LINEAR REGRESSION

WITH GRADIENT DESCENT

Group 3

List Of Content

1. **Pick a sample** (x_1, x_2, x_3, y) **from training data**

2. **Compute output** $\hat{y}$

$$\hat{y} = w_1 * TV + w_2 * R + w_3 * N + b$$

$$\hat{y} = w_1 * x_1 + w_2 * x_2 + w_3 * x_3 + b$$

3. **Compute loss** $L = (\hat{y} - y)^2$

4. **Compute derivative**

$$\frac{\partial L}{\partial w_1} = 2x_1(\hat{y} - y) \qquad \frac{\partial L}{\partial w_3} = 2x_3(\hat{y} - y)$$

$$\frac{\partial L}{\partial w_2} = 2x_2(\hat{y} - y) \qquad \frac{\partial L}{\partial b} = 2(\hat{y} - y)$$

5. **Update parameters**

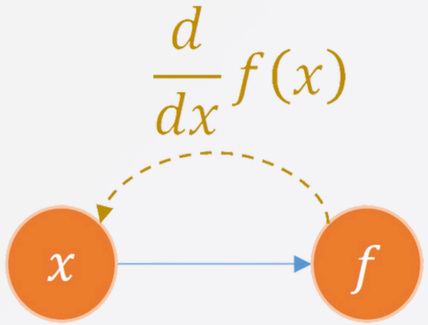
$$w_1 = w_1 - \eta \frac{\partial L}{\partial w_1} \qquad w_3 = w_3 - \eta \frac{\partial L}{\partial w_3}$$

$$w_2 = w_2 - \eta \frac{\partial L}{\partial w_2} \qquad b = b - \eta \frac{\partial L}{\partial b}$$

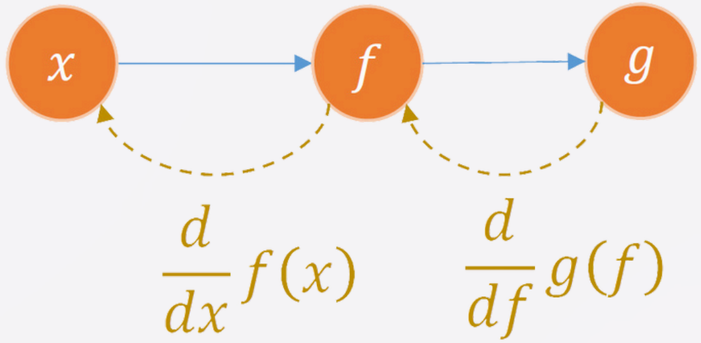
Chain Rules

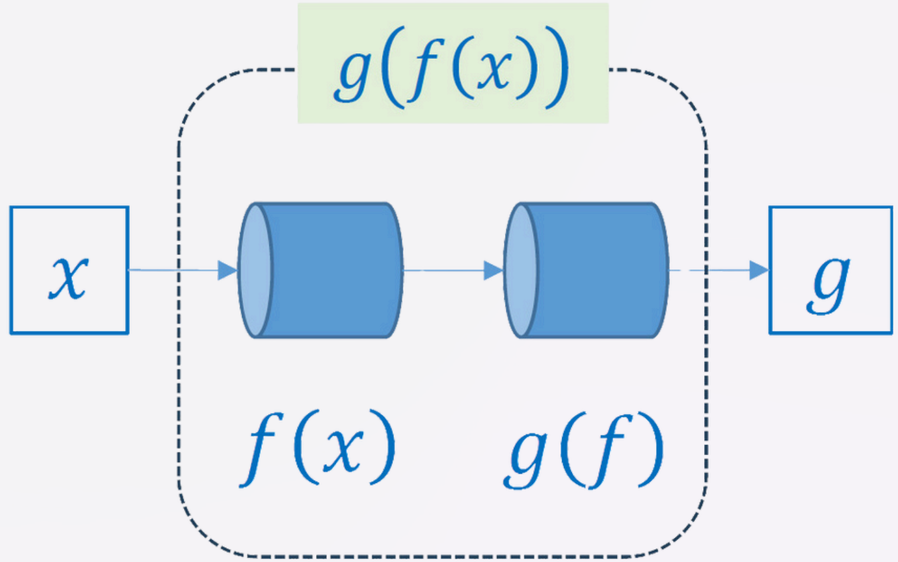
Composite function

$f(x)$



$g(f(x))$



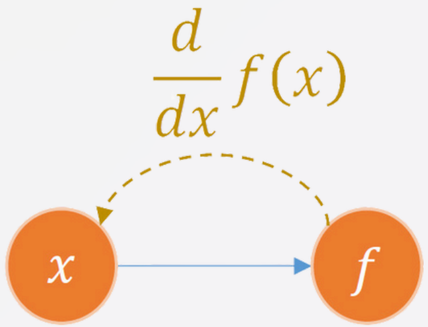


$$\frac{d}{d(x)}g(f(x)) = \left[\frac{d}{df}g(f)\right] \times \left[\frac{d}{dx}f(x)\right]$$

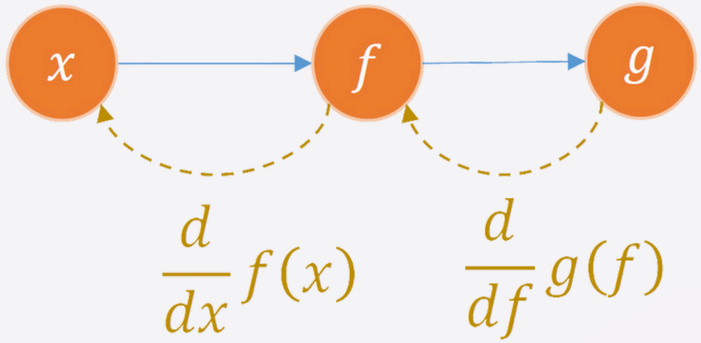
Chain Rules

Composite function

$f(x)$



$g(f(x))$



$$\frac{d}{d(x)}g(f(x)) = \left[\frac{d}{df}g(f)\right] \times \left[\frac{d}{dx}f(x)\right]$$

Example:

$$f(x) = 2x + 1$$

$$g(f) = (f-2)^2$$

$$\Rightarrow f'(x) = 2$$

$$\Rightarrow g'(f) = 2(f-2)$$

$$\Rightarrow \frac{dg}{dx} = \frac{dg}{df} \times \frac{df}{dx}$$

$$= 2(f-2) \times 2$$

$$= 8x - 4$$

Chain Rules

Problem 1

$$f(x) = ax$$

$$\eta = 0.1$$

$$g(f) = (f-3)^2$$

$$\Rightarrow g'(f) = 2(f-3)$$

$$\frac{dg}{da} = \frac{dg}{df} \cdot \frac{df}{da}$$

$$\frac{dg}{da} = 2x(f-3)$$

$$a_t = a_{t-1} - \eta g'(a_{t-1})$$

Problem 2

$$f(x) = 2x + b$$

$$\eta = 0.1$$

$$g(f) = (f-3)^2$$

$$\Rightarrow g'(f) = 2(f-3)$$

$$\frac{dg}{db} = \frac{dg}{df} \cdot \frac{df}{da}$$

$$\frac{dg}{db} = 2(f-3)$$

$$b_t = b_{t-1} - \eta g'(b_{t-1})$$

Chain Rules

Problem 3

$$f(x) = ax + b$$

$$\eta = 0.1$$

$$g(f) = (f-3)^2$$

$$\frac{\partial g}{\partial a} = \frac{\partial g}{\partial f} \cdot \frac{\partial f}{\partial a}$$

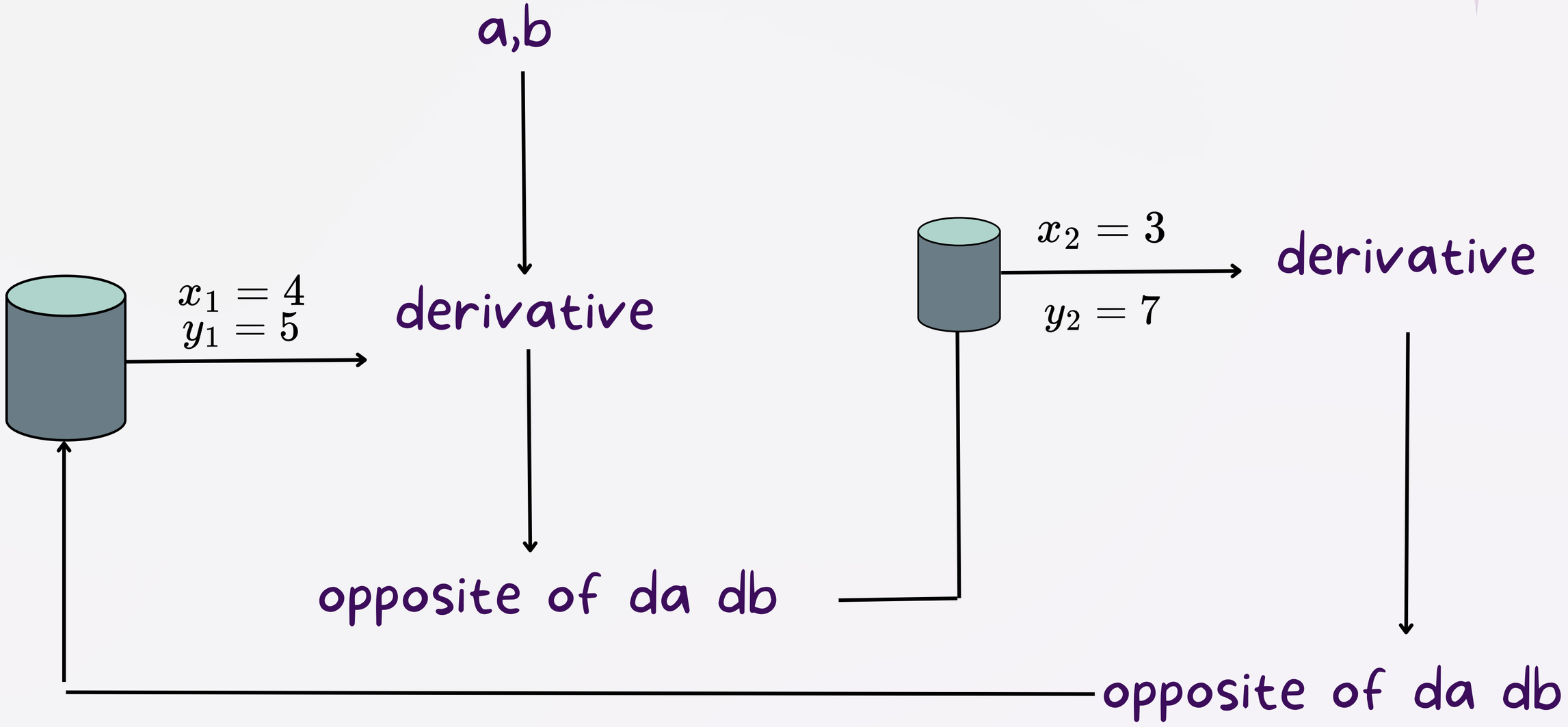
$$= 2x\,(f-3)$$

$$\frac{\partial g}{\partial b} = \frac{\partial g}{\partial f} \cdot \frac{\partial f}{\partial b}$$

$$= 2\,(f-3)$$

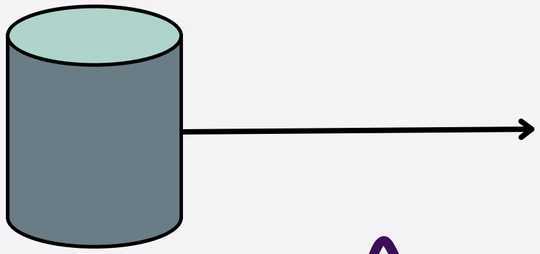
$$\theta_t = \theta_{t-1} - \eta g'(\theta_{t-1})$$

