CS 260: Foundations of Data Science

Prof. Sara Mathieson

Spring 2025

HAVERFORD COLLEGE

Admin

- **Lab 3** due Wednesday night
- **Next TA Hours** TODAY 8-10pm in H110
- **Next Office Hours** Monday 12-1pm in L302
- **Video on** if possible!

Outline

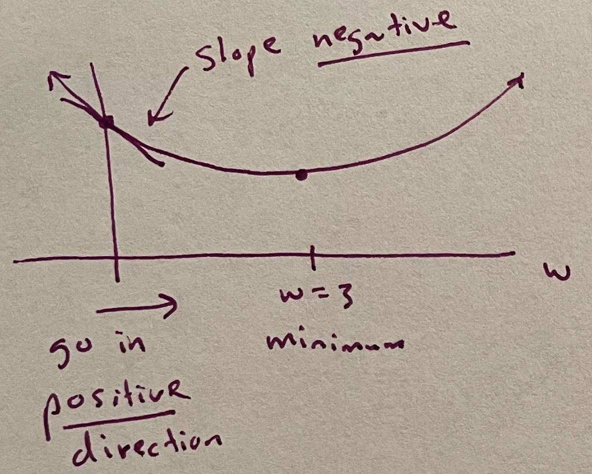
- SGD (Stochastic Gradient Descent)
- Handout 6 (SGD solution example)
- Analytic vs. SGD (pros and cons)
- (if time) Polynomial regression

Outline

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

Small example from class on Tues

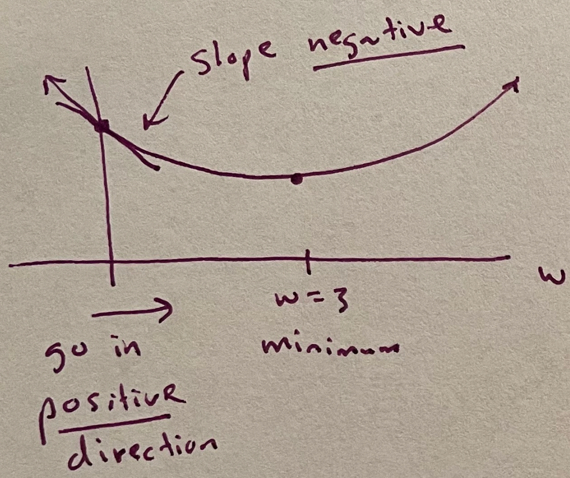
Goal: minimize the function $f(w) = w^2 - 6w + 11$



$$f(w) = (w-3)^2 + 2$$
$$= w^2 - 6w + 11$$
$$\Rightarrow \boxed{f'(w) = 2w - 6}$$

Small example from class on Tues

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



$$f(w) = (w-3)^2 + 2$$
$$= w^2 - 6w + 11$$
$$\Rightarrow \boxed{f'(w) = 2w - 6}$$

① $w \leftarrow 0 - 0.1(2 \cdot 0 - 6)$

$w \leftarrow 0 + 0.6$

$\boxed{w \leftarrow 0.6}$

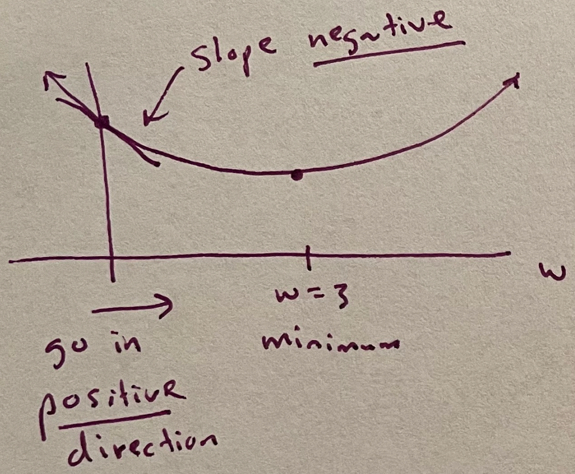
② $w \leftarrow 0.6 - 0.1(2 \cdot 0.6 - 6)$

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

$\boxed{w \leftarrow 1.08}$

Small example from class on Tues

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



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

$$= w^2 - 6w + 11$$

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

① $w \leftarrow 0 - 0.1(2 \cdot 0 - 6)$

$w \leftarrow 0 + 0.6$

$\boxed{w \leftarrow 0.6}$

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

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

$\boxed{w \leftarrow 1.08}$

Stop when:

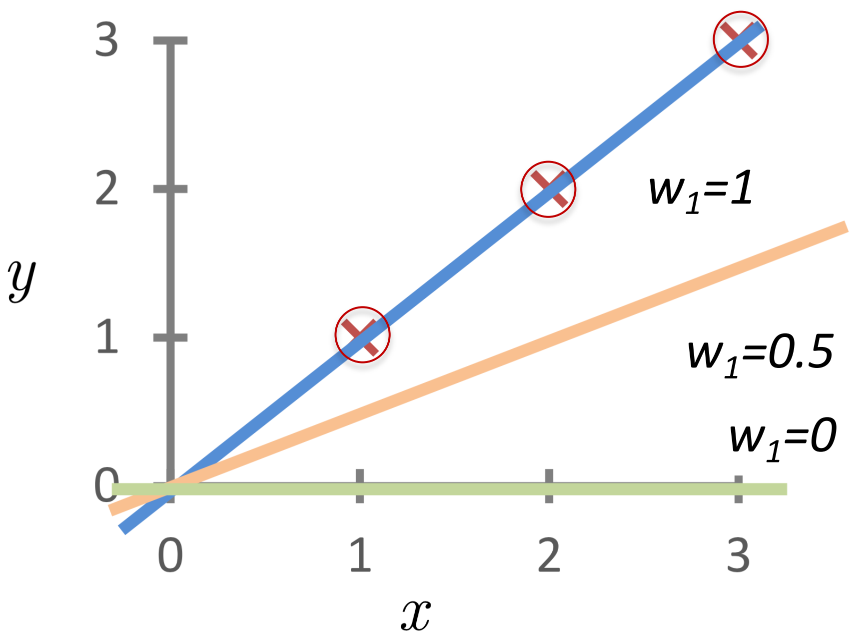
$$|f(w^t) - f(w^{t-1})| < \varepsilon, \quad \varepsilon = 1 \times 10^{-8}$$

(for example)

