Deep Learning - IIT Ropar

Abhishek Thakur
Prime Minister Research Fellow (IIT Hyderabad)

Prof. Mitesh M. Khapra, IIT Madras
Prof. Sudarshan Iyengar, IIT Ropar

Week 2
August 6, 2022

Multi layer Perceptron

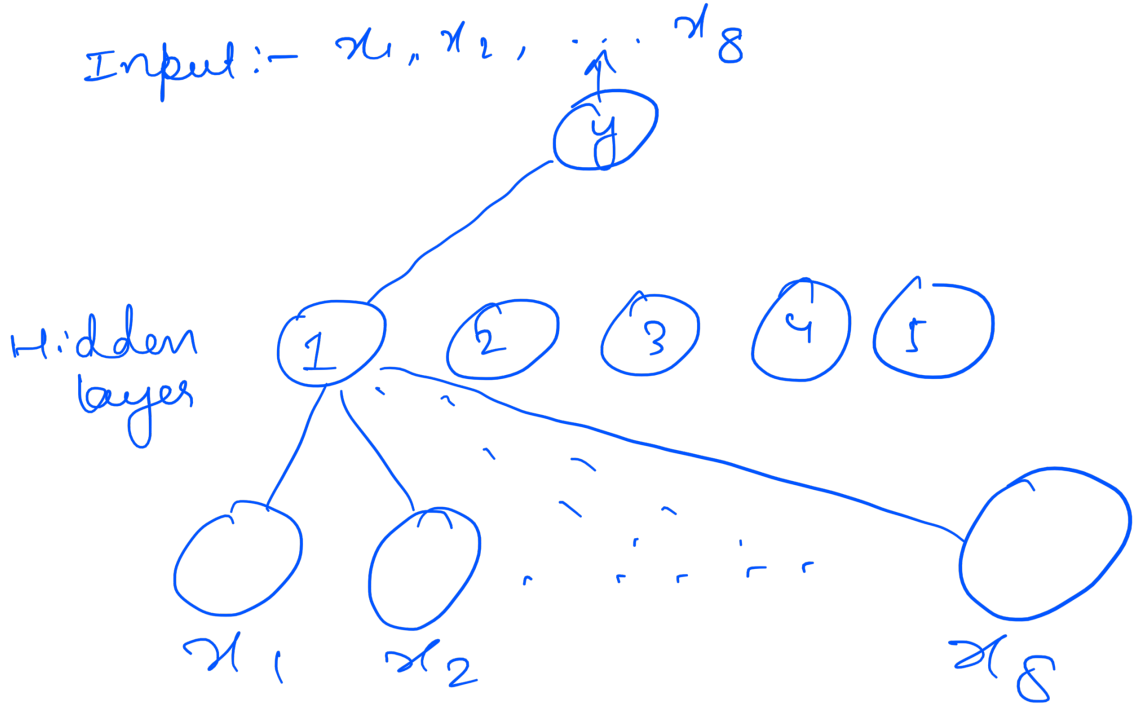
Q1: Suppose the number of neurons in the input layer is 8 and the consecutive hidden layer is 5 (In MLP). Pick out the maximum number of connections that can exist from the input layer to the hidden layer