Multi sample: way 1



Multi sample: way 2

a,b



derivative

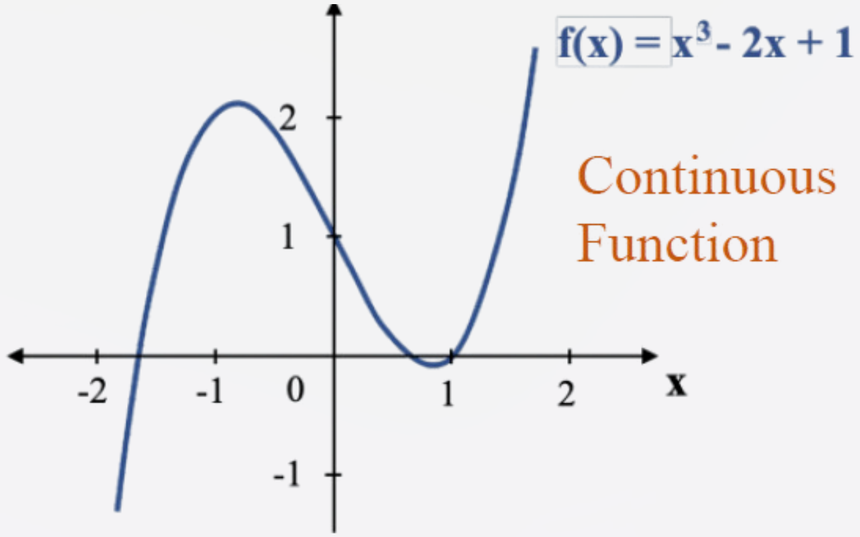
$$\frac{1}{i}\sum_i \frac{\partial g_i}{\partial a} = \frac{1}{i}\left(\frac{\partial g_1}{\partial f_1}\frac{\partial f_1}{\partial a} + \cdots + \frac{\partial g_i}{\partial f_i}\frac{\partial f_i}{\partial a}\right)$$

Average the gradient

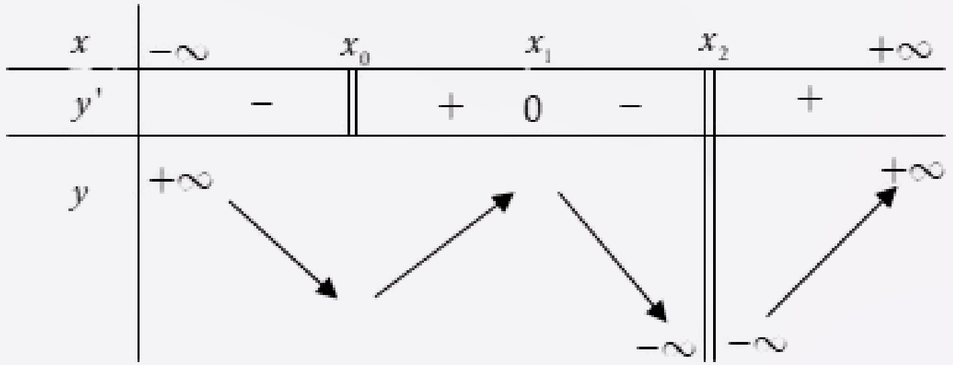
$$\frac{1}{i}\sum_i \frac{\partial g_i}{\partial b} = \frac{1}{i}\left(\frac{\partial g_1}{\partial f_1}\frac{\partial f_1}{\partial b} + \cdots + \frac{\partial g_i}{\partial f_i}\frac{\partial f_i}{\partial b}\right)$$

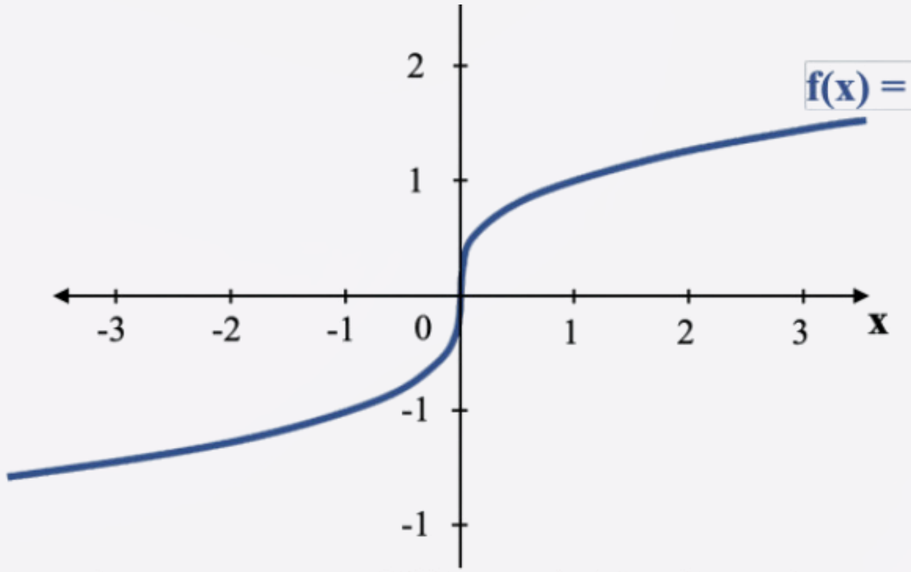
opposite of da db

Loss Functions (cost functions)

