CS 66: Machine Learning

Prof. Sara Mathieson

Spring 2019

Admin

- Office hours **TODAY**! 1-3pm (Sci Center 249)

- Lab 2 due **Tuesday night**
  - See Piazza for dictionary example
  - See Piazza for lab suggestions

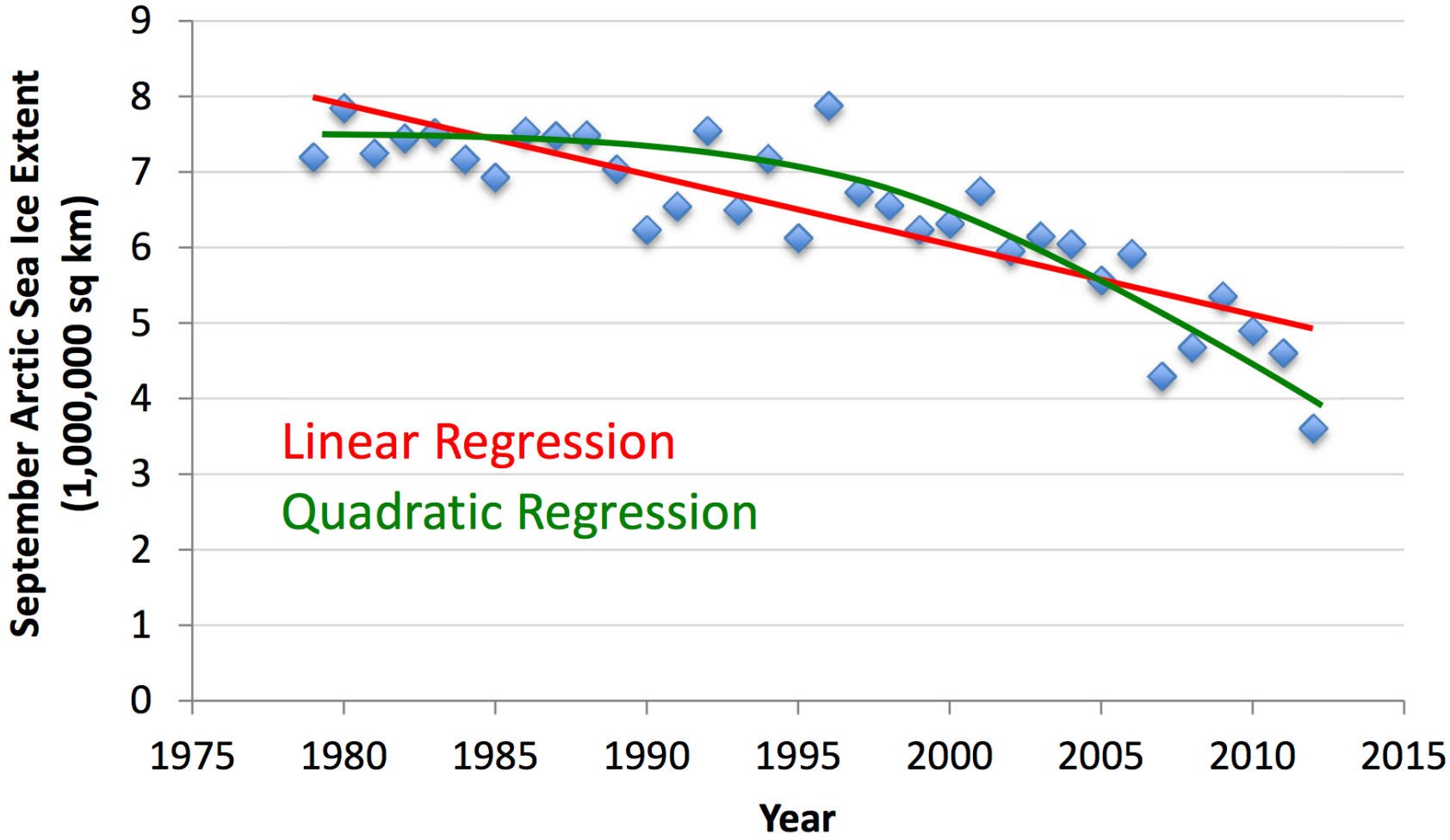
Outline for February 8

- Where we are with regression so far
- Cost function and multiple linear regression
- SGD (Stochastic Gradient Descent)
- Normal equations solution

Outline for February 8

- Where we are with regression so far
- Cost function and multiple linear regression
- SGD (Stochastic Gradient Descent)
- Normal equations solution

Regression Example



Slide: modified from Jessica Wu

[Data from G. Witt. Journal of Statistics Education, Volume 21, Number 1 (2013)]

Linear Regression so far…

- Output (y) is continuous, not a discrete label
- Learned model: *linear function* mapping input to output (a *weight* for each feature + *bias*)
- Goal: minimize the RSS (residual sum of squared errors) or SSE (sum of squared errors)

