Q1)

1) Generated eq.

$$\left\{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, y\right\}_0, \left\{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, y\right\}_1, \ldots \left\{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, y\right\}_{399}$$

$$\Rightarrow \hat{y}_i = w^T x_i = w_1 x_{1i} + w_2 x_{2i}$$

$$J = \frac{1}{400} \sum_{i=0}^{399} \left(\hat{y}_i - y_i\right)^2$$

$$J = \frac{1}{400} \sum_{i=0}^{399} \left(w_1 x_{1i} + w_2 x_{2i} - y_i\right)^2$$

2) J v/s $\underset{[-20,20]}{w_1}$ | J v/s $\underset{[-20,20]}{w_2}$

$\Rightarrow$ J v/s $\left(\underset{[-20,20]}{w_1} \text{ and } \underset{[-20,20]}{w_2}\right)$

3) $\eta_{opt} = Hes^{-1}$ | $\eta_{opt}(w^*) = Hes(w^*)^{-1}$

$$\frac{\partial J}{\partial w_1} = \frac{2}{400} \sum_{i=0}^{399} \left(w_1 x_{1i} + w_2 x_{2i} - y_i\right) x_{1i}$$

$$\frac{\partial^2 J}{\partial w_1^2} = \frac{2}{400} \sum_{i=0}^{399} x_{1i}^2 \quad \Big| \quad \frac{\partial^2 J}{\partial w_2^2} = \frac{2}{400} \sum_{i=0}^{399} x_{2i}^2$$

$$\eta_{opt}(w_1) = \frac{400}{2\sum_{i=0}^{399} x_{1,i}^2} \quad \Big| \quad \eta_{opt}(w_2) = \frac{400}{2\sum_{i=0}^{399} x_{2,i}^2}$$

4) $$w_1 = w_1 - \eta \frac{\partial J}{\partial w_1} \quad \Big| \quad w_2 = w_2 - \eta \frac{\partial J}{\partial w_2}$$

$\eta_{opt}(w_1) = 0.09075$, $\eta_{opt}(w_2) = 0.08933$

$\eta_{opt} = \eta_{opt}(w_2) = 0.08933$ (Smaller)

$\Rightarrow$ (i) Assuming uniformly $[-4, 4]^2$ $\Rightarrow$ $x_1 \in \mathcal{U}(-4, 4)$

$x_2 \in \mathcal{U}(-4, 4)$

Input is $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$.

Q2>

① $f(x,y) = x^2 + 100\,(y - x^2)^2$

$\Rightarrow \eta = 0.0001$ or 0.001

$$\frac{\partial f}{\partial x} = 2x + 200\,(y - x^2)(-2x) = 2x - 400\,x\,(y - x^2)$$

$$\frac{\partial f}{\partial y} = 200\,(y - x^2)$$

$\Rightarrow (x_0, y_0) = (-20, 10)$

$\Rightarrow$ Update

$$x = x - \eta \frac{\partial f}{\partial x} \quad \Big| \quad y = y - \eta \frac{\partial f}{\partial y}$$

② <u>Update :-</u> <u>Polyak's momentum</u>

$$\Rightarrow \Delta x = -\eta \frac{\partial f}{\partial x} + \beta \Delta x \quad ; \quad \beta = 0.9$$

$$x = x + \Delta x.$$

$$\Rightarrow \Delta y = -\eta \frac{\partial f}{\partial y} + \beta \Delta y \quad ; \quad \beta = 0.9$$

$$y = y + \Delta y$$

③ Update :- Nesterov Accelerated Gradient

$$\Delta x = \beta \Delta x - \eta \frac{\partial f}{\partial x}\left(x + \beta \Delta x, y\right)$$

$$x = x + \Delta x$$

⇒ Updating

$$\frac{\partial f}{\partial x} = 2x - 400x(y - x) \quad \Big| \quad \frac{\partial f}{\partial y} = 200(y - x^2)$$

$$\frac{\partial^2 f}{\partial x^2} = 2 + 800x \quad \Big| \quad \frac{\partial^2 f}{\partial y^2} = 200$$

