CS 260: Foundations of Data Science

Prof. Thao Nguyen

Fall 2026

HAVERFORD
COLLEGE

Materials by Sara Mathieson

Admin

- **Sit somewhere new**

- **Lab 1** grades & feedback posted on Moodle

Cost Function (mini-quiz)

$$h_w(x) = w_1 x$$

(assume $w_0=0$ for this example)

1. What is the cost function for this model?

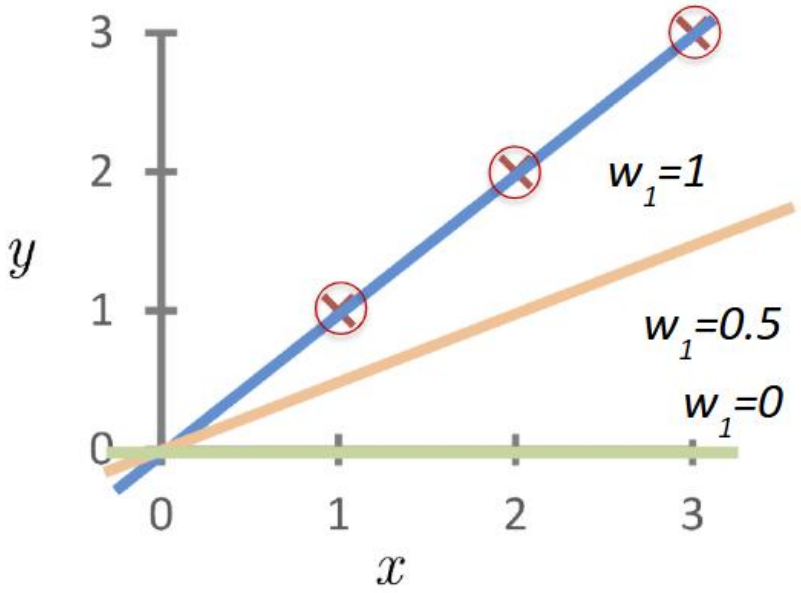
2. Given the datapoints $(x_1, y_1) = (1,1)$, $(x_2, y_2) = (2,2)$, $(x_3, y_3) = (3,3)$, compute J(0), J(0.5), and J(1)

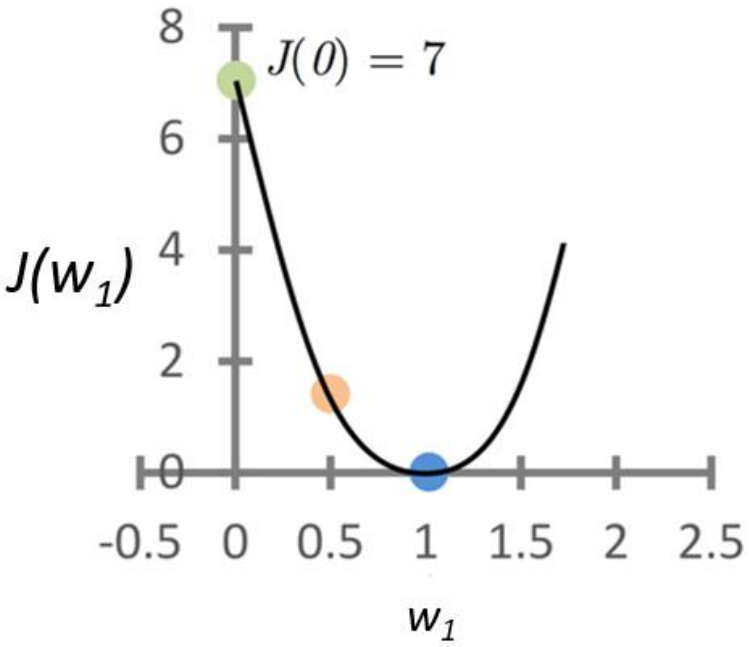
Cost Function (mini-quiz)

$h_w(x) = w_1 x$

(assume $w_0=0$ for this example)

$J(w_1)$





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

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

Outline for today

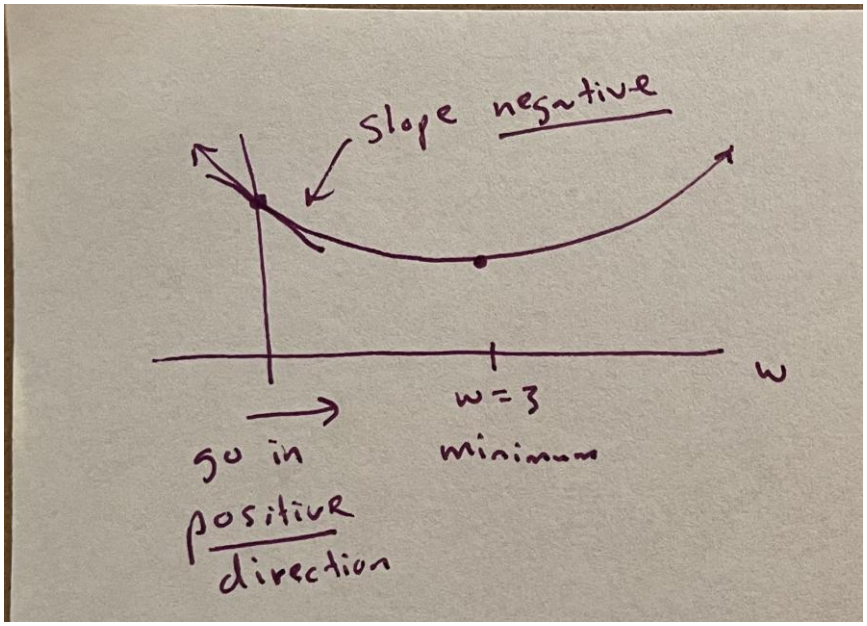
- SGD (Stochastic Gradient Descent)
- Handout 6 (SGD solution example)
- Analytic vs. SGD (pros and cons)
- (if time) Polynomial regression

Outline for today

- SGD (Stochastic Gradient Descent)
- Handout 6 (SGD solution example)
- Analytic vs. SGD (pros and cons)
- (if time) Polynomial regression