Simple Linear Regression

- *X* only contains one feature
- Linear model can be described by a *slope* and a *y-intercept* (bias)

$$h_{\boldsymbol{b}}(\boldsymbol{x}) = b_0 + b_1 x$$

Simple Linear Regression

- *X* only contains one feature
- Linear model can be described by a *slope* and a *y-intercept* (bias)

$$\boxed{h_{\boldsymbol{b}}(\boldsymbol{x}) = b_0 + b_1 x}$$

- Solution from Wed

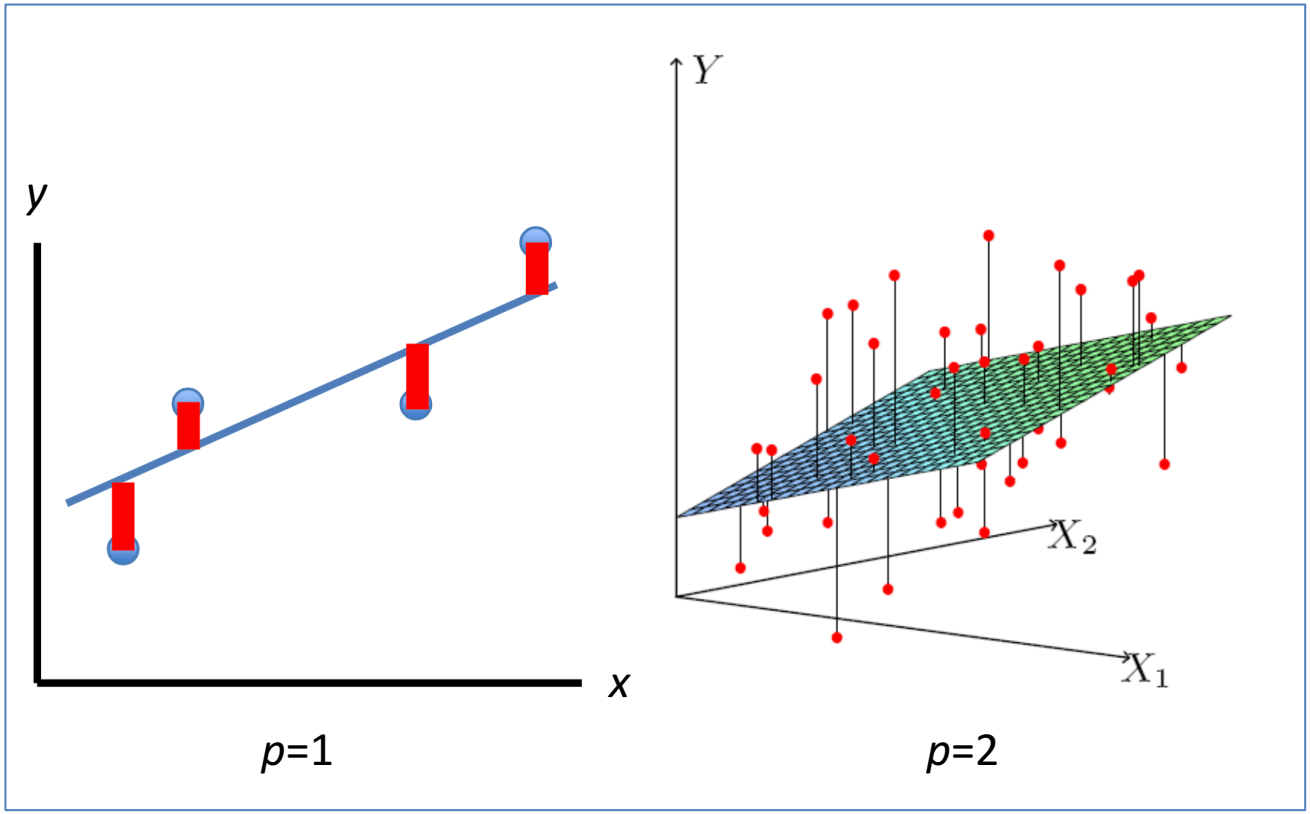
$$\hat{b}_1 = \frac{\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y})}{\mathrm{Var}(\boldsymbol{x})} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$$

Outline for February 8

- Where we are with regression so far
- Cost function and multiple linear regression
- SGD (Stochastic Gradient Descent)
- Normal equations solution

Cost Function: sum of squared errors



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

Slide: modified from Jessica Wu
Original: Eric Eaton

Multiple Linear Regression

$x_0 = 1$

$$\hat{y} = h_{\vec{b}}(\vec{x}) = b_0 x_0 + b_1 x_1 + b_2 x_2 + \cdots b_p x_p$$

$$\vec{b} \cdot \vec{x}$$

$$\boxed{h_{\vec{b}}(\vec{x}) = \vec{b}^T \vec{x}}$$

$$\underbrace{\begin{bmatrix} b_0 & b_1 & \cdots & b_p \end{bmatrix}}_{\vec{b}^T} \quad 1 \times (p+1) \qquad \underbrace{\begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}}_{\vec{x}} \quad (p+1) \times 1 = [\ \ ]$$

