Excersise 1 :

i) $\sigma(x) = \frac{1}{1+e^{-x}}$

$$\frac{\partial \sigma(x)}{\partial x} = \frac{0 - 1 * e^{-x}(-1)}{(1+e^{-x})^2}$$

$$= \frac{e^{-x}}{(1+e^{-x})^2} \Rightarrow \frac{\cancel{(1+e^{-x})}}{(1+e^{-x})^{\cancel{2}}} - \frac{1}{(1+e^{-x})^2}$$

$$= \sigma(x) - \sigma^2(x) \quad \therefore \frac{1}{1+e^{-x}} = \sigma(x)$$

ii) Taylor Series ⇒

$$P(x) = f(a) + \frac{1}{1!}\frac{\partial f}{\partial x}(a)(x-a)^1 + \frac{1}{2!}\frac{\partial^2 f}{\partial x^2}(a)(x-a)^2 + \dots$$

Here $a=0$

$$P(x) = f(0) + \frac{1}{1!}\frac{\partial f}{\partial x}(0)\cdot x + \frac{1}{2!}\frac{\partial^2 f}{\partial x^2}(0)\cdot x^2$$

$$= \frac{1}{2} + \left(\frac{1}{2} - \frac{1}{4}\right)\cdot x + \frac{1}{2!}\cdot 0\cdot x^2 = \frac{1}{2} + \frac{1}{4}x.$$

$$\frac{\partial^2 f}{\partial x^2} = \frac{\partial}{\partial x}\Big(\sigma(x)\,(1-\sigma(x))\Big) = \sigma(x)\cdot(-1)(\sigma(x) - \sigma^2(x))$$
$$+ \sigma(x)(1-\sigma(x))(1-\sigma(x))$$
$$= \sigma(x)(1-\sigma(x))^2 - \sigma^2(x)(1-\sigma(x))$$
$$= \sigma(x)(1-\sigma(x))\Big((1-\sigma(x)) - \sigma(x)\Big)$$

At $x=0$: $\Rightarrow \frac{1}{2} * \frac{1}{2}\left(\frac{1}{2} - \frac{1}{2}\right) = 0$

(c)

Lets approximate the f function at w_n:

$$f(x) = f(w_n) + \frac{1}{1!}\frac{\partial f}{\partial w}(w_n)(x - w_n) + \dots$$

$$f(x) = f(w_n) + g(n)\cdot(x - w_n)$$

Lets approximate it at $x = w_{n+1}$

$$f(w_{n+1}) = f(w_n) + g(n)(w_{n+1} - w_n)$$

$$\therefore w_{n+1} = w_n - \epsilon g(n)$$

$$f(w_{n+1}) = f(w_n) + g(n)(\cancel{w_n} - \epsilon g(n) - \cancel{w_n})$$

$$\left[f(w_{n+1}) = f(w_n) - \epsilon(\nabla f(w_n))^2\right]$$

This is Similar to Gradient descent step, but instead of gradient we are using the Square of gradient.

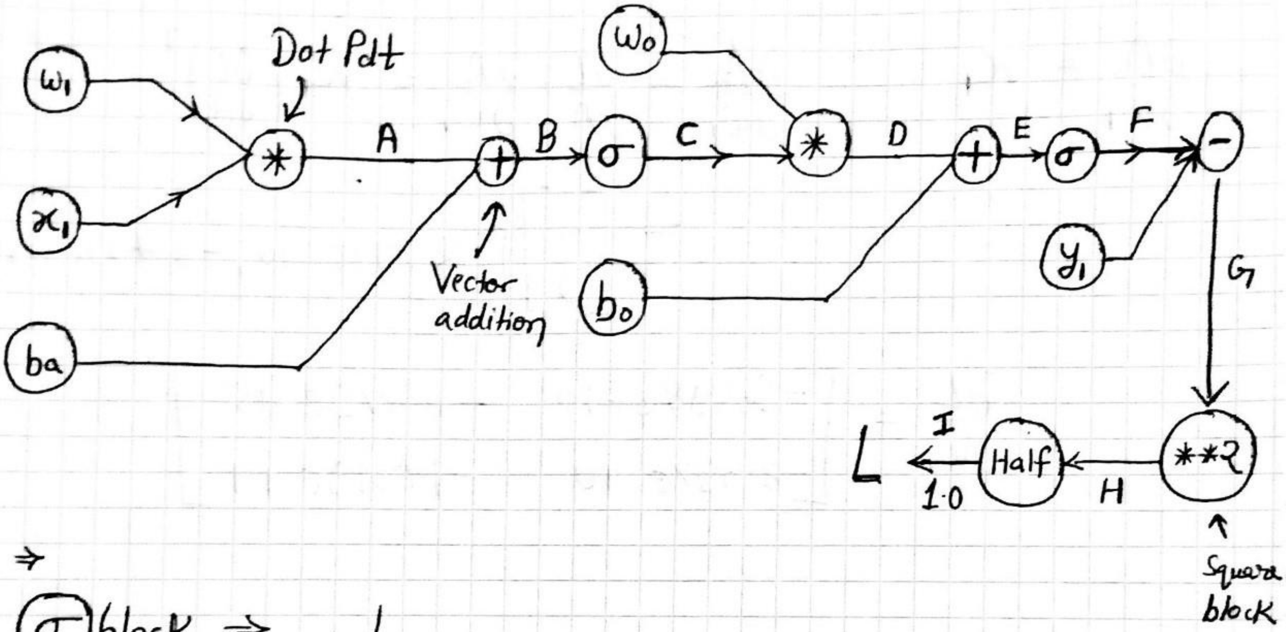
Above Equation shows that the value of function at next step will decrease, and suddenly hit local/Global minima.

