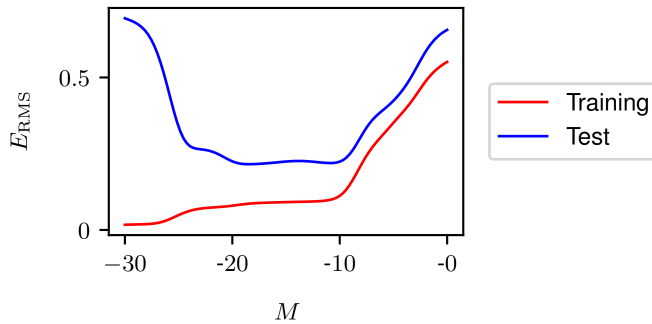
Figure 1.10 Graph of the root-mean-square error (1.3) versus $\ln \lambda$ for the $M = 9$ polynomial.



1.2.6 Model selection

The quantity λ is an example of a *hyperparameter* whose values are fixed during the minimization of the error function to determine the model parameters $\mathbf{w}$. We cannot simply determine the value of λ by minimizing the error function jointly with respect to $\mathbf{w}$ and λ since this will lead to $\lambda \to 0$ and an over-fitted model with small or zero training error. Similarly, the order M of the polynomial is a hyperparameter of the model, and simply optimizing the training set error with respect to M will lead to large values of M and associated over-fitting. We therefore need to find a way to determine suitable values for hyperparameters. The results above suggest a simple way of achieving this, namely by taking the available data and partitioning it into a training set, used to determine the coefficients $\mathbf{w}$, and a separate *validation* set, also called a *hold-out* set or a *development set*. We then select the model having the lowest error on the validation set. If the model design is iterated many times using a data set of limited size, then some over-fitting to the validation data can occur, and so it may be necessary to keep aside a third *test set* on which the performance of the selected model can finally be evaluated.

For some applications, the supply of data for training and testing will be limited. To build a good model, we should use as much of the available data as possible for training. However, if the validation set is too small, it will give a relatively noisy estimate of predictive performance. One solution to this dilemma is to use *cross-*

Table 1.2 Table of the coefficients $\mathbf{w}^{\star}$ for $M = 9$ polynomials with various values for the regularization parameter λ. Note that $\ln \lambda = -\infty$ corresponds to a model with no regularization, i.e., to the graph at the bottom right in Figure 1.6. We see that, as the value of λ increases, the magnitude of a typical coefficient gets smaller.

	$\ln \lambda = -\infty$	$\ln \lambda = -18$	$\ln \lambda = 0$
$w_0^{\star}$	0.26	0.26	0.11
$w_1^{\star}$	-66.13	0.64	-0.07
$w_2^{\star}$	$1,665.69$	43.68	-0.09
$w_3^{\star}$	$-15,566.61$	-144.00	-0.07
$w_4^{\star}$	$76,321.23$	57.90	-0.05
$w_5^{\star}$	$-217,389.15$	117.36	-0.04
$w_6^{\star}$	$370,626.48$	9.87	-0.02
$w_7^{\star}$	$-372,051.47$	-90.02	-0.01
$w_8^{\star}$	$202,540.70$	-70.90	-0.01
$w_9^{\star}$	$-46,080.94$	75.26	0.00