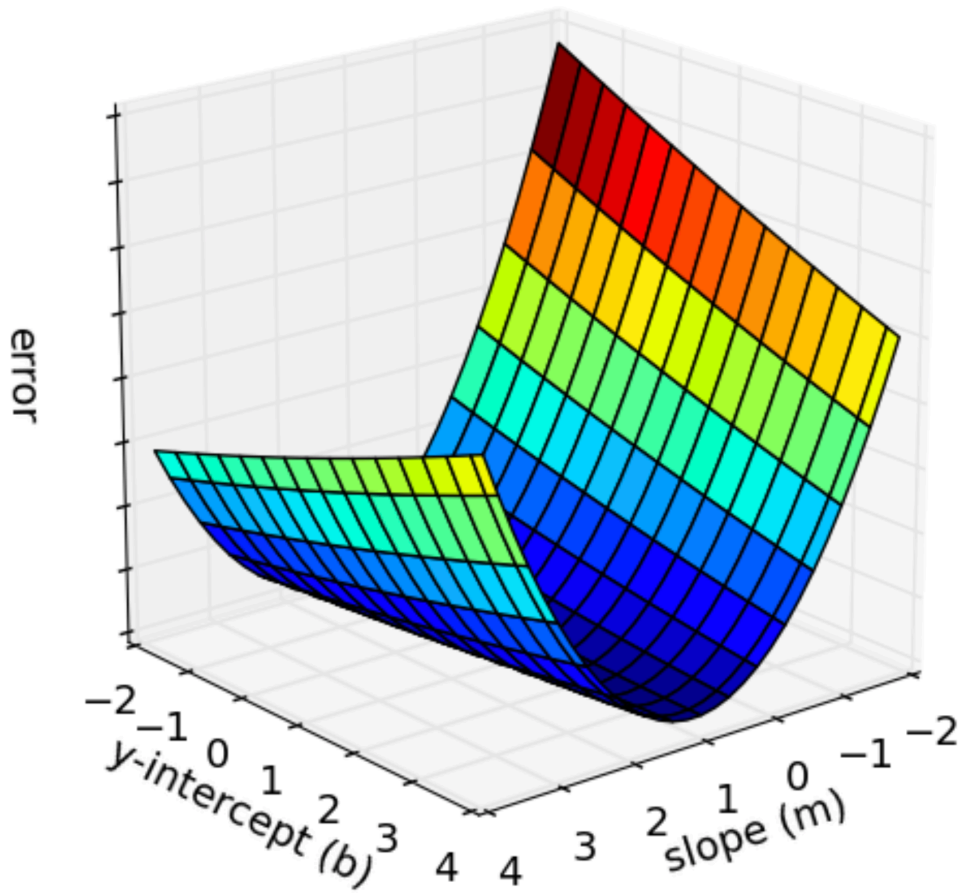
$$X = \underbrace{\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{i1} & x_{i2} & \cdots & x_{ip} \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}}_{p+1} \quad n$$

minimize

$$\boxed{J(\vec{b}) = \frac{1}{2} \sum_{i=1}^{n} \left(\vec{b}^T \vec{x}_i - y_i \right)^2}$$