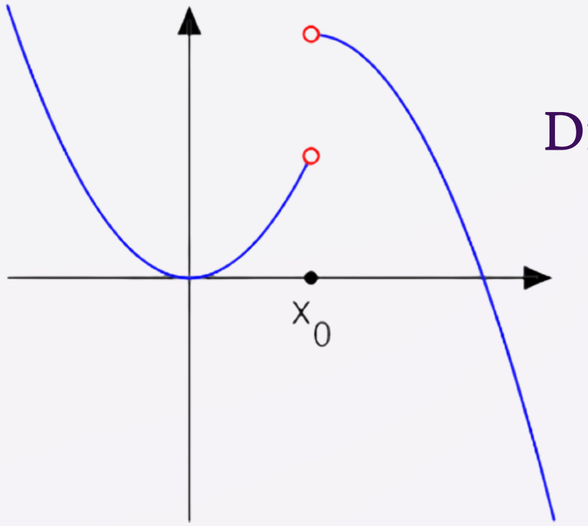


Continuous Function



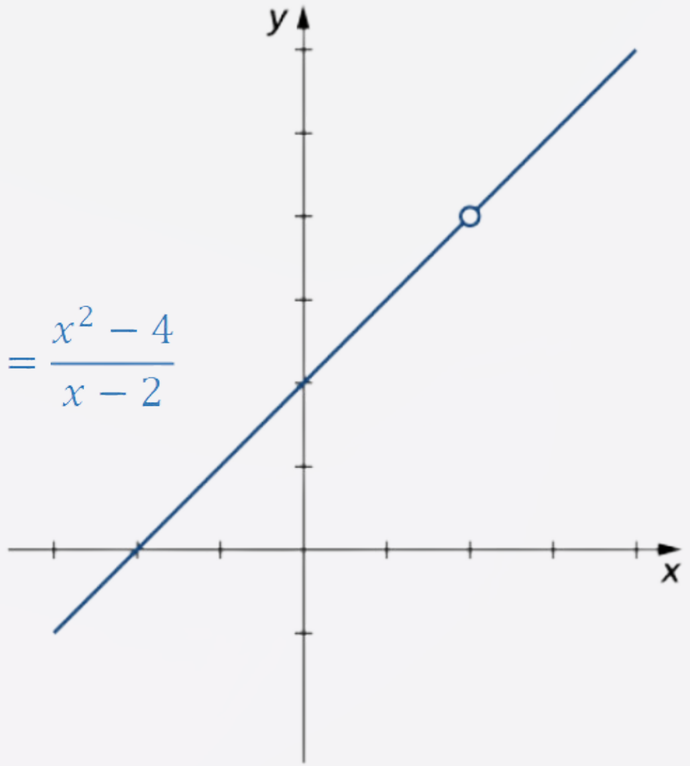


Continuous non-differentiable functions



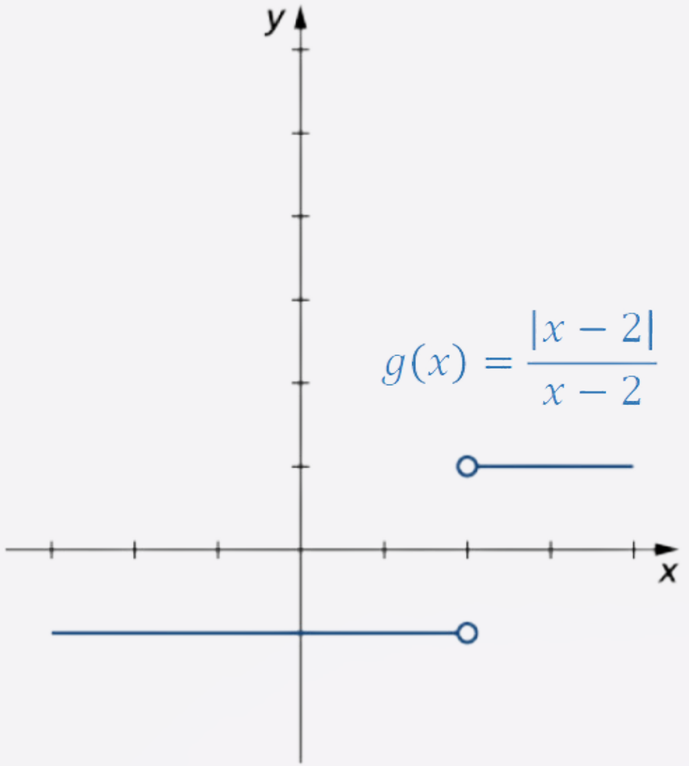
Discontinuous Functions

Conditions for Loss Function



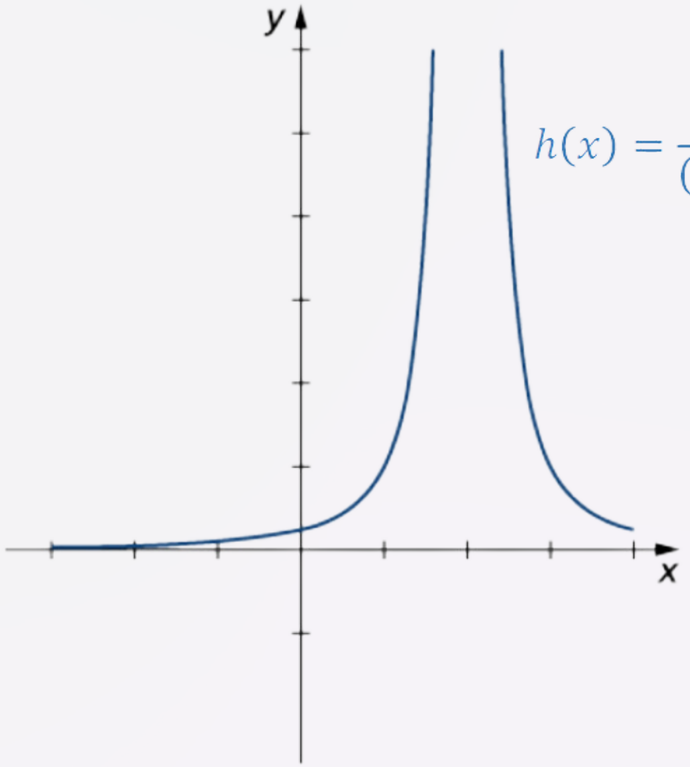
$$f(x) = \frac{x^2 - 4}{x - 2}$$

$$\lim_{x \to 2} f(x) = ?$$



$$g(x) = \frac{|x - 2|}{x - 2}$$

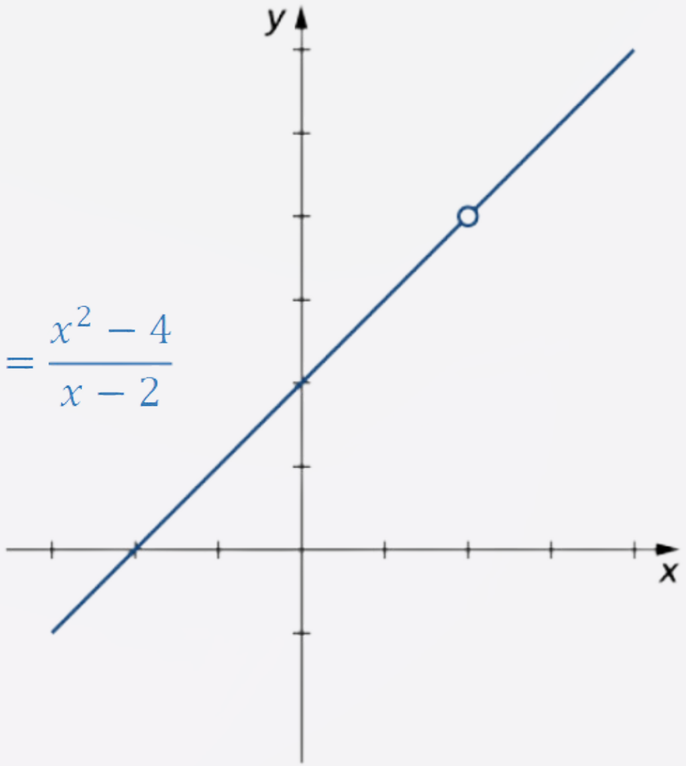
$$\lim_{x \to 2} g(x) = ?$$



$$h(x) = \frac{1}{(x - 2)^2}$$

$$\lim_{x \to 2} h(x) = ?$$

Conditions for Loss Function



$f(x) = \dfrac{x^2 - 4}{x - 2}$

$\lim_{x \to 2} f(x) = ?$

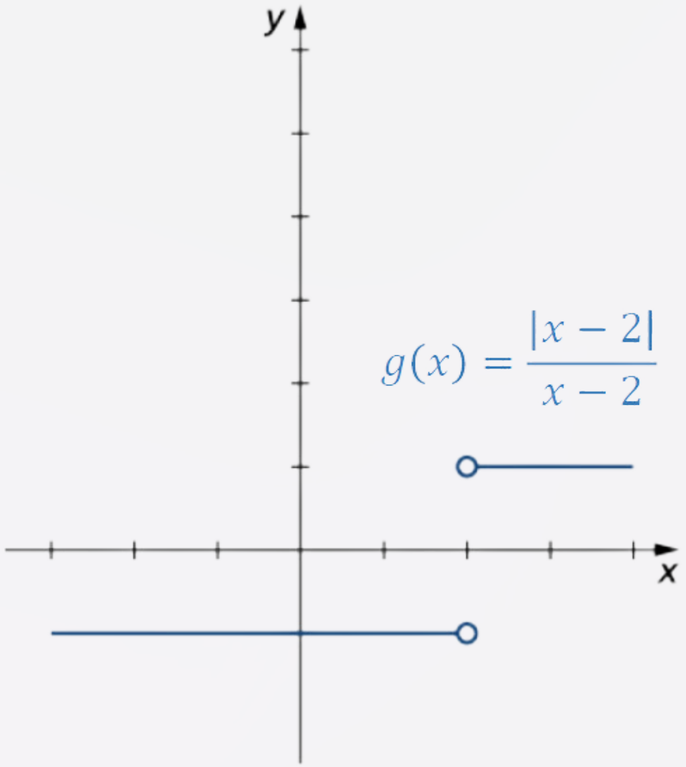
we can simplify $f(x)$:

$$\frac{x^2 - 4}{x - 2} = \frac{(x - 2)(x + 2)}{x - 2}$$

Simplifying gives $f(x) = x + 2$ for $x \neq 2$. Therefore, the limit as $x \to 2$ is:

$$\lim_{x \to 2} f(x) = 4$$

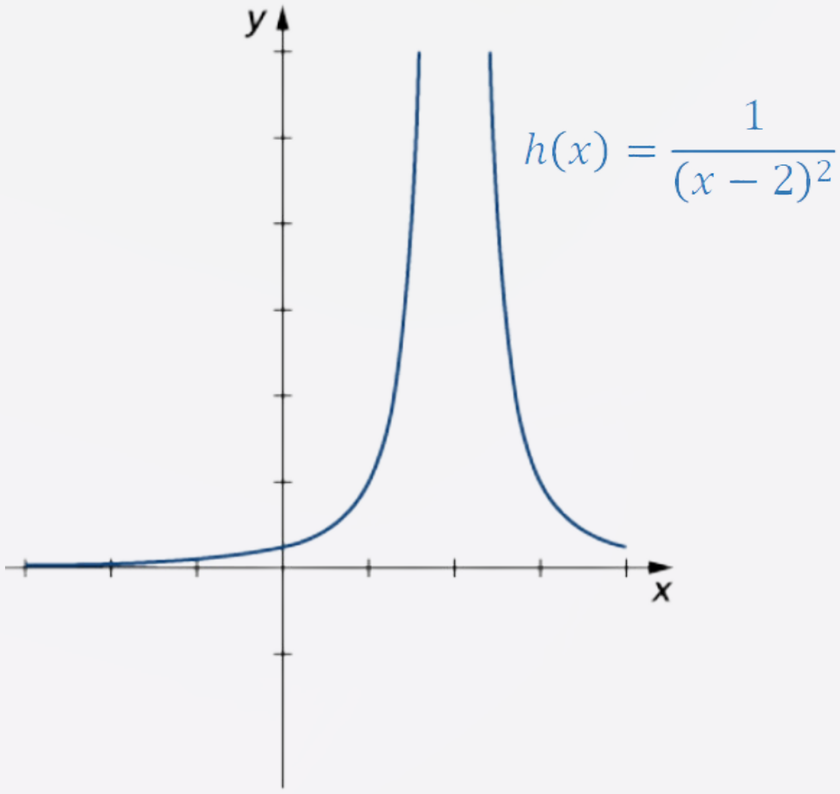
Conditions for Loss Function



$\lim_{x \to 2} g(x) = ?$

$\lim_{x \to 2} g(x)$ does not exist.

Conditions for Loss Function

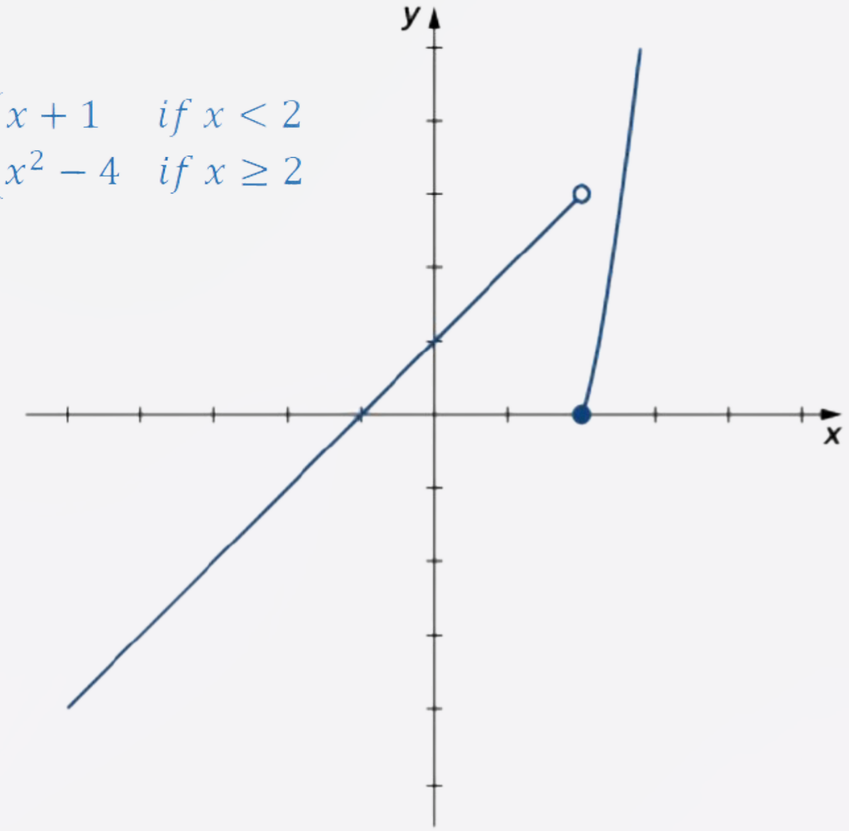


$$\lim_{x \to 2} h(x) = ?$$

$$\lim_{x \to 2} h(x) = \infty$$

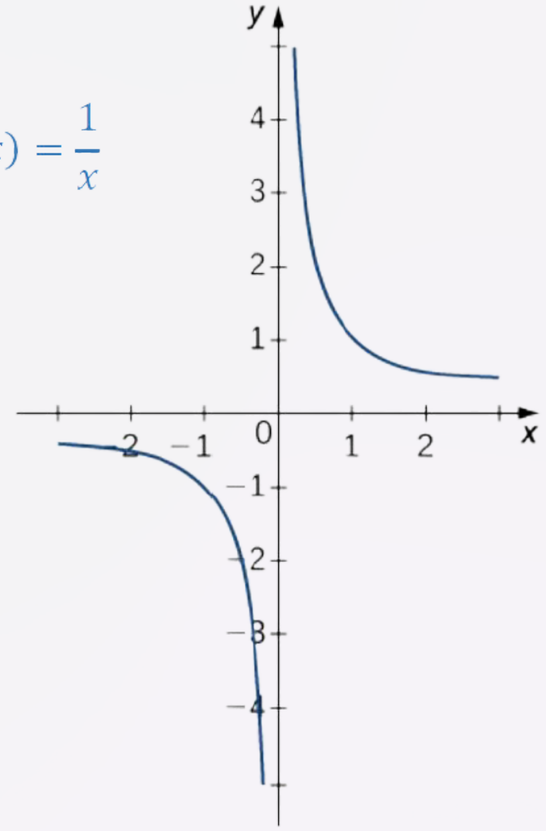
Conditions for Loss Function

$$f(x) = \begin{cases} x + 1 & if\ x < 2 \\ x^2 - 4 & if\ x \geq 2 \end{cases}$$



$$\lim_{x \to 2} f(x) = ?$$

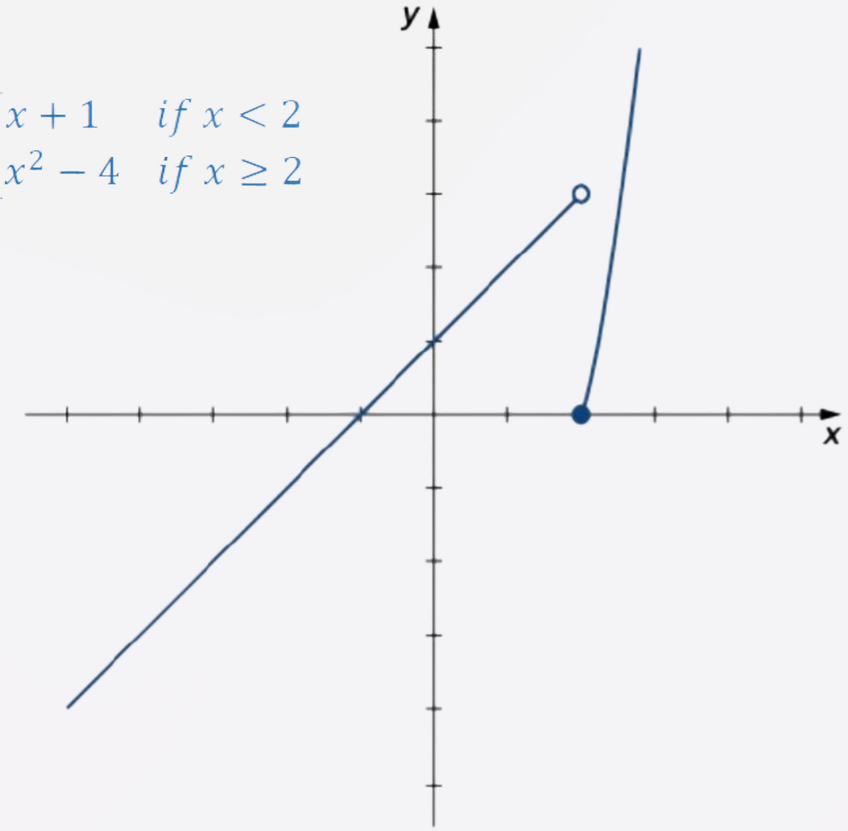
$$g(x) = \frac{1}{x}$$



$$\lim_{x \to 0} g(x) = ?$$

Conditions for Loss Function

$$f(x) = \begin{cases} x+1 & if\ x < 2 \\ x^2 - 4 & if\ x \geq 2 \end{cases}$$



$$\lim_{x \to 2} f(x) = ?$$

From the left of 2 ($x \to 2^-$):

When $x < 2$, $f(x) = x + 1$. So,

$$\lim_{x \to 2^-} f(x) = 2 + 1 = 3$$

From the right of 2 ($x \to 2^+$):

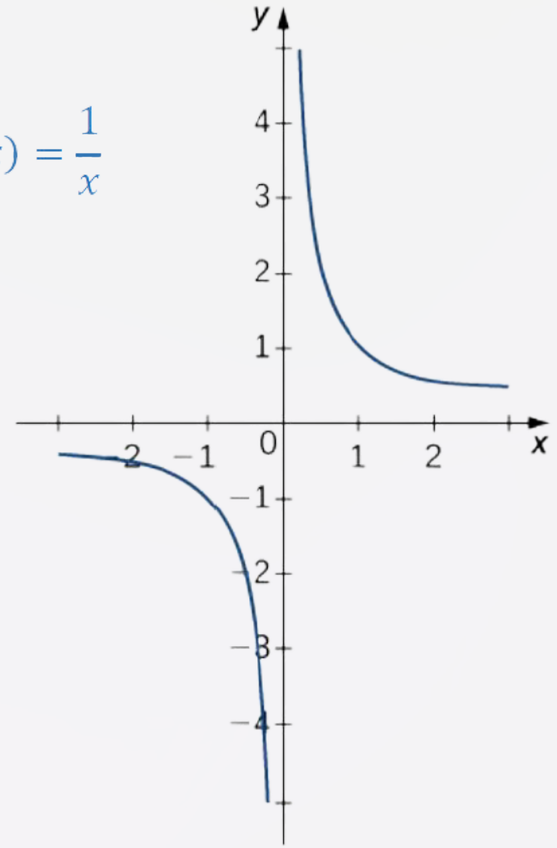
When $x \geq 2$, $f(x) = x^2 - 4$. So,

$$\lim_{x \to 2^+} f(x) = 2^2 - 4 = 0$$

$$\lim_{x \to 2} f(x) \text{ does not exist.}$$

Conditions for Loss Function

$g(x) = \frac{1}{x}$



$\lim_{x \to 0} g(x) = ?$

As $x \to 0^-$ (from the left):

$g(x) = \frac{1}{x}$, so when x approaches 0 from the negative side, $g(x) \to -\infty$.

