CS 260: Foundations of Data Science

Prof. Sara Mathieson

Fall 2023

HAVERFORD
COLLEGE

Admin

- **Lab 1** grades posted on Moodle
- **Lab 3** due Wednesday night
  - I'll check in with everyone today about Lab 3 during lab
- **Lab 4** posted today

Outline for September 26

- Recap SGD (stochastic gradient descent)

- Introduction to classification
  - Decision tree models
  - Probabilistic interpretation

- Evaluation Metrics
  - Confusion matrices
  - Precision and recall
  - ROC curves

Outline for September 26

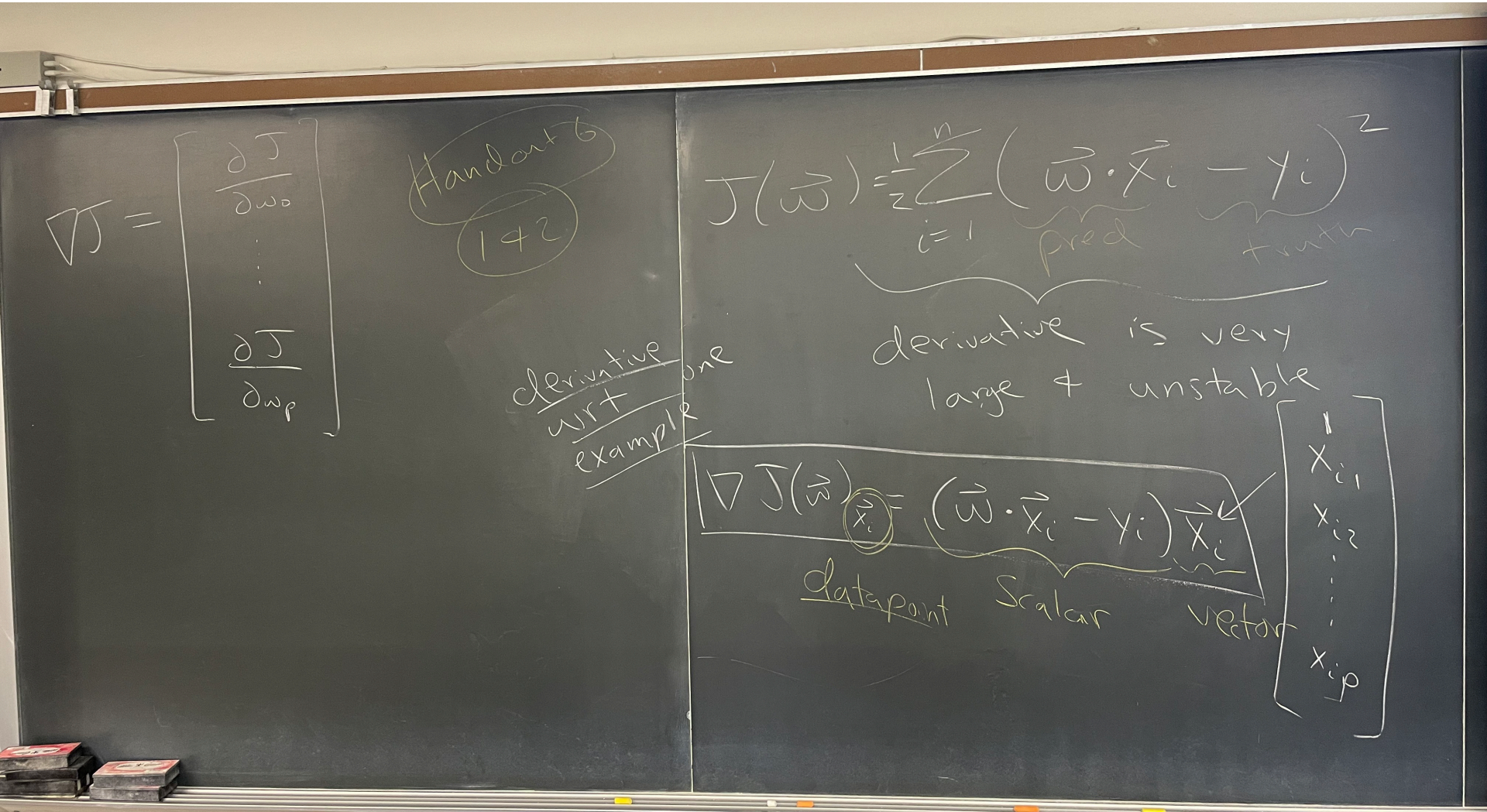
- Recap SGD (stochastic gradient descent)

- Introduction to classification
  - Decision tree models
  - Probabilistic interpretation

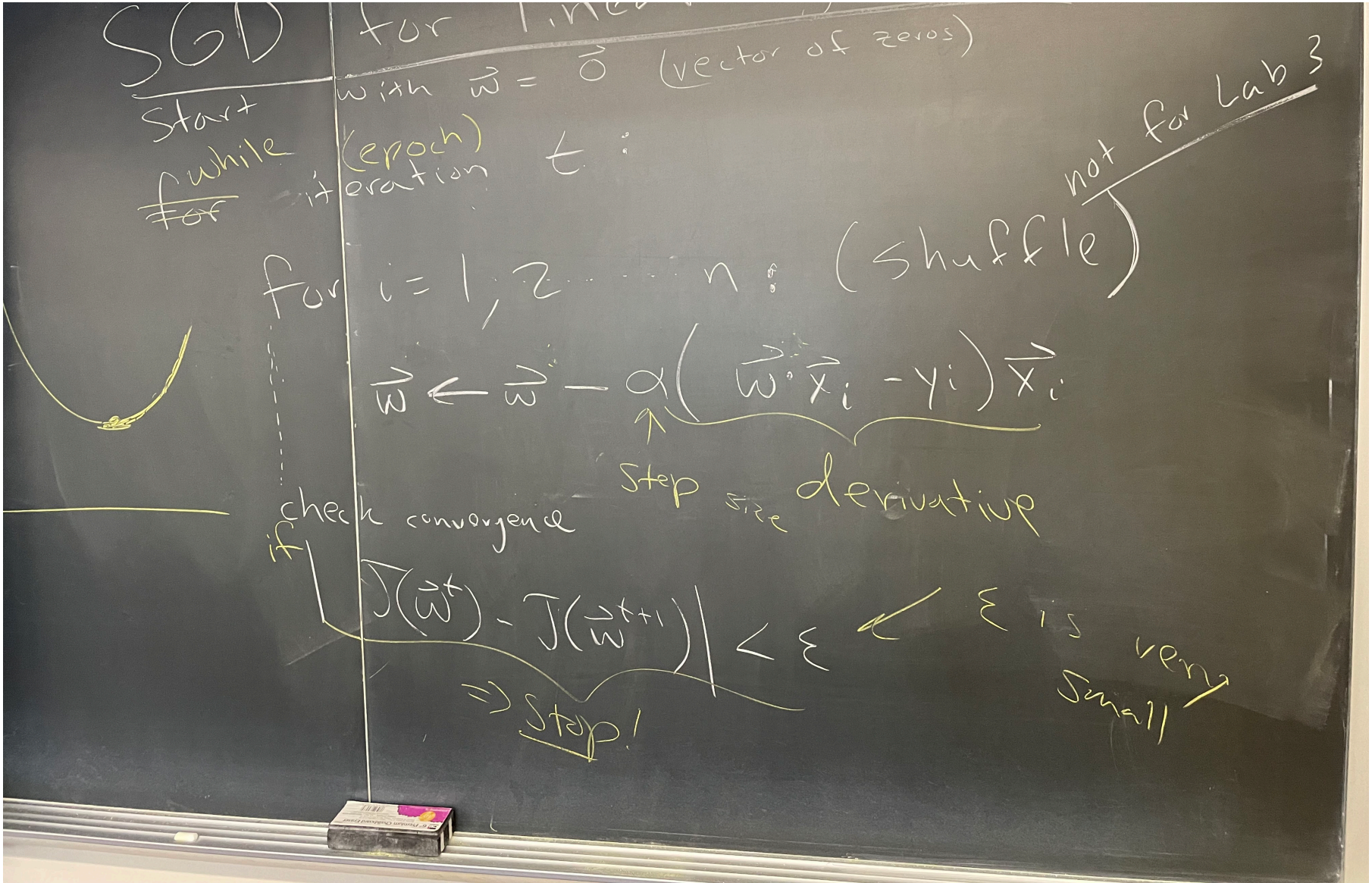
- Evaluation Metrics
  - Confusion matrices
  - Precision and recall
  - ROC curves

