Learning from data

How long will this delivery take?

Regression, training and honest evaluation



Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

Today’s learning goals

01 Predict a delivery time and measure its error.

02 Train regression with gradient descent.

Mohammad Sabokrou | New Uzbekistan University

03 Explain underfitting, overfitting and model complexity.

04 Use weight decay and choose settings with validation.

Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

Three common machine-learning tasks

REGRESSION	CLASSIFICATION	CLUSTERING
22.5 minutes	**Late / on time**	**Groups A, B, C**
Predict a numerical value. Training examples include known delivery times.	Predict a category. Training examples include known category labels.	Find similar cases. Training examples have no supplied group labels.

Mohammad Sabokrou | New Uzbekistan University

Regression and classification are supervised. Clustering is unsupervised.

Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

One delivery, two prediction questions

REGRESSION

How many minutes?

22.5

A numerical prediction can include decimals.

CLASSIFICATION

Will it arrive late?

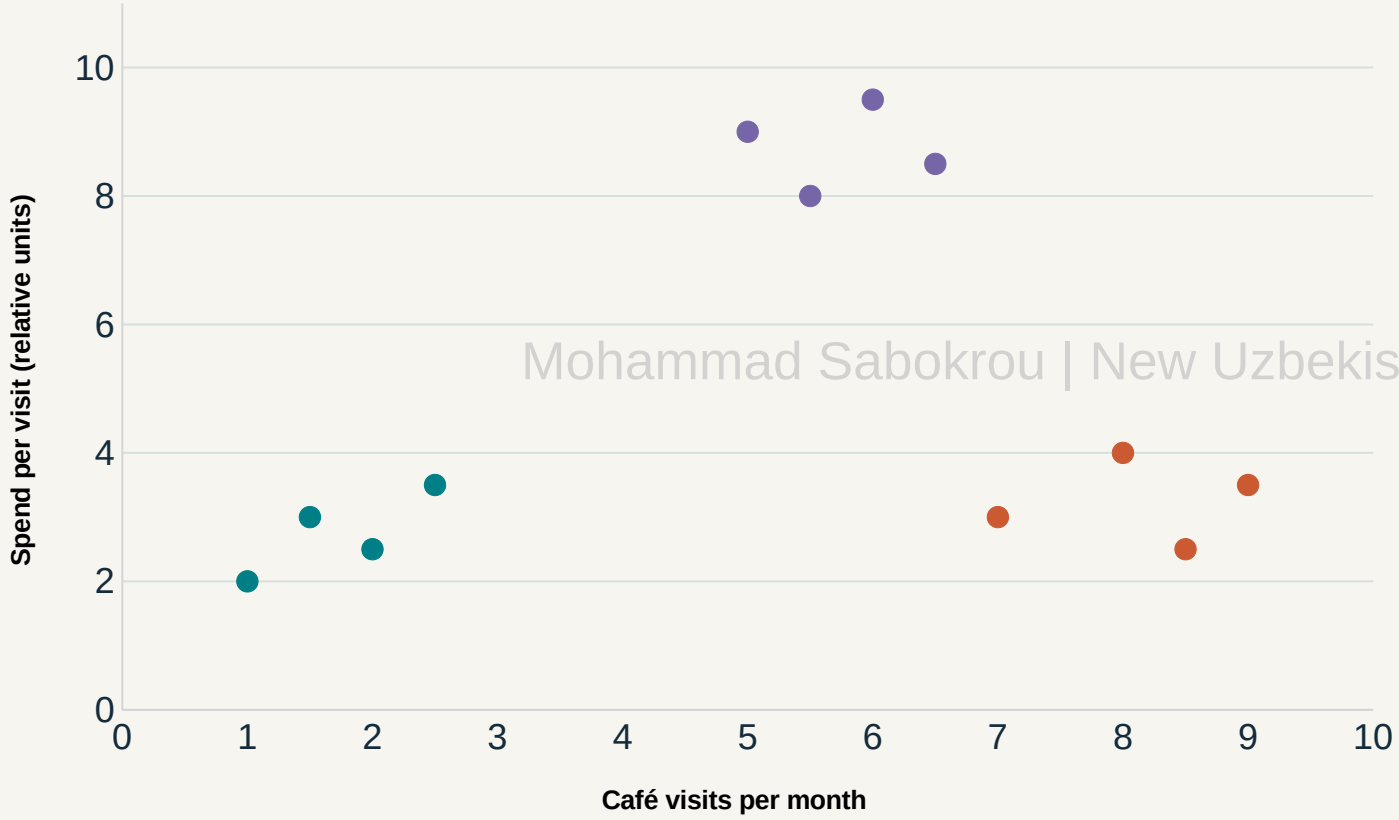
On time

A category uses a defined label. Here, “late” means more than 30 minutes.

Mohammad Sabokrou | New Uzbekistan University

Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

Clustering: discover groups without group labels



Mohammad Sabokrou | New Uzbekistan University

Each dot is a customer.

Inputs: visits and spending.

No “customer type” labels are supplied for training.

We interpret the groups after finding them.

Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

A prediction needs a clear moment

At order confirmation

Mohammad Sabokrou | New Uzbekistan University

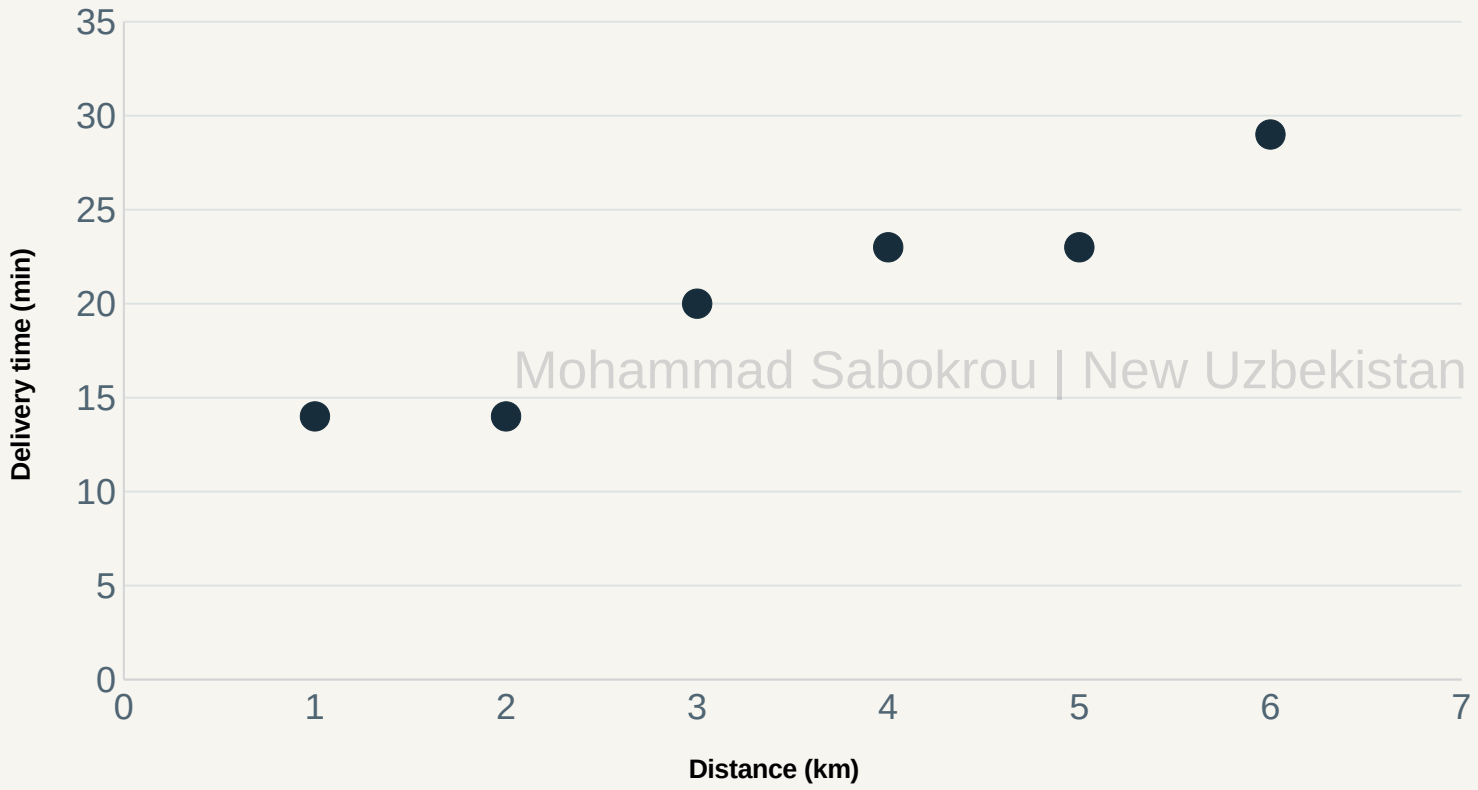
We know the distance.
We do not know the
final delivery time yet.

Target: minutes until arrival



Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

Six deliveries we can inspect



Mohammad Sabokrou | New Uzbekistan University

Each dot is one delivery.

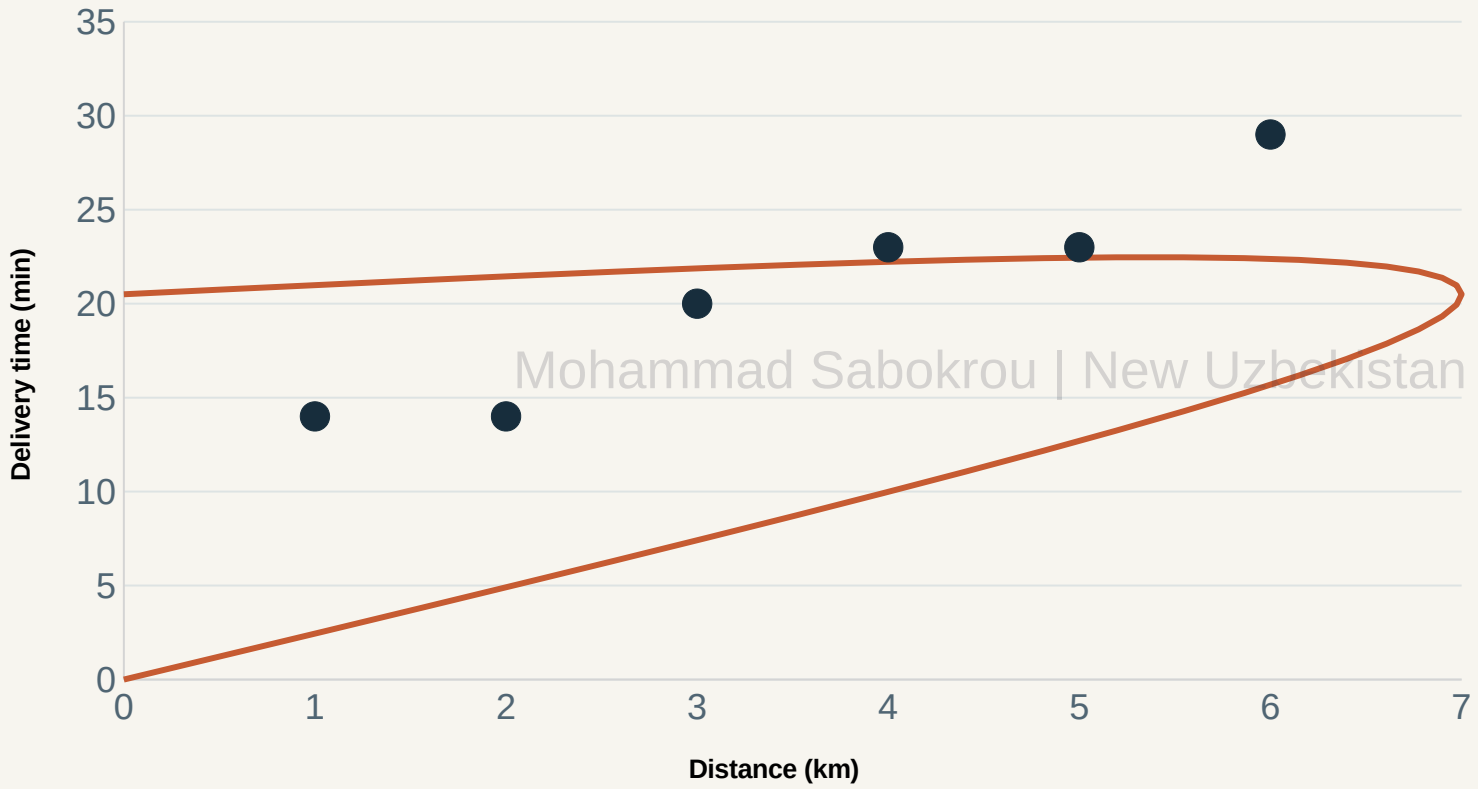
Horizontal: distance.

Vertical: total time.

Original synthetic teaching data; six rows are for arithmetic, not deployment.

Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

Before fitting: a simple baseline



Mohammad Sabokrou | New Uzbekistan University

Ignore distance.

Always predict the training average: 20.5 min.

A useful model should improve on this.

Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

The regression equation

$$\hat{y} = b + wx$$

ŷ: predicted time x: distance in kilometres

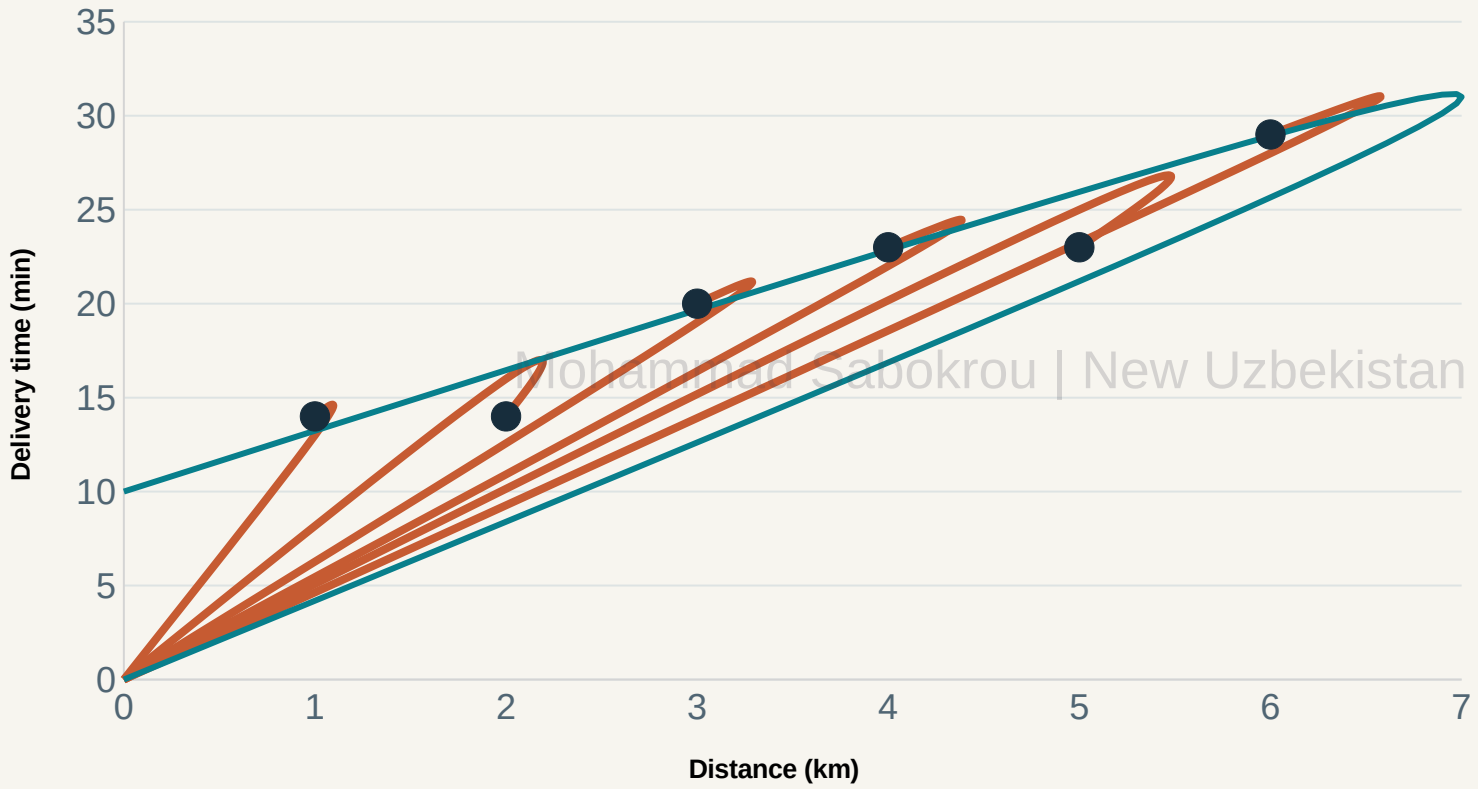
Mohammad Sabokrou | New Uzbekistan University

b: starting value w: extra minutes per kilometre

At 4 km: ŷ = 10 + 3 × 4 = 22 minutes

Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

A residual is the vertical gap



Mohammad Sabokrou | New Uzbekistan University

Residual = observed − predicted

At 5 km: 23 − 25 = −2

Negative: prediction was too high.

Visual approach: Berkeley Data 8, §15.5. Original chart.

Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

Math recap: average, square and square root

AVERAGE	SQUARE	SQUARE ROOT
$(2 + 4) \div 2 = 3$	$(-3)^2 = 9$	$\sqrt{9} = 3$
Add the values. Divide by how many.	Multiply by itself. $(-3) \times (-3) = 9$	Undo the square. $3 \times 3 = 9$

Mohammad Sabokrou | New Uzbekistan University

MSE averages squared errors. RMSE takes the square root of MSE.

Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

Three ways to summarize prediction errors

Residuals: +2, −2, +4 minutes

MAE	MSE	RMSE
$(2 + 2 + 4) / 3 \approx 2.67$	$(4 + 4 + 16) / 3 = 8$	$\sqrt{8} \approx 2.83$
Typical absolute miss	Training loss in min²	Larger misses count more

Mohammad Sabokrou | New Uzbekistan University

Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

One training cycle

01	**Predict**	Use the current b and w.
02	**Measure loss**	Compare predictions with targets.
03	**Find gradients**	Which changes reduce the loss?
04	**Update**	Take a small step, then repeat.

