where δ_{kl} are the elements of the identity matrix and are defined by

$$\delta_{kl} = \begin{cases} 1, & \text{if } k = l, \\ 0, & \text{otherwise.} \end{cases} \tag{8.32}$$

Section 3.4

If we have individual logistic sigmoid activation functions at each output unit, then

$$\frac{\partial y_k}{\partial a_l} = \delta_{kl}\sigma'(a_l) \tag{8.33}$$

Section 3.4

whereas for softmax outputs, we have

$$\frac{\partial y_k}{\partial a_l} = \delta_{kl}y_k - y_k y_l. \tag{8.34}$$

We can summarize the procedure for calculating the Jacobian matrix as follows. Apply the input vector corresponding to the point in input space at which the Jacobian matrix is to be evaluated, and forward propagate in the usual way to obtain the states of all the hidden and output units in the network. Next, for each row k of the Jacobian matrix, corresponding to the output unit k, backpropagate using the recursive relation (8.30), starting with (8.31), (8.33) or (8.34), for all the hidden units in the network. Finally, use (8.29) for the backpropagation to the inputs. The Jacobian can also be evaluated using an alternative *forward* propagation formalism, which can be derived in an analogous way to the backpropagation approach given here. Exercise 8.5

Again, the implementation of such algorithms can be checked using numerical differentiation in the form

$$\frac{\partial y_k}{\partial x_i} = \frac{y_k(x_i + \epsilon) - y_k(x_i - \epsilon)}{2\epsilon} + \mathcal{O}(\epsilon^2), \tag{8.35}$$

which involves $2D$ forward propagation passes for a network having D inputs and therefore requires $\mathcal{O}(DW)$ steps in total.

8.1.6 The Hessian matrix

We have shown how backpropagation can be used to obtain the first derivatives of an error function with respect to the weights in the network. Backpropagation can also be used to evaluate the second derivatives of the error, which are given by

$$\frac{\partial^2 E}{\partial w_{ji}\partial w_{lk}}. \tag{8.36}$$

It is often convenient to consider all the weight and bias parameters as elements w_i of a single vector, denoted $\mathbf{w}$, in which case the second derivatives form the elements H_{ij} of the *Hessian* matrix $\mathbf{H}$:

$$H_{ij} = \frac{\partial^2 E}{\partial w_i \partial w_j} \tag{8.37}$$