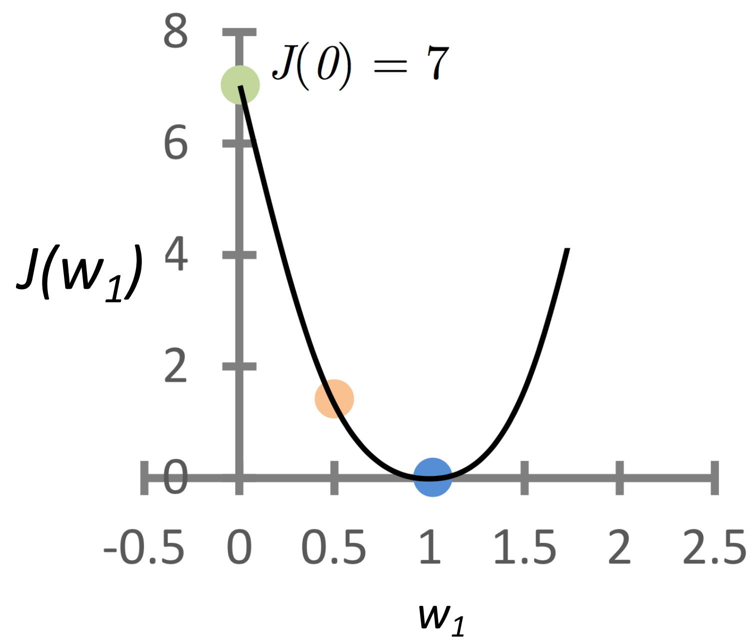
Cost Function (mini-quiz)

$h_w(x) = w_1 x$

(assume $w_0=0$ for this example)



$J(w_1)$



TODO: Compute/verify the cost function $J(w_1)$ for

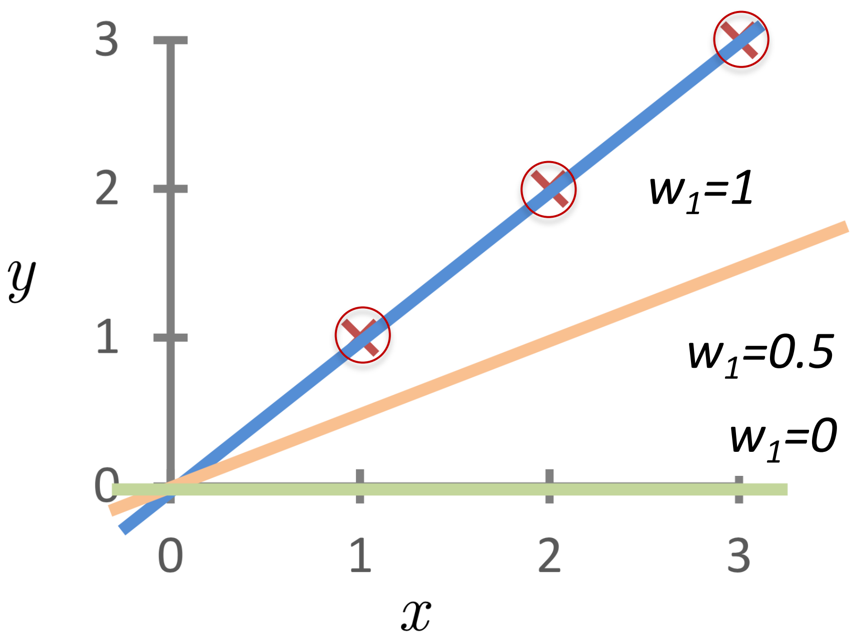
- $w_1 = 0$
- $w_1 = 0.5$
- $w_1 = 1$

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

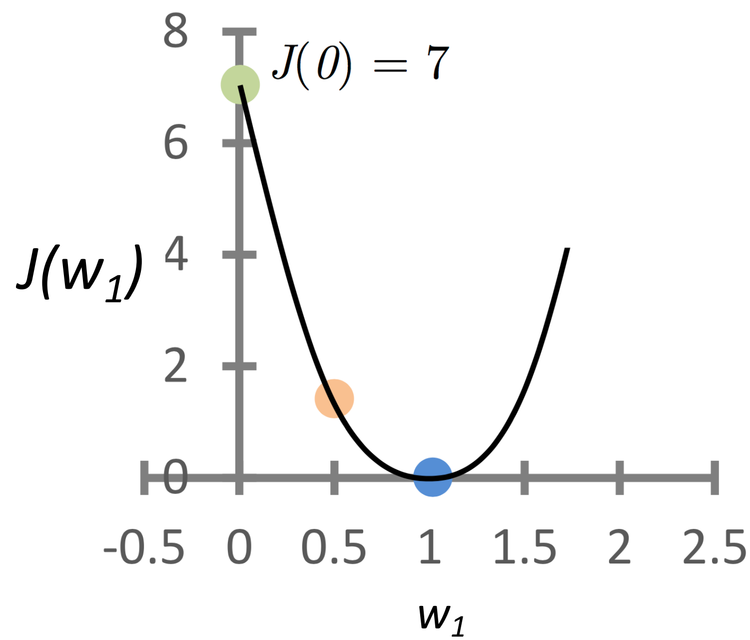
Cost Function (mini-quiz)

$h_w(x) = w_1 x$

(assume w_0=0 for this example)



$J(w_1)$



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

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

Stochastic Gradient Descent for Linear Regression

$J(\vec{w}) = \frac{1}{2}\sum (y_i - \hat{y}_i)^2$ $\qquad \hat{y} = \vec{w}\cdot\vec{x}$

$\nabla J_{x_i} = (y_i - \vec{w}\cdot\vec{x}_i)(-\vec{x}_i)$

$\vec{w} = \vec{0}$

for t in range(T):

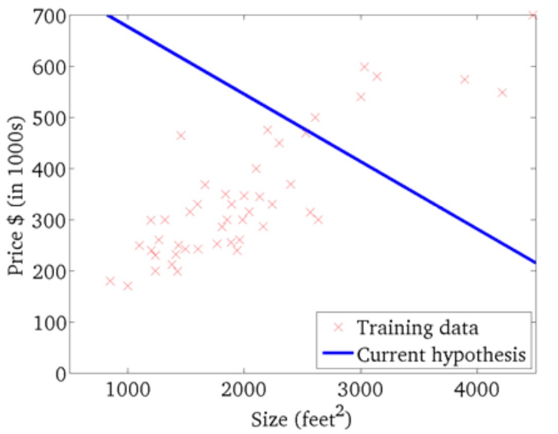
  for $i = 1, 2, \dots n$ shuffled

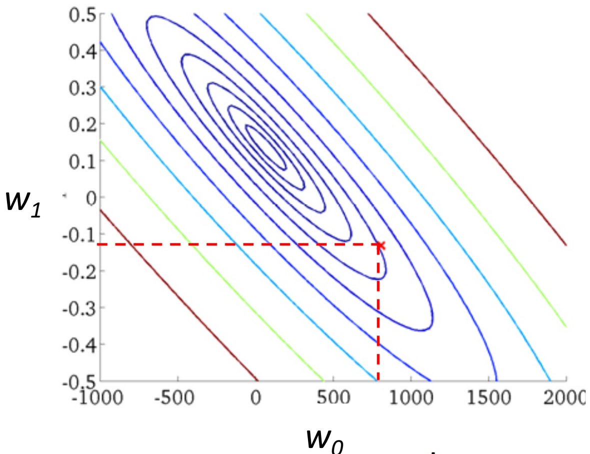
    $\vec{w} \leftarrow \vec{w} - \alpha(\vec{w}\cdot\vec{x}_i - y_i)\,\vec{x}_i$

converge?