(A) 14

(B) 40 $5 \times 8 = 40$

(C) 18

(D) 10



Perceptron

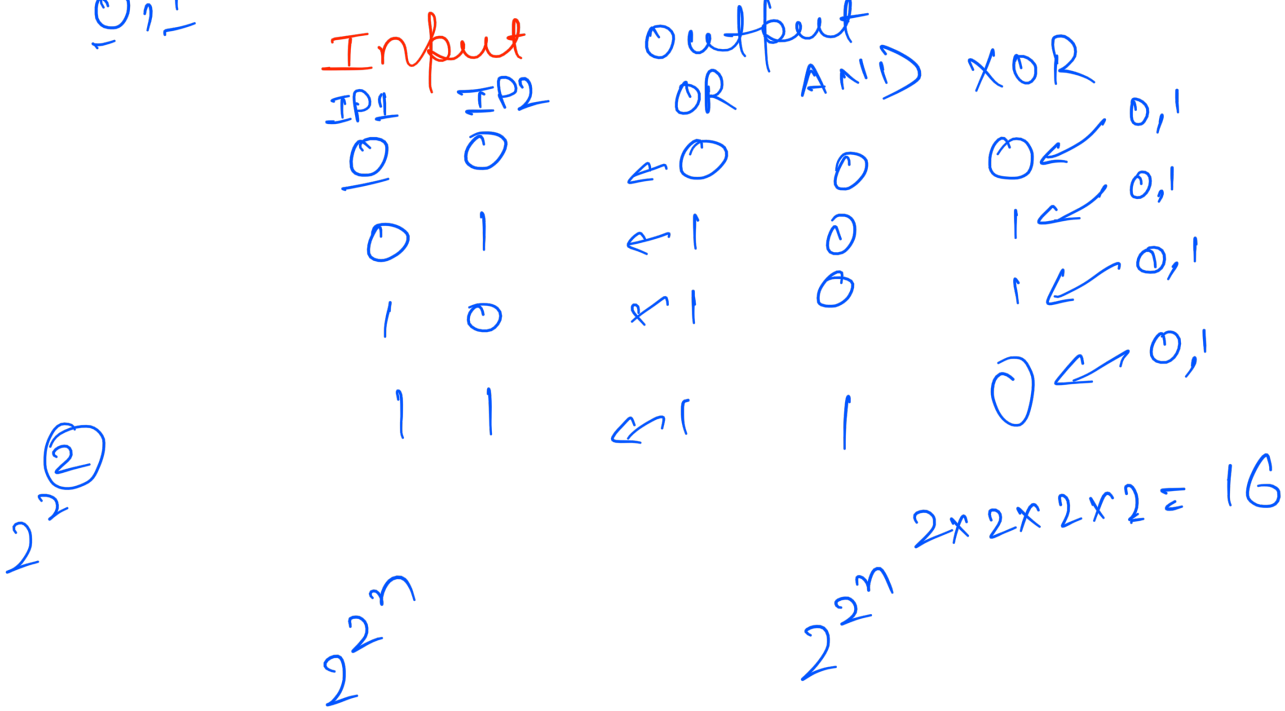
Q2: Given that you have n inputs, how many boolean functions can be designed?

(A) 2^{2^n}

(B) 2^{n^2}

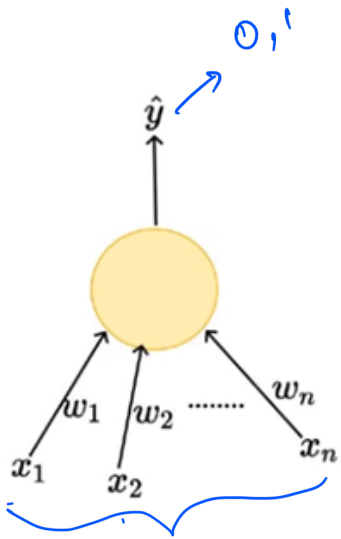
(C) $2^2 n$

(D) n^{2^2}



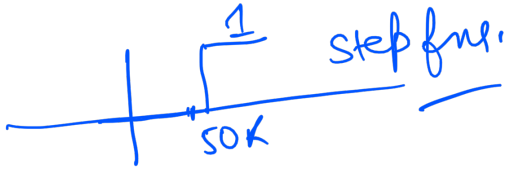
Perceptron Vs Sigmoid

Perceptron



$\hat{y} = 1$ if $\sum_{i=1}^{n} w_i x_i \geq b$

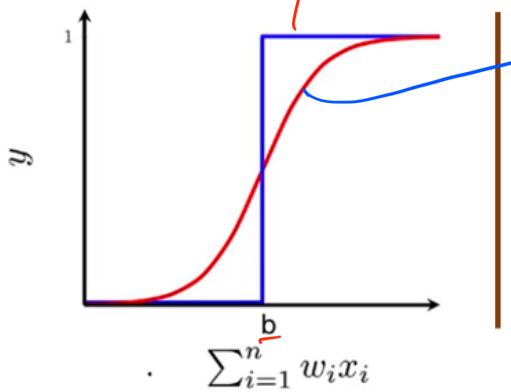
$\hat{y} = 0$ otherwise



mobile → 50K

you have 49K

Sigmoid Neuron



$b^- + 0 \; b^+$ $\lim_{\delta \to 0} b - \delta$ $\{0 + 0\; 1\}$

$$y = \frac{1}{1+e^{-(w^T x+b)}}$$

Perceptron Vs Sigmoid

Q3: Why do we need sigmoid neuron over perceptron?

