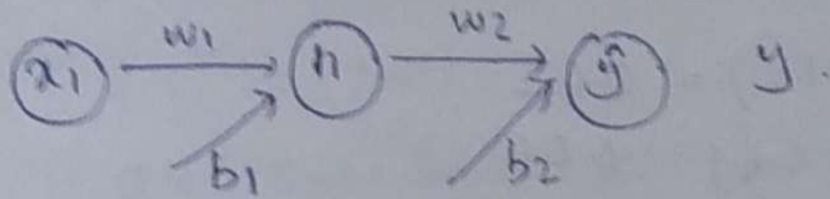
Assignment 02

①



activation = ReLU(z) = max$(0, z)$

loss function. $\frac{1}{2}(y-\hat{y})^2$

$x_1 = 2, y = 5$	$\partial L/\partial w_1 = -1.2$	$\partial L/\partial w_1 = -0.8$
$w_1 = 0.5, b_1 = 0$	$\partial L/\partial b_1 = -0.6$	$\partial L/\partial b_1 = -0.4$
$w_2 = 0.3, b_2 = 0$	$\partial L/\partial w_2 = -2$	$\partial L/\partial w_2 = -1.5$
	$\partial L/\partial b_2 = -1$	$\partial L/\partial b_2 = -0.5$

Part A

① $h = \text{ReLU}(w_1 x_1 + b_1)$

$= \max(0, 2\cdot(0.5) + 0) = \boxed{1}$

$\hat{y} = \text{ReLU}(h w_2 + b_2)$

$= \max(0, 1\cdot(0.3) + 0) = \boxed{0.3}$

② Loss

$L = \frac{1}{2}(y-\hat{y})^2$.

$= \frac{1}{2}(5-0.3)^2 = 4.7^2/2 = \boxed{11.045}$

① SGD

$$\theta_t = \theta_{t-1} - \eta g_t \quad (\eta = 0.1)$$

Iteration 1

$w_1 = 0.5 - 0.1(-1.2) = 0.62$

$b_1 = 0 - 0.1(-0.6) = 0.06$

$w_2 = 0.3 - 0.1(-2) = 0.50$

$b_2 = 0 - 0.1(-1) = 0.10$

Iteration 2

$w_1 = 0.62 - 0.1(-0.8) = 0.70$

$b_1 = 0.06 - 0.1(-0.4) = 0.10$

$w_2 = 0.5 - 0.1(-15) = 0.65$

$b_2 = 0.1 - 0.1(-0.5) = 0.15$

② Momentum

$$v_t = \beta v_{t-1} + \eta g_t, \quad \theta_t = \theta_{t-1} - v_t \quad (v_0 = 0, \beta = 0.9, \eta = 0.1)$$

Iteration 1

$v_1 = (0.9)(0) + (0.1)(g_1) = (0.1)g_1$

$w_1 = 0.5 - 0.1(-1.2) = 0.62$

$b_1 = 0 - 0.1(-0.6) = 0.06$

$w_2 = 0.3 - 0.1(-2) = 0.50$

$b_2 = 0 - 0.1(-1) = 0.10$

Iteration 2

$$V_2 = 0.9 V_1 + 0.1 g_2 = (0.9)(0.1) g_1 + (0.1)\cdot(g_2) = (0.1)(0.9 g_1 + g_2)$$

$$V^2_{w_1} = 0.9(-1.2) + \cancel{0.1}(-0.8) = (0.1)(-1.88) = -0.188$$

$$V^2_{b_1} = 0.9(-0.6) + \cancel{0.1}(-0.4) = (0.1)(-0.94) = -0.094$$

$$V^2_{w_2} = 0.9(-2) + \cancel{0.1}(-1.5) = (0.1)(-3.30) = -0.33$$

$$V^2_{b_2} = 0.9(-1) + \cancel{0.1}(-0.5) = (0.1)(-1.40) = -0.014$$

$$w_1 = 0.62 + 0.188 = 0.808$$

$$b_1 = 0.06 + 0.094 = 0.154$$

$$w_2 = 0.5 + 0.33 = 0.83$$

$$b_2 = 0.1 + 0.14 = 0.24$$

③ RMS Prop

$$s_t = \beta s_{t-1} + (1-\beta) g_t^2 \qquad \theta_t = \theta_{t-1} - \frac{\eta}{\sqrt{s_t + \epsilon}} g_t$$