$|J(w_{prev}) - J(w_{curr})| < \varepsilon$

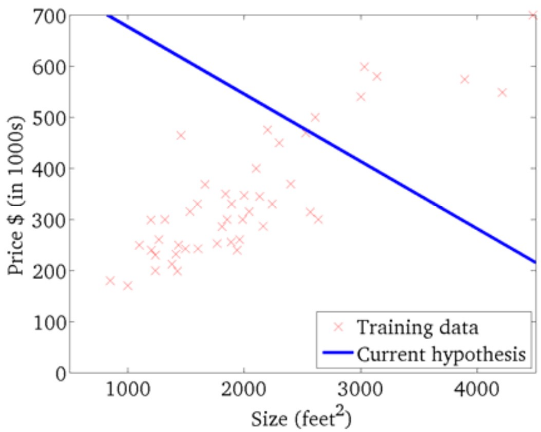
Linear Model and Cost Function J

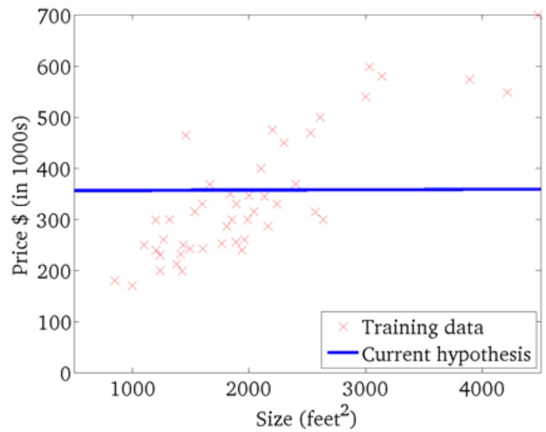




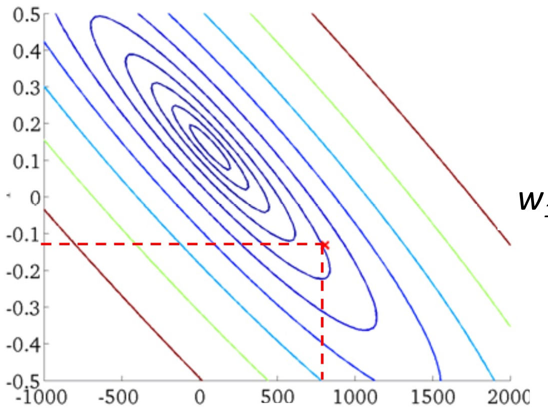
contour map

Linear Model and Cost Function J





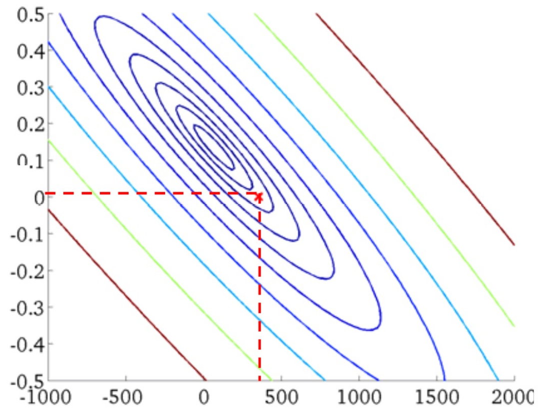
w_1



w_0

contour map

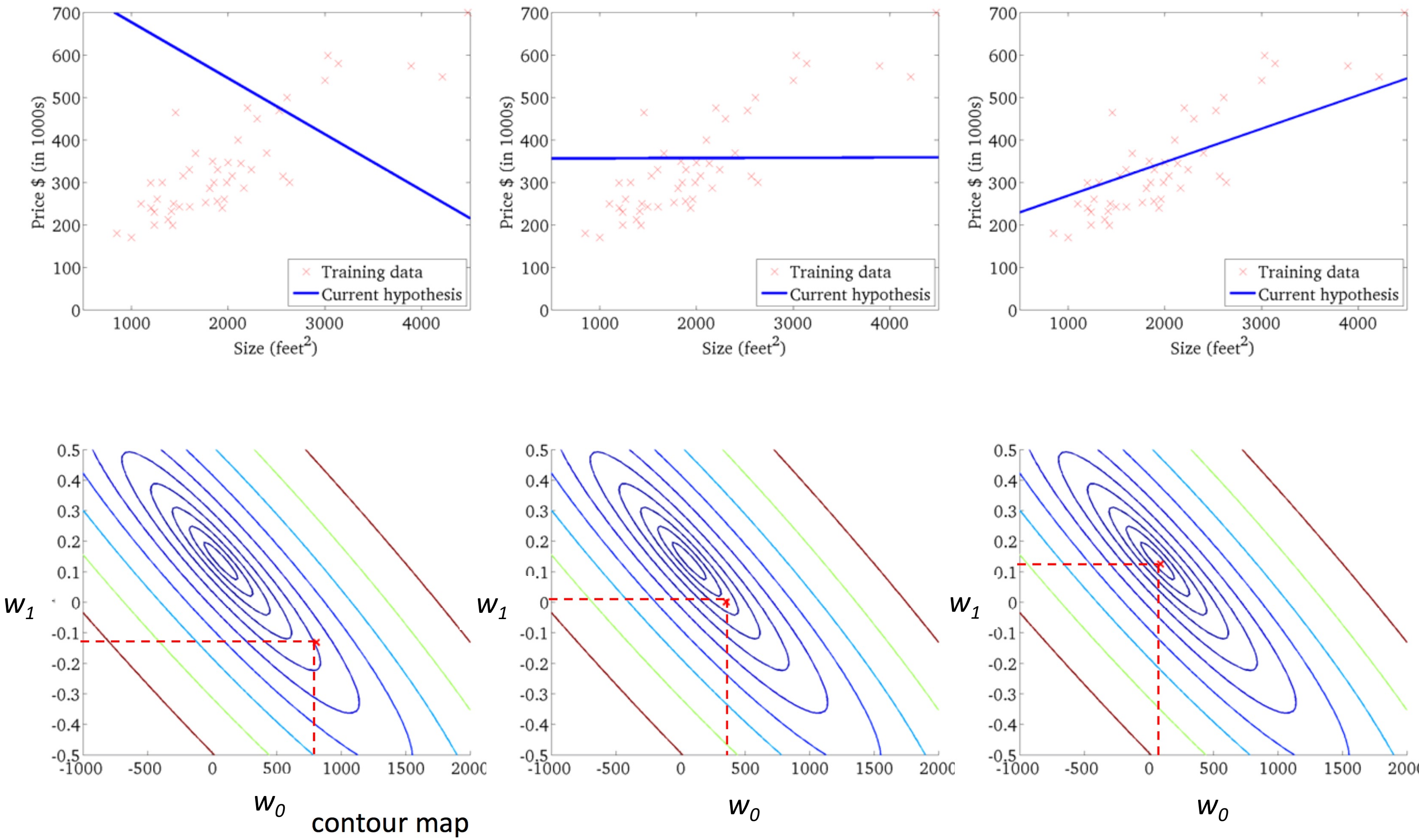
w_1



w_0

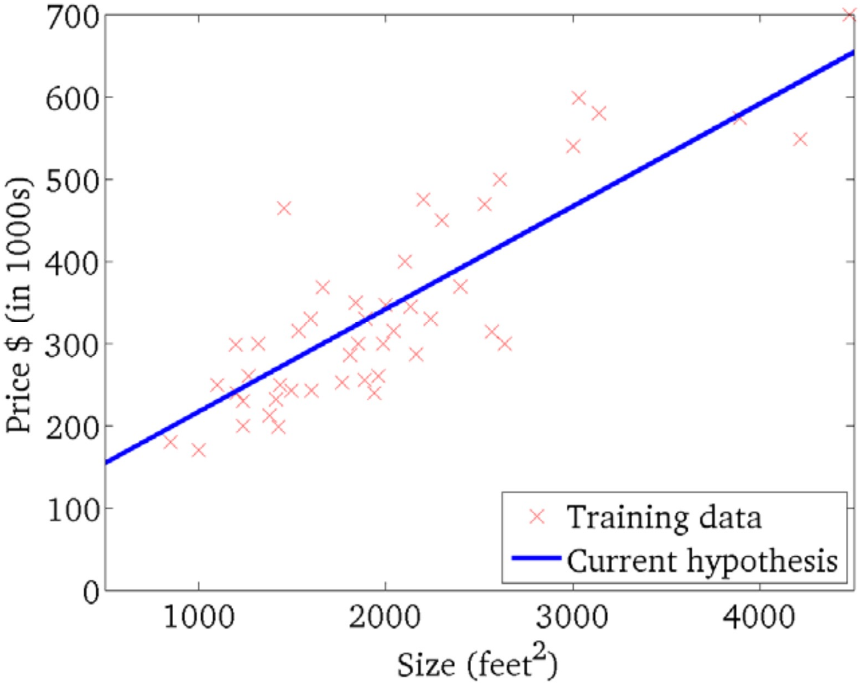
Slide: modified from Jessica Wu
Based on slide by Andrew Ng