Stochastic Gradient Descent for Linear Regression

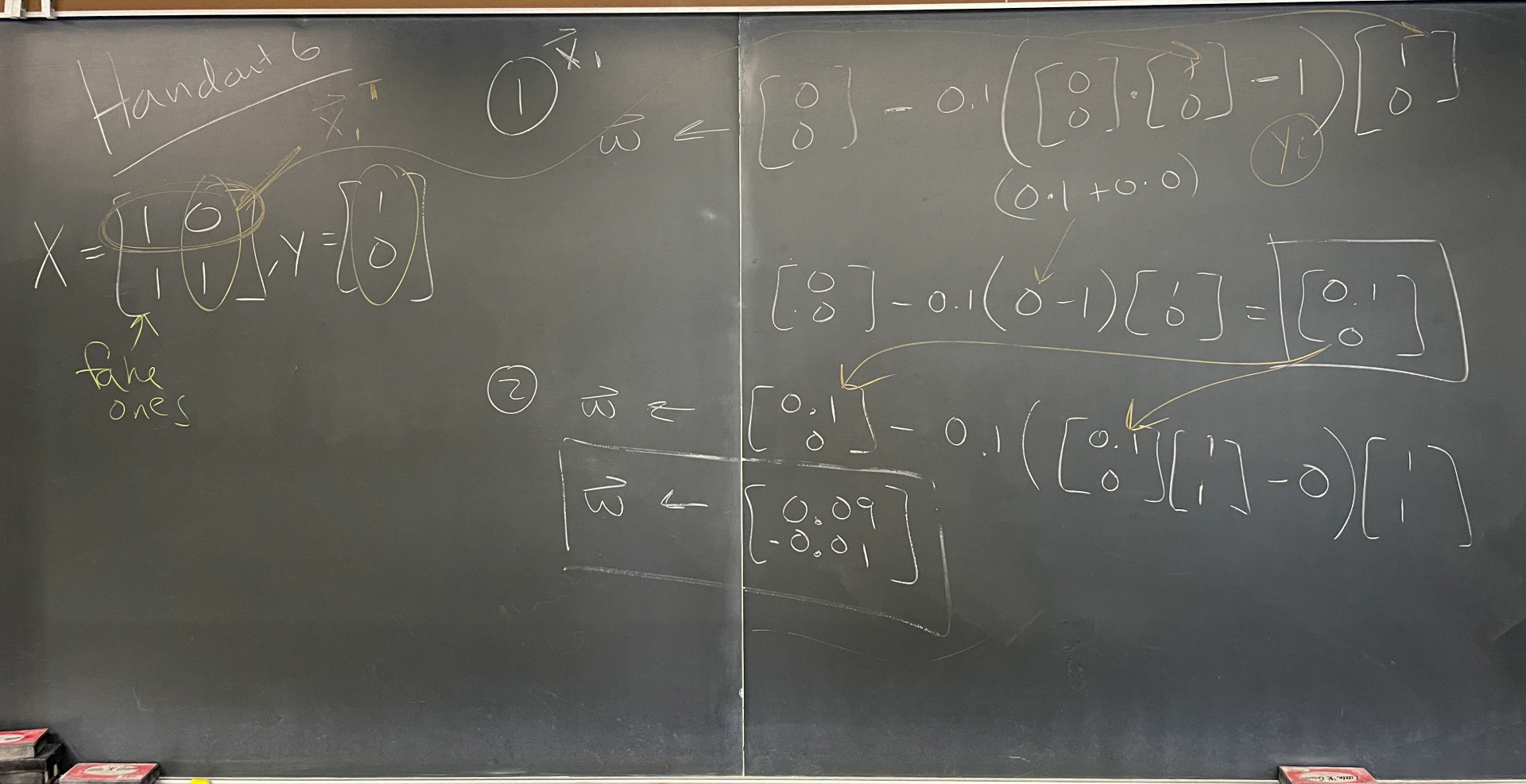
Key Idea: take the derivative of **one datapoint** at a time and use that to update w



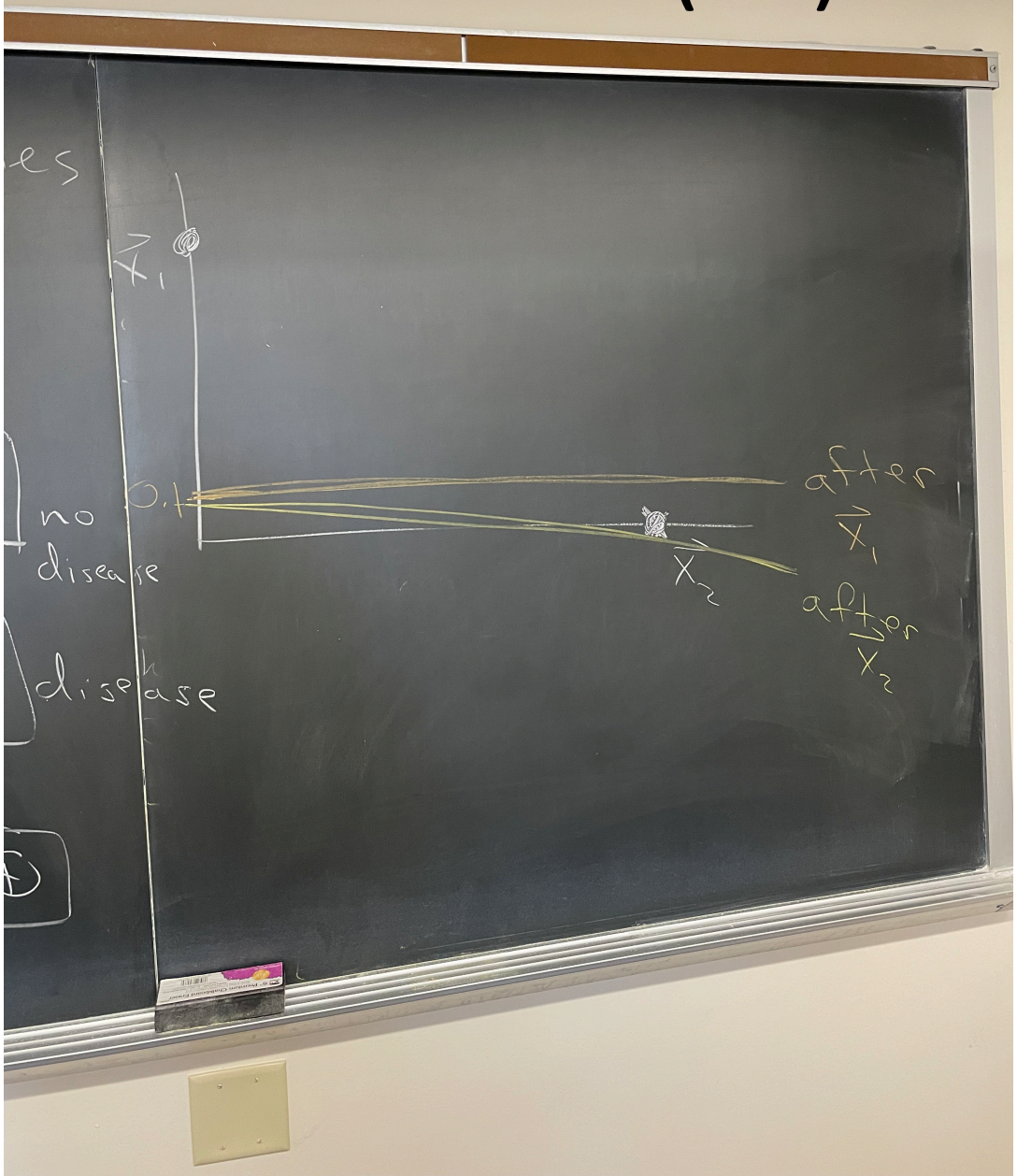
Stochastic Gradient Descent for Linear Regression



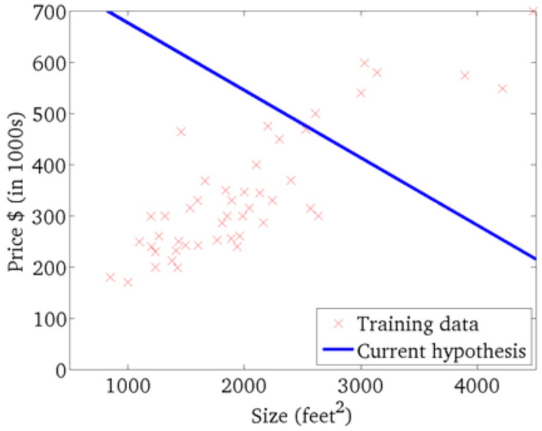
Handout 6 (#1, #2)

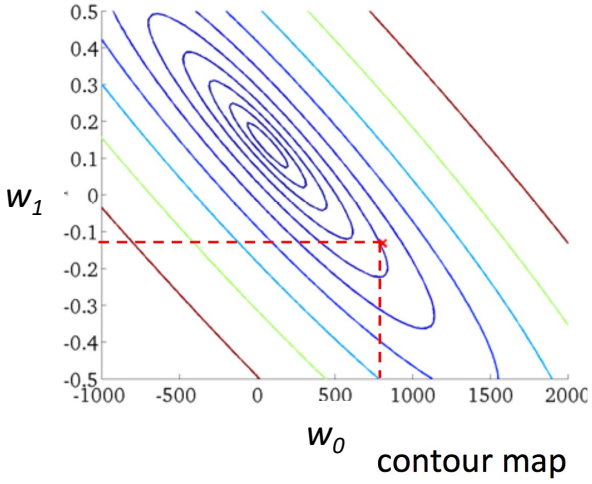


Handout 6 (#4)