$s_0 = 0$, $\eta = 0.1$, $\beta = 0.9$, $\epsilon = 10^{-8}$

Iteration 1

$$s' = \beta(0) + (1-0.9)(g_1)^2 = 0.1 (g_1)^2$$

$$s'_{w_1} = 0.1(1.2)^2 = 0.144$$

$$s'_{b_1} = 0.1(0.6)^2 = 0.036$$

$$s'_{w_2} = 0.1(2)^2 = 0.4$$

$$s'_{b_2} = 0.1(1)^2 = 0.1$$

$$w_1 = 0.5 - \frac{0.1(-1.2)}{\sqrt{0.144 + 10^{-8}}} = 0.5 + \frac{(0.1)(1.2)}{0.3794} = 0.5 + 0.3162$$

$$= 0.8162,$$

$$b_1 = 0 - \frac{(0.1)(-0.6)}{\cancel{0.1897}\ \sqrt{0.036 + 10^{-8}}} = 0 + \frac{(0.1)(0.6)}{0.1897} = 0 + 0.3162$$

$$= 0.3162 //$$

$$w_2 = 0.3 - \frac{(0.1)(-2)}{\sqrt{0.4 + 10^{-8}}} = 0.3 + \frac{(2)(0.1)}{0.6324} = 0.3 + 0.3162$$

$$= 0.6162 //$$

$$b_2 = 0 - \frac{(0.1)(-1)}{\sqrt{0.1 + 10^{-8}}} = 0 + \frac{(0.1)(1)}{0.3162} = 0 + 0.3162$$

$$= 0.3162 //$$

Iteration 2

$$S^2 = (0.9)(S^1) + (0.1)(g_2)^2$$

$$S_{w1} = (0.9)(0.144) + (0.1)(0.8^2)$$

$$= 0.1296 + 0.064 = 0.44$$

$$S_{b1} = (0.9)(0.036) + (0.1)(0.4)^2$$

$$= 0.0324 + 0.016 = 0.484$$

$$S_{w2} = (0.9)(0.4) + (0.1)(1.5^2)$$

$$= 0.36 + 0.225 = 0.0585$$

$$S_{b2} = (0.9)(0.1) + (0.1)(0.5^2)$$

$$= 0.09 + 0.025 = 0.115$$

$$w_1^2 = 0.8162 - \frac{(0.1)(-0.8)}{\sqrt{0.1936}} = 0.8162 + \frac{(0.1)(0.8)}{0.44}$$

$$= 0.8162 + 0.1818 = 0.9980$$

$$b_1 = 0.3162 - \frac{(0.1)(-0.4)}{\sqrt{0.484}} = 0.3162 + \frac{(0.1)(0.4)}{0.22}$$

$$= 0.3162 + 0.1818 = 0.4980$$

$$w2 = 0.6162 - \frac{(0.1)(-1.5)}{\sqrt{0.585}} = 0.6162 + \frac{(0.1)(1.5)}{0.7645}$$

$$= 0.6162 + 0.1961 = 0.8123$$

$$b2 = 0.3162 - \frac{(0.1)(-0.5)}{\sqrt{0.115}} = 0.3162 + \frac{(0.1)(0.5)}{0.3391}$$

$$= 0.3162 + 0.1474 = 0.4636$$

① Adam

$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$

$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$

$\hat{m}_t = \frac{m_t}{(1-\beta_1^t)}$ $\quad \hat{v}_t = \frac{v_t}{1-\beta_2^t}$

$\beta_1 = 0.9$ $\quad m_0 = 0$

$\beta_2 = 0.99$ $\quad v_0 = 0$

$\epsilon = 10^{-18}$

$\theta_t = \theta_{t-1} - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}$

