1) Gradient descent and backpropagation in a simple neural network

1. $z(x_1) = 1\times 2 + 1\times 1 + 0\times 1 = 3$ $\qquad z(x_2) = 1\times(-3) + 1\times 2 + 0\times 1 = -1$

$h(x_1) = g(z(x_1)) = g(3) = 3$ $\qquad h(x_2) = g(-1) = 0$

$\hat{y}(x_1) = 1\times h(x_1) + 0\times 1 = 1\times 3 = 3$ $\qquad \hat{y}(x_2) = 1\times 0 + 0\times 1 = 0$

2. $\mathcal{L}(D) = \frac{1}{2}\left(L(x_1, y_1) + L(x_2, y_2)\right)$

$L(x_1, y_1) = (\hat{y}(x_1) - y_1)^2 = (3-1.3)^2 = 2.89$

$L(x_2, y_2) = (0-1.9)^2 = 3.61$

$\mathcal{L}(D) = \frac{1}{2}(2.89 + 3.61) = 3.25$

3. L is a function of $\hat{y}(x)$, etc.

$\hat{y}$ " " " " $h(x)$, w_3, w_4, etc.

$h(x)$ " " " " $z(x)$, etc.

$z(x)$ " " " " w_0, w_1, w_2, etc.

So, L is function of z and z is function of w_i $i\in\{0,1,2\}$. Therefore:

$$\frac{\partial L}{\partial w_i} = \frac{\partial L}{\partial z}\cdot\frac{\partial z}{\partial w_i} \quad i\in\{0,1,2\} \qquad (1)$$

Scanned by CamScanner

and, we had $\delta_h(x,y) = \frac{\partial L}{\partial z}$ (2)

(1) and (2) $\Rightarrow \frac{\partial L}{\partial w_i} = \delta_h(x,y) \cdot \frac{\partial Z}{\partial w_i} \quad i \in \{0,1,2\}$

$\frac{\partial Z}{\partial w_0} = x[0] \qquad \frac{\partial Z}{\partial w_1} = x[1] \qquad \frac{\partial Z}{\partial w_2} = 1$

$\boxed{\frac{\partial L}{\partial w_0} = \delta_h(x,y) \cdot x[0]} \quad \boxed{\frac{\partial L}{\partial w_1} = \delta_h(x,y) \cdot x[1]} \quad \boxed{\frac{\partial L}{\partial w_2} = \delta_h(x,y)}$

L is a function of $\hat{y}$ and $\hat{y}$ is a function of w_i $i \in \{3,4\}$.

Therefore: $\frac{\partial L}{\partial w_i} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_i} \quad i \in \{3,4\}$ (3)

and, we had $\delta_{\hat{y}}(x,y) = \frac{\partial L}{\partial \hat{y}}$ (4)

(3) and (4) $\Rightarrow \frac{\partial L}{\partial w_i} = \delta_{\hat{y}}(x,y) \cdot \frac{\partial \hat{y}}{\partial w_i} \quad i \in \{3,4\}$

$\frac{\partial \hat{y}}{\partial w_3} = h(x) \qquad \frac{\partial \hat{y}}{\partial w_4} = 1$

$\boxed{\frac{\partial L}{\partial w_3} = \delta_{\hat{y}}(x,y) \cdot h(x)} \qquad \boxed{\frac{\partial L}{\partial w_4} = \delta_{\hat{y}}(x,y)}$

4- $\nabla_\theta \mathcal{L}(D) = \frac{1}{2}\left(\nabla_\theta L(x_1,y_1) + \nabla_\theta L(x_2,y_2)\right)$

$\nabla_\theta L(x_1,y_1) = \left(\delta_h(x_1,y_1) \times 2,\ \delta_h(x_1,y_1) \times 1,\ \delta_h(x_1,y_1),\ \delta_{\hat{y}}(x_1,y_1) \cdot h(x_1),\ \delta_{\hat{y}}(x,y_1)\right)$

$\nabla_\theta L(x_2,y_2) = \left(\delta_h(x_2,y_2) \times (-3),\ \delta_h(x_2,y_2) \times 2,\ \delta_h(x_2,y_2),\ \delta_{\hat{y}}(x_2,y_2) \cdot h(x_2),\ \delta_{\hat{y}}(x_2,y_2)\right)$

$$\delta_{\hat{y}}(x,y) = \frac{\partial L(x,y)}{\partial \hat{y}} = 2(\hat{y}(x) - y)$$

$$\delta_h(x,y) = \delta_{\hat{y}}(x,y) \cdot \omega_3 \cdot g'(z(x)) = \begin{cases} \delta_{\hat{y}}(x,y) \cdot \omega_3 & z(x) > 0 \\ 0 & z(x) \leq 0 \end{cases}$$

$$\delta_{\hat{y}}(x_1, y_1) = 2(\hat{y}(x_1) - y_1) = 2(3 - 1.3) = 3.4$$

$$\delta_{\hat{y}}(x_2, y_2) = 2(\hat{y}(x_2) - y_2) = 2(0 - 1.9) = -3.8$$

$z(x_1) = 3 > 0$

$$\delta_h(x_1, y_1) = \delta_{\hat{y}}(x_1, y_1) \cdot \omega_3 = 3.4 \times 1 = 3.4$$

$z(x_2) = -1 < 0$

$$\delta_h(x_2, y_2) = 0$$

Therefore:

$$\nabla_\theta L(x_1, y_1) = (3.4 \times 2, 3.4 \times 1, 3.4, 3.4 \times h(x_1), 3.4)$$

$$= (6.8, 3.4, 3.4, 10.2, 3.4)$$

$$\nabla_\theta L(x_2, y_2) = (0 \times (-3), 0 \times 2, 0, (-3.8) \times h(x_2), -3.8)$$

$$= (0, 0, 0, 0, -3.8)$$

$$\nabla_\theta \mathcal{L}(D) = (3.4, 1.7, 1.7, 5.1, -0.2)$$

5- $\theta = (1, 1, 0, 1, 0)$

$- 0.01 \cdot (3.4, 1.7, 1.7, 5.1, -0.2)$

$= (0.966, 0.983, -0.017, 0.949, 0.002)$

6- $Z(x_1) = 0.966 \times 2 + 0.983 \times 1 + (-0.017) \times 1 = 2.898$

$h(x_1) = g(2.898) = 2.898$

$\hat{y}(x_1) = 0.949 \times 2.898 + 0.002 \times 1 = 2.752$

$Z(x_2) = 0.966 \times (-3) + 0.983 \times 2 + (-0.017) \times 1 = -0.949$

$h(x_2) = g(-0.949) = 0$

$\hat{y}(x_2) = 0.949 \times 0 + 0.002 \times 1 = 0.002$

$L(x_1, y_1) = (2.752 - 1.3)^2 = 2.108$

$L(x_2, y_2) = (0.002 - 1.9)^2 = 3.602$

$\mathcal{L}(D) = \frac{1}{2}(2.108 + 3.602) = 2.855$

After updating θ, the loss of the network on D is less.

7- η controls the amount of changes on θ_l base on the gradient of the network's loss function). Larger η can help to faster convergence to the optimum θ; however, large η can also lead to a failure for convergence. Generally speaking, relatively large η for the first iterations of learning and relatively small η for the final " " " can be appropriate for the process of learning.