Machine Learning Tutorial

by Paulius Sasnauskas
sasnausk@ualberta.ca

2025-10-02

Part I: Fundamentals

Part II: Example Problems

Machine Learning

Machine Learning, Wikipedia

Machine Learning

- learn from data

Machine Learning, Wikipedia

Machine Learning

- learn from data
- generalize to unseen data

Machine Learning, Wikipedia

Machine Learning

- learn from data
- generalize to unseen data
- without explicit instructions

```
def has_ears(image):
    # find oval in image center
    circle = image.matchOval(image, x=100, y=100, w=30, h=20)

    # find two triangles
    triangle1 = image.matchTriangle(circle.x-20, circle.y)
    triangle2 = image.matchTriangle(circle.x+20, circle.y)

    if triangle1 and triangle2:
        return True
    else:
        return False
```

Machine Learning, Wikipedia

What Problems Can ML Solve?

- Many labeled examples
- Patterns in the data
- Complex relationships

Model

Model – a simplification of a real process or system.

Model

Model – a simplification of a real process or system.

x	y
2	19.6
4.5	44.1
10	98

Model

Model – a simplification of a real process or system.

x	y
2	19.6
4.5	44.1
10	98

y = ?

Model

Model – a simplification of a real process or system.

Usually we want to know what is the process that generates the data.

x	y
2	19.6
4.5	44.1
10	98

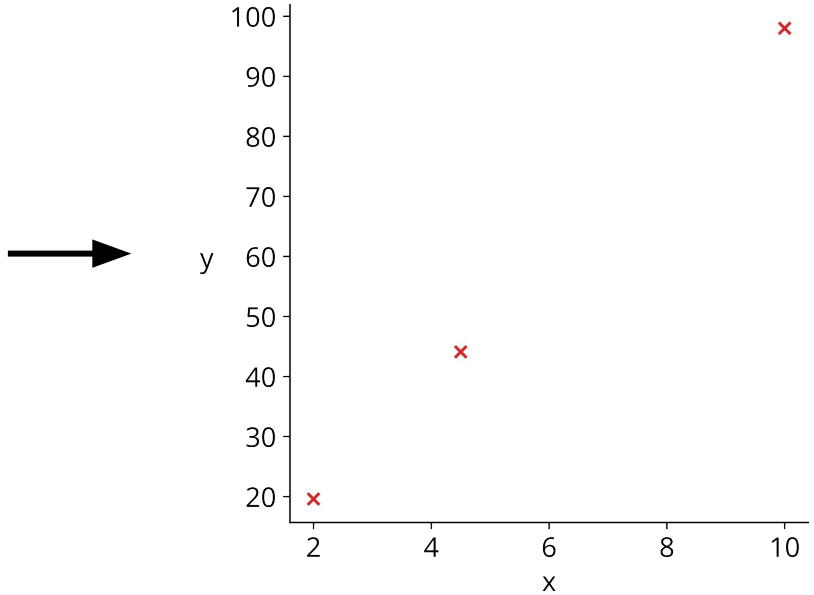
y = ?

Model

Model – a simplification of a real process or system.

Usually we want to know what is the process that generates the data.

x	y
2	19.6
4.5	44.1
10	98



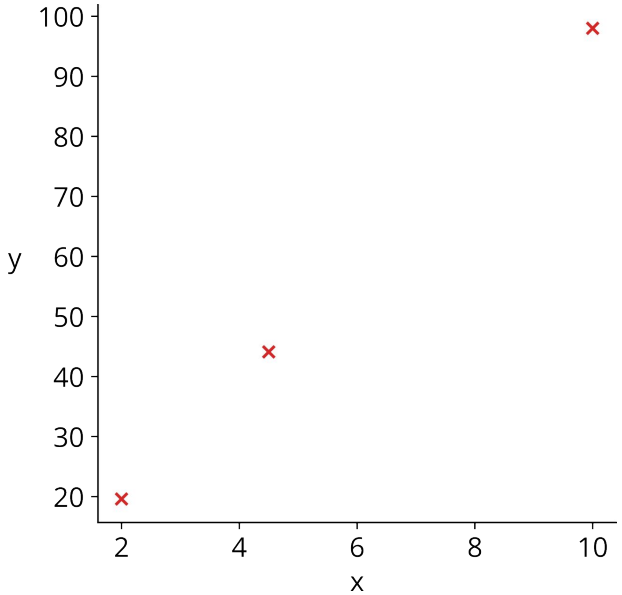
y = ?

Model

Model – a simplification of a real process or system.

Usually we want to know what is the process that generates the data.

x	y
2	19.6
4.5	44.1
10	98



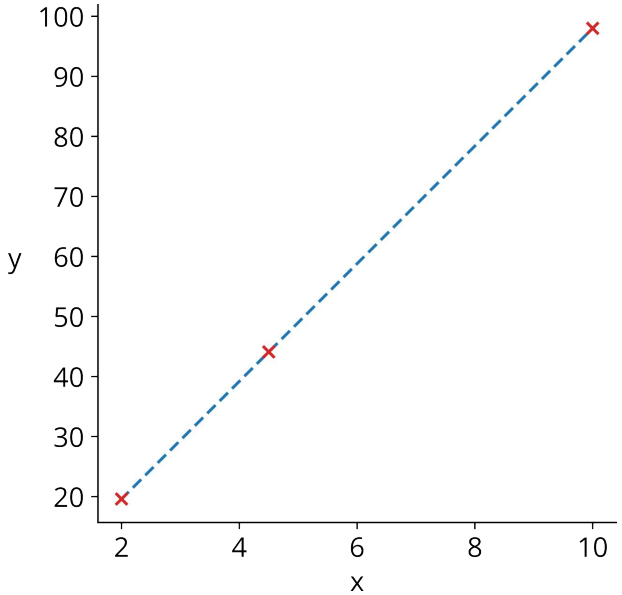
y = ax

Model

Model – a simplification of a real process or system.

Usually we want to know what is the process that generates the data.

x	y
2	19.6
4.5	44.1
10	98



$y = ax$

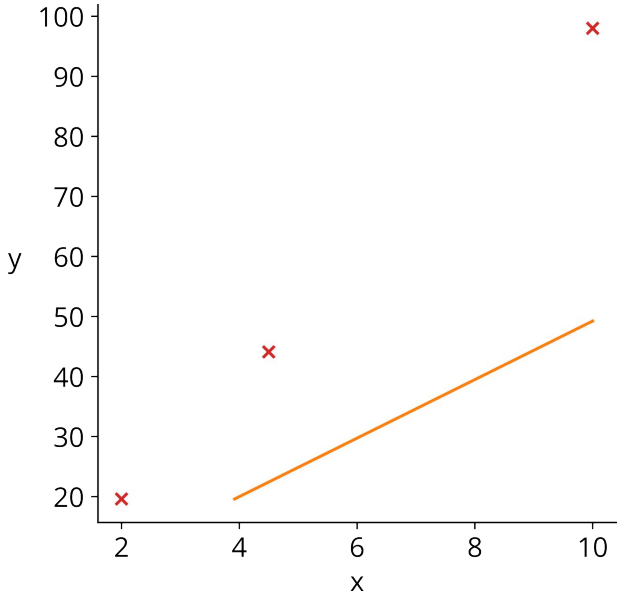
$a = ?$

Model

Model – a simplification of a real process or system.

Usually we want to know what is the process that generates the data.

x	y
2	19.6
4.5	44.1
10	98



$y = ax$

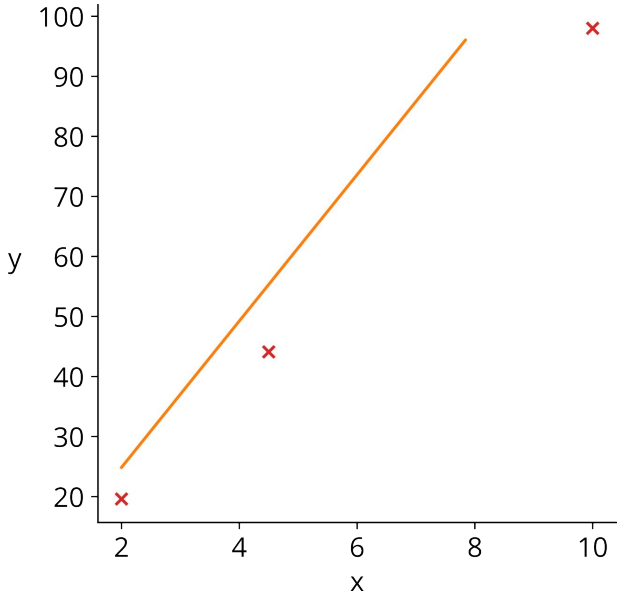
$a = 5?$

Model

Model – a simplification of a real process or system.

Usually we want to know what is the process that generates the data.

x	y
2	19.6
4.5	44.1
10	98



$y = ax$

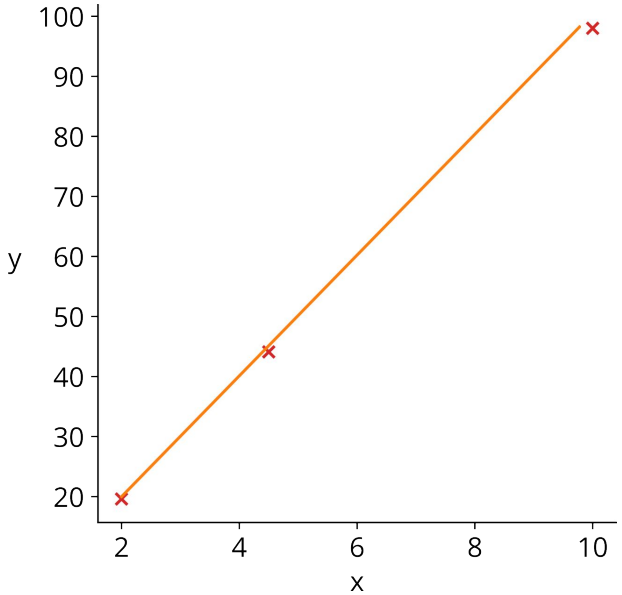
$a = 12?$

Model

Model – a simplification of a real process or system.

Usually we want to know what is the process that generates the data.

x	y
2	19.6
4.5	44.1
10	98



$y = ax$

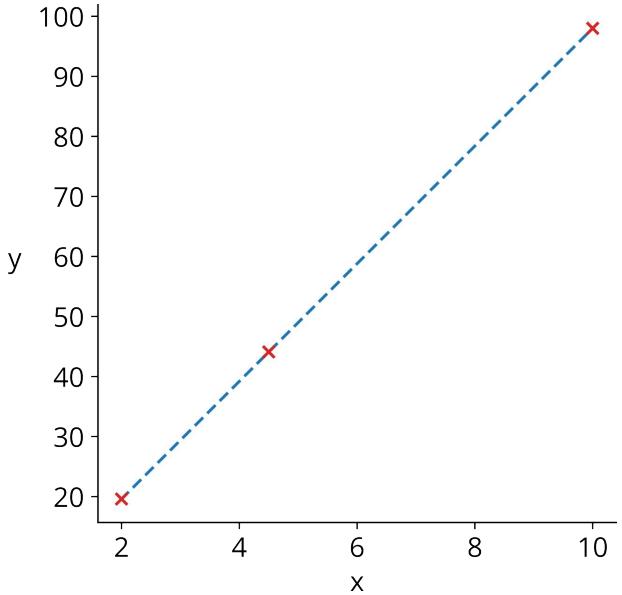
$a = 10?$

Model

Model – a simplification of a real process or system.

Usually we want to know what is the process that generates the data.

x	y
2	19.6
4.5	44.1
10	98



$y = ax$

$a = 9.8$

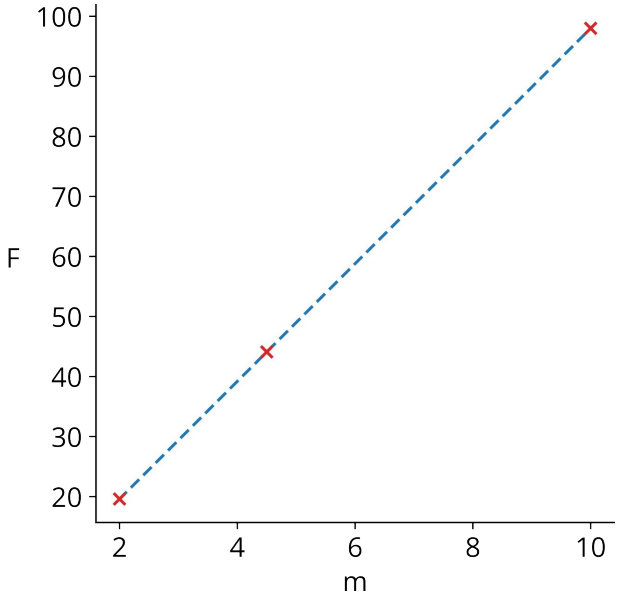
$y = 9.8\,x$

Model

Model – a simplification of a real process or system.

Usually we want to know what is the process that generates the data.

m	F
2	19.6
4.5	44.1
10	98



F = mg

g = 9.8

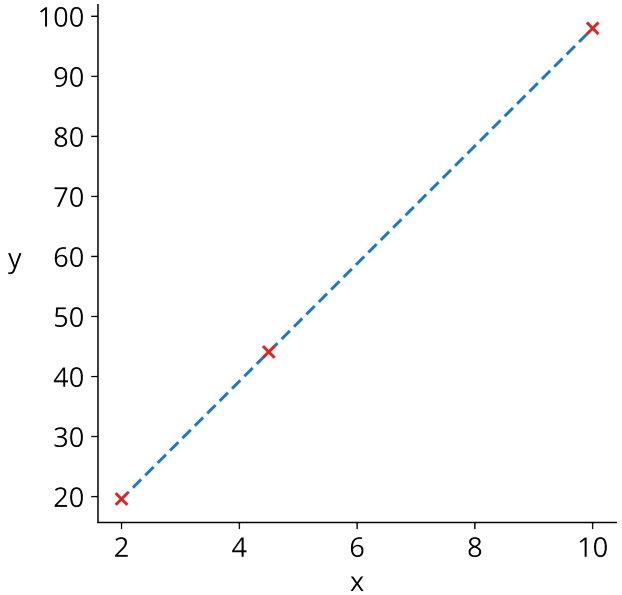
F = 9.8 m

Model

Model – a simplification of a real process or system.

Usually we want to know what is the process that generates the data.

x	y
2	19.6
4.5	44.1
10	98



$y = ax$

$a = 9.8$

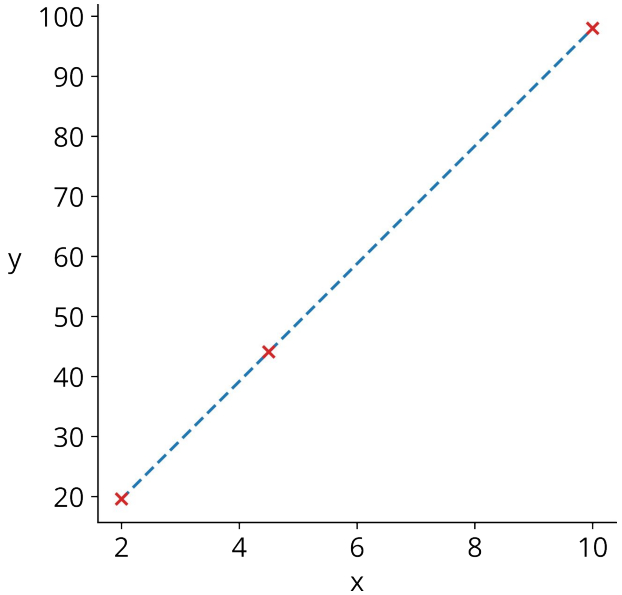
$y = 9.8\,x$

Model

Model – a simplification of a real process or system.

Usually we want to know what is the process that generates the data.

x	y
2	19.6
4.5	44.1
10	98



$\hat{y} = ax$

$a = 9.8$

$\hat{y} = 9.8\,x$

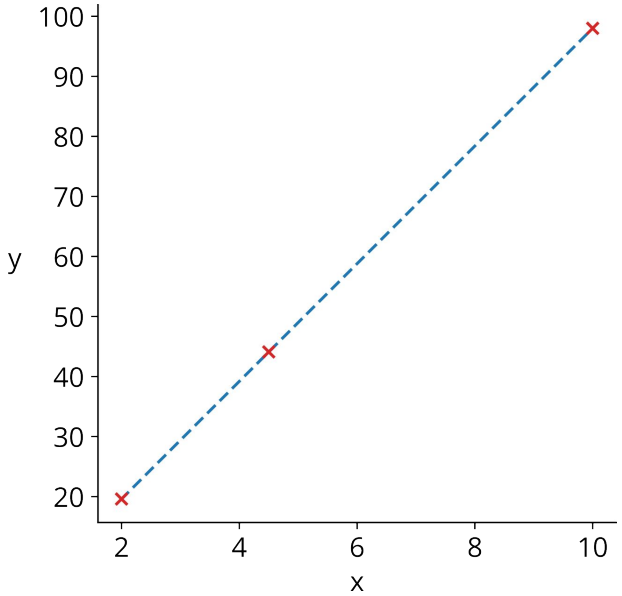
Model

(x, y) – data
$\hat{y}$ – prediction
y_p

Model – a simplification of a real process or system.

Usually we want to know what is the process that generates the data.

x	y
2	19.6
4.5	44.1
10	98



$\hat{y} = ax$

$a = 9.8$

$\hat{y} = 9.8\,x$

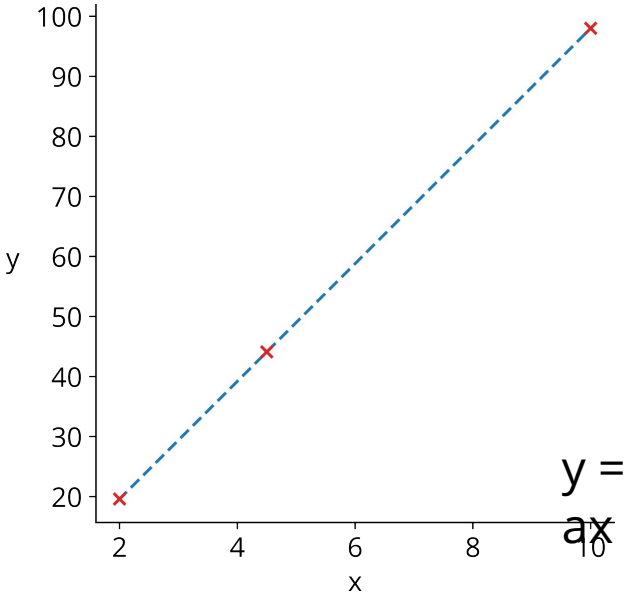
Model

(x, y) – data
$\hat{y}$ – prediction

Model – a simplification of a real process or system.

Usually we want to know what is the process that generates the data.

x	y	z
2	19.6	?
4.5	44.1	?
10	98	?



$\hat{y} = ax$

$a = 9.8$

$\hat{y} = 9.8\,x$

$y = (1+1/(1 - z^2)\, z\, z^T/b^2)(1 - z^2)^{-1/2} ax$

$b \approx 2.99 \times 10^8$

Model

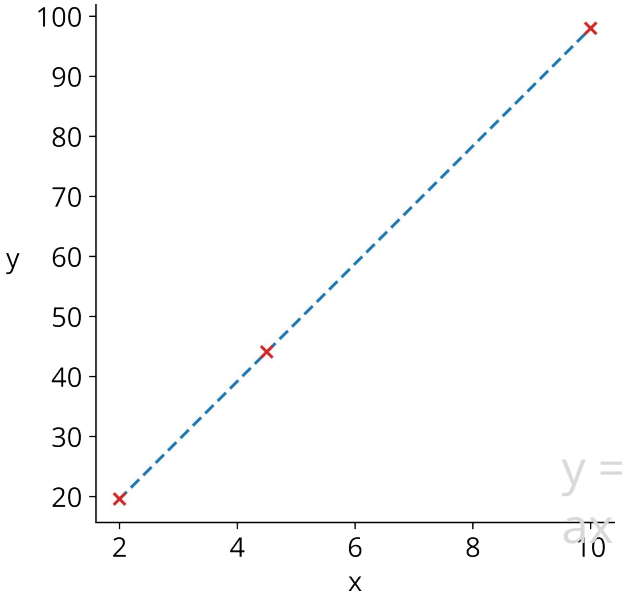
(x, y) – data
$\hat{y}$ – prediction

Model – a simplification of a real process or system.

Usually we want to know what is the process that generates the data.

x	y	z
2	19.6	?
4.5	44.1	?
10	98	?

$\hat{y} \approx y$



$\hat{y} = ax$

$a = 9.8$

$\hat{y} = 9.8\,x$

$y = (1+1/(1 - z^2)\, z\, z^T/b^2)(1 - z^2)^{-1/2}$

ax

$b \approx 2.99 \times 10^8$

Model

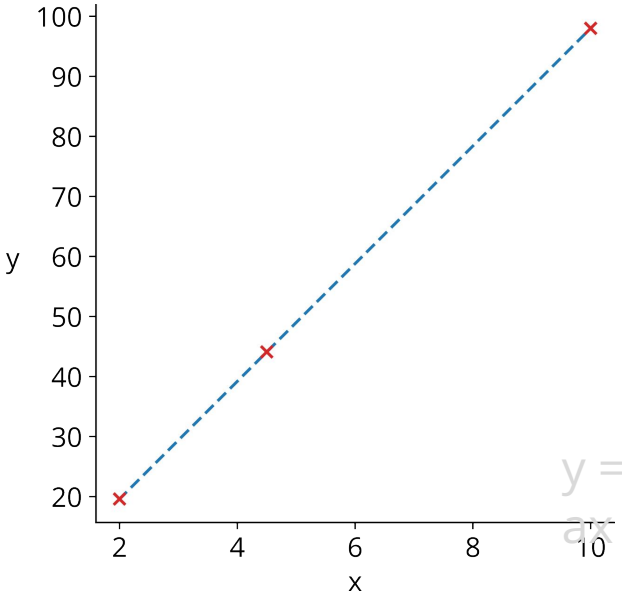
(x, y) – data
$\hat{y}$ – prediction

Model – a simplification of a real process or system.

Usually *it suffices to approximate* what is the process that generates the data.

x	y	z
2	19.6	?
4.5	44.1	?
10	98	?

$\hat{y} \approx y$



$\hat{y} = ax$

$a = 9.8$

$\hat{y} = 9.8\,x$

$y = (1+1/(1 - z^2)\, z\, z^T/b^2)(1 - z^2)^{-1/2}$

ax

$b \approx 2.99 \times 10^8$

Loss Function

What is a good approximation?

Loss Function

What is a good approximation?

Is it possible to evaluate “goodness”?

Loss Function

What is a good approximation?

Is it possible to evaluate “goodness”?

x	y
2	19.6
4.5	44.1
10	98

Loss Function

What is a good approximation?

Is it possible to evaluate “goodness”?

Distance from the data to the model:

x	y
2	19.6
4.5	44.1
10	98

$$y_i - \hat{y}(x_i)$$

Loss Function

What is a good approximation?

Is it possible to evaluate “goodness”?

Distance from the data to the model:

x	y
2	19.6
4.5	44.1
10	98

$$\left| y_i - \hat{y}(x_i) \right|$$

Loss Function

What is a good approximation?

Is it possible to evaluate “goodness”?

Distance from the data to the model:

x	y
2	19.6
4.5	44.1
10	98

$$\left(y_i - \hat{y}(x_i)\right)^2$$

Loss Function

What is a good approximation?

Is it possible to evaluate “goodness”?

Distance from the data to the model:

x	y
2	19.6
4.5	44.1
10	98

$$\sum_i \left(y_i - \hat{y}(x_i) \right)^2$$

Loss Function

What is a good approximation?

Is it possible to evaluate “goodness”?

Distance from the data to the model:

x	y
2	19.6
4.5	44.1
10	98

$$\mathcal{L} = \sum_i \Big(y_i - \hat{y}(x_i) \Big)^2$$

Loss Function

What is a good approximation?

Is it possible to evaluate “goodness”?

Distance from the data to the model:

x	y
2	19.6
4.5	44.1
10	98

$$\mathcal{L} = \sum_i \left(y_i - \hat{y}(x_i) \right)^2$$

If $\mathcal{L} = 0$ then $y_i - \hat{y}(x_i) = 0$ for all i

Loss Function

What is a good approximation?

Is it possible to evaluate “goodness”?

Distance from the data to the model:

x	y
2	19.6
4.5	44.1
10	98

$$\mathcal{L} = \sum_i \left(y_i - \hat{y}(x_i) \right)^2$$

If $\mathcal{L} = 0$ then $y_i = \hat{y}(x_i)$ for all i

Loss Function

What is a good approximation?

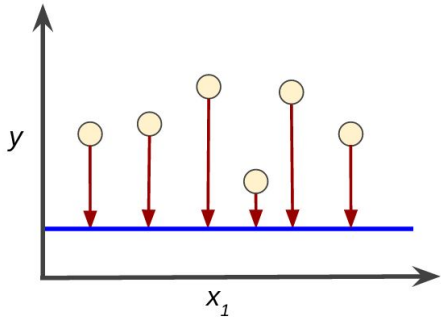
Is it possible to evaluate “goodness”?

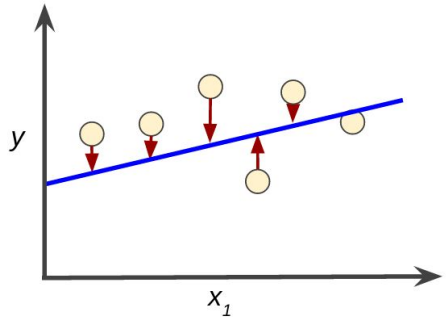
Distance from the data to the model:

x	y
2	19.6
4.5	44.1
10	98

$$\mathcal{L} = \sum_i \left(y_i - \hat{y}(x_i) \right)^2$$

If $\mathcal{L} = 0$ then $y_i = \hat{y}(x_i)$ for all i





Loss Function

What is a good approximation?

Is it possible to evaluate “goodness”?

Distance from the data to the model:

x	y
2	19.6
4.5	44.1
10	98

$$\mathcal{L} = \sum_i \left(y_i - \hat{y}(x_i) \right)^2$$ **L2 Loss**

Loss Function

What is a good approximation?

Is it possible to evaluate “goodness”?

Distance from the data to the model:

x	y
2	19.6
4.5	44.1
10	98

$$\mathcal{L} = \sum_i \left(y_i - \hat{y}(x_i) \right)^2$$ **L2 Loss**

$\hat{y} = a\ x$

Loss Function

What is a good approximation?

Is it possible to evaluate “goodness”?

Distance from the data to the model:

x	y
2	19.6
4.5	44.1
10	98

$$\mathcal{L} = \sum_i \left(y_i - \hat{y}(x_i) \right)^2 \qquad \textbf{L2 Loss}$$

$\hat{y} = a\,x \qquad \mathcal{L} = (\,y_1 - a \cdot x_1)^2 + (\,y_2 - a \cdot x_2\,)^2 + (\,y_2 - a \cdot x_2)^2$

Loss Function

What is a good approximation?

Is it possible to evaluate “goodness”?

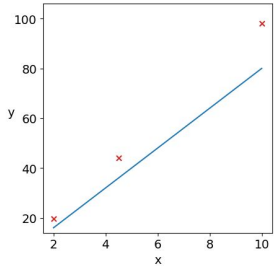
Distance from the data to the model:

x	y
2	19.6
4.5	44.1
10	98

$$\mathcal{L} = \sum_i \left(y_i - \hat{y}(x_i) \right)^2$$ **L2 Loss**

$\hat{y} = a\,x$ $\mathcal{L} = (19.6 - a \cdot 2)^2 + (44.1 - a \cdot 4.5)^2 + (98 - a \cdot 10)^2$

Loss Function



What is a good approximation?

Is it possible to evaluate “goodness”?

Distance from the data to the model:

x	y
2	19.6
4.5	44.1
10	98

$$\mathcal{L} = \sum_i \left(y_i - \hat{y}(x_i) \right)^2 \qquad \textbf{L2 Loss}$$

$\hat{y} = 8\,x$ $\quad \mathcal{L} = (19.6 - 8 \cdot 2)^2 + (44.1 - 8 \cdot 4.5)^2 + (98 - 8 \cdot 10)^2$

Loss Function

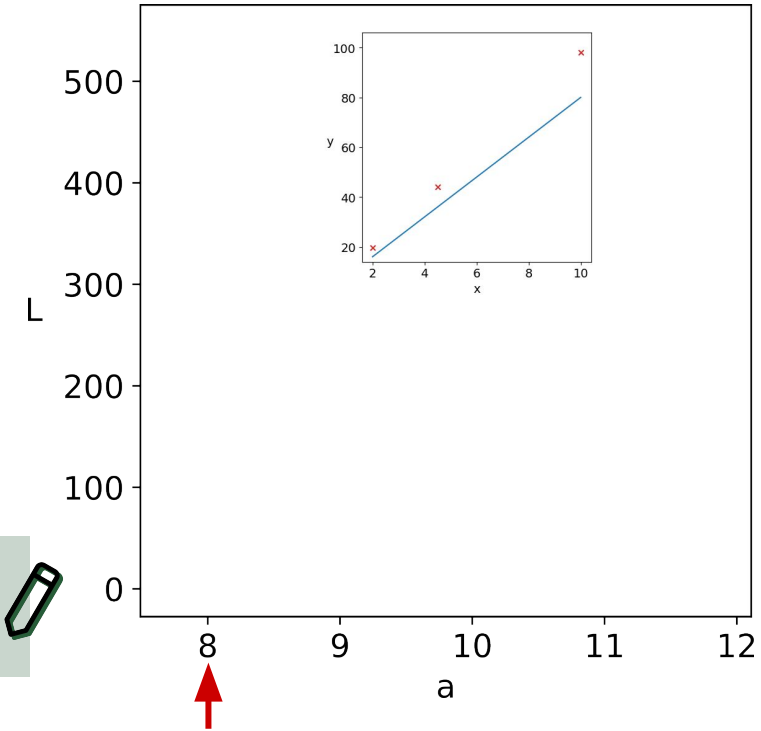
What is a good approximation?

Is it possible to evaluate “goodness”?

Distance from the data to the model:

x	y
2	19.6
4.5	44.1
10	98

$$\mathcal{L} = \sum_i \left(y_i - \hat{y}(x_i) \right)^2$$ **L2 Loss**



$\hat{y} = 8\,x \qquad \mathcal{L} = (19.6 - 8 \cdot 2)^2 + (44.1 - 8 \cdot 4.5)^2 + (98 - 8 \cdot 10)^2$

Loss Function

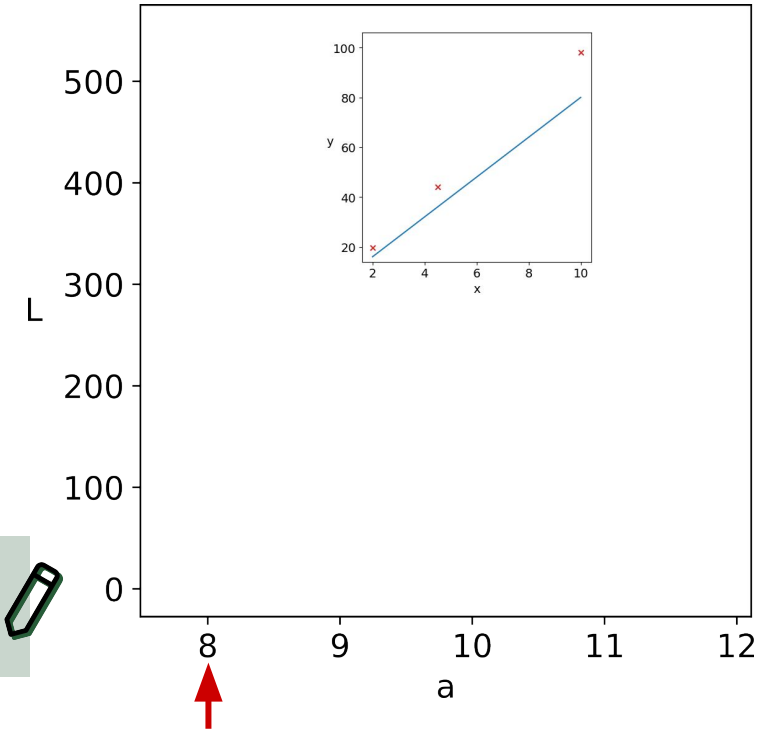
What is a good approximation?

Is it possible to evaluate “goodness”?