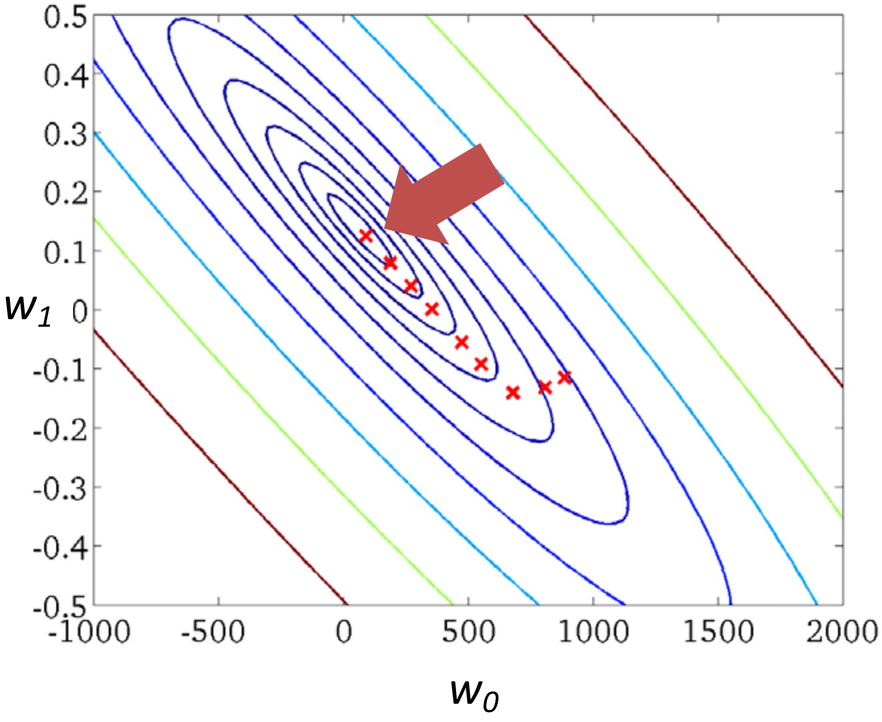
Linear Model and Cost Function J



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

Gradient Descent: walking toward the minimum

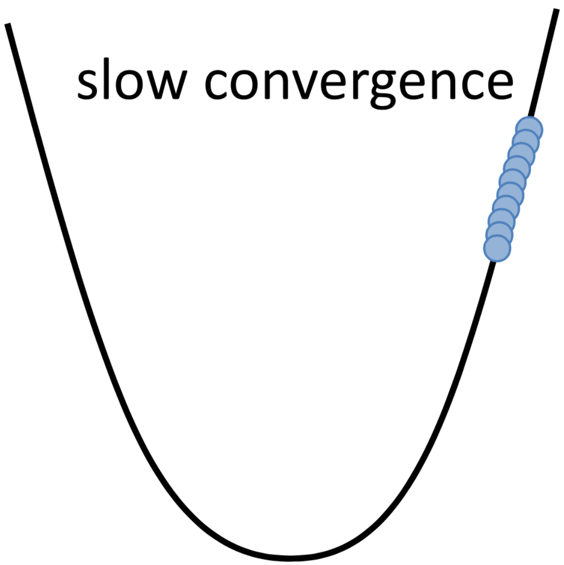




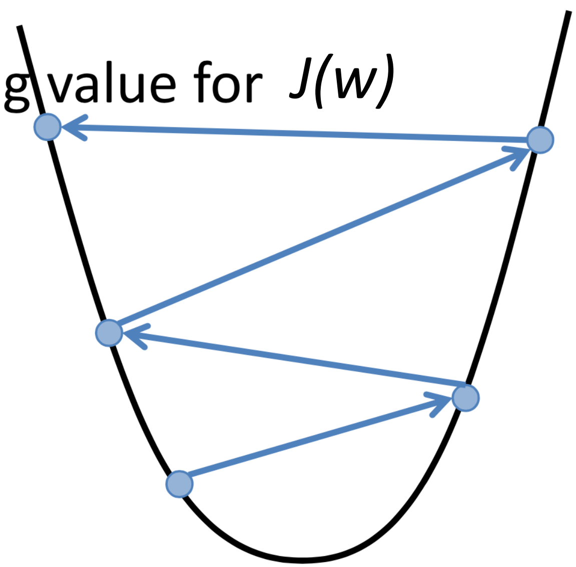
Slide: modified from Jessica Wu
Based on slide by Andrew Ng

Choosing the step size alpha

α too small



α too large



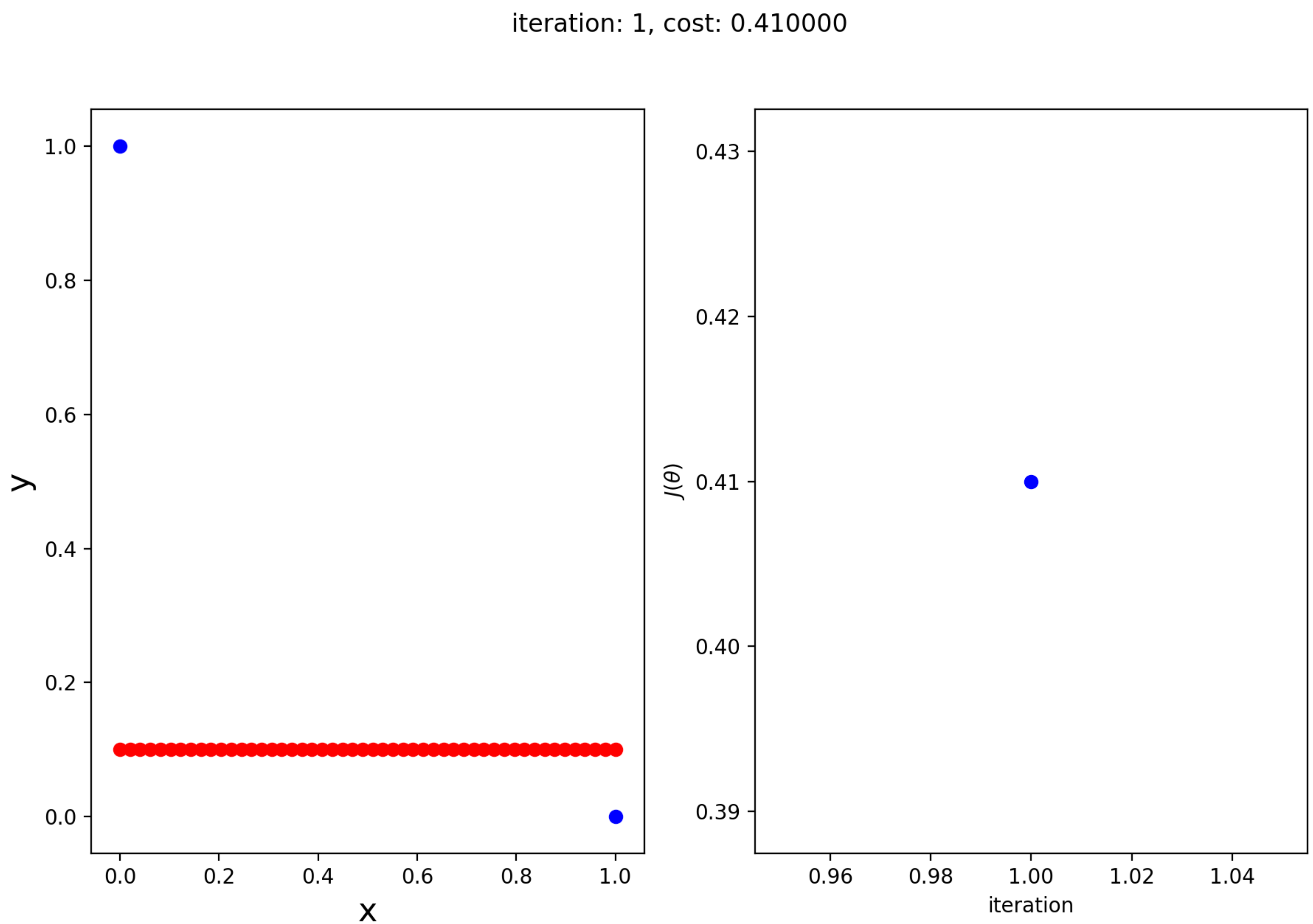
- may overshoot minimum
- may fail to converge (may even diverge)