cost function

Error as a function of slope & y-intercept



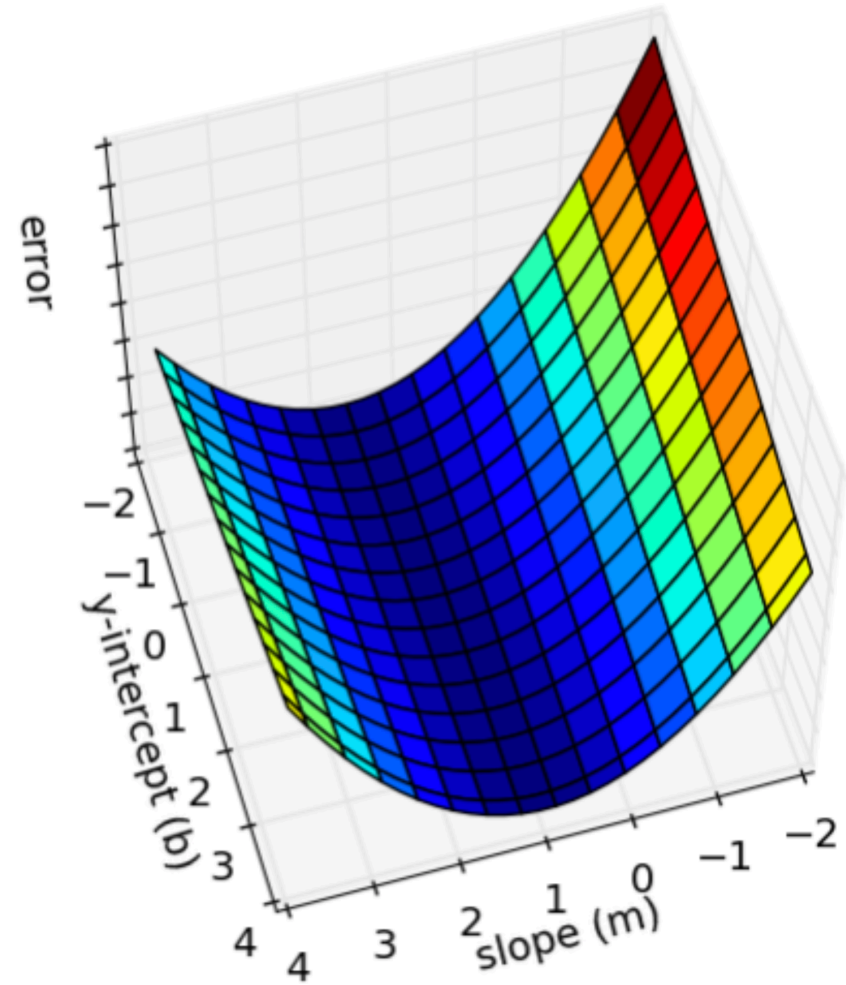
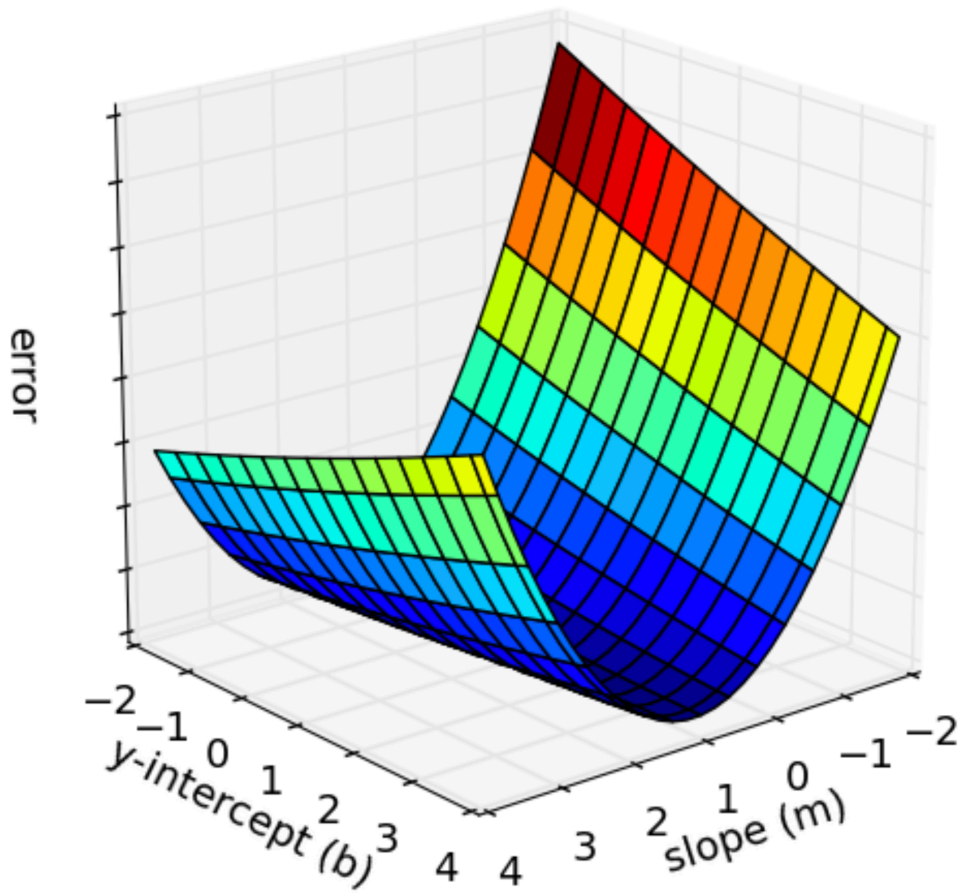
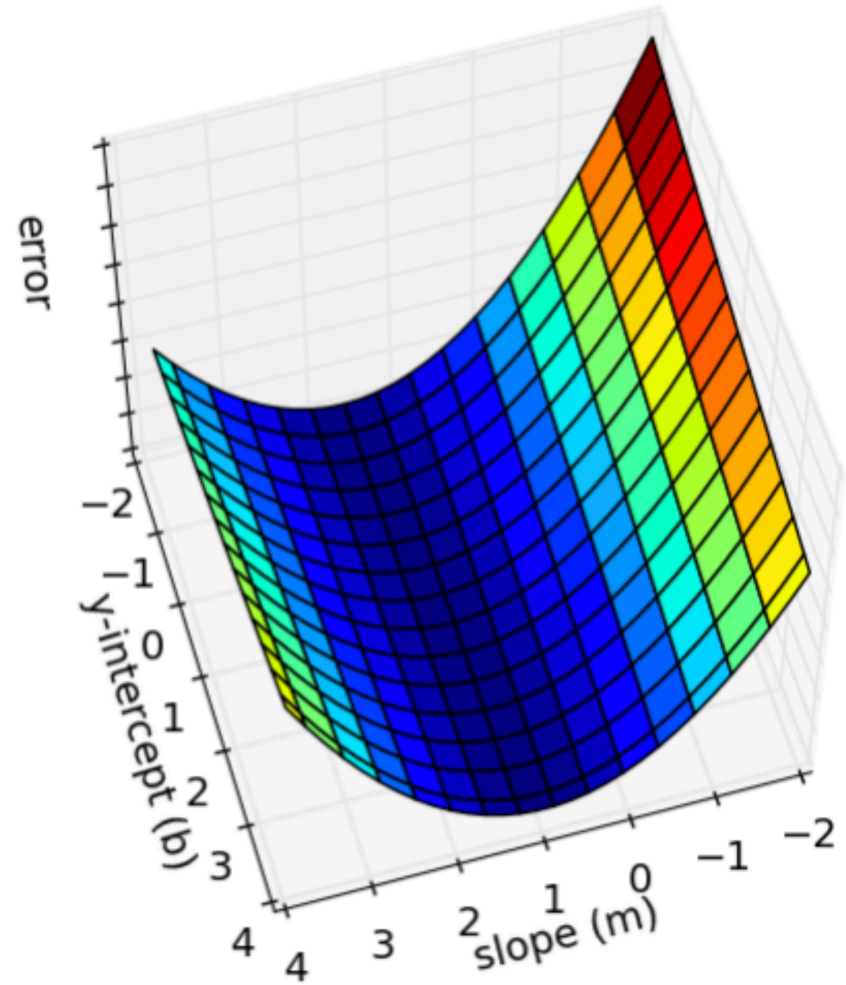


