log-likelihood \( -\log p(X; \theta) \). The latter is a common practical choice since for independently drawn data, the product of probabilities decomposes into the sum of the logarithms of individual likelihoods. This leads to the following objective function to be minimized
\[ -\log p(X; \theta) = m [g(\theta) - \langle \theta, \mu[X] \rangle] \] (2.54)

Since \( g(\theta) \) is convex and \( \langle \theta, \mu[X] \rangle \) is linear in \( \theta \), it follows that minimization of (2.54) is a convex optimization problem. Using Theorem 2.14 and the first order optimality condition \( \partial g(\theta) = \mu[X] \) for (2.54) implies that
\[ \theta = [\partial g(\theta)]^{-1} \langle \mu[X] \rangle \] or equivalently \( \mathbb{E}_{x \sim p(x; \theta)}[\phi(x)] = \partial g(\theta) = \mu[X] \). (2.55)

Put another way, the above conditions state that we aim to find the distribution \( p(x; \theta) \) which has the same expected value of \( \phi(x) \) as what we observed empirically via \( \mu[X] \). Under very mild technical conditions a solution to (2.55) exists.

In general, (2.55) cannot be solved analytically. In certain special cases, though, this is easily possible. We discuss two such choices in the following: Multinomial and Poisson distributions.

**Example 2.6 (Poisson Distribution)** For the Poisson distribution\(^1\) where \( p(x; \theta) = \frac{1}{x!} \exp(\theta x - e^\theta) \) it follows that \( g(\theta) = e^\theta \) and \( \phi(x) = x \). This allows us to solve (2.55) in closed form using
\[ \partial g(\theta) = e^\theta = \frac{1}{m} \sum_{i=1}^{m} x_i \] and hence \( \theta = \log \sum_{i=1}^{m} x_i - \log m \). (2.56)

\(^1\) Often the Poisson distribution is specified using \( \lambda := \log \theta \) as its rate parameter. In this case we have \( p(x; \lambda) = \lambda^x e^{-\lambda}/x! \) as its parametrization. The advantage of the natural parametrization using \( \theta \) is that we can directly take advantage of the properties of the log-partition function as generating the cumulants of \( x \).
Example 2.7 (Multinomial Distribution) For the multinomial distribution the log-partition function is given by $g(\theta) = \log \sum_{i=1}^{N} e^{\theta_i}$, hence we have that

$$\partial_i g(\theta) = \frac{e^{\theta_i}}{\sum_{j=1}^{N} e^{\theta_j}} = \frac{1}{m} \sum_{j=1}^{m} \{x_j = i\}.$$  \hfill (2.57)

It is easy to check that (2.57) is satisfied for $e^{\theta_i} = \frac{1}{m} \sum_{j=1}^{m} \{x_j = i\}$. In other words, the MLE for a discrete distribution simply given by the empirical frequencies of occurrence.

The multinomial setting also exhibits two rather important aspects of exponential families: firstly, choosing $\theta_i = c + \log \sum_{i=1}^{m} \{x_j = i\}$ for any $c \in \mathbb{R}$ will lead to an equivalent distribution. This is the case since the sufficient statistic $\phi(x)$ is not minimal. In our context this means that the coordinates of $\phi(x)$ are linearly dependent — for any $x$ we have that $\sum_j [\phi(x)]_j = 1$, hence we could eliminate one dimension. This is precisely the additional degree of freedom which is reflected in the scaling freedom in $\theta$.

Secondly, for data where some events do not occur at all, the expression $\log \left[ \sum_{j=1}^{m} \{x_j = i\} \right] = \log 0$ is ill defined. This is due to the fact that this particular set of counts occurs on the boundary of the convex set within which the natural parameters $\theta$ are well defined. We will see how different types of priors can alleviate the issue.

Using the MLE is not without problems. As we saw in Figure 2.1, convergence can be slow, since we are not using any side information. The latter can provide us with problems which are both numerically better conditioned and which show better convergence, provided that our assumptions are accurate. Before discussing a Bayesian approach to estimation, let us discuss basic statistical properties of the estimator.

2.4.2 Bias, Variance and Consistency

When designing any estimator $\hat{\theta}(X)$ we would like to obtain a number of desirable properties: in general it should not be biased towards a particular solution unless we have good reason to believe that this solution should be preferred. Instead, we would like the estimator to recover, at least on average, the “correct” parameter, should it exist. This can be formalized in the notion of an unbiased estimator.