$$\eta_{opt}(x) = \frac{1}{2 + 800x} \quad \Big| \quad \eta_{opt}(y) = \frac{1}{200} = 0.005$$

$$x = 20 \Rightarrow \frac{1}{2 + 16000}$$

$$= 0.0000625$$

$$x = 1 \Rightarrow .00125$$

(3) $f(x,y) = \frac{50}{9}(x^2+y^2)^3 - \frac{209}{18}(x^2+y^2)^2 + \frac{59}{9}(x^2+y^2)$

$$\frac{\partial f}{\partial x} = \frac{50}{3}(x^2+y^2)^2(2x) \quad \frac{-209}{9}(x^2+y^2)(2x) + \frac{59}{9}(2x)$$

$$\frac{\partial f}{\partial x} = \frac{100}{3}x(x^2+y^2)^2 - \frac{418}{9}x(x^2+y^2) + \frac{118}{9}x$$

$$\frac{\partial f}{\partial y} = \frac{100}{3}y(x^2+y^2)^2 - \frac{418}{9}y(x^2+y^2) + \frac{118}{9}y$$

Q20) d) <u>Adam</u>

$\eta = 0.001$, $\rho_1 = 0.9$, $\rho_2 = 0.999$

$\delta = 10^{-8}$

$s = 0$, $r = 0$

first moment of gradient — second moment of gradient

$t = 0$

time step.

while stopping criteria:

scalar → $g = \frac{1}{m} \nabla_\theta \sum_i L\left(f(x^{(i)}; \theta), y^{(i)}\right)$ for each param

$t = t + 1$

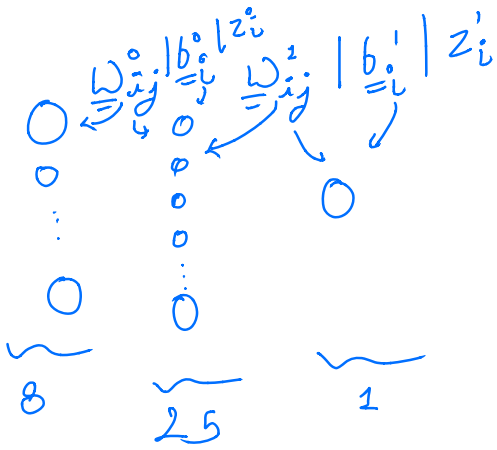
biased momentum update
$$\begin{cases} s = \rho_1 s + (1 - \rho_1) g \\ r = \rho_2 r + (1 - \rho_2) g^2 \end{cases}$$

bias correction
$$\begin{cases} \hat{s} = \dfrac{s}{1 - \rho_1^t} \\ \hat{r} = \dfrac{r}{1 - \rho_2^t} \end{cases}$$

Param update
$$\begin{cases} \Delta\theta = -\eta \dfrac{\hat{s}}{\sqrt{\hat{r}} + \delta} \end{cases}$$

$$\theta = \theta + \Delta\theta$$

5)



tanh a) $f(x) = \tanh(x) = \dfrac{2}{1+e^{-2x}} - 1$ | $f'(x) = 1 - f(x)^2$ (element-wise)

Relu b) $f(x) = \begin{cases} 0, & x < 0 \\ x, & x \geq 0 \end{cases}$ | $f'(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases}$

$\Rightarrow$

$$z^0_{25\times1} = W^0_{25\times8}\, x_{(8\times1)} + b^0_{25\times1}$$

$$y^0_{25\times1} = \tanh(z^0)_{25\times1}$$

$$z^1_{1\times1} = W^1_{1\times25}\, y^0_{25\times1} + b^1_{1\times1}$$

Training eg.

$$= \text{MSE} \quad \frac{1}{N}\sum_{i=1}^{N} (z^1_i - y_i)^2$$

$$L(f(x;\theta), y)$$

Training Sample

(x, y)