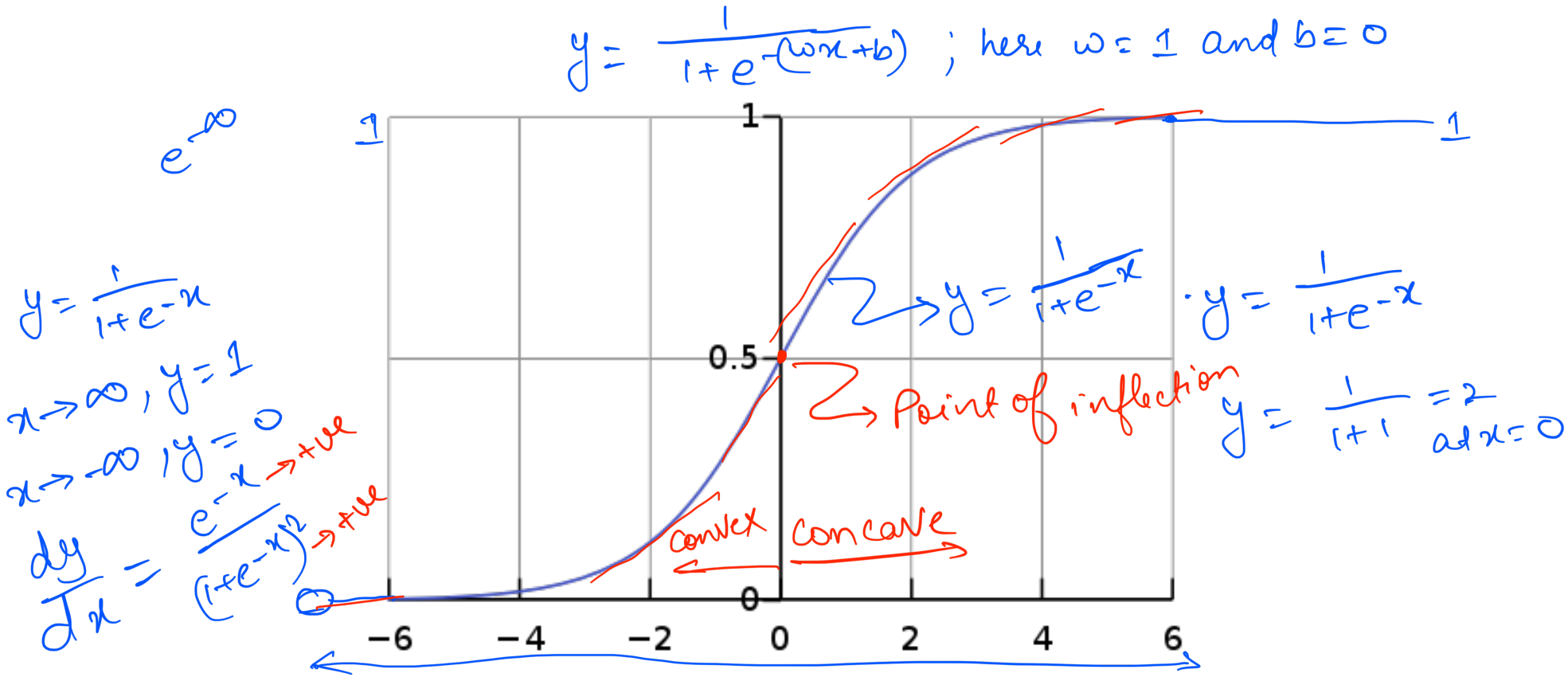
(A) The small change in the input to a perceptron can sometimes cause the output to completely flip, say from 0 to 1

(B) output function is much smoother than the step function

(C) None

(D) Both

Sigmoid Neuron



A sigmoid function is a bounded, differentiable, real function that is defined for all real input values and has a non-negative derivative at each point and exactly one inflection point.

Sigmoid Neuron

Q4: What is the Characteristics of sigmoid function?