As $x \to 0^+$ (from the right):

When $x \to 0^+$, $g(x) \to +\infty$.

$\lim_{x \to 0} g(x)$ does not exist.

Conditions for Loss Function

1. Continuity

DEFINITION:

A function **f(x)** is **continuous** at a point a ***if and only if*** the following three conditions are satisfied:

1. $f(a)$ is defined.
2. $\lim_{x \to a} f(x)$ exists.
3. $\lim_{x \to a} f(x) = f(a)$.

A function is **discontinuous** at a point a if it fails to be continuous at a.

Conditions for Loss Function

1. Continuity

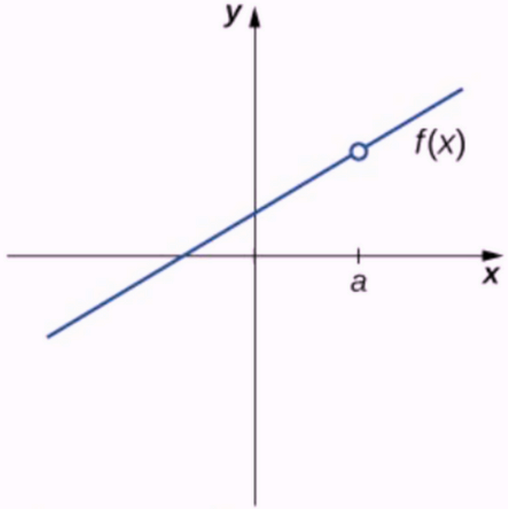


Figure 2.32 The function $f(x)$ is not continuous at a because $f(a)$ is undefined.

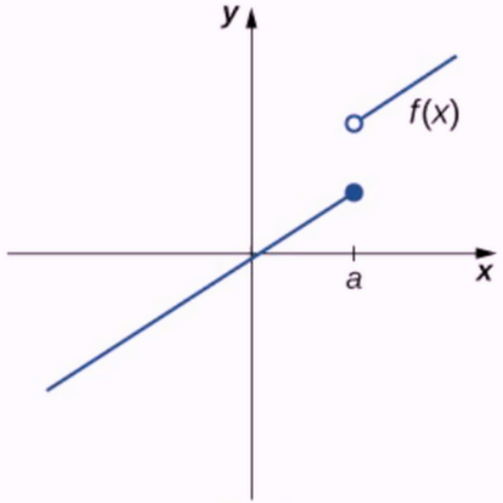


Figure 2.33 The function $f(x)$ is not continuous at a because $\lim_{x \to a} f(x)$ does not exist.

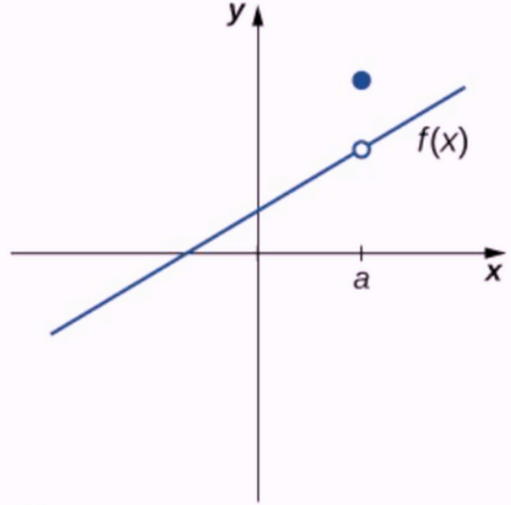


Figure 2.34 The function $f(x)$ is not continuous at a because $\lim_{x \to a} f(x) \neq f(a)$.

Conditions for Loss Function

2. Differentiability

DEFINITION:

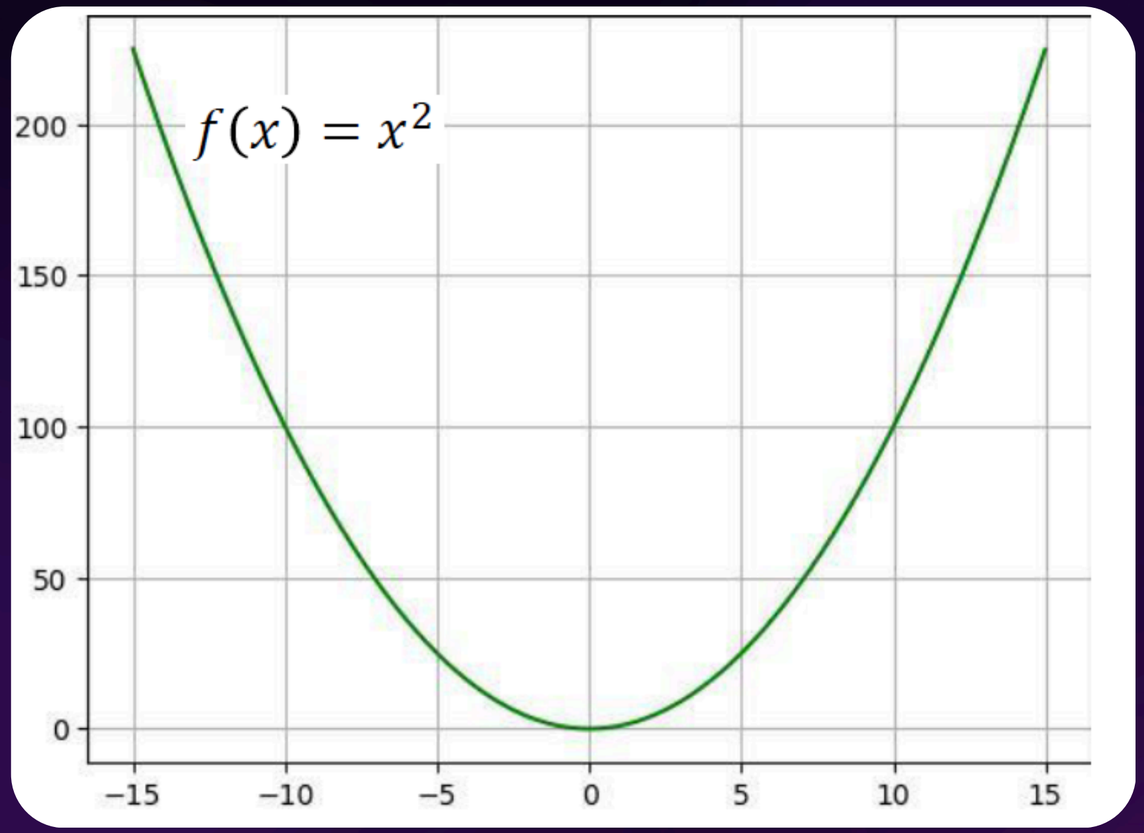
Let f be a function. The *derivative function*, denoted by f', is the function whose domain consists of the values of x such that the following limit exists:

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}.$$

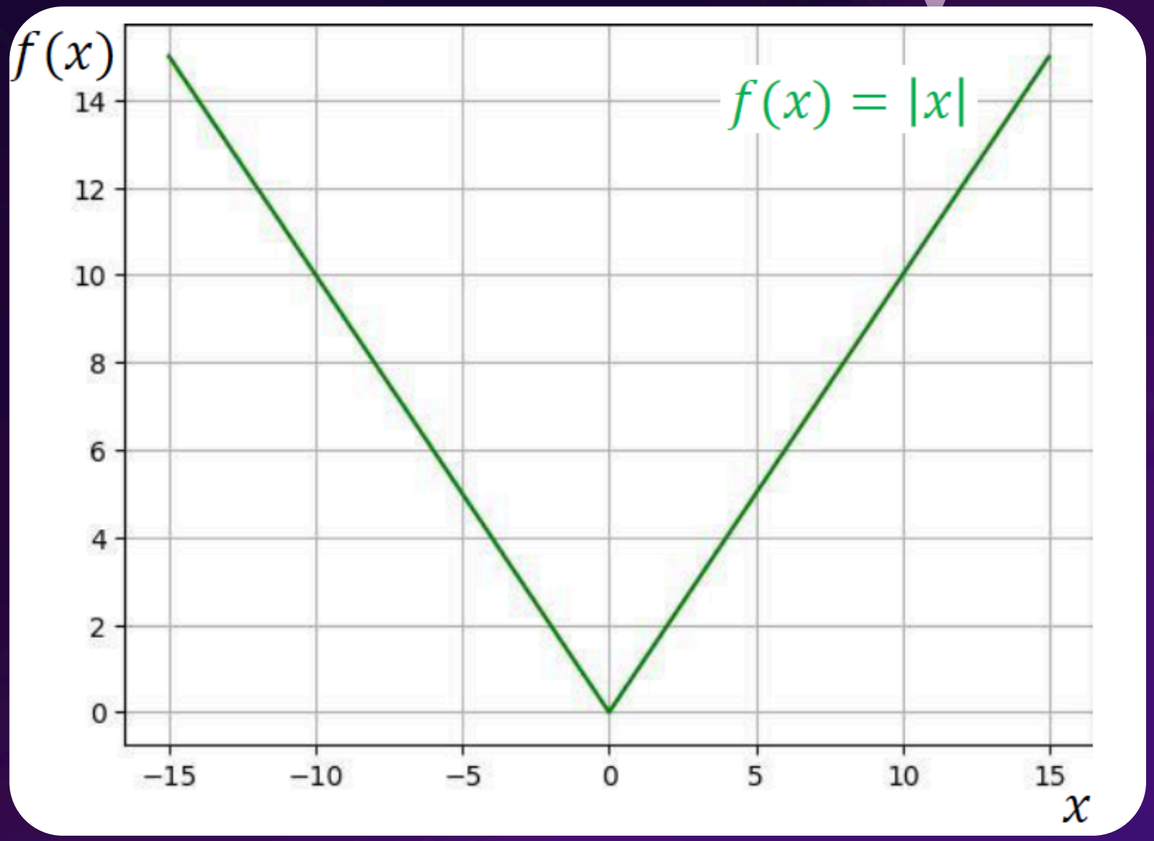
A function $f(x)$ is said to be *differentiable at* a if $f'(a)$ exists. More generally, a function is said to be *differentiable on* S if it is differentiable at every point in an open set S, and a *differentiable function* is one in which $f'(x)$ exists on its domain.

