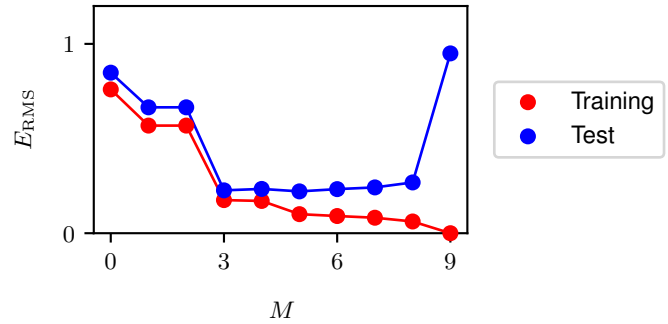


Figure 1.7 Graphs of the root-mean-square error, defined by (1.3), evaluated on the training set, and on an independent test set, for various values of M .



RMS errors are shown, for various values of M , in Figure 1.7. The test set error is a measure of how well we are doing in predicting the values of t for new data observations of x . Note from Figure 1.7 that small values of M give relatively large values of the test set error, and this can be attributed to the fact that the corresponding polynomials are rather inflexible and are incapable of capturing the oscillations in the function $\sin(2\pi x)$. Values of M in the range $3 \leq M \leq 8$ give small values for the test set error, and these also give reasonable representations of the generating function $\sin(2\pi x)$, as can be seen for $M = 3$ in Figure 1.6.

For $M = 9$, the training set error goes to zero, as we might expect because this polynomial contains 10 degrees of freedom corresponding to the 10 coefficients w_0, \dots, w_9 , and so can be tuned exactly to the 10 data points in the training set. However, the test set error has become very large and, as we saw in Figure 1.6, the corresponding function $y(x, \mathbf{w}^*)$ exhibits wild oscillations.

This may seem paradoxical because a polynomial of a given order contains all lower-order polynomials as special cases. The $M = 9$ polynomial is therefore capable of generating results at least as good as the $M = 3$ polynomial. Furthermore, we might suppose that the best predictor of new data would be the function $\sin(2\pi x)$ from which the data was generated (and we will see later that this is indeed the case). We know that a power series expansion of the function $\sin(2\pi x)$ contains terms of all orders, so we might expect that results should improve monotonically as we increase M .

We can gain some insight into the problem by examining the values of the coefficients \mathbf{w}^* obtained from polynomials of various orders, as shown in Table 1.1. We see that, as M increases, the magnitude of the coefficients typically gets larger. In particular for the $M = 9$ polynomial, the coefficients have become finely tuned to the data. They have large positive and negative values so that the corresponding polynomial function matches each of the data points exactly, but between data points (particularly near the ends of the range) the function exhibits the large oscillations observed in Figure 1.6. Intuitively, what is happening is that the more flexible polynomials with larger values of M are increasingly tuned to the random noise on the target values.

Further insight into this phenomenon can be gained by examining the behaviour of the learned model as the size of the data set is varied, as shown in Figure 1.8. We see that, for a given model complexity, the over-fitting problem become less severe