Iteration 1

$m^1 = \beta_1(0) + (1-0.9) g_1 = 0.1 g$

$m^1_{w1} = 0.1(-1.2) = -0.12$

$m^1_{b1} = 0.1(-0.66) = -0.06$

$m^1_{w2} = 0.1(-2) = -0.2$

$m^1_{b2} = 0.1(-1) = -0.1$ //

$v^1 = \beta_2(0) + (1-0.99)(g^2) = (0.01) g^2$

$v_{w1} = 0.01 * 1.44 = 0.0144$

$v_{b1} = 0.01 \times 0.36 = 0.0036$

$v_{w2} = 0.01 * 4 = 0.04$

$v_{b2} = 0.01 * 1 = 0.01$

$$\hat{m} = m/1-0.9 = m/0.1 = 10 * m$$

$$\hat{m}_{w1} = -0.12 * 10 = -1.2$$

$$\hat{m}_{b1} = -0.06 * 10 = -0.6$$

$$\hat{m}_{w2} = -0.2 * 10 = -2$$

$$\hat{m}_{b2} = -0.1 * 10 = -1$$

$$\hat{v} = v/1-0.99 = 100 * v$$

$$\hat{v}_{w1} = 100 * 0.0144 = 1.44$$

$$\hat{v}_{b1} = 100 * 0.0036 = 0.36$$

$$\hat{v}_{w2} = 100 * 0.04 = 4$$

$$\hat{v}_{b2} = 100 * 0.01 = 1$$

$$w_1 = 0.5 - \frac{(0.1)(-1.2)}{\sqrt{1.44} + \epsilon} = 0.5 + 0.1 = 0.6$$

$$b_1 = 0 - \frac{(0.1)(-0.6)}{\sqrt{0.36} + \epsilon} = 0 + 0.1 = 0.1$$

$$w_2 = 0.3 - \frac{(0.1)(-2.0)}{\sqrt{4} + \epsilon} = 0.3 + 0.1 = 0.4$$

$$b_2 = 0 - \frac{(0.1)(-1)}{\sqrt{1} + \epsilon} = 0 + 0.1 = 0.1$$

- Iteration 2

$m^2 = 0.9\, m^1 + 0.1\, g^2$

$m^2_{w1} = 0.9(\cancel{0}\,{-0.12}) + 0.1(-0.8) = -0.188$ //

$m^2_{b1} = 0.9(-0.06) + 0.1(-0.4) = -0.094$ //

$m^2_{w2} = 0.9(-0.2) + 0.1(-1.5) = -0.33$ //

$m^2_{b2} = 0.9(-0.1) + 0.1(-0.5) = -0.14$ //

$v^2 = 0.99\, v^1 + 0.01 (g^2)^2.$

$v^2_{w1} = 0.99(0.0144) + 0.01(0.64) = \cancel{0.02068}$ 0.0206

$v^2_{b1} = 0.99(0.0036) + 0.01(0.16) = 0.0051$

$v^2_{w2} = 0.99(0.04) + 0.01(2.25) = 0.0621$

$v^2_{b2} = 0.99(0.01) + 0.01(0.25) = 0.0124$

$\hat{m} = m^1 / 1-(0.9)^2 = m^1/1-0.81 = m^1/0.19$ //

$\hat{m}_{w1} = -0.188/0.19 = -0.98$

$\hat{m}_{b1} = -0.094/0.19 = -0.49$

$\hat{m}_{w2} = -0.33/0.19 = -1.73$

$\hat{m}_{b2} = -0.14/0.19 = -0.73$

$\hat{v} = v^1 / (1-(0.99)^2) = v^1/0.0199$ //

$\hat{v}_{w1} = 0.0206 / 0.0199 = 1.03$

$\hat{v}_{b1} = 0.0051 / 0.0199 = 0.02$

$\hat{v}_{w2} = 0.0621 / 0.0199 = 0.31$

$\hat{v}_{b2} = 0.0124 / 0.0199 = 0.06$

$$\theta_t = \theta_{t-1} - \frac{0.1}{\sqrt{\hat{v}_t} + \epsilon} \hat{m}_t$$

$$w_1 = 0.6 - \frac{0.1}{\sqrt{1.03} + \epsilon} * (-0.98) = 0.6971$$

$$b_1 = 0.1 - \frac{0.1}{\sqrt{0.02} + \epsilon} * (-0.49) = 0.1971$$

$$w_2 = 0.4 - \frac{0.1}{\sqrt{0.31} + \epsilon} * (-0.73) = 0.4983$$

$$b_2 = 0.1 - \frac{0.1}{\sqrt{0.06} + \epsilon} * (-0.73) = 0.1933$$

Q more stable optimizer?

Ans: Adam is the most stable, ~~since~~ because it is the combination of Rms prop and momentum. momentum gives smooth velocity and Rmsprop enable adaptive scaling of parameters.

Comparison

sgd

sgd uses the current gradient only for updating weights. so it converges slowly, may fall into local minima.

momentum uses past gradients to smooth and accelerate. It becomes faster (but chances are there for overshooting)

Rmsprop scales learning rate per parameter. It scales down steps if ~~they are~~ gradients are larger. It is fast, but there is no chance of overshooting.

Adam combines momentum and Rmsprop. It moves with momentum, but automatically adapts steps ~~is~~ and slows down if requires. It is smoother and stable compared to other optimisers.