Conditions for Loss Function

Check if the function is continuous and differentiable

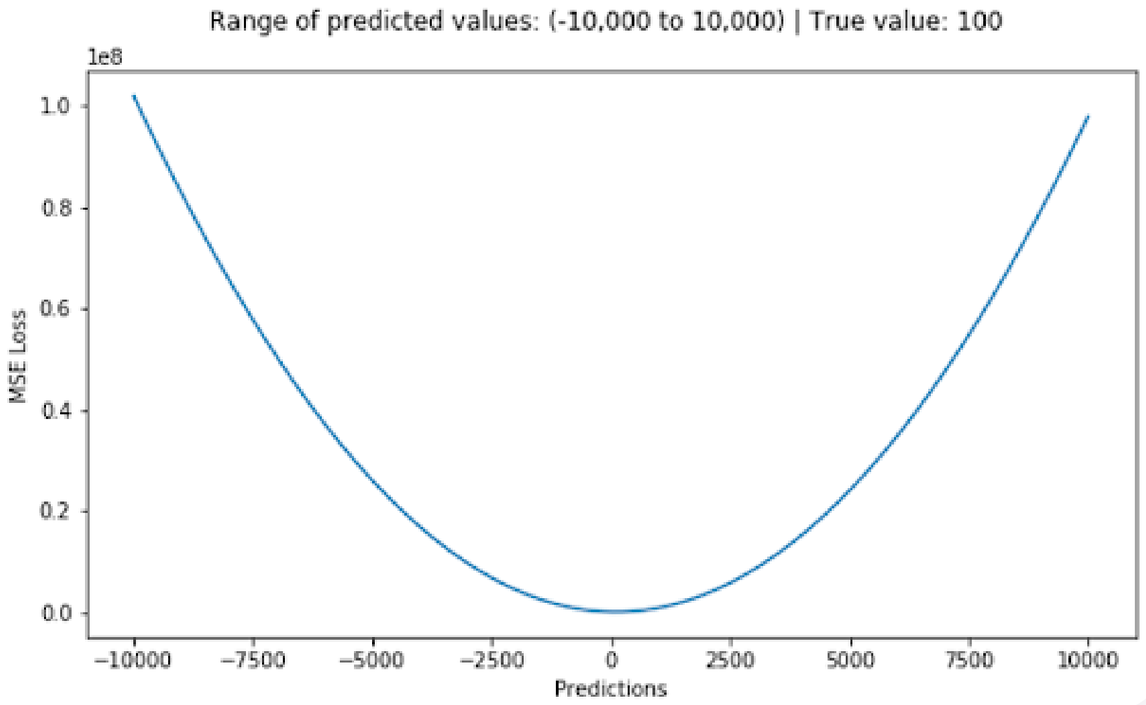


The function $f(x) = x^2$ is both continuous and differentiable everywhere in its domain $(-\infty, \infty)$.



The function $f(x) = |x|$ is continuous everywhere on $(-\infty, \infty)$, but it is not differentiable at $x = 0$ due to the sharp corner at that point.

Mean Square Error (MSE)

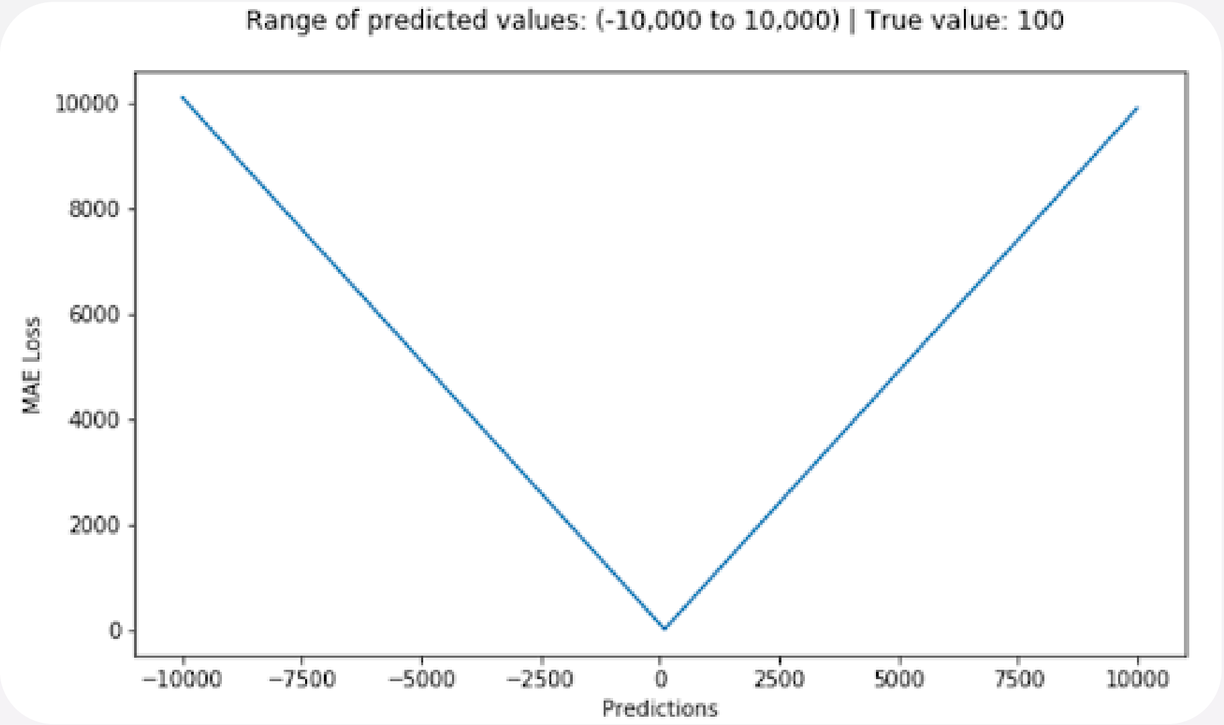


One example $L(\hat{y}, y) = (\hat{y} - y)^2$

N example $L(\hat{y}, y) = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$

- n is the number of observations (data points).
- y_i is the actual value for observation i.
- $\hat{y}_i$ is the predicted value for observation i.

Mean Absolute Error (MAE)

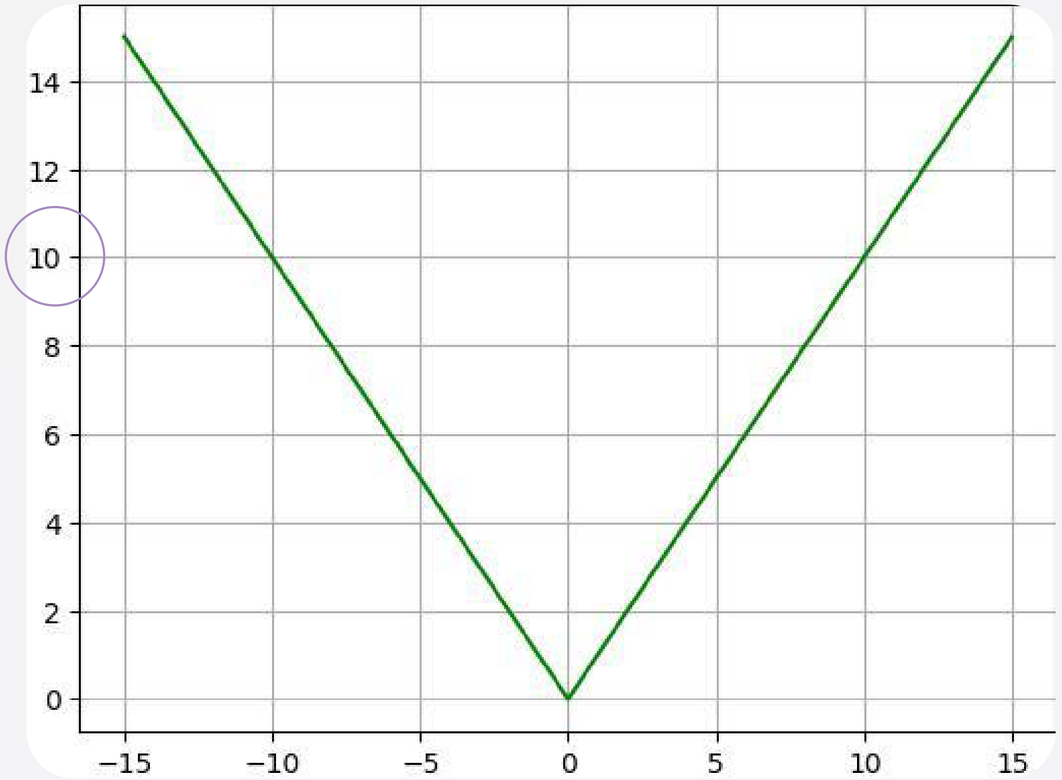


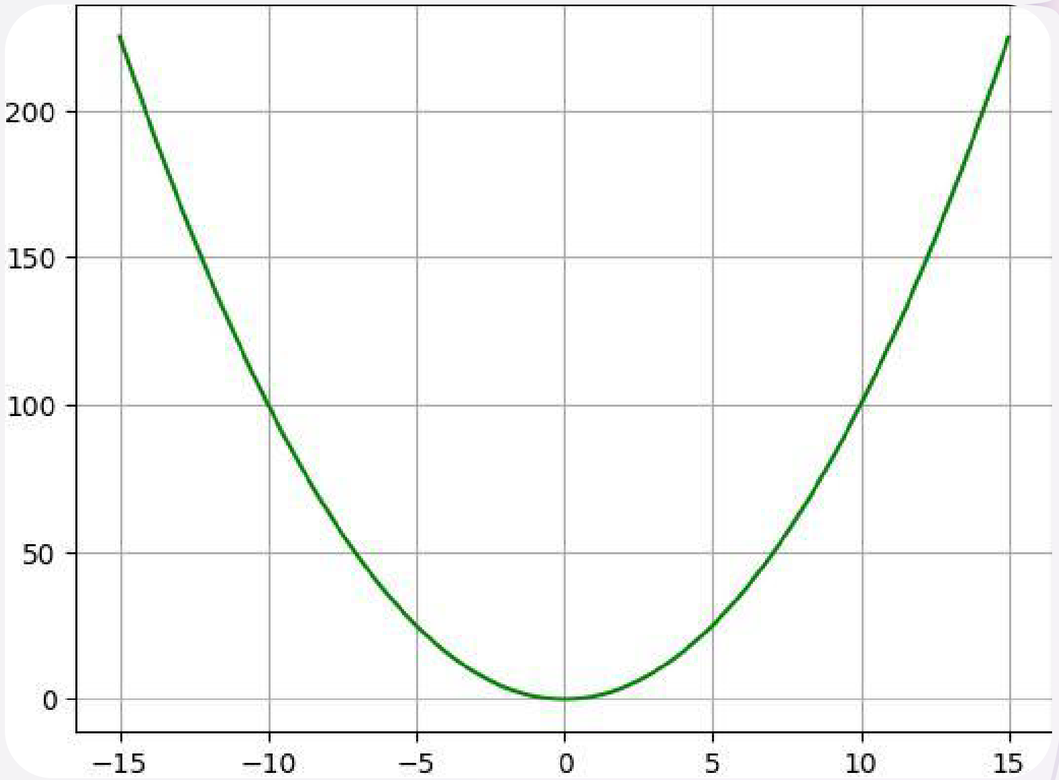
One example $L(\hat{y}, y) = |\hat{y} - y|$

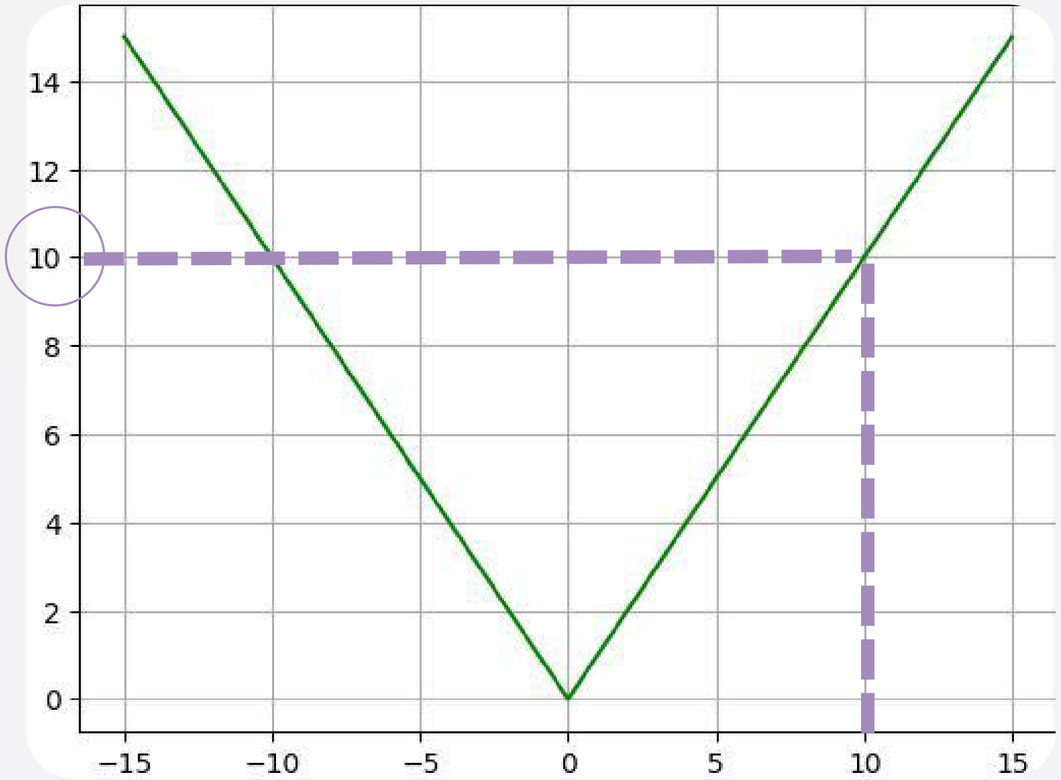
N example $L(\hat{y}, y) = \frac{1}{N} \sum_{i=1}^{N} |\hat{y}_i - y_i|$

