DL for Visual Recognization. ①

Exercise 3

Yash, Gaurav

1

ⓐ $f(x) = 100(x_2 - x_1^2)^2 + (1-x_1)^2$

$$f'(x_1) = \frac{\partial f(x)}{\partial x_1} = -400(x_1 x_2 - x_1^3) - 2(1-x_1)$$

$$f'(x_2) = \frac{\partial f(x)}{\partial x_2} = 200(x_2 - x_1^2)$$

Now $\nabla f(x) = (-400(x_1x_2 - x_1^3) - 2(1-x_1))\hat{i} + (200(x_2 - x_1^2))\hat{j}$

where $\hat{i}$ & $\hat{j}$ are unit vectors.

Now $\nabla^2 f(x)$,

$$\left.\begin{aligned} \frac{\partial f'(x_1)}{\partial x_1} &= -400(x_2 - 3x_1^2) + 2 \\ \frac{\partial f'(x_1)}{\partial x_2} &= -400x_1 \\ \frac{\partial f'(x_2)}{\partial x_1} &= -400x_1 \\ \frac{\partial f'(x_2)}{\partial x_2} &= 200 \end{aligned}\right\} \Rightarrow \nabla^2 f(x) = \begin{bmatrix} -400(x_2 - 3x_1^2) + 2 & -400x_1 \\ -400x_1 & 200 \end{bmatrix}$$

Now, we want to prove $x^* = (1,1)^T$ as lonely minimizer of the given function. For that we need $\nabla f(x) = 0$.

$$\Rightarrow -400(x_1x_2 - x_1^3) - 2(1-x_1) = 0 \;\&\; 200(x_2 - x_1^2) = 0.$$

$$\Rightarrow x_2 = x_1^2 \quad \text{—} ①$$

By ① $-400(x_1^3 - x_1^3) - 2(1-x_1) = 0$

$$\Rightarrow x_1 = 1 \quad \Rightarrow x_2 = 1.$$

→ Hence, $(1,1)^T$ is a critical pt..

Now, $\nabla^2 f(x)_{(1,1)} = \begin{bmatrix} 802 & -400 \\ -400 & 200 \end{bmatrix}$

To prove $\nabla^2 f(x)_{(1,1)}$ as +ve definite, we need to prove for any positive column vector z so that $z^T \nabla^2 f(x)_{(1,1)} z$ is strictly +ve.

Let $z = \begin{bmatrix} a \\ b \end{bmatrix}$, where $a, b > 0$.

$$\therefore \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} 802 & -400 \\ -400 & 200 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}$$

$$= \begin{bmatrix} 802a - 400b & -400a + 200b \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}$$

$$= 802a^2 - 400ab - 400ab + 200b^2$$

$$= 802a^2 - 800ab + 200b^2 > 0, \text{ as } a, b > 0.$$

Hence proved. □

(b) $f(x) = 8x_1 + 12x_2 + x_1^2 - 2x_2^2$

$f'(x_1) = \dfrac{\partial f(x)}{\partial x_1} = 8 + 2x_1$

$f'(x_2) = \dfrac{\partial f(x)}{\partial x_2} = 12 - 4x_2$

$\therefore \nabla f(x) = (8 + 2x_1)\hat{i} + (12 - 4x_2)\hat{j}$, where $\hat{i}$ & $\hat{j}$ are unit vectors.

Now, for a stationary / critical pt. $\nabla f(x) = 0$.

$\Rightarrow x_1 = -4$ & $x_2 = 3$, $\therefore (-4, 3)$ is only one critical pt.