Distance from the data to the model:

x	y
2	19.6
4.5	44.1
10	98

$$\mathcal{L} = \sum_i \left(y_i - \hat{y}(x_i) \right)^2$$ **L2 Loss**



$\hat{y} = 8\,x$ $\mathcal{L} = (19.6 - 8 \cdot 2)^2 + (44.1 - 8 \cdot 4.5)^2 + (98 - 8 \cdot 10)^2 = 402.57$

Loss Function

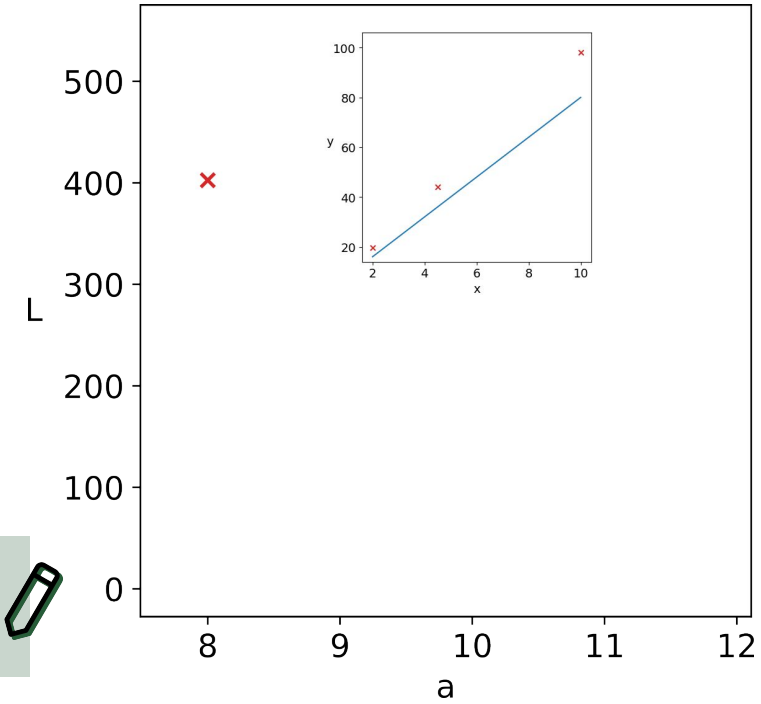
What is a good approximation?

Is it possible to evaluate “goodness”?

Distance from the data to the model:

x	y
2	19.6
4.5	44.1
10	98

$$\mathcal{L} = \sum_i \left(y_i - \hat{y}(x_i) \right)^2$$ **L2 Loss**



$\hat{y} = 8\,x$ $\mathcal{L} = (19.6 - 8 \cdot 2)^2 + (44.1 - 8 \cdot 4.5)^2 + (98 - 8 \cdot 10)^2 = 402.57$

Loss Function

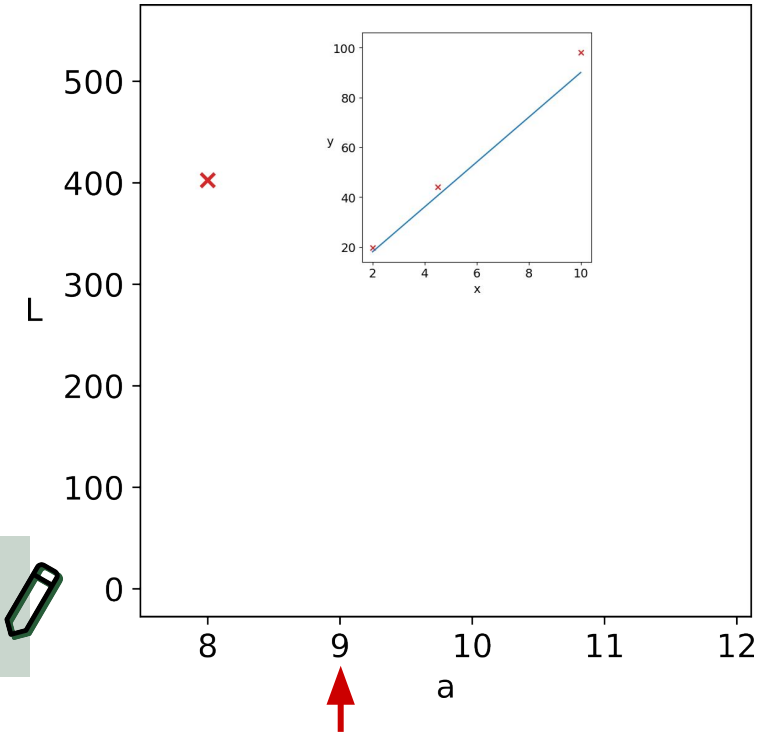
What is a good approximation?

Is it possible to evaluate “goodness”?

Distance from the data to the model:

x	y
2	19.6
4.5	44.1
10	98

$$\mathcal{L} = \sum_i \left(y_i - \hat{y}(x_i) \right)^2$$ **L2 Loss**



$\hat{y} = 8\,x$ $\quad \mathcal{L} = (19.6 - 8 \cdot 2)^2 + (44.1 - 8 \cdot 4.5)^2 + (98 - 8 \cdot 10)^2 = 402.57$

$\hat{y} = 9\,x$

Loss Function

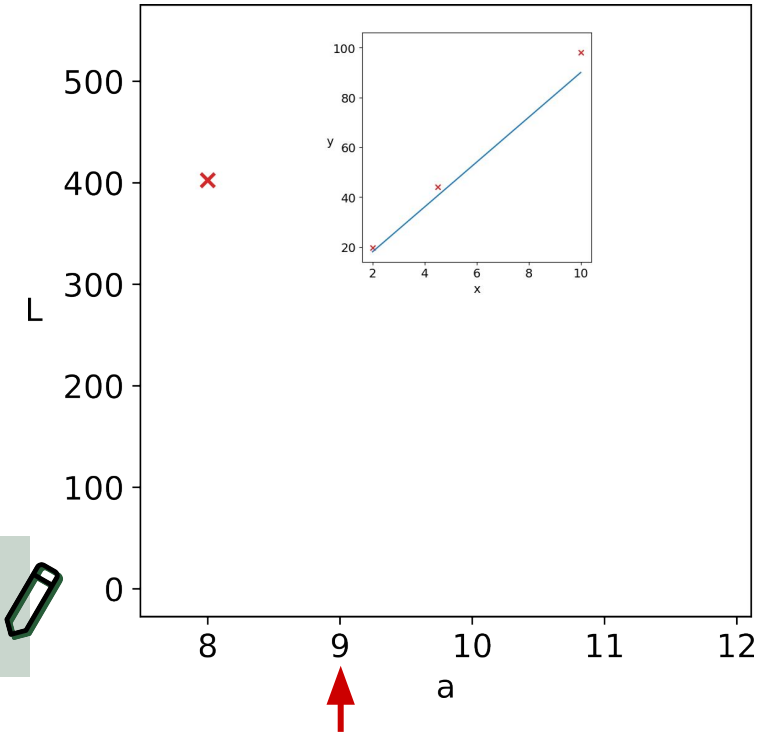
What is a good approximation?

Is it possible to evaluate “goodness”?

Distance from the data to the model:

x	y
2	19.6
4.5	44.1
10	98

$$\mathcal{L} = \sum_i \left(y_i - \hat{y}(x_i) \right)^2$$ **L2 Loss**



$\hat{y} = 8\,x \quad \mathcal{L} = (19.6 - 8 \cdot 2)^2 + (44.1 - 8 \cdot 4.5)^2 + (98 - 8 \cdot 10)^2 = 402.57$

$\hat{y} = 9\,x \quad \mathcal{L} = (19.6 - 9 \cdot 2)^2 + (44.1 - 9 \cdot 4.5)^2 + (98 - 9 \cdot 10)^2$

Loss Function

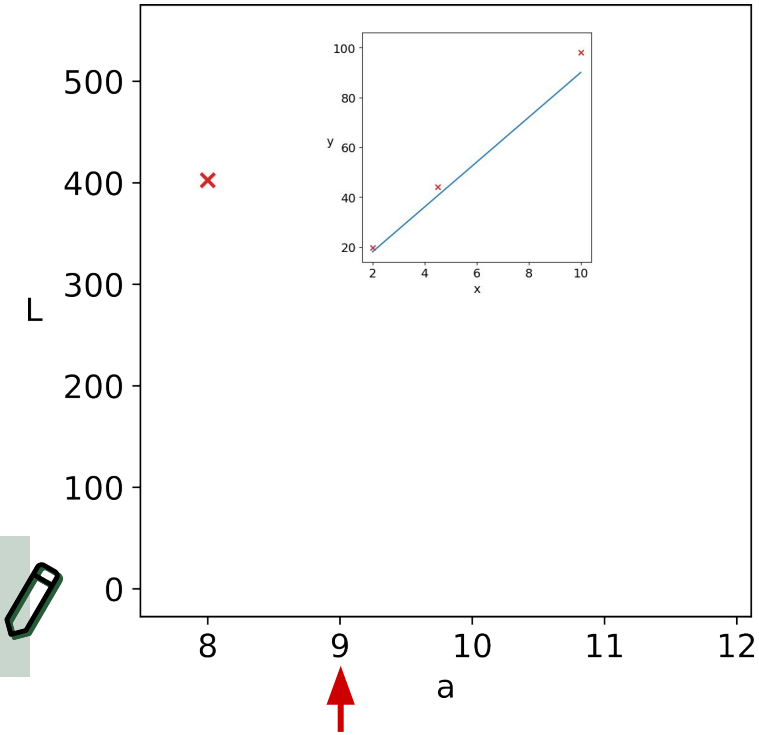
What is a good approximation?

Is it possible to evaluate “goodness”?

Distance from the data to the model:

x	y
2	19.6
4.5	44.1
10	98

$$\mathcal{L} = \sum_i \left(y_i - \hat{y}(x_i) \right)^2$$ **L2 Loss**



$\hat{y} = 8\,x \quad \mathcal{L} = (19.6 - 8 \cdot 2)^2 + (44.1 - 8 \cdot 4.5)^2 + (98 - 8 \cdot 10)^2 = 402.57$

$\hat{y} = 9\,x \quad \mathcal{L} = (19.6 - 9 \cdot 2)^2 + (44.1 - 9 \cdot 4.5)^2 + (98 - 9 \cdot 10)^2 = 79.52$

Loss Function

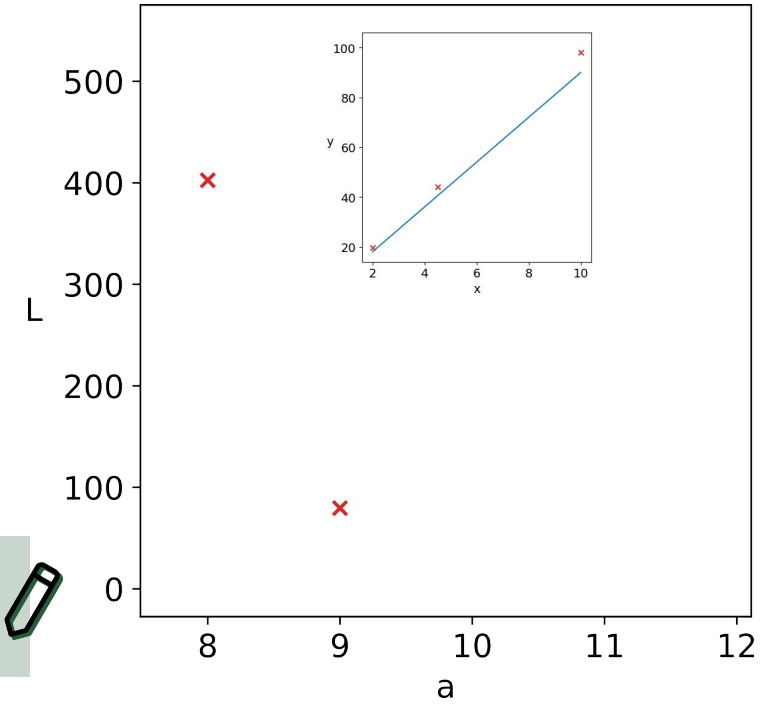
What is a good approximation?

Is it possible to evaluate “goodness”?

Distance from the data to the model:

x	y
2	19.6
4.5	44.1
10	98

$$\mathcal{L} = \sum_i \left(y_i - \hat{y}(x_i) \right)^2$$ **L2 Loss**



$\hat{y} = 8\,x \quad \mathcal{L} = (19.6 - 8 \cdot 2)^2 + (44.1 - 8 \cdot 4.5)^2 + (98 - 8 \cdot 10)^2 = 402.57$

$\hat{y} = 9\,x \quad \mathcal{L} = (19.6 - 9 \cdot 2)^2 + (44.1 - 9 \cdot 4.5)^2 + (98 - 9 \cdot 10)^2 = 79.52$

Loss Function

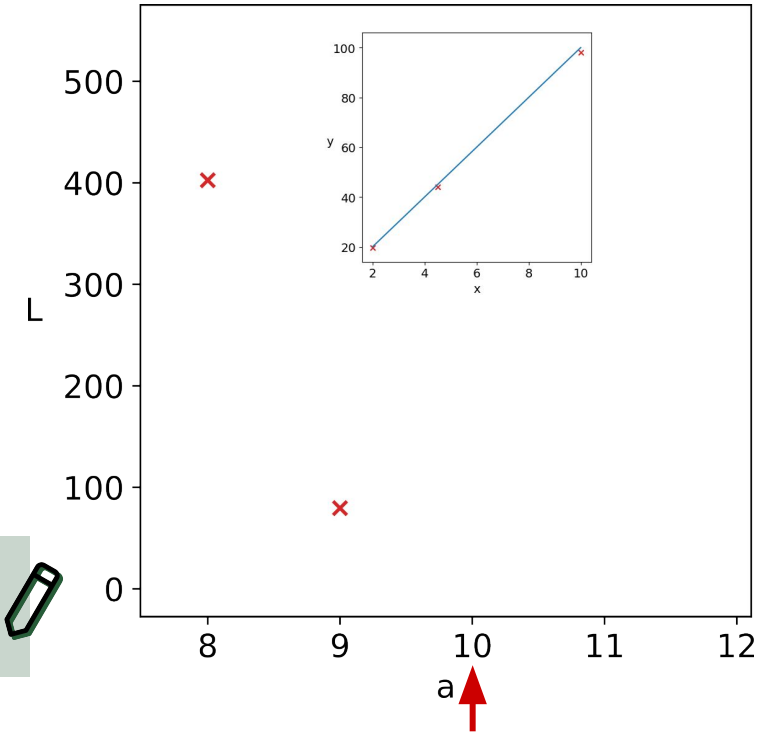
What is a good approximation?

Is it possible to evaluate “goodness”?

Distance from the data to the model:

x	y
2	19.6
4.5	44.1
10	98

$$\mathcal{L} = \sum_i \left(y_i - \hat{y}(x_i) \right)^2$$ **L2 Loss**



$\hat{y} = 8\,x \quad \mathcal{L} = (19.6 - 8 \cdot 2)^2 + (44.1 - 8 \cdot 4.5)^2 + (98 - 8 \cdot 10)^2 = 402.57$

$\hat{y} = 9\,x \quad \mathcal{L} = (19.6 - 9 \cdot 2)^2 + (44.1 - 9 \cdot 4.5)^2 + (98 - 9 \cdot 10)^2 = 79.52$

$\hat{y} = 10\,x$

Loss Function

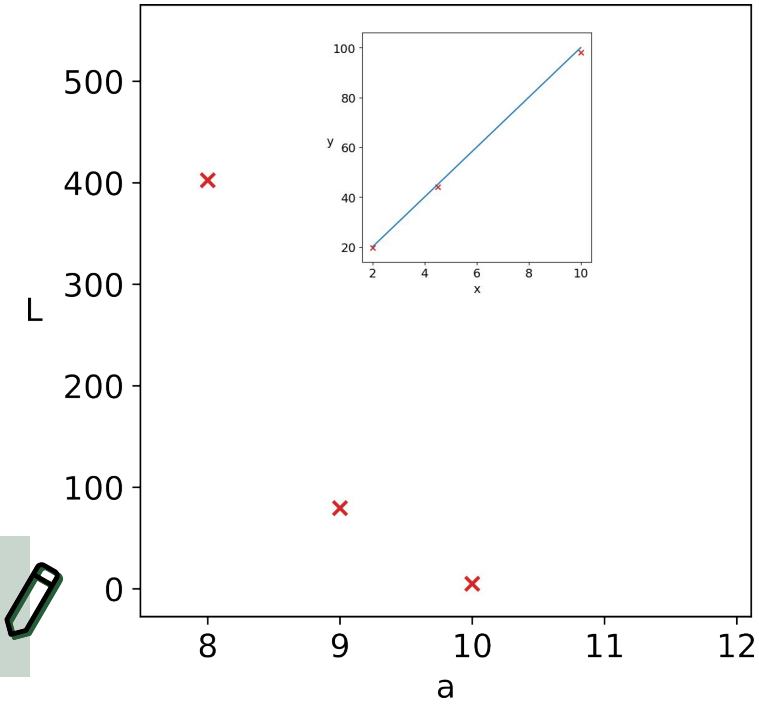
What is a good approximation?

Is it possible to evaluate “goodness”?

Distance from the data to the model:

x	y
2	19.6
4.5	44.1
10	98

$$\mathcal{L} = \sum_i \left(y_i - \hat{y}(x_i) \right)^2$$ **L2 Loss**



$\hat{y} = 8\,x \quad \mathcal{L} = (19.6 - 8 \cdot 2)^2 + (44.1 - 8 \cdot 4.5)^2 + (98 - 8 \cdot 10)^2 = 402.57$

$\hat{y} = 9\,x \quad \mathcal{L} = (19.6 - 9 \cdot 2)^2 + (44.1 - 9 \cdot 4.5)^2 + (98 - 9 \cdot 10)^2 = 79.52$

$\hat{y} = 10\,x \quad \mathcal{L} = (19.6 - 10 \cdot 2)^2 + (44.1 - 10 \cdot 4.5)^2 + (98 - 10 \cdot 10)^2 = 4.97$

Loss Function

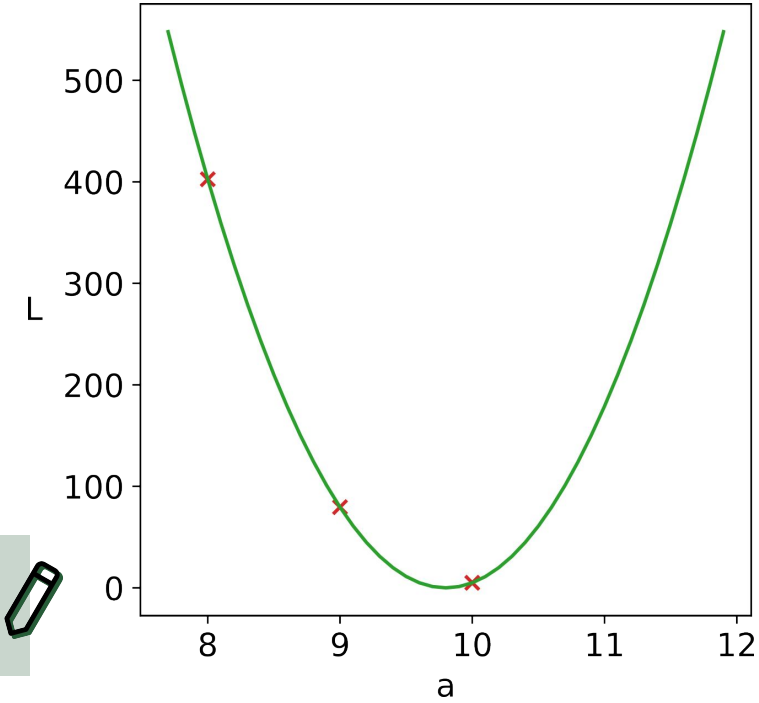
What is a good approximation?

Is it possible to evaluate “goodness”?

Distance from the data to the model:

x	y
2	19.6
4.5	44.1
10	98

$$\mathcal{L} = \sum_i \left(y_i - \hat{y}(x_i) \right)^2$$ **L2 Loss**



$\hat{y} = 8\,x \quad \mathcal{L} = (19.6 - 8 \cdot 2)^2 + (44.1 - 8 \cdot 4.5)^2 + (98 - 8 \cdot 10)^2 = 402.57$

$\hat{y} = 9\,x \quad \mathcal{L} = (19.6 - 9 \cdot 2)^2 + (44.1 - 9 \cdot 4.5)^2 + (98 - 9 \cdot 10)^2 = 79.52$

$\hat{y} = 10\,x \quad \mathcal{L} = (19.6 - 10 \cdot 2)^2 + (44.1 - 10 \cdot 4.5)^2 + (98 - 10 \cdot 10)^2 = 4.97$

Loss Function

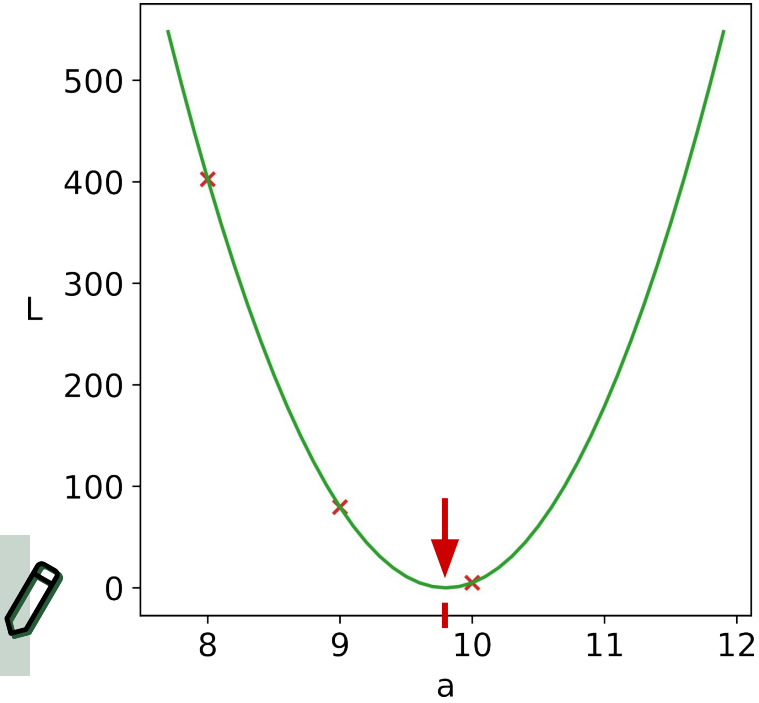
What is a good approximation?

Is it possible to evaluate “goodness”?

Distance from the data to the model:

x	y
2	19.6
4.5	44.1
10	98

$$\mathcal{L} = \sum_i \left(y_i - \hat{y}(x_i) \right)^2$$ **L2 Loss**



$\hat{y} = 8\,x \quad \mathcal{L} = (19.6 - 8 \cdot 2)^2 + (44.1 - 8 \cdot 4.5)^2 + (98 - 8 \cdot 10)^2 = 402.57$

$\hat{y} = 9\,x \quad \mathcal{L} = (19.6 - 9 \cdot 2)^2 + (44.1 - 9 \cdot 4.5)^2 + (98 - 9 \cdot 10)^2 = 79.52$

$\hat{y} = 10\,x \quad \mathcal{L} = (19.6 - 10 \cdot 2)^2 + (44.1 - 10 \cdot 4.5)^2 + (98 - 10 \cdot 10)^2 = 4.97$

Gradient Descent

How do we find the value for a?

$$\hat{y} = ax$$

Gradient Descent

How to find the minimum? $\mathcal{L}$

Gradient Descent

How to find the minimum? $\min \mathcal{L}$

Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$

Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$ $a^* = \arg\min_a \mathcal{L}(a)$

Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$

Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$

Derivatives show “influence” to functions

Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$

Derivatives show “influence” to functions

$\mathcal{L}'$

Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$

Derivatives show “influence” to functions

$\mathcal{L}'$ $\mathcal{L}'_x?$ $\mathcal{L}'_a?$

Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$

Derivatives show “influence” to functions

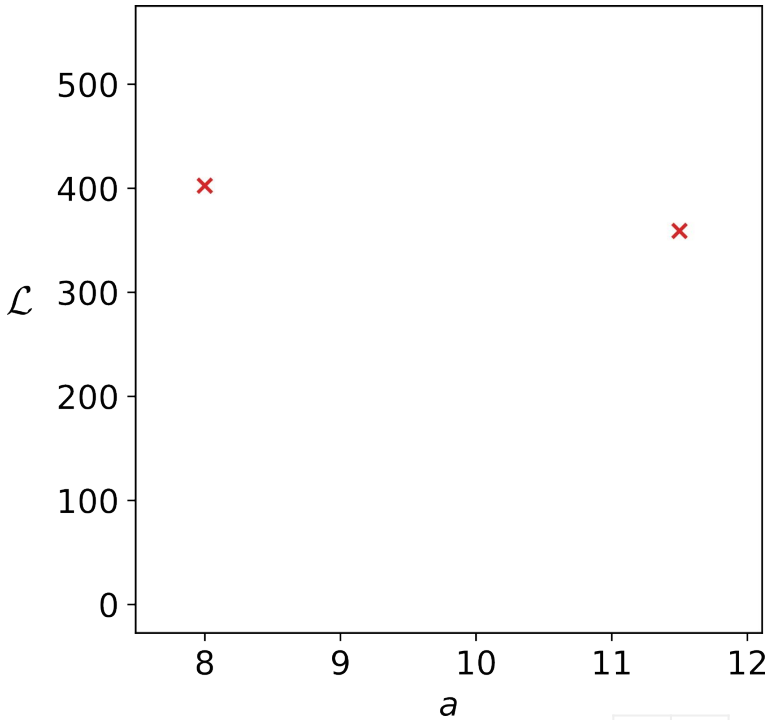
$$\frac{\partial \mathcal{L}}{\partial a}$$

Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$

Derivatives show “influence” to functions

$$\frac{\partial \mathcal{L}}{\partial a}$$



$$\hat{y} = ax$$

$$\mathcal{L} = \sum_i (\hat{y}_i - y_i)^2$$

x	y
2	19.6
4.5	44.1
10	98

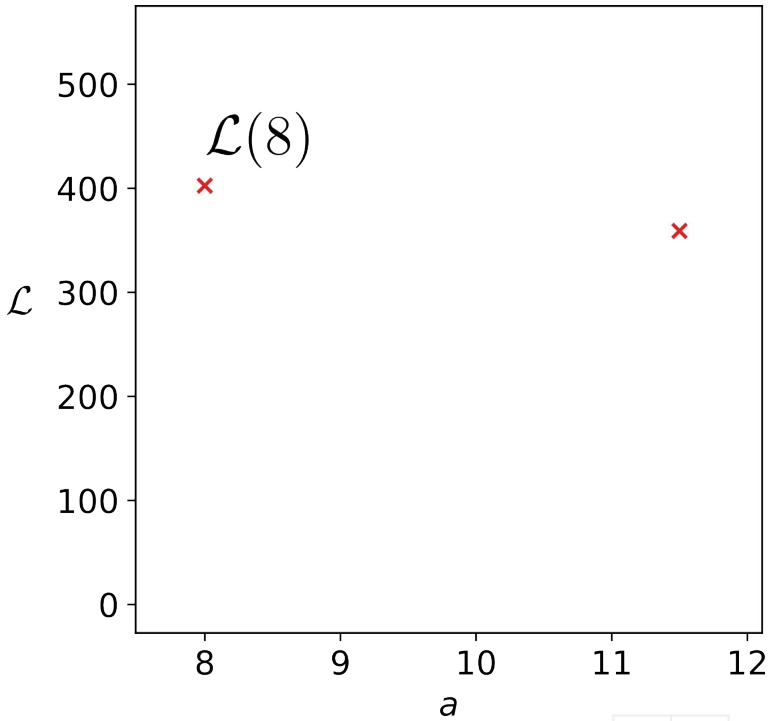
Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$

Derivatives show “influence” to functions

$$\frac{\partial \mathcal{L}}{\partial a} \qquad \mathcal{L} = \mathcal{L}(a)$$

$$\mathcal{L}(8) \approx 402$$



$$\hat{y} = ax$$

$$\mathcal{L} = \sum_i (\hat{y}_i - y_i)^2$$

x	y
2	19.6
4.5	44.1
10	98

Gradient Descent

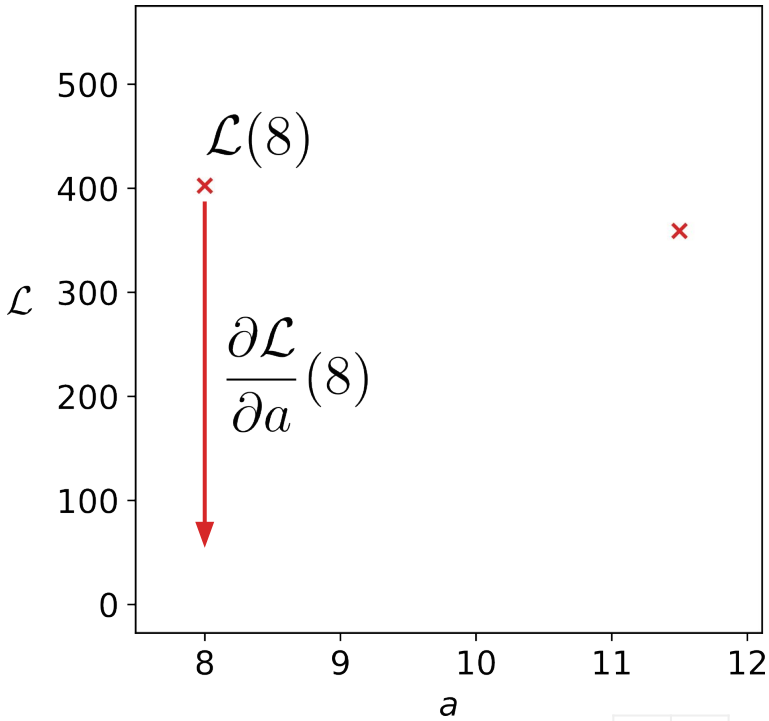
How to find the minimum? $\min_a \mathcal{L}$

Derivatives show “influence” to functions

$$\frac{\partial \mathcal{L}}{\partial a} \qquad \mathcal{L} = \mathcal{L}(a)$$

$$\mathcal{L}(8) \approx 402$$

$$\frac{\partial \mathcal{L}}{\partial a}(8) \approx -447$$



$$\hat{y} = ax$$

$$\mathcal{L} = \sum_i (\hat{y}_i - y_i)^2$$

x	y
2	19.6
4.5	44.1
10	98

Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$

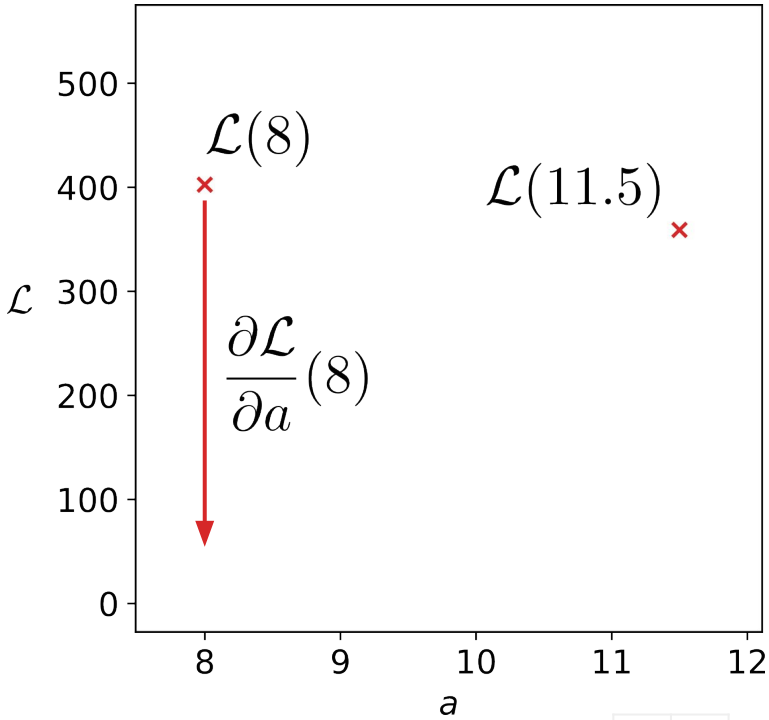
Derivatives show “influence” to functions

$$\frac{\partial \mathcal{L}}{\partial a} \qquad \mathcal{L} = \mathcal{L}(a)$$

$$\mathcal{L}(8) \approx 402$$

$$\frac{\partial \mathcal{L}}{\partial a}(8) \approx -447$$

$$\mathcal{L}(11.5) \approx 359$$



$$\hat{y} = ax$$

$$\mathcal{L} = \sum_i (\hat{y}_i - y_i)^2$$

x	y
2	19.6
4.5	44.1
10	98

Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$

Derivatives show “influence” to functions

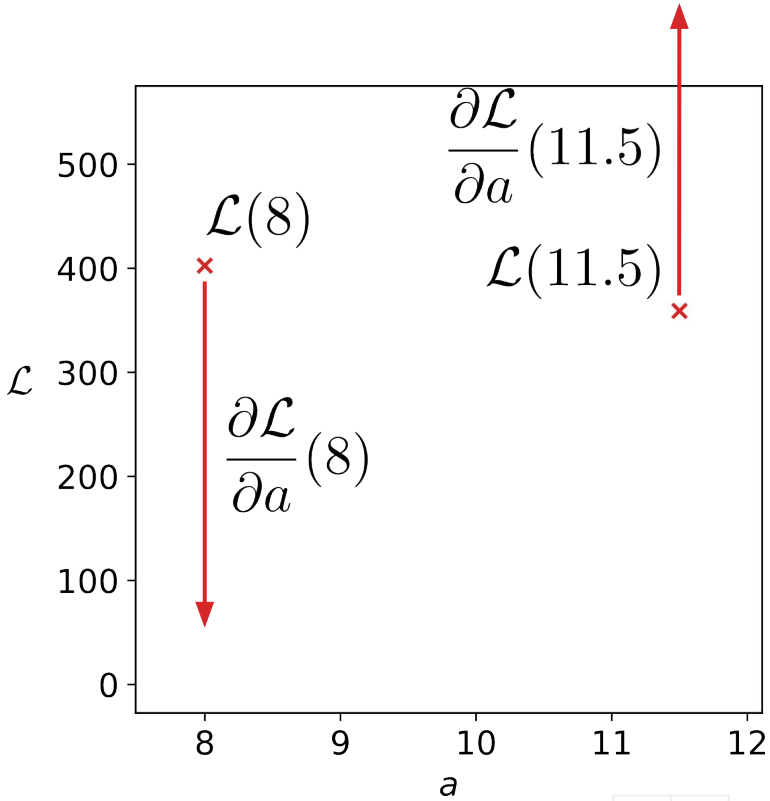
$$\frac{\partial \mathcal{L}}{\partial a} \qquad \mathcal{L} = \mathcal{L}(a)$$

$$\mathcal{L}(8) \approx 402$$

$$\frac{\partial \mathcal{L}}{\partial a}(8) \approx -447$$

$$\mathcal{L}(11.5) \approx 359$$

$$\frac{\partial \mathcal{L}}{\partial a}(8) \approx 422$$



$$\hat{y} = ax$$

$$\mathcal{L} = \sum_i (\hat{y}_i - y_i)^2$$

x	y
2	19.6
4.5	44.1
10	98

Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$

Derivatives show “influence” to functions

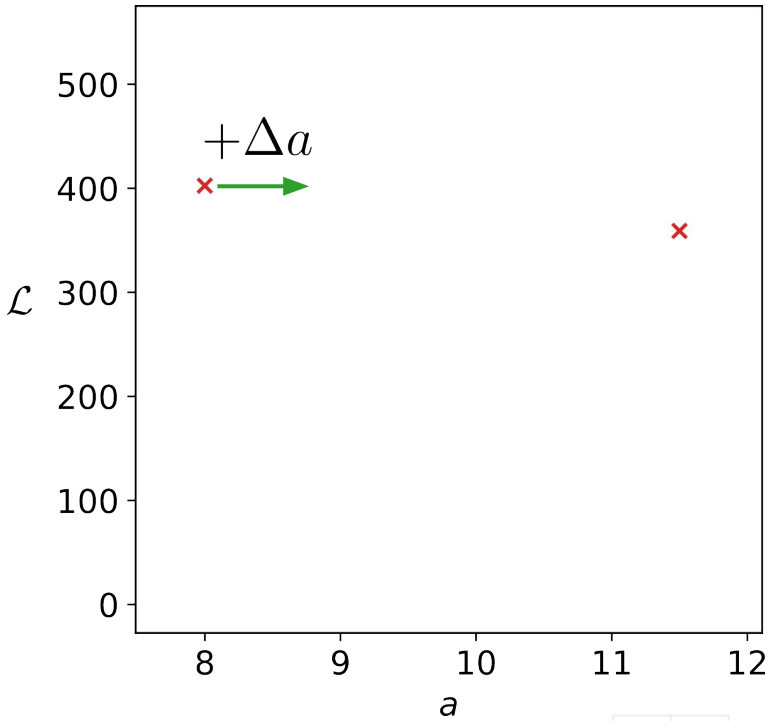
$\frac{\partial \mathcal{L}}{\partial a}$ $\mathcal{L} = \mathcal{L}(a)$ $\Delta a > 0$