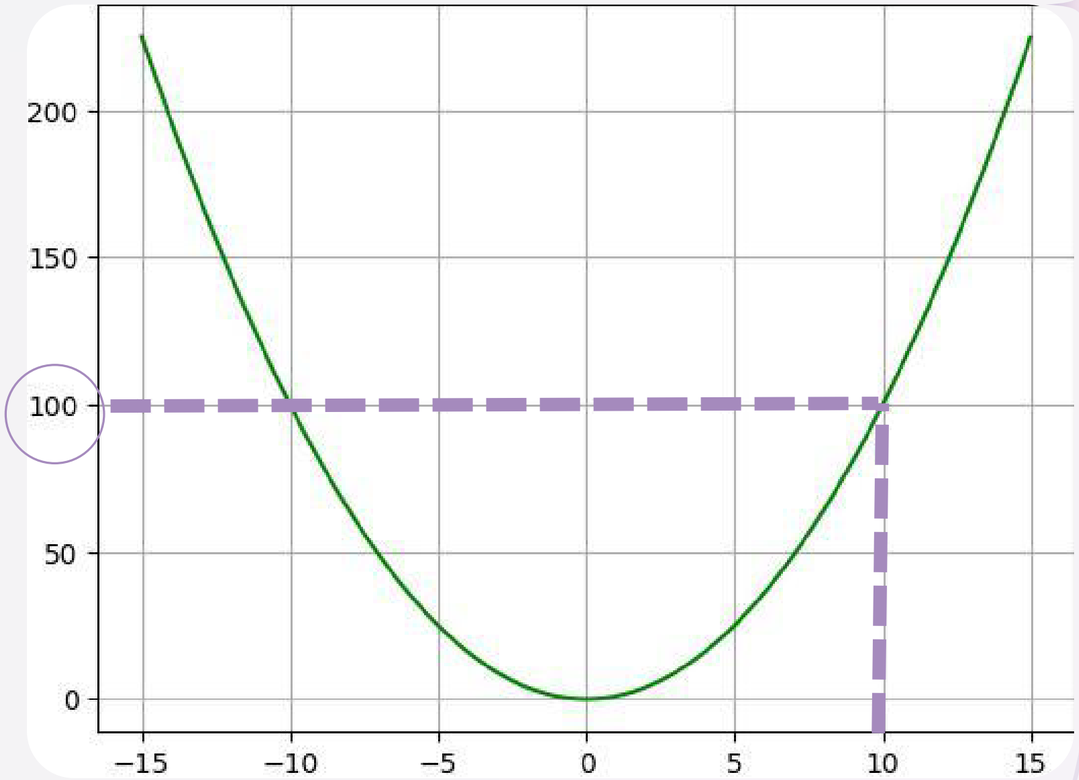
- n is the number of observations (data points).
- y_i is the actual value for observation i.
- $\hat{y}_i$ is the predicted value for observation i.
- $|y_i - \hat{y}_i|$ represents the absolute error for each observation

The pros and cons of MSE and MAE when data contain outliers?


14
12
10
8
6
4
2
0
−15
−10
−5
0
5
10
15


200
150
100
50
0
−15
−10
−5
0
5
10
15





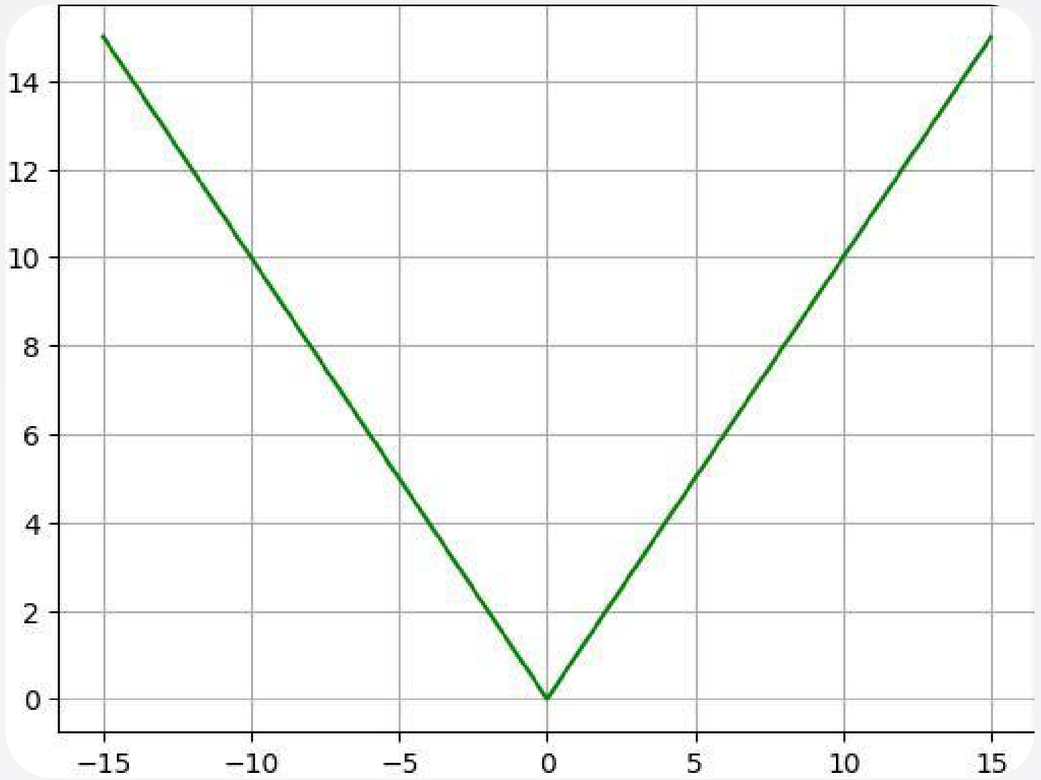
$g(10) \gg f(10)$

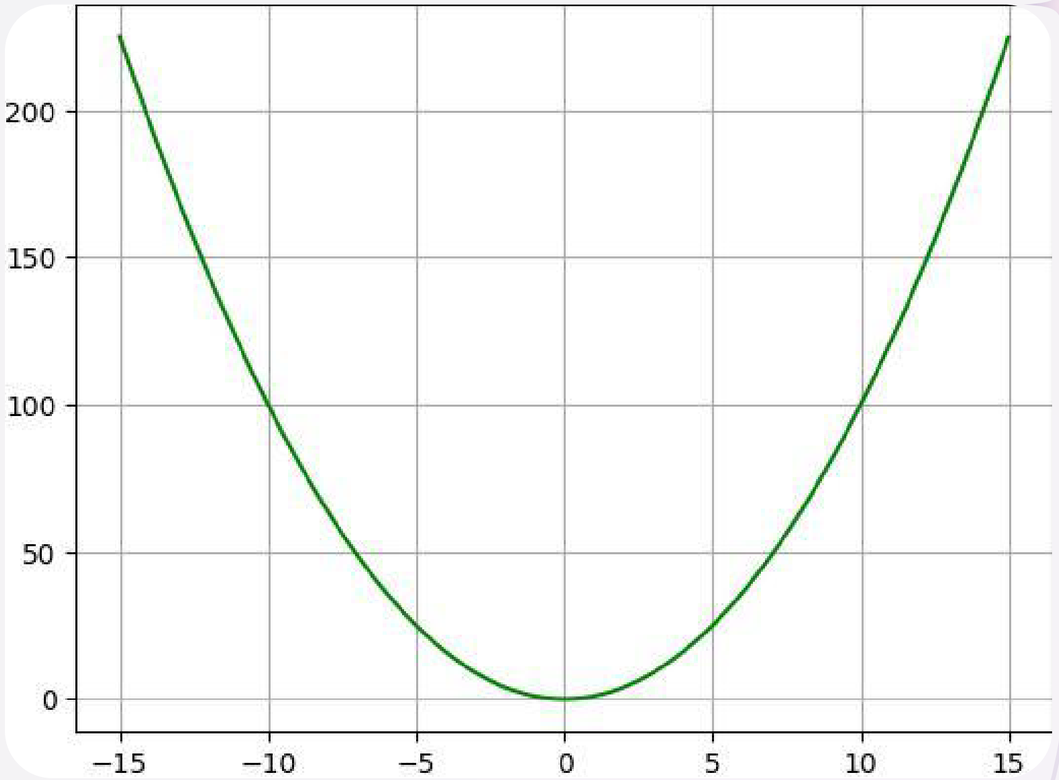
If $x = 10$ is an outlier, $g(10)$ has more negative effect

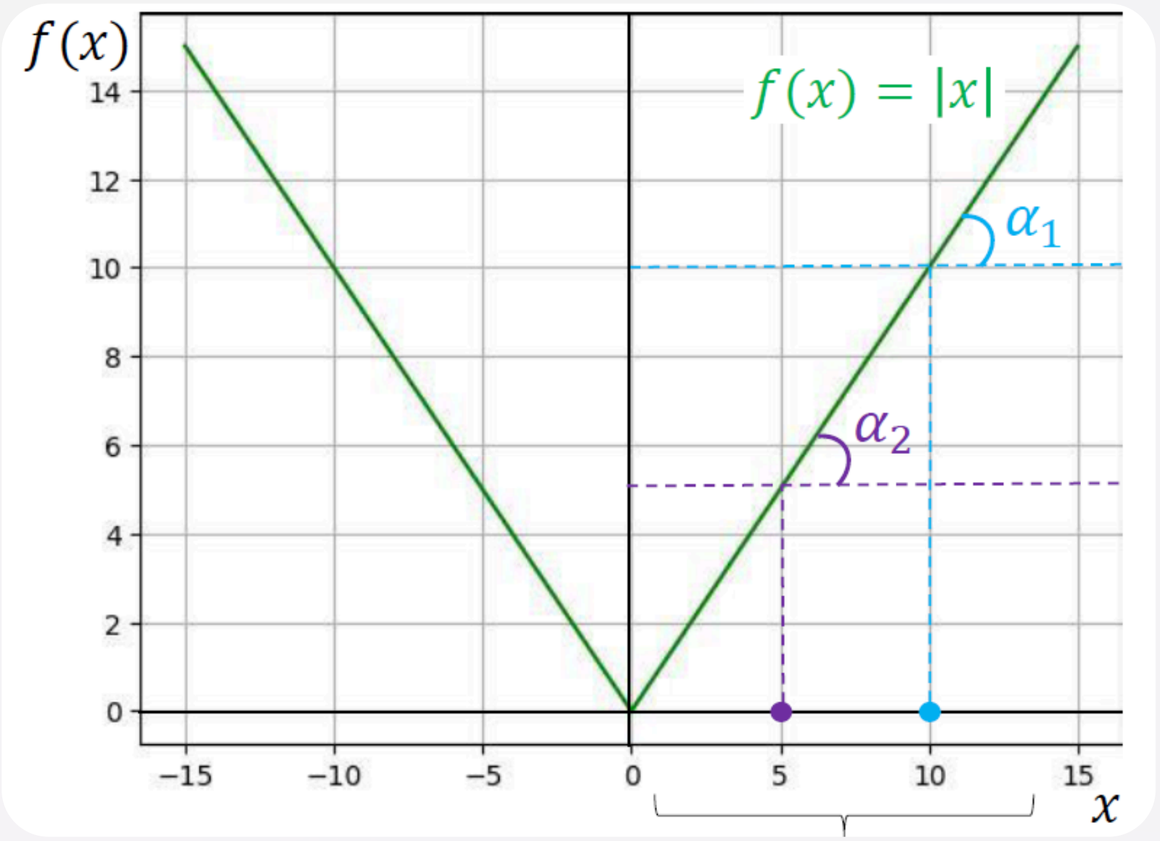
$\Rightarrow$ MAE is better to tolerate outliers