Image: Bryce Wiedenbeck

Error as a function of slope & y-intercept

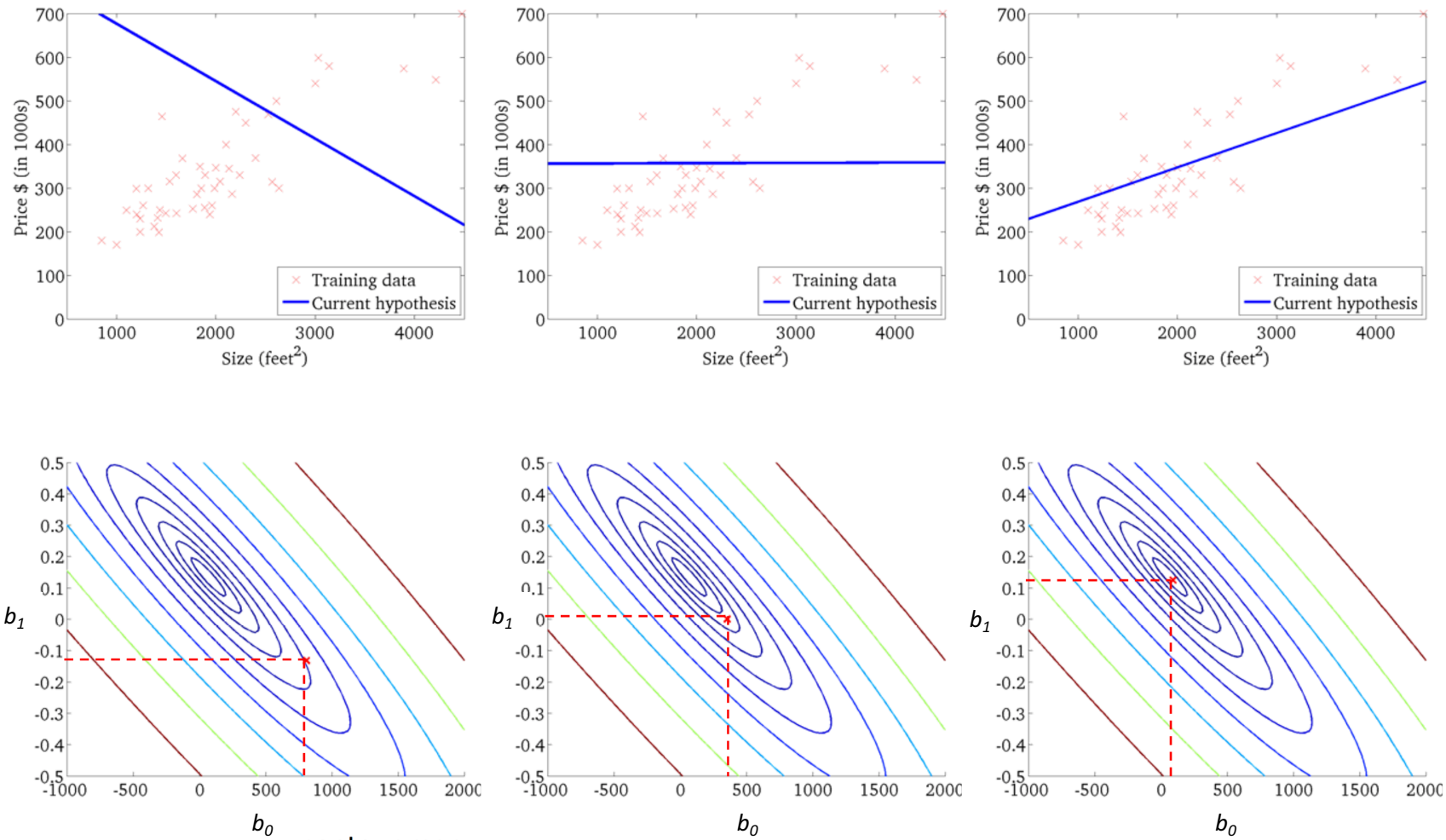




Convex => single global optimum (can prove it is a minimum with second derivative!

