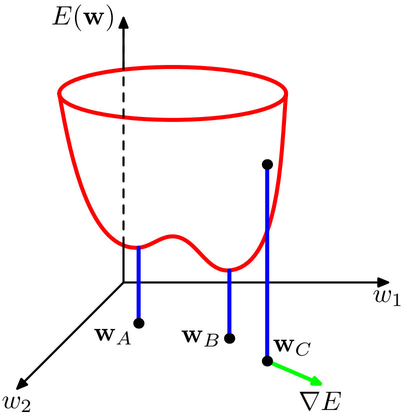

Figure 7.1 Geometrical view of the error function $E(\mathbf{w})$ as a surface sitting over weight space. Point $\mathbf{w}_A$ is a local minimum and $\mathbf{w}_B$ is the global minimum, so that $E(\mathbf{w}_\mathrm{A}) > E(\mathbf{w}_\mathrm{B})$. At any point $\mathbf{w}_C$, the local gradient of the error surface is given by the vector ∇E.

We will aim to find a vector $\mathbf{w}$ such that $E(\mathbf{w})$ takes its smallest value. However, the error function typically has a highly nonlinear dependence on the weights and bias parameters, and so there will be many points in weight space at which the gradient vanishes (or is numerically very small). Indeed, for any point $\mathbf{w}$ that is a local minimum, there will generally be other points in weight space that are equivalent minima. For instance, in a two-layer network of the kind shown in Figure 6.9, with M hidden units, each point in weight space is a member of a family of $M!\,2^M$ equivalent points.

Section 6.2.4

Furthermore, there may be multiple non-equivalent stationary points and in particular multiple non-equivalent minima. A minimum that corresponds to the smallest value of the error function across the whole of $\mathbf{w}$-space is said to be a *global minimum*. Any other minima corresponding to higher values of the error function are said to be *local minima*. The error surfaces for deep neural networks can be very complex, and it was thought that gradient-based methods might become trapped in poor local minima. In practice, this seems not to be the case, and large networks can reach solutions with similar performance under a variety of initial conditions.

Section 9.3.2

7.1.1 Local quadratic approximation

Insight into the optimization problem and into the various techniques for solving it can be obtained by considering a local quadratic approximation to the error function. The Taylor expansion of $E(\mathbf{w})$ around some point $\widehat{\mathbf{w}}$ in weight space is given by

$$E(\mathbf{w}) \simeq E(\widehat{\mathbf{w}}) + (\mathbf{w}-\widehat{\mathbf{w}})^\mathrm{T}\mathbf{b} + \frac{1}{2}(\mathbf{w}-\widehat{\mathbf{w}})^\mathrm{T}\mathbf{H}(\mathbf{w}-\widehat{\mathbf{w}}) \tag{7.3}$$

where cubic and higher terms have been omitted. Here $\mathbf{b}$ is defined to be the gradient of E evaluated at $\widehat{\mathbf{w}}$

$$\mathbf{b} \equiv \nabla E|_{\mathbf{w}=\widehat{\mathbf{w}}}\,. \tag{7.4}$$

The *Hessian* is defined to be the corresponding matrix of second derivatives

$$\mathbf{H}(\widehat{\mathbf{w}}) = \nabla\nabla E(\mathbf{w})|_{\mathbf{w}=\widehat{\mathbf{w}}}\,. \tag{7.5}$$