Algorithm 7.1: Stochastic gradient descent

```
Input: Training set of data points indexed by $n \in \{1, \ldots, N\}$
       Error function per data point $E_n(\mathbf{w})$
       Learning rate parameter $\eta$
       Initial weight vector $\mathbf{w}$
Output: Final weight vector $\mathbf{w}$
--------------------------------------------------------------
$n \leftarrow 1$
repeat
    $\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla E_n(\mathbf{w})$ // update weight vector
    $n \leftarrow n + 1 (\mathrm{mod}\ N)$ // iterate over data
until convergence
return $\mathbf{w}$
```

to be processed. To find a more efficient approach, note that error functions based on maximum likelihood for a set of independent observations comprise a sum of terms, one for each data point:

$$E(\mathbf{w}) = \sum_{n=1}^{N} E_n(\mathbf{w}). \tag{7.17}$$

The most widely used training algorithms for large data sets are based on a sequential version of gradient descent known as *stochastic gradient descent* (Bottou, 2010), or SGD, which updates the weight vector based on one data point at a time, so that

$$\mathbf{w}^{(\tau)} = \mathbf{w}^{(\tau-1)} - \eta \nabla E_n(\mathbf{w}^{(\tau-1)}). \tag{7.18}$$

This update is repeated by cycling through the data. A complete pass through the whole training set is known as a training *epoch*. This technique is also known as *online gradient descent*, especially if the data arises from a continuous stream of new data points. Stochastic gradient descent is summarized in Algorithm 7.1.

A further advantage of stochastic gradient descent, compared to batch gradient descent, is that it handles redundancy in the data much more efficiently. To see this, consider an extreme example in which we take a data set and double its size by duplicating every data point. Note that this simply multiplies the error function by a factor of 2 and so is equivalent to using the original error function, if the value of the learning rate is adjusted to compensate. Batch methods will require double the computational effort to evaluate the batch error function gradient, whereas stochastic gradient descent will be unaffected. Another property of stochastic gradient descent is the possibility of escaping from local minima, since a stationary point with respect to the error function for the whole data set will generally not be a stationary point for each data point individually.