Linear Model and Cost Function J

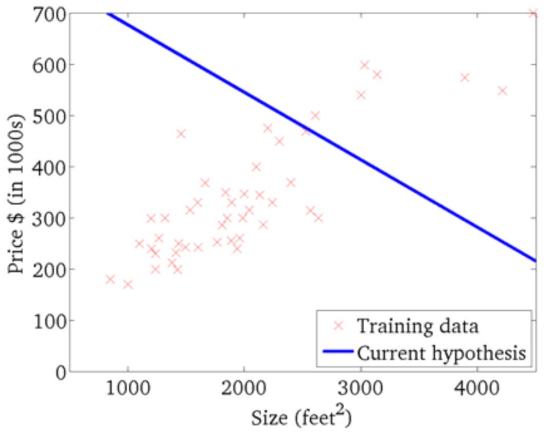


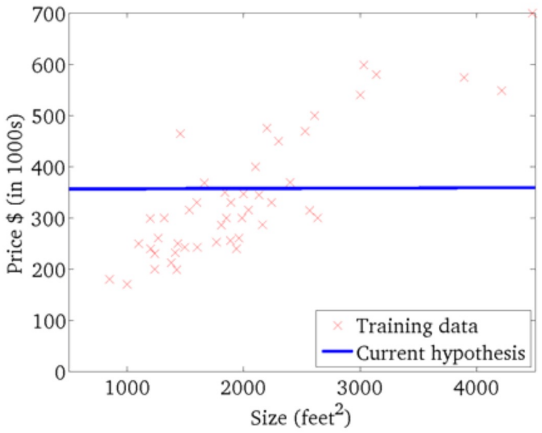


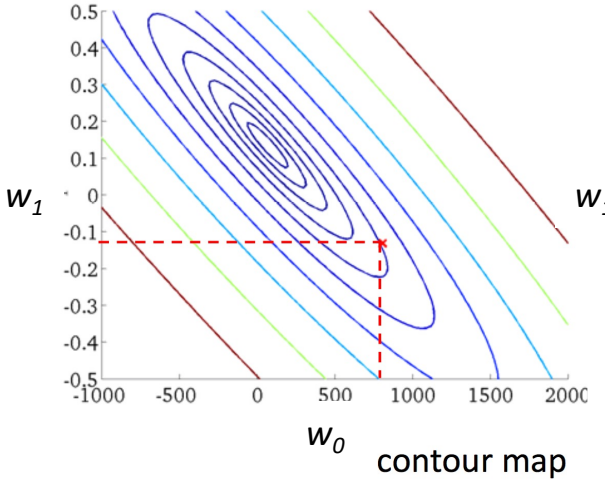
contour map

Slide: modified from Jessica Wu
Based on slide by Andrew Ng

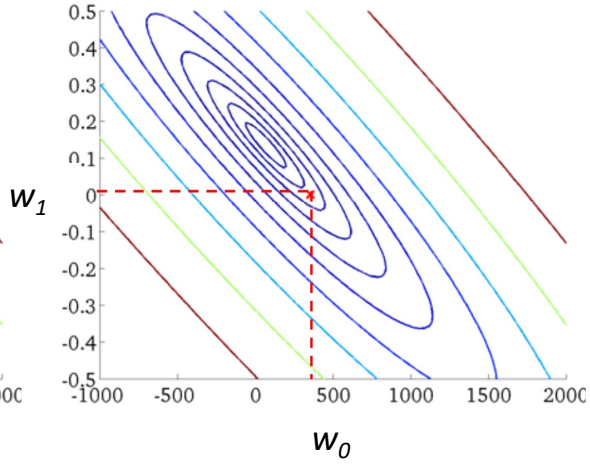
Linear Model and Cost Function J





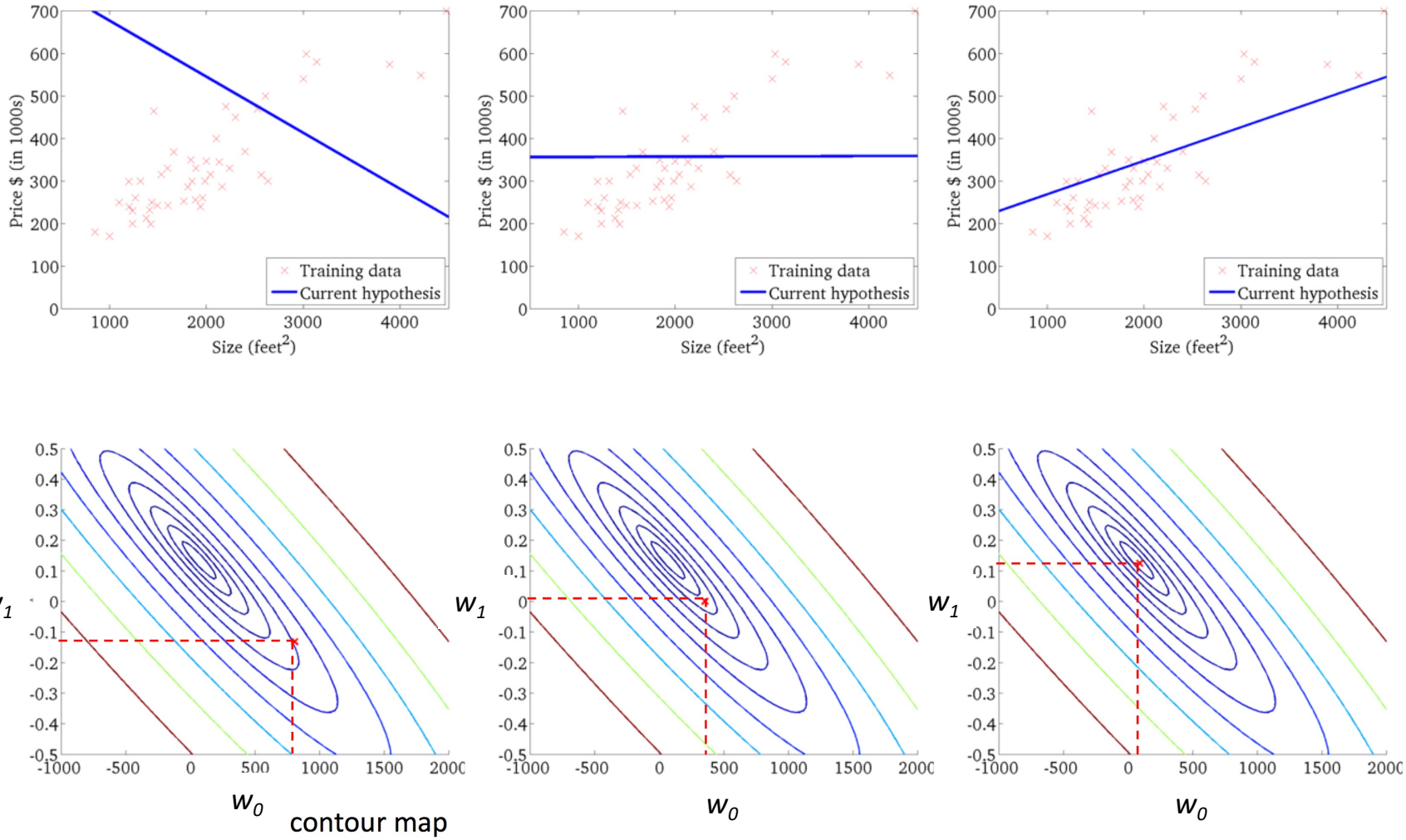


contour map



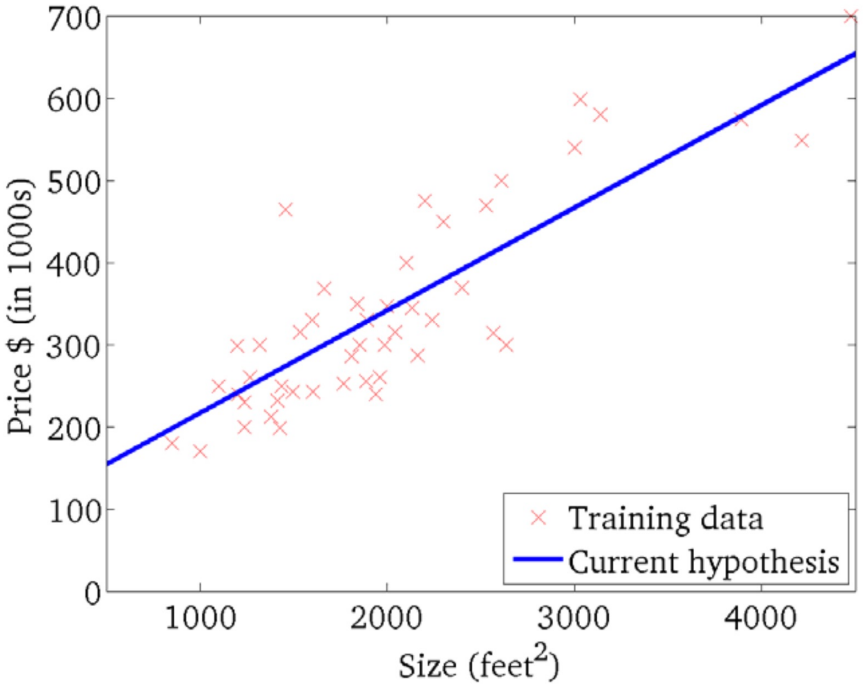
Slide: modified from Jessica Wu
Based on slide by Andrew Ng