Stochastic gradient descent example

Goal: minimize the function $f(w) = w^2 - 6w + 11$



$$f(w) = w^2 - 6w + 11$$
$$= (w-3)^2 + 2$$

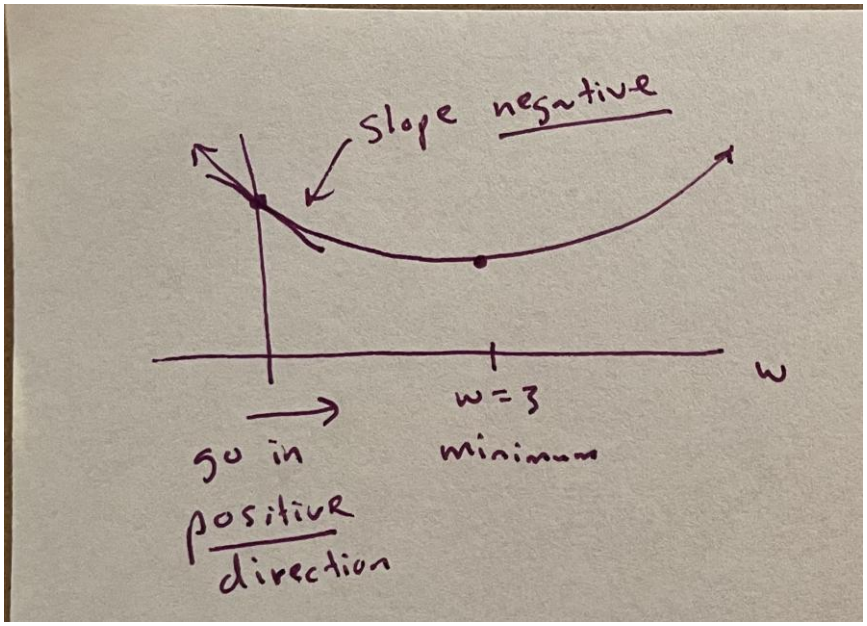
$$\boxed{f'(w) = 2w - 6}$$

Update rule: $w \leftarrow w - \alpha f'(w)$

α means step size

Stochastic gradient descent example

Goal: minimize the function $f(w) = w^2 - 6w + 11$



$$f(w) = w^2 - 6w + 11$$
$$= (w-3)^2 + 2$$

$$\boxed{f'(w) = 2w - 6}$$

$w \leftarrow 0 - 0.1(2 \times 0 - 6)$

$w \leftarrow 0 + 0.6$

$w \leftarrow 0.6$

$w \leftarrow 0.6 - 0.1(2 \times 0.6 - 6)$

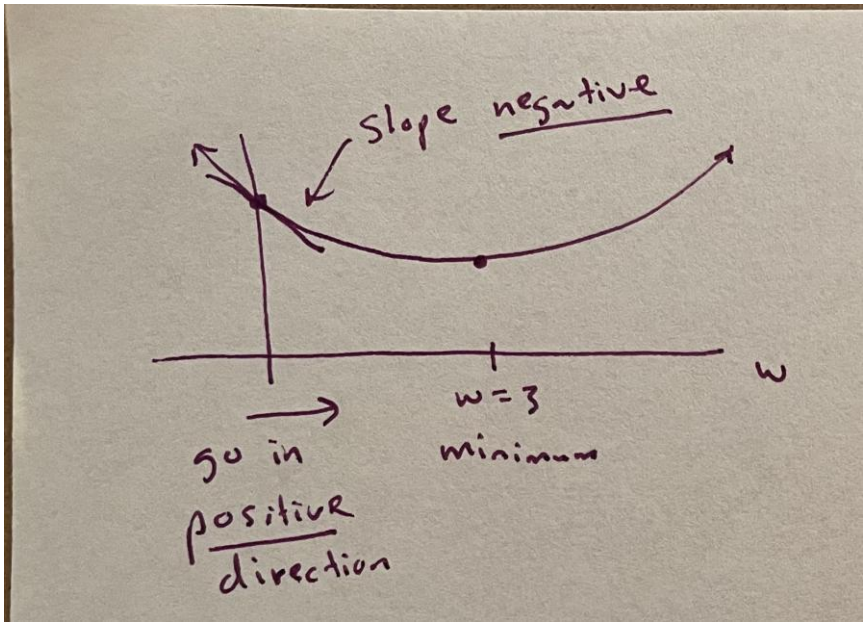
$w \leftarrow 0.6 - 0.1(-4.8)$

$w \leftarrow 1.08$

$$w \leftarrow w - \alpha f'(w)$$

Stochastic gradient descent example

Goal: minimize the function $f(w) = w^2 - 6w + 11$



$$f(w) = w^2 - 6w + 11$$
$$= (w-3)^2 + 2$$

$$\boxed{f'(w) = 2w - 6}$$

$$w \leftarrow 0 - 0.1(2 \times 0 - 6)$$
$$w \leftarrow 0 + 0.6$$
$$w \leftarrow 0.6$$

$$w \leftarrow 0.6 - 0.1(2 \times 0.6 - 6)$$
$$w \leftarrow 0.6 - 0.1(-4.8)$$
$$w \leftarrow 1.08$$

Stop when: $|f(w^t) - f(w^{t-1})| < \varepsilon$

For example:
$\varepsilon = 1 \times 10^{-8}$

Stochastic Gradient Descent for Linear Regression

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

$$J(\vec{w}) = \frac{1}{2}\sum_{i=1}^{n}(\vec{w}\cdot\vec{x}_i - y_i)^2$$

Gradient with respect to one data point (i.e., $\vec{x}_i$):

$$\nabla J_{\vec{x}_i} = \frac{\partial J(\vec{w})_{\vec{x}_i}}{\partial \vec{w}} = (\vec{w}\cdot\vec{x}_i - y_i)\vec{x}_i$$

Stochastic Gradient Descent for Linear Regression

Note: The iteration is also called the "epoch"

for iteration t:

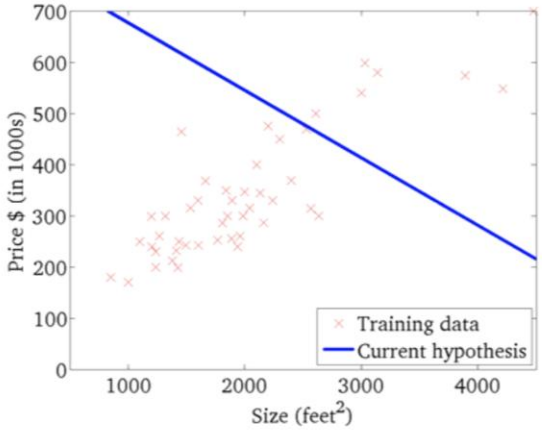
Note: We usually shuffle the order of the data points