$\mathcal{L}(8) \approx 402$

$\frac{\partial \mathcal{L}}{\partial a}(8) \approx -447$

$\mathcal{L}(11.5) \approx 359$

$\frac{\partial \mathcal{L}}{\partial a}(8) \approx 422$



$\hat{y} = ax$

$\mathcal{L} = \sum_i (\hat{y}_i - y_i)^2$

x	y
2	19.6
4.5	44.1
10	98

Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$

Derivatives show “influence” to functions

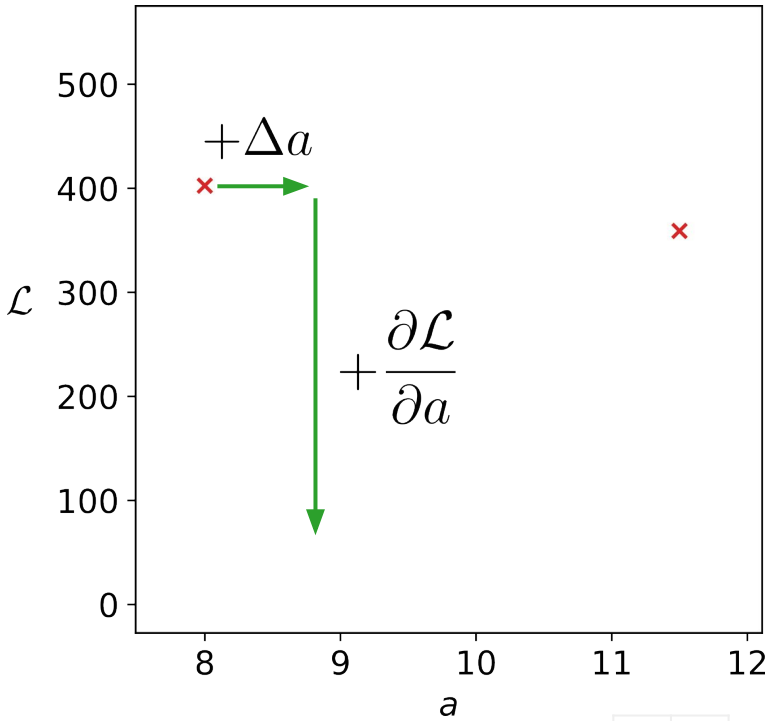
$$\frac{\partial \mathcal{L}}{\partial a} \qquad \mathcal{L} = \mathcal{L}(a) \qquad \Delta a > 0$$

$$\mathcal{L}(8) \approx 402$$

$$\frac{\partial \mathcal{L}}{\partial a}(8) \approx -447$$

$$\mathcal{L}(11.5) \approx 359$$

$$\frac{\partial \mathcal{L}}{\partial a}(8) \approx 422$$



$$\hat{y} = ax$$

$$\mathcal{L} = \sum_i (\hat{y}_i - y_i)^2$$

x	y
2	19.6
4.5	44.1
10	98

Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$

Derivatives show “influence” to functions

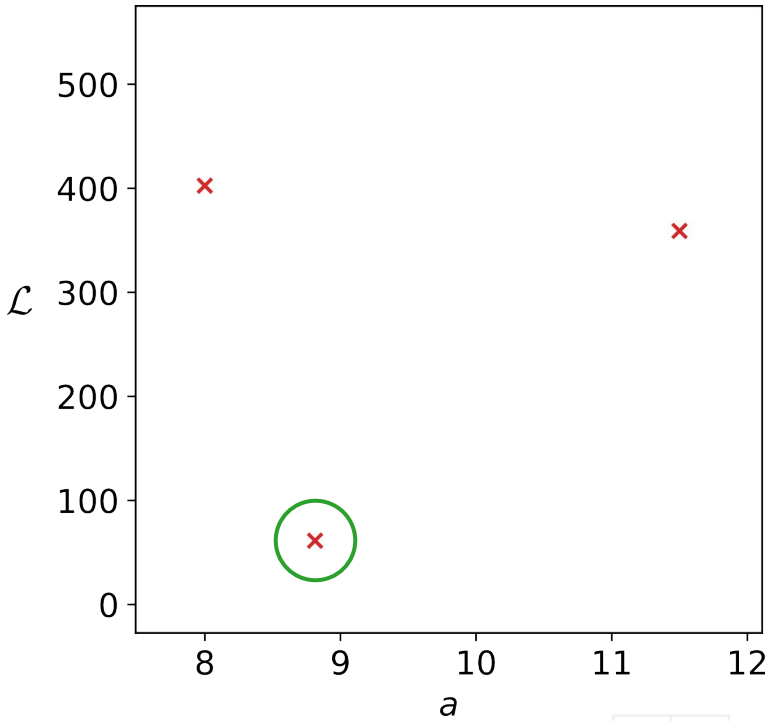
$$\frac{\partial \mathcal{L}}{\partial a} \qquad \mathcal{L} = \mathcal{L}(a) \qquad \Delta a > 0$$

$$\mathcal{L}(8) \approx 402$$

$$\frac{\partial \mathcal{L}}{\partial a}(8) \approx -447$$

$$\mathcal{L}(11.5) \approx 359$$

$$\frac{\partial \mathcal{L}}{\partial a}(8) \approx 422$$



$$\hat{y} = ax$$

$$\mathcal{L} = \sum_i (\hat{y}_i - y_i)^2$$

x	y
2	19.6
4.5	44.1
10	98

Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$

Derivatives show “influence” to functions

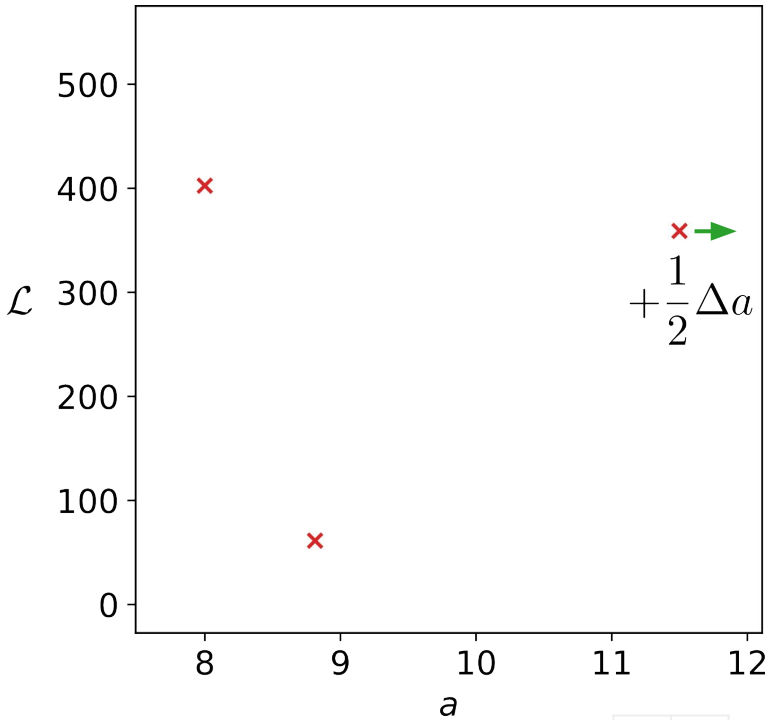
$$\frac{\partial \mathcal{L}}{\partial a} \qquad \mathcal{L} = \mathcal{L}(a) \qquad \Delta a > 0$$

$$\mathcal{L}(8) \approx 402$$

$$\frac{\partial \mathcal{L}}{\partial a}(8) \approx -447$$

$$\mathcal{L}(11.5) \approx 359$$

$$\frac{\partial \mathcal{L}}{\partial a}(8) \approx 422$$



$$\hat{y} = ax$$

$$\mathcal{L} = \sum_i (\hat{y}_i - y_i)^2$$

x	y
2	19.6
4.5	44.1
10	98

Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$

Derivatives show “influence” to functions

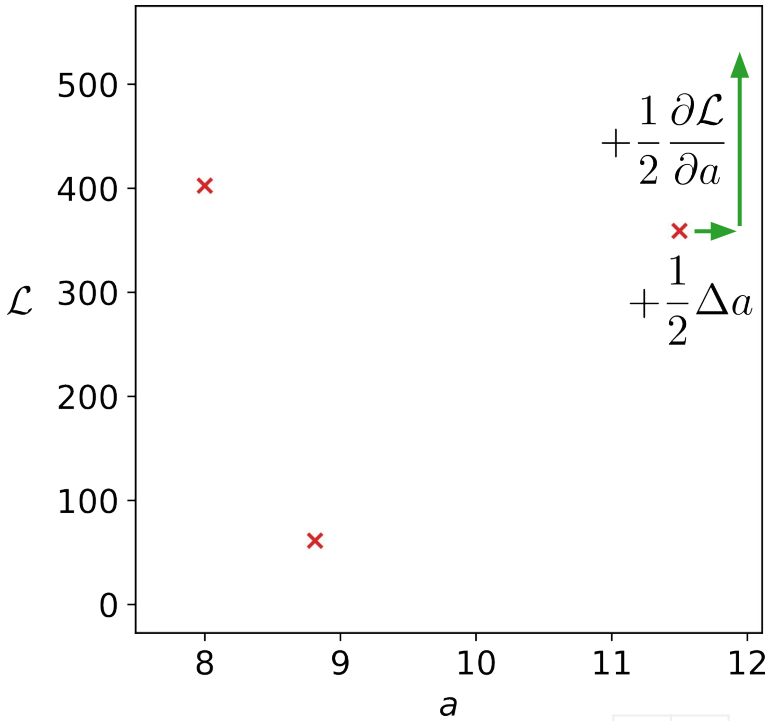
$$\frac{\partial \mathcal{L}}{\partial a} \qquad \mathcal{L} = \mathcal{L}(a) \qquad \Delta a > 0$$

$$\mathcal{L}(8) \approx 402$$

$$\frac{\partial \mathcal{L}}{\partial a}(8) \approx -447$$

$$\mathcal{L}(11.5) \approx 359$$

$$\frac{\partial \mathcal{L}}{\partial a}(8) \approx 422$$



$$\hat{y} = ax$$

$$\mathcal{L} = \sum_i (\hat{y}_i - y_i)^2$$

x	y
2	19.6
4.5	44.1
10	98

Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$

Derivatives show “influence” to functions

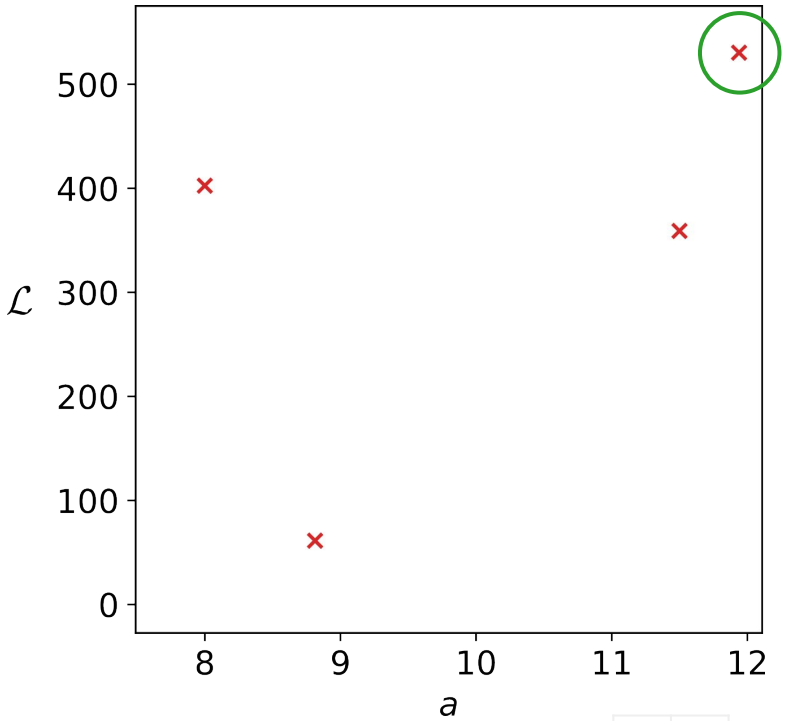
$$\frac{\partial \mathcal{L}}{\partial a} \qquad \mathcal{L} = \mathcal{L}(a) \qquad \Delta a > 0$$

$$\mathcal{L}(8) \approx 402$$

$$\frac{\partial \mathcal{L}}{\partial a}(8) \approx -447$$

$$\mathcal{L}(11.5) \approx 359$$

$$\frac{\partial \mathcal{L}}{\partial a}(8) \approx 422$$



$$\hat{y} = ax$$

$$\mathcal{L} = \sum_i (\hat{y}_i - y_i)^2$$

x	y
2	19.6
4.5	44.1
10	98

Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$

Derivatives show “influence” to functions

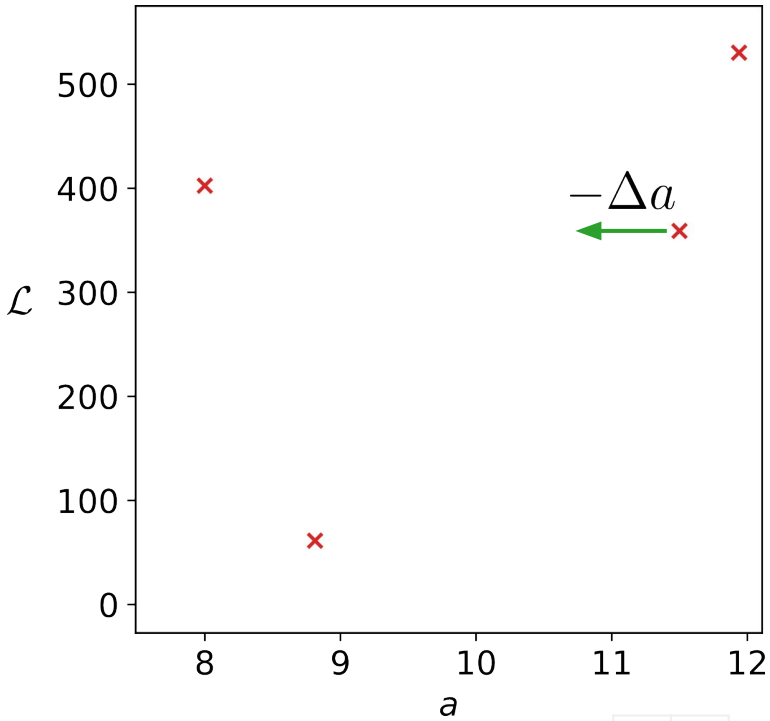
$$\frac{\partial \mathcal{L}}{\partial a} \qquad \mathcal{L} = \mathcal{L}(a) \qquad \Delta a > 0$$

$$\mathcal{L}(8) \approx 402$$

$$\frac{\partial \mathcal{L}}{\partial a}(8) \approx -447$$

$$\mathcal{L}(11.5) \approx 359$$

$$\frac{\partial \mathcal{L}}{\partial a}(8) \approx 422$$



$$\hat{y} = ax$$

$$\mathcal{L} = \sum_i (\hat{y}_i - y_i)^2$$

x	y
2	19.6
4.5	44.1
10	98

Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$

Derivatives show “influence” to functions

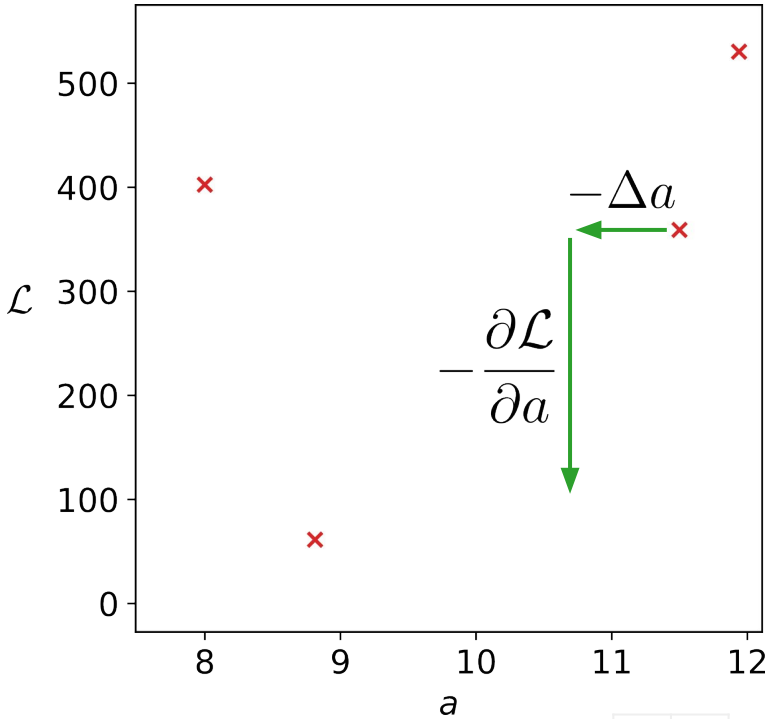
$\frac{\partial \mathcal{L}}{\partial a}$ $\mathcal{L} = \mathcal{L}(a)$ $\Delta a > 0$

$\mathcal{L}(8) \approx 402$

$\frac{\partial \mathcal{L}}{\partial a}(8) \approx -447$

$\mathcal{L}(11.5) \approx 359$

$\frac{\partial \mathcal{L}}{\partial a}(8) \approx 422$



$\hat{y} = ax$

$\mathcal{L} = \sum_i (\hat{y}_i - y_i)^2$

x	y
2	19.6
4.5	44.1
10	98

Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$

Derivatives show “influence” to functions

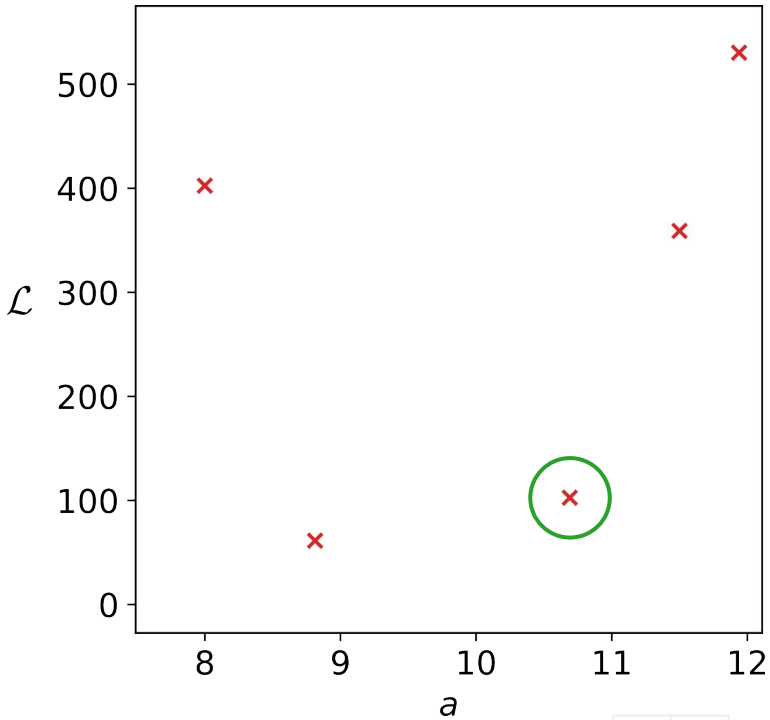
$$\frac{\partial \mathcal{L}}{\partial a} \qquad \mathcal{L} = \mathcal{L}(a) \qquad \Delta a > 0$$

$$\mathcal{L}(8) \approx 402$$

$$\frac{\partial \mathcal{L}}{\partial a}(8) \approx -447$$

$$\mathcal{L}(11.5) \approx 359$$

$$\frac{\partial \mathcal{L}}{\partial a}(8) \approx 422$$



$$\hat{y} = ax$$

$$\mathcal{L} = \sum_i (\hat{y}_i - y_i)^2$$

x	y
2	19.6
4.5	44.1
10	98

Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$

Derivatives show “influence” to functions

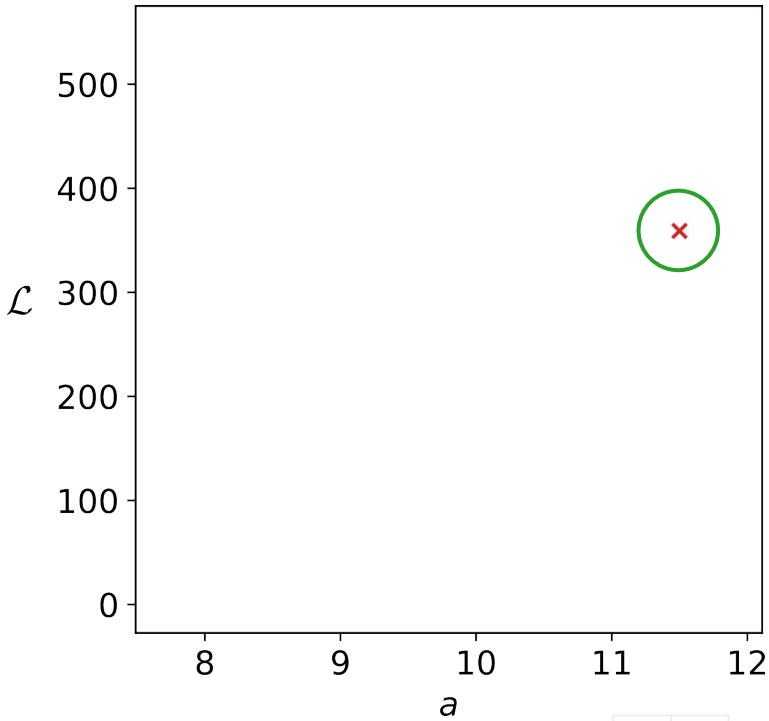
$$\frac{\partial \mathcal{L}}{\partial a} \qquad \mathcal{L} = \mathcal{L}(a)$$

$$\mathcal{L}(8) \approx 402$$

$$\frac{\partial \mathcal{L}}{\partial a}(8) \approx -447$$

$$\mathcal{L}(11.5) \approx 359$$

$$\frac{\partial \mathcal{L}}{\partial a}(8) \approx 422$$



$$\hat{y} = ax$$

$$\mathcal{L} = \sum_i (\hat{y}_i - y_i)^2$$

x	y
2	19.6
4.5	44.1
10	98

Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$

Derivatives show “influence” to functions

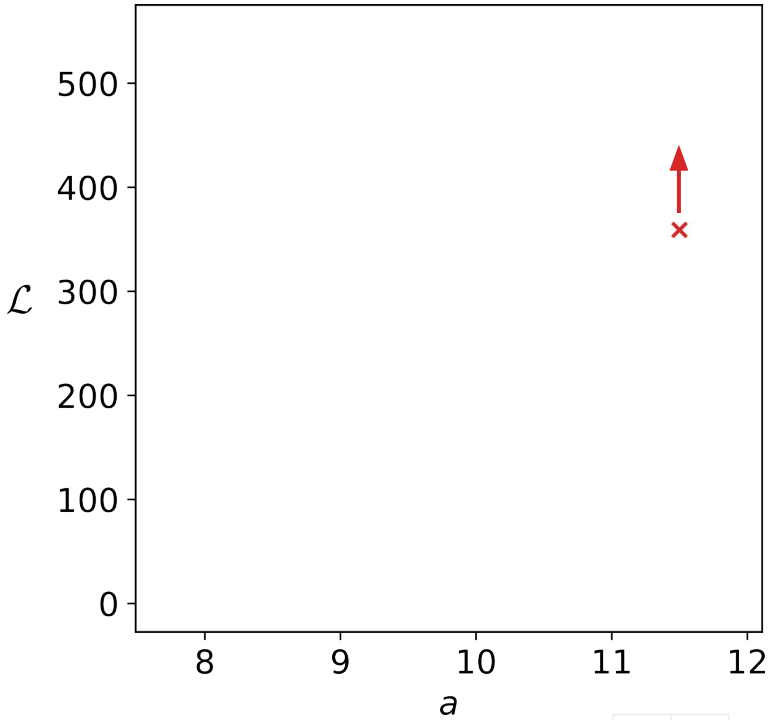
$\frac{\partial \mathcal{L}}{\partial a}$ $\mathcal{L} = \mathcal{L}(a)$

$\mathcal{L}(8) \approx 402$

$\frac{\partial \mathcal{L}}{\partial a}(8) \approx -447$

$\mathcal{L}(11.5) \approx 359$

$\frac{\partial \mathcal{L}}{\partial a}(8) \approx 422$



$\hat{y} = ax$

$\mathcal{L} = \sum_i (\hat{y}_i - y_i)^2$

x	y
2	19.6
4.5	44.1
10	98

Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$

Derivatives show “influence” to functions

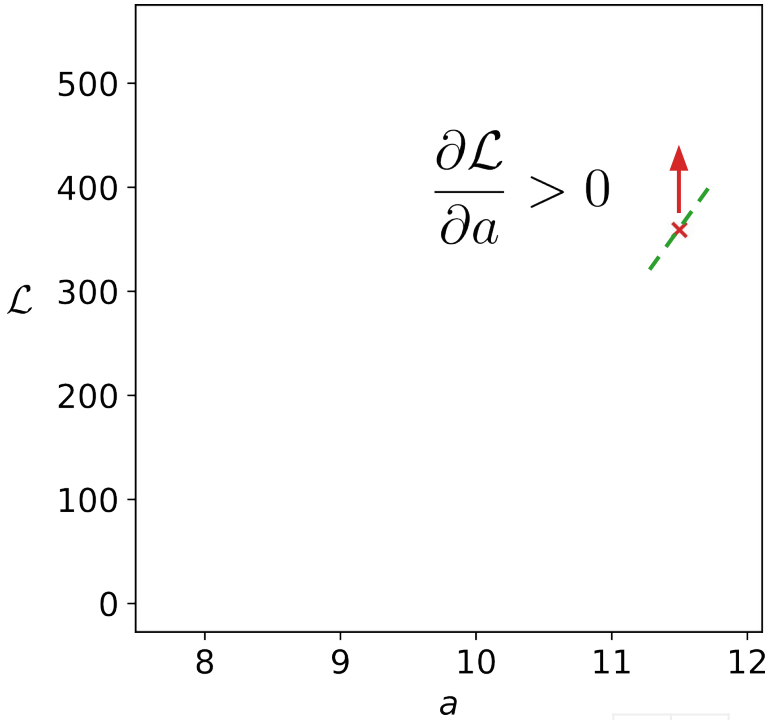
$$\frac{\partial \mathcal{L}}{\partial a} \qquad \mathcal{L} = \mathcal{L}(a)$$

$$\mathcal{L}(8) \approx 402$$

$$\frac{\partial \mathcal{L}}{\partial a}(8) \approx -447$$

$$\mathcal{L}(11.5) \approx 359$$

$$\frac{\partial \mathcal{L}}{\partial a}(8) \approx 422$$



$$\hat{y} = ax$$

$$\mathcal{L} = \sum_i (\hat{y}_i - y_i)^2$$

x	y
2	19.6
4.5	44.1
10	98

Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$

Derivatives show “influence” to functions

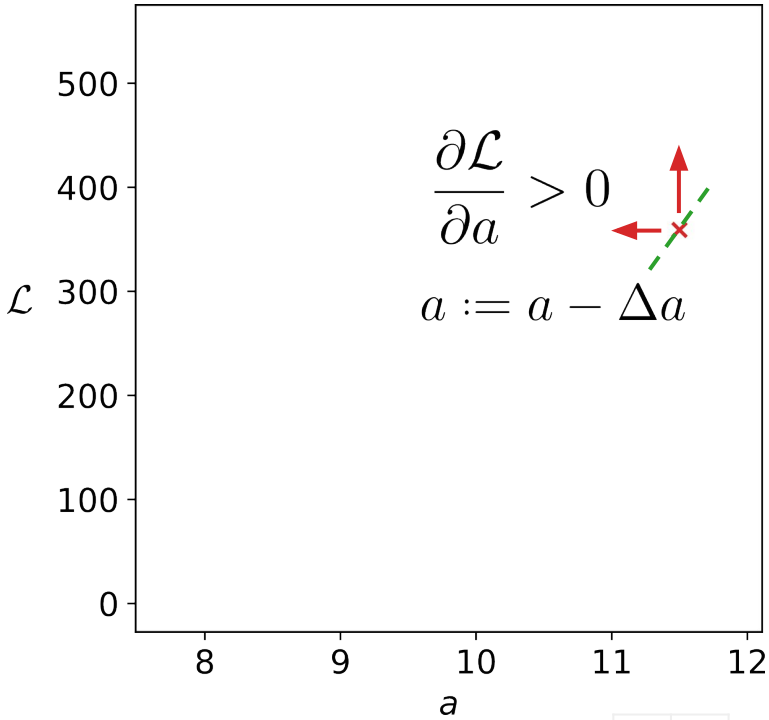
$$\frac{\partial \mathcal{L}}{\partial a} \qquad \mathcal{L} = \mathcal{L}(a)$$

$$\mathcal{L}(8) \approx 402$$

$$\frac{\partial \mathcal{L}}{\partial a}(8) \approx -447$$

$$\mathcal{L}(11.5) \approx 359$$

$$\frac{\partial \mathcal{L}}{\partial a}(8) \approx 422$$



$$\hat{y} = ax$$

$$\mathcal{L} = \sum_i (\hat{y}_i - y_i)^2$$

x	y
2	19.6
4.5	44.1
10	98

Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$

Derivatives show “influence” to functions

$$\frac{\partial \mathcal{L}}{\partial a} \qquad \mathcal{L} = \mathcal{L}(a)$$

$$\mathcal{L}(8) \approx 402$$

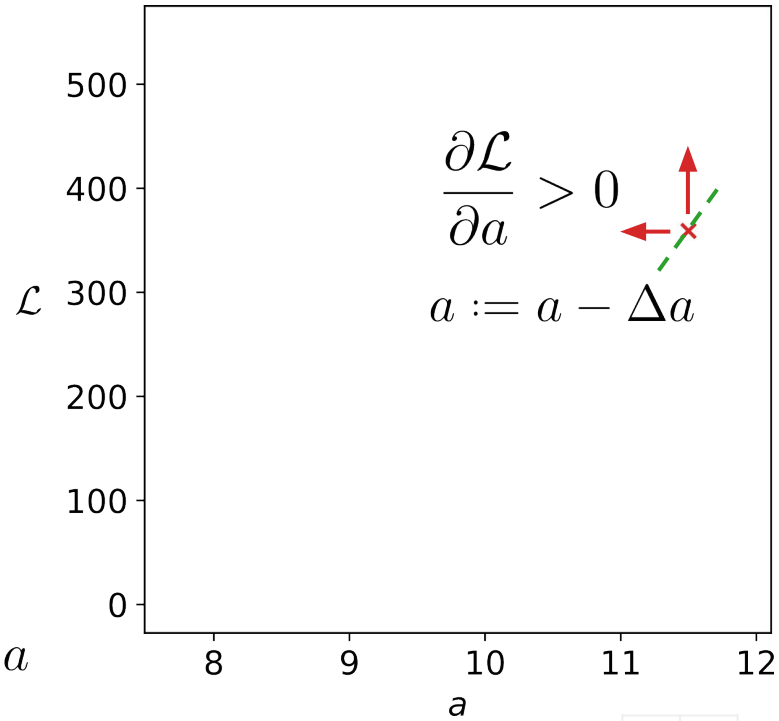
$$\frac{\partial \mathcal{L}}{\partial a}(8) \approx -447$$

$$\mathcal{L}(11.5) \approx 359$$

$$\frac{\partial \mathcal{L}}{\partial a}(8) \approx 422$$

$$\frac{\partial \mathcal{L}}{\partial a} > 0$$

$$a := a - \Delta a$$



$$\hat{y} = ax$$

$$\mathcal{L} = \sum_i (\hat{y}_i - y_i)^2$$

x	y
2	19.6
4.5	44.1
10	98

Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$

Derivatives show “influence” to functions

$\frac{\partial \mathcal{L}}{\partial a}$ $\mathcal{L} = \mathcal{L}(a)$

$\mathcal{L}(8) \approx 402$

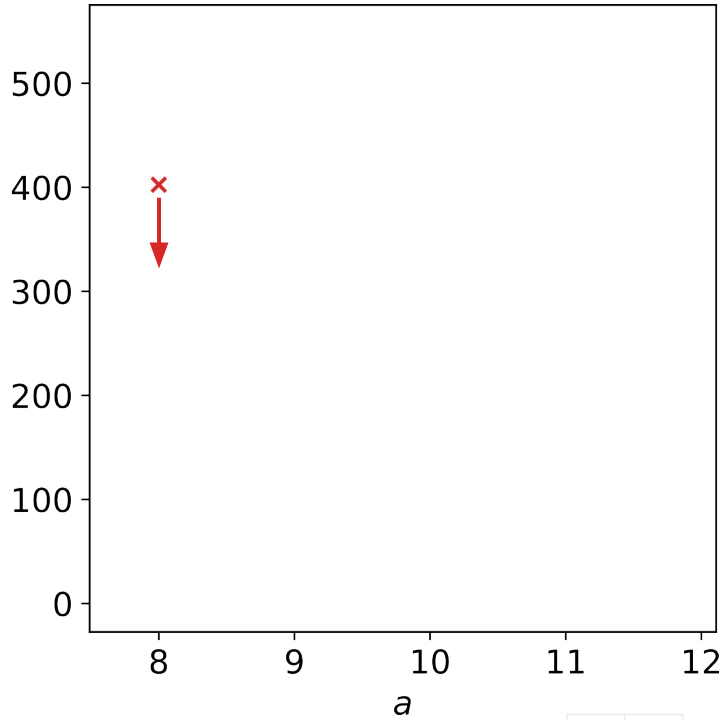
$\frac{\partial \mathcal{L}}{\partial a}(8) \approx -447$