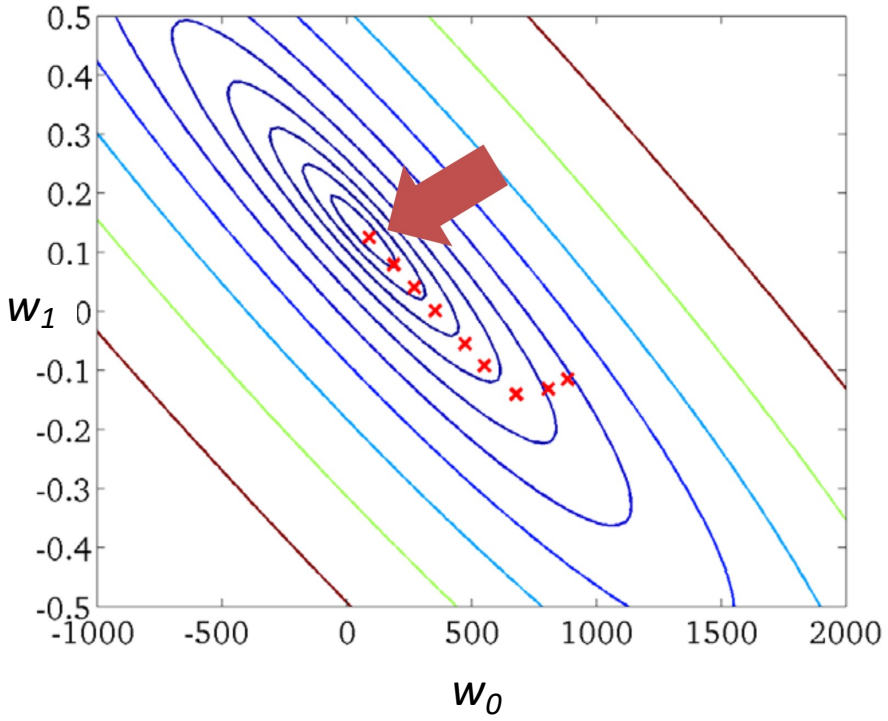
Linear Model and Cost Function J



Slide: modified from Jessica Wu
Based on slide by Andrew Ng

Gradient Descent: walking toward the minimum



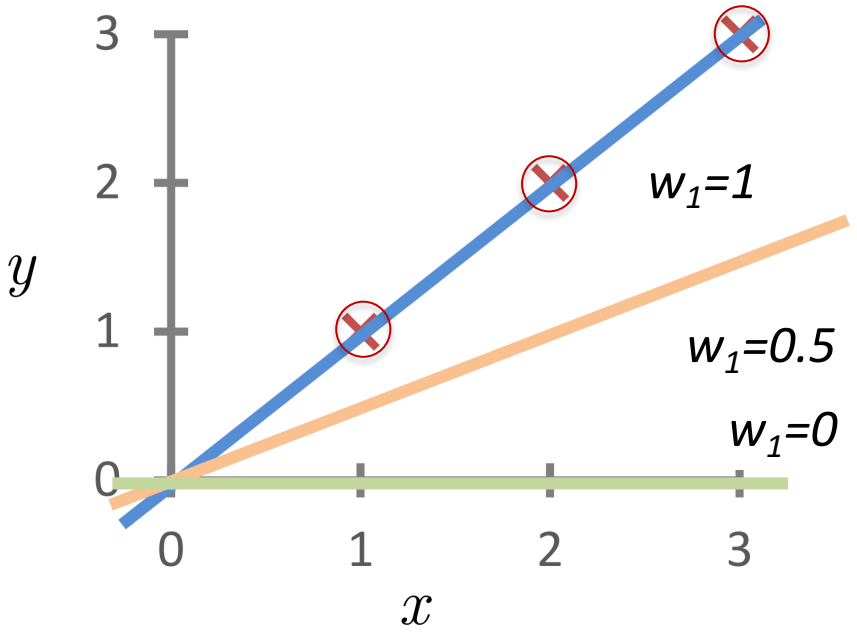


Slide: modified from Jessica Wu
Based on slide by Andrew Ng

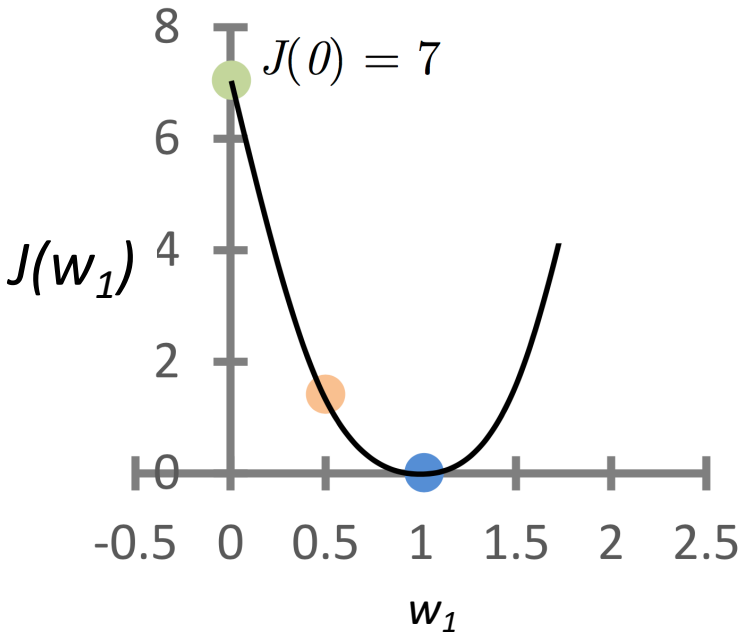
Cost Function (extra practice)

$h_w(x) = w_1 x$

(assume $w_0=0$ for this example)



$J(w_1)$



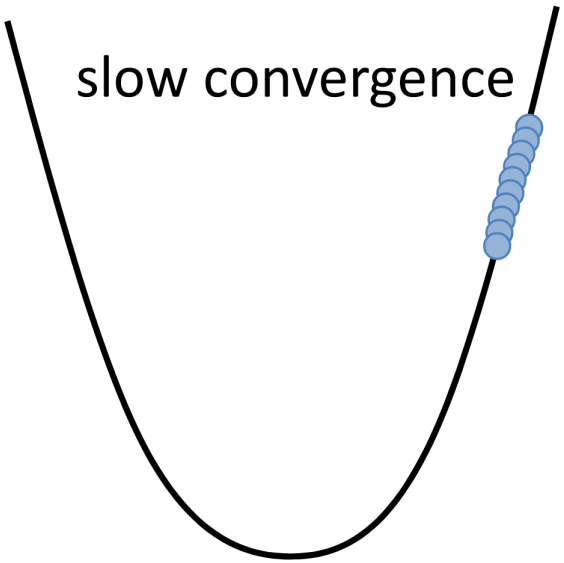
$$J(0.5) = \frac{1}{2}\left[(0.5-1)^2 + (1-2)^2 + (1.5-3)^2\right] = 1.75$$

Slide: modified from Jessica Wu
Based on slide by Eric Eaton
[Example by Andrew Ng]

Choosing the step size alpha

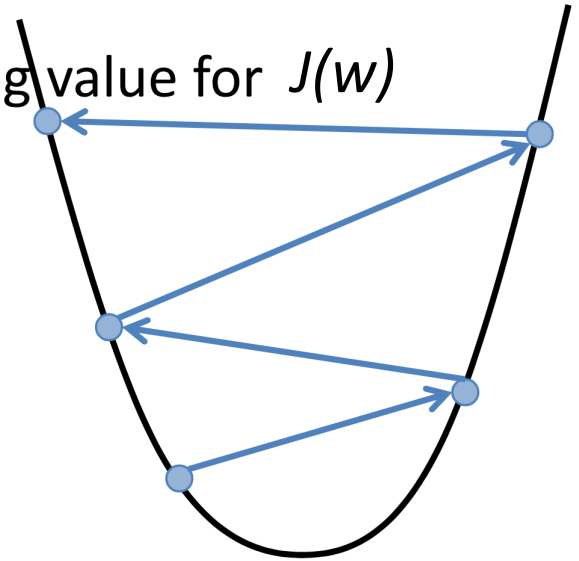
α too small

slow convergence



α too large

increasing value for $J(w)$



- may overshoot minimum
- may fail to converge (may even diverge)