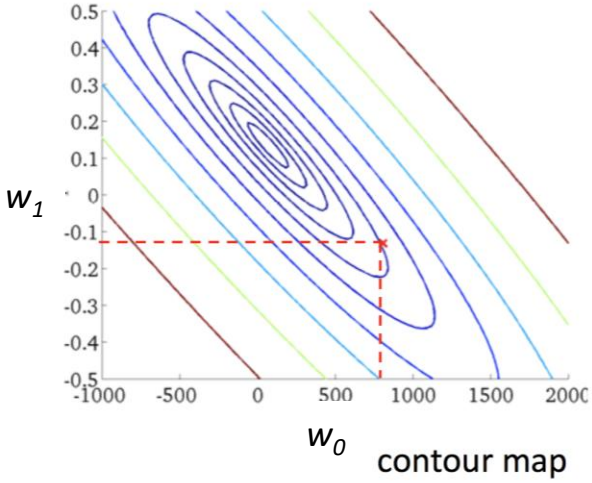
for i=1, 2, 3…n:

$$\vec{w} \leftarrow \vec{w} - \alpha(\vec{w} \cdot \vec{x}_i - y_i)\vec{x}_i$$

check for convergence: $|J(\vec{w}^t) - J(\vec{w}^{t-1})| < \varepsilon$

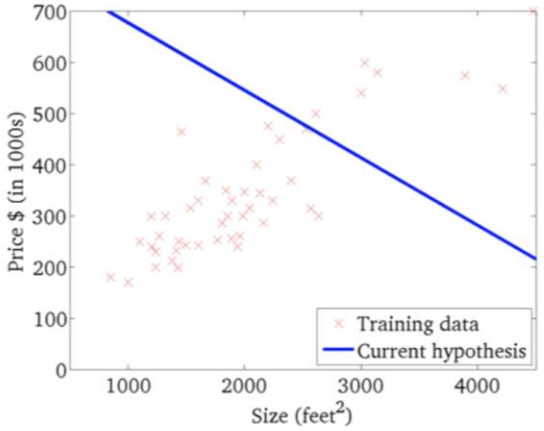
Linear Model and Cost Function J

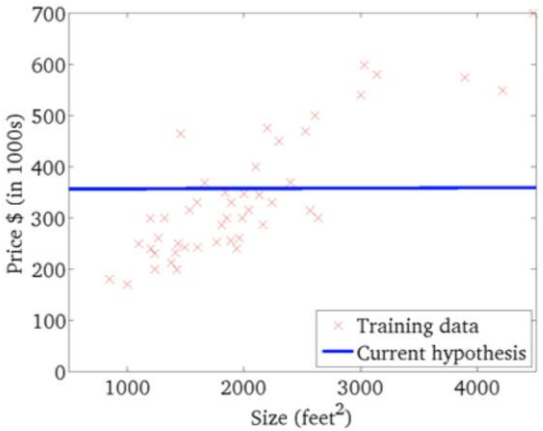


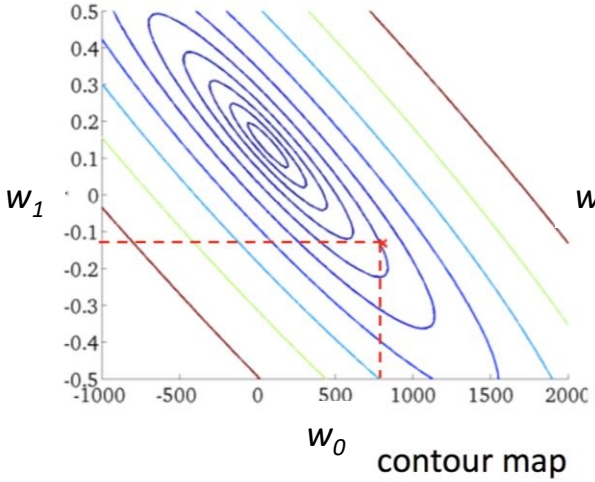


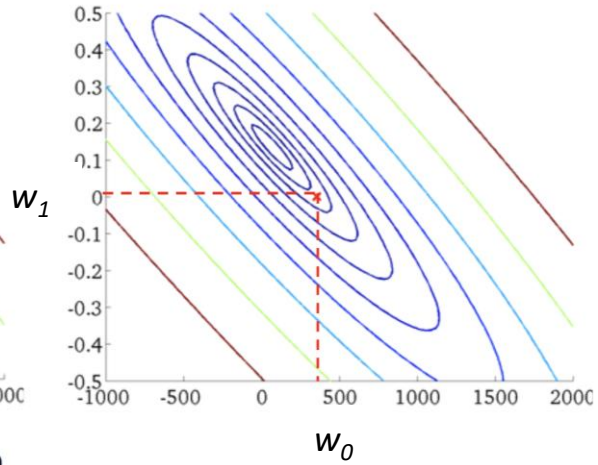
Slide: modified from Jessica Wu
Based on slide by Andrew Ng