Image: Bryce Wiedenbeck

Cost Function



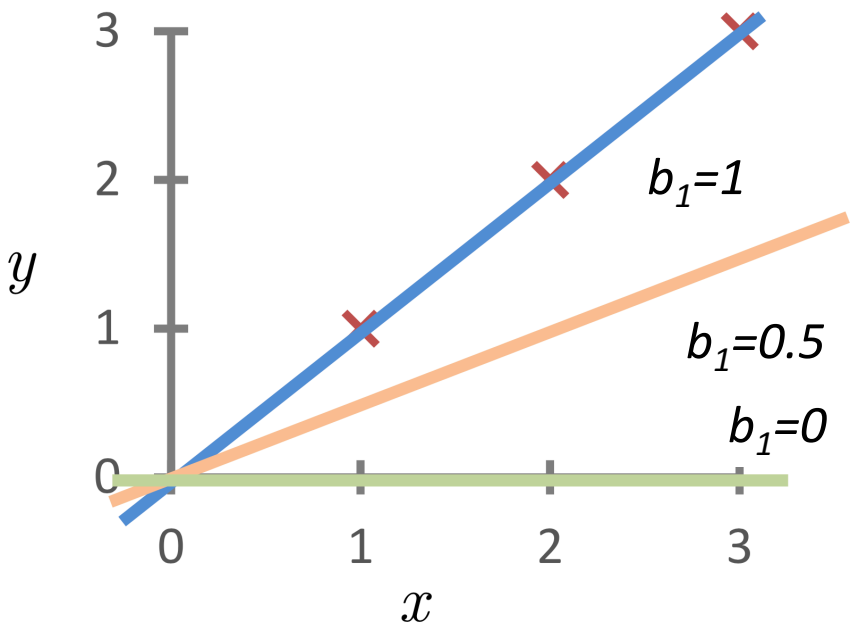
contour map

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

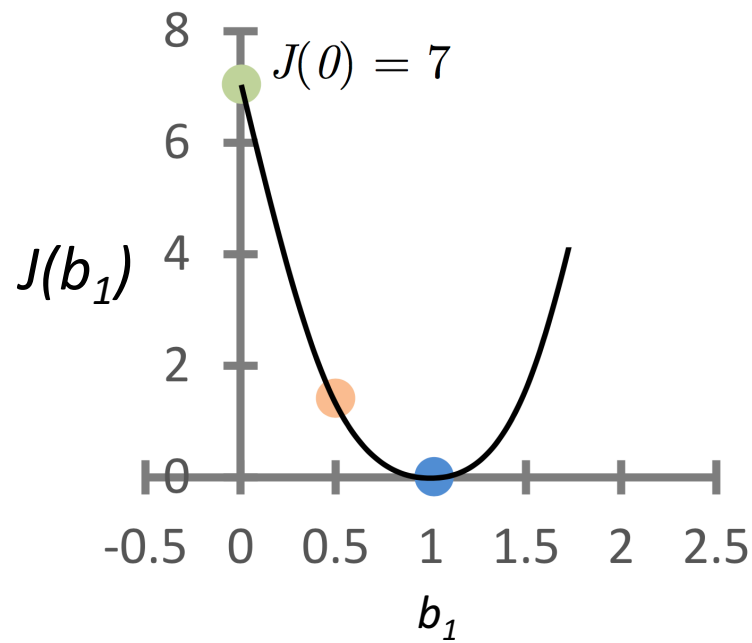
Cost Function (extra example)

$h_b(x)$

(assume $b_0=0$ for this example)



$J(b_1)$



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

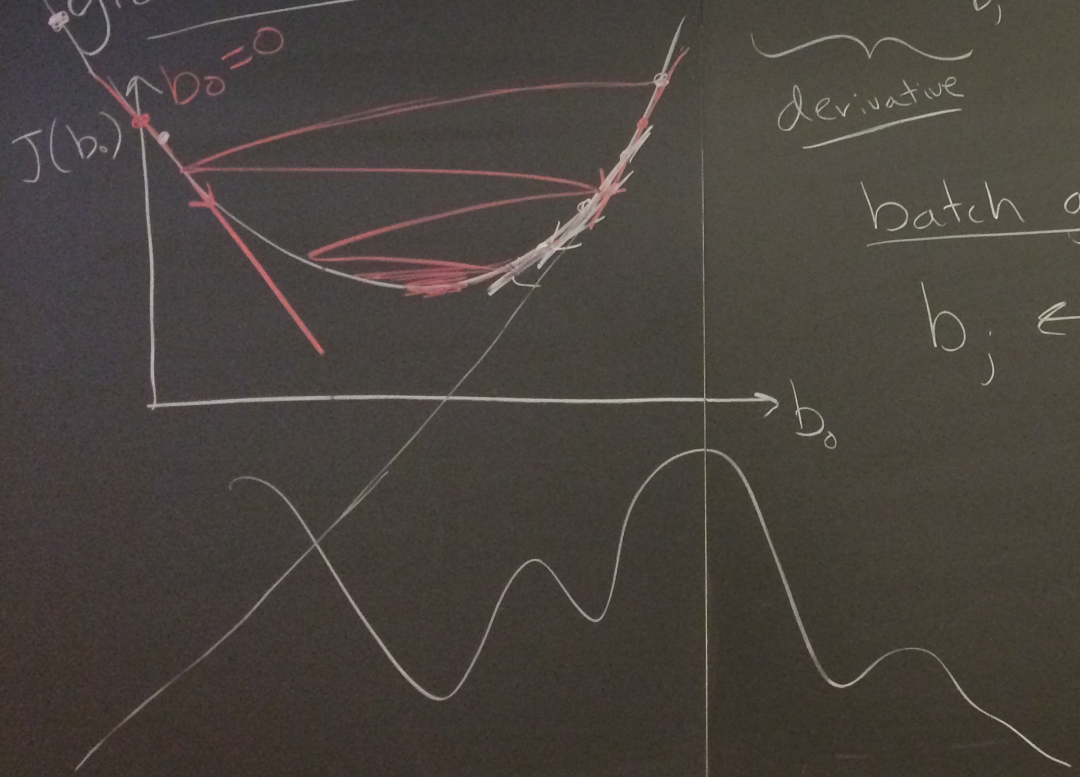
Slide: modified from Jessica Wu
Based on slide by Eric Eaton
[Example by Andrew Ng]

Outline for February 8

- Where we are with regression so far
- Cost function and multiple linear regression
- SGD (Stochastic Gradient Descent)
- Normal equations solution

Method 1

gradient descent



$$\underbrace{J(\vec{b})\Big|_{b_j}}_{\text{derivative}} = \sum_{i=1}^{n} \left(\vec{b}^T \vec{x}_i - y_i \right) x_{ij}$$

batch gradient descent

for all $j = 0 \ldots p$

$$b_j \leftarrow b_j - \alpha \sum_{i=1}^{n} \left(\vec{b}^T \vec{x}_i - y_i \right) x_{ij}$$

slow!

Stocastic gradient descent

shuffle the n data points

for $i = 1 \dots n$:

for $j = 0 \dots p$:

$$\boxed{b_j \leftarrow b_j - \alpha\left(\vec{b}^T \vec{x}_i - y_i\right) x_{ij}}$$

