Example: SGD on Quadratics

$$R(\theta) = \frac{1}{N}\sum_{i=1}^{N} R_i(\theta), \qquad R_i(\theta) = \frac{1}{2}\left(\theta - \theta^{(i)}\right)^2.$$

$$\frac{1}{N}\sum_{i=1}^{N}\theta^{(i)} = 0, \quad \frac{1}{N}\sum_{i=1}^{N}\left[\theta^{(i)}\right]^2 = 1$$

$$R(\theta) = \frac{1}{2N}\sum_{i=1}^{N}\left(\theta - \theta^{(i)}\right)^2 = \frac{1}{2}\theta^2 + \frac{1}{2}$$

GD iterates: $\theta_k = (1-\varepsilon)^k \theta_0$

What about SGD?

$$\nabla R_i(\theta) = \theta - \theta^{(i)}$$

uniform RV in $\{1, 2, \ldots, N\}$

SGD iterates:
$$\theta_{k+1} = \theta_k - \varepsilon\left(\theta_k - \theta^{(\gamma_k)}\right)$$
$$= (1-\varepsilon)\theta_k + \varepsilon\,\theta^{(\gamma_k)}.$$

Then,
$$\theta_k = (1-\varepsilon)\theta_{k-1} + \varepsilon\,\theta^{(\gamma_{k-1})}$$
$$= (1-\varepsilon)\left[(1-\varepsilon)\theta_{k-2} + \varepsilon\,\theta^{(\gamma_{k-2})}\right] + \varepsilon\,\theta^{(\gamma_{k-1})}$$
$$= \quad \vdots$$

$$\theta_k = \underbrace{(1-\varepsilon)^k\theta_0}_{\text{same as GD}} + \underbrace{\varepsilon\sum_{j=1}^{k}(1-\varepsilon)^{j-1}\,\theta^{(\gamma_{k-j})}}_{\text{random}}$$

$$\because \quad \mathbb{E}\,\theta^{(\gamma_{k-j})} = \frac{1}{N}\sum_{i=1}^{N}\theta^{(i)} = 0$$

$$\therefore \quad \mathbb{E}\,\theta_k = (1-\varepsilon)^k\theta_0$$

What about second moments?

$$\mathbb{E}\left[\theta_k^2\right] = (1-\varepsilon)^{2k}\theta_0^2 + \overbrace{2\varepsilon(1-\varepsilon)^k\theta_0\sum_{j=1}^{k}(1-\varepsilon)^{j-1}\,\theta^{(\gamma_{k-j})}}^{=0}$$
$$+\ \varepsilon^2\sum_{j,l=1}^{k}(1-\varepsilon)^{j+l-2}\,\theta^{(\gamma_{k-j})}\theta^{(\gamma_{k-l})}$$

$$+ \varepsilon^2 \sum_{j,l=1}^{k} (1-\varepsilon)^{j+l-2} \theta^{(\gamma_{k-j})} \theta^{(\gamma_{k-l})}$$

$$= (1-\varepsilon)^{2k} \theta_0^2 + \varepsilon^2 \sum_{j,l=1}^{k} (1-\varepsilon)^{j+l-2} \mathbb{E}\left[\theta^{(\gamma_{k-j})} \theta^{(\gamma_{k-l})}\right]$$

$$\left[\text{IID, } \mathbb{E}\left[\theta^{(\gamma_{k-j})} \theta^{(\gamma_{k-l})}\right] = \begin{cases} 0 & j \neq l \\ 1 & j = l \end{cases}\right]$$

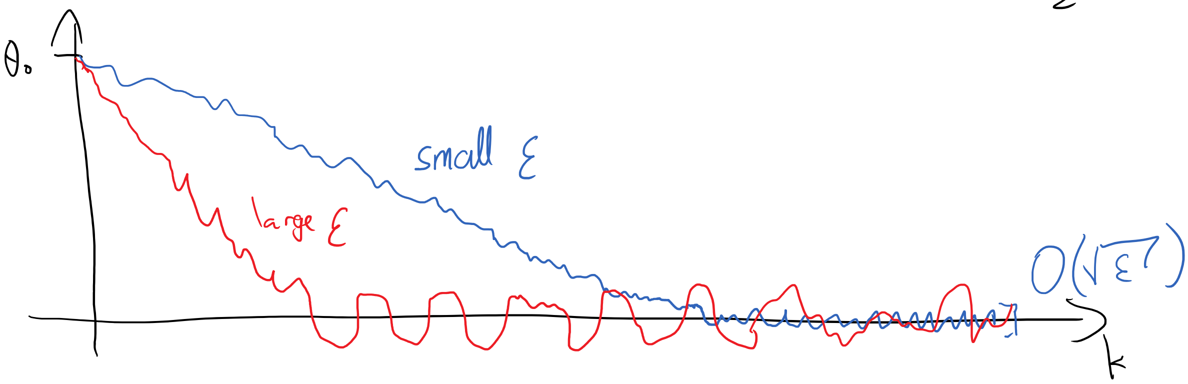
$$= (1-\varepsilon)^{2k} \theta_0^2 + \varepsilon^2 \sum_{j=1}^{k} (1-\varepsilon)^{2j-2} \longrightarrow \text{Geometric series}$$

