where all of these quantities are viewed as functions of $\mathbf{w}$. The denominator in (2.111) is the normalization constant, which ensures that the posterior distribution on the left-hand side is a valid probability density and integrates to one. Indeed, by integrating both sides of (2.111) with respect to $\mathbf{w}$, we can express the denominator in Bayes' theorem in terms of the prior distribution and the likelihood function:

$$p(\mathcal{D}) = \int p(\mathcal{D}|\mathbf{w})p(\mathbf{w})\,\mathrm{d}\mathbf{w}. \tag{2.113}$$

In both the Bayesian and frequentist paradigms, the likelihood function $p(\mathcal{D}|\mathbf{w})$ plays a central role. However, the manner in which it is used is fundamentally different in the two approaches. In a frequentist setting, $\mathbf{w}$ is considered to be a fixed parameter, whose value is determined by some form of 'estimator', and error bars on this estimate are determined (conceptually, at least) by considering the distribution of possible data sets $\mathcal{D}$. By contrast, from the Bayesian viewpoint there is only a single data set $\mathcal{D}$ (namely the one that is actually observed), and the uncertainty in the parameters is expressed through a probability distribution over $\mathbf{w}$.

2.6.2 Regularization

Section 1.2.5

We can use this Bayesian perspective to gain insight into the technique of regularization that was used in the sine curve regression example to reduce over-fitting. Instead of choosing the model parameters by maximizing the likelihood function with respect to $\mathbf{w}$, we can maximize the posterior probability (2.111). This technique is called the *maximum a posteriori* estimate, or simply *MAP* estimate. Equivalently, we can minimize the negative log of the posterior probability. Taking negative logs of both sides of (2.111), we have

$$-\ln p(\mathbf{w}|\mathcal{D}) = -\ln p(\mathcal{D}|\mathbf{w}) - \ln p(\mathbf{w}) + \ln p(\mathcal{D}). \tag{2.114}$$

The first term on the right-hand side of (2.114) is the usual log likelihood. The third term can be omitted since it does not depend on $\mathbf{w}$. The second term takes the form of a function of $\mathbf{w}$, which is added to the log likelihood, and we can recognize this as a form of regularization. To make this more explicit, suppose we choose the prior distribution $p(\mathbf{w})$ to be the product of independent zero-mean Gaussian distributions for each of the elements of $\mathbf{w}$ such that each has the same variance s^2 so that

$$p(\mathbf{w}|s) = \prod_{i=0}^{M} \mathcal{N}(w_i|0, s^2) = \prod_{i=0}^{M} \left(\frac{1}{2\pi s^2}\right)^{1/2} \exp\left\{-\frac{w_i^2}{2s^2}\right\}. \tag{2.115}$$

Substituting into (2.114), we obtain

$$-\ln p(\mathbf{w}|\mathcal{D}) = -\ln p(\mathcal{D}|\mathbf{w}) + \frac{1}{2s^2}\sum_{i=0}^{M} w_i^2 + \text{const}. \tag{2.116}$$

If we consider the particular case of the linear regression model whose log likelihood is given by (2.66), then we find that maximizing the posterior distribution is equivalent to minimizing the function

Exercise 2.41