7 Gradient Descent

In the previous chapter we saw that neural networks are a very broad and flexible class of functions and are able in principle to approximate any desired function to arbitrarily high accuracy given a sufficiently large number of hidden units. Moreover, we saw that deep neural networks can encode inductive biases corresponding to hierarchical representations, which prove valuable in a wide range of practical applications. We now turn to the task of finding a suitable setting for the network parameters (weights and biases), based on a set of training data.

As with the regression and classification models discussed in earlier chapters, we choose the model parameters by optimizing an error function. We have seen how to define a suitable error function for a particular application by using maximum likelihood. Although in principle the error function could be minimized numerically through a series of direct error function evaluations, this turns out to be very inefficient. Instead, we turn to another core concept that is used in deep learning, which

Section 6.4

C. M. Bishop, H. Bishop, *Deep Learning*, https://doi.org/10.1007/978-3-031-45468-4_7