Statement I. Differentiable

Statement II. Smooth and continuous

Statement III. It is monotonic

Statement IV. derivative is always positive

(A) I

(B) I II

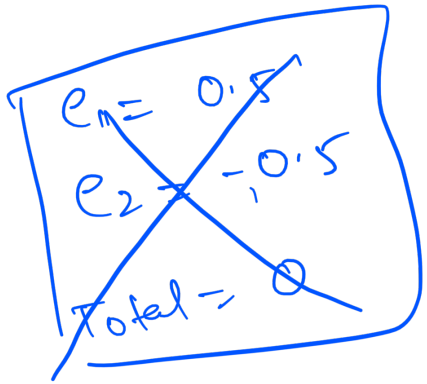
(C) II II III

(D) I II III IV

Error Function

Sigmoid function is given as

$$\hat{y} = \frac{1}{1 + e^{-(\mathbf{w}^T \mathbf{x})}}$$



- **Parameter:** $\mathbf{w}$
- **Learning algorithm:** Gradient Descent [we will see soon]
- **Objective/Loss/Error function:** One possibility is

$$\mathscr{L}(\mathbf{w}) = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

$\hat{y}$ → estimated
y → actual value

The learning algorithm should aim to find a w which minimizes the above function (squared error between y and $\hat{y}$)

Error Function

Q5: Squared error function is preferred over simple difference between the actual and predicted value of output. Identify the appropriate reason for this statement.

(A) square blows up the error

(B) In the second case, the positive and negative errors cancel out each other

(C) It is not differentiable

(D) The statement cannot be justified

Non Linear Separable logic gate

Q6: Identify the statements that are True about learning algorithm.

Statement I. Maximizes objective function

Statement II. Learns the parameters from data

Statement III. minimize the objective function

(A) I

(B) I II

(C) II III

(D) III

objective : $\mathcal{L}(w) = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$

minimum

we have to find w,

$\mathcal{L}(w) = \left[\sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \right]$

Loss Function

Q7: (x,y) = (0.5, 0.2), (2.5, 0.9), w = 0.5, b = 0 and the function is logistic sigmoid function. calculate the loss function.

w = 0.5
b = 0
$f(x) \rightarrow y$
0.5 → 0.2
2.5 → 0.9

Given:

$$\mathscr{L}(w,b) = \frac{1}{2} * \sum_{i=1}^{N} (y_i - f(x_i))^2$$

$$= \frac{1}{2} * (y_1 - f(x_1))^2 + (y_2 - f(x_2))^2$$

y_1 - $\hat{y}_1$

$$= \frac{1}{2} * (0.9 - f(2.5))^2 + (0.2 - f(0.5))^2$$

$$f(x) = \hat{y} = \frac{1}{1+e^{-(wx+b)}} = \frac{1}{1+e^{-0.5x}}$$ (x = 2.5)

Loss Function

$$L = \frac{1}{2} \times \left[(0.9 - 0.77)^2 + (0.2 - 0.56)^2 \right]$$

$$L = \frac{1}{2} \left[0.0169 + 0.13 \right]$$

$$L = \frac{1}{2} \left[0.14 \right] = 0.07$$

$$\frac{1}{1 + e^{-0.25}}$$

0.13

0.0169

Sigmoid Neuron

Q8: Which of the following results when the value of b is changed in a sigmoid function given below?

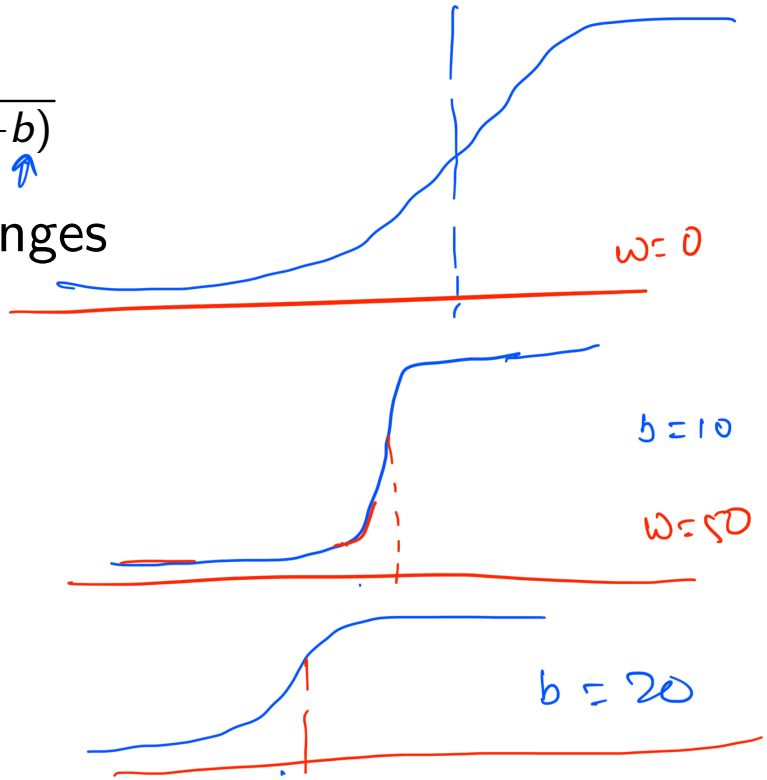
$$\sigma(x) = \frac{1}{1 + e^{-(wx+b)}}$$