$\mathcal{L}(11.5) \approx 359$

$\frac{\partial \mathcal{L}}{\partial a}(8) \approx 422$

$$\frac{\partial \mathcal{L}}{\partial a} > 0$$

$$a := a - \Delta a$$



$\hat{y} = ax$

$\mathcal{L} = \sum_i (\hat{y}_i - y_i)^2$

x	y
2	19.6
4.5	44.1
10	98

Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$

Derivatives show “influence” to functions

$$\frac{\partial \mathcal{L}}{\partial a} \qquad \mathcal{L} = \mathcal{L}(a)$$

$$\mathcal{L}(8) \approx 402$$

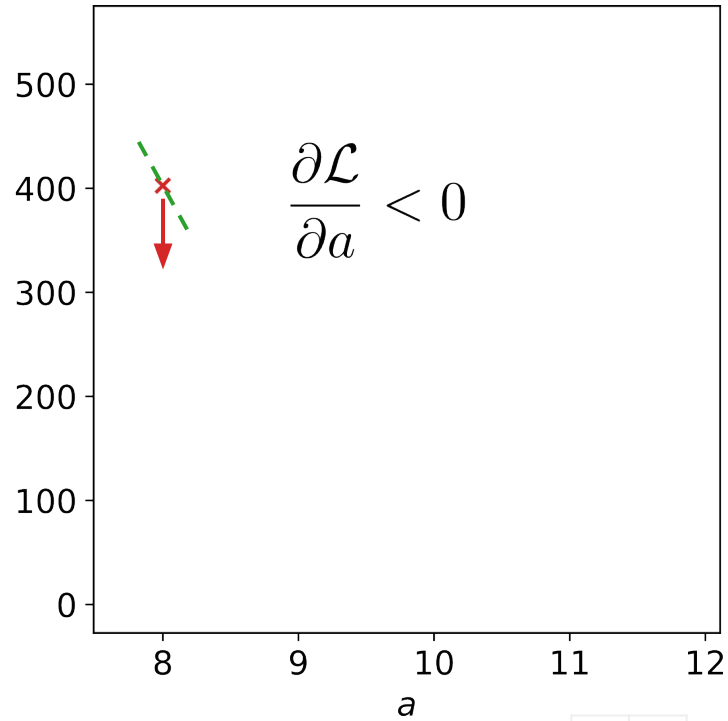
$$\frac{\partial \mathcal{L}}{\partial a}(8) \approx -447$$

$$\mathcal{L}(11.5) \approx 359$$

$$\frac{\partial \mathcal{L}}{\partial a}(8) \approx 422$$

$$\frac{\partial \mathcal{L}}{\partial a} > 0$$

$$a := a - \Delta a$$



$$\hat{y} = ax$$

$$\mathcal{L} = \sum_i (\hat{y}_i - y_i)^2$$

x	y
2	19.6
4.5	44.1
10	98

Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$

Derivatives show “influence” to functions

$$\frac{\partial \mathcal{L}}{\partial a} \qquad \mathcal{L} = \mathcal{L}(a)$$

$$\mathcal{L}(8) \approx 402$$

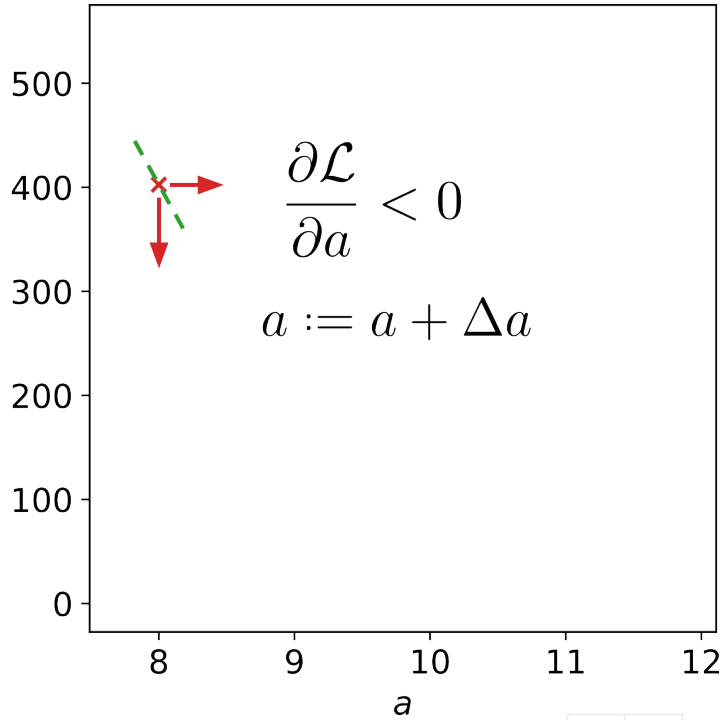
$$\frac{\partial \mathcal{L}}{\partial a}(8) \approx -447$$

$$\mathcal{L}(11.5) \approx 359$$

$$\frac{\partial \mathcal{L}}{\partial a}(8) \approx 422$$

$$\frac{\partial \mathcal{L}}{\partial a} > 0$$

$$a := a - \Delta a$$



$$\hat{y} = ax$$

$$\mathcal{L} = \sum_i (\hat{y}_i - y_i)^2$$

x	y
2	19.6
4.5	44.1
10	98

Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$

Derivatives show “influence” to functions

$\frac{\partial \mathcal{L}}{\partial a}$ $\mathcal{L} = \mathcal{L}(a)$

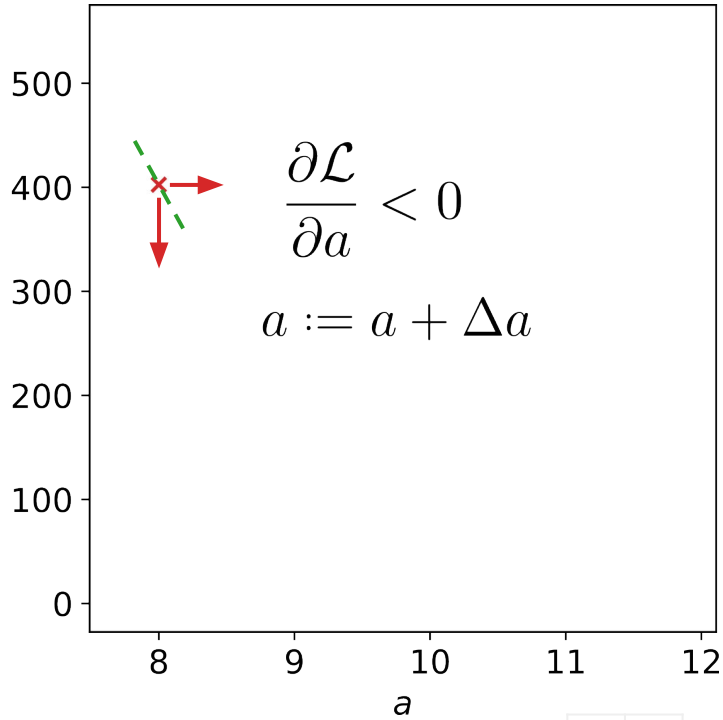
$\mathcal{L}(8) \approx 402$

$\frac{\partial \mathcal{L}}{\partial a}(8) \approx -447$

$\mathcal{L}(11.5) \approx 359$

$\frac{\partial \mathcal{L}}{\partial a}(8) \approx 422$

$\frac{\partial \mathcal{L}}{\partial a} < 0$	$\frac{\partial \mathcal{L}}{\partial a} > 0$
$a := a + \Delta a$	$a := a - \Delta a$



$\hat{y} = ax$

$\mathcal{L} = \sum_i (\hat{y}_i - y_i)^2$

x	y
2	19.6
4.5	44.1
10	98

Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$

Derivatives show “influence” to functions

$\frac{\partial \mathcal{L}}{\partial a}$ $\mathcal{L} = \mathcal{L}(a)$

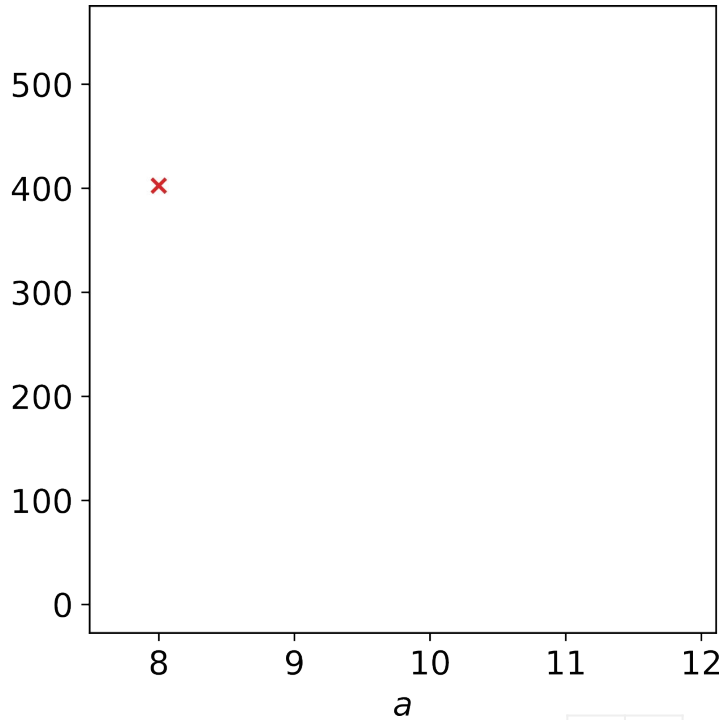
$\mathcal{L}(8) \approx 402$

$\frac{\partial \mathcal{L}}{\partial a}(8) \approx -447$

$\mathcal{L}(11.5) \approx 359$

$\frac{\partial \mathcal{L}}{\partial a}(8) \approx 422$

$\frac{\partial \mathcal{L}}{\partial a} < 0$	$\frac{\partial \mathcal{L}}{\partial a} > 0$
$a := a + \Delta a$	$a := a - \Delta a$



$\hat{y} = ax$

$\mathcal{L} = \sum_i (\hat{y}_i - y_i)^2$

x	y
2	19.6
4.5	44.1
10	98

Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$

Derivatives show “influence” to functions

$\frac{\partial \mathcal{L}}{\partial a}$ $\mathcal{L} = \mathcal{L}(a)$

$\mathcal{L}(8) \approx 402$

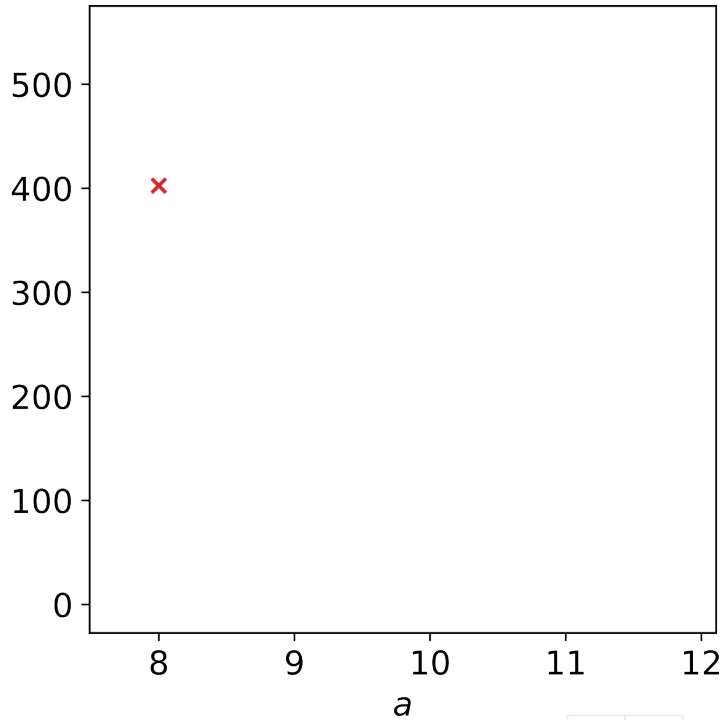
$\frac{\partial \mathcal{L}}{\partial a}(8) \approx -447$

$\mathcal{L}(11.5) \approx 359$

$\frac{\partial \mathcal{L}}{\partial a}(8) \approx 422$

$\frac{\partial \mathcal{L}}{\partial a} < 0$	$\frac{\partial \mathcal{L}}{\partial a} > 0$
$a := a + \Delta a$	$a := a - \Delta a$

$$a := a - \frac{\partial \mathcal{L}}{\partial a}$$



$\hat{y} = ax$

$\mathcal{L} = \sum_i (\hat{y}_i - y_i)^2$

x	y
2	19.6
4.5	44.1
10	98

Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$

Derivatives show “influence” to functions

$\frac{\partial \mathcal{L}}{\partial a}$

$\mathcal{L} = \mathcal{L}(a)$

$\mathcal{L}(8) \approx 402$

$\frac{\partial \mathcal{L}}{\partial a}(8) \approx -447$

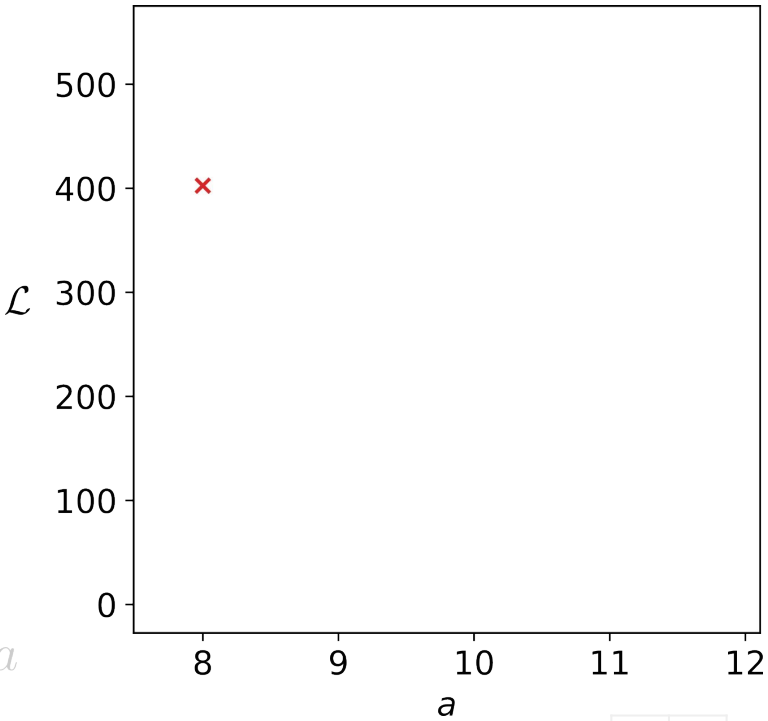
$\mathcal{L}(11.5) \approx 359$

$\frac{\partial \mathcal{L}}{\partial a}(8) \approx 422$

$\frac{\partial \mathcal{L}}{\partial a} < 0$	$\frac{\partial \mathcal{L}}{\partial a} > 0$
$a := a + \Delta a$	$a := a - \Delta a$

$$a := a - \gamma \frac{\partial \mathcal{L}}{\partial a}$$

γ chosen by us



$\hat{y} = ax$

$\mathcal{L} = \sum_i (\hat{y}_i - y_i)^2$

x	y
2	19.6
4.5	44.1
10	98

Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$

Derivatives show “influence” to functions

$\frac{\partial \mathcal{L}}{\partial a}$

$\mathcal{L} = \mathcal{L}(a)$

$\mathcal{L}(8) \approx 402$

$\frac{\partial \mathcal{L}}{\partial a}(8) \approx -447$

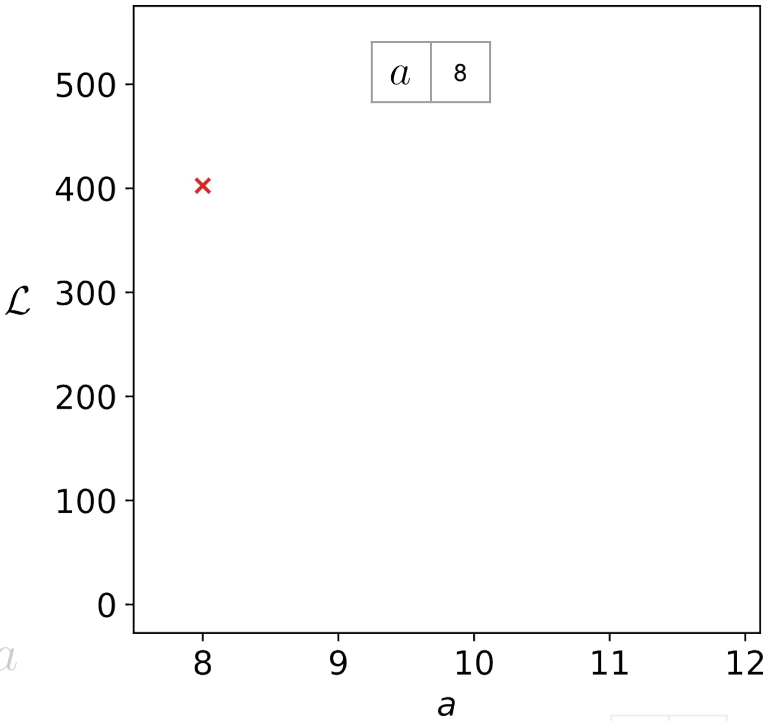
$\mathcal{L}(11.5) \approx 359$

$\frac{\partial \mathcal{L}}{\partial a}(8) \approx 422$

$\frac{\partial \mathcal{L}}{\partial a} < 0$	$\frac{\partial \mathcal{L}}{\partial a} > 0$
$a := a + \Delta a$	$a := a - \Delta a$

$$a := a - \gamma \frac{\partial \mathcal{L}}{\partial a}$$

γ chosen by us



$\hat{y} = ax$

$\mathcal{L} = \sum_i (\hat{y}_i - y_i)^2$

x	y
2	19.6
4.5	44.1
10	98

Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$

Derivatives show “influence” to functions

$\frac{\partial \mathcal{L}}{\partial a}$ $\mathcal{L} = \mathcal{L}(a)$

$\mathcal{L}(8) \approx 402$

$\frac{\partial \mathcal{L}}{\partial a}(8) \approx -447$

$\mathcal{L}(11.5) \approx 359$

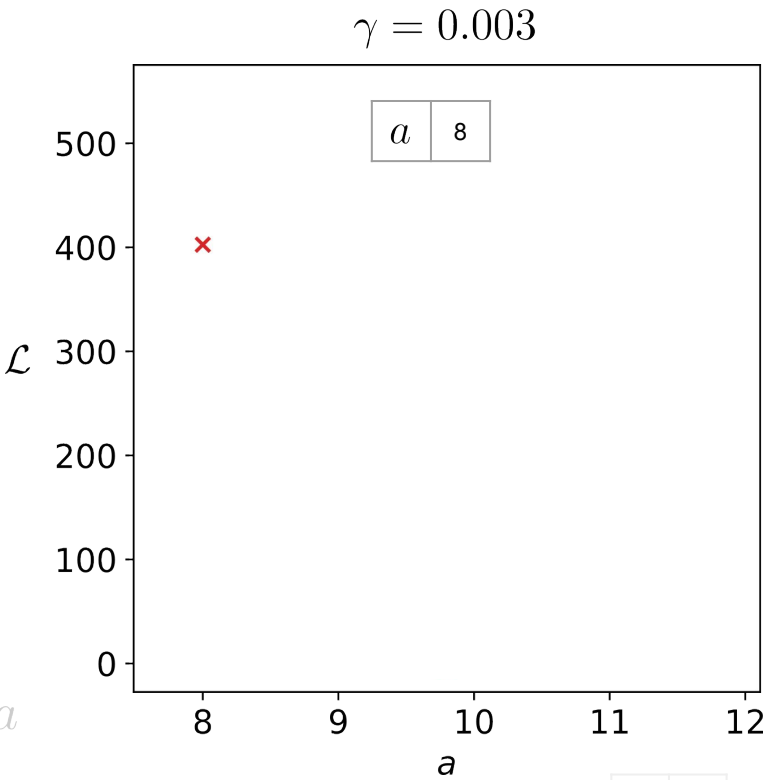
$\frac{\partial \mathcal{L}}{\partial a}(8) \approx 422$

$\frac{\partial \mathcal{L}}{\partial a} < 0$ | $\frac{\partial \mathcal{L}}{\partial a} > 0$

$a := a + \Delta a$ | $a := a - \Delta a$

$$a := a - \gamma \frac{\partial \mathcal{L}}{\partial a}$$

γ chosen by us



$\hat{y} = ax$

$\mathcal{L} = \sum_i (\hat{y}_i - y_i)^2$

x	y
2	19.6
4.5	44.1
10	98

Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$

Derivatives show “influence” to functions

$\frac{\partial \mathcal{L}}{\partial a}$ $\mathcal{L} = \mathcal{L}(a)$

$\mathcal{L}(8) \approx 402$

$\frac{\partial \mathcal{L}}{\partial a}(8) \approx -447$

$\mathcal{L}(11.5) \approx 359$

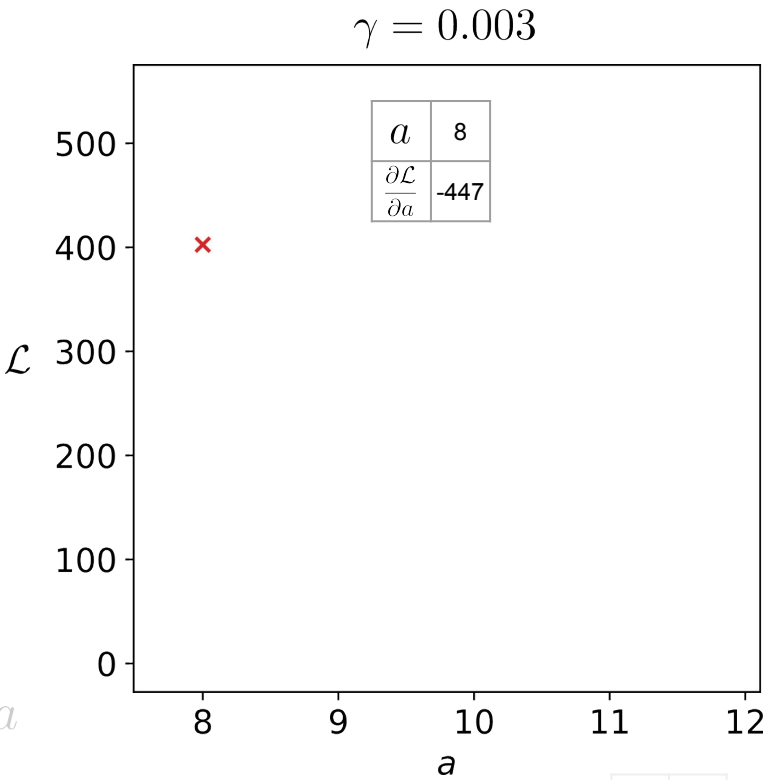
$\frac{\partial \mathcal{L}}{\partial a}(8) \approx 422$

$\frac{\partial \mathcal{L}}{\partial a} < 0$ | $\frac{\partial \mathcal{L}}{\partial a} > 0$

$a := a + \Delta a$ | $a := a - \Delta a$

$$a := a - \gamma \frac{\partial \mathcal{L}}{\partial a}$$

γ chosen by us



$\hat{y} = ax$

$\mathcal{L} = \sum_i (\hat{y}_i - y_i)^2$

x	y
2	19.6
4.5	44.1
10	98

Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$

Derivatives show “influence” to functions

$\frac{\partial \mathcal{L}}{\partial a}$ $\mathcal{L} = \mathcal{L}(a)$

$\mathcal{L}(8) \approx 402$

$\frac{\partial \mathcal{L}}{\partial a}(8) \approx -447$

$\mathcal{L}(11.5) \approx 359$

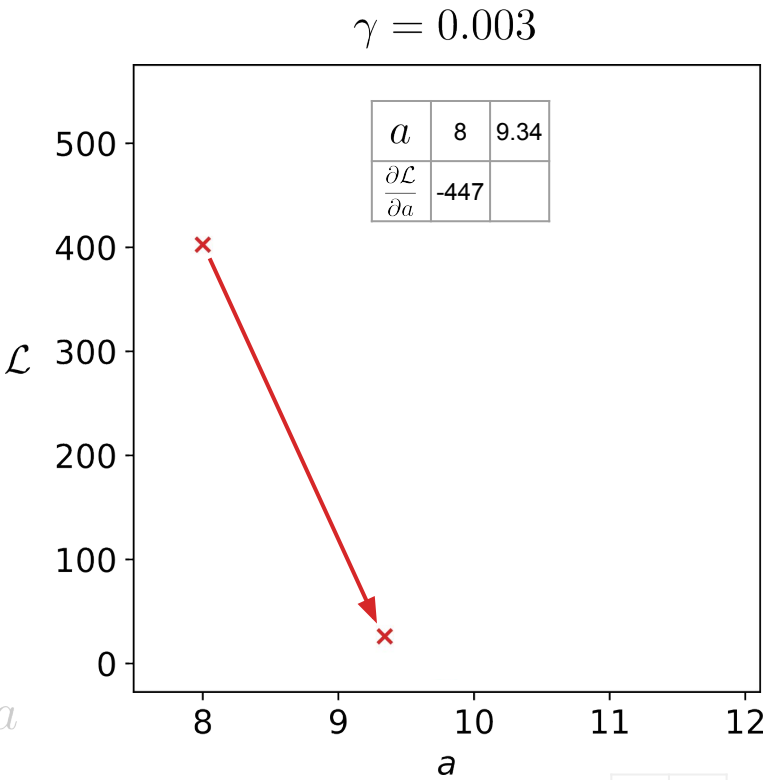
$\frac{\partial \mathcal{L}}{\partial a}(8) \approx 422$

$\frac{\partial \mathcal{L}}{\partial a} < 0$ | $\frac{\partial \mathcal{L}}{\partial a} > 0$

$a := a + \Delta a$ | $a := a - \Delta a$

$$a := a - \gamma \frac{\partial \mathcal{L}}{\partial a}$$

γ chosen by us



$\hat{y} = ax$

$\mathcal{L} = \sum_i (\hat{y}_i - y_i)^2$

x	y
2	19.6
4.5	44.1
10	98

Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$

Derivatives show “influence” to functions

$\frac{\partial \mathcal{L}}{\partial a}$ $\mathcal{L} = \mathcal{L}(a)$

$\mathcal{L}(8) \approx 402$

$\frac{\partial \mathcal{L}}{\partial a}(8) \approx -447$

$\mathcal{L}(11.5) \approx 359$

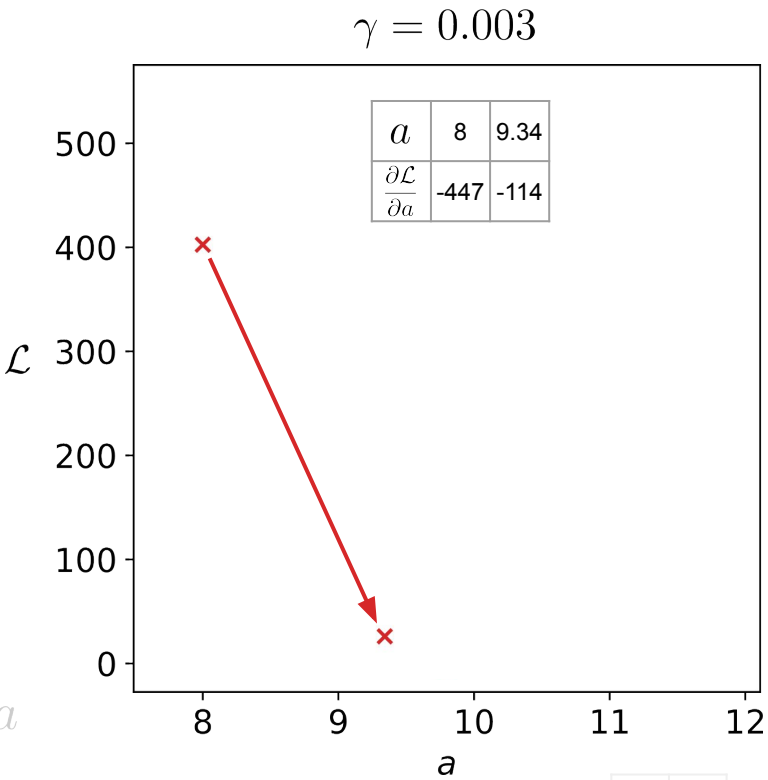
$\frac{\partial \mathcal{L}}{\partial a}(8) \approx 422$

$\frac{\partial \mathcal{L}}{\partial a} < 0$ | $\frac{\partial \mathcal{L}}{\partial a} > 0$

$a := a + \Delta a$ | $a := a - \Delta a$

$$a := a - \gamma \frac{\partial \mathcal{L}}{\partial a}$$

γ chosen by us



$\hat{y} = ax$

$\mathcal{L} = \sum_i (\hat{y}_i - y_i)^2$

x	y
2	19.6
4.5	44.1
10	98

Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$

Derivatives show “influence” to functions

$\frac{\partial \mathcal{L}}{\partial a}$ $\mathcal{L} = \mathcal{L}(a)$

$\mathcal{L}(8) \approx 402$

$\frac{\partial \mathcal{L}}{\partial a}(8) \approx -447$

$\mathcal{L}(11.5) \approx 359$

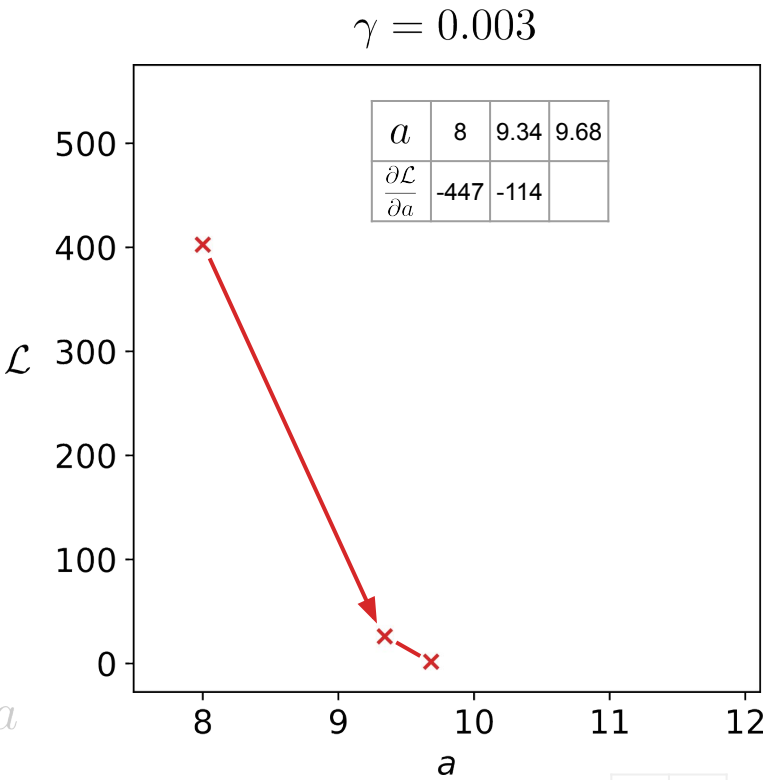
$\frac{\partial \mathcal{L}}{\partial a}(8) \approx 422$

