Standard gradient descend: We wish to find $w^* = \arg\min L(w)$. Assume L is differentiable and convex, in a classic GD approach with fixed learning rate η we update w_t accordingly

$$w_{t+1} = w_t - \eta \nabla L(w_t) \equiv w_t - \eta g_t \tag{1}$$

We can look at picking a learning rate as an online learning problem. At each step the online meta-learner needs to pick $\eta_t \in (0, B)$, update w_t as $w_{t+1} = w_t - \eta_t g_t$ and suffer a loss according to the function $\ell_t(\eta) = L(w_t - \eta g_t)$.

If we perform gradient descent on η_t, with a fixed meta-learning rate α, we get $\eta_{t+1} = \eta_t - \alpha \frac{d\ell_t}{d\eta}|_{\eta_t}$.

Taking the derivative $\frac{d\ell_t}{d\eta}|_{\eta_t}$ it is not hard to see from the chain rule that

$$\frac{d\ell_t}{d\eta}|_{\eta_t} = -\langle g_{t+1}, g_t \rangle = -\langle \nabla L(w_{t+1}), \nabla L(w_t) \rangle \tag{2}$$

The intuition behind this gradient is easy - if we continue in a similar direction then we increase the learning rate, if we backtrack then we decrease. Problem is that the derivative can very by the magnitude of the derivative, for example in a platou the change will be minimal. Another problem is that the algorithm in not scale invariant anymore - it will behave differently for $L'(w) = \lambda L(w)$ unlike gradient descend.

One possibility is to have $w_{t+1} = w_t - f(\eta) g_t$ where f is some function like the sigmoid for example. In this case we get

$$\frac{d\ell}{d\eta}|_{\eta_t} = -\langle g_t, g_{t+1} \rangle \cdot f'(\eta_t) \tag{3}$$

And this can help bound the learning rate into a reasonable domain, but will not help with scale-invariance.