is that optimizing the error function can be done much more efficiently by making use of gradient information, in other words by evaluating the derivatives of the error function with respect to the network parameters. This is why we took care to ensure that the function represented by the neural network is differentiable by design. Likewise, the error function itself also needs to be differentiable.

The required derivatives of the error function with respect to each of the parameters in the network can be evaluated efficiently using a technique called *back-propagation*, which involves successive computations that flow backwards through the network in a way that is analogous to the forward flow of function computations during the evaluation of the network outputs.

Chapter 8

Although the likelihood is used to define an error function, the goal when optimizing the error function in a neural network is to achieve good generalization on test data. In classical statistics, maximum likelihood is used to fit a parametric model to a finite data set, in which the number of data points typically far exceeds the number of parameters in the model. The optimal solution has the maximum value of the likelihood function, and the values found for the fitted parameters are of direct interest. By contrast, modern deep learning works with very rich models containing huge numbers of learnable parameters, and the goal is never simply exact optimization. Instead, the properties and behaviour of the learning algorithm itself, along with various methods for regularization, are important in determining how well the solution generalizes to new data.

Section 9.3.2

Chapter 9

7.1. Error Surfaces

Our goal during training is to find values for the weights and biases in the neural network that will allow it to make effective predictions. For convenience we will group these parameters into a single vector $\mathbf{w}$, and we will optimize $\mathbf{w}$ by using a chosen error function $E(\mathbf{w})$. At this point, it is useful to have a geometrical picture of the error function, which we can view as a surface sitting over 'weight space', as shown in Figure 7.1.

First note that if we make a small step in weight space from $\mathbf{w}$ to $\mathbf{w} + \delta\mathbf{w}$ then the change in the error function is given by

$$\delta E \simeq \delta\mathbf{w}^{\mathrm{T}} \nabla E(\mathbf{w}) \tag{7.1}$$

where the vector $\nabla E(\mathbf{w})$ points in the direction of the greatest rate of increase of the error function. Provided the error $E(\mathbf{w})$ is a smooth, continuous function of $\mathbf{w}$, its smallest value will occur at a point in weight space such that the gradient of the error function vanishes, so that

$$\nabla E(\mathbf{w}) = 0 \tag{7.2}$$

as otherwise we could make a small step in the direction of $-\nabla E(\mathbf{w})$ and thereby further reduce the error. Points at which the gradient vanishes are called stationary points and may be further classified into minima, maxima, and saddle points.

Section 7.1.1