Secondly, we would like that, even if no correct parameter can be found, e.g. when we are trying to fit a Gaussian distribution to data which is not
normally distributed, that we will converge to the best possible parameter choice as we obtain more data. This is what is understood by *consistency*.

Finally, we would like the estimator to achieve low bias and near-optimal estimates as quickly as possible. The latter is measured by the *efficiency* of an estimator. In this context we will encounter the Cramér-Rao bound which controls the best possible rate at which an estimator can achieve this goal. Figure 2.11 gives a pictorial description.

Fig. 2.11. Left: unbiased estimator; the estimates, denoted by circles have as mean the true parameter, as denoted by a star. Middle: consistent estimator. While the true model is not within the class we consider (as denoted by the ellipsoid), the estimates converge to the white star which is the best model within the class that approximates the true model, denoted by the solid star. Right: different estimators have different regions of uncertainty, as made explicit by the ellipses around the true parameter (solid star).

**Definition 2.17 (Unbiased Estimator)** An estimator \( \hat{\theta}[X] \) is unbiased if for all \( \theta \) where \( X \sim p(X; \theta) \) we have \( E_X[\hat{\theta}[X]] = \theta \).

In other words, in expectation the parameter estimate matches the true parameter. Note that this only makes sense if a true parameter actually *exists*. For instance, if the data is Poisson distributed and we attempt modeling it by a Gaussian we will obviously not obtain unbiased estimates.

For finite sample sizes MLE is often *biased*. For instance, for the normal distribution the variance estimates carry bias \( O(m^{-1}) \). See problem 2.18 for details. In general, under fairly mild conditions, MLE is asymptotically unbiased [DGL96]. We prove this for exponential families. For more general settings the proof depends on the dimensionality and smoothness of the family of densities that we have at our disposal.

**Theorem 2.18 (MLE for Exponential Families)** Assume that \( X \) is an \( m \)-sample drawn iid from \( p(x; \theta) \). The estimate \( \hat{\theta}[X] = g^{-1}(\mu[X]) \) is asymp-
2.4 Estimation

totally normal with
\[
m^{-\frac{1}{2}}[\hat{\theta}[X] - \theta] \to N(0, [\partial^2_\theta g(\theta)]^{-1}). \tag{2.58}
\]

In other words, the estimate \( \hat{\theta}[X] \) is asymptotically normal, it converges to the true parameter \( \theta \), and moreover, the variance at the correct parameter is given by the inverse of the covariance matrix of the data, as given by the second derivative of the log-partition function \( \partial^2_\theta g(\theta) \).

**Proof** Denote by \( \mu = \partial_\theta g(\theta) \) the true mean. Moreover, note that \( \partial^2_\theta g(\theta) \) is the covariance of the data drawn from \( p(x; \theta) \). By the central limit theorem (Theorem 2.3) we have that \( n^{-\frac{1}{2}}[\mu[X] - \mu] \to N(0, \partial^2_\theta g(\theta)) \).

Now note that \( \hat{\theta}[X] = [\partial_\theta g]^{-1}(\mu[X]) \). Therefore, by the delta method (Theorem 2.5) we know that \( \hat{\theta}[X] \) is also asymptotically normal. Moreover, by the inverse function theorem the Jacobian of \( g^{-1} \) satisfies \( \partial_\mu [\partial_\theta g]^{-1}(\mu) = [\partial^2_\theta g(\theta)]^{-1} \). Applying Slutsky’s theorem (Theorem 2.4) proves the claim. 

Now that we established the asymptotic properties of the MLE for exponential families it is only natural to ask how much variation one may expect in \( \hat{\theta}[X] \) when performing estimation. The Cramer-Rao bound governs this.

**Theorem 2.19 (Cramér and Rao [Rao73])** Assume that \( X \) is drawn from \( p(X; \theta) \) and let \( \hat{\theta}[X] \) be an asymptotically unbiased estimator. Denote by \( I \) the Fisher information matrix and by \( B \) the variance of \( \hat{\theta}[X] \) where
\[
I := \text{Var}[\partial_\theta \log p(X; \theta)] \text{ and } B := \text{Var}[\hat{\theta}[X]]. \tag{2.59}
\]

In this case \( \det IB \geq 1 \) for all estimators \( \hat{\theta}[X] \).

