Brief Note on Optimization Techniques

To minimize a scalar function $E(\mathbf{w})$ with respect to a vector of parameters $\mathbf{w}$, we iteratively move from $\mathbf{w}_t$ to $\mathbf{w}_{t+1}$ in step number t. The change $\Delta\mathbf{w}_t = \mathbf{w}_{t+1} - \mathbf{w}_t$ is determined by the technique used, and the initial point $\mathbf{w}_1$ is usually chosen randomly. The simplest technique is gradient descent where the gradient $\mathbf{g}_t = \nabla E$ (at $\mathbf{w} = \mathbf{w}_t$) determines the direction

$$\Delta\mathbf{w}_t = -\eta\mathbf{g}_t$$

where η is the learning rate.

Adaptive Learning Rate. In this method, gradient descent is used

$$\Delta\mathbf{w}_t = -\eta_t\mathbf{g}_t$$

except that the learning rate η_t is now allowed to change from step to step according to

$$\eta_{t+1} = \begin{cases} \alpha\eta_t & \text{if} \quad \Delta E < 0 \quad \text{at step } t \\ \beta\eta_t & \text{if} \quad \Delta E \geq 0 \quad \text{at step } t \end{cases}$$

where $\alpha > 1$ (typically $\alpha = 1.1$), and $\beta < 1$ (typically $\beta = 0.5$). ΔE is the change in E which is $E(\mathbf{w}_{t+1}) - E(\mathbf{w}_t)$. In the case where $\Delta E \geq 0$, the step is undone ($\Delta\mathbf{w}_t$ is set to zero, hence $\mathbf{w}_{t+1} = \mathbf{w}_t$), then in step $t+1$, the new learning rate (which is decreased by the β factor) is then used with $\mathbf{g}_{t+1}$ (which is the already calculated gradient $\mathbf{g}_t$).

Momentum. Simple gradient descent is modified by adding a momentum term

$$\Delta\mathbf{w}_t = -\eta\mathbf{g}_t + \mu\Delta\mathbf{w}_{t-1}$$

where $0 \leq \mu < 1$. For the initial step, $\Delta\mathbf{w}_0$ is assumed zero.

Line Search. This is a general method for deciding how far to go in a certain direction, where the direction has been determined by some technique. The idea is to keep going until a minimum has been reached in that direction. If the direction is the negative of the gradient and line search is used, the method is sometimes called *steepest descent*. With line search, no predetermined learning rate is used. We will describe a simple method for line search called binary search.

Let $\mathbf{d}$ be the search direction starting from the point $\mathbf{w}$. We evaluate E at a sequence of points $\mathbf{w}_0 = \mathbf{w}$ and, recursively, $\mathbf{w}_{n+1} = \mathbf{w}_n + 2^n\epsilon\mathbf{d}$ (doubling the step in each move). It is assumed that the direction $\mathbf{d}$ and the parameter ϵ are such that $E(\mathbf{w}_1) < E(\mathbf{w}_0)$ (e.g., the direction is the negative of the gradient and ϵ is very small). The move continues from point to point until $E(\mathbf{w}_{n+1}) \geq E(\mathbf{w}_n)$. Now the triple $\mathbf{w}_{n-1}$, $\mathbf{w}_n$, and $\mathbf{w}_{n+1}$ are taken as the initial range where the minimum occurs.

The range is then narrowed down iteratively until it is so small that one can simply take the middle point as a good approximation of the minimum. We will describe a method for exponentially narrowing down the range. Evaluate E at the midpoint between $\mathbf{w}_{n-1}$ and $\mathbf{w}_n$, and at the midpoint between $\mathbf{w}_n$ and $\mathbf{w}_{n+1}$. The point where E is the smallest (among the five points) together with its two neighbors are taken as the new triple, and the step is repeated until the range spanning the three points is small enough. The middle point becomes the output of the line search.

Conjugate Gradient. This technique uses line search in conjunction with search directions that are modified versions of steepest descent.

$$\mathbf{d}_{t+1} = -\mathbf{g}_{t+1} + \beta_t \mathbf{d}_t$$

where $\mathbf{d}_t$ is the search direction at step t (starting from $\mathbf{w}_t$). β_t is considered zero at $t = 0$, and afterwards calculated by

$$\beta_t = \frac{\mathbf{g}'_{t+1}(\mathbf{g}_{t+1} - \mathbf{g}_t)}{\mathbf{g}'_t \mathbf{g}_t},$$

where $\mathbf{g}'$ denotes the transpose of the (column) vector $\mathbf{g}$. Periodically (every $t = \mathcal{T}$ steps), β_t is reset to zero again.