Mohammad Sabokrou | New Uzbekistan University

Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

A tiny dataset for learning by hand

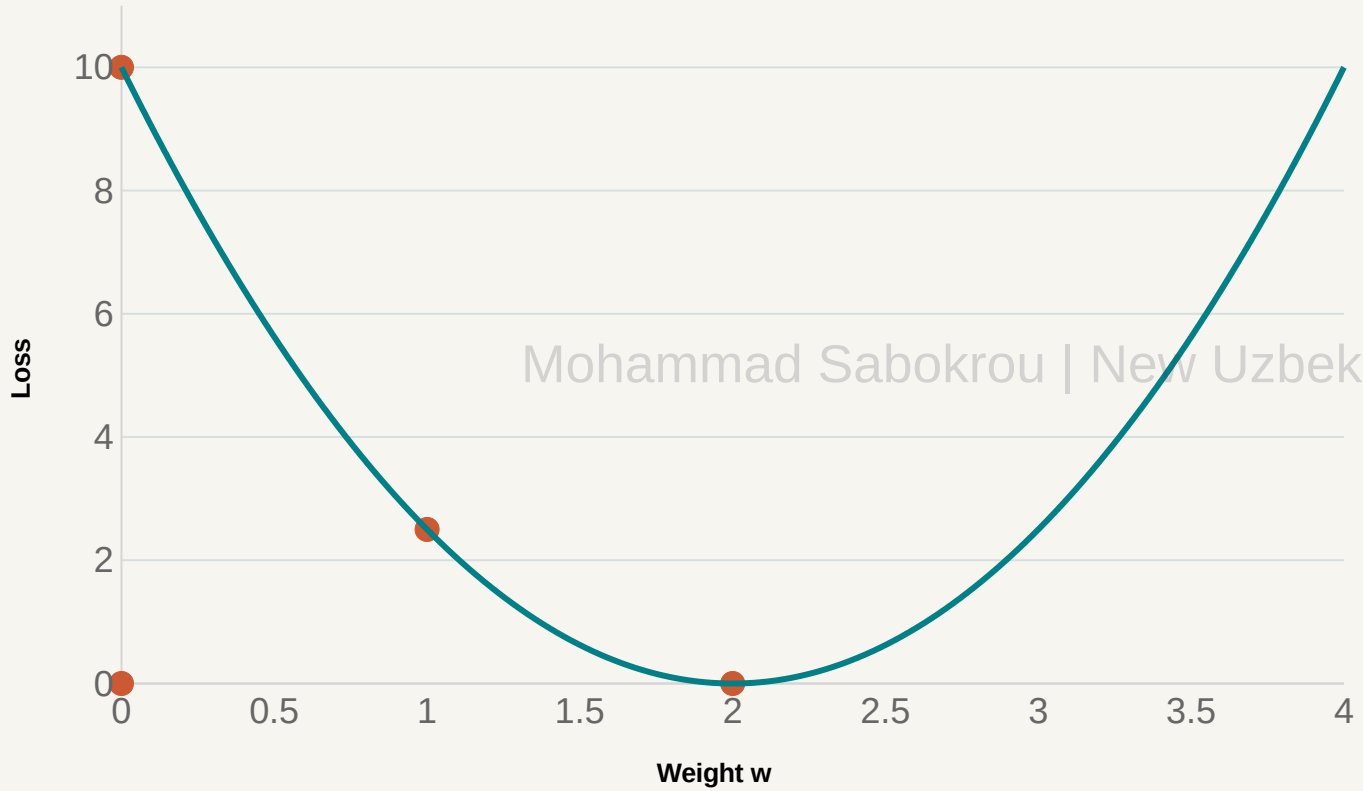
Input x	Target y	Prediction with w = 0
1	2	0
2	4	0

Mohammad Sabokrou | New Uzbekistan University

For this toy example only: predicted y = w × x; intercept fixed at zero.

Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

A loss measures how badly a weight fits



Mohammad Sabokrou | New Uzbekistan University

Toy data: (1, 2), (2, 4)

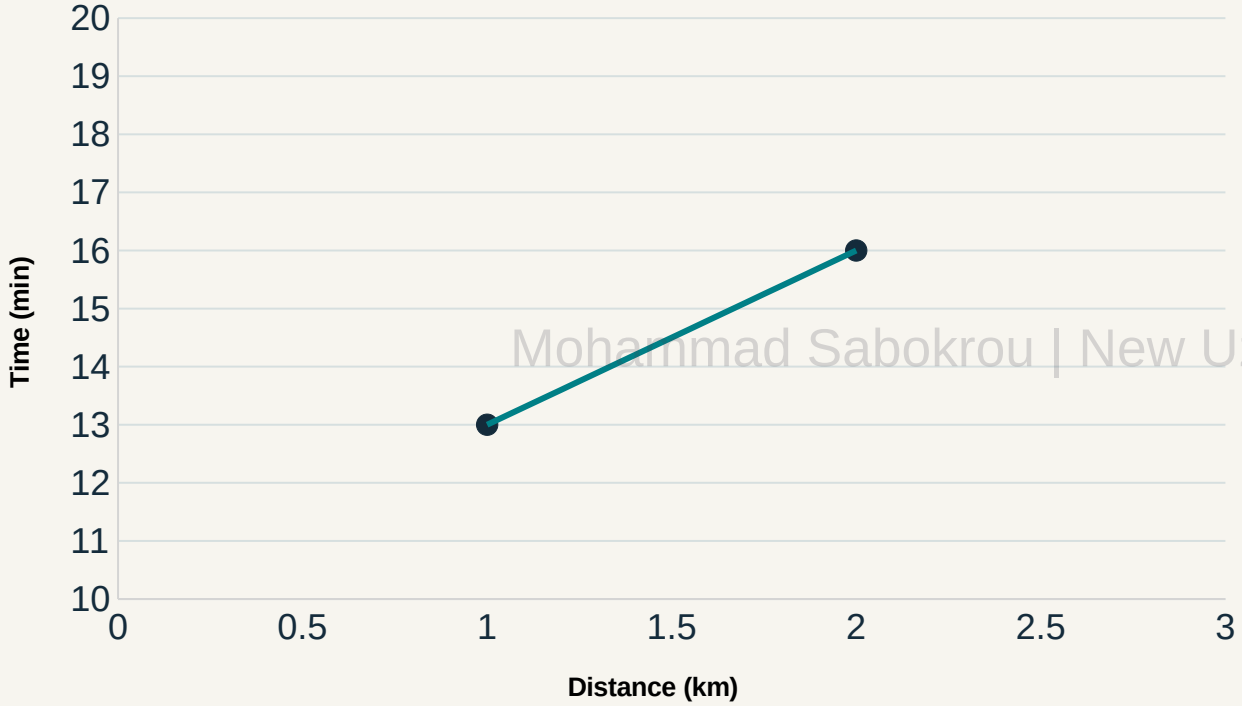
$\hat{y} = wx, \quad b = 0$

w = 0: loss 10
w = 1: loss 2.5
w = 2: loss 0

Loss = average squared prediction error

Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

Math recap: slope means “change per step”