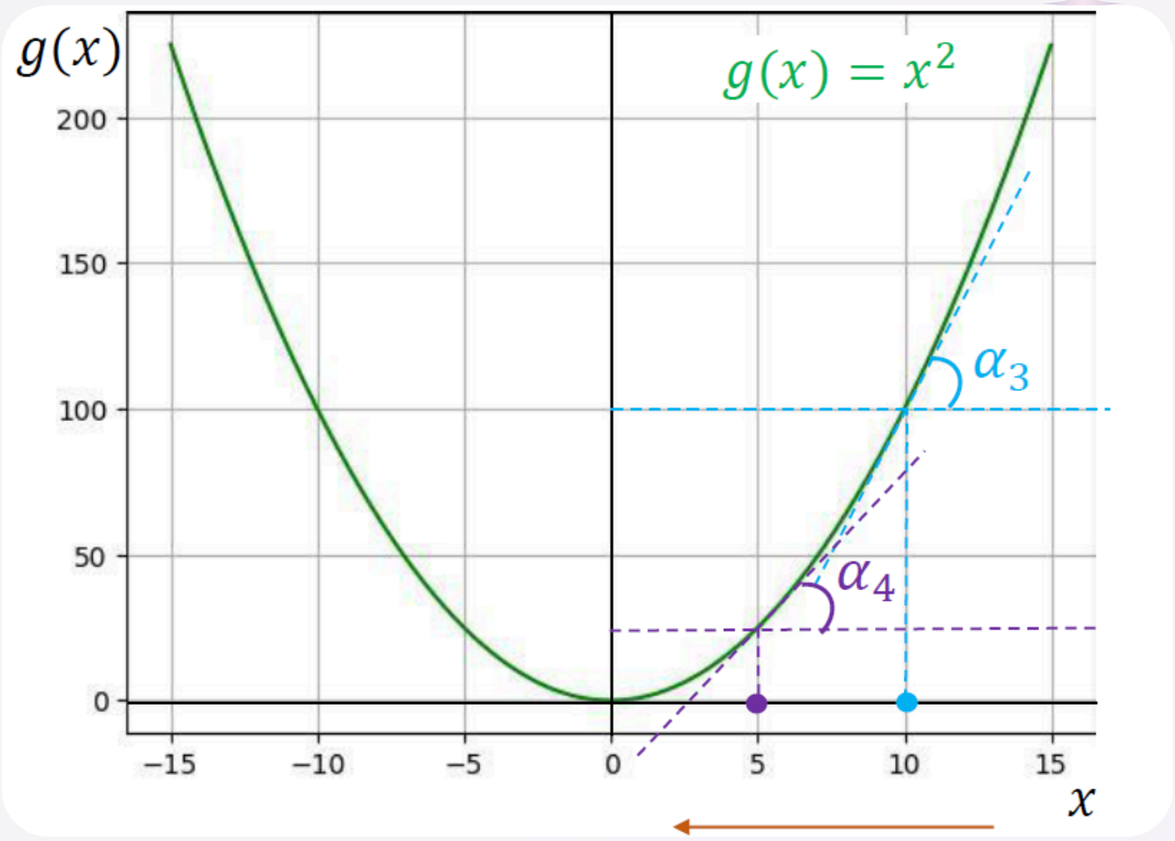
The pros and cons of MSE and MAE

with a fixed learning rate η?


14
12
10
8
6
4
2
0
−15
−10
−5
0
5
10
15


200
150
100
50
0
−15
−10
−5
0
5
10
15





$\alpha 1 = \alpha 2$

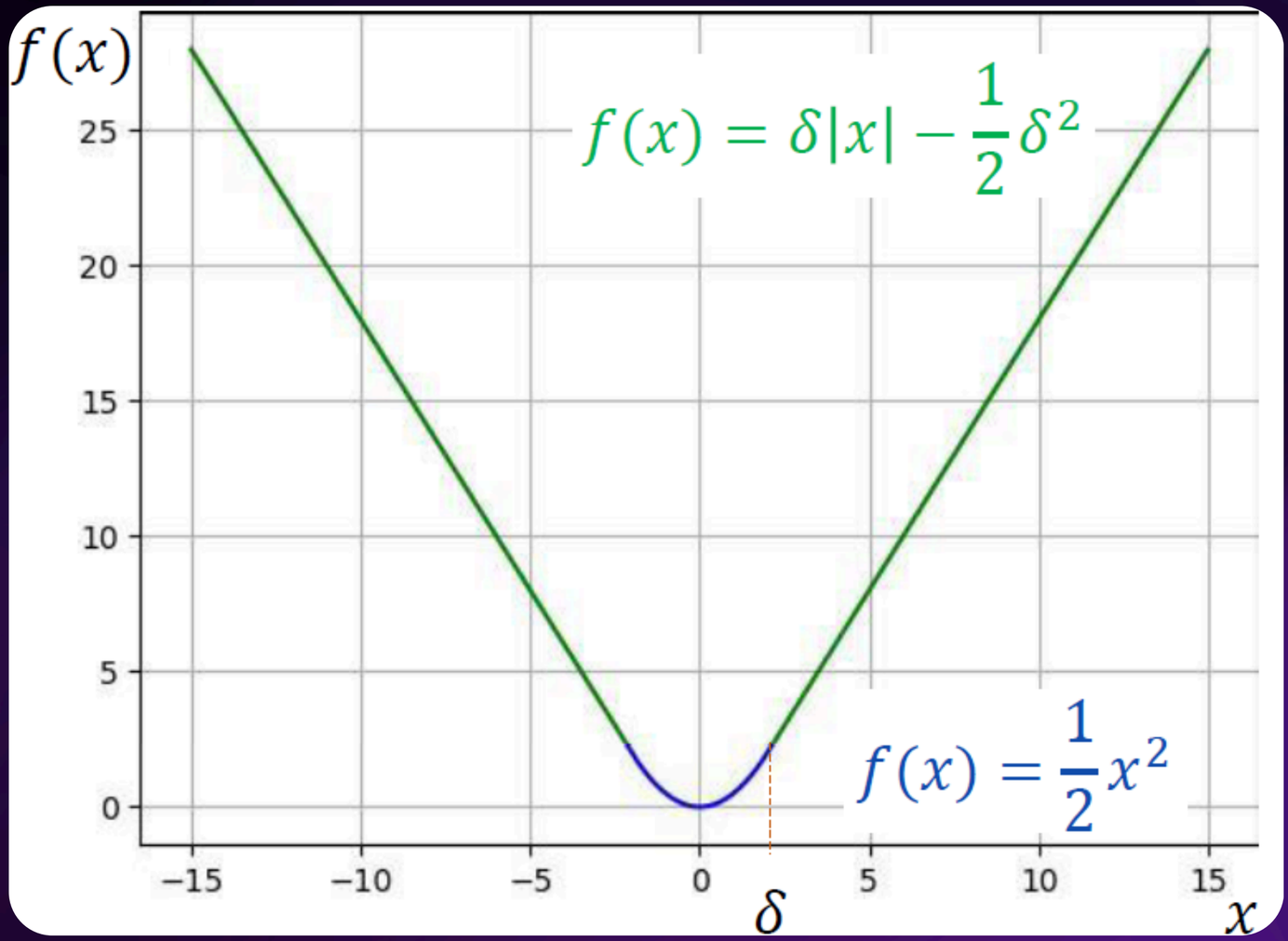
$\eta f'(x)$ values are constants

$\alpha 3 = \alpha 4$

$\eta f'(x)$ values reduce

$\Rightarrow$MSE is better when working with a fixed learning rate

Huber loss


$f(x)$
$f(x) = \delta|x| - \frac{1}{2}\delta^2$
$f(x) = \frac{1}{2}x^2$
25
20
15
10
5
0
−15
−10
−5
0
5
10
15
δ
x

Huber loss

1. **Pick all the N samples from training data**
2. **Compute output** $\hat{y} = wx + b$
3. **Compute loss**

$$L(\hat{y}, y) = \begin{cases} \frac{1}{2}(\hat{y} - y)^2 & \text{for } |\hat{y} - y| \leq \delta \\ \delta|\hat{y} - y| - \frac{1}{2}\delta^2 & \text{otherwise} \end{cases}$$

4. **Compute derivative**

$$L'_w = \begin{cases} x(\hat{y} - y) & \text{for } |\hat{y} - y| \leq \delta \\ \frac{\delta x(\hat{y} - y)}{|\hat{y} - y|} & \text{otherwise} \end{cases}$$

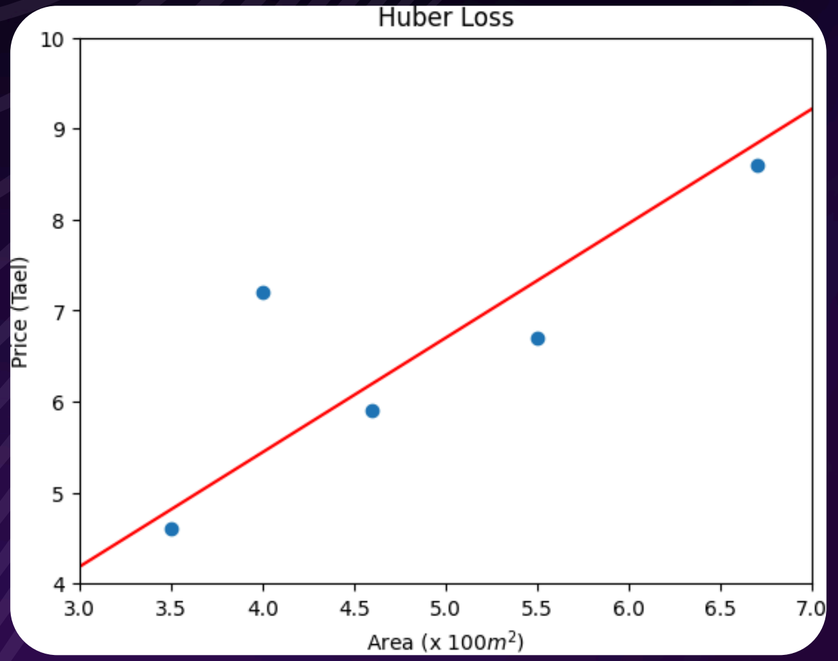
$$L'_b = \begin{cases} (\hat{y} - y) & \text{for } |\hat{y} - y| \leq \delta \\ \frac{\delta(\hat{y} - y)}{|\hat{y} - y|} & \text{otherwise} \end{cases}$$

Where:

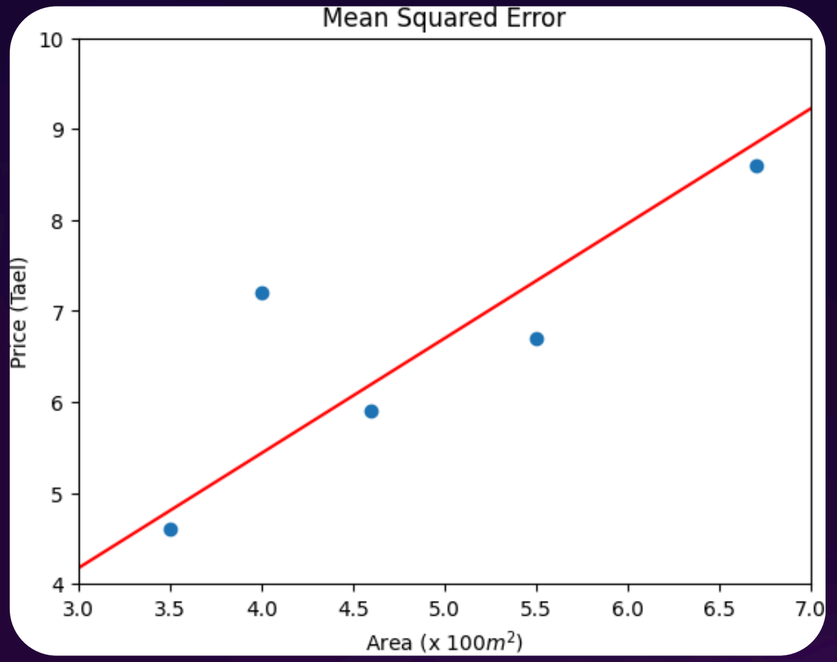
- $L(\hat{y}, y)$ is the loss function.
- L'_w and L'_b are the derivatives with respect to w and b
- $\hat{y}$ is the predicted value.
- y is the true value.
- δ is a threshold.

