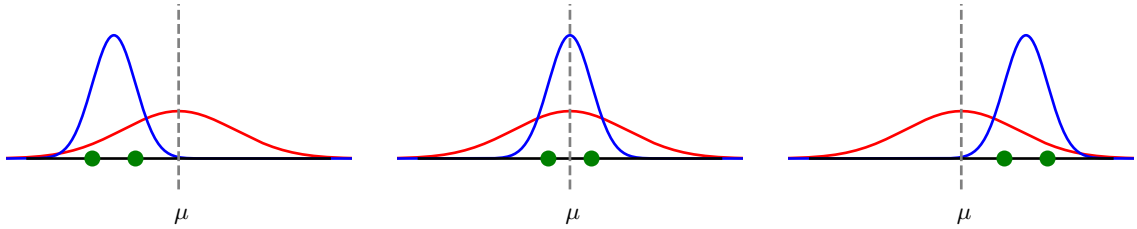

Figure 2.10 Illustration of how bias arises when using maximum likelihood to determine the mean and variance of a Gaussian. The red curves show the true Gaussian distribution from which data is generated, and the three blue curves show the Gaussian distributions obtained by fitting to three data sets, each consisting of two data points shown in green, using the maximum likelihood results (2.57) and (2.58). Averaged across the three data sets, the mean is correct, but the variance is systematically underestimated because it is measured relative to the sample mean and not relative to the true mean.

Correcting for the bias of maximum likelihood in complex models such as neural networks is not so easy, however.

Note that the bias of the maximum likelihood solution becomes less significant as the number N of data points increases. In the limit $N \to \infty$ the maximum likelihood solution for the variance equals the true variance of the distribution that generated the data. In the case of the Gaussian, for anything other than small N, this bias will not prove to be a serious problem. However, throughout this book we will be interested in complex models with many parameters, for which the bias problems associated with maximum likelihood will be much more severe. In fact, the issue of bias in maximum likelihood is closely related to the problem of *over-fitting*.

Section 2.6.3

2.3.4 Linear regression

Section 1.2

We have seen how the problem of linear regression can be expressed in terms of error minimization. Here we return to this example and view it from a probabilistic perspective, thereby gaining some insights into error functions and regularization.

The goal in the regression problem is to be able to make predictions for the target variable t given some new value of the input variable x by using a set of training data comprising N input values $\mathbf{x} = (x_1, \ldots, x_N)$ and their corresponding target values $\mathbf{t} = (t_1, \ldots, t_N)$. We can express our uncertainty over the value of the target variable using a probability distribution. For this purpose, we will assume that, given the value of x, the corresponding value of t has a Gaussian distribution with a mean equal to the value $y(x, \mathbf{w})$ of the polynomial curve given by (1.1), where $\mathbf{w}$ are the polynomial coefficients, and a variance σ^2. Thus, we have

$$p(t|x, \mathbf{w}, \sigma^2) = \mathcal{N}\left(t|y(x, \mathbf{w}), \sigma^2\right). \tag{2.64}$$

This is illustrated schematically in Figure 2.11.

We now use the training data $\{\mathbf{x}, \mathbf{t}\}$ to determine the values of the unknown parameters $\mathbf{w}$ and σ^2 by maximum likelihood. If the data is assumed to be drawn