Linear Model and Cost Function J



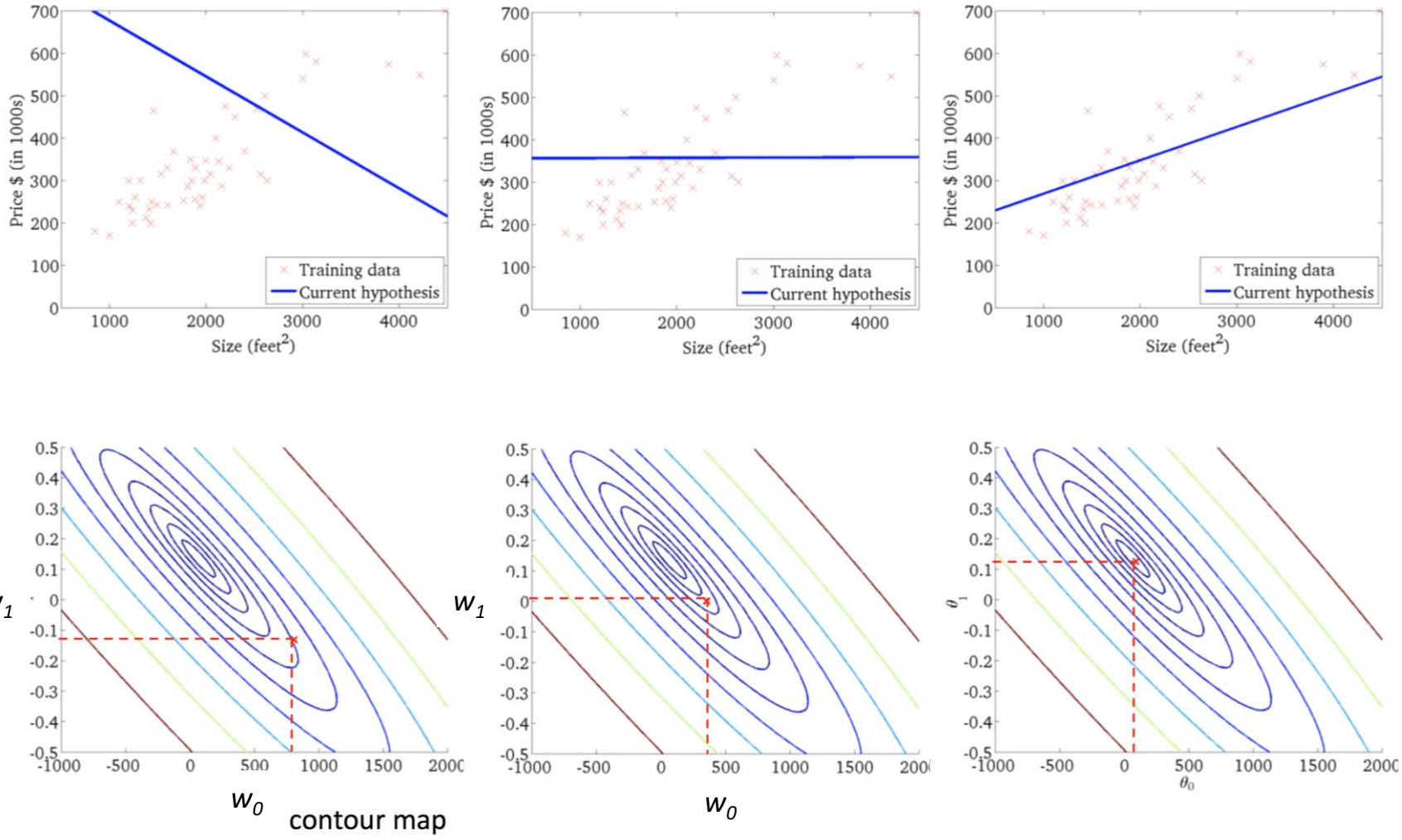






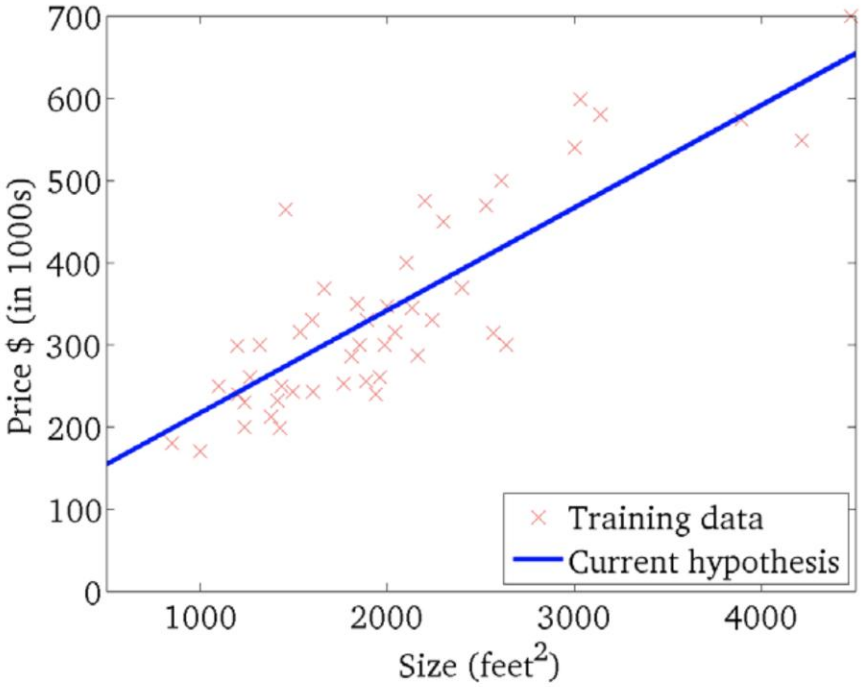
Slide: modified from Jessica Wu
Based on slide by Andrew Ng

