If there is a total of W weights and biases in the network, then $\mathbf{w}$ and $\mathbf{b}$ have length W and $\mathbf{H}$ has dimensionality $W \times W$. From (7.3), the corresponding local approximation to the gradient is given by

$$\nabla E(\mathbf{w}) = \mathbf{b} + \mathbf{H}(\mathbf{w} - \widehat{\mathbf{w}}). \tag{7.6}$$

For points $\mathbf{w}$ that are sufficiently close to $\widehat{\mathbf{w}}$, these expressions will give reasonable approximations for the error and its gradient.

Consider the particular case of a local quadratic approximation around a point $\mathbf{w}^{\star}$ that is a minimum of the error function. In this case there is no linear term, because $\nabla E = 0$ at $\mathbf{w}^{\star}$, and (7.3) becomes

$$E(\mathbf{w}) = E(\mathbf{w}^{\star}) + \frac{1}{2}(\mathbf{w} - \mathbf{w}^{\star})^{\mathrm{T}}\mathbf{H}(\mathbf{w} - \mathbf{w}^{\star}) \tag{7.7}$$

where the Hessian $\mathbf{H}$ is evaluated at $\mathbf{w}^{\star}$. To interpret this geometrically, consider the eigenvalue equation for the Hessian matrix:

$$\mathbf{H}\mathbf{u}_i = \lambda_i \mathbf{u}_i \tag{7.8}$$

Appendix A

where the eigenvectors $\mathbf{u}_i$ form a complete orthonormal set so that

$$\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j = \delta_{ij}. \tag{7.9}$$

We now expand $(\mathbf{w} - \mathbf{w}^{\star})$ as a linear combination of the eigenvectors in the form

$$\mathbf{w} - \mathbf{w}^{\star} = \sum_i \alpha_i \mathbf{u}_i. \tag{7.10}$$

Appendix A

Exercise 7.1

This can be regarded as a transformation of the coordinate system in which the origin is translated to the point $\mathbf{w}^{\star}$ and the axes are rotated to align with the eigenvectors through the orthogonal matrix whose columns are $\{\mathbf{u}_1, \ldots, \mathbf{u}_W\}$. By substituting (7.10) into (7.7) and using (7.8) and (7.9), the error function can be written in the form

$$E(\mathbf{w}) = E(\mathbf{w}^{\star}) + \frac{1}{2}\sum_i \lambda_i \alpha_i^2. \tag{7.11}$$

Suppose we set all $\alpha_i = 0$ for $i \neq j$ and then vary α_j, corresponding to moving $\mathbf{w}$ away from $\mathbf{w}^{\star}$ in the direction of $\mathbf{u}_j$. We see from (7.11) that the error function will increase if the corresponding eigenvalue λ_j is positive and will decrease if it is negative. If all eigenvalues are positive then $\mathbf{w}^{\star}$ corresponds to a local minimum of the error function, whereas if they are all negative then $\mathbf{w}^{\star}$ corresponds to a local maximum. If we have a mix of positive and negative eigenvalues then $\mathbf{w}^{\star}$ represents a saddle point.

A matrix $\mathbf{H}$ is said to be *positive definite* if, and only if,

$$\mathbf{v}^{\mathrm{T}}\mathbf{H}\mathbf{v} > 0, \qquad \text{for all } \mathbf{v}. \tag{7.12}$$