**Proof** We prove the claim for the scalar case. The extension to matrices is straightforward. Using the Cauchy-Schwarz inequality we have
\[
\text{Cov}^2 [\partial_\theta \log p(X; \theta), \hat{\theta}[X]] \leq \text{Var} [\partial_\theta \log p(X; \theta)] \text{Var}[\hat{\theta}[X]] = IB. \tag{2.60}
\]

Note that at the true parameter the expected log-likelihood score vanishes
\[
\mathbb{E}_X[\partial_\theta \log p(X; \theta)] = \partial_\theta \int p(X; \theta)dX = \partial_\theta 1 = 0. \tag{2.61}
\]

Hence we may simplify the covariance formula by dropping the means via
\[
\text{Cov} [\partial_\theta \log p(X; \theta), \hat{\theta}[X]] = \mathbb{E}_X [\partial_\theta \log p(X; \theta)\hat{\theta}[X]]
= \int p(X; \theta)\hat{\theta}(X)\partial_\theta \log p(X; \theta)d\theta
= \partial_\theta \int p(X; \theta)\hat{\theta}(X)dX = \partial_\theta 1 = 1.
\]
Here the last equality follows since we may interchange integration by $X$ and the derivative with respect to $\theta$.

The Cramér-Rao theorem implies that there is a limit to how well we may estimate a parameter given finite amounts of data. It is also a yardstick by which we may measure how efficiently an estimator uses data. Formally, we define the efficiency as the quotient between actual performance and the Cramér-Rao bound via

$$e := 1 / \det IB.$$  \hfill (2.62)

The closer $e$ is to 1, the lower the variance of the corresponding estimator $\hat{\theta}(X)$. Theorem 2.18 implies that for exponential families MLE is asymptotically efficient. It turns out to be generally true.

**Theorem 2.20 (Efficiency of MLE [Cra46, GW92, Ber85])** The maximum likelihood estimator is asymptotically efficient ($e = 1$).

So far we only discussed the behavior of $\hat{\theta}(X)$ whenever there exists a true $\theta$ generating $p(\theta; X)$. If this is not true, we need to settle for less: how well $\hat{\theta}(X)$ approaches the best possible choice of within the given model class. Such behavior is referred to as consistency. Note that it is not possible to define consistency per se. For instance, we may ask whether $\hat{\theta}$ converges to the optimal parameter $\theta^*$, or whether $p(x; \hat{\theta})$ converges to the optimal density $p(x; \theta^*)$, and with respect to which norm. Under fairly general conditions this turns out to be true for finite-dimensional parameters and smoothly parametrized densities. See [DGL96, vdG00] for proofs and further details.

### 2.4.3 A Bayesian Approach

The analysis of the Maximum Likelihood method might suggest that inference is a solved problem. After all, in the limit, MLE is unbiased and it exhibits as small variance as possible. Empirical results using a finite amount of data, as present in Figure 2.1 indicate otherwise.

While not making any assumptions can lead to interesting and general theorems, it ignores the fact that in practice we almost always have some idea about what to expect of our solution. It would be foolish to ignore such additional information. For instance, when trying to determine the voltage of a battery, it is reasonable to expect a measurement in the order of 1.5V or less. Consequently such prior knowledge should be incorporated into the estimation process. In fact, the use of side information to guide estimation
turns out to be the tool to building estimators which work well in high dimensions.

Recall Bayes’ rule (1.9) which states that \( p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \). In our context this means that if we are interested in the posterior probability of \( \theta \) assuming a particular value, we may obtain this using the likelihood (often referred to as evidence) of \( x \) having been generated by \( \theta \) via \( p(x|\theta) \) and our prior belief \( p(\theta) \) that \( \theta \) might be chosen in the distribution generating \( x \). Observe the subtle but important difference to MLE: instead of treating \( \theta \) as a parameter of a density model, we treat \( \theta \) as an unobserved random variable which we may attempt to infer given the observations \( X \).

This can be done for a number of different purposes: we might want to infer the most likely value of the parameter given the posterior distribution \( p(\theta|X) \). This is achieved by

\[
\hat{\theta}_{\text{MAP}}(X) := \arg\max_{\theta} p(\theta|X) = \arg\min_{\theta} -\log p(X|\theta) - \log p(\theta).
\] (2.63)

The second equality follows since \( p(X) \) does not depend on \( \theta \). This estimator is also referred to as the Maximum a Posteriori, or MAP estimator. It differs from the maximum likelihood estimator by adding the negative log-prior to the optimization problem. For this reason it is sometimes also referred to as Penalized MLE. Effectively we are penalizing unlikely choices \( \theta \) via \( -\log p(\theta) \).