(A) value of x at which transition occurs changes

(B) value of w changes

(C) the slope becomes steeper

(D) the transition point is found



Sigmoid Neuron

Q9: Which of the following results when the value of w is changed in a sigmoid function given below?

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

(A) value of x at which transition occurs changes

(B) value of w changes

(C) the slope becomes steeper

(D) the transition point is found

w ↑ → slope becomes more steeper

Gradient Descent Rule

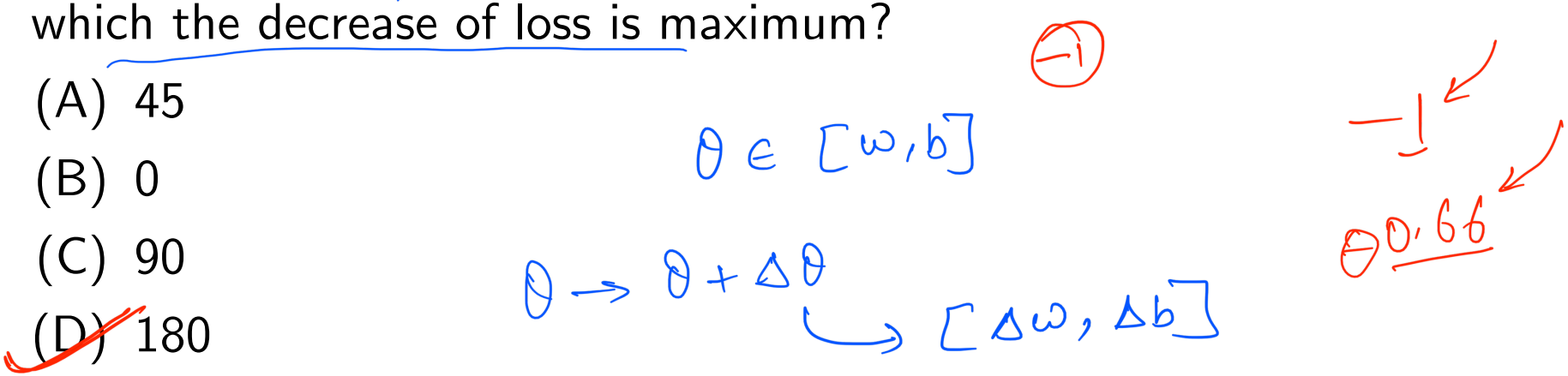
Q10: The condition for the new loss to be less than the current loss is $u^T L(\theta) < 0$. What is the direction of u with respect to the Gradient for which the decrease of loss is maximum?

(A) 45

(B) 0

(C) 90

(D) 180



Gradient Descent Rule

For ease of notation, let $\Delta\theta = u$, then from Taylor series, we have,

$$\mathscr{L}(\theta + \eta u) = \mathscr{L}(\theta) + \eta * u^T \nabla_\theta \mathscr{L}(\theta) + \frac{\eta^2}{2!} * u^T \nabla^2 \mathscr{L}(\theta) u + \frac{\eta^3}{3!} * \ldots + \frac{\eta^4}{4!} * \ldots$$
$$= \mathscr{L}(\theta) + \eta * u^T \nabla_\theta \mathscr{L}(\theta) \; [\eta \text{ is typically small, so } \eta^2, \eta^3, .. \to 0]$$

Note that the move (ηu) would be favorable only if,

$$\mathscr{L}(\theta + \eta u) - \mathscr{L}(\theta) < 0 \; [\textit{i.e., if the new loss is less than the previous loss}]$$

