



**Figure 1.9** Plots of  $M = 9$  polynomials fitted to the data set shown in Figure 1.4 using the regularized error function (1.4) for two values of the regularization parameter  $\lambda$  corresponding to  $\ln \lambda = -18$  and  $\ln \lambda = 0$ . The case of no regularizer, i.e.,  $\lambda = 0$ , corresponding to  $\ln \lambda = -\infty$ , is shown at the bottom right of Figure 1.6.

term takes the form of the sum of the squares of all of the coefficients, leading to a modified error function of the form

$$\tilde{E}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2 \quad (1.4)$$

where  $\|\mathbf{w}\|^2 \equiv \mathbf{w}^T \mathbf{w} = w_0^2 + w_1^2 + \dots + w_M^2$ , and the coefficient  $\lambda$  governs the relative importance of the regularization term compared with the sum-of-squares error term. Note that often the coefficient  $w_0$  is omitted from the regularizer because its inclusion causes the results to depend on the choice of origin for the target variable (Hastie, Tibshirani, and Friedman, 2009), or it may be included but with its own regularization coefficient. Again, the error function in (1.4) can be minimized exactly in closed form. Techniques such as this are known in the statistics literature as *shrinkage* methods because they reduce the value of the coefficients. In the context of neural networks, this approach is known as *weight decay* because the parameters in a neural network are called weights and this regularizer encourages them to decay towards zero.

Figure 1.9 shows the results of fitting the polynomial of order  $M = 9$  to the same data set as before but now using the regularized error function given by (1.4). We see that, for a value of  $\ln \lambda = -18$ , the over-fitting has been suppressed and we now obtain a much closer representation of the underlying function  $\sin(2\pi x)$ . If, however, we use too large a value for  $\lambda$  then we again obtain a poor fit, as shown in Figure 1.9 for  $\ln \lambda = 0$ . The corresponding coefficients from the fitted polynomials are given in Table 1.2, showing that regularization has the desired effect of reducing the magnitude of the coefficients.

The impact of the regularization term on the generalization error can be seen by plotting the value of the RMS error (1.3) for both training and test sets against  $\ln \lambda$ , as shown in Figure 1.10. We see that  $\lambda$  now controls the effective complexity of the model and hence determines the degree of over-fitting.