$(x_1, y_1), (x_2, y_2), \ldots (x_m, y_m)$

8 d vector (IR8)

$$\Rightarrow \frac{\partial L}{\partial z^1}$$

Matrix Computation

$$(x_0, y_0), (x_1, y_1), (x_2, y_2), \dots, (x_{N-1}, y_{N-1}) \; ; \; x_i \in \mathbb{R}^8, \; y_i \in \mathbb{R}$$

$$X = [x_0, x_1, x_2, \dots, x_{N-1}]_{8 \times N}$$

$$Y = [y_0, y_1, y_2, \dots, y_{N-1}]_{1 \times N}$$

Forward pass

$$\underset{[25 \times N]}{Z^0} = \underset{[25 \times 8]}{W^0} \; \underset{[8 \times N]}{X} + \underset{[25 \times 1]}{b^0}$$

Broadcasting $\rightarrow [25 \times N]$

$$\underset{(25 \times N)}{Y^0} = \tanh(Z^0)$$

$$\underset{[1 \times N]}{\hat{Y}} = \underset{[1 \times 25]}{W^1} \; \underset{[25 \times N]}{Y^0} + \underset{[1 \times 1]}{b^1}$$

$$\underset{[1 \times 1]}{L(f(x;\theta), Y)} = \frac{1}{N} \| \underset{[1 \times N]}{Y - \hat{Y}} \|_2^2 = \frac{1}{N} (Y - \hat{Y})^T (Y - \hat{Y})$$

$$\frac{\partial L}{\partial \hat{Y}} = \frac{2}{N} (\hat{Y} - Y)_{[1 \times N]}$$

$$\underset{[1 \times 25]}{\frac{\partial L}{\partial W^1}} = \underset{[1 \times N]}{\frac{\partial L}{\partial \hat{Y}}} \cdot \underset{[N \times 25]}{\frac{\partial \hat{Y}}{\partial W^1}} \quad \Bigg| \quad \frac{\partial \hat{Y}}{\partial W^1} = \underset{[N \times 25]}{Y^{0\,T}}$$

$$= \frac{2}{N} \underset{[1 \times N]}{(Y - \hat{Y})} \; \underset{[N \times 25]}{(Y^0)^T}$$

$$\underset{[1\times 1]}{\frac{\partial L}{\partial b^1}} = \underset{[1\times N]}{\frac{\partial L}{\partial \hat{Y}}} \cdot \underset{[N\times 1]}{\frac{\partial \hat{Y}}{\partial b^1}}$$

$$= \frac{2}{N} \underset{[1\times N]}{(Y-\hat{Y})} \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{N\times 1}$$

$$\Rightarrow \underset{[25\times N]}{\frac{\partial L}{\partial Y^0}} = \underset{[25\times 1]}{\frac{\partial \hat{Y}}{\partial Y^0}} \cdot \underset{[1\times N]}{\frac{\partial L}{\partial \hat{Y}}}$$

$$= \underset{[25\times 1]}{\left(W^{1T}\right)} \underset{[1\times N]}{\left(\frac{2}{N}(Y-\hat{Y})\right)}$$

$$\Rightarrow \underset{[25\times N]}{\frac{\partial L}{\partial Z^0}} = \underset{[25\times N]}{\frac{\partial L}{\partial Y^0}} \underbrace{.*}_{\text{Element wise}} \underset{[25\times N]}{\tanh'(Z^0)}$$

elementwise

$$\frac{\partial L}{\partial Z^0} = \underset{[25\times N]}{\frac{\partial L}{\partial Y_0}} .* \underbrace{\left(\mathbb{1} - \tanh^2(Z^0)\right)}_{[25\times N]} \; ; \; \mathbb{1} = \text{np.ones}(25, N)$$

$$\Rightarrow \underset{[25\times 8]}{\frac{\partial L}{\partial W^0}} = \underset{[25\times N]}{\frac{\partial L}{\partial Z^0}} \cdot \underbrace{\frac{\partial Z^0}{\partial W^0}}_{X^T \; [N\times 8]}$$

$$\underset{[25\times 1]}{\frac{\partial L}{\partial b^0}} = \underset{[25\times N]}{\frac{\partial L}{\partial Z^0}} \cdot \underset{[N\times 1]}{\frac{\partial Z^0}{\partial b^0}} \; \mathbb{1}$$