$\frac{\partial \mathcal{L}}{\partial a} < 0$ | $\frac{\partial \mathcal{L}}{\partial a} > 0$

$a := a + \Delta a$ | $a := a - \Delta a$

$$a := a - \gamma \frac{\partial \mathcal{L}}{\partial a}$$

γ chosen by us



$\hat{y} = ax$

$\mathcal{L} = \sum_i (\hat{y}_i - y_i)^2$

x	y
2	19.6
4.5	44.1
10	98

Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$

Derivatives show “influence” to functions

$\frac{\partial \mathcal{L}}{\partial a}$ $\mathcal{L} = \mathcal{L}(a)$

$\mathcal{L}(8) \approx 402$

$\frac{\partial \mathcal{L}}{\partial a}(8) \approx -447$

$\mathcal{L}(11.5) \approx 359$

$\frac{\partial \mathcal{L}}{\partial a}(8) \approx 422$

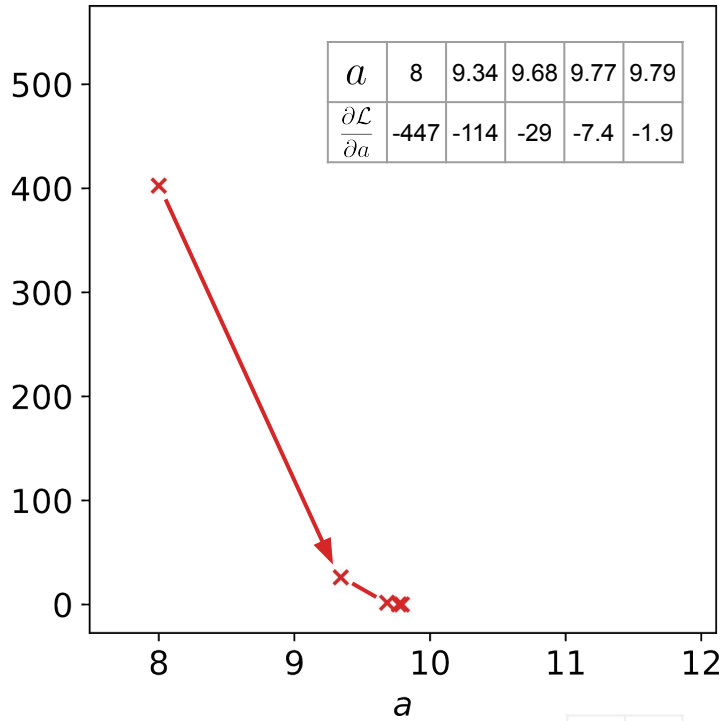
$\frac{\partial \mathcal{L}}{\partial a} < 0$ | $\frac{\partial \mathcal{L}}{\partial a} > 0$

$a := a + \Delta a$ | $a := a - \Delta a$

$$a := a - \gamma \frac{\partial \mathcal{L}}{\partial a}$$

γ chosen by us

$\gamma = 0.003$



a	8	9.34	9.68	9.77	9.79
$\frac{\partial \mathcal{L}}{\partial a}$	-447	-114	-29	-7.4	-1.9

$\hat{y} = ax$

$\mathcal{L} = \sum_i (\hat{y}_i - y_i)^2$

x	y
2	19.6
4.5	44.1
10	98

Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$

Derivatives show “influence” to functions

$\frac{\partial \mathcal{L}}{\partial a}$ $\mathcal{L} = \mathcal{L}(a)$

$\mathcal{L}(8) \approx 402$

$\frac{\partial \mathcal{L}}{\partial a}(8) \approx -447$

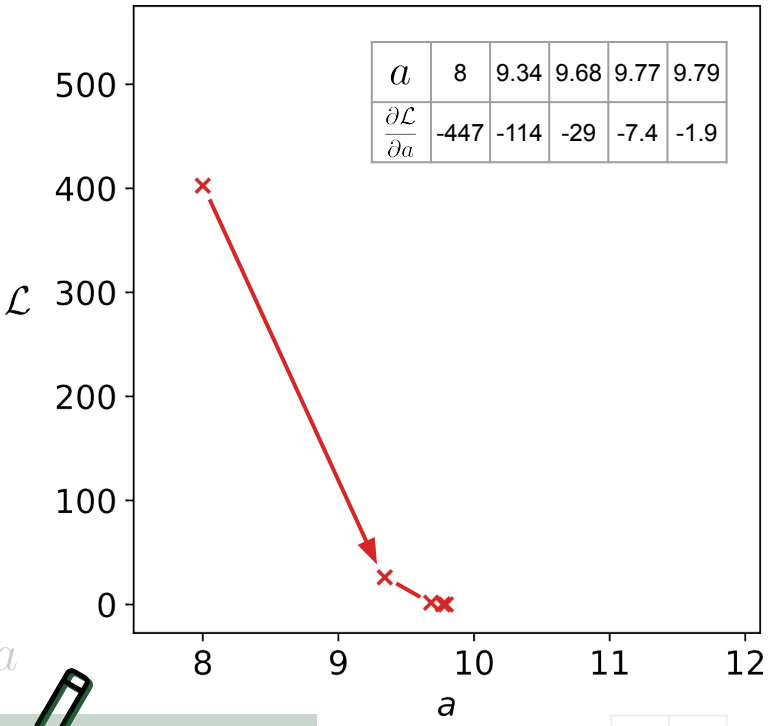
$\mathcal{L}(11.5) \approx 359$

$\frac{\partial \mathcal{L}}{\partial a}(8) \approx 422$

$\frac{\partial \mathcal{L}}{\partial a} < 0$ | $\frac{\partial \mathcal{L}}{\partial a} > 0$

$a := a + \Delta a$ | $a := a - \Delta a$

$\gamma = 0.003$



a	8	9.34	9.68	9.77	9.79
$\frac{\partial \mathcal{L}}{\partial a}$	-447	-114	-29	-7.4	-1.9

repeat:

$$a := a - \gamma \frac{\partial \mathcal{L}}{\partial a}$$

until minimum is reached

$(\hat{y}_i - y_i)^2$

x	y
2	19.6
4.5	44.1
10	98

Gradient Descent

How to find the minimum? $\min_a \mathcal{L}$

Derivatives show “influence” to functions

$$\frac{\partial \mathcal{L}}{\partial a} \qquad \mathcal{L} = \mathcal{L}(a)$$

$\mathcal{L}(8) \approx 402$

$\frac{\partial \mathcal{L}}{\partial a}(8) \approx -447$

$\mathcal{L}(11.5) \approx 359$

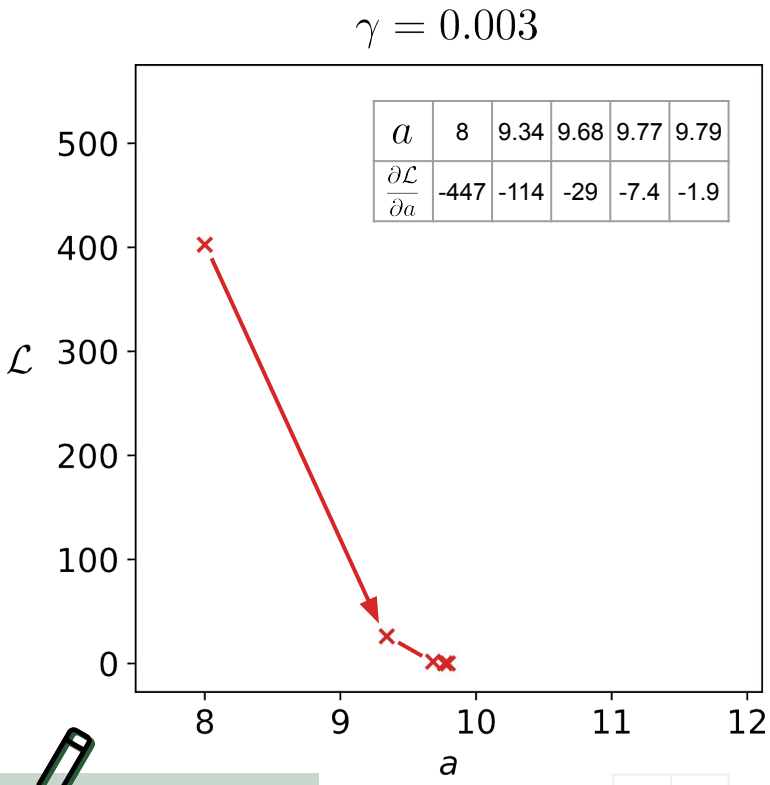
$\frac{\partial \mathcal{L}}{\partial a}(8) \approx 422$

γ – *learning rate*

repeat:

$$a := a - \gamma \frac{\partial \mathcal{L}}{\partial a}$$

until minimum is reached



a	8	9.34	9.68	9.77	9.79
$\frac{\partial \mathcal{L}}{\partial a}$	-447	-114	-29	-7.4	-1.9

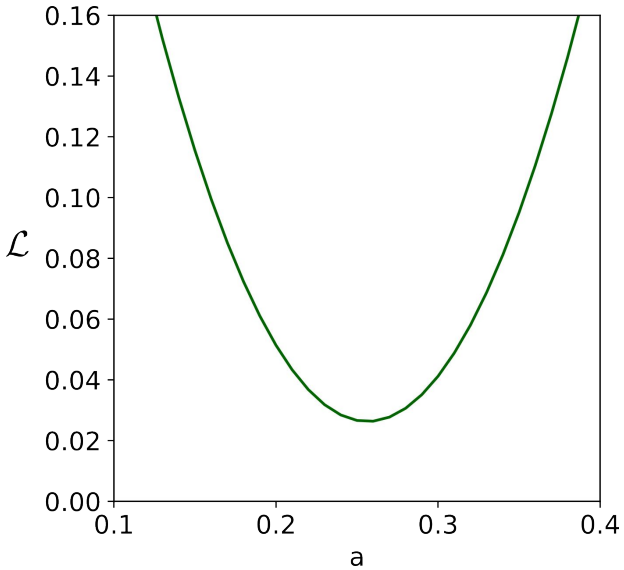
$\sum (\hat{y}_i - y_i)^2$

x	y
2	19.6
4.5	44.1
10	98

Gradient Descent

x	y
-0.3	1.9
-0.2	1.85
0.2	1.85
0.3	1.85

$$\hat{y} = ax + b - 2a^2$$

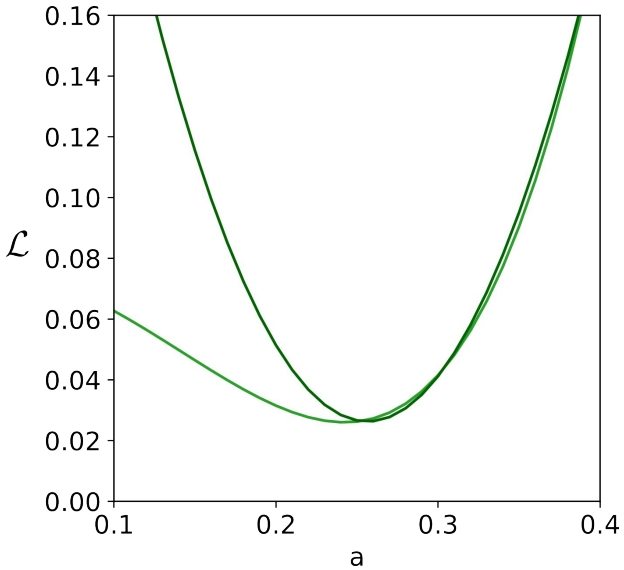


b = 2

Gradient Descent

x	y
-0.3	1.9
-0.2	1.85
0.2	1.85
0.3	1.85

$$\hat{y} = ax + b - 2a^2$$



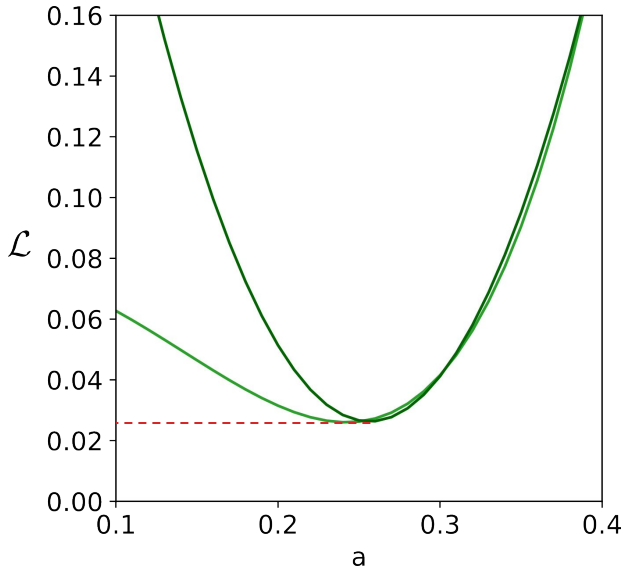
$$\hat{y} = -5.5(a - 0.246)x + b - 0.217$$

b = 2

Gradient Descent

x	y
-0.3	1.9
-0.2	1.85
0.2	1.85
0.3	1.85

$$\hat{y} = ax + b - 2a^2$$



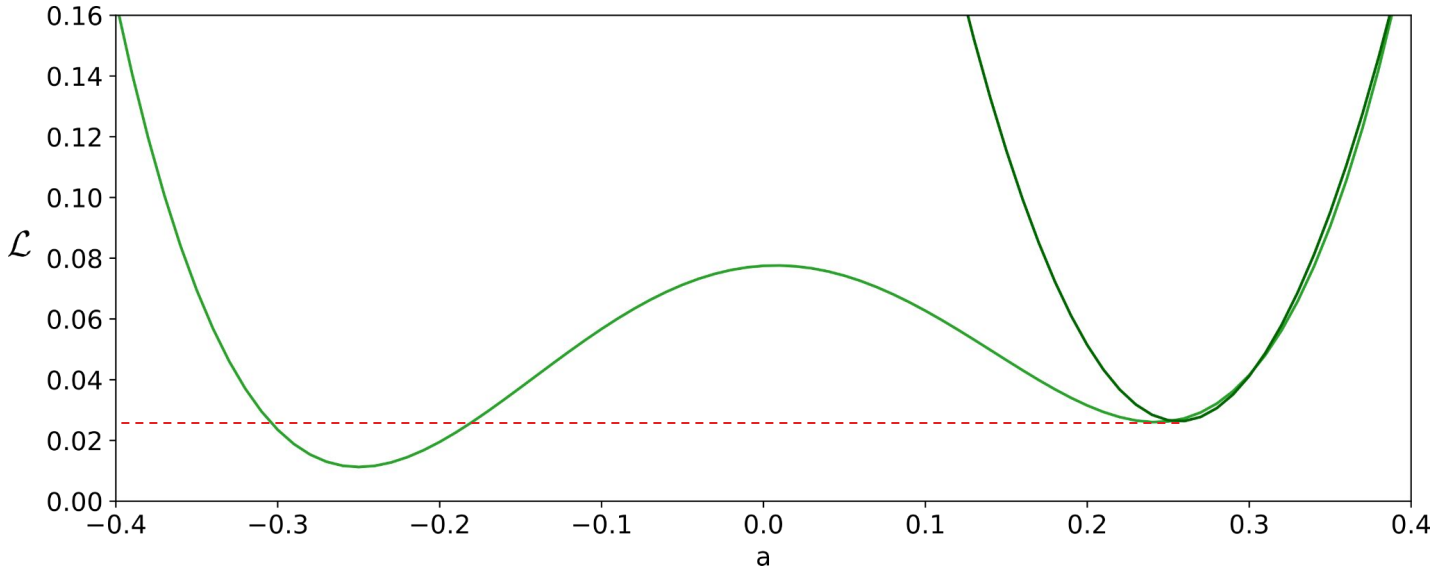
$$\hat{y} = -5.5(a - 0.246)x + b - 0.217$$

b = 2

Gradient Descent

x	y
-0.3	1.9
-0.2	1.85
0.2	1.85
0.3	1.85

$$\hat{y} = ax + b - 2a^2$$



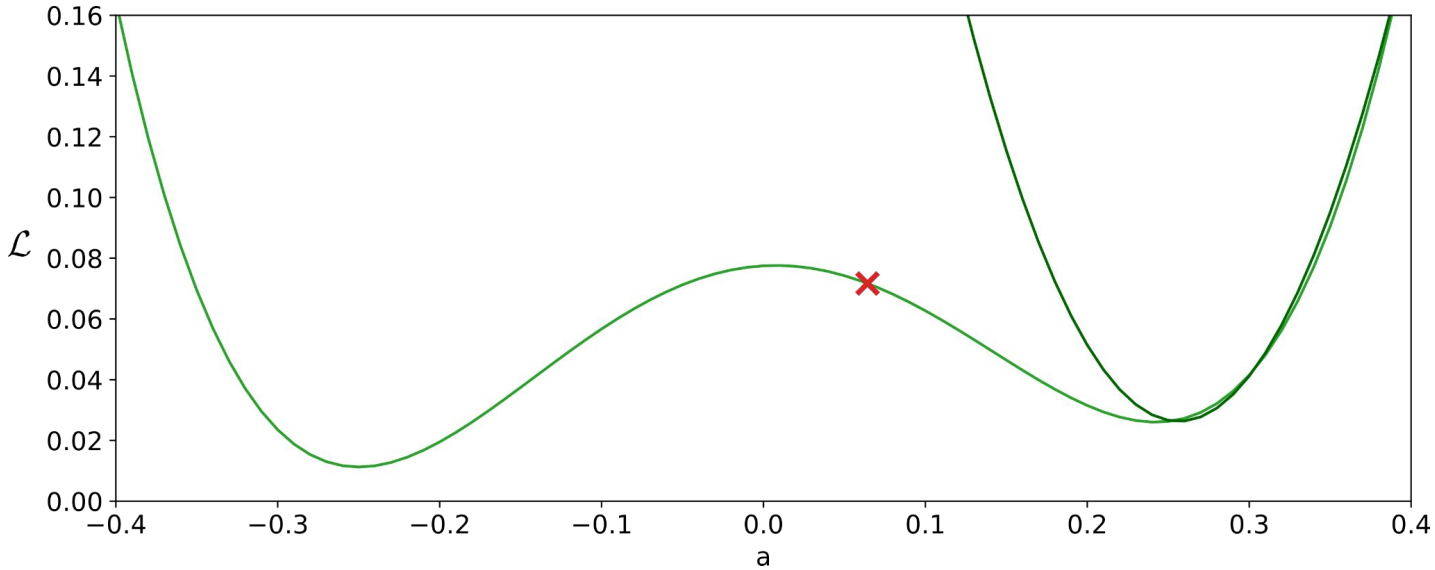
$$\hat{y} = -5.5(a - 0.246)x + b - 0.217$$

b = 2

Gradient Descent

x	y
-0.3	1.9
-0.2	1.85
0.2	1.85
0.3	1.85

$$\hat{y} = ax + b - 2a^2$$



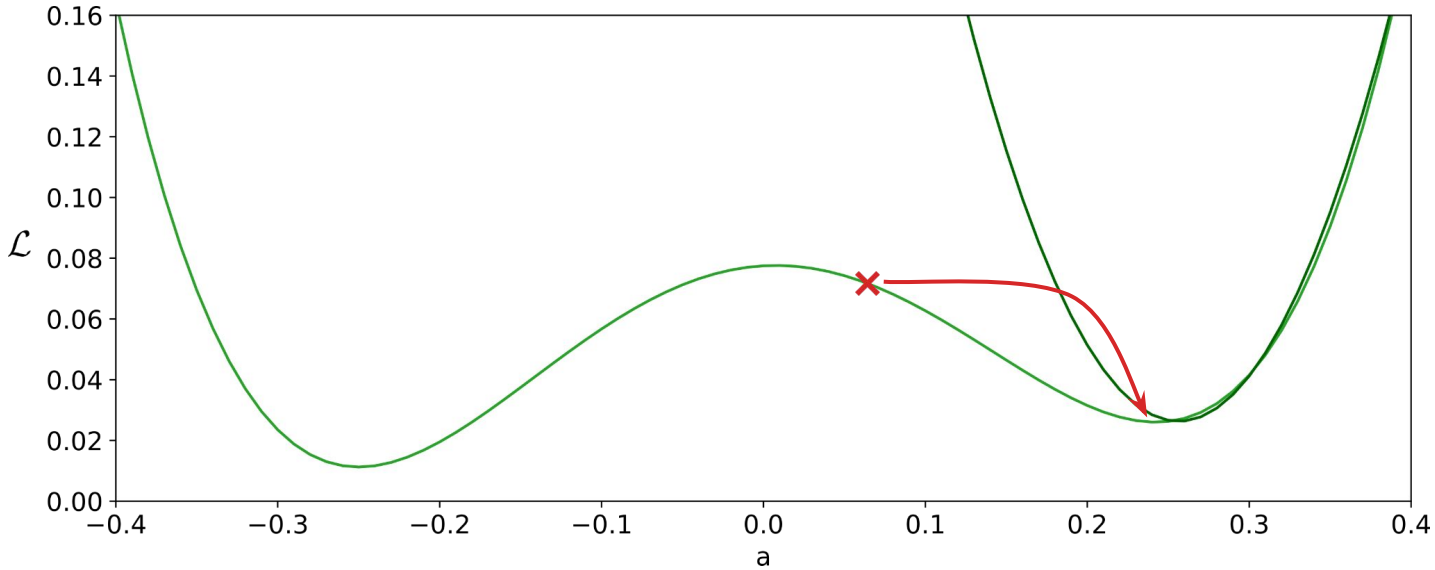
$$\hat{y} = -5.5(a - 0.246)x + b - 0.217$$

b = 2

Gradient Descent

x	y
-0.3	1.9
-0.2	1.85
0.2	1.85
0.3	1.85

$$\hat{y} = ax + b - 2a^2$$



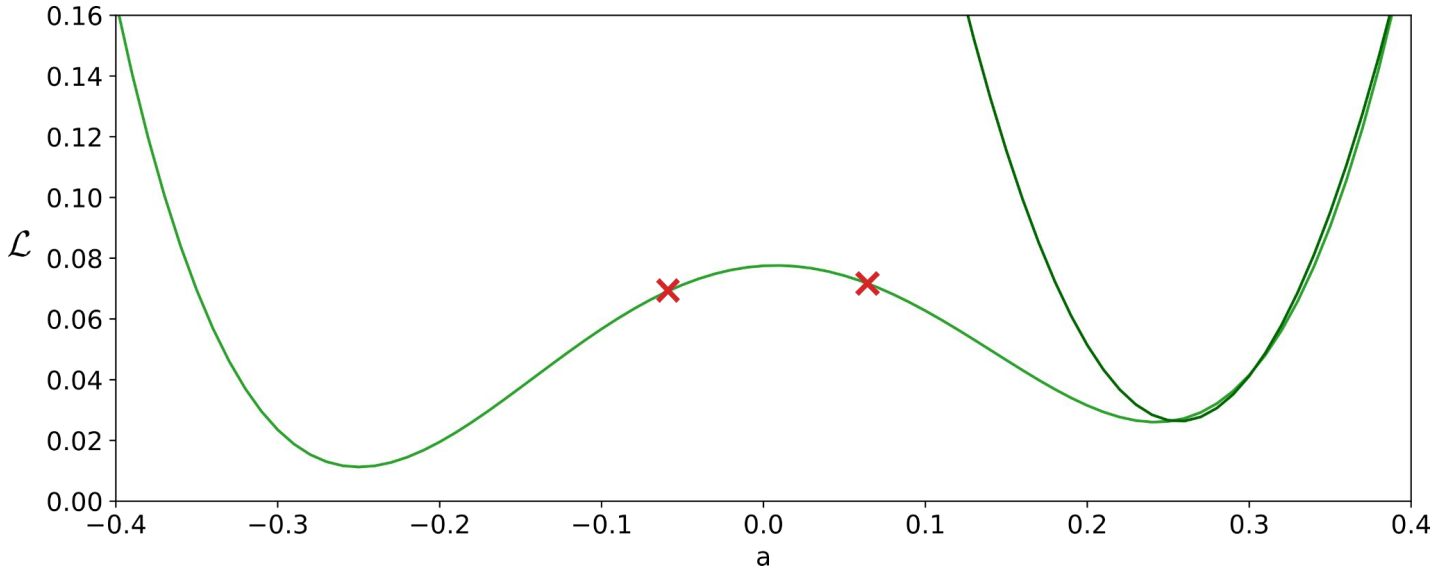
$$\hat{y} = -5.5(a - 0.246)x + b - 0.217$$

b = 2

Gradient Descent

x	y
-0.3	1.9
-0.2	1.85
0.2	1.85
0.3	1.85

$$\hat{y} = ax + b - 2a^2$$



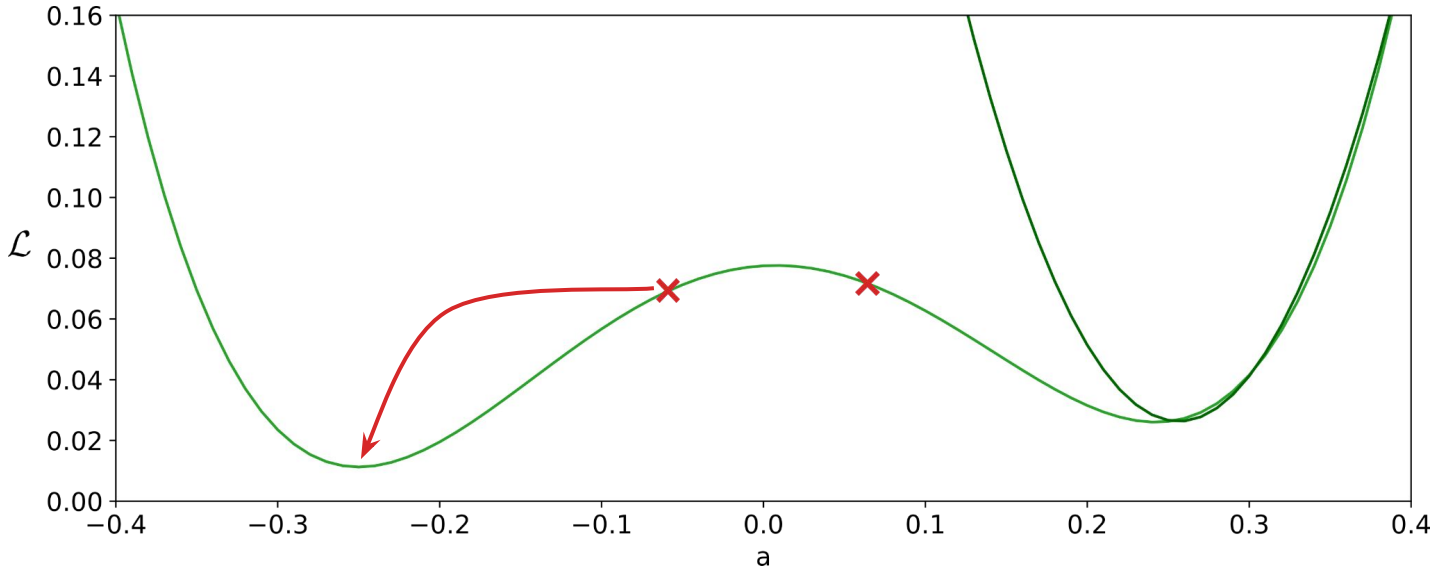
$$\hat{y} = -5.5(a - 0.246)x + b - 0.217$$

b = 2

Gradient Descent

x	y
-0.3	1.9
-0.2	1.85
0.2	1.85
0.3	1.85

$$\hat{y} = ax + b - 2a^2$$



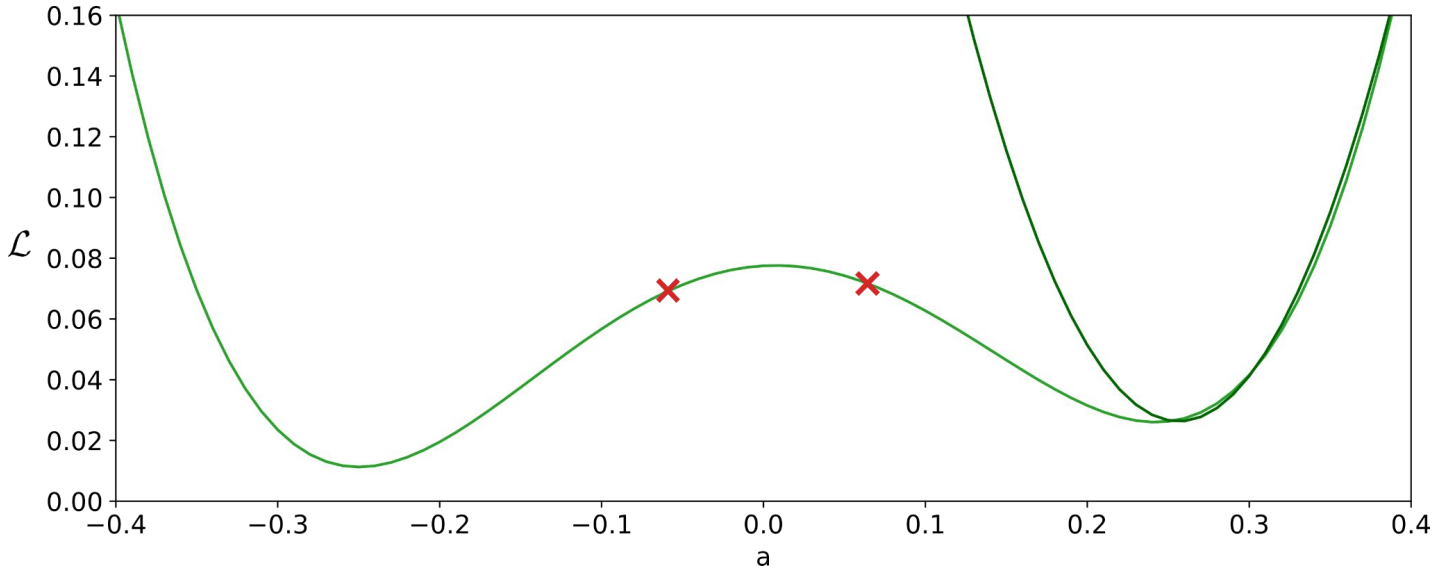
$$\hat{y} = -5.5(a - 0.246)x + b - 0.217$$