Excersise 2 ⇒

(a) $$L = \frac{1}{2} \sum_{i=1}^{N} \left(\sigma\left(w_o^T\left(\sigma\left(w_i^T x_n + ba\right)\right) + bo\right) - y_n\right)^2$$

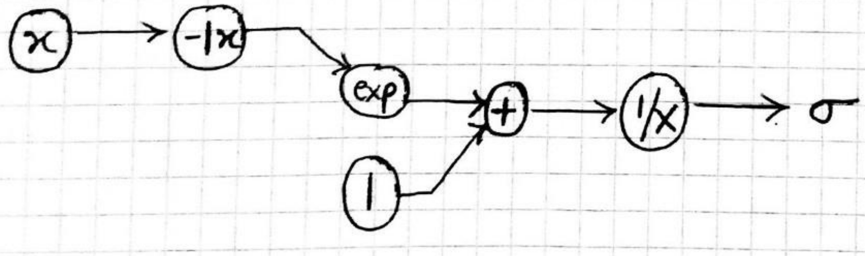
For a Single instance

$$= \frac{1}{2} \left(\sigma\left(w_o^T\left(\sigma\left(w_1^T x_1 + ba\right)\right) + b_o\right) - y_1\right)^2$$



Note ⇒

σ block ⇒ $\frac{1}{1+e^{-x}}$



2.b : Lets assume some Values → We could also assume Constants.

$$W_1 = \begin{bmatrix}1 & 2\\ 3 & 4\end{bmatrix}_{2\times2} \qquad W_o = \begin{bmatrix}1\\ 2\end{bmatrix}_{2\times1} \qquad b_a = \begin{bmatrix}2\\ 4\end{bmatrix} \qquad b_o = \begin{bmatrix}3\\ \bullet\end{bmatrix}_{1\times1}$$

Forward Pass →

$$x_1 = \begin{bmatrix}2\\ 3\end{bmatrix}$$

$$A = W_1 x_1 = \begin{bmatrix}1 & 2\\ 3 & 4\end{bmatrix}\begin{bmatrix}2\\ 3\end{bmatrix} = \begin{bmatrix}8\\ 18\end{bmatrix}$$

$\therefore y_1 = 1$ Assuming

$$B = W_1 x_1 + b_a = \begin{bmatrix}8\\ 18\end{bmatrix} + \begin{bmatrix}2\\ 4\end{bmatrix} = \begin{bmatrix}10\\ 22\end{bmatrix}$$

$$C = \sigma(B) = \begin{bmatrix}0.999\\ 1\end{bmatrix}.$$

- $D = W_o \cdot C = [2.999]$
- $E = W_o \cdot C + b_o = [5.999]$
- $F = \sigma(E) = [0.9975]$
- $G = [0.9975 - 1] = -0.002499$
- $H = G^2 = 6.24e-06$

$I = 3.124e-06 = L$

Backward Pass →

- $\frac{\partial L}{\partial L} = \frac{\partial I}{\partial L} = 1 \quad \therefore I = L$
- $I = \frac{H}{2} \Rightarrow \frac{\partial I}{\partial H} = \frac{1}{2} \quad \therefore \frac{\partial I}{\partial I} * \frac{\partial I}{\partial H} \Rightarrow \left[\frac{\partial L}{\partial H} = \frac{1}{2}\right]$ ✓
- $H = G^2 \quad \therefore \frac{\partial H}{\partial G} = 2*G \quad \therefore \frac{\partial L}{\partial H} * \frac{\partial H}{\partial G} \Rightarrow \frac{\partial L}{\partial G} = \frac{1}{\cancel{2}} * \cancel{2}G$

$$\left[\frac{\partial L}{\partial G} = G\right]$$

- $G_7 = F - y_1$

$$\frac{\partial G_7}{\partial F} = 1 \qquad \therefore \frac{\partial L}{\partial G_7} \frac{\partial G_7}{\partial F} = G_7 * 1 = G_7 \qquad \left[\therefore \frac{\partial L}{\partial F} = G_7 \right]$$

