$$E(\mathbf{w}) = \frac{1}{2\sigma^2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{1}{2s^2}\mathbf{w}^{\mathrm{T}}\mathbf{w}. \tag{2.117}$$

We see that this takes the form of the regularized sum-of-squares error function encountered earlier in the form (1.4).

2.6.3 Bayesian machine learning

The Bayesian perspective has allowed us to motivate the use of regularization and to derive a specific form for the regularization term. However, the use of Bayes' theorem alone does not constitute a truly Bayesian treatment of machine learning since it is still finding a single solution for $\mathbf{w}$ and does not therefore take account of uncertainty in the value of $\mathbf{w}$. Suppose we have a training data set $\mathcal{D}$ and our goal is to predict some target variable t given a new input value x. We are therefore interested in the distribution of t given both x and $\mathcal{D}$. From the sum and product rules of probability, we have

$$p(t|x, \mathcal{D}) = \int p(t|x, \mathbf{w})p(\mathbf{w}|\mathcal{D})\,\mathrm{d}\mathbf{w}. \tag{2.118}$$

We see that the prediction is obtained by taking a weighted average $p(t|x, \mathbf{w})$ over all possible values of $\mathbf{w}$ in which the weighting function is given by the posterior probability distribution $p(\mathbf{w}|\mathcal{D})$. The key difference that distinguishes Bayesian methods is this integration over the space of parameters. By contrast, conventional frequentist methods use point estimates for parameters obtained by optimizing a loss function such as a regularized sum-of-squares.

This fully Bayesian treatment of machine learning offers some powerful insights. For example, the problem of over-fitting, encountered earlier in the context of polynomial regression, is an example of a pathology arising from the use of maximum likelihood, and does not arise when we marginalize over parameters using the Bayesian approach. Similarly, we may have multiple potential models that we could use to solve a given problem, such as polynomials of different orders in the regression example. A maximum likelihood approach simply picks the model that gives the highest probability of the data, but this favours ever more complex models, leading to over-fitting. A fully Bayesian treatment involves averaging over all possible models, with the contribution of each model weighted by its posterior probability. Moreover, this probability is typically highest for models of intermediate complexity. Very simple models (such as polynomials of low order) have low probability as they are unable to fit the data well, whereas very complex models (such as polynomials of very high order) also have low probability because the Bayesian integration over parameters automatically and elegantly penalizes complexity. For a comprehensive overview of Bayesian methods applied to machine learning, including neural networks, see Bishop (2006).

Section 1.2

Section 9.6

Unfortunately, there is a major drawback with the Bayesian framework, and this is apparent in (2.118), which involves integrating over the space of parameters. Modern deep learning models can have millions or billions of parameters and even simple approximations to such integrals are typically infeasible. In fact, given a