Delta-Bar-Delta (Jacobs)

Since the cost surface for multi-layer networks can be complex, choosing a learning rate can be difficult. What works in one location of the cost surface may not work well in another location. Delta-Bar-Delta is a heuristic algorithm for modifying the learning rate as training progresses:

- Each weight has its own learning rate.
- For each weight: the gradient at the current timestep is compared with the gradient at the previous step (actually, previous gradients are averaged)
- If the gradient is in the same direction the learning rate is increased
- If the gradient is in the opposite direction the learning rate is decreased
- Should be used with batch only.

Let

$g_{ij}(t)$ = gradient of E wrt w_{ij} at time t

then define

$$\bar{g}_{ij}(t) = (1 - \beta)\, g_{ij}(t) + \beta\, \bar{g}_{ij}(t - 1) \qquad \text{where} \qquad 0 < \beta < 1$$

Then the learning rate μ_{ij} for weight w_{ij} at time t+1 is given by

$$\mu_{ij}(t + 1) = \begin{cases} \mu_{ij}(t) + \kappa & \bar{g}_{ij}(t - 1)g_{ij}(t) > 0 \\ (1 - \gamma)\mu_{ij}(t) & \bar{g}_{ij}(t - 1)g_{ij}(t) < 0 \\ \mu_{ij}(t) & \text{otherwise} \end{cases}$$

where β, γ, and κ are chosen by the hand.

Downsides:

- Knowing how to choose the parameters β, γ, and κ is not easy.
- Doesn't work for online.