Slide: modified from Jessica Wu
Based on slide by Eric Eaton
[Originally by Andrew Ng]

Handout 6

Linear Regression: SGD solution *(find and work with a partner)*

In linear regression, we seek to minimize the sum of squared errors between the actual response and our prediction. We often call this RSS (residual sum of squares) or SSE (sum of squared errors). As an objective function, we often call it J and include a $\frac{1}{2}$ in front to make the derivatives work out nicely.

$$J(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{n}\left(h_{\boldsymbol{w}}(\boldsymbol{x_i}) - y_i\right)^2$$

For linear regression in general, one iteration of stochastic gradient descent includes the following updates (usually with the data points shuffled):

```
for i = 1,2,...,n:
```
$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \alpha(\boldsymbol{w}\cdot\boldsymbol{x}_i - y_i)\boldsymbol{x}_i$$

We will begin with our same data from the previous two handouts: $(x_1, y_1) = (0, 1)$ and $(x_2, y_2) = (1, 0)$, except we will reverse the order of the points to make the progress of gradient descent a bit clearer. So in this case our matrix/vector formulation is:

$$\boldsymbol{X} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}, \qquad \boldsymbol{y} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Handout 6

$$\boldsymbol{X} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}, \qquad \boldsymbol{y} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

1. Assuming $\alpha = 0.1$ and our initial values are $w_0 = 0$ and $w_1 = 0$, what are w_0 and w_1 after the just the first data point is used to update the gradient?

$$\vec{w} \leftarrow \begin{bmatrix} 0 \\ 0 \end{bmatrix} - 0.1\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix} - 1 \right) \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$\vec{w} \leftarrow \begin{bmatrix} 0 \\ 0 \end{bmatrix} + 0.1 \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$\boxed{\vec{w} \leftarrow \begin{bmatrix} 0.1 \\ 0 \end{bmatrix}}$$

2. What are w_0 and w_1 after the second data point is used? Since we only have two examples here, your result would be the weight vector after the first iteration of SGD.

$$\vec{w} \leftarrow \begin{bmatrix} 0.1 \\ 0 \end{bmatrix} - 0.1\left(\begin{bmatrix} 0.1 \\ 0 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \end{bmatrix} - 0 \right) \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\boxed{\vec{w} \leftarrow \begin{bmatrix} 0.09 \\ -0.01 \end{bmatrix}}$$

3. What is the value of the objective function (cost) after this initial iteration?

$$\hat{y} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0.09 \\ -0.01 \end{bmatrix} = \begin{bmatrix} 0.09 \\ 0.08 \end{bmatrix}$$

$$J(\vec{w}) = \frac{1}{2} \begin{bmatrix} 0.91 \\ -0.08 \end{bmatrix} \cdot \begin{bmatrix} 0.91 \\ -0.08 \end{bmatrix}$$

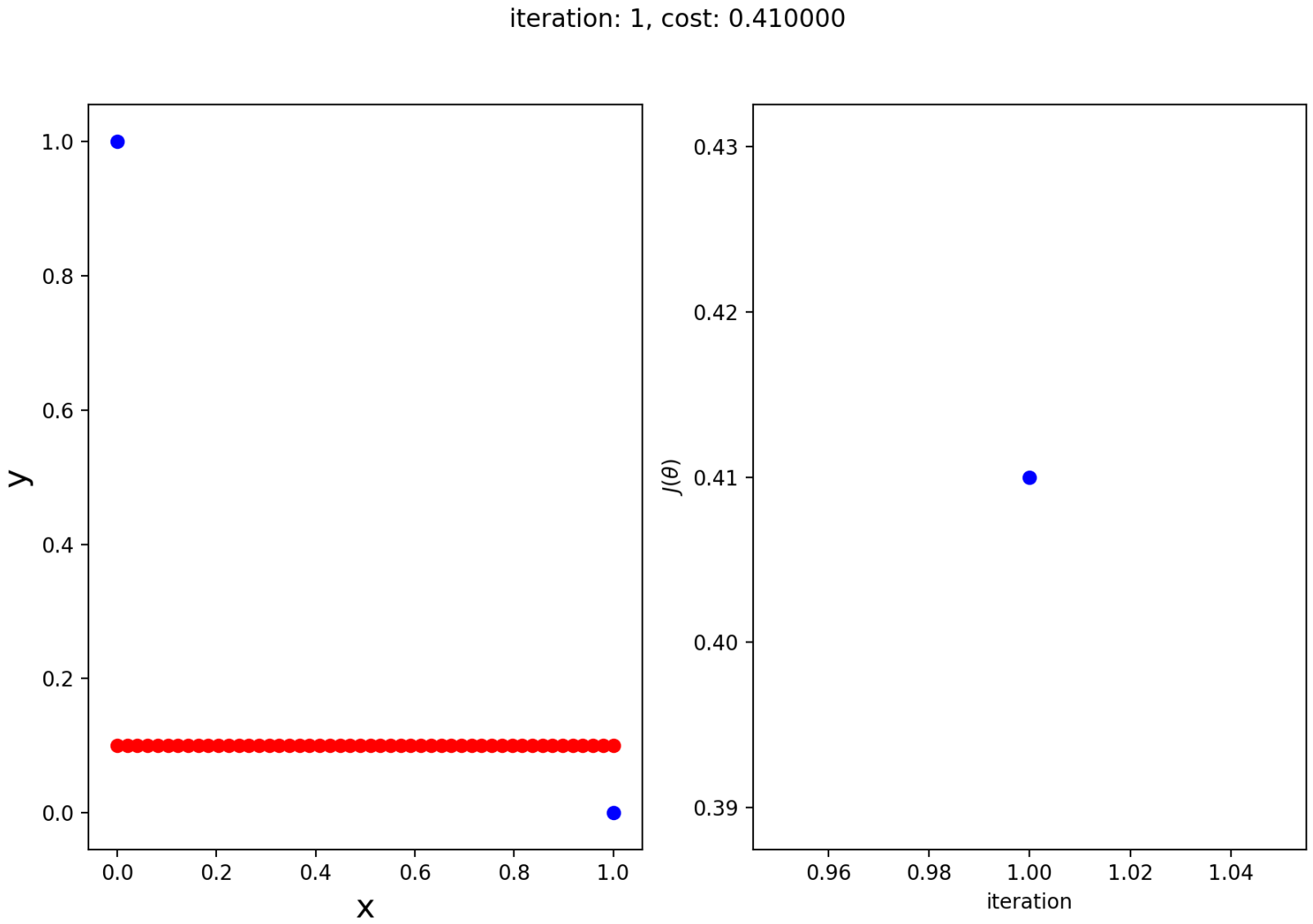
$$\vec{y} - \hat{y} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} - \begin{bmatrix} 0.09 \\ 0.08 \end{bmatrix} = \begin{bmatrix} 0.91 \\ -0.08 \end{bmatrix}$$