Mohammad Sabokrou | New Uzbekistan University

1 km farther

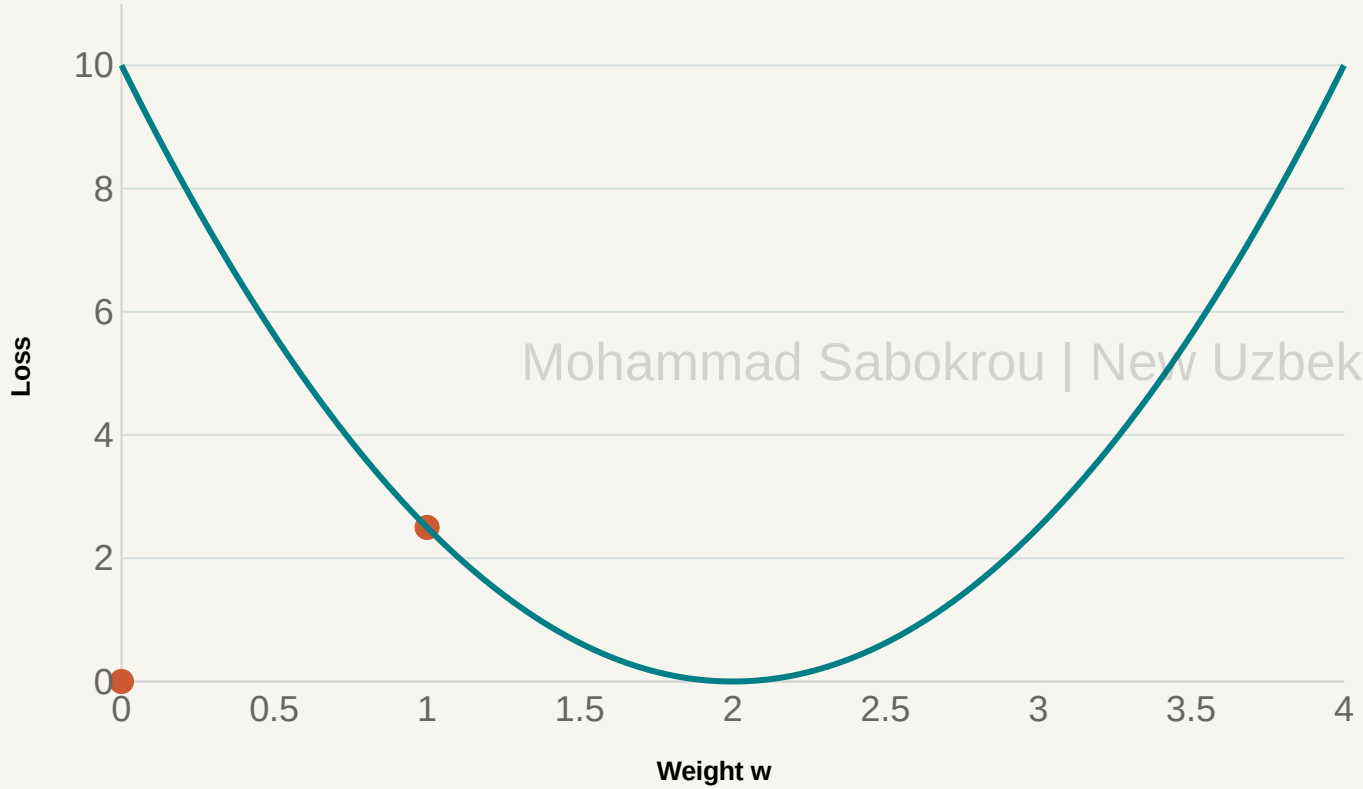
3 minutes more

Slope = 3 ÷ 1 = 3

A gradient uses the same idea: how loss changes when a parameter changes a little.

Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

Gradient means local slope of the loss



Mohammad Sabokrou | New Uzbekistan University

At w = 1

$g = dL/dw = -5$

Increase w a little.
Loss falls here.

Negative g: move right.
Positive g: move left.

For this toy: $L = 2.5(w - 2)^2$ and $g = 5(w - 2)$

Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

One gradient-descent update

$$w_{t+1} = w_t - \alpha g_t$$

old weight step size × gradient

Mohammad Sabokrou | New Uzbekistan University

$$w_1 = 0 - 0.1 \times (-10) = 1$$

t counts updates. α (alpha) is the learning rate.

After this step, the toy MSE falls from 10 to 2.5.

Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

Your turn: the next update

$$w_t = 1, \quad g_t = -5, \quad \alpha = 0.1$$

1. Calculate the next weight.

Mohammad Sabokrou | New Uzbekistan University

2. Predict y for x = 1 and x = 2.

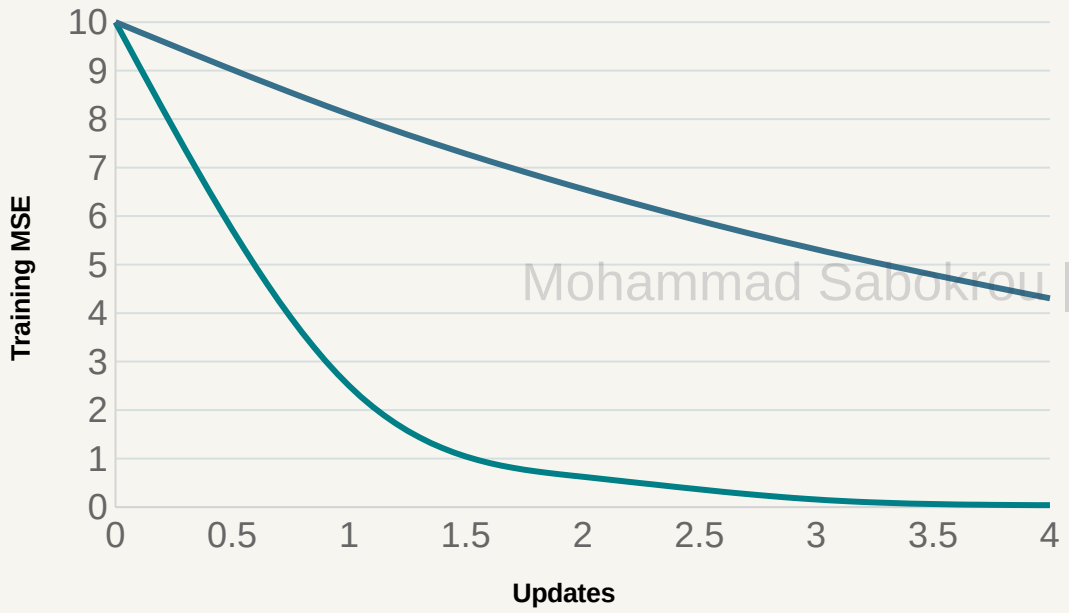
3. Calculate MSE against targets 2 and 4.