Note that using \( \hat{\theta}_{\text{MAP}}(X) \) as the parameter of choice is not quite accurate. After all, we can only infer a distribution over \( \theta \) and in general there is no guarantee that the posterior is indeed concentrated around its mode. A more accurate treatment is to use the distribution \( p(\theta|X) \) directly via

\[
p(x|X) = \int p(x|\theta)p(\theta|X)d\theta.
\] (2.64)

In other words, we integrate out the unknown parameter \( \theta \) and obtain the density estimate directly. As we will see, it is generally impossible to solve (2.64) exactly, an important exception being conjugate priors. In the other cases one may resort to sampling from the posterior distribution to approximate the integral.

While it is possible to design a wide variety of prior distributions, this book focuses on two important families: norm-constrained prior and conjugate priors. We will encounter them throughout, the former sometimes in the guise of regularization and Gaussian Processes, the latter in the context of exchangeable models such as the Dirichlet Process.
Norm-constrained priors take on the form

\[ p(\theta) \propto \exp(-\lambda \|\theta - \theta_0\|^d_p) \] for \( p, d \geq 1 \) and \( \lambda > 0 \) \hspace{1cm} (2.65)

That is, they restrict the deviation of the parameter value \( \theta \) from some guess \( \theta_0 \). The intuition is that extreme values of \( \theta \) are much less likely than more moderate choices of \( \theta \) which will lead to more smooth and even distributions \( p(x|\theta) \).

A popular choice is the Gaussian prior which we obtain for \( p = d = 1 \) and \( \lambda = 1/2\sigma^2 \). Typically one sets \( \theta_0 = 0 \) in this case. Note that in (2.65) we did not spell out the normalization of \( p(\theta) \) — in the context of MAP estimation this is not needed since it simply becomes a constant offset in the optimization problem (2.63). We have

\[ \hat{\theta}_{\text{MAP}}[X] = \arg\min_{\theta} \left[ g(\theta) - \langle \theta, \mu[X]\rangle + \lambda \|\theta - \theta_0\|^d_p \right] \hspace{1cm} (2.66) \]

For \( d, p \geq 1 \) and \( \lambda \geq 0 \) the resulting optimization problem is \textit{convex} and it has a unique solution. Moreover, very efficient algorithms exist to solve this problem. We will discuss this in detail in Chapter 5. Figure 2.12 shows the regions of equal prior probability for a range of different norm-constrained priors.

As can be seen from the diagram, the choice of the norm can have profound consequences on the solution. That said, as we will show in Chapter ??, the estimate \( \hat{\theta}_{\text{MAP}} \) is well concentrated and converges to the optimal solution under fairly general conditions.

An alternative to norm-constrained priors are \textit{conjugate} priors. They are designed such that the posterior \( p(\theta|X) \) has the same functional form as the
2.4 Estimation

prior $p(\theta)$. In exponential families such priors are defined via

$$p(\theta|n, \nu) = \exp \left( (n\nu, \theta) - ng(\theta) - h(\nu, n) \right) \quad \text{where} \quad (2.67)$$

$$h(\nu, n) = \log \int \exp \left( (n\nu, \theta) - ng(\theta) \right) d\theta. \quad (2.68)$$

Note that $p(\theta|n, \nu)$ itself is a member of the exponential family with the feature map $\phi(\theta) = (\theta, -g(\theta))$. Hence $h(\nu, n)$ is convex in $(n\nu, n)$. Moreover, the posterior distribution has the form

$$p(\theta|X) \propto p(X|\theta)p(\theta|n, \nu) \propto \exp \left( \langle m\mu[X] + n\nu, \theta \rangle - (m + n)g(\theta) \right). \quad (2.69)$$

That is, the posterior distribution has the same form as a conjugate prior with parameters $\frac{m\mu[X]+n\nu}{m+n}$ and $m + n$. In other words, $n$ acts like a phantom sample size and $\nu$ is the corresponding mean parameter. Such an interpretation is reasonable given our desire to design a prior which, when combined with the likelihood remains in the same model class: we treat prior knowledge as having observed virtual data beforehand which is then added to the actual set of observations. In this sense data and prior become completely equivalent — we obtain our knowledge either from actual observations or from virtual observations which describe our belief into how the data generation process is supposed to behave.