$$\boxed{J(\vec{w}) = 0.417}$$

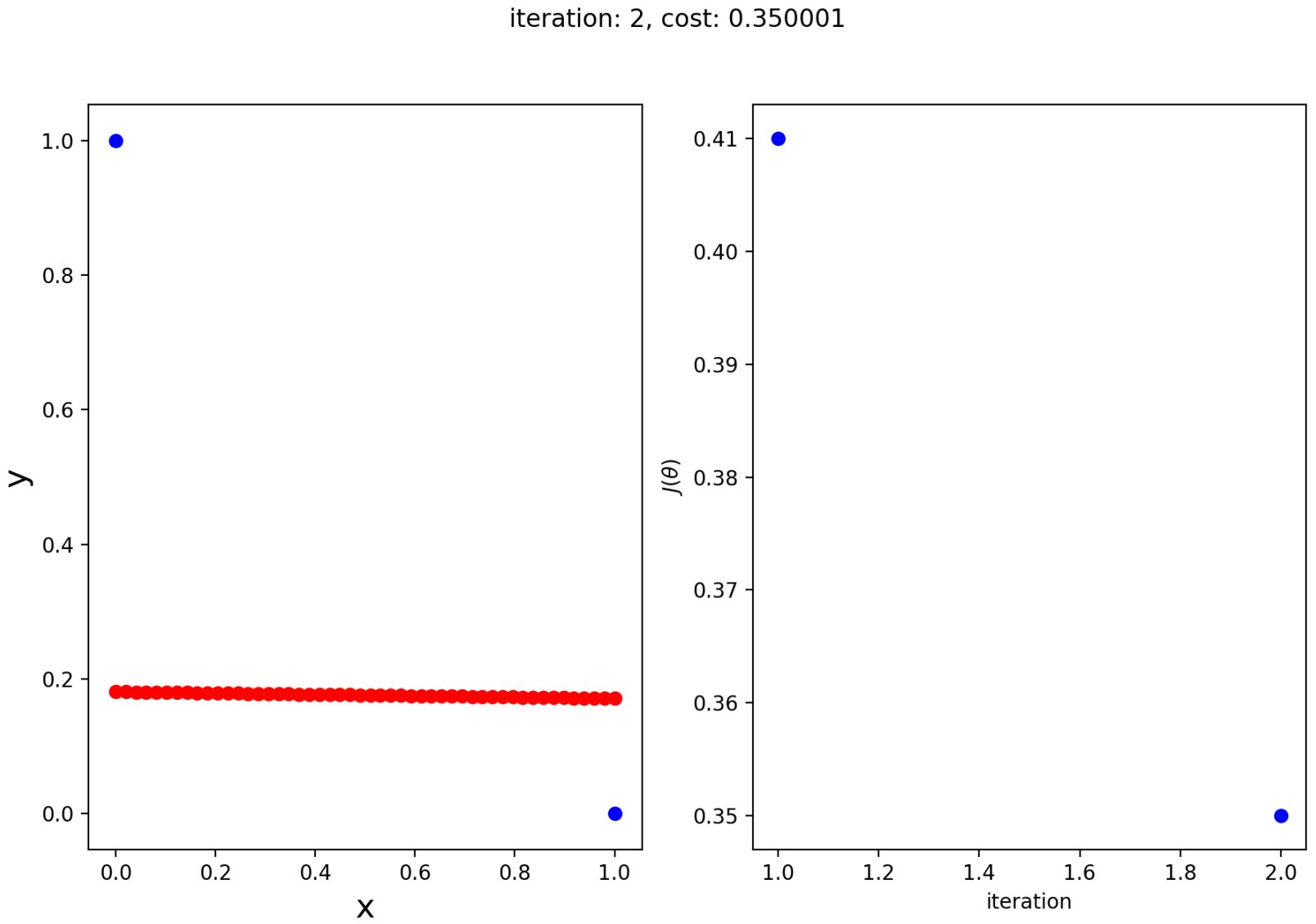
SGD with our small dataset from the handouts