b = 2

Gradient Descent

x	y
-0.3	1.9
-0.2	1.85
0.2	1.85
0.3	1.85

$$\hat{y} = ax + b - 2a^2$$

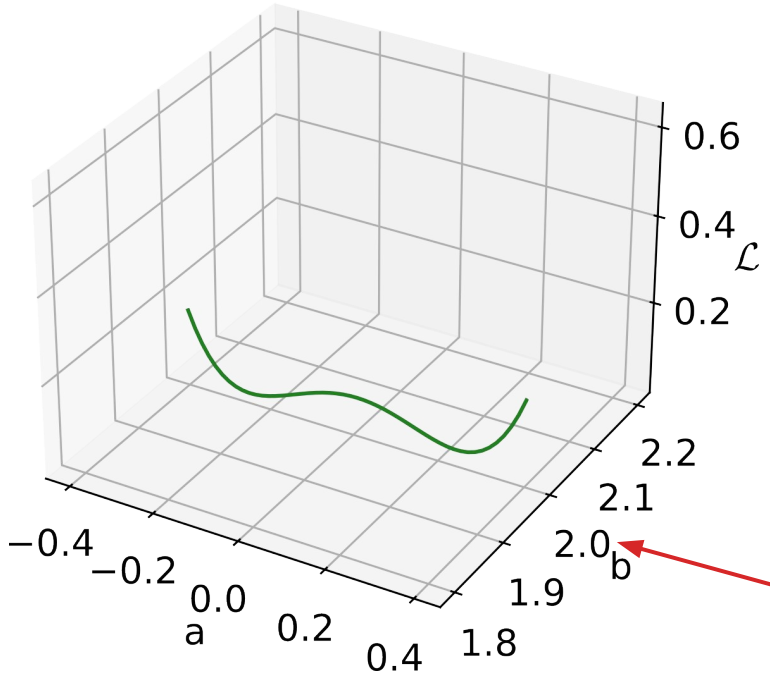


$$\hat{y} = -5.5(a - 0.246)x + b - 0.217$$

b = 2

Gradient Descent

x	y
-0.3	1.9
-0.2	1.85
0.2	1.85
0.3	1.85

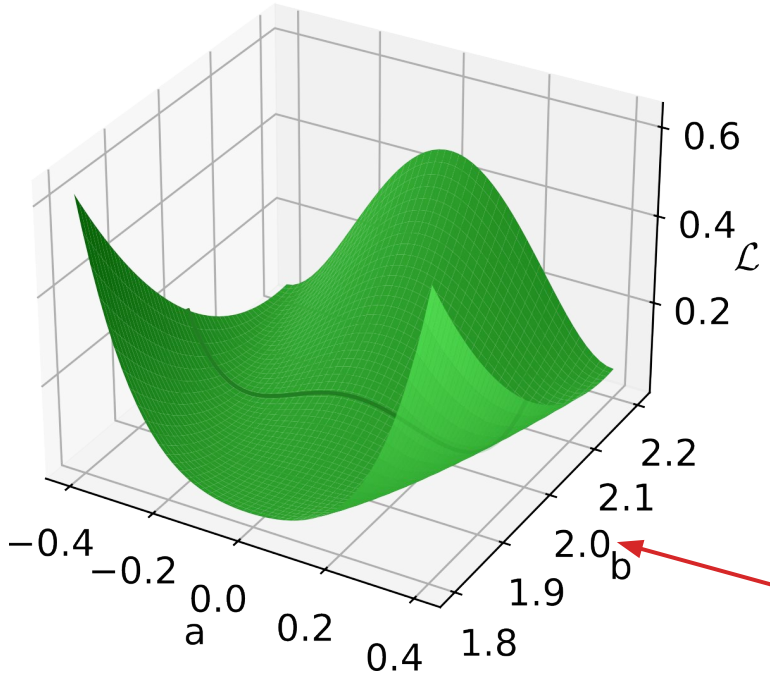


$$\hat{y} = -5.5(a - 0.246)x + b - 0.217$$

b = 2

Gradient Descent

x	y
-0.3	1.9
-0.2	1.85
0.2	1.85
0.3	1.85

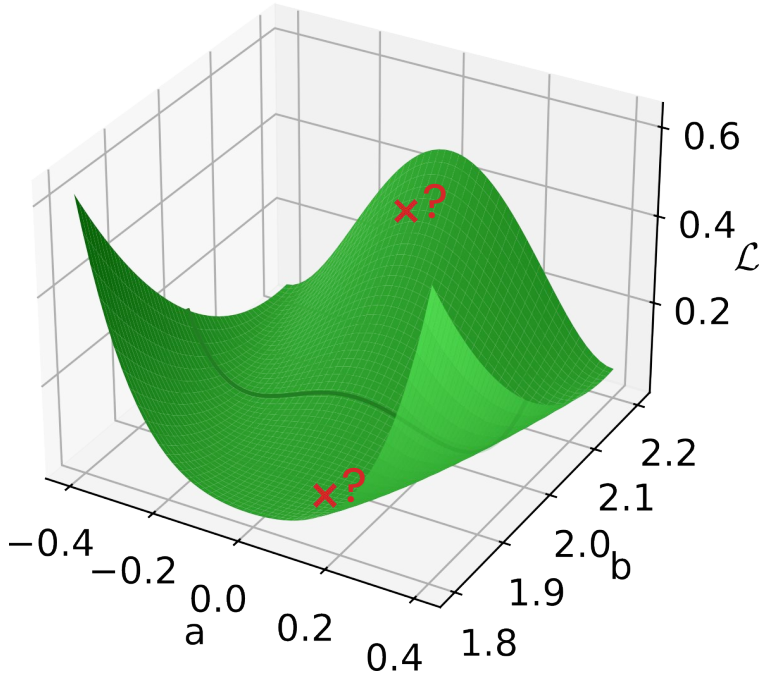


$$\hat{y} = -5.5(a - 0.246)x + b - 0.217$$

b = 2

Gradient Descent

x	y
-0.3	1.9
-0.2	1.85
0.2	1.85
0.3	1.85



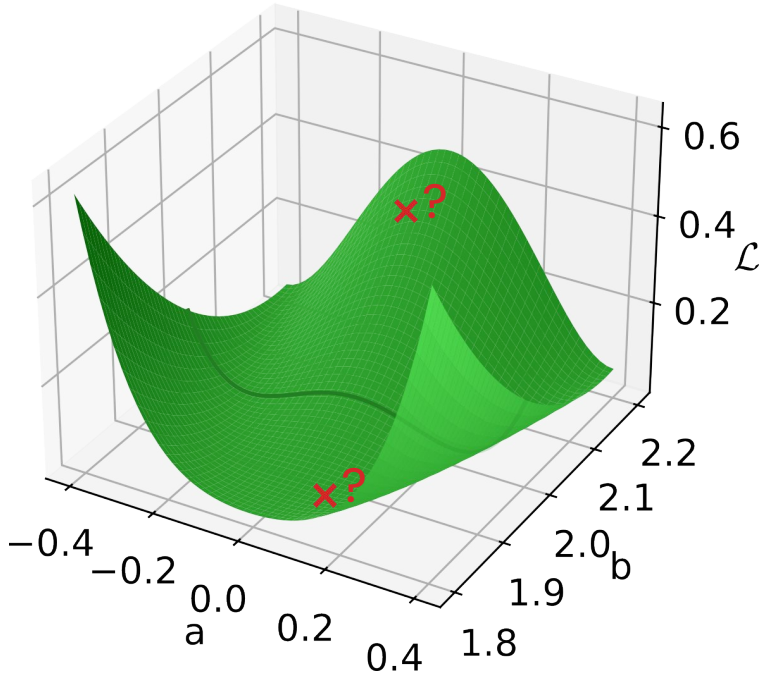
$$\hat{y} = -5.5(a - 0.246)x + b - 0.217$$

b = 2

Gradient Descent

x	y
-0.3	1.9
-0.2	1.85
0.2	1.85
0.3	1.85

$$\theta = \{a, b\}$$

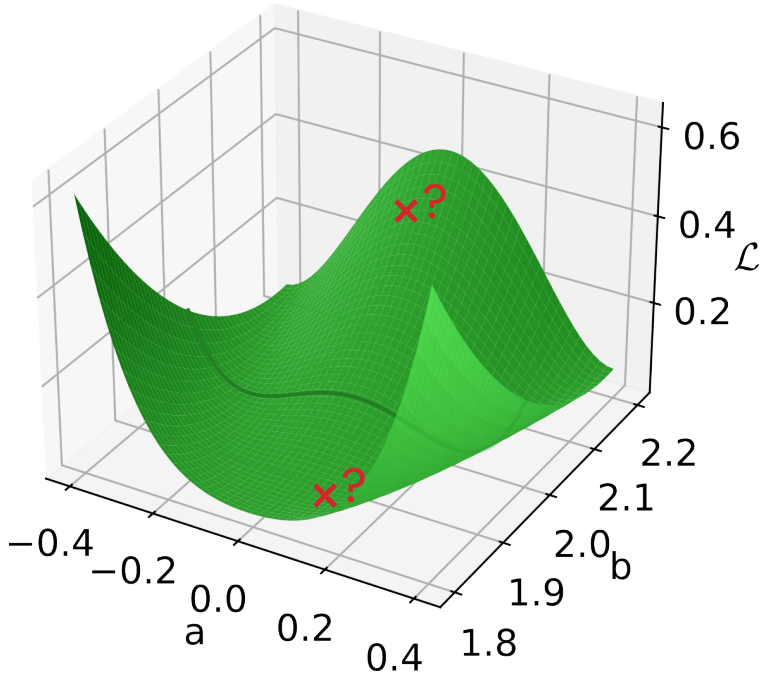


$$\hat{y} = -5.5(a - 0.246)x + b - 0.217$$

b = 2

Gradient Descent

x	y
-0.3	1.9
-0.2	1.85
0.2	1.85
0.3	1.85



$$\theta = \{a, b\}$$

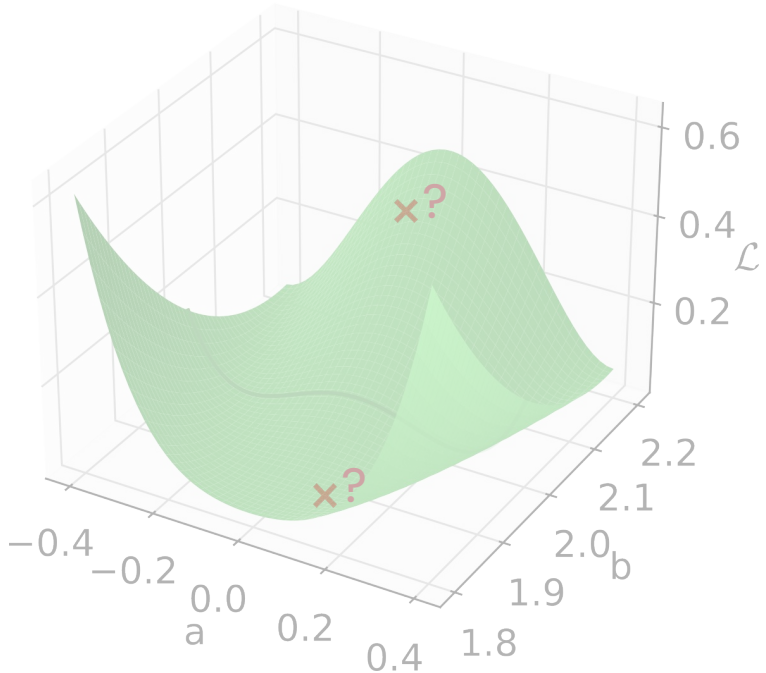
$$\theta = \{a, b, c, \dots\}$$

$$\hat{y} = -5.5(a - 0.246)x + b - 0.217$$

b = 2

Gradient Descent

x	y
-0.3	1.9
-0.2	1.85
0.2	1.85
0.3	1.85



$$\theta = \{a, b\}$$

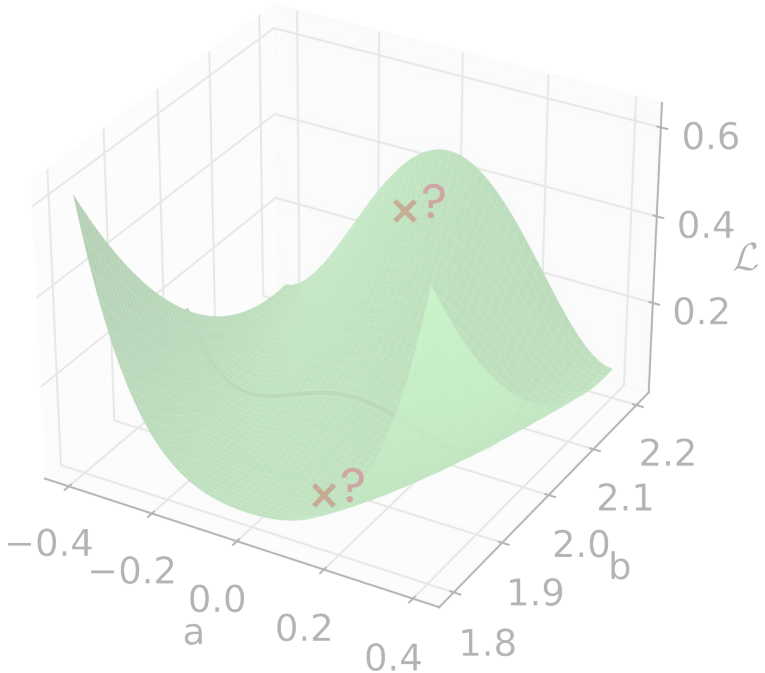
$$\theta = \{a, b, c, \dots\}$$

$$\hat{y} = -5.5(a - 0.246)x + b - 0.217$$

b = 2

Gradient Descent

x	y
-0.3	1.9
-0.2	1.85
0.2	1.85
0.3	1.85
...	...
...	...
...	...



$$\theta = \{a, b\}$$

$$\theta = \{a, b, c, \dots\}$$

$$\hat{y} = -5.5(a - 0.246)x + b - 0.217$$

b = 2

Loading the Data

size of dataset $\gg$ *computer memory*

Loading the Data

size of dataset $\gg$ *computer memory*

→ *batching*

Stochastic Gradient Descent

$$\mathcal{L}_\theta = \frac{1}{N}\sum_{i}^{N} \ell_\theta(x_i, y_i) \qquad \text{e.g., } \ell_\theta(x_i, y_i) = \left(\hat{y}(x_i) - y_i\right)^2$$

Stochastic Gradient Descent

$$\mathcal{L}_\theta = \frac{1}{N}\sum_i^N \ell_\theta(x_i, y_i) \qquad \text{e.g., } \ell_\theta(x_i, y_i) = \left(\hat{y}(x_i) - y_i\right)^2$$

Gradient Descent $\quad \theta_{t+1} = \theta_t - \gamma \frac{\partial \mathcal{L}}{\partial \theta_t}$

Stochastic Gradient Descent

$$\mathcal{L}_\theta = \frac{1}{N}\sum_i^N \ell_\theta(x_i, y_i) \qquad \text{e.g., } \ell_\theta(x_i, y_i) = \left(\hat{y}(x_i) - y_i\right)^2$$

(Full) Batch Gradient Descent $\theta_{t+1} = \theta_t - \gamma \frac{\partial \mathcal{L}}{\partial \theta_t}$

Stochastic Gradient Descent

$$\mathcal{L}_\theta = \frac{1}{N}\sum_i^N \ell_\theta(x_i, y_i) \qquad \text{e.g., } \ell_\theta(x_i, y_i) = \left(\hat{y}(x_i) - y_i\right)^2$$

(Full) Batch Gradient Descent $\theta_{t+1} = \theta_t - \gamma \frac{\partial \mathcal{L}}{\partial \theta_t}$

Stochastic Gradient Descent $\theta_{t+1} = \theta_t - \gamma \frac{\partial \ell}{\partial \theta_t}$

Stochastic Gradient Descent

$$\mathcal{L}_\theta = \frac{1}{N}\sum_i^N \ell_\theta(x_i, y_i) \qquad \text{e.g., } \ell_\theta(x_i, y_i) = \left(\hat{y}(x_i) - y_i\right)^2$$

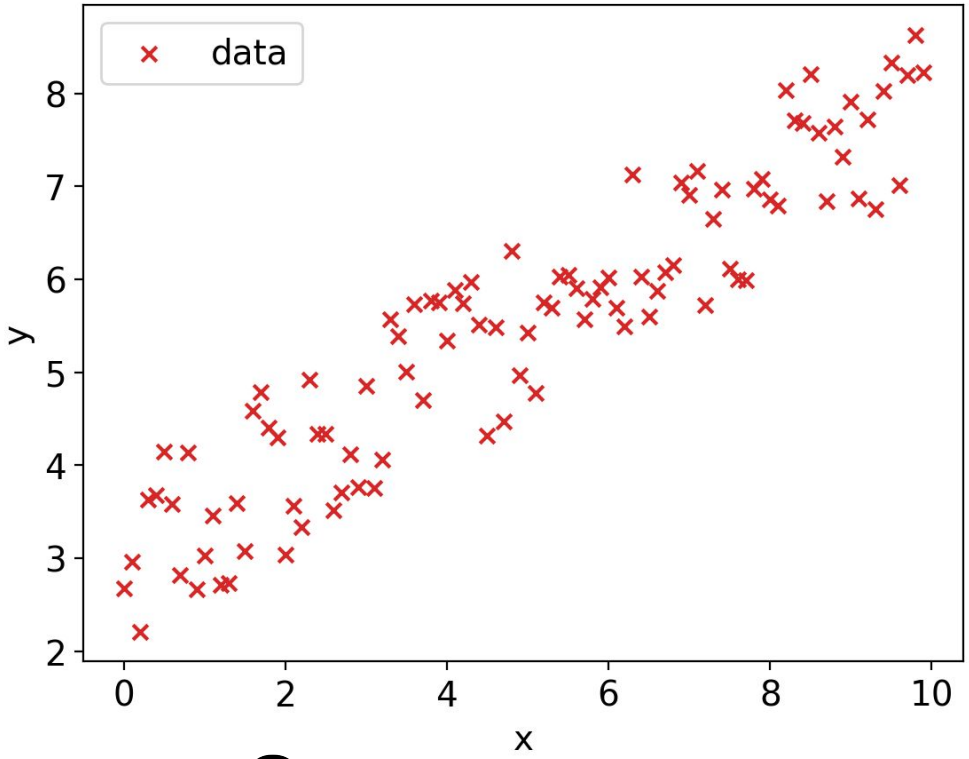
(Full) Batch Gradient Descent $\theta_{t+1} = \theta_t - \gamma \frac{\partial \mathcal{L}}{\partial \theta_t}$

Stochastic Gradient Descent $\theta_{t+1} = \theta_t - \gamma \frac{\partial \ell}{\partial \theta_t}$

Mini-Batch Gradient Descent $\theta_{t+1} = \theta_t - \gamma \frac{\partial \frac{1}{|B|}\sum_{b \in B} \ell(x_b, y_b)}{\partial \theta_t}$

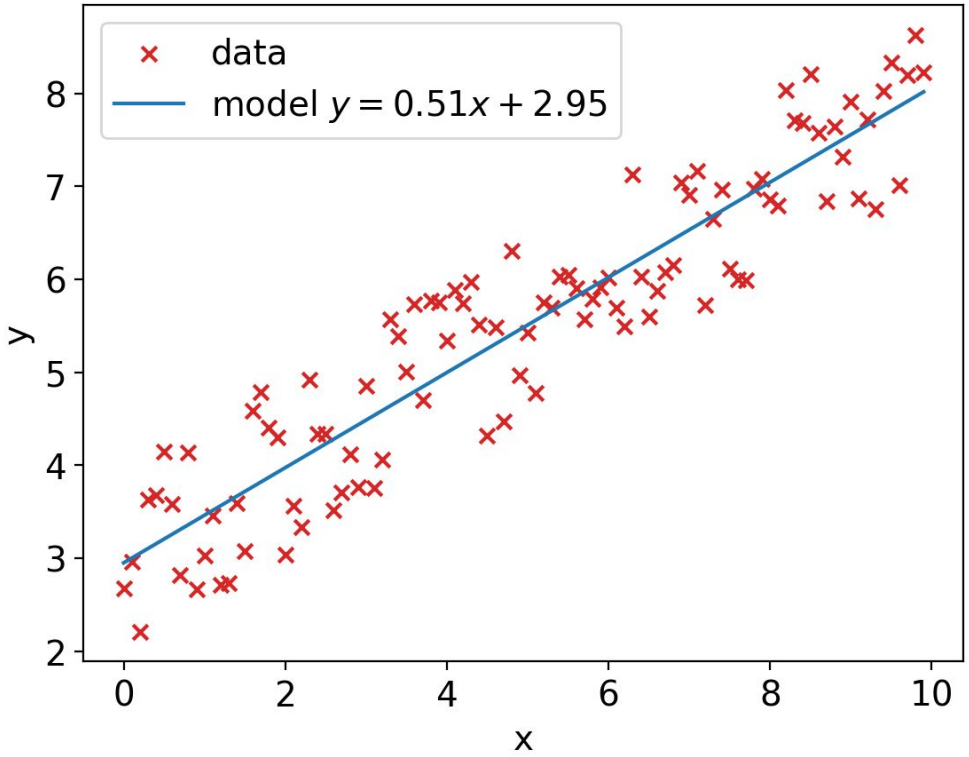
Models

Models



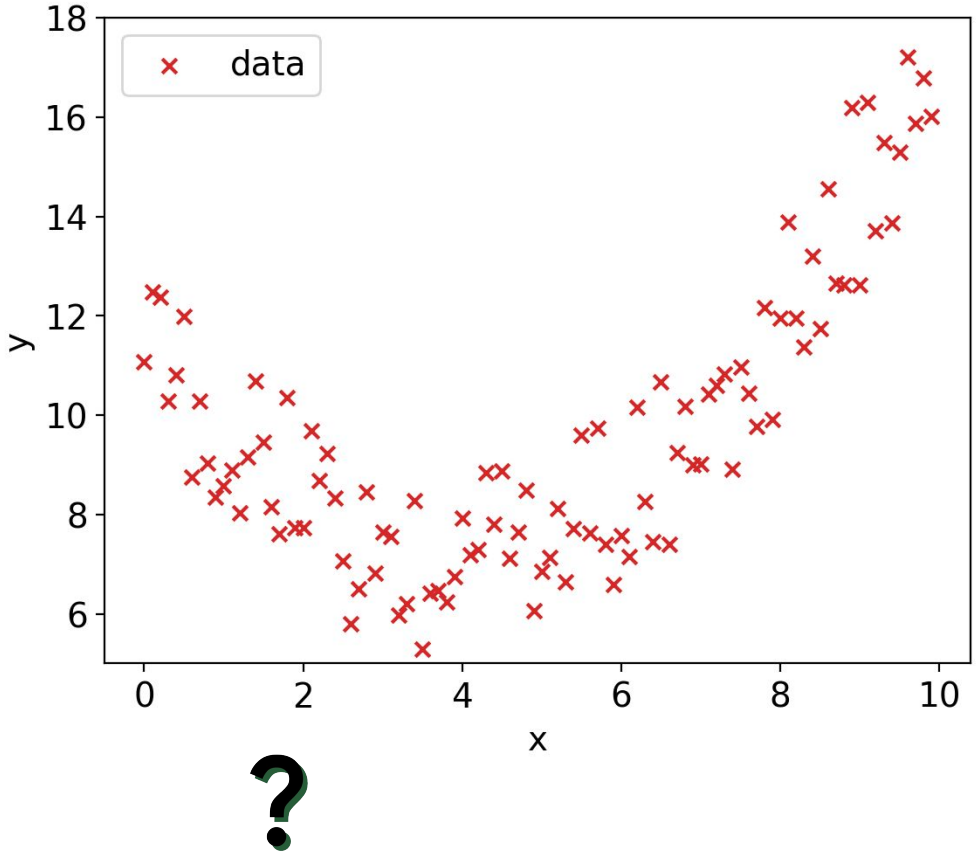
?

Models

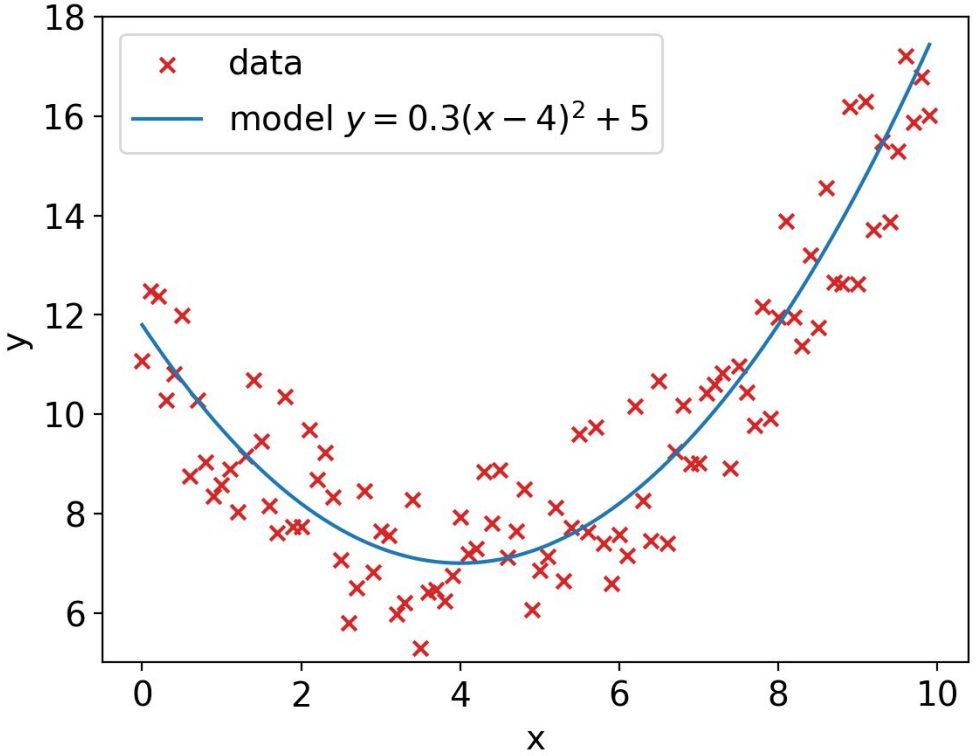


⇒ *Linear model*

Models

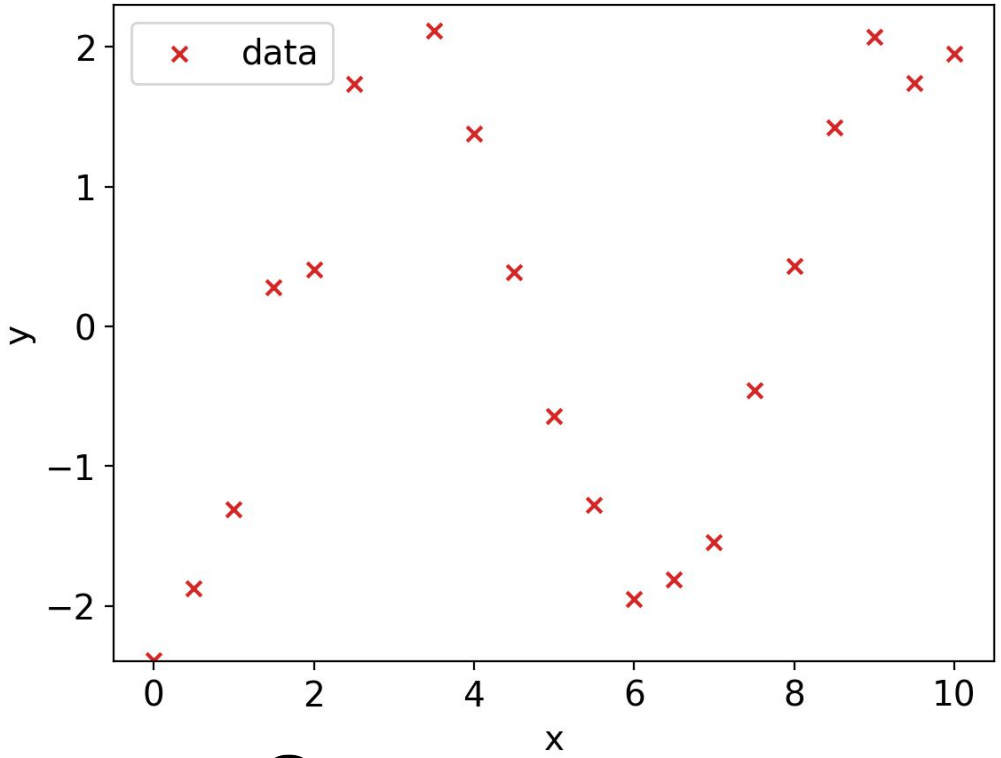


Models



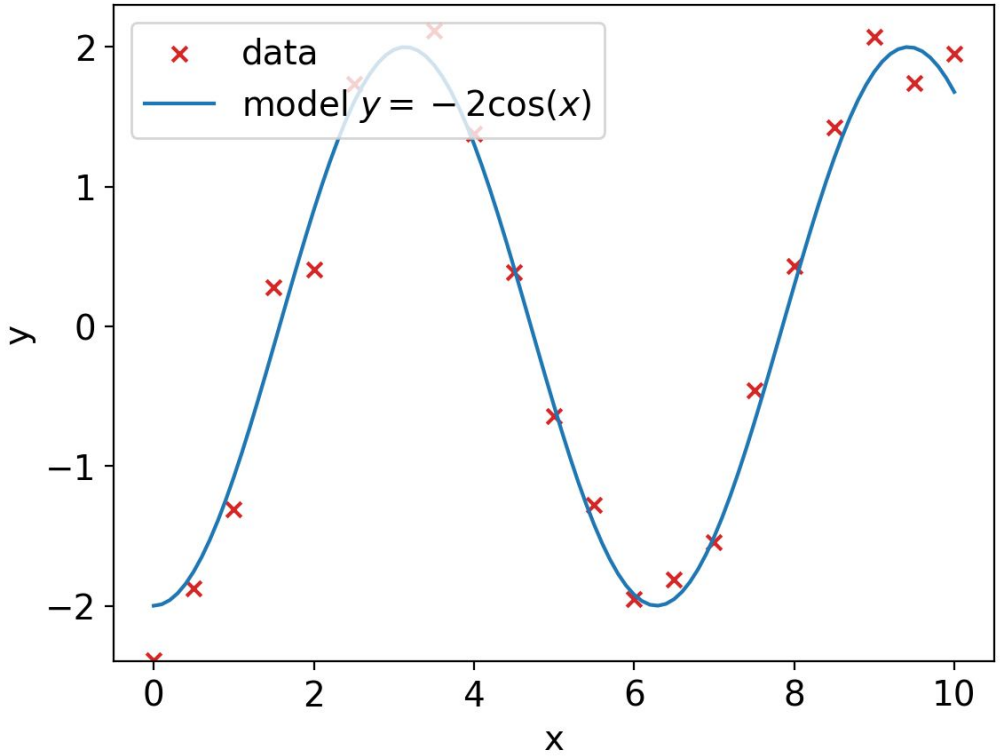
⇒ *Quadratic model*

Models



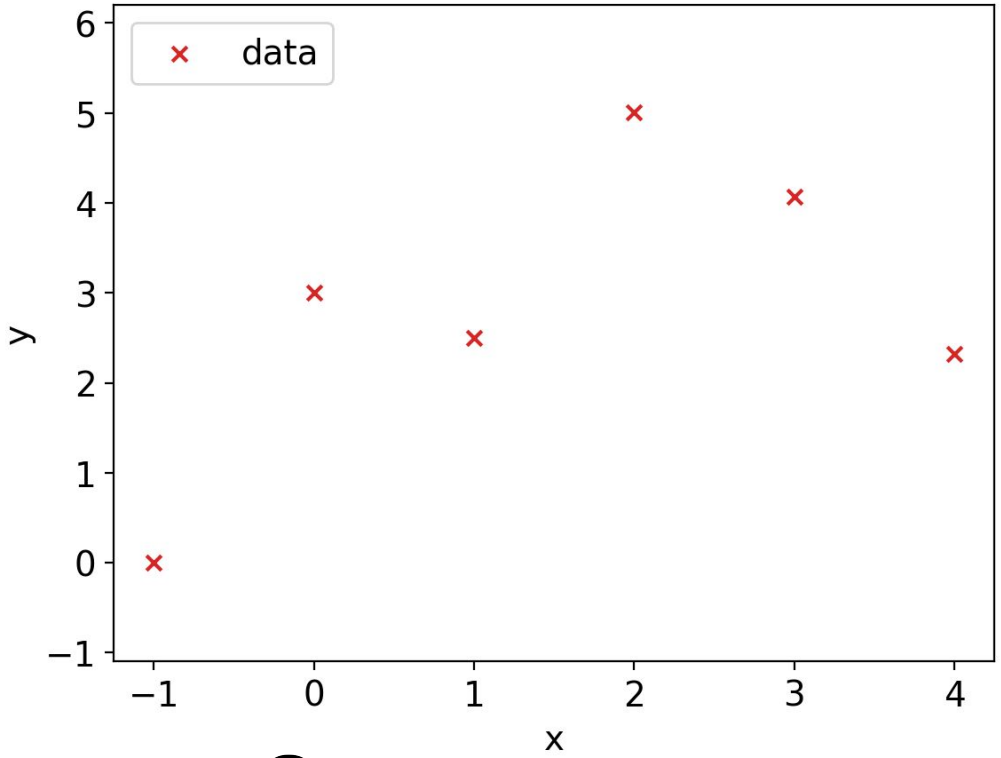
?

Models



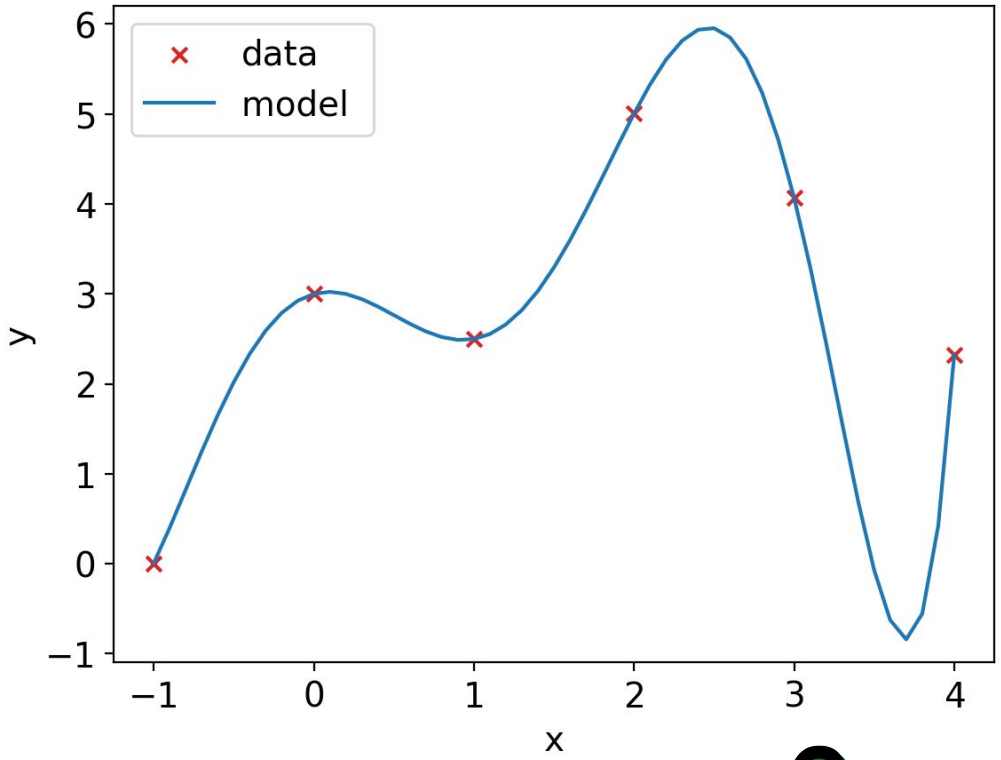
⇒ *Cosine model*

Models



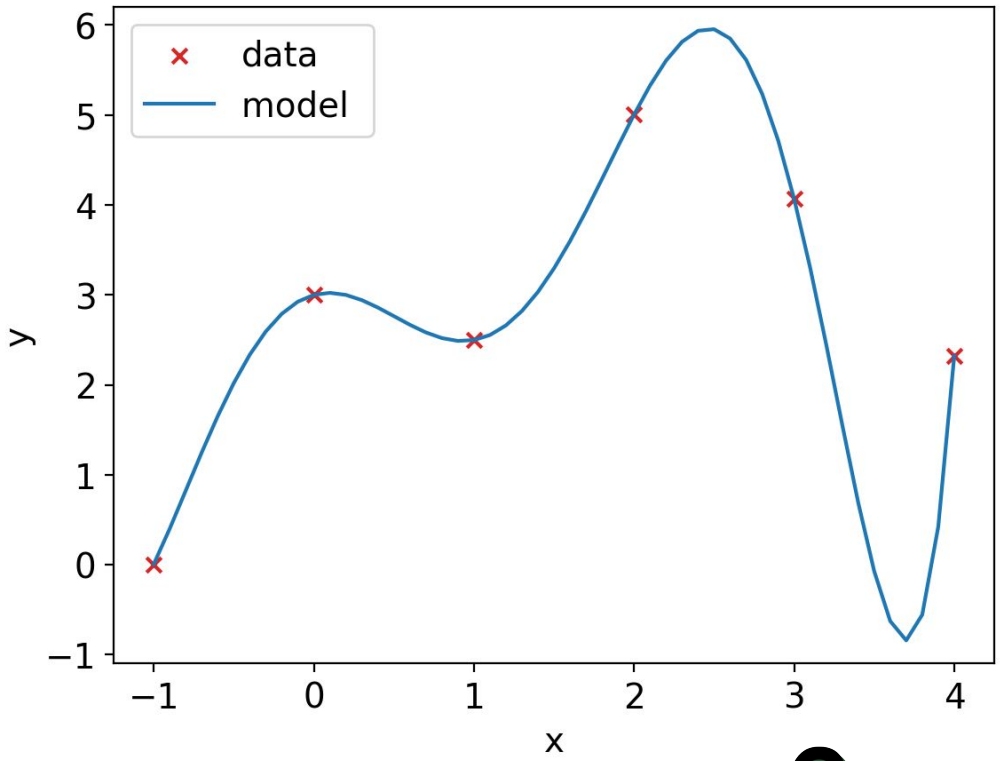
?