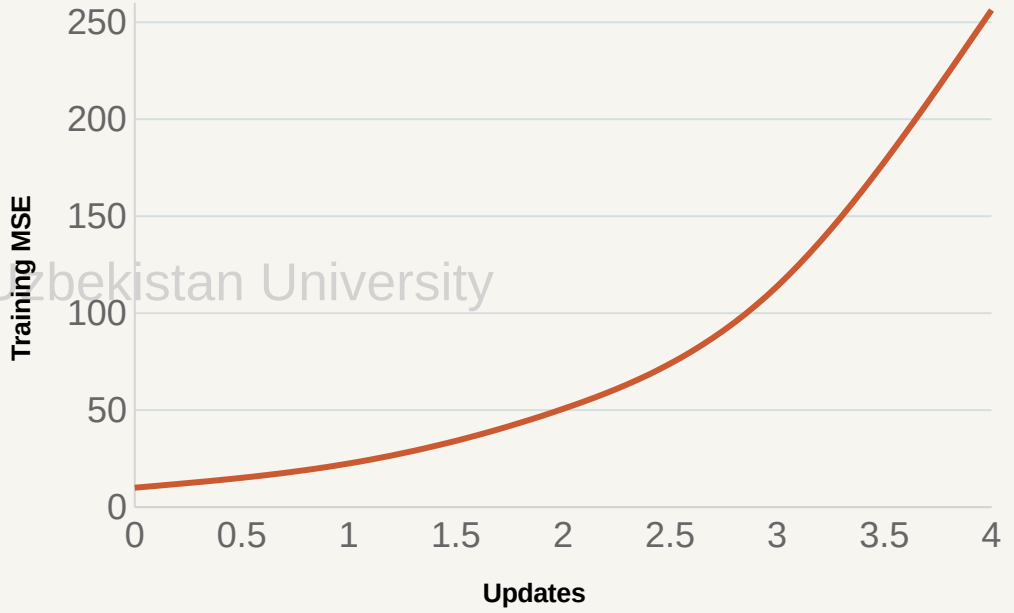
Use: $w_{t+1} = w_t - \alpha g_t$

Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

The learning rate controls the size of each step

Same toy • same start: w = 0 • same initial MSE = 10





Mohammad Sabokrou | New Uzbekistan University

α = 0.02: slow progress

α = 0.1: loss falls quickly

α = 0.5: loss grows

Notice the different vertical scales.

Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

Learning slope and intercept together

$$\hat{y} = b + wx \qquad e = \hat{y} - y$$

Prediction minus observation gives the training error e.

Mohammad Sabokrou | New Uzbekistan University

$$\text{Gradient for } b = 2 \times \text{average}(e)$$

$$\text{Gradient for } w = 2 \times \text{average}(e \times x)$$

Compute both gradients, then update b and w together.

Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

Feature scaling helps gradient descent

Step	Training-only calculation
Center	Subtract the training mean
Scale	Divide by the training standard deviation
Predict new rows	Reuse the same stored mean and standard deviation

Mohammad Sabokrou | New Uzbekistan University

For our distances: z = (x − 3.5) ÷ 1.708.

Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

Python: training is a repeated calculation

```
import numpy as np
x = np.arange(1., 7.)
y = np.array([14., 14., 20., 23., 23., 29.])
mu, sd = x.mean(), x.std()
z = (x - mu) / sd
b, w, alpha = 0., 0., 0.1
for epoch in range(100):
    e = b + w*z - y
    gb, gw = 2*e.mean(), 2*(e*z).mean()
    b, w = b-alpha*gb, w-alpha*gw
slope = w/sd
intercept = b - slope*mu
print(round(intercept, 3), round(slope, 3))
```

Mohammad Sabokrou | New Uzbekistan University

Output: 10.0, 3.0 **Try one epoch. What changes?**

Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

Five-minute break

When we return

Compare simple and flexible models.

Then control their weights.



Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

Model complexity means flexibility

Constant	$\hat{y} = b$	1 parameter
Straight line	$\hat{y} = b + w_1 x$	2 parameters
Quadratic curve	$\hat{y} = b + w_1 x + w_2 x^2$	3 parameters

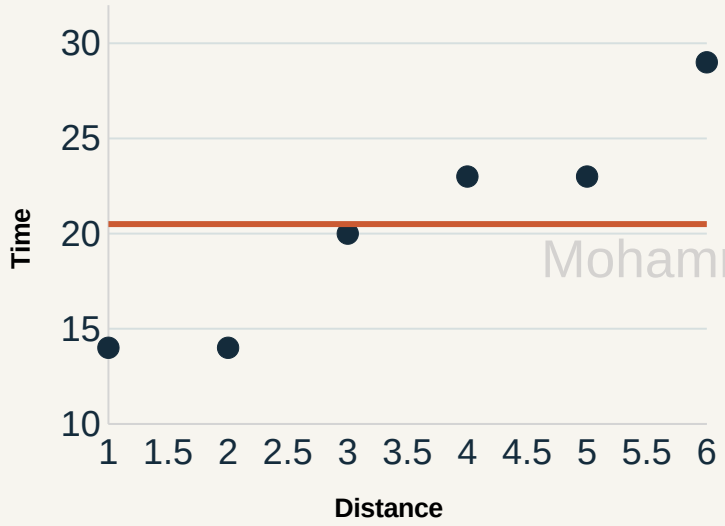
Mohammad Sabokrou | New Uzbekistan University

More flexibility can capture a pattern or fit noise.

Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

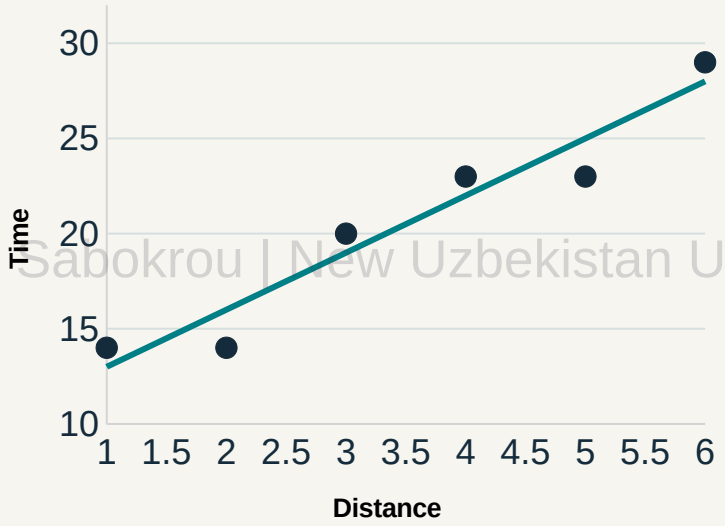
Underfitting, a useful fit, and overfitting

Underfitting



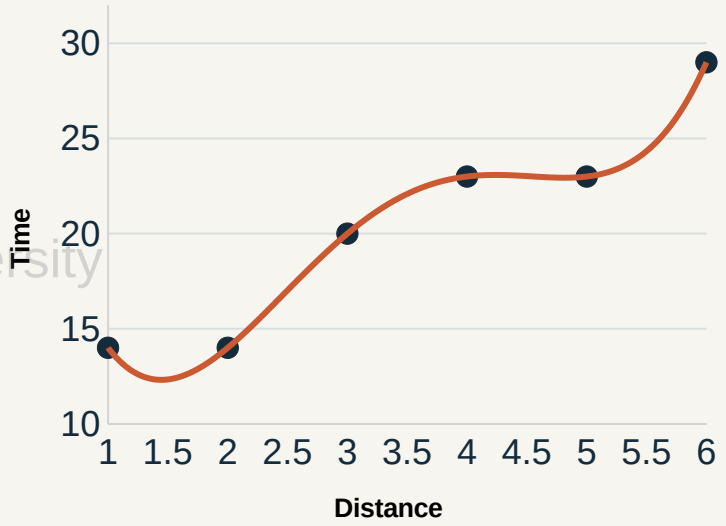
Misses the distance pattern.

Useful fit



Captures the main trend.