Note: this is with the original order of the points

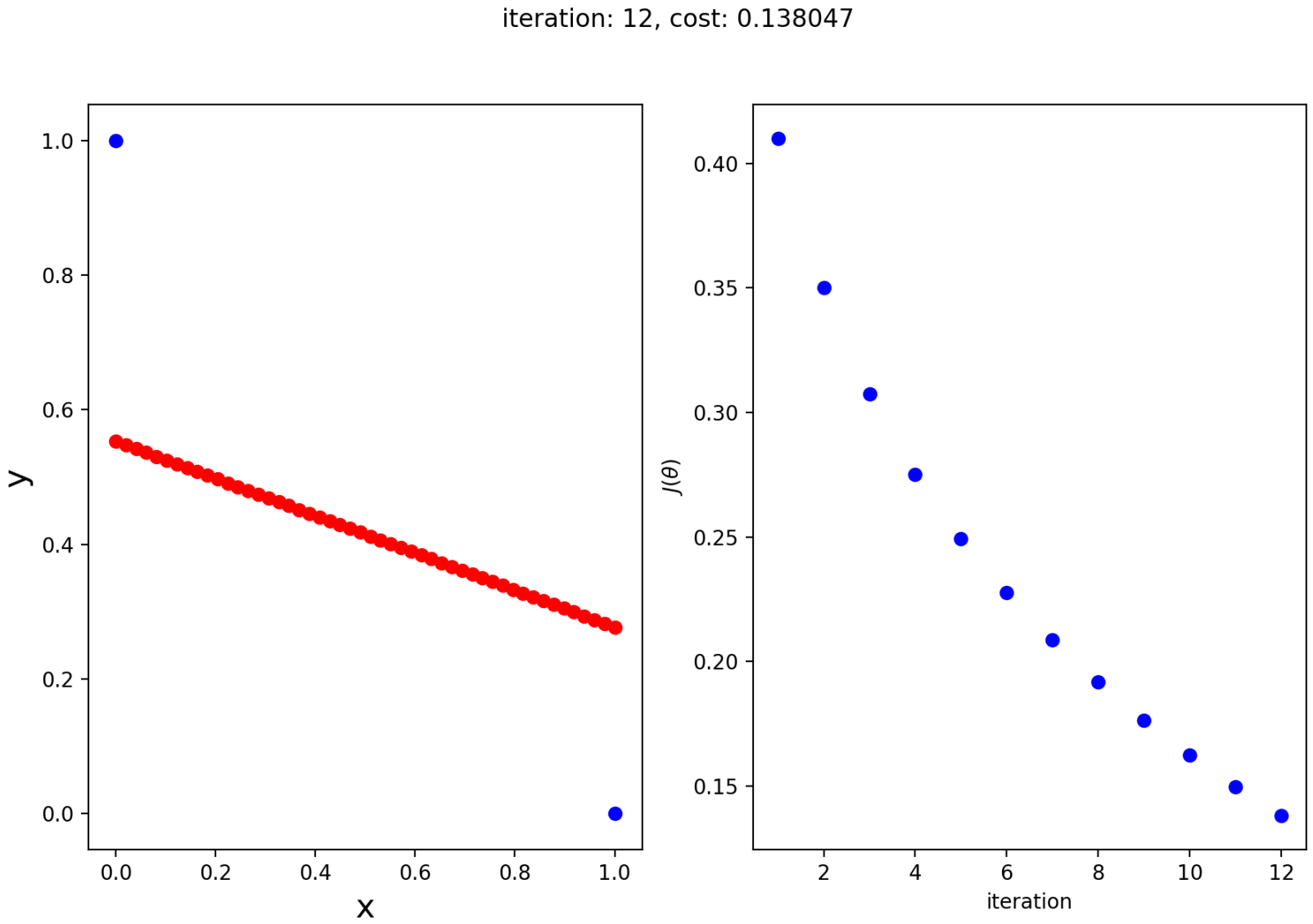
Small example, iteration 1



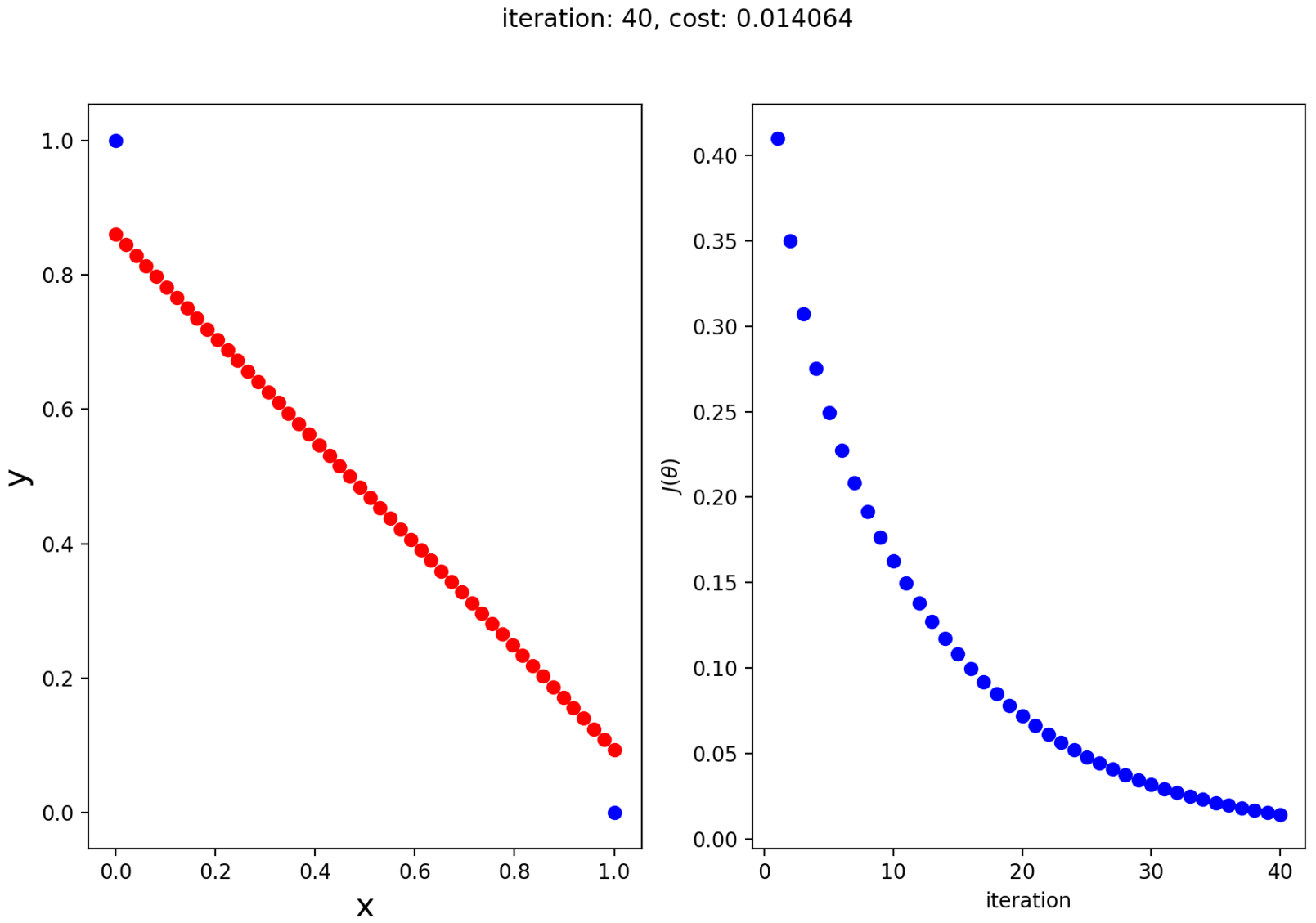
Small example, iteration 2



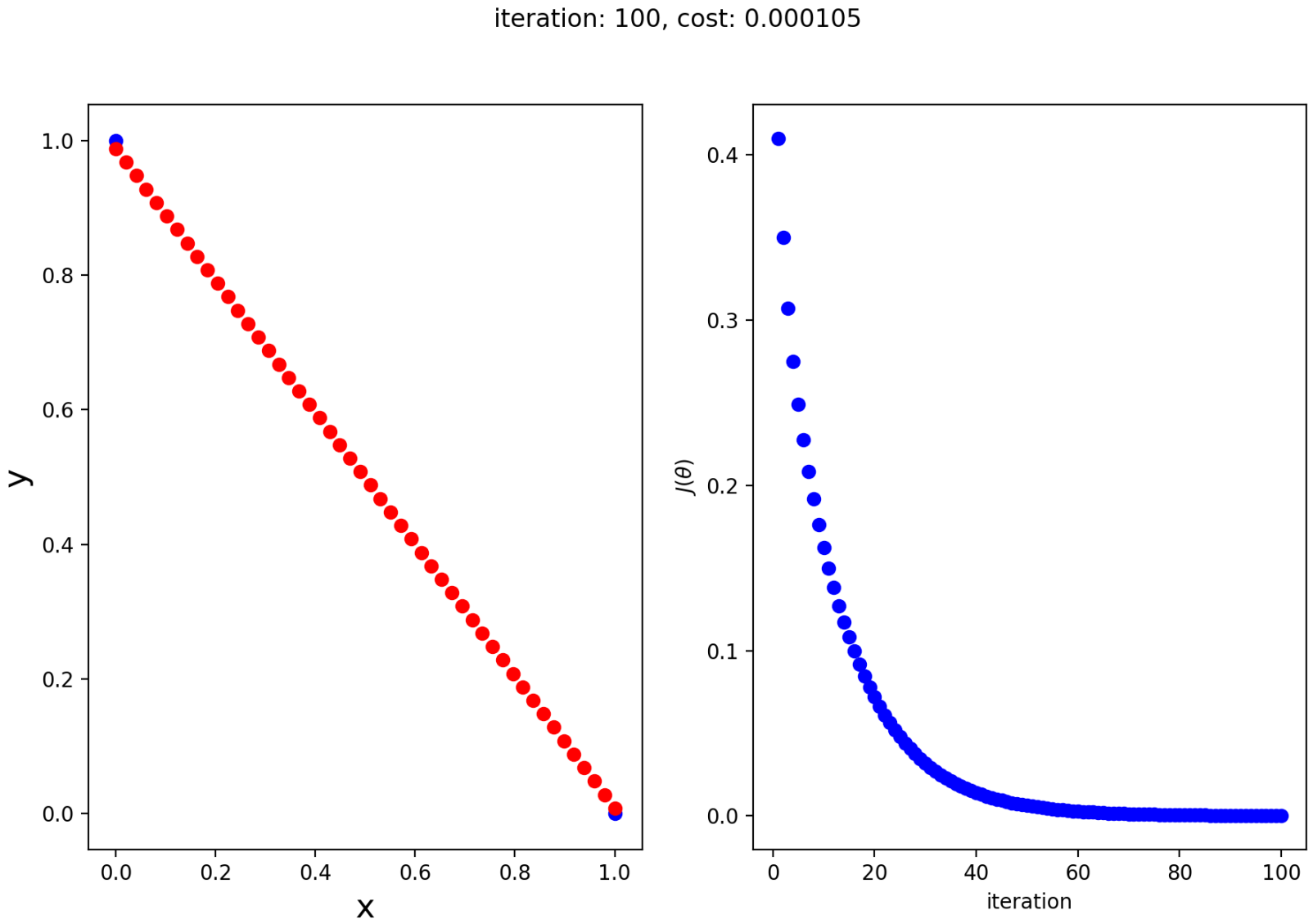
Small example, iteration 12



Small example, iteration 40



Small example, iteration 100



Pros and Cons

Gradient Descent

- requires multiple iterations
- need to choose α
- works well when p is large
- can support online learning

(Analytic Solution)

Normal Equations

- non-iterative
- no need for α
- slow if p is large
  - matrix inversion is $O(p^3)$

Slide: modified from Jessica Wu
Based on slide by Eric Eaton

Linear Regression Runtime

- T = # iterations of SGD
- n = # examples
- p = # features

1) What is the runtime of SGD?
2) What is the runtime of the analytic solution?

Challenge outside of class!

Outline for September 26

- Recap SGD (stochastic gradient descent)

- Introduction to classification
  - Decision tree models
  - Probabilistic interpretation

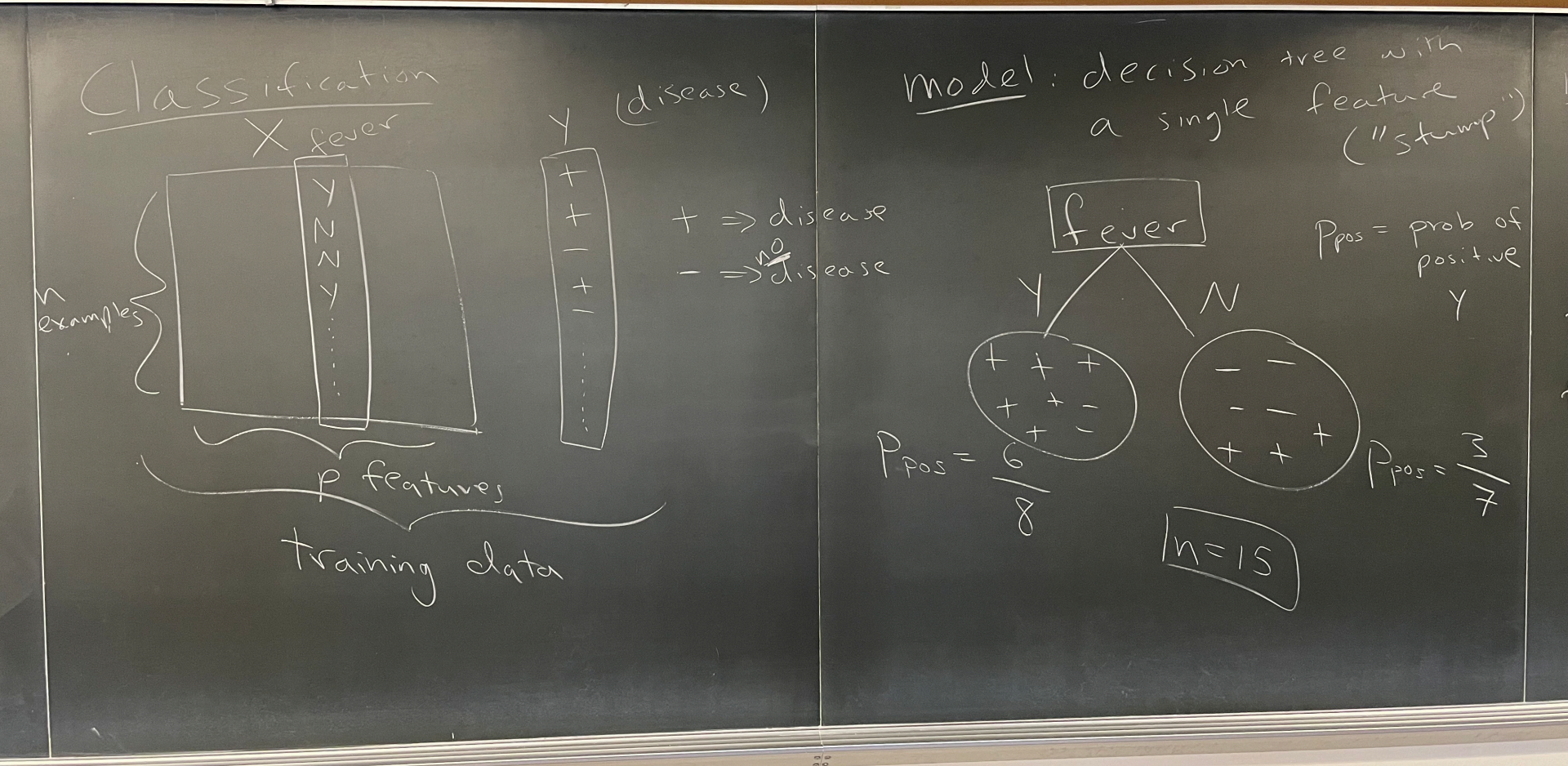
- Evaluation Metrics
  - Confusion matrices
  - Precision and recall
  - ROC curves

