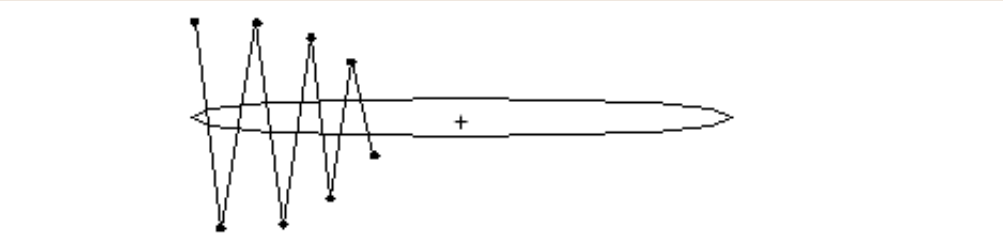
Momentum

We saw that if the cost surface is not spherical, learning can be quite slow because the learning rate must be kept small to prevent divergence along the steep curvature directions



One way to solve this is to use the inverse Hession (= correlation matrix for linear nets) as the learning rate matrix. This can be problematic because the Hessian can be a large matrix that is difficult to invert. Also, for multilayer networks, the Hessian is not constant (i.e. it changes as the weights change). Recomputing the inverse Hessian at each iteration would be prohibitively expensive and not worth the extra computation. However a much simpler approach is to use the addition of a momentum term.

$$w(t+1) = w(t) - \mu \frac{\partial E}{\partial w} + \beta(w(t) - w(t-1))$$

where w(t) is the weight at the tth iteration. Written another way

$$\Delta w(t+1) = -\mu \frac{\partial E}{\partial w} + \beta \, \Delta w(t)$$

where Dw(t) = w(t)-w(t-1). Thus, the amount you change the weight is proportional to the negative gradient plus the previous weight change.
b is called the momentum parameter. and must satisfy 0 <= b < 1.

Momentum Example

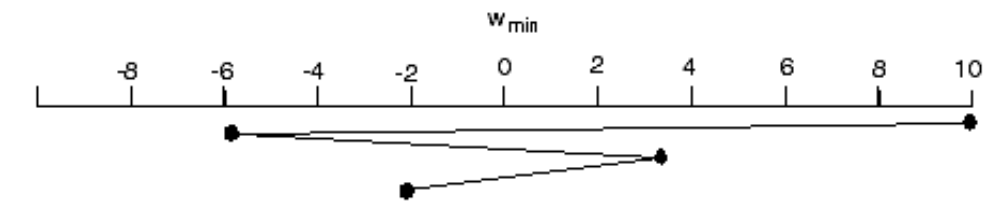
Consider the oscillatory behavior shown above. The gradient changes sign at each step. By adding in a small amount of the previous weight change, we can lessen the oscillations. Suppose m = .8, w(0)=10
$E = w^2 \Rightarrow w_{min} = 0$ and dE/dx = 2w

No Momentum b = 0:

t = 0: $Dw(1) = -.8 \frac{\partial E(1)}{\partial w} = -.8\ (20) = -16$, w(1) = 10-16 = -6

t = 2: Dw(1) = -.8 $\frac{\partial E(2)}{\partial w}$ = -.8 (-12) = 9.6, w(2) = -6+9.6 = 3.6

t = 3: Dw(1) = -.8 $\frac{\partial E(3)}{\partial w}$ = - .8(7.2) = -5.76, w(2) = 3.6 - 5.76 = -2.16

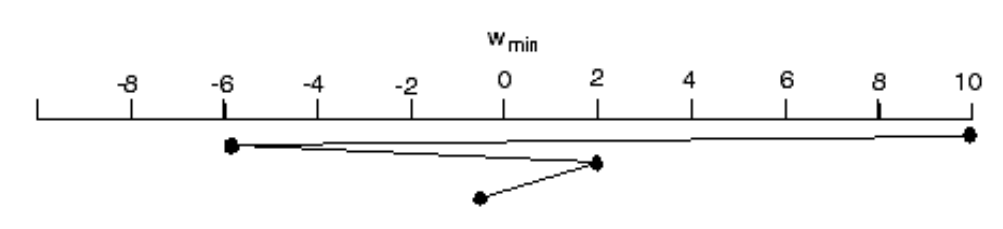


With Momentum b = .1:

t = 0: Dw(1) = -.8 $\frac{\partial E(1)}{\partial w}$ + b Dw(0) = -.8 (20) + .1*0 = -16, w(1) = 10-16 = -6

t = 2: Dw(1) = -.8 $\frac{\partial E(2)}{\partial w}$ + b Dw(1) = -.8 (-12) + .1*(-16) = 8, w(2) = -6+8 = 2

t = 3: Dw(1) = -.8 $\frac{\partial E(3)}{\partial w}$ + b Dw(2) = - .8(4) + .1*(8) = -2.4, w(2) = 2-2.4 = -.4



[Top] [Next: DeltaBarDelta] [Back to the first page]