Eq. (2.69) has the added benefit of allowing us to provide an exact normalized version of the posterior. Using (2.67) we obtain that

$$p(\theta|X) = \exp \left( \langle m\mu[X] + n\nu, \theta \rangle - (m + n)g(\theta) \right) \times$$

$$\exp \left( \langle m\mu[X] + n\nu, \theta \rangle - (m + n)g(\theta) - h \left( \frac{m\mu[X]+n\nu}{m+n}, m + n \right) \right) d\theta.$$

The main remaining challenge is to compute the normalization $h$ for a range of important conjugate distributions. The table on the following page provides details. Besides attractive algebraic properties, conjugate priors also have a second advantage — the integral (2.64) can be solved exactly:

$$p(x|X) = \int \exp \left( \langle \phi(x), \theta \rangle - g(\theta) \right) \times$$

$$\exp \left( \langle m\mu[X] + n\nu, \theta \rangle - (m + n)g(\theta) - h \left( \frac{m\mu[X]+n\nu}{m+n}, m + n \right) \right) d\theta.$$

Combining terms one may check that the integrand amounts to the normalization in the conjugate distribution, albeit $\phi(x)$ added. This yields

$$p(x|X) = \exp \left( h \left( \frac{m\mu[X]+n\nu+n\phi(x)}{m+n+1}, m + n + 1 \right) - h \left( \frac{m\mu[X]+n\nu}{m+n}, m + n \right) \right)$$

Such an expansion is very useful whenever we would like to draw $x$ from $p(x|X)$ without the need to obtain an instantiation of the latent variable $\theta$. We provide explicit expansions in appendix 2. [?] use the fact that $\theta$ can be
2 Density Estimation

integrated out to obtain what is called a collapsed Gibbs sampler for topic models [BNJ03].

2.4.4 An Example

Assume we would like to build a language model based on available documents. For instance, a linguist might be interested in estimating the frequency of words in Shakespeare’s collected works, or one might want to compare the change with respect to a collection of webpages. While models describing documents by treating them as bags of words which all have been obtained independently of each other are exceedingly simple, they are valuable for quick-and-dirty content filtering and categorization, e.g. a spam filter on a mail server or a content filter for webpages.

Hence we model a document $d$ as a multinomial distribution: denote by $w_i$ for $i \in \{1, \ldots, m_d\}$ the words in $d$. Moreover, denote by $p(w|\theta)$ the probability of occurrence of word $w$, then under the assumption that the words are independently drawn, we have

$$p(d|\theta) = \prod_{i=1}^{m_d} p(w_i|\theta). \tag{2.70}$$

It is our goal to find parameters $\theta$ such that $p(d|\theta)$ is accurate. For a given collection $D$ of documents denote by $m_w$ the number of counts for word $w$ in the entire collection. Moreover, denote by $m$ the total number of words in the entire collection. In this case we have

$$p(D|\theta) = \prod_i p(d_i|\theta) = \prod_w p(w|\theta)^{m_w}. \tag{2.71}$$

Finding suitable parameters $\theta$ given $D$ proceeds as follows: In a maximum likelihood model we set

$$p(w|\theta) = \frac{m_w}{m}. \tag{2.72}$$

In other words, we use the empirical frequency of occurrence as our best guess and the sufficient statistic of $D$ is $\phi(w) = e_w$, where $e_w$ denotes the unit vector which is nonzero only for the “coordinate” $w$. Hence $\mu[D]_w = \frac{m_w}{m}$.

We know that the conjugate prior of the multinomial model is a Dirichlet model. It follows from (2.69) that the posterior mode is obtained by replacing $\mu[D]$ by $\frac{m_w[D]+n_w}{m+n}$. Denote by $n_w := \nu_w \cdot n$ the pseudo-counts arising from the conjugate prior with parameters $(\nu, n)$. In this case we will estimate the
probability of the word \( w \) as

\[
p(w|\theta) = \frac{m_w + n_w}{m + n} = \frac{m_w + n \nu_w}{m + n}. \tag{2.73}
\]

In other words, we add the pseudo counts \( n_w \) to the actual word counts \( m_w \). This is particularly useful when the document we are dealing with is brief, that is, whenever we have little data: it is quite unreasonable to infer from a webpage of approximately 1000 words that words not occurring in this page have zero probability. This is exactly what is mitigated by means of the conjugate prior \((\nu, n)\).