$\alpha = \frac{1}{\text{iteration}}$

$\boxed{\alpha = 0.01}$

iterate til convergence

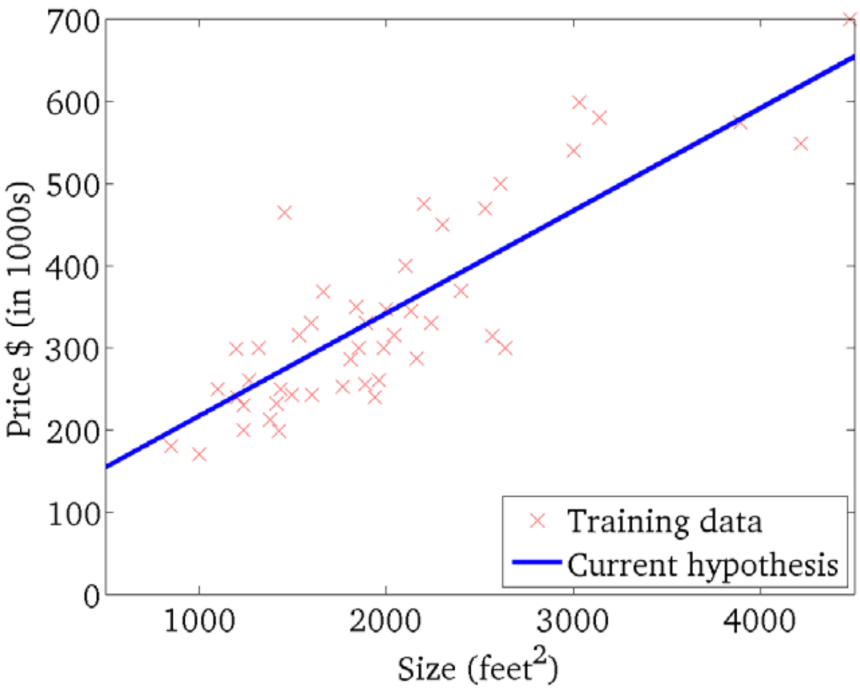
"J doesn't change"

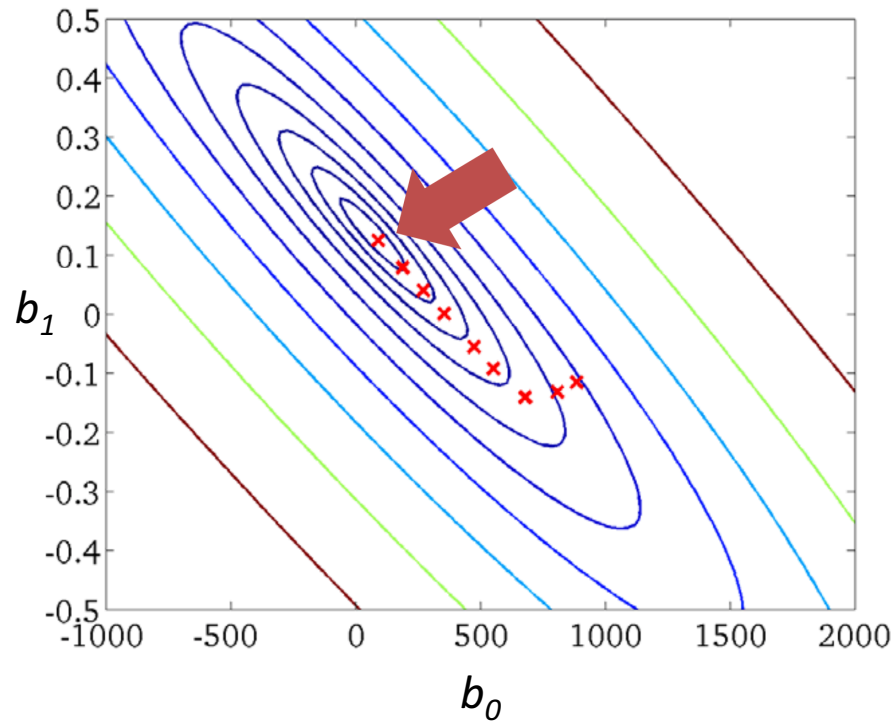
or all

$j = 0 \dots p$

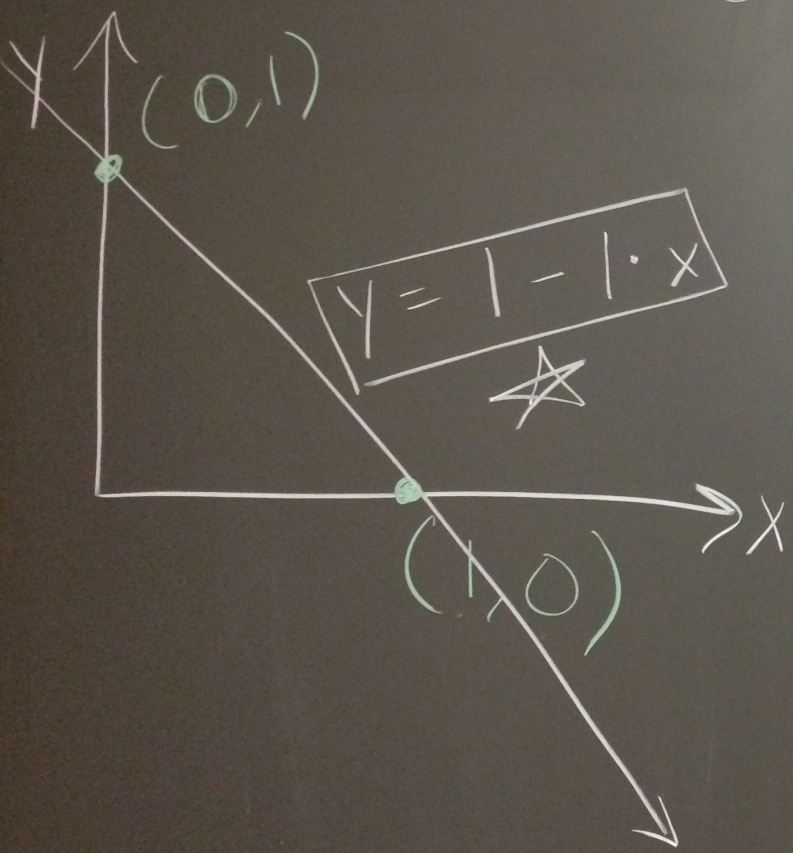
$- y_i) x_{ij}$

Gradient Descent: walking toward the minimum





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



① $$\frac{\mathrm{cov}(x,y)}{\mathrm{var}(x)} = \frac{\cancel{\tfrac{1}{2}}\left[\left(1-\tfrac{1}{2}\right)\left(0-\tfrac{1}{2}\right)+\left(0-\tfrac{1}{2}\right)\left(1-\tfrac{1}{2}\right)\right]}{\cancel{\tfrac{1}{2}}\left[\left(1-\tfrac{1}{2}\right)^2+\left(0-\tfrac{1}{2}\right)^2\right]}$$