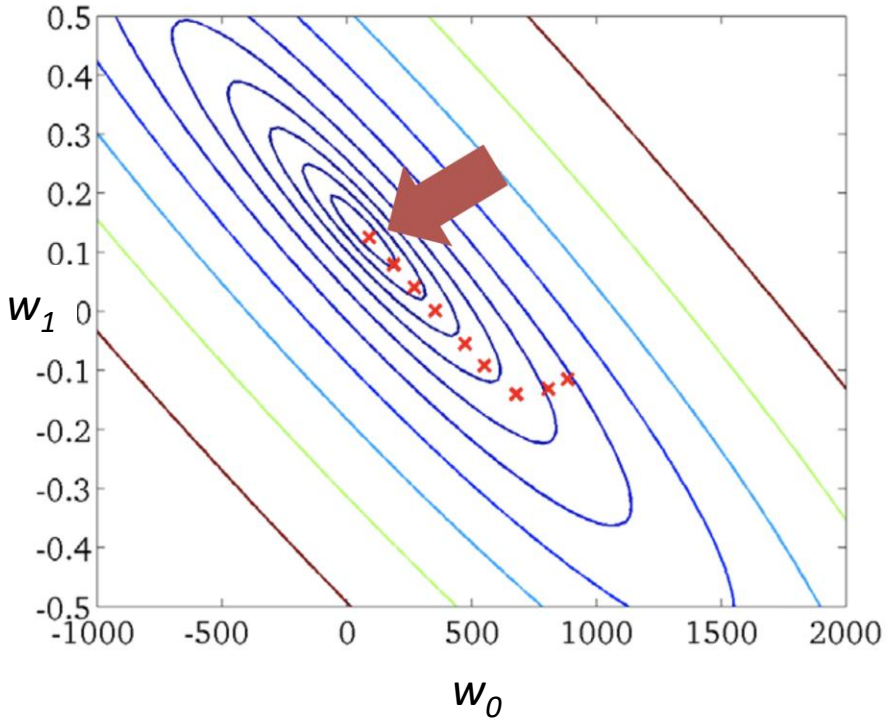
Linear Model and Cost Function J



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

Gradient Descent: walking toward the minimum



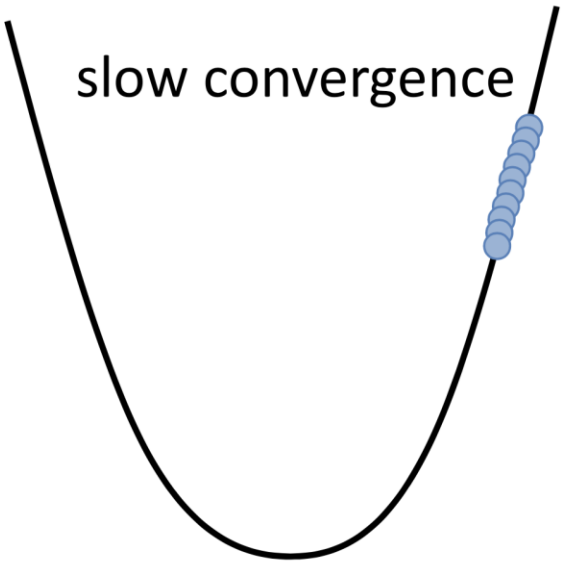


Slide: modified from Jessica Wu
Based on slide by Andrew Ng

Choosing the step size alpha

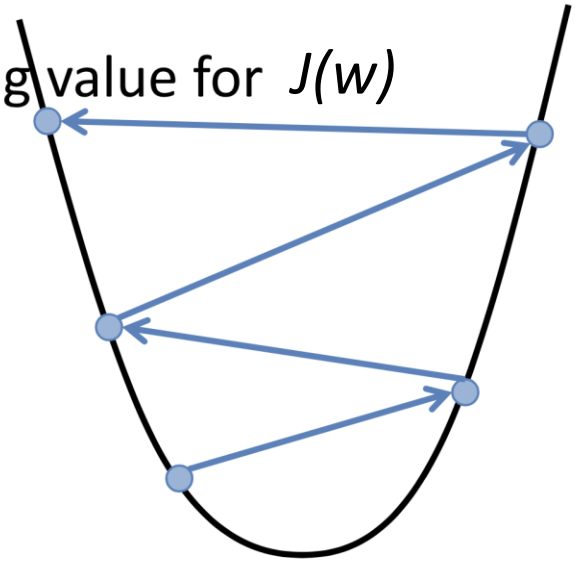