Finally, let us consider norm-constrained priors of the form \((2.65)\). In this case, the integral required for

\[
p(D) = \int p(D|\theta)p(\theta)d\theta
\]

\[
\propto \int \exp \left( -\lambda \|\theta - \theta_0\|^d_p + m \langle \mu[D], \theta \rangle - mg(\theta) \right) d\theta
\]

is intractable and we need to resort to an approximation. A popular choice is to replace the integral by \( p(D|\theta^*) \) where \( \theta^* \) maximizes the integrand. This is precisely the MAP approximation of \((2.63)\). Hence, in order to perform estimation we need to solve

\[
\min_{\theta} g(\theta) - \langle \mu[D], \theta \rangle + \frac{\lambda}{m} \|\theta - \theta_0\|^d_p. \tag{2.74}
\]

A very simple strategy for minimizing \((2.74)\) is gradient descent. That is for a given value of \( \theta \) we compute the gradient of the objective function and take a fixed step towards its minimum. For simplicity assume that \( d = p = 2 \) and \( \lambda = 1/2\sigma^2 \), that is, we assume that \( \theta \) is normally distributed with variance \( \sigma^2 \) and mean \( \theta_0 \). The gradient is given by

\[
\partial_{\theta} \left[ -\log p(D, \theta) \right] = \mathbf{E}_{x \sim p(x|\theta)}[\phi(x)] - \mu[D] + \frac{1}{m \sigma^2}[\theta - \theta_0]. \tag{2.75}
\]

In other words, it depends on the discrepancy between the mean of \( \phi(x) \) with respect to our current model and the empirical average \( \mu[X] \), and the difference between \( \theta \) and the prior mean \( \theta_0 \).

Unfortunately, convergence of the procedure \( \theta \leftarrow \theta - \eta \partial_{\theta} \ldots \) is usually very slow, even if we adjust the steplength \( \eta \) efficiently. The reason is that the gradient need not point towards the minimum as the space is most likely distorted. A better strategy is to use Newton’s method (see Chapter 5 for a detailed discussion and a convergence proof). It relies on a second order
Taylor approximation

\[- \log p(D, \theta + \delta) \approx - \log p(D, \theta) + \langle \delta, G \rangle + \frac{1}{2} \delta^T H \delta \quad (2.76)\]

where $G$ and $H$ are the first and second derivatives of $- \log p(D, \theta)$ with respect to $\theta$. The quadratic expression can be minimized with respect to $\delta$ by choosing $\delta = -H^{-1}G$ and we can fashion an update algorithm from this by letting $\theta \leftarrow \theta - H^{-1}G$. One may show (see Chapter 5) that Algorithm 2.1 is quadratically convergent. Note that the prior on $\theta$ ensures that $H$ is well conditioned even in the case where the variance of $\phi(x)$ is not. In practice this means that the prior ensures fast convergence of the optimization algorithm.

**Algorithm 2.1** Newton method for MAP estimation

NewtonMAP($D$)

Initialize $\theta = \theta_0$

while not converged do

Compute $G = \mathbb{E}_{x \sim p(x|\theta)}[\phi(x)] - \mu[D] + \frac{1}{m\sigma^2}[^{\theta - \theta_0}]$

Compute $H = \text{Var}_{x \sim p(x|\theta)}[\phi(x)] + \frac{1}{m\sigma^2} I$

Update $\theta \leftarrow \theta - H^{-1}G$

end while

return $\theta$

2.5 Generating Random Numbers

WORK IN PROGRESS.

So far we considered the problem of estimating the underlying probability density, given a set of samples drawn from that density. Now let us turn to the converse problem, that is, how to generate random variables given the underlying probability density. In other words, we want to design a random variable generator. This is very useful in a number of situations. Let us illustrate with an example.

Suppose you are interested in testing the performance of a network router under different load conditions. Instead of introducing the under-development router in a live network and wreaking havoc, one could estimate the probability density of the network traffic under various load conditions and build a model. The behavior of the network can then be simulated by simulating our probabilistic model. This involves drawing random variables from an estimated probability distribution.

To carry this illustration further, suppose that you generate data packets