Models



⇒ *Polynomial model* ?

Models

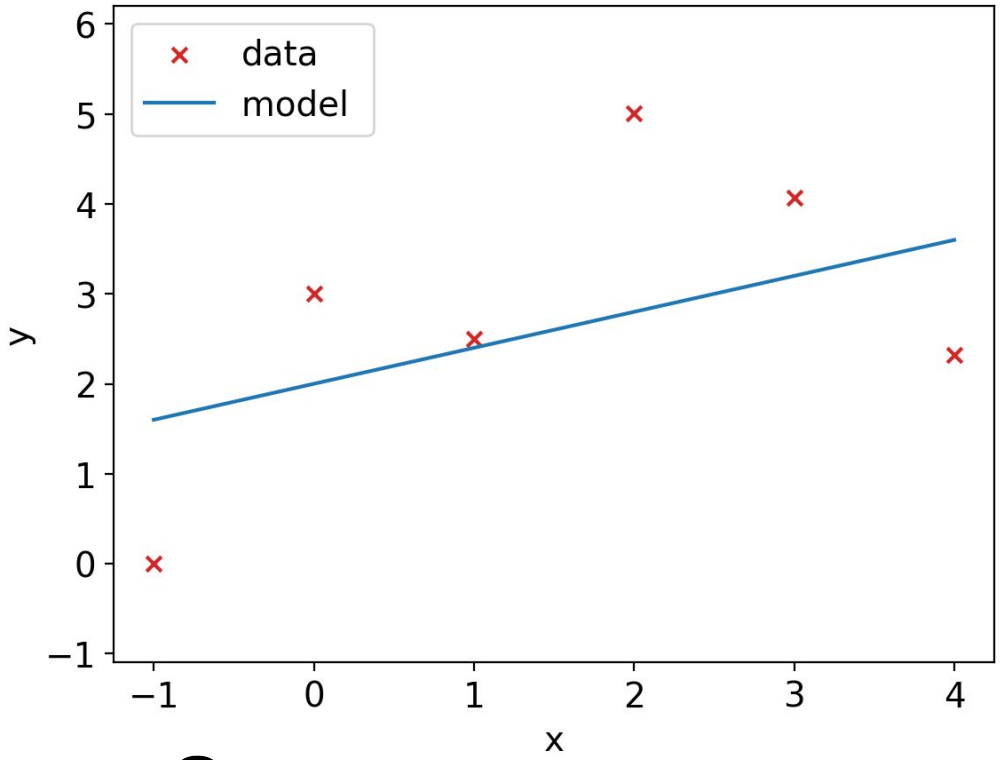


⇒ *Overfitting*

⇒ *Polynomial model*

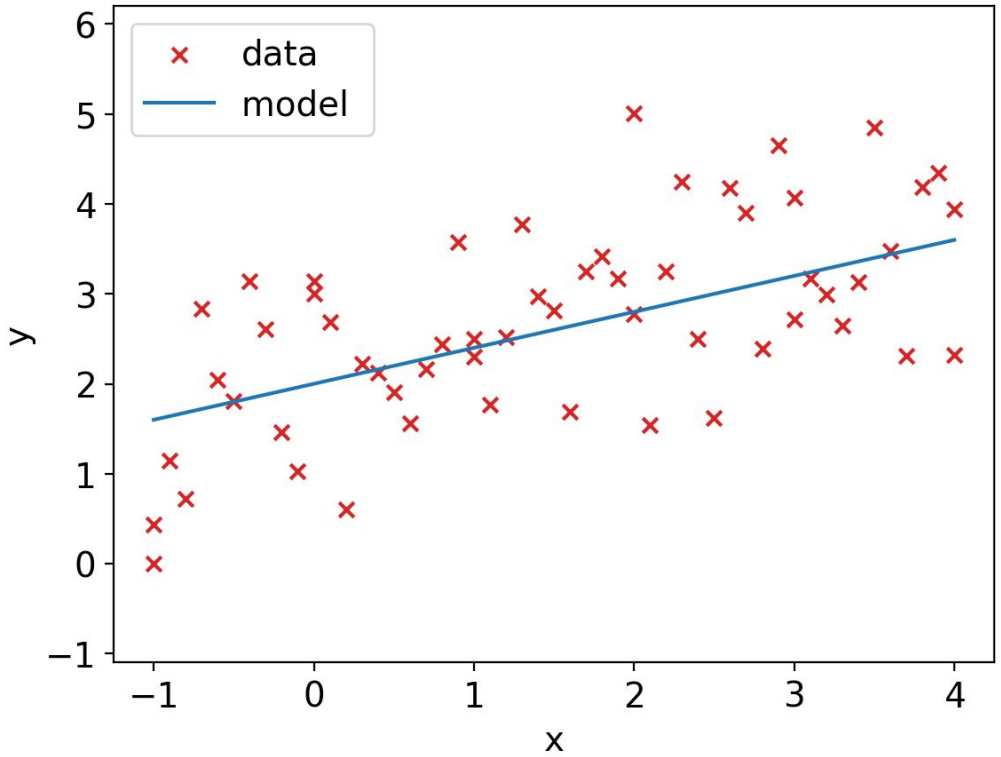
?

Models



? ⇒ *Linear model?*

Models



⇒ *Linear model*

Models

data:

CIFAR-10 dataset

Models

data:

Linear model?
Quadratic model?

CIFAR-10 dataset

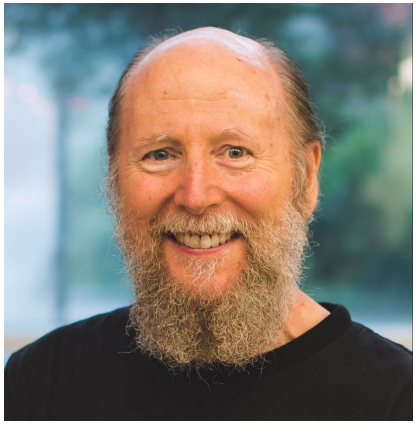
Python

linear_model.py

How to Choose a Model?

How to Choose a Model?

The Bitter Lesson (2019)



(Distilled version)

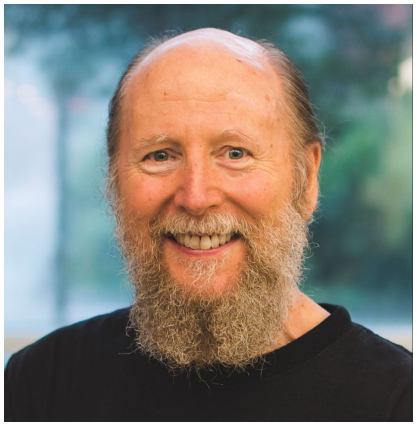
General methods that leverage computation and learning ultimately prove more effective than approaches relying on human knowledge and domain-specific expertise.

— Rich Sutton, ACM A. M. Turing Award 2024 recipient

http://www.incompleteideas.net/IncIdeas/BitterLesson.html

How to Choose a Model?

The Bitter Lesson (2019)



(Distilled version)

General methods that leverage computation and learning ultimately prove more effective than approaches relying on human knowledge and domain-specific expertise.

— Rich Sutton, ACM A. M. Turing Award 2024 recipient

http://www.incompleteideas.net/IncIdeas/BitterLesson.html

(⇒ we should not assume models (e.g., linear or not), we should just have a general learning method leveraging computation, e.g., NNs + GD)

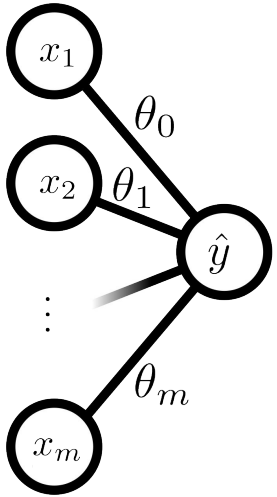
Neural Networks

$$\hat{y} = \boldsymbol{x}\,\boldsymbol{\theta} + \boldsymbol{b} = \sum_{i=0}^{m} x_i \theta_i + b$$

$$\boldsymbol{x} = \left[x_1, x_2, \ldots, x_m\right]$$

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_m \end{bmatrix}$$

$$\boldsymbol{b} \in \mathbb{R}$$



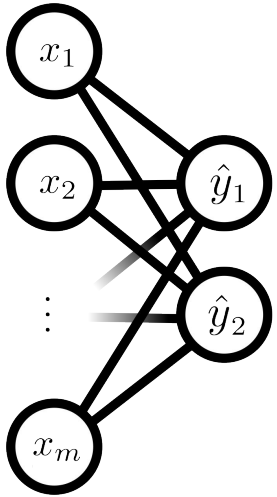
Neural Networks

$$\hat{y} = \boldsymbol{x}\,\boldsymbol{\theta} + \boldsymbol{b} = \begin{bmatrix} \sum_{i=0}^{m} x_i \theta_{1i} + b_1 & \sum_{i=0}^{m} x_i \theta_{2i} + b_2 \end{bmatrix}$$

$$\boldsymbol{x} = \begin{bmatrix} x_1, x_2, \ldots, x_m \end{bmatrix}$$

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_{11} & \theta_{21} \\ \vdots & \vdots \\ \theta_{1m} & \theta_{2m} \end{bmatrix}$$

$$\boldsymbol{b} = \begin{bmatrix} b_1 & b_2 \end{bmatrix}$$



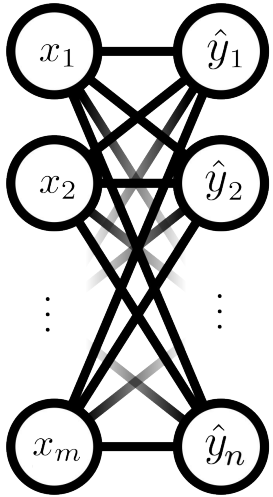
Neural Networks

$$\hat{y} = \boldsymbol{x}\,\boldsymbol{\theta} + \boldsymbol{b} = \left[\sum_{i=0}^{m} x_i\theta_{1i} + b_1 \quad \dots \quad \sum_{i=0}^{m} x_i\theta_{ni} + b_n\right]$$

$$\boldsymbol{x} = \left[x_1, x_2, \dots, x_m\right]$$

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_{11} & \theta_{21} & \cdots & \theta_{n1} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{1m} & \theta_{2m} & \cdots & \theta_{nm} \end{bmatrix}$$

$$\boldsymbol{b} = \begin{bmatrix} b_1 & b_2 & \dots & b_n \end{bmatrix}$$



(n → m)

Neural Networks

?

$$\hat{y} = \sigma(\boldsymbol{x}\,\boldsymbol{\theta}^{(1)} + \boldsymbol{b}^{(1)})\,\boldsymbol{\theta}^{(2)} + \boldsymbol{b}^{(2)}$$

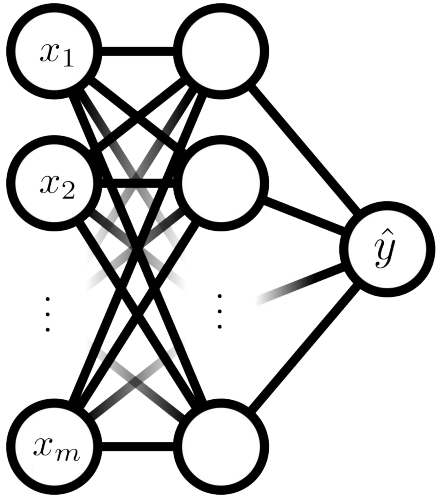
(nonlinearities)

$$\boldsymbol{x} = \left[x_1, x_2, \ldots, x_m\right]$$

$$\boldsymbol{\theta} = \left(\boldsymbol{\theta}^{(1)}, \boldsymbol{b}^{(1)}, \boldsymbol{\theta}^{(2)}, \boldsymbol{b}^{(2)}\right)$$

$\boldsymbol{\theta^{(1)}}$ – $M \times N$ $\quad$ $\boldsymbol{b}^{(1)}$ – $N \times 1$

$\boldsymbol{\theta^{(2)}}$ – $N \times 1$ $\quad$ $\boldsymbol{b}^{(2)} \in \mathbb{R}$



Neural Networks

$$\hat{y} = \sigma(\ldots\sigma(\boldsymbol{x}\,\boldsymbol{\theta}^{(1)} + \boldsymbol{b}^{(1)})\,\boldsymbol{\theta}^{(2)} + \boldsymbol{b}^{(2)}\ldots)\,\boldsymbol{\theta}^{(L)} + \boldsymbol{b}^{(L)}$$

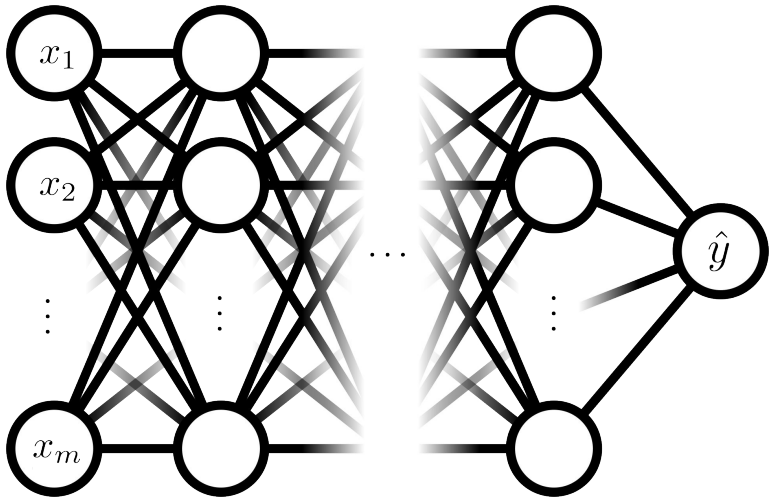
$$\boldsymbol{x} = \left[x_1, x_2, \ldots, x_m\right]$$

$$\boldsymbol{\theta} = \left(\boldsymbol{\theta}^{(1)}, \boldsymbol{b}^{(1)}, \ldots, \boldsymbol{\theta}^{(L)}, \boldsymbol{b}^{(L)}\right)$$

$\boldsymbol{\theta^{(1)}}$ – $M \times N$ $\quad$ $\boldsymbol{b}^{(1)}$ – $N \times 1$

$\boldsymbol{\theta^{(2)}}$ – $N \times 1$ $\quad$ $\boldsymbol{b}^{(2)} \in \mathbb{R}$

...



Neural Networks

$$\hat{y} = \sigma(\dots\sigma(\boldsymbol{x}\,\boldsymbol{\theta}^{(1)} + \boldsymbol{b}^{(1)})\,\boldsymbol{\theta}^{(2)} + \boldsymbol{b}^{(2)}\dots)\,\boldsymbol{\theta}^{(L)} + \boldsymbol{b}^{(L)}$$

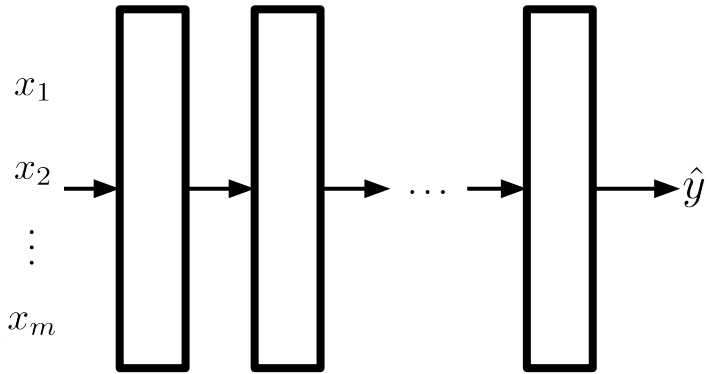
$$\boldsymbol{x} = \left[x_1, x_2, \dots, x_m\right]$$

$$\boldsymbol{\theta} = \left(\boldsymbol{\theta}^{(1)}, \boldsymbol{b}^{(1)}, \dots, \boldsymbol{\theta}^{(L)}, \boldsymbol{b}^{(L)}\right)$$

$\boldsymbol{\theta^{(1)}}$ – $M \times N$ $\quad$ $\boldsymbol{b}^{(1)}$ – $N \times 1$

$\boldsymbol{\theta^{(2)}}$ – $N \times 1$ $\quad$ $\boldsymbol{b}^{(2)} \in \mathbb{R}$

...



Backpropagation Example

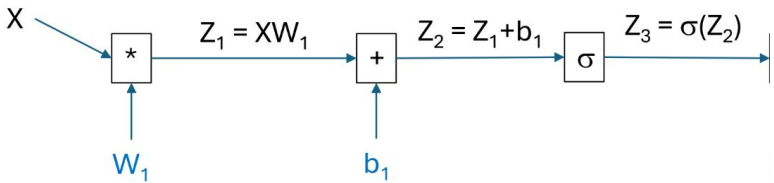
$$Y^p = \sigma(XW_1 + b_1)W_2 + b_2$$

Backpropagation Example

$$Y^p = \sigma(XW_1 + b_1)W_2 + b_2 \qquad L(Y^p, Y) = \frac{1}{2}(Y^p - Y)^2$$

Backpropagation Example

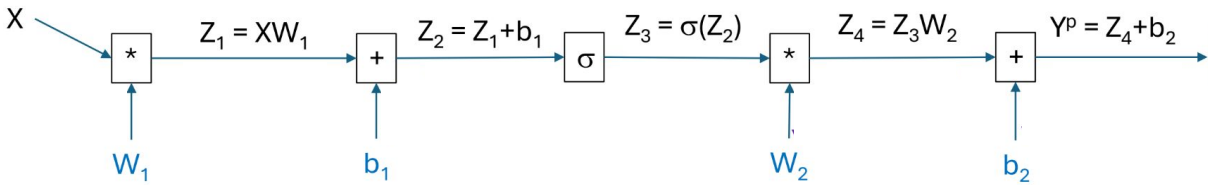
$$Y^p = \sigma(XW_1 + b_1)W_2 + b_2 \qquad L(Y^p, Y) = \frac{1}{2}(Y^p - Y)^2$$



Backpropagation Example

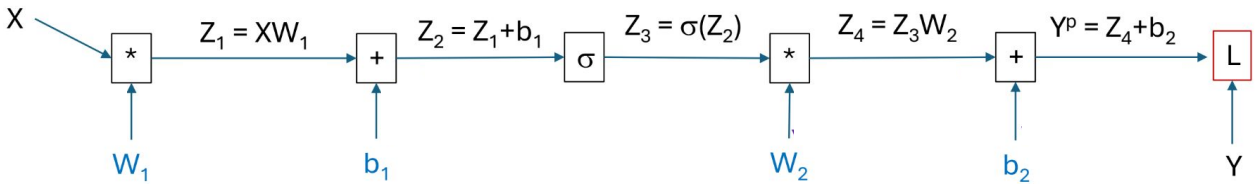
$$Y^p = \sigma(XW_1 + b_1)W_2 + b_2$$

$$L(Y^p, Y) = \frac{1}{2}(Y^p - Y)^2$$



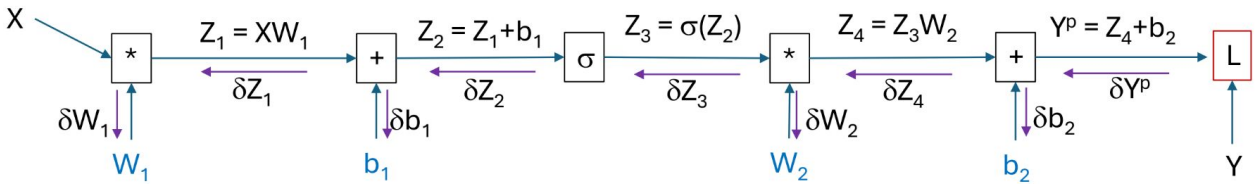
Backpropagation Example

$$Y^p = \sigma(XW_1 + b_1)W_2 + b_2 \qquad L(Y^p, Y) = \frac{1}{2}(Y^p - Y)^2$$



Backpropagation Example

$$Y^p = \sigma(XW_1 + b_1)W_2 + b_2 \qquad L(Y^p, Y) = \frac{1}{2}(Y^p - Y)^2$$

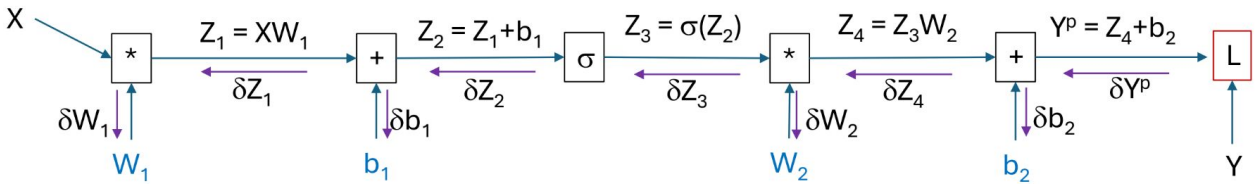


Backpropagation Implementation

See Andrej Karpathy's micrograd and backpropagation lecture material

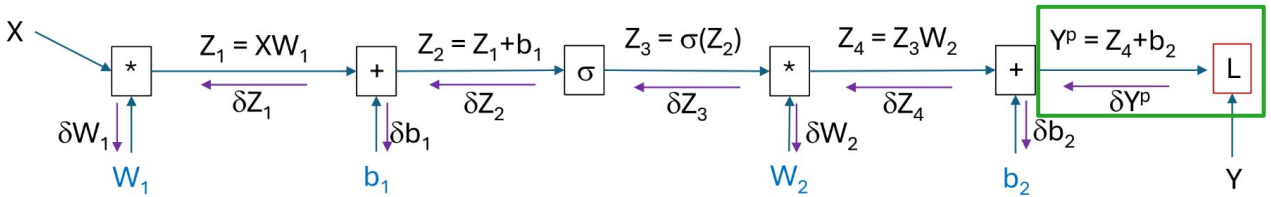
Backpropagation Example

$$Y^p = \sigma(XW_1 + b_1)W_2 + b_2 \qquad L(Y^p, Y) = \frac{1}{2}(Y^p - Y)^2$$



Backpropagation Example

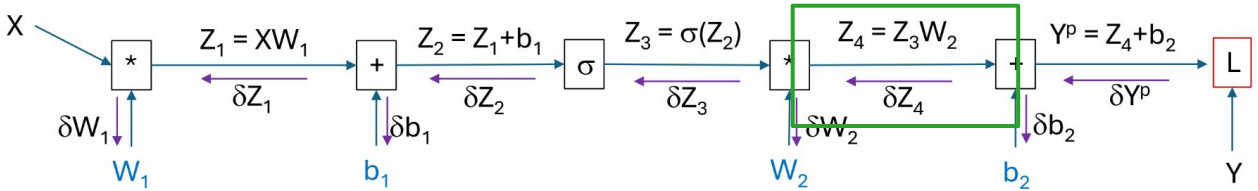
$$Y^p = \sigma(XW_1 + b_1)W_2 + b_2 \qquad L(Y^p, Y) = \frac{1}{2}(Y^p - Y)^2$$



1 $$\delta Y^p \equiv \frac{\partial L}{\partial Y^p} = \frac{\partial}{\partial Y^p}\left[\frac{1}{2}||Y^p - Y||^2\right] = y^p - y$$

Backpropagation Example

$$Y^p = \sigma(XW_1 + b_1)W_2 + b_2 \qquad L(Y^p, Y) = \frac{1}{2}(Y^p - Y)^2$$



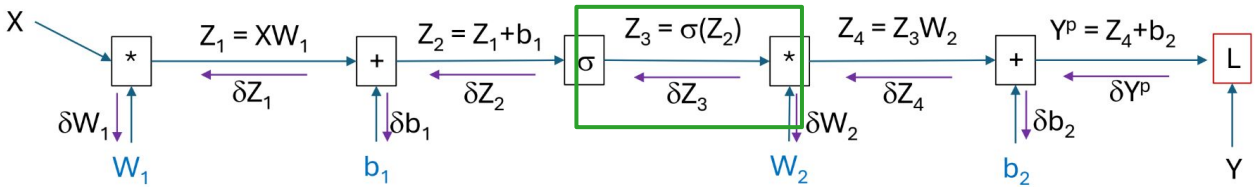
1 $$\delta Y^p \equiv \frac{\partial L}{\partial Y^p} = \frac{\partial}{\partial Y^p}\left[\frac{1}{2}||Y^p - Y||^2\right] = y^p - y$$

2 $$\delta Z_4 \equiv \frac{\partial L}{\partial Z_4} = \frac{\partial Y^p}{\partial Z_4}\frac{\partial L}{\partial Y^p} = \delta Y^p$$

Because $Y^p = Z_4 + b_2, \frac{\partial Y^p}{\partial Z_4} = 1$

Backpropagation Example

$$Y^p = \sigma(XW_1 + b_1)W_2 + b_2 \qquad L(Y^p, Y) = \frac{1}{2}(Y^p - Y)^2$$



1 $$\delta Y^p \equiv \frac{\partial L}{\partial Y^p} = \frac{\partial}{\partial Y^p}\left[\frac{1}{2}||Y^p - Y||^2\right] = y^p - y$$

2 $$\delta Z_4 \equiv \frac{\partial L}{\partial Z_4} = \frac{\partial Y^p}{\partial Z_4}\frac{\partial L}{\partial Y^p} = \delta Y^p$$

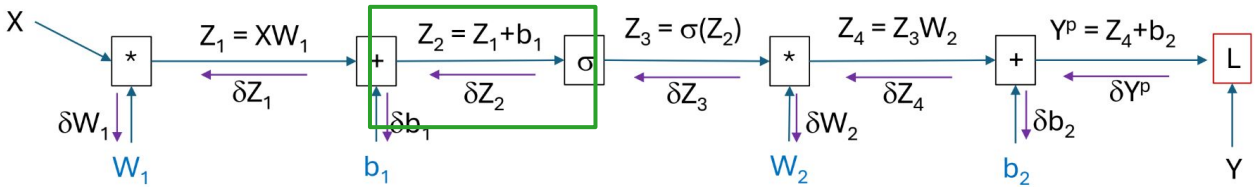
Because $Y^p = Z_4 + b_2, \frac{\partial Y^p}{\partial Z_4} = 1$

3 $$\delta Z_3 \equiv \frac{\partial L}{\partial Z_3} = \frac{\partial Z_4}{\partial Z_3}\frac{\partial L}{\partial Z_4} = \delta Z_4 W_2^T$$

Because $Z_4 = Z_3 W_2, \frac{\partial Z_4}{\partial Z_3} = W_2^T$

Backpropagation Example

$$Y^p = \sigma(XW_1 + b_1)W_2 + b_2 \qquad L(Y^p, Y) = \frac{1}{2}(Y^p - Y)^2$$



1 $$\delta Y^p \equiv \frac{\partial L}{\partial Y^p} = \frac{\partial}{\partial Y^p}\left[\frac{1}{2}||Y^p - Y||^2\right] = y^p - y$$

2 $$\delta Z_4 \equiv \frac{\partial L}{\partial Z_4} = \frac{\partial Y^p}{\partial Z_4}\frac{\partial L}{\partial Y^p} = \delta Y^p$$

Because $Y^p = Z_4 + b_2, \frac{\partial Y^p}{\partial Z_4} = 1$

3 $$\delta Z_3 \equiv \frac{\partial L}{\partial Z_3} = \frac{\partial Z_4}{\partial Z_3}\frac{\partial L}{\partial Z_4} = \delta Z_4 W_2^T$$

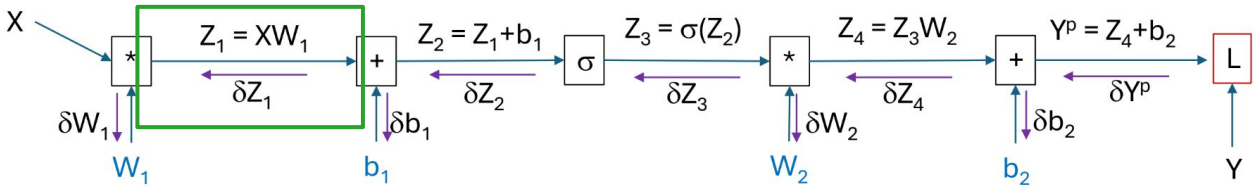
Because $Z_4 = Z_3 W_2, \frac{\partial Z_4}{\partial Z_3} = W_2^T$

4 $$\delta Z_2 \equiv \frac{\partial L}{\partial Z_2} = \frac{\partial Z_3}{\partial Z_2}\frac{\partial L}{\partial Z_3} = \sigma'(Z_2).\delta Z_3$$

Because $Z_3 = \sigma(Z_2), \frac{\partial Z_3}{\partial Z_2} = \sigma'(Z_2)$

Backpropagation Example

$$Y^p = \sigma(XW_1 + b_1)W_2 + b_2 \qquad L(Y^p, Y) = \frac{1}{2}(Y^p - Y)^2$$



1 $$\delta Y^p \equiv \frac{\partial L}{\partial Y^p} = \frac{\partial}{\partial Y^p}\left[\frac{1}{2}||Y^p - Y||^2\right] = y^p - y$$

2 $$\delta Z_4 \equiv \frac{\partial L}{\partial Z_4} = \frac{\partial Y^p}{\partial Z_4}\frac{\partial L}{\partial Y^p} = \delta Y^p$$

Because $Y^p = Z_4 + b_2, \frac{\partial Y^p}{\partial Z_4} = 1$

3 $$\delta Z_3 \equiv \frac{\partial L}{\partial Z_3} = \frac{\partial Z_4}{\partial Z_3}\frac{\partial L}{\partial Z_4} = \delta Z_4 W_2^T$$

Because $Z_4 = Z_3 W_2, \frac{\partial Z_4}{\partial Z_3} = W_2^T$

4 $$\delta Z_2 \equiv \frac{\partial L}{\partial Z_2} = \frac{\partial Z_3}{\partial Z_2}\frac{\partial L}{\partial Z_3} = \sigma'(Z_2).\delta Z_3$$

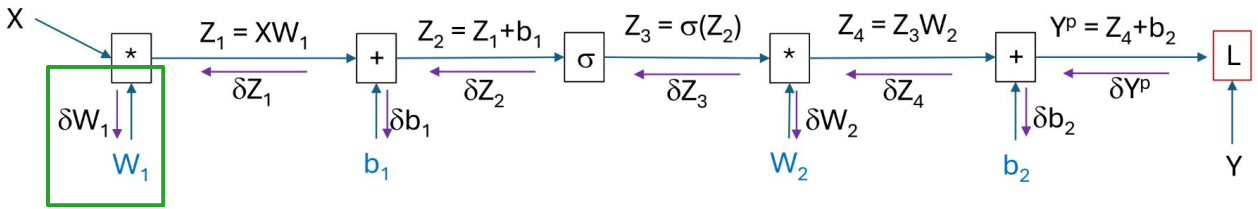
Because $Z_3 = \sigma(Z_2), \frac{\partial Z_3}{\partial Z_2} = \sigma'(Z_2)$

5 $$\delta Z_1 \equiv \frac{\partial L}{\partial Z_1} = \frac{\partial Z_2}{\partial Z_1}\frac{\partial L}{\partial Z_2} = \delta Z_2$$