α too small

slow convergence



α too large

increasing value for $J(w)$



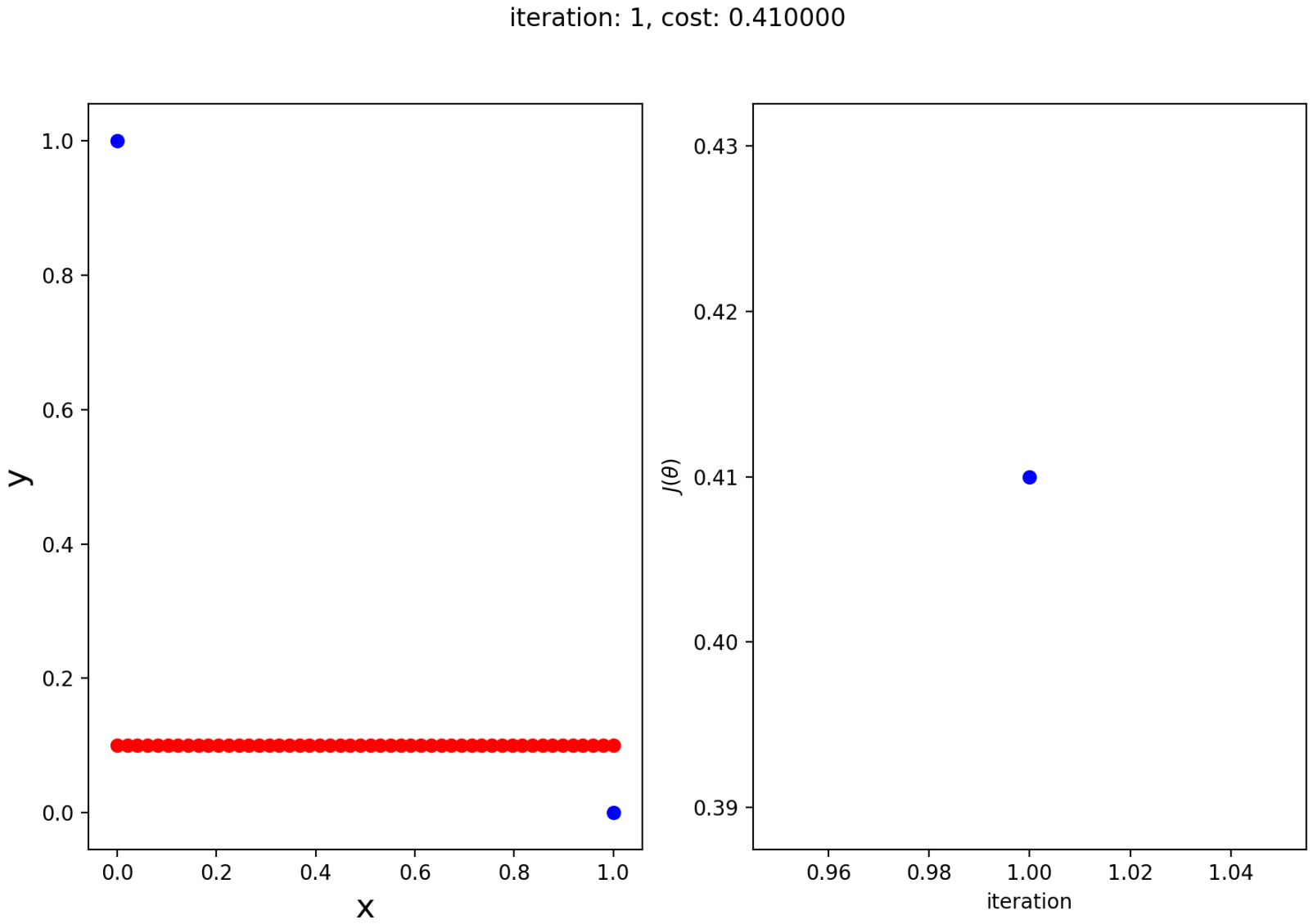
- may overshoot minimum
- may fail to converge (may even diverge)

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

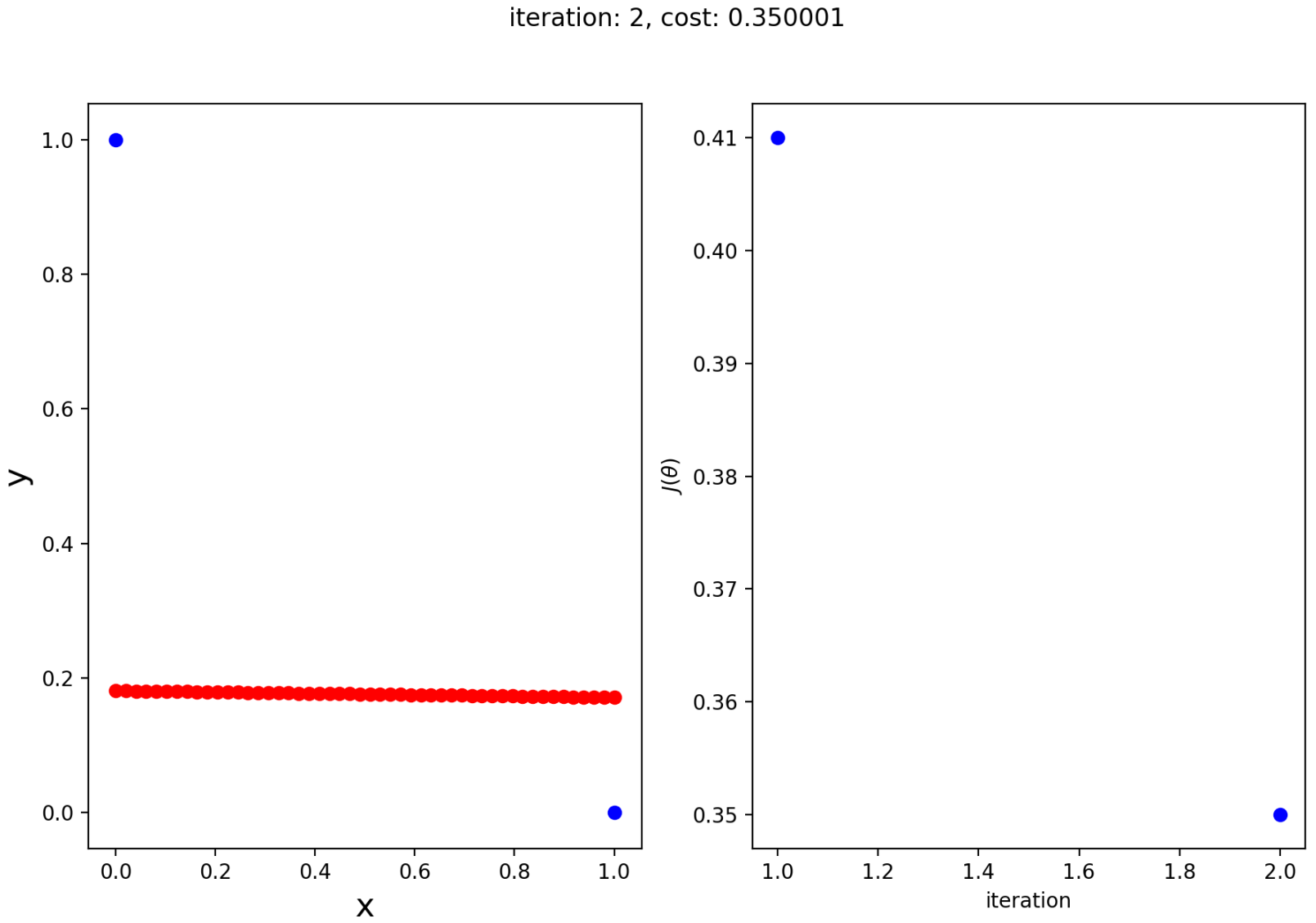
SGD with our small dataset from the handouts