$$\mathbb{E}[\theta_k^2] = (1-\varepsilon)^{2k} \theta_0^2 + \frac{\varepsilon}{2-\varepsilon}\left[1-(1-\varepsilon)^{2k}\right]$$

$$\mathbb{E}[\theta_k] = (1-\varepsilon)^k \theta_0$$

$$\mathrm{Var}[\theta_k] = \frac{\varepsilon}{2-\varepsilon}\left[1-(1-\varepsilon)^{2k}\right] \xrightarrow[k\to\infty]{0<\varepsilon<1} \frac{\varepsilon}{2-\varepsilon} > 0$$

$$\approx \frac{\varepsilon}{2} \quad (\text{if } \varepsilon \text{ small})$$



Example (backprop)

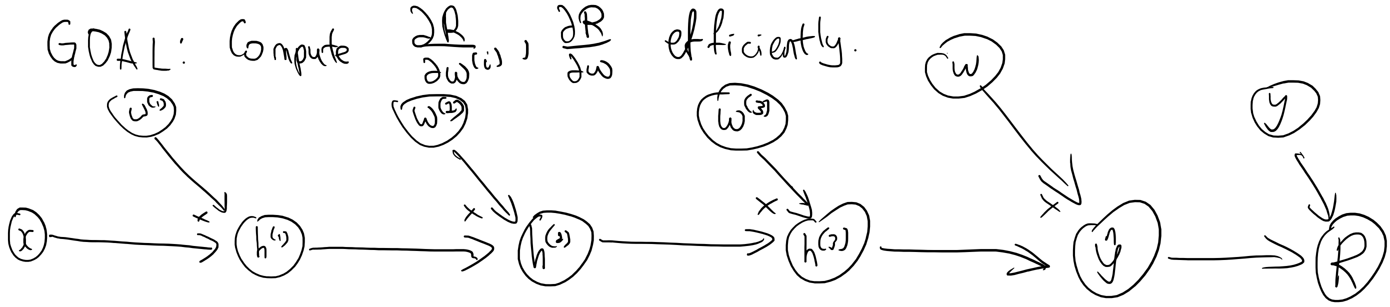
- $h^{(1)} = w^{(1)} \cdot x$
- $h^{(2)} = w^{(2)} \cdot h^{(1)}$
- $h^{(3)} = w^{(3)} \cdot h^{(2)}$
- $\hat{y} = w \cdot h^{(3)}$

$$\theta = \{w^{(1)}, w^{(2)}, w^{(3)}, w\}$$

$$\text{Data} = \{x, y\}$$

$$R(\theta) = L(\hat{y}, y)$$

GOAL: Compute $\frac{\partial R}{\partial w^{(i)}}, \frac{\partial R}{\partial w}$ efficiently.



Step 1: Forward Propagation

Given (x, y), $\{W^{(i)}\}$, w

Compute: $h^{(1)} = W^{(1)} \cdot x$

$h^{(2)} = W^{(2)} \cdot h^{(1)}$

$h^{(3)} = W^{(3)} \cdot h^{(2)}$

$\hat{y} = w \cdot h^{(3)}$

Store: $h^{(1)}, h^{(2)}, h^{(3)}, \hat{y}$

Step 2: Backpropagation

$$\frac{\partial R}{\partial \hat{y}} = \frac{\partial L(y, \hat{y})}{\partial \hat{y}} \xrightarrow{\text{Store}} \hat{p}$$

Then,

$$\frac{\partial R}{\partial h^{(3)}} = \frac{\partial R}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial h^{(3)}} = \hat{p} \cdot w \xrightarrow{\text{Store}} p^{(3)}$$

$$\frac{\partial R}{\partial h^{(2)}} = \frac{\partial R}{\partial h^{(3)}} \times \frac{\partial h^{(3)}}{\partial h^{(2)}} = p^{(3)} \cdot W^{(3)} \xrightarrow{\text{Store}} p^{(2)}$$

$$\frac{\partial R}{\partial h^{(1)}} = \frac{\partial R}{\partial h^{(2)}} \times \frac{\partial h^{(2)}}{\partial h^{(1)}} = p^{(2)} \cdot W^{(2)} \xrightarrow{\text{store}} p^{(1)}$$

Step 3: Compute gradients

$$\frac{\partial R}{\partial w} = \frac{dR}{d\hat{y}} \cdot \frac{d\hat{y}}{dw} = \hat{p} \times h^{(3)}$$

$$\frac{dR}{dW^{(3)}} = \frac{dR}{dh^{(3)}} \times \frac{dh^{(3)}}{dW^{(3)}} = p^{(3)} \times h^{(2)}$$

$$\frac{dR}{dW^{(2)}} = \frac{dR}{dh^{(2)}} \times \frac{dh^{(2)}}{dW^{(2)}} = p^{(2)} \times h^{(1)}$$

$$\frac{dR}{dW^{(1)}} = \frac{dR}{dh^{(1)}} \times \frac{dh^{(1)}}{dW^{(1)}} = p^{(1)} \times x$$