Because $Z_2 = Z_1 + b_1, \frac{\partial Z_2}{\partial Z_1} = 1$

Backpropagation Example

$$Y^p = \sigma(XW_1 + b_1)W_2 + b_2 \qquad L(Y^p, Y) = \frac{1}{2}(Y^p - Y)^2$$

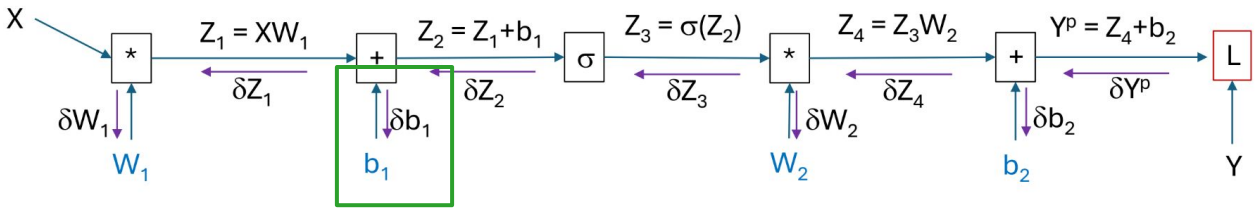


1 $\delta W_1 \equiv \frac{\partial L}{\partial W_1} = \frac{\partial Z_1}{\partial W_1}\frac{\partial L}{\partial Z_1} = X^T \delta Z_1$

Because $Z_1 = XW_1, \frac{\partial Z_1}{\partial W_1} = X^T$

Backpropagation Example

$$Y^p = \sigma(XW_1 + b_1)W_2 + b_2 \qquad L(Y^p, Y) = \frac{1}{2}(Y^p - Y)^2$$



1 $$\delta W_1 \equiv \frac{\partial L}{\partial W_1} = \frac{\partial Z_1}{\partial W_1}\frac{\partial L}{\partial Z_1} = X^T \delta Z_1$$

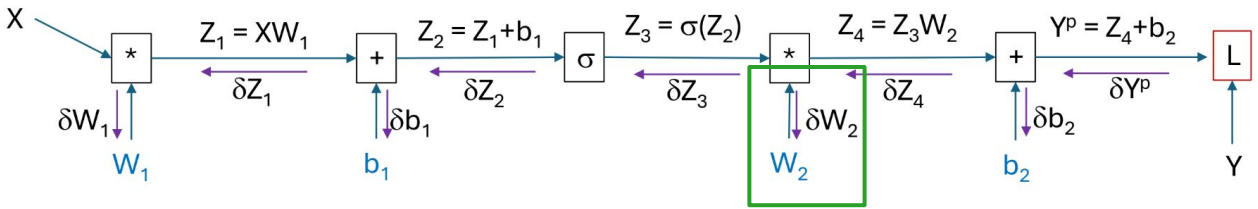
Because $Z_1 = XW_1, \frac{\partial Z_1}{\partial W_1} = X^T$

2 $$\delta b_1 \equiv \frac{\partial L}{\partial b_1} = \frac{\partial Z_2}{\partial b_1}\frac{\partial L}{\partial Z_2} = \sum_k (\delta Z_2)_{k,:}$$

Because $Z_2 = Z_1 + b_1, \frac{\partial Z_2}{\partial b_1} = [1, \dots, 1]$

Backpropagation Example

$$Y^p = \sigma(XW_1 + b_1)W_2 + b_2 \qquad L(Y^p, Y) = \frac{1}{2}(Y^p - Y)^2$$



1 $\delta W_1 \equiv \frac{\partial L}{\partial W_1} = \frac{\partial Z_1}{\partial W_1}\frac{\partial L}{\partial Z_1} = X^T \delta Z_1$

Because $Z_1 = XW_1, \frac{\partial Z_1}{\partial W_1} = X^T$

2 $\delta b_1 \equiv \frac{\partial L}{\partial b_1} = \frac{\partial Z_2}{\partial b_1}\frac{\partial L}{\partial Z_2} = \sum_k (\delta Z_2)_{k,:}$

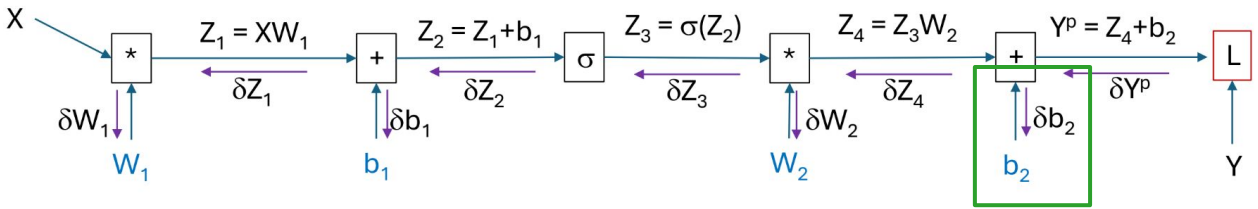
Because $Z_2 = Z_1 + b_1, \frac{\partial Z_2}{\partial b_1} = [1, \dots, 1]$

3 $\delta W_2 \equiv \frac{\partial L}{\partial W_2} = \frac{\partial Z_4}{\partial W_2}\frac{\partial L}{\partial Z_4} = Z_3^T \delta Z_4$

Because $Z_4 = Z_3 W_2, \frac{\partial Z_4}{\partial W_2} = Z_3^T$

Backpropagation Example

$$Y^p = \sigma(XW_1 + b_1)W_2 + b_2 \qquad L(Y^p, Y) = \frac{1}{2}(Y^p - Y)^2$$



1 $\delta W_1 \equiv \frac{\partial L}{\partial W_1} = \frac{\partial Z_1}{\partial W_1}\frac{\partial L}{\partial Z_1} = X^T \delta Z_1$

Because $Z_1 = XW_1, \frac{\partial Z_1}{\partial W_1} = X^T$

2 $\delta b_1 \equiv \frac{\partial L}{\partial b_1} = \frac{\partial Z_2}{\partial b_1}\frac{\partial L}{\partial Z_2} = \sum_k (\delta Z_2)_{k,:}$

Because $Z_2 = Z_1 + b_1, \frac{\partial Z_2}{\partial b_1} = [1, \dots, 1]$

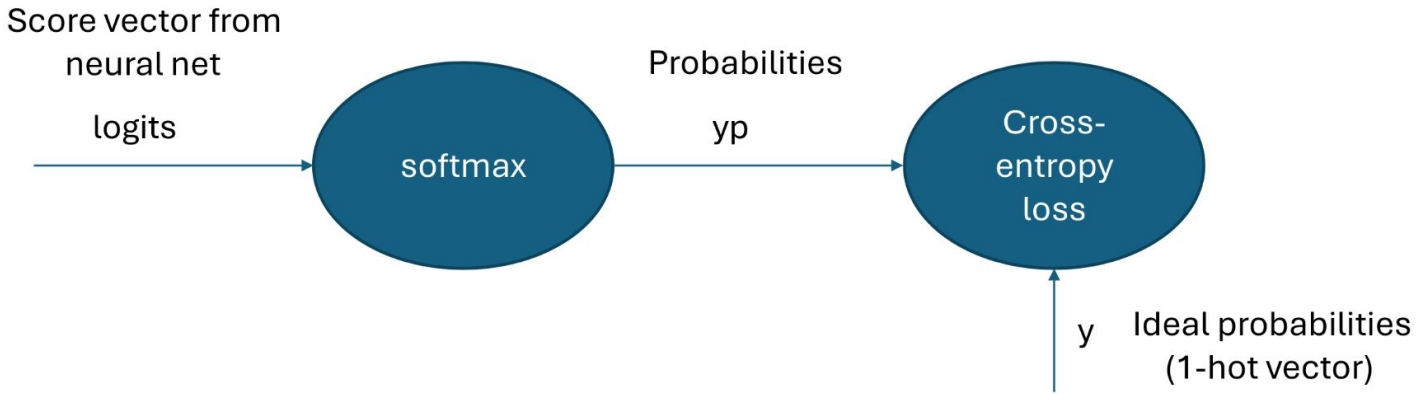
3 $\delta W_2 \equiv \frac{\partial L}{\partial W_2} = \frac{\partial Z_4}{\partial W_2}\frac{\partial L}{\partial Z_4} = Z_3^T \delta Z_4$

Because $Z_4 = Z_3 W_2, \frac{\partial Z_4}{\partial W_2} = Z_3^T$

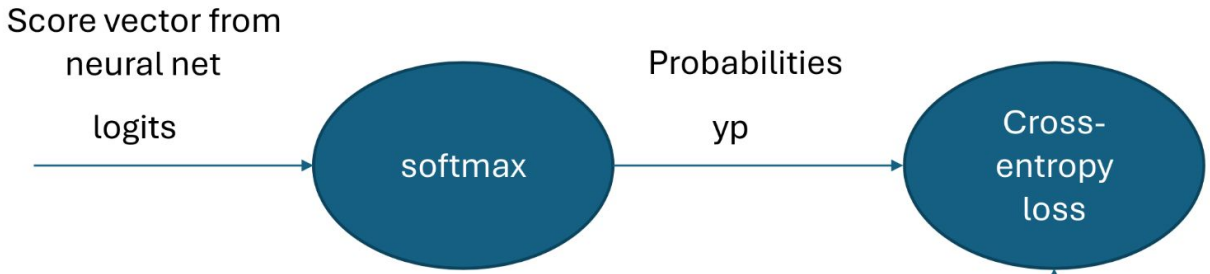
4 $\delta b_2 \equiv \frac{\partial L}{\partial b_2} = \frac{\partial Y^p}{\partial b_2}\frac{\partial L}{\partial Y^p} = \sum_k (\delta Y^p)_{k,:}$

Because $Y^p = Z_4 + b_2, \frac{\partial Y^p}{\partial b_2} = [1, \dots, 1]$

Backpropagation (Softmax & CE Loss)

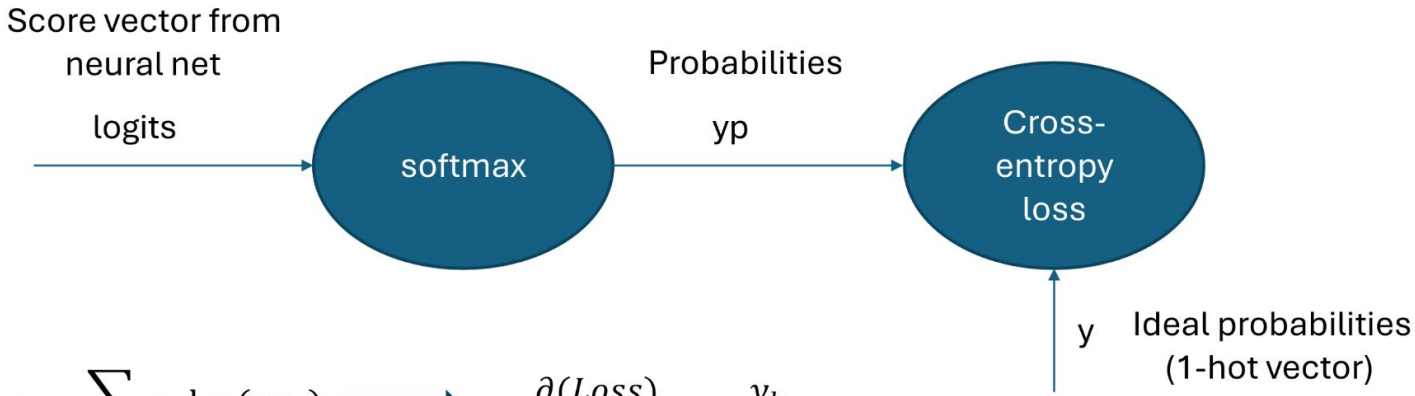


Backpropagation (Softmax & CE Loss)



$$Loss = -\sum_k y_k \log(yp_k) \Rightarrow \frac{\partial(Loss)}{\partial(yp)_k} = -\frac{y_k}{yp_k}$$

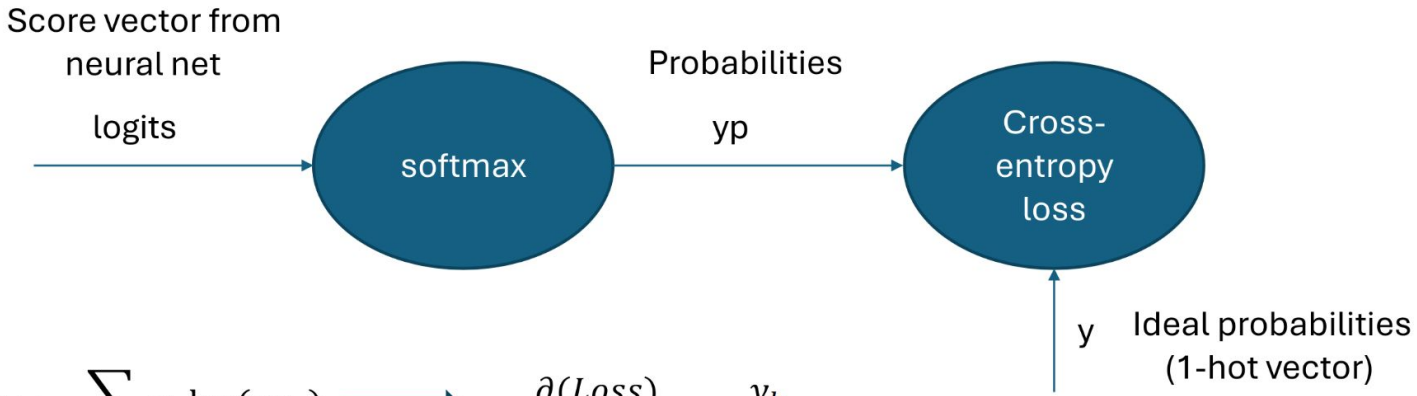
Backpropagation (Softmax & CE Loss)



$$Loss = -\sum_k y_k \log(yp_k) \Rightarrow \frac{\partial(Loss)}{\partial(yp)_k} = -\frac{y_k}{yp_k}$$

$$yp_i = \frac{\exp(logits_i)}{\sum_k \exp(logits_k)} \Rightarrow \frac{\partial(yp)_k}{\partial(logits)_i} = \begin{cases} yp_i(1 - yp_i), & \text{if } i = k, \\ -yp_i yp_k, & \text{otherwise.} \end{cases}$$

Backpropagation (Softmax & CE Loss)



$$Loss = -\sum_k y_k \log(yp_k) \quad \Rightarrow \quad \frac{\partial(Loss)}{\partial(yp)_k} = -\frac{y_k}{yp_k}$$

$$yp_i = \frac{\exp(logits_i)}{\sum_k \exp(logits_k)} \quad \Rightarrow \quad \frac{\partial(yp)_k}{\partial(logits)_i} = \begin{cases} yp_i(1 - yp_i), & \text{if } i = k, \\ -yp_i yp_k, & \text{otherwise.} \end{cases}$$

Using the above two results in the chain rule, $\delta(logits)_i \equiv \frac{\partial(Loss)}{\partial(logits)_i} = \sum_k \frac{\partial(yp)_k}{\partial(logits)_i} \frac{\partial(Loss)}{\partial(yp)_k} = yp_i - y_i$

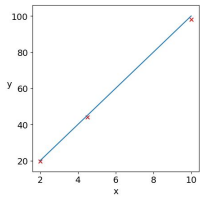
Backpropagation (More Explanations)

See the Back Propagation Practice folder

Recap

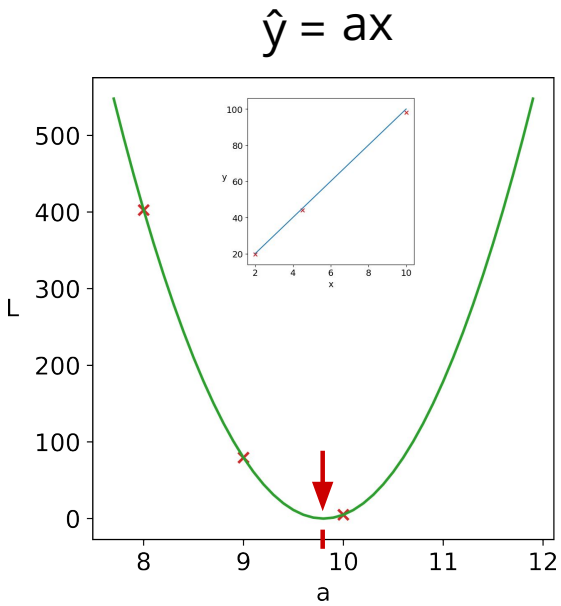
Recap

$$\hat{y} = ax$$



Recap

$$\mathcal{L} = \sum_i \left(y_i - \hat{y}(x_i) \right)^2 \qquad \textbf{L2 Loss}$$



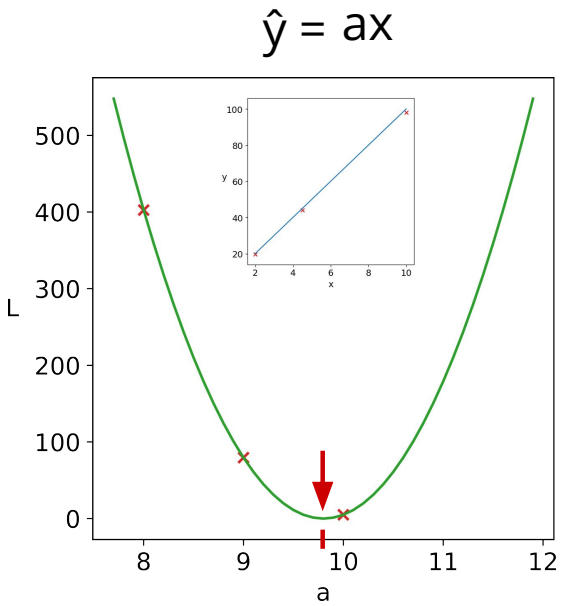
Recap

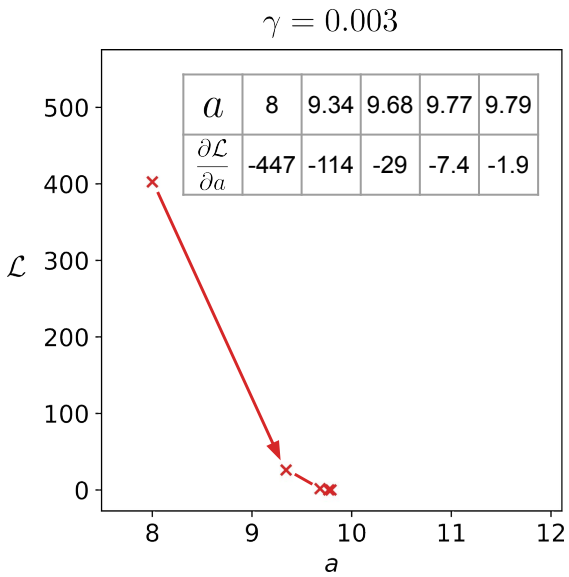
$$\mathcal{L} = \sum_i \Big(y_i - \hat{y}(x_i)\Big)^2$$ **L2 Loss**

repeat:

$a := a - \gamma \frac{\partial \mathcal{L}}{\partial a}$

until minimum is reached





a	8	9.34	9.68	9.77	9.79
$\frac{\partial \mathcal{L}}{\partial a}$	-447	-114	-29	-7.4	-1.9

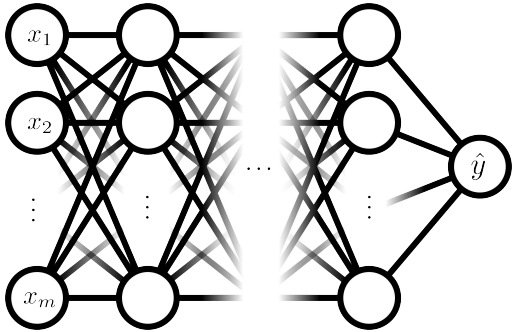
Recap

$$\mathcal{L} = \sum_i \Big(y_i - \hat{y}(x_i)\Big)^2 \qquad \textbf{L2 Loss}$$

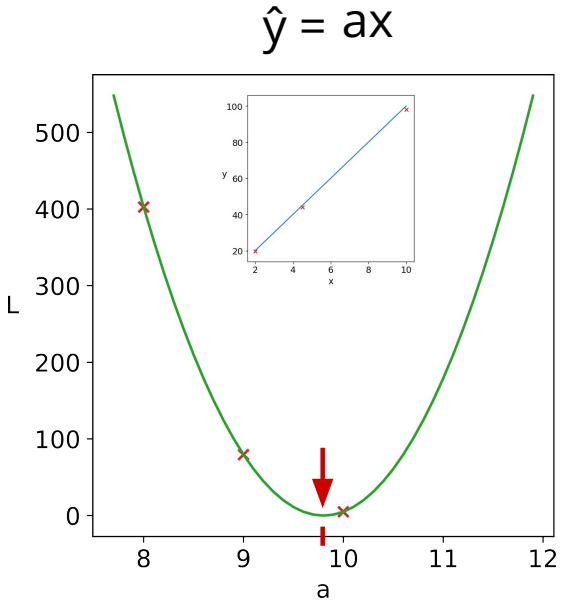
repeat:

$a := a - \gamma \frac{\partial \mathcal{L}}{\partial a}$

until minimum is reached

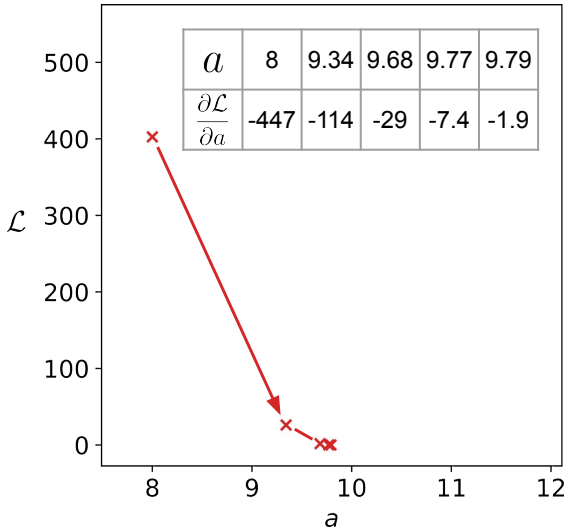


$$\hat{y} = \sigma(\dots\sigma(\boldsymbol{x}\,\boldsymbol{\theta}^{(1)} + \boldsymbol{b}^{(1)})\,\boldsymbol{\theta}^{(2)} + \boldsymbol{b}^{(2)}\dots)\,\boldsymbol{\theta}^{(L)} + \boldsymbol{b}^{(L)}$$



$\gamma = 0.003$

a	8	9.34	9.68	9.77	9.79
$\frac{\partial \mathcal{L}}{\partial a}$	-447	-114	-29	-7.4	-1.9



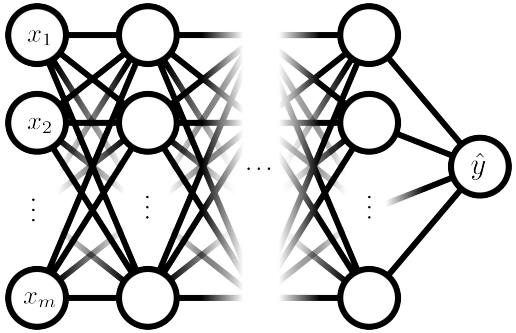
Recap

$$\mathcal{L} = \sum_i \Big(y_i - \hat{y}(x_i)\Big)^2 \qquad \textbf{L2 Loss}$$

repeat:

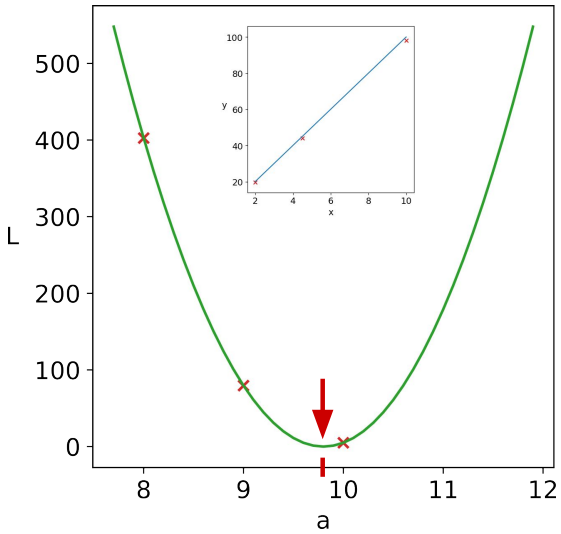
$a := a - \gamma \frac{\partial \mathcal{L}}{\partial a}$

until minimum is reached



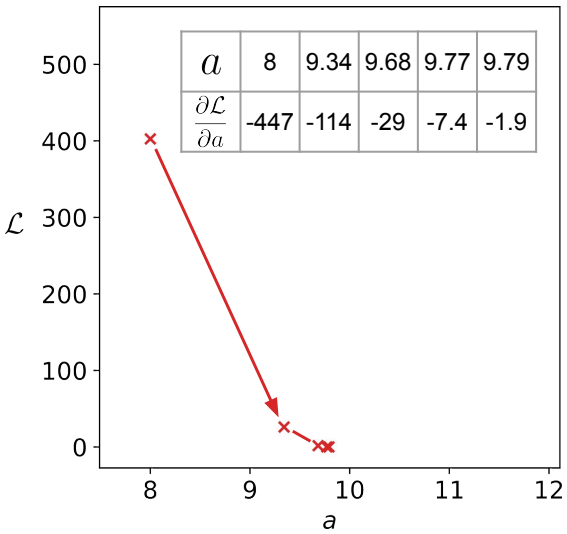
$$\hat{y} = \sigma(\dots\sigma(\boldsymbol{x}\,\boldsymbol{\theta}^{(1)} + \boldsymbol{b}^{(1)})\,\boldsymbol{\theta}^{(2)} + \boldsymbol{b}^{(2)}\dots)\,\boldsymbol{\theta}^{(L)} + \boldsymbol{b}^{(L)}$$

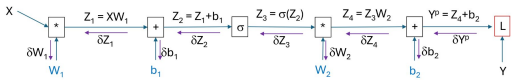
$\hat{y} = ax$



$\gamma = 0.003$

a	8	9.34	9.68	9.77	9.79
$\frac{\partial \mathcal{L}}{\partial a}$	-447	-114	-29	-7.4	-1.9





Part I: Fundamentals

Part II: Example Problems

Q1 (30). Consider an autoencoder and a classifier acting together for semi-supervised learning. Let us denote the symbols first. X is a batch of input images. W1, b1, W2, b2, W3, b3 are learnable parameters. The latent code is Z = ReLU(X * W1 + b1), where * denotes matrix-matrix multiplication and + denotes broadcast addition. The decoder output is Y = Z * W2 + b2. The classifier is attached to the latent code as C = softmax(Z * W3 + b3). The reconstruction loss for the autoencoder is the MSE loss, L2(X, Y), whereas the classification loss is cross entropy, CE(C, C_gt), where C_gt is the ground truth class probabilities with 1-hot encoding. Write the expressions for dW1, db1, dW2, db2, dW3 and db3, where these symbols mean gradient of the combined loss, L = L2(X, Y) + CE(C, C_gt), with respect to W1, b1, W2, b2, W3, and b3, respectively. Draw a computational graph for the entire model with loss nodes. Unlike a realistic image batch in a semi-supervised learning, for the sake of simplicity, the entire batch of images X here has ground truth C_gt available.

Q1 distilled:

Problem:

$Z = ReLU(XW_1 + b_1)$

$Y = ZW_2 + b_2$

$C = softmax(ZW_3 + b_3)$

$L2(X,Y) = \frac{1}{2}(X-Y)^2$

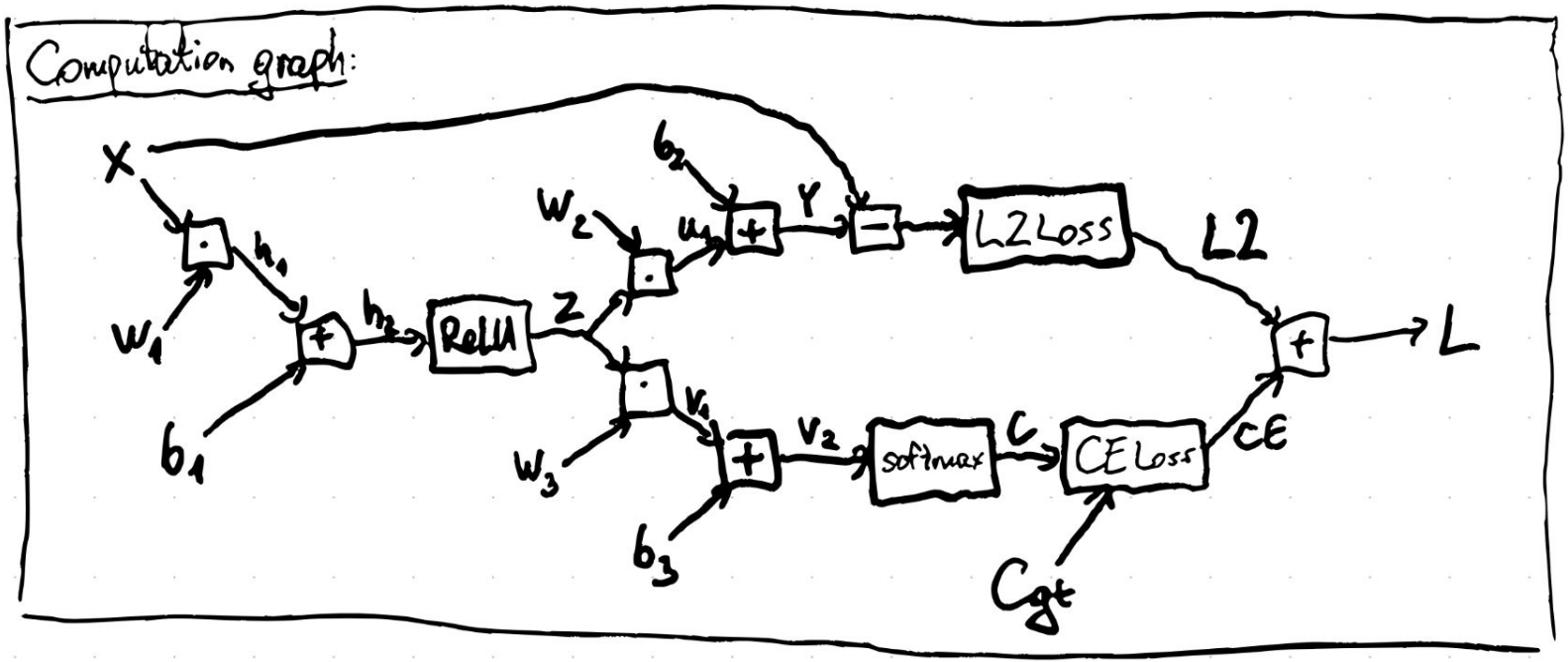
$CE(C, C_{gt})$

$L = L2(X,Y) + CE(C, C_{gt})$

Tasks: – draw a computation graph for the model with loss nodes

– find $\delta W_1, \delta b_1, \delta W_2, \delta b_2, \delta W_3, \delta b_3$, where $\delta A = \frac{\partial L}{\partial A}$ for some A

Q1 solution (1)



^ one of multiple possible

Q1 solution (2)

Answer: (also expanding δZ and h_2)

$$\delta W_1 = \begin{cases} X\left(W_2(Y-X) + W_3(C-C_{gt})\right), & \text{if } XW_1+b_1 > 0, \\ 0, & \text{otherwise} \end{cases}$$

$$\delta b_1 = \begin{cases} W_2(Y-X) + W_3(C-C_{gt}), & \text{if } XW_1+b_1 > 0, \\ 0, & \text{otherwise} \end{cases}$$

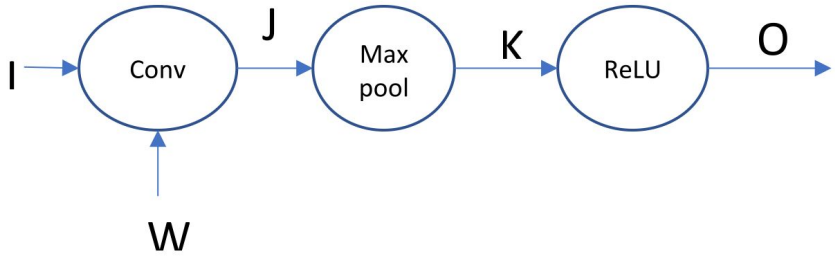
$$\delta W_2 = Z(Y-X)$$

$$\delta b_2 = Y-X$$

$$\delta W_3 = Z(C-C_{gt})$$

$$\delta b_3 = C-C_{gt}$$

Q3 (30). Compute the forward pass of the following model:



If I is the following image patch

1	3	2	4	6	4
4	8	3	1	0	2
2	1	4	3	9	1
4	7	2	3	9	2

and W is the following 3-by-3 filter matrix

1	2	1
0	0	0
-1	-2	-1

Compute J, K and O. Do not zero pad while doing the convolution. Assume stride size 1 for the convolution. Assume a 2-by-2 max pooling with stride 2. Write J, K and O below.

Q3 solution

Problem:

$$J = I \underset{\substack{\text{no padding} \\ \text{stride}=1}}{*} W \qquad I = \begin{bmatrix} 1 & 3 & 2 & 4 & 6 & 4 \\ 4 & 8 & 3 & 1 & 0 & 2 \\ 2 & 1 & 4 & 3 & 9 & 1 \\ 4 & 7 & 2 & 3 & 9 & 2 \end{bmatrix}$$

$$K = \underset{\substack{2\times 2 \\ \text{stride}=2}}{\text{max pool}}(J)$$

$$O = \text{ReLU}(K) \qquad W = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}$$

Tasks:

- Write J, K, O

Answer:

$$J = \begin{bmatrix} 1 & -1 & -3 & -2 \\ 3 & 1 & -12 & -20 \end{bmatrix} \qquad K = \begin{bmatrix} 3 & -2 \end{bmatrix} \qquad O = \begin{bmatrix} 3 & 0 \end{bmatrix}$$

sasnausk@ualberta.ca