Note: this is with the original order of the points

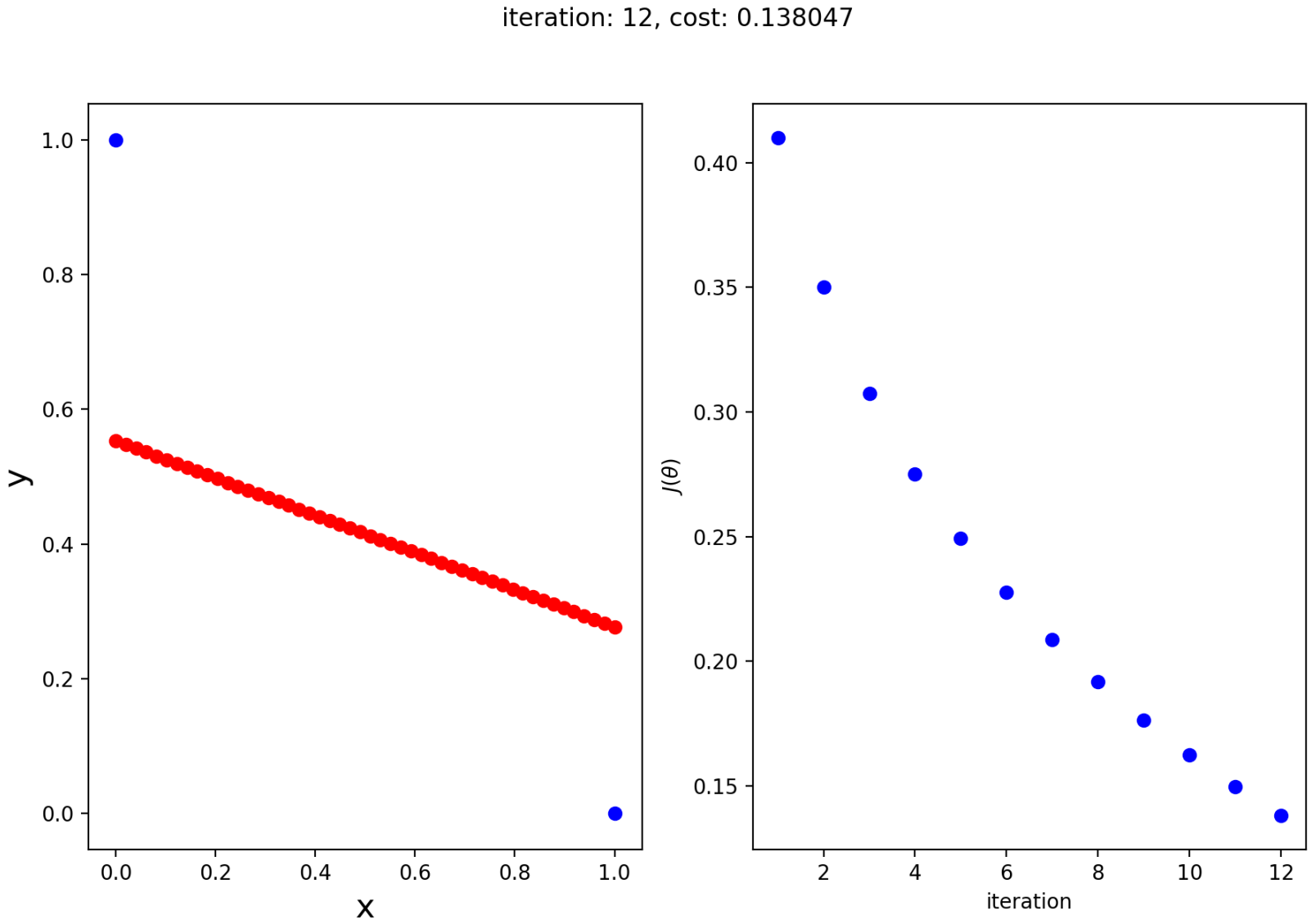
Small example, iteration 1



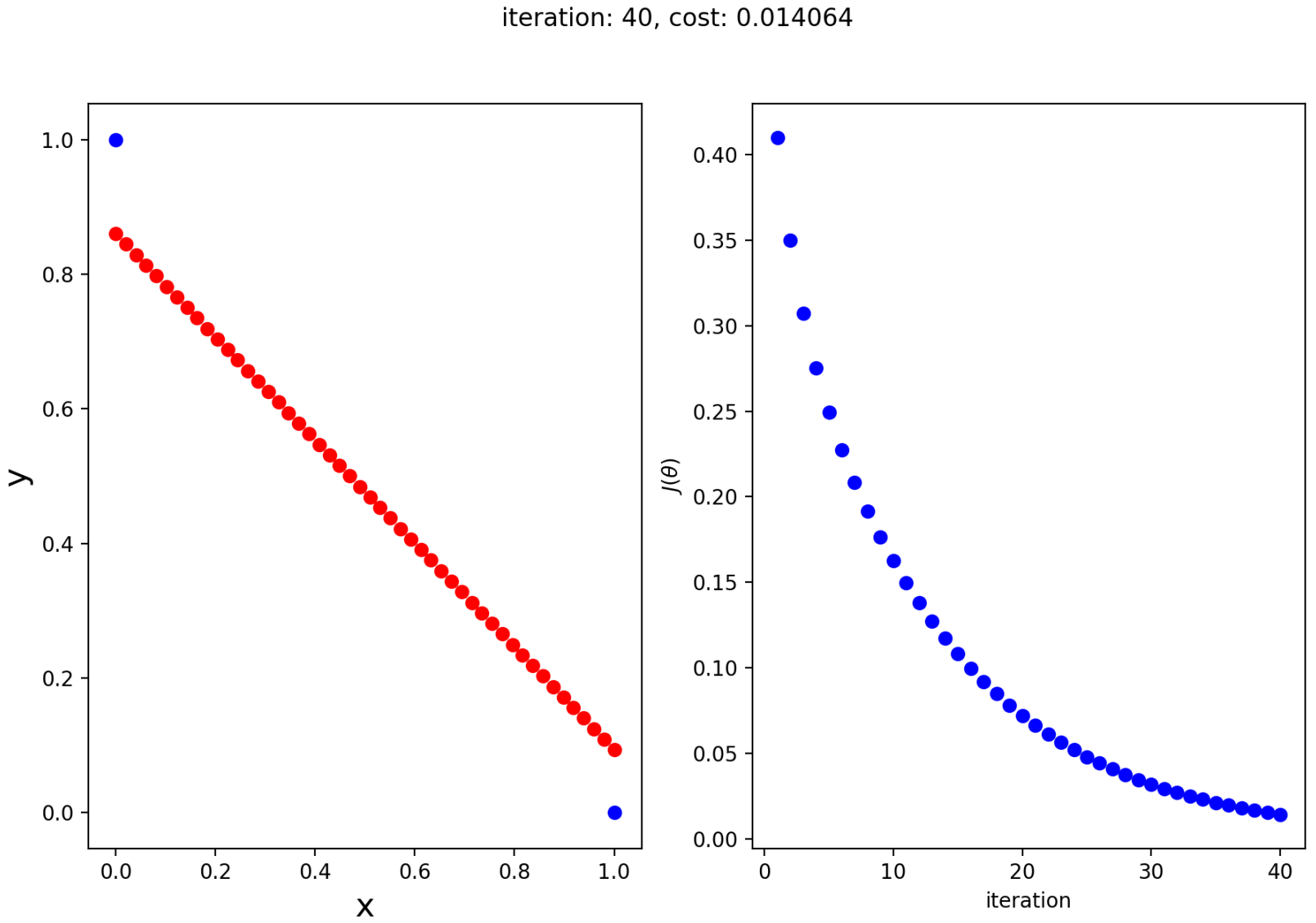
Small example, iteration 2



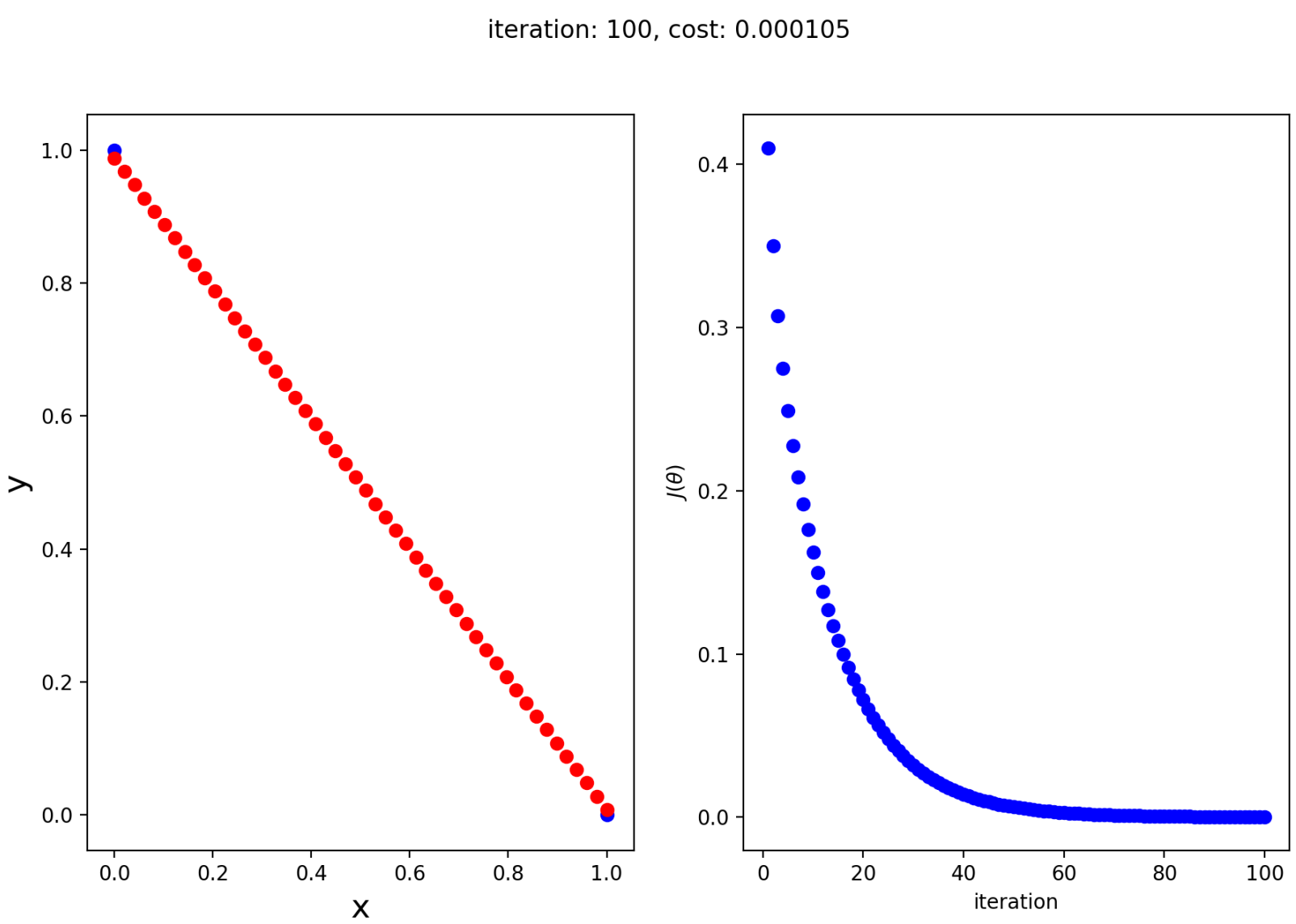
Small example, iteration 12



Small example, iteration 40



Small example, iteration 100



Outline for today

- SGD (Stochastic Gradient Descent)
- Handout 6 (SGD solution example)
- Analytic vs. SGD (pros and cons)
- (if time) Polynomial regression

Handout 6

Handout 6

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

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 just the first data point is used to update the gradient?

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

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

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

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.

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

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

Handout 6

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?