→ Now, in order to prove it as a saddle pt., its Hessian matrix's eigenvalues should be both +ve as well as -ve.

$$\frac{\partial f'(x_1)}{\partial x_1} = 2 \qquad \frac{\partial f'(x_2)}{\partial x_1} = 0$$

$$\frac{\partial f'(x_1)}{\partial x_2} = 0 \qquad \frac{\partial f'(x_2)}{\partial x_2} = -4$$

$$\therefore \nabla^2 f(x) = \begin{bmatrix} 2 & 0 \\ 0 & -4 \end{bmatrix}$$

Now, in order to get eigenvalues we need to do $|A - \lambda I| = 0$.

$$\Rightarrow |\nabla^2 f(x) - \lambda I_2| = 0$$

$$\Rightarrow \begin{vmatrix} 2-\lambda & 0 \\ 0 & -4-\lambda \end{vmatrix} = 0.$$

$$\Rightarrow (\lambda + 4)(\lambda - 2) = 0$$

$$\Rightarrow \lambda = -4, 2$$

Hence, as eigen values are both +ve & -ve, we conclude that the only critical pt. $(-4,3)$ that we have is the saddle pt.

2)a) For many high dimensional nonconvex functⁿ (in our case study also), local minima & maxima are rare when compared to saddle pts (where it is both local minima & a local maxima for different cross-sections of the function.

→ We can also posit that in higher dimensions, the expected ratio of saddle pts to local minima grows exponentially with n. functions $f: \mathbb{R}^n \to \mathbb{R}$.

→ Also for random functions, eigenvalues of Hessian ~~are~~ are more likely to be +ve as we reach regions with lower cost.

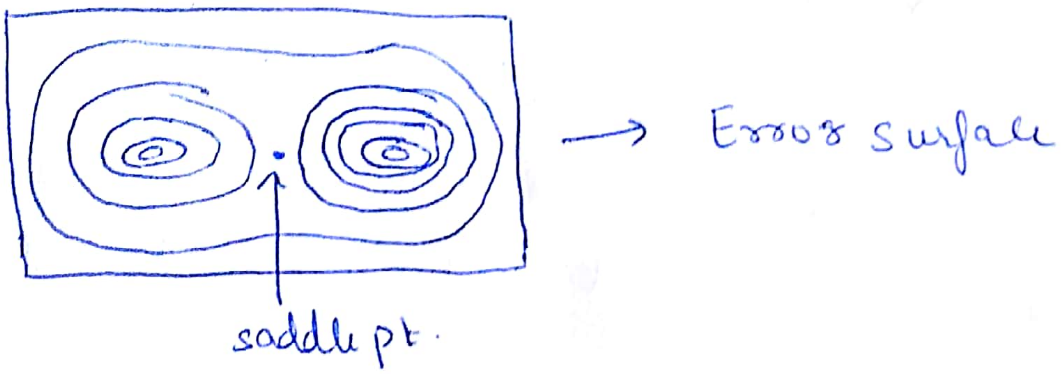
→ This ~~means~~ means that for local minima, they are more likely to have low cost and critical pts. with higher cost are far more likely to be saddle pt's. (very high cost implies local maxima)

→ But Dauphin et al. (2014) showed emperically that real neural networks also have loss functions that contain very many high cost saddle pt's.

→ Now, regarding our case study, the gradient is very small in the ~~plot~~ iteratn's b/w 100 to 10,000. (i.e. the learning rate is also not good). (Refer Fig. 8.2 Pg. 280 in Deep Learning book)

→ And, as a result the network might stuck at a saddle pt. ∵ the gradient being small.

→ The saddle pt. would be surrounded by high error plateau and thus decreasing the learning rate.



b) Since saddle pt's are often surrounded by plateau's of small curvature, gradient descent slows near them. Therefore for Newton's method, saddle pt's constitute a problem ∵ it is designed to solve for a pt. where gradient is zero. Hence, without apt modificatn, it can jump to a saddle pt. In our case, ~~at~~ b/w 100th & 10,000th iteratn, it is nearly a flat region. In these scenario, gradient & Hessian are zero. And hence at 100th iteratn, the determinant of Hessian matrix $\neq 0$. (i.e. $|H| \neq 0$)