- $F = \sigma(E)$

$$\frac{\partial F}{\partial E} = \sigma(E)(1 - \sigma(E)) \Rightarrow F(1-F).$$

$$\therefore \frac{\partial L}{\partial F} * \frac{\partial F}{\partial E} = G_7 F(1-F) = [-0.00249] \cdot [0.9975] (1 - 0.9975)$$

$$\therefore \left[\frac{\partial L}{\partial E} = -6.1665e-06 \right]$$

- $E = D + b_0$

$$\frac{\partial E}{\partial D} = 1 \qquad \therefore \left[\frac{\partial L}{\partial D} = \frac{\partial L}{\partial E} * \frac{\partial E}{\partial D} = -6.166e-06. \right]$$

$$\frac{\partial E}{\partial b_0} = 1 \qquad \therefore \left[\frac{\partial L}{\partial b_0} = \frac{\partial L}{\partial E} \cdot \frac{\partial E}{\partial b_0} = -6.166e-06. \right]$$

$D = W_0 \cdot C$

$$\frac{\partial D}{\partial W_0} = C. \qquad \frac{\partial L}{\partial W_0} = \frac{\partial L}{\partial D} * \frac{\partial D}{\partial W_0} = -6.166e-06 * C.$$

$$\left[\frac{\partial L}{\partial W_0} = \begin{bmatrix} -6.165e-06 \\ -6.166e-06 \end{bmatrix} \right] ✓$$

$$\frac{\partial D}{\partial C} = W_0$$

$$\therefore \frac{\partial L}{\partial C} = -6.166e-06 * W_0 = \begin{bmatrix} -6.166* e-06 \\ -1.23e-05 \end{bmatrix}_{2\times 1} ✓$$

$$C = \sigma(B). \qquad \therefore \frac{\partial C}{\partial B} = C(1-C)$$

$$\frac{\partial L}{\partial B} = \frac{\partial L}{\partial C} * \frac{\partial C}{\partial B} = C(1-C) \overline{\frac{\partial L}{\partial C}} = \begin{bmatrix} -2.799e-10 \\ -3.439e-15 \end{bmatrix}$$

$> \quad B = A + ba$

$$\frac{\partial B}{\partial ba} = 1 \qquad \frac{\partial L}{\partial ba} = \frac{\partial L}{\partial B} \cdot \frac{\partial B}{\partial ba} = \begin{bmatrix} -2.799e-10 \\ -3.439e-15 \end{bmatrix}$$

$$\frac{\partial B}{\partial A} = 1 \qquad \therefore \frac{\partial L}{\partial A} = \begin{bmatrix} -2.799e-10 \\ -3.439e-15 \end{bmatrix}$$

Because of Consistency

$$A = \omega_1^{T} \cdot x_1$$

$$\frac{\partial A}{\partial \omega_1} = x_1 \qquad \frac{\partial A}{\partial \omega_1} = \frac{\partial L}{\partial A} \cdot \frac{\partial A}{\partial \omega_1} = x_1 \cdot \left(\frac{\partial L}{\partial A} \right)^{T}$$

$$= \begin{bmatrix} 2 \\ 3 \end{bmatrix} \left[-2.799e-10, -3.439e-15 \right]$$

$$\frac{\partial L}{\partial \omega_1} = \begin{bmatrix} -5.598e-10 & -6.879e-15 \\ -8.397e-10 & -1.0319e-14 \end{bmatrix}$$

Problem 6.3

A) The eigenvalues of the Hessian characterize the local curvature of the loss which, for example, determine how fast models can be optimized via first-order methods (atleast for convex problems), and is also conjectured to influence the generalization properties.
So if the training error decreases so does the fraction of negative eigen value decreases. It happens because we know that with the training error decreases, we close toward the minimum where we get all positive eigenvalues.

B) In Gradient Descent:

Step of this method points in the right direction close to a saddle point and if an eigenvalue is positive then step will move away in the direction of eigenvector and achieve minimum value. So eigenvalues are very small hence the steps taken are very small. Gradient descent drawback is the small size of the step along with each eigenvector.

In Newton method:

Newton methods solve the slowness problem by rescaling the gradients in each direction with the inverse of the corresponding eigenvalue. However, this approach can result in moving in the wrong direction. So for the negative eigenvalues the newton step moves in the direction opposite to the gradient descent step, and thus moves in the direction of increasing error.

In Trust region:

In Trust Region if the minimum eigenvalues is very low then the damping factor has to be large and this results in potentially small step size.

C) To achieve the optimal rescaling as in the case of Newton’s Method the gradient descent was scaled along each eigen direction by 1/ λi where λi represents eigenvalues. This also preserves the direction, which was not possible in the case of Newton’s Method .

# Team Members:

Priyanka Upadhyay, Email Id : s8prupad@stud.uni-saarland.de
Gopal Bhattrai, Email Id : gobh00001@stud.uni-saarland.de

Tutor: Redion Xhepa

Matriculation Number :

Priyanka : 2581714
Gopal : 7013547