R Prop

$\alpha = 1.2$, $\beta = 0.5$

For each layer l, for each i, j:

Initialize $w_{l,i,j}$, $\Delta w_{l,i,j} > 0$

$$prevD(l,i,j) = \frac{\partial E}{\partial w_{l,i,j}}$$

$$\Delta w_{l,i,j} = sign(prevD(l,i,j))\,\Delta w_{l,i,j}$$

while not converge:

$$w_{lij} = w_{lij} - \Delta w_{lij}$$

$$D(l,i,j) = \frac{\partial E}{\partial w_{lij}}$$

if $(sign(prevD(l\,ij)) == sign\ D(l,i,j))$:

$$\Delta w_{lij} = \min(\alpha \Delta w_{lij}, \Delta_{max})$$

$$prevD(l,i,j) = D(l,i,j)$$

else:

$$w_{lij} = w_{lij} + \Delta w_{lij}$$

$$\Delta w_{lij} = \beta \Delta w_{lij}$$

Backprop

$\eta = 0.00001$

while convergence :

$$\frac{\partial L}{\partial w^0}, \frac{\partial L}{\partial b^0}, \frac{\partial L}{\partial w^1}, \frac{\partial L}{\partial b^1} = \text{grad}\left(w^0, b^0, w^1, b^1\right)$$

$$w^0 = w^0 - \eta \frac{\partial L}{\partial w^0}$$

$$b^0 = b^0 - \eta \frac{\partial L}{\partial b^0}$$

$$w^1 = w^1 - \eta \frac{\partial L}{\partial w^1}$$

$$b^1 = b^1 - \eta \frac{\partial L}{\partial b^1}$$

Quickprop :

while convergence :

$$\frac{\partial L}{\partial w^0}, \frac{\partial L}{\partial b^0}, \frac{\partial L}{\partial w^1}, \frac{\partial L}{\partial b^1} = \text{grad}\left(w^0, b^0, w^1, b^1\right)$$

For each l, i, j

iteration no. $\longrightarrow (k)$

$$\Delta w_{l,i,j} = \frac{\Delta w^{(k-1)}_{l,i,j} \left(w^k_{lij} - w^{k-1}_{lij}\right) E'\left(w^k_{lij}\right)}{E'\left(w^k_{lij}\right) - E'\left(w^{k-1}_{lij}\right)}$$

$$w^{k+1} = w^k_{l,i,j} - \Delta w^k_{l,i,j}$$

Q1 ⇒

$$\underset{[1\times 4]}{C} = \overset{A}{[1\times 2]}\overset{B}{[2\times 4]} \;,$$

$$\underset{\mathbb{R}^{1\times 1}}{L} = \|\mathcal{C}\|_2^2 \;,$$

$$\frac{\partial L}{\partial C} \in \mathbb{R}^{1\times 4}$$

[not scalar multiplcn, matrix multn
Order matters]

$$\Rightarrow \underset{[1\times 2]}{\frac{\partial L}{\partial A}} = \underset{[1\times 4]}{\frac{\partial L}{\partial C}} \cdot \frac{\partial C}{\partial A}$$

$$\underset{[1\times 2]}{\frac{\partial L}{\partial A}} = \underset{[1\times 4]}{\frac{\partial L}{\partial C}} \cdot \underset{[4\times 2]}{B^T}$$

$$\left.\begin{aligned} \underset{[2\times 4]}{\frac{\partial L}{\partial B}} &= \underset{[1\times 4]}{\frac{\partial L}{\partial C}} \cdot \frac{\partial C}{\partial B} \\ \underset{[2\times 4]}{\frac{\partial L}{\partial B}} &= \underbrace{\frac{\partial C}{\partial B}}_{\substack{A^T \\ [2\times 1]}} \cdot \underset{[1\times 4]}{\frac{\partial L}{\partial C}} \\ &= \end{aligned}\right\} \times$$

$$\Rightarrow (x-a)^2 = (a-x)^2$$

$$\frac{\partial (x-a)^2}{\partial x} = \frac{\partial (a-x)^2}{\partial x}$$

$$2(x-a) = -2(a-x)$$

$$= 2(a-x)$$