X $\quad$ $\vec{w}$

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

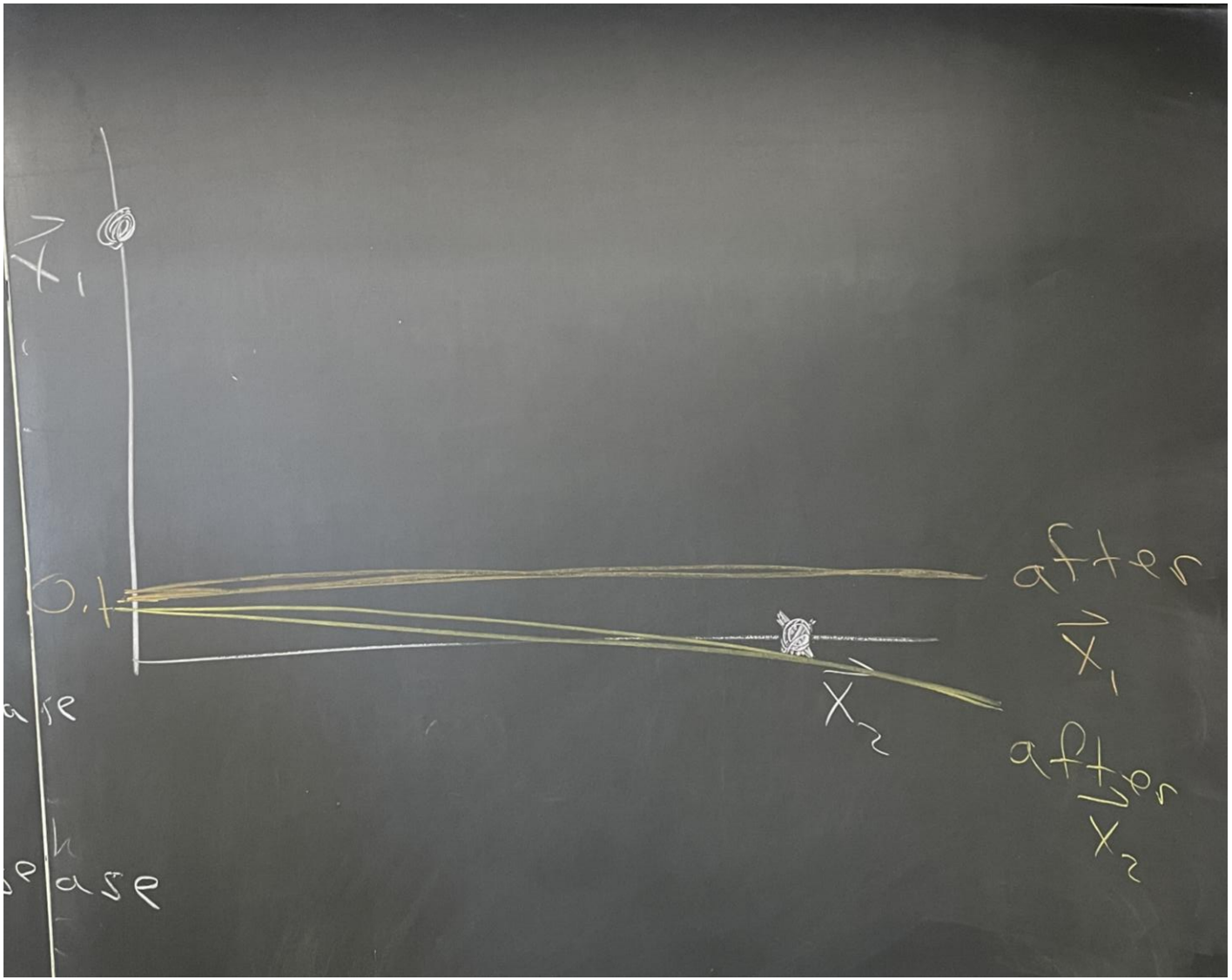
$$\vec{y} - \hat{y} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} - \begin{bmatrix} 0.09 \\ 0.08 \end{bmatrix} = \begin{bmatrix} 0.91 \\ -0.08 \end{bmatrix}$$

residuals

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

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

Handout 6 (#4)



Outline for today

- SGD (Stochastic Gradient Descent)
- Handout 6 (SGD solution example)
- Analytic vs. SGD (pros and cons)
- (if time) Polynomial regression

Pros and Cons

Gradient Descent

- Requires multiple iterations
- Need to choose α
- Works well when p is large
- Can support online learning

(normal equation)

Analytic Solution

- 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?

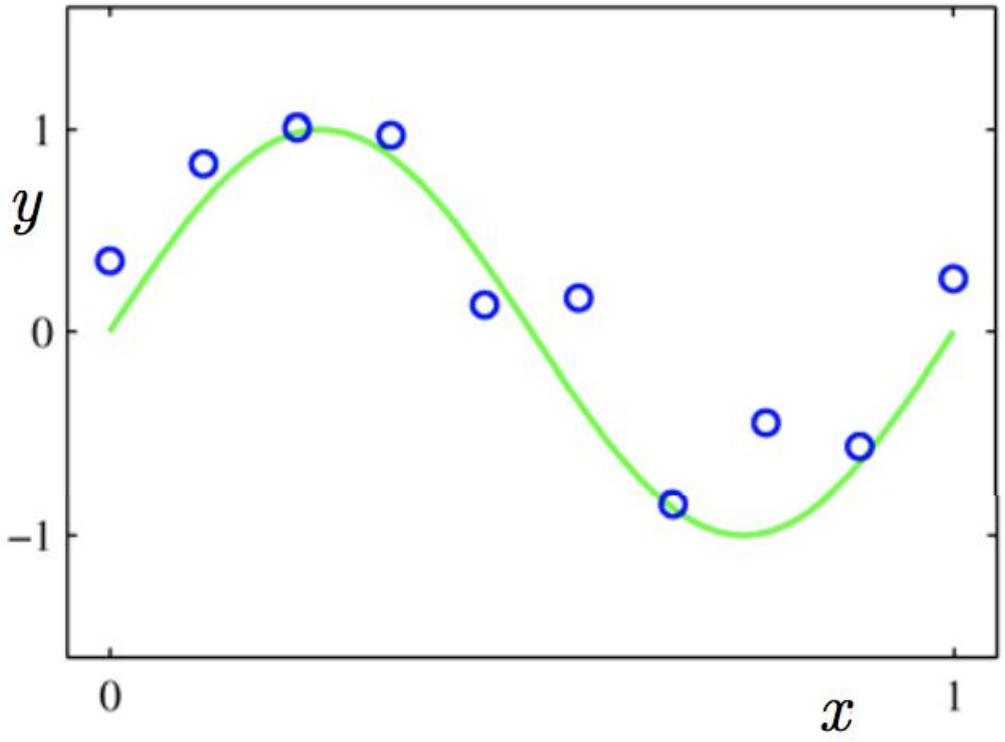
Challenge outside of class!

Outline for today

- SGD (Stochastic Gradient Descent)
- Handout 6 (SGD solution example)
- Analytic vs. SGD (pros and cons)
- (if time) Polynomial regression

Polynomial Regression

- Can be thought of as regular linear regression with a change of basis



Plot: Jessica Wu

Polynomial Regression

$$\boldsymbol{X} = \begin{bmatrix} x_1^0 & x_1^1 & x_1^2 & \cdots & x_1^d \\ & & \vdots & & \\ x_n^0 & x_n^1 & x_n^2 & \cdots & x_n^d \end{bmatrix}$$