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

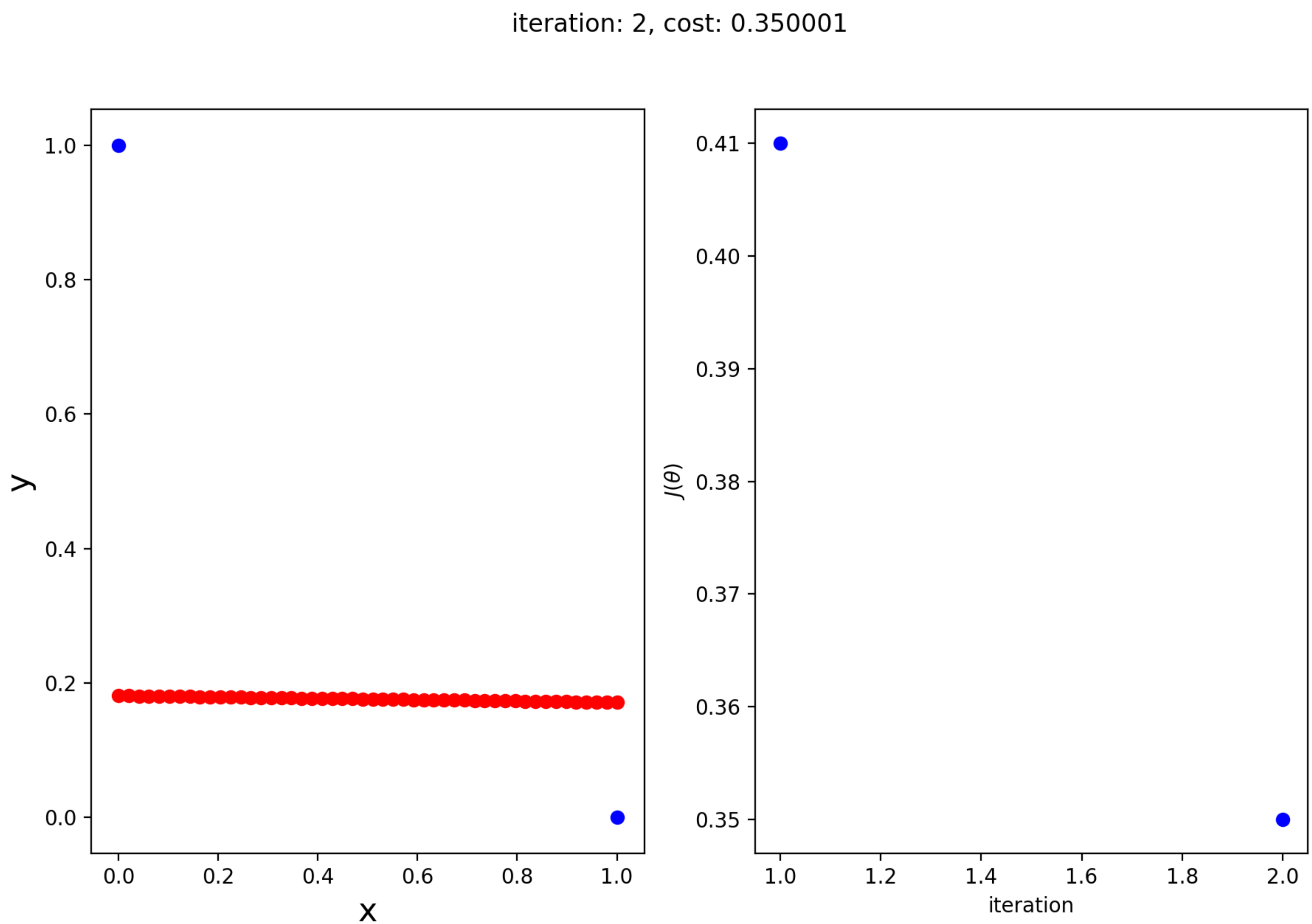
SGD with our small dataset from the handouts