Comparison (outliers)

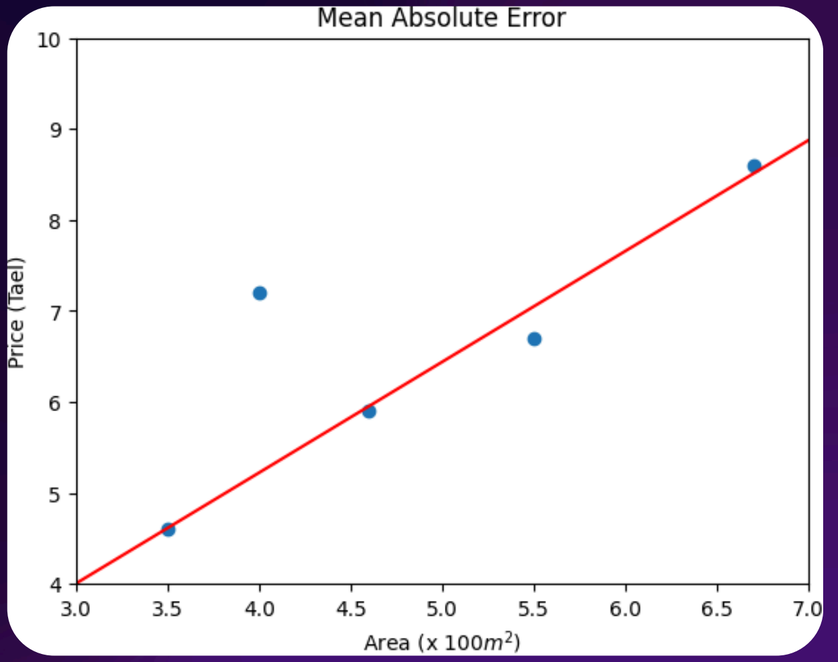
Batch training



Huber Loss



MSE



MAE

Data Normalization

Features			Label
TV	Radio	Newspaper	Sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	12
151.5	41.3	58.5	16.5
180.8	10.8	58.4	17.9

Unnormalized data

$$x_1 = 230.1$$
$$x_2 = 37.8$$
$$x_3 = 69.2$$
$$y = 22.1$$
$$\hat{y} = 4.0947591112215855$$

$$\frac{\partial L}{\partial w_1} = -8286.011857015827$$
$$\frac{\partial L}{\partial w_2} = -1361.196211196482$$
$$\frac{\partial L}{\partial w_3} = -2491.9253390069933$$
$$\frac{\partial L}{\partial b} = -36.01048177755683$$

$$w_1 = 0.98995518326565295$$
$$w_2 = 0.019689747124005393$$
$$w_3 = 0.027253710779249984$$
$$b = 0.00061048177755684$$

Normalized data

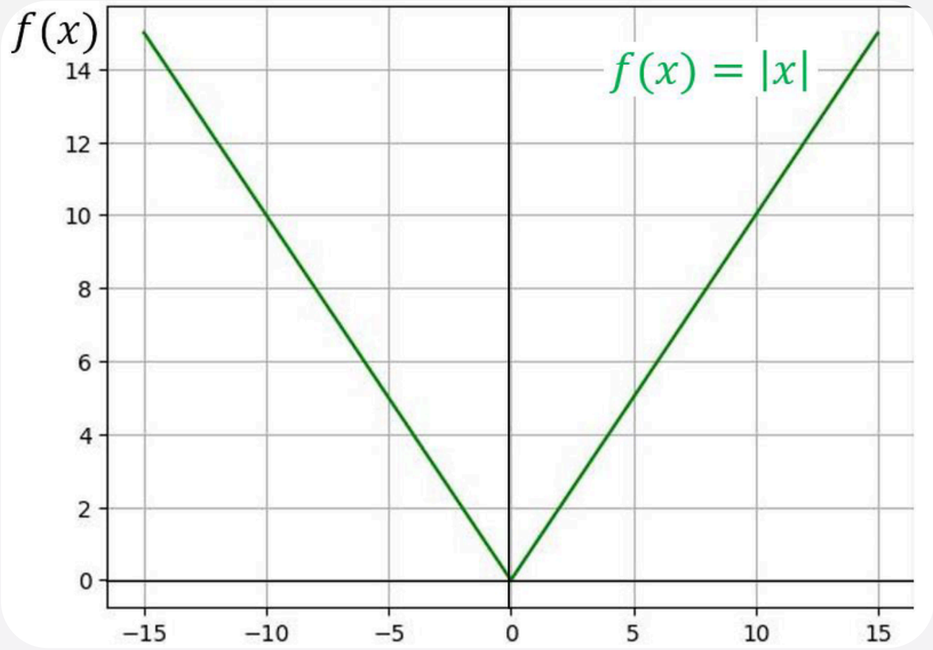
$$X_{\text{norm}} = \frac{X - X_{\text{mean}}}{X_{\text{max}} - X_{\text{min}}}$$

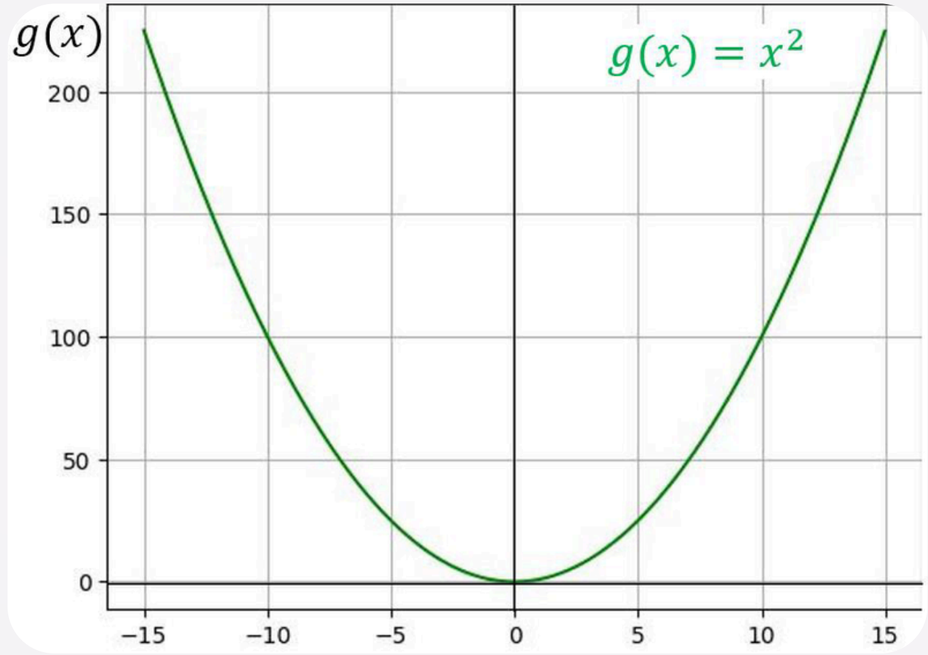
$$w_1 = 0.01609506469549467$$
$$w_2 = 0.006077785012088 91$$
$$w_3 = 0.0023344573891806507$$
$$b = 0$$

$$x_1 = 0.5504267881241568$$
$$x_2 = -0.09835863697705782$$
$$x_3 = 0.007579284750337614$$
$$y = 22.1$$
$$y\_hat = 0.008279045632653116$$

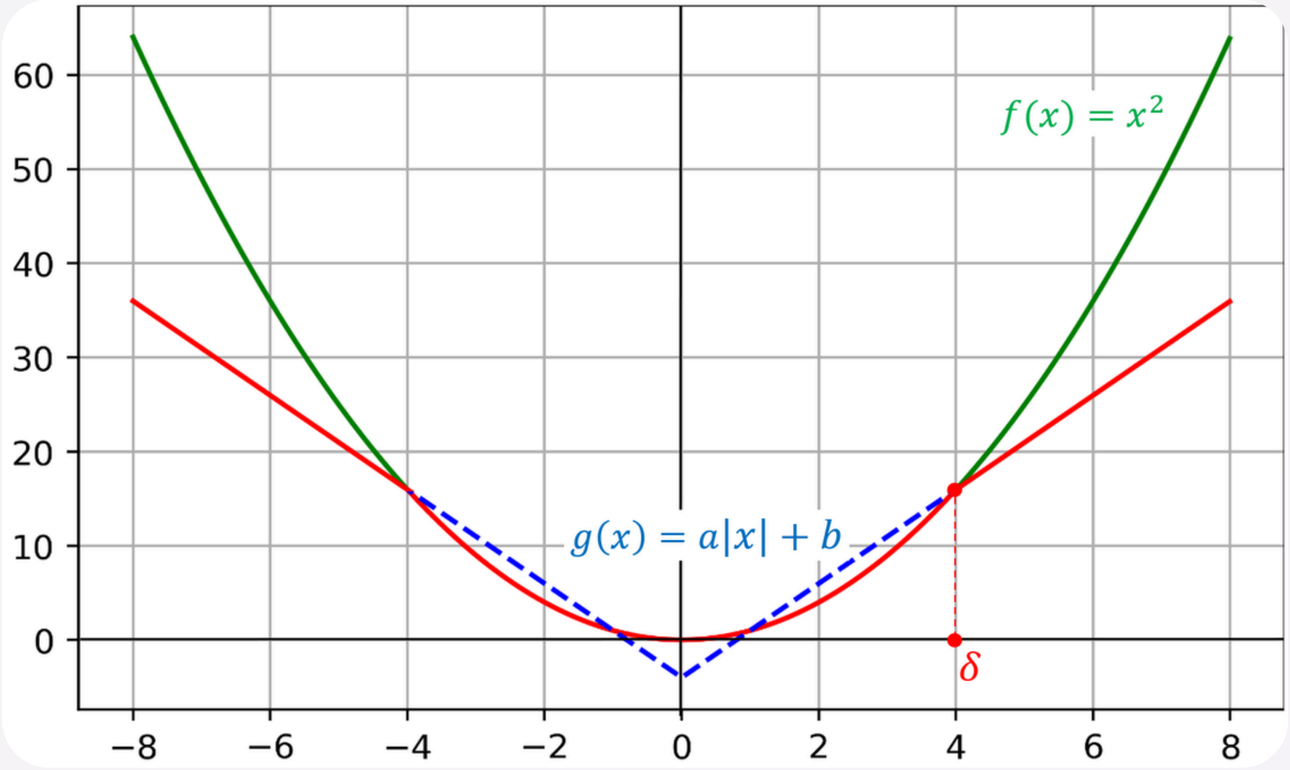
$$\frac{\partial L}{\partial w_1} = -24.319750018095103$$
$$\frac{\partial L}{\partial w_2} = 4.34582132908159$$
$$\frac{\partial L}{\partial w_3} = -0.33487888747630074$$
$$\frac{\partial L}{\partial b} = -44.1834419087347$$

$$w_1 = 0.2592925648764457$$
$$w_2 = 0.037380846218892686$$
$$w_3 = 0.005683246263943658$$
$$b = 0.441834419087347$$


$f(x)$
$f(x) = |x|$
14
12
10
8
6
4
2
0
−15
−10
−5
0
5
10
15


$g(x)$
$g(x) = x^2$
200
150
100
50
0
−15
−10
−5
0
5
10
15

Summary


$f(x) = x^2$
$g(x) = a|x| + b$
δ
60
50
40
30
20
10
0
−8
−6
−4
−2
0
2
4
6
8

1. **Pick all the N samples from training data**
2. **Compute output** $\hat{y}$

$$\hat{y} = X\theta$$

3. Compute loss

$$L = \frac{1}{N}(\hat{y} - y)^T(\hat{y} - y)$$

4. Compute gradient

$$k = 2(\hat{y} - y)$$

$$\nabla_\theta L = X^T k$$

5. Update parameters

$$\theta = \theta - \eta \frac{\nabla_\theta L}{N}$$ η is learning rate

Q&A SESSION