This implies,

$$u^T \nabla_\theta \mathscr{L}(\theta) < 0$$

$u = \Delta\theta$

$\mathscr{L}(\theta + \eta u) \simeq \mathscr{L}(\theta) + \eta * u^T \nabla_\theta \mathscr{L}(\theta)$

$t=2$ $\quad$ $t=1$ $\quad$ $= \eta u^T \nabla_\theta \mathscr{L}(\theta) < 0$

$\mathscr{L}(\theta + \eta u) - \mathscr{L}(\theta) < 0$

Gradient Descent Rule

Okay, so we have,

$$u^T \nabla_\theta \mathscr{L}(\theta) < 0$$

u , $\nabla_\theta \mathscr{L}(\theta) = \beta$

$u^T \nabla_\theta \mathscr{L}(\theta) = ||u|| \cdot ||\nabla \mathscr{L}|| \cdot \cos\beta$

But, what is the range of $u^T \nabla_\theta \mathscr{L}(\theta)$? Let us see....
Let β be the angle between u and $\nabla_\theta \mathscr{L}(\theta)$, then we know that,

$$-1 \leq cos(\beta) = \frac{u^T \nabla_\theta \mathscr{L}(\theta)}{||u|| * ||\nabla_\theta \mathscr{L}(\theta)||} \leq 1$$

$\beta = 180°$ $\beta = 0°$

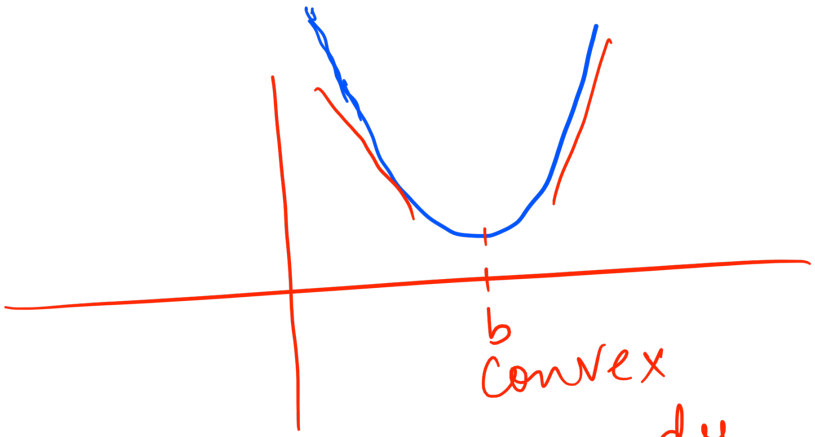
multiply throughout by $k = ||u|| * ||\nabla_\theta \mathscr{L}(\theta)||$

$$-k \leq k * cos(\beta) = u^T \nabla_\theta \mathscr{L}(\theta) \leq k$$

Thus, $\mathscr{L}(\theta + \eta u) - \mathscr{L}(\theta) = u^T \nabla_\theta \mathscr{L}(\theta) = k * cos(\beta)$ will be most negative when $cos(\beta) = -1$ *i.e.*, when β is $180°$

Gradient Descent Rule

- The direction u that we intend to move in should be at 180° w.r.t. the gradient
- In other words, move in a direction opposite to the gradient



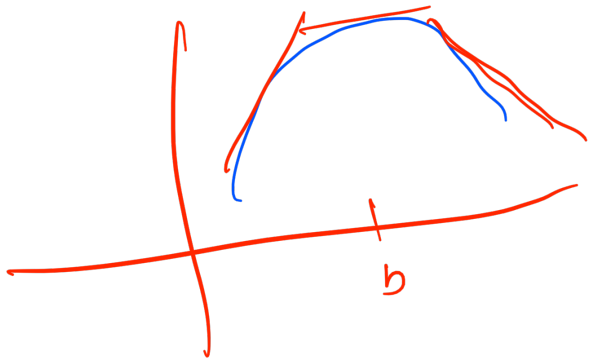
Convex

$x < b, \frac{dy}{dx} < 0$

$x = b, \frac{dy}{dx} = 0$

$x > b, \frac{dy}{dx} > 0$

b



$x < b, \frac{dy}{dx} > 0$

$x = b, \frac{dy}{dx} = 0$

$x > b, \frac{dy}{dx} < 0$


NPTEL

THANK YOU