Note: this is with the original order of the points

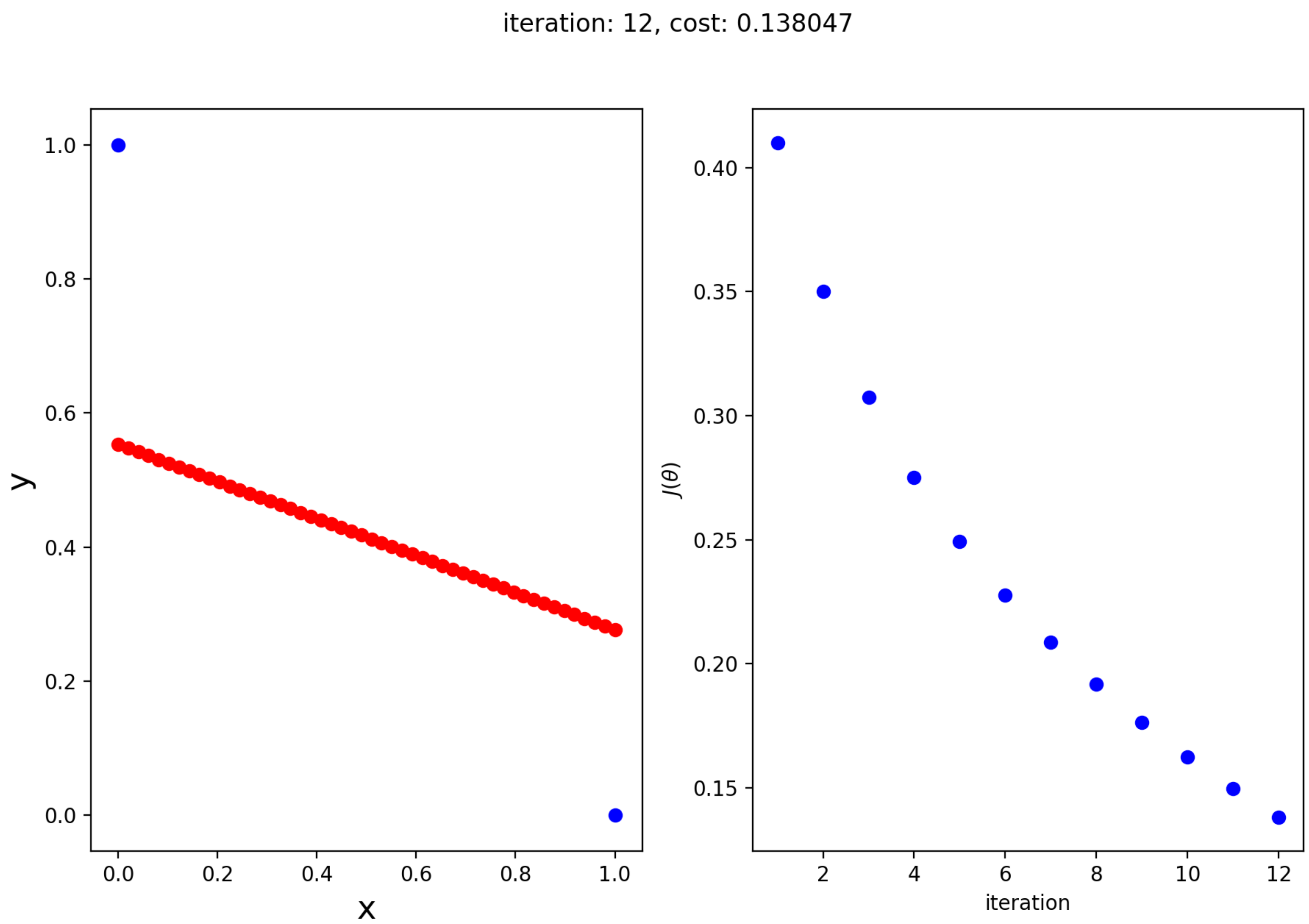
Small example, iteration 1



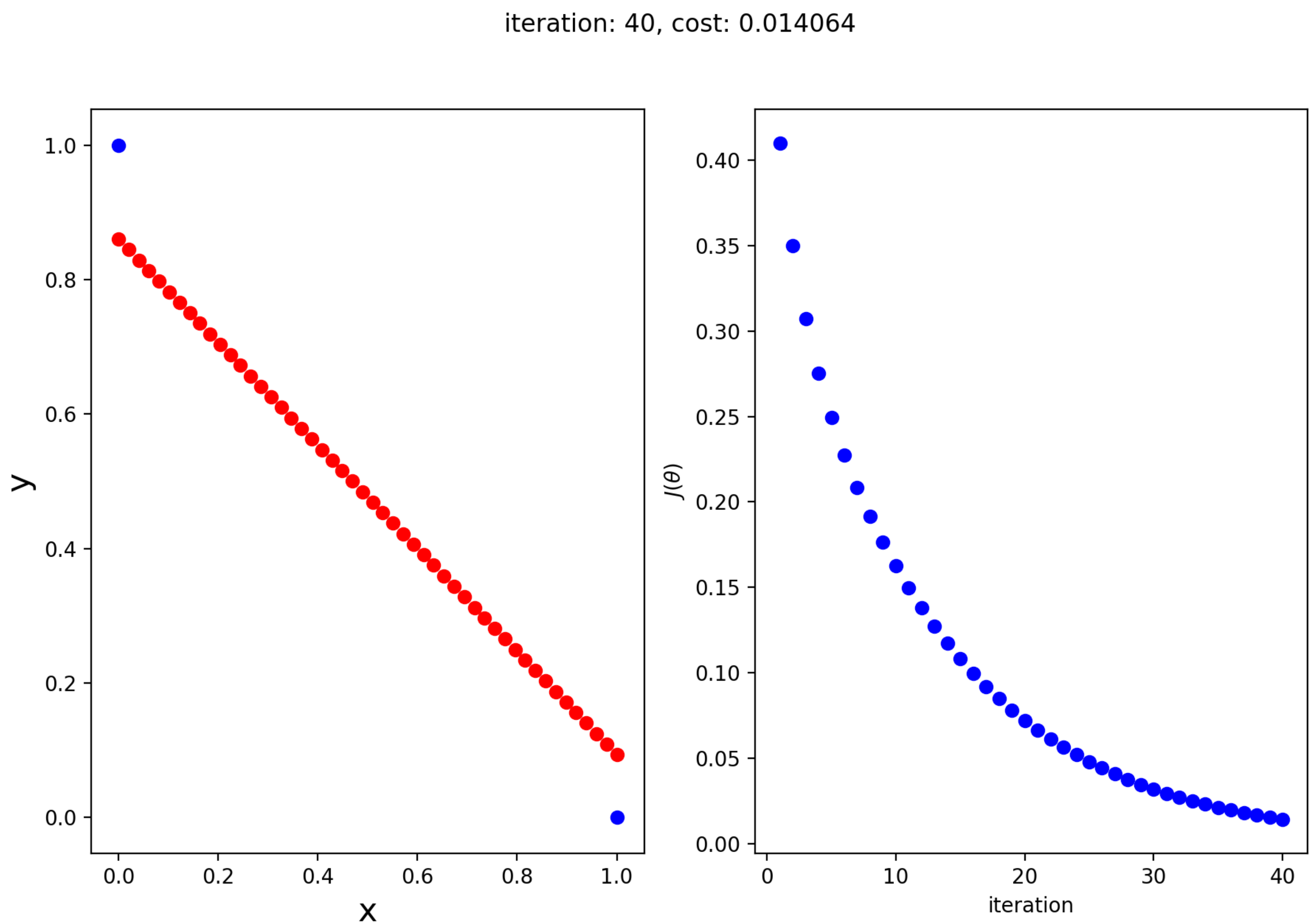
Small example, iteration 2



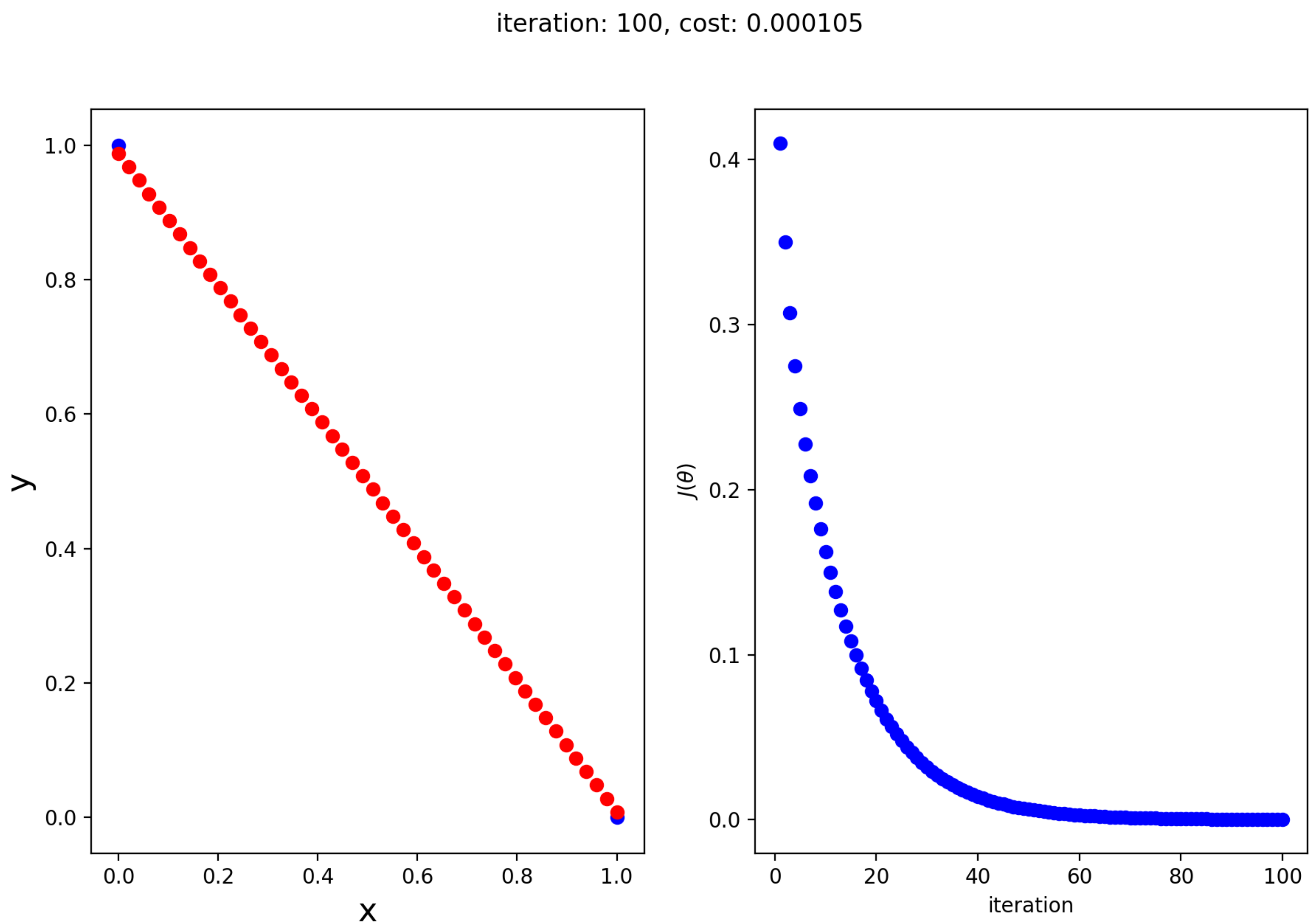
Small example, iteration 12



Small example, iteration 40



Small example, iteration 100



Outline

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

Handout 6

$\boxed{a \cdot b = a^T b}$

$$\vec{w} \leftarrow \begin{bmatrix} 0 \\ 0 \end{bmatrix} - \alpha \left(\overbrace{\vec{w} \cdot \vec{x_i}}^{\vec{w}^T \vec{x}_i} - y \right) \vec{x}_i$$

$$x_i = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

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

col of 1's

x_1 + x_2

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? Si

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?

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

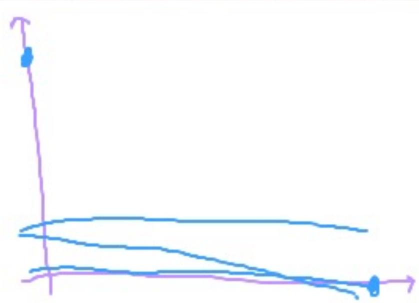