Overfitting



Chases each training point.

Mohammad Sabokrou | New Uzbekistan University

New-data performance decides whether extra flexibility helps.

Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

New deliveries: a fair comparison

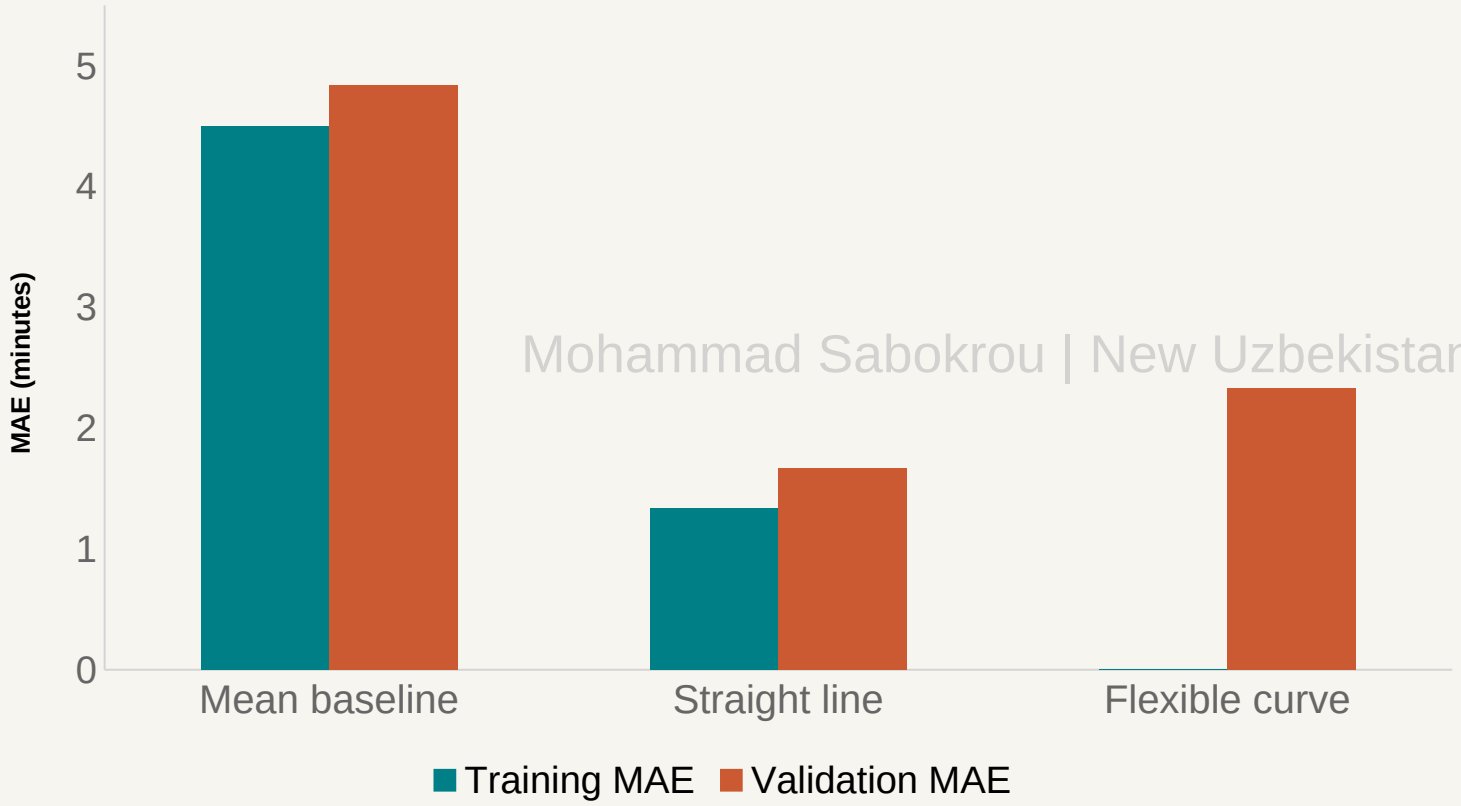
Distance	Observed	Baseline: 20.5	Line: 10 + 3x
2 km	17 min	20.5 min	16 min
4 km	21 min	20.5 min	22 min
6 km	31 min	20.5 min	28 min

Mohammad Sabokrou | New Uzbekistan University

Use the fixed training baseline and fitted line. Do not refit on these outcomes.

Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

Training perfection can hide worse predictions



0 training error

does not guarantee
the best new-data result.

Mohammad Sabokrou | New Uzbekistan University

Validation MAE Line: 1.67 min
Curve: 2.33 min

Same three validation deliveries for every model. Small illustration, not a deployment claim.

Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

Underfitting and overfitting suggest different fixes

UNDERFITTING

Training error is high.
Validation error is also high.

Try better features, a richer model,
or check whether training converged.

OVERFITTING

Training error is low.
Validation error is much higher.

Try less flexibility, regularization,
or more representative training data.

Mohammad Sabokrou | New Uzbekistan University

Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

Regularization adds a preference during training

$$J = \text{prediction loss} + \text{weight penalty}$$

Fit the training targets while discouraging large weights.

Mohammad Sabokrou | New Uzbekistan University

λ (lambda) controls how strongly we apply the penalty.

λ = 0 gives ordinary unregularized fitting.

Linear regression with an L_2 penalty is ridge regression.

Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

The L_2 penalty, without vector notation

$$J = \text{MSE} + (\lambda/2)(w_1^2 + w_2^2)$$

Square each weight, add the squares, then multiply by λ/2.

Mohammad Sabokrou | New Uzbekistan University

$$w_1 = 2, \quad w_2 = -1, \quad \lambda = 0.4$$

$$\text{penalty} = 0.2 \times (4 + 1) = 1$$

Here b has no penalty. Scale the input features before comparing weights.

Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

Why the update is called weight decay

$$w_{t+1} = w_t - \alpha(g_t + \lambda w_t)$$

The penalty adds λw to the data-loss gradient.

Mohammad Sabokrou | New Uzbekistan University

$$w_{t+1} = (1 - \alpha\lambda)w_t - \alpha g_t$$

Shrink the old weight, then take the data-gradient step.

This equivalence is for ordinary gradient descent / SGD.

Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

One update with and without weight decay

Same starting point: $w = 1$, $g = -5$, $\alpha = 0.1$

WITHOUT DECAY

$$1 - 0.1(-5) = 1.5$$

WITH $\lambda = 2$

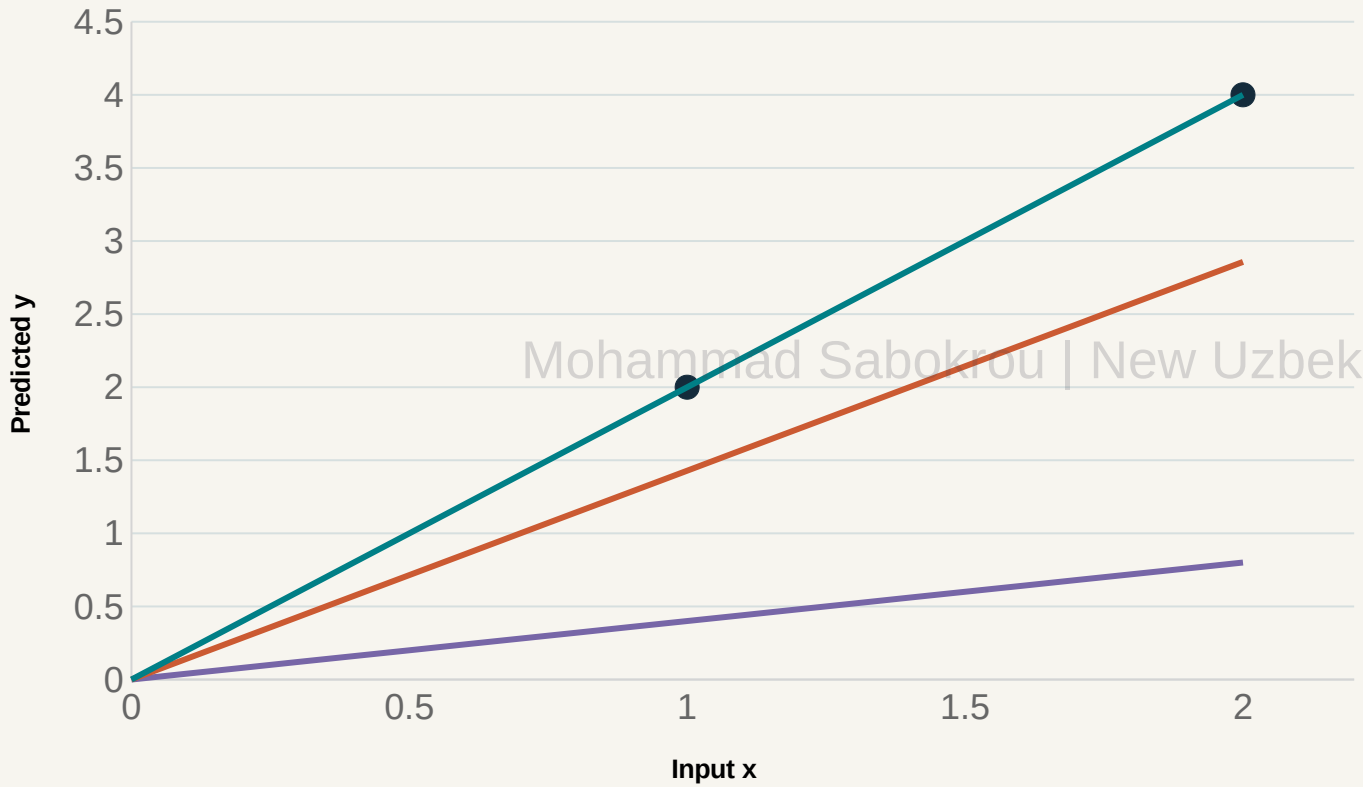
$$1 - 0.1(-5 + 2) = 1.3$$

Mohammad Sabokrou | New Uzbekistan University

The weight still grows, but the penalty restrains the increase.

Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

Too much decay can underfit



Mohammad Sabokrou | New Uzbekistan University

λ = 0: w = 2

λ = 2: w ≈ 1.43

λ = 20: w = 0.4

Same two-row toy. Each line minimizes its own penalized objective.

Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

Learning rate and regularization do different jobs

α: learning rate	How large is each training step?
λ: penalty strength	How strongly do we prefer smaller weights?

Mohammad Sabokrou | New Uzbekistan University

Degree: model complexity	How much can the fitted curve bend?
Epochs: training duration	How many passes through the data?

Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

Practice questions differ from an exam

Training	Validation	Test
Learn model parameters	Choose the approach	Final evaluation

Mohammad Sabokrou | New Uzbekistan University

The same data must not play every role at once.

Teaching approach: Stanford STATS 202, Validation.

Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

Five-fold cross-validation

Each column is a group of 20 rows from development data.

	A	B	C	D	E
Round 1	Validate	Train	Train	Train	Train
Round 2	Train	Validate	Train	Train	Train
Round 3	Train	Train	Validate	Train	Train
Round 4	Train	Train	Train	Validate	Train
Round 5	Train	Train	Train	Train	Validate

Mohammad Sabokrou | New Uzbekistan University

Fresh model each round. The final test set stays separate.

Adapted concept: Stanford STATS 202, K-fold CV. Original diagram.

Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

Choosing degree and λ with cross-validation

01 Reserve the final test set.

02 For each degree and λ, refit on every training fold.

Mohammad Sabokrou | New Uzbekistan University

03 Compare average validation MAE, without the penalty.

04 Refit the selected approach, then evaluate the test.

Scaling belongs inside each training fold.

Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

Leakage gives the model an unfair preview

Tempting shortcut	Why it misleads
Use actual arrival time as an input	Unavailable at order time
Choose predictors using all outcomes	Validation answers influence choices
Fill missing values before splitting	Held-out values influence preprocessing

Mohammad Sabokrou | New Uzbekistan University

Fit preprocessing and feature selection only on each training fold.

Stanford STATS 202 and scikit-learn, Common pitfalls.

Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

Exit check: explain the complete learning process

01 At $w = 1$, $g = -5$ and $\alpha = 0.1$, what is the next weight?

02 With $\lambda = 2$, what does that update become?

Mohammad Sabokrou | New Uzbekistan University

03 Why can a perfect training fit fail on new deliveries?

04 Where do we fit scaling and choose model complexity?

Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

Teaching sources: regression and evaluation

UC Berkeley Data 8: regression and residuals
inferentialthinking.com/chapters/15/Prediction.html

Stanford CS229: model selection and cross-validation
cs229.stanford.edu/notes_archive/cs229-notes5.pdf

Mohammad Sabokrou | New Uzbekistan University

University of Toronto CSC311: gradient descent
teach.cs.toronto.edu/~csc311notes/linear_regression/gd.html

Copyright Mohammad Sabokrou | New Uzbekistan University | 2026

Teaching sources: regularization and weight decay

Dive into Deep Learning, §3.7: weight decay
www.d2l.ai/chapter_linear-regression/weight-decay.html

Goodfellow, Bengio & Courville, Chapter 7: regularization
www.deeplearningbook.org/contents/regularization.html

Mohammad Sabokrou | New Uzbekistan University

Loshchilov & Hutter, ICLR 2019: decoupled weight decay
arxiv.org/abs/1711.05101

Copyright Mohammad Sabokrou | New Uzbekistan University | 2026