Binary classification examples

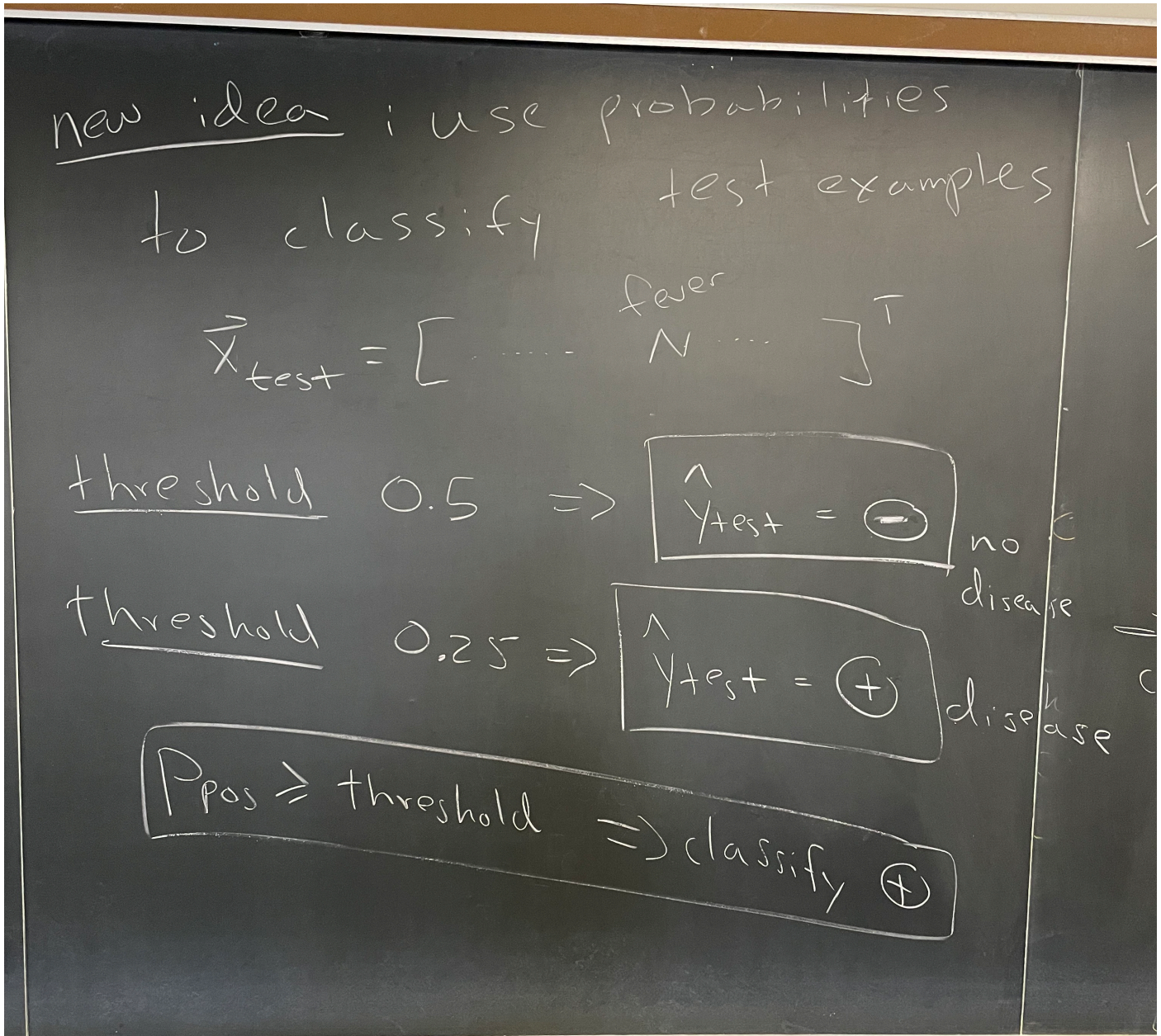
- Transactions that indicate credit card fraud
- Accounts that are bots
- Detecting which scans show tumors
- Prenatal test for Down's Syndrome
- Finding genes under natural selection
- Finding regions of the genome with high recombination rate ("hotspots")

In all these examples, we are trying to find unusual items ("needle in a haystack") -- we call these *positives*

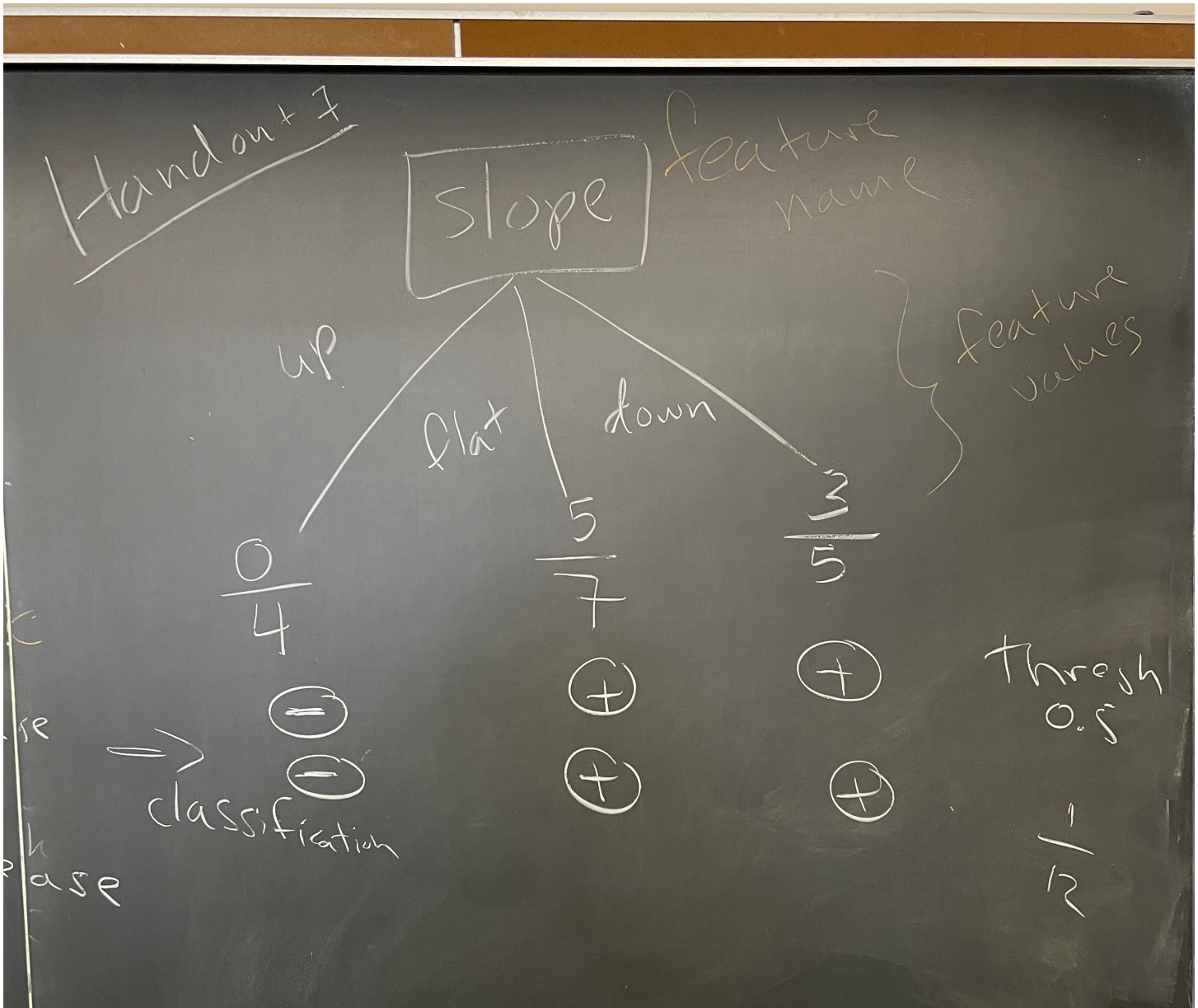
Introduction to Classification



Introduction to Classification



Handout 7


Handout 7
slope
feature name
feature values
up
flat
down
0/4
5/7
3/5
classification
through origin

Outline for September 26

- Recap SGD (stochastic gradient descent)

- Introduction to classification
  - Decision tree models
  - Probabilistic interpretation

- Evaluation Metrics
  - Confusion matrices
  - Precision and recall
  - ROC curves

Next time!