$$\vec{w} \leftarrow \begin{bmatrix} 0 \\ 0 \end{bmatrix} + 0.1 \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

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

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.

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

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



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?

$$\hat{y} = \underbrace{\begin{bmatrix}1 & 0\\1 & 1\end{bmatrix}}_{X}\underbrace{\begin{bmatrix}0.09\\-0.01\end{bmatrix}}_{\vec{w}} = \underbrace{\begin{bmatrix}0.09\\0.08\end{bmatrix}}_{\hat{y}}$$

residuals

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

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

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

Outline

- 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

(Analytic Solution)

Normal Equations

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

$$(\underbrace{X^T}_{(p+1)\times n} \underbrace{X}_{n \times (p+1)})^{-1}$$

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

Linear Regression Runtime

$A \Rightarrow n \times m$

$B \Rightarrow m \times p$

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

(linear)

$O(m)$

$$AB = \underbrace{\begin{bmatrix} \end{bmatrix}}_{n \times m} \underbrace{\begin{bmatrix} \end{bmatrix}}_{m \times p} = \underbrace{\begin{bmatrix} O \ \cdots \end{bmatrix}}_{n \times p}$$

entries

$O(np)$

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

matrix mult

$\rightarrow O(mnp)$

if $m \approx n \approx p$ $\rightarrow \boxed{O(n^3)}$

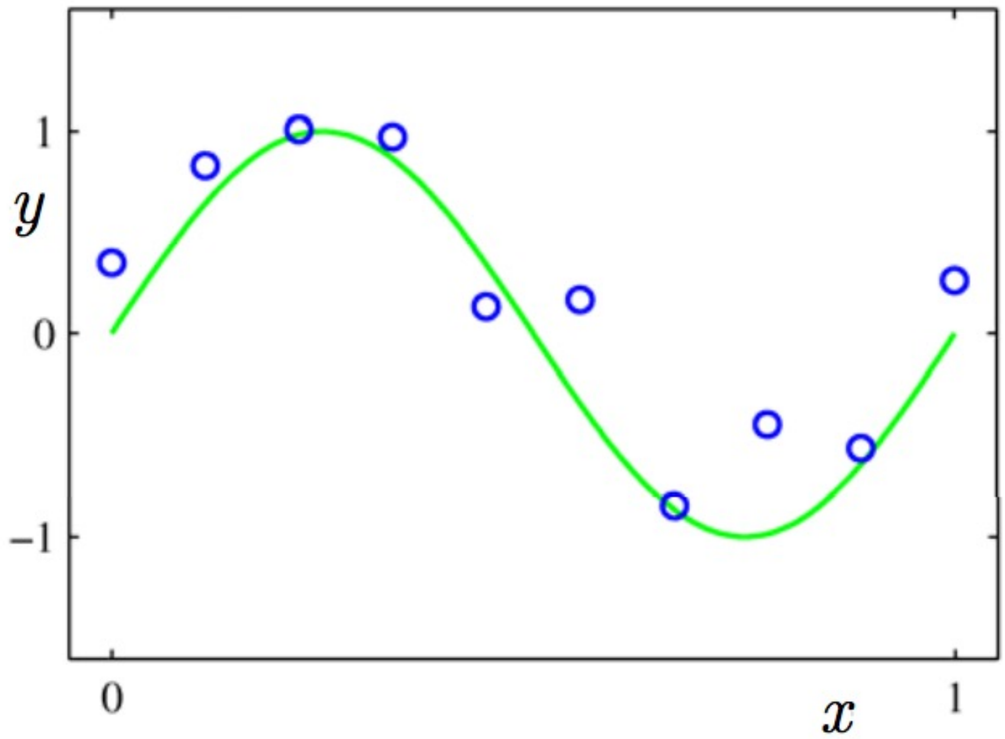
Challenge outside of class!

Outline

- 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

$p = 1 \quad , \deg = d$

$$h_{\vec{w}}(x) = w_0 + w_1 x + w_2 x^2 + \cdots w_d x^d$$

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 & \cdots & x_1^d \\ & \vdots & & & & \\ 1 & x_n & x_n^2 & x_n^3 & \cdots & x_n^d \end{bmatrix}$$