$$\boxed{\hat{b}_1 = \frac{-\frac{1}{4}-\frac{1}{4}}{\frac{1}{4}+\frac{1}{4}} = -1}$$

$$\hat{b}_0 = \frac{1}{2} - (-1)\cdot\frac{1}{2}$$

$$\boxed{\hat{b}_0 = 1}$$

Handout 2, Question 1

$i = 1$: $(x_1, y_1) = (1, 0)$

$$\left.\begin{aligned} b_0 &\leftarrow 0 - 0.1(0 - 0)\cdot 1 \\ b_1 &\leftarrow 0 - 0.1(0 - 0)\cdot 1 \end{aligned}\right\} \text{all together}$$

$\boxed{b_0 = 0}$, $\boxed{b_1 = 0}$

$i = 2$: $(x_2, y_2) = (0, 1)$

$$b_0 \leftarrow 0 - 0.1(0 - 1)\cdot 1$$

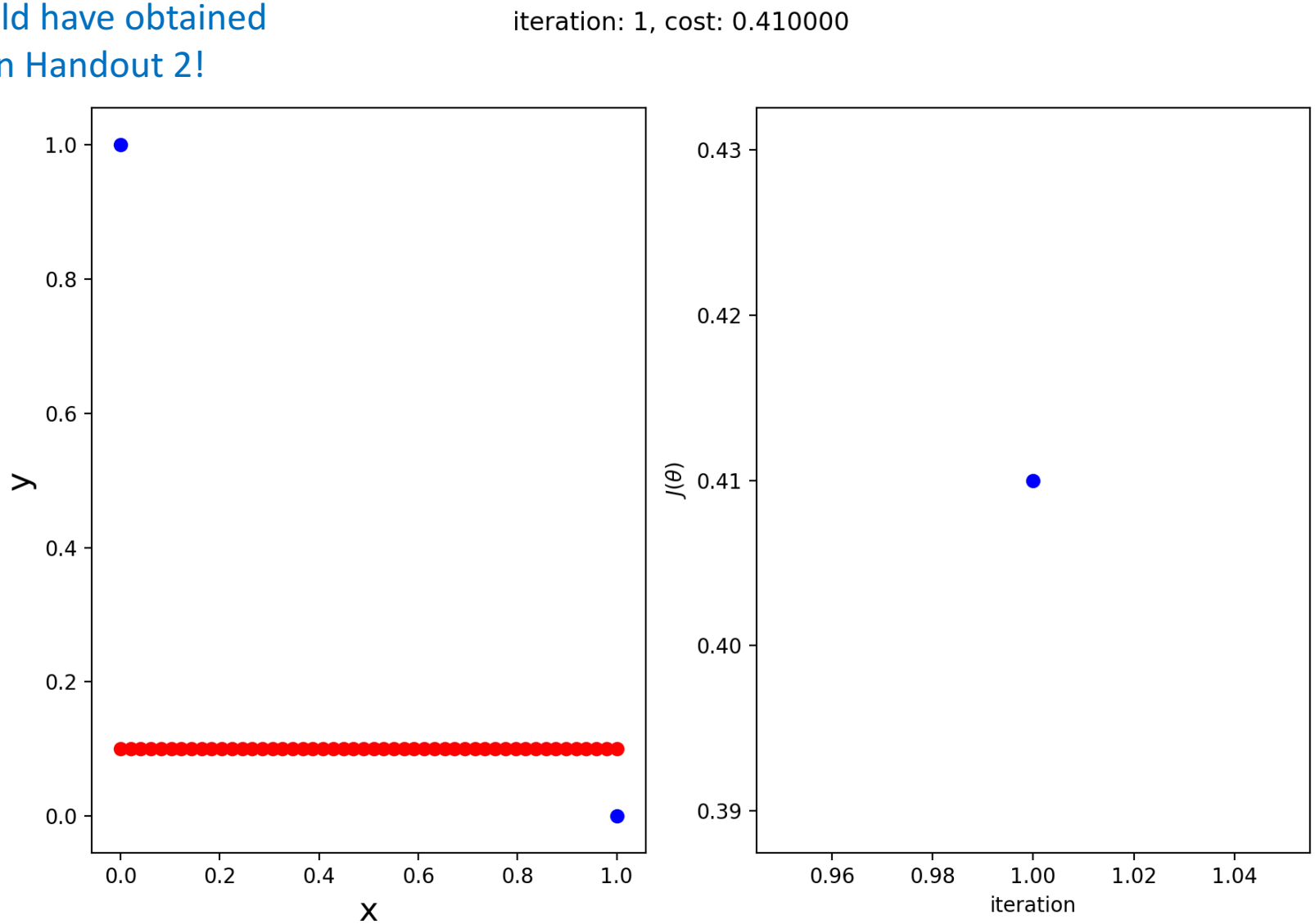
$$b_1 \leftarrow 0 - 0.1(0 - 1)\cdot 0$$

$$\boxed{\begin{aligned} b_0 &= 0.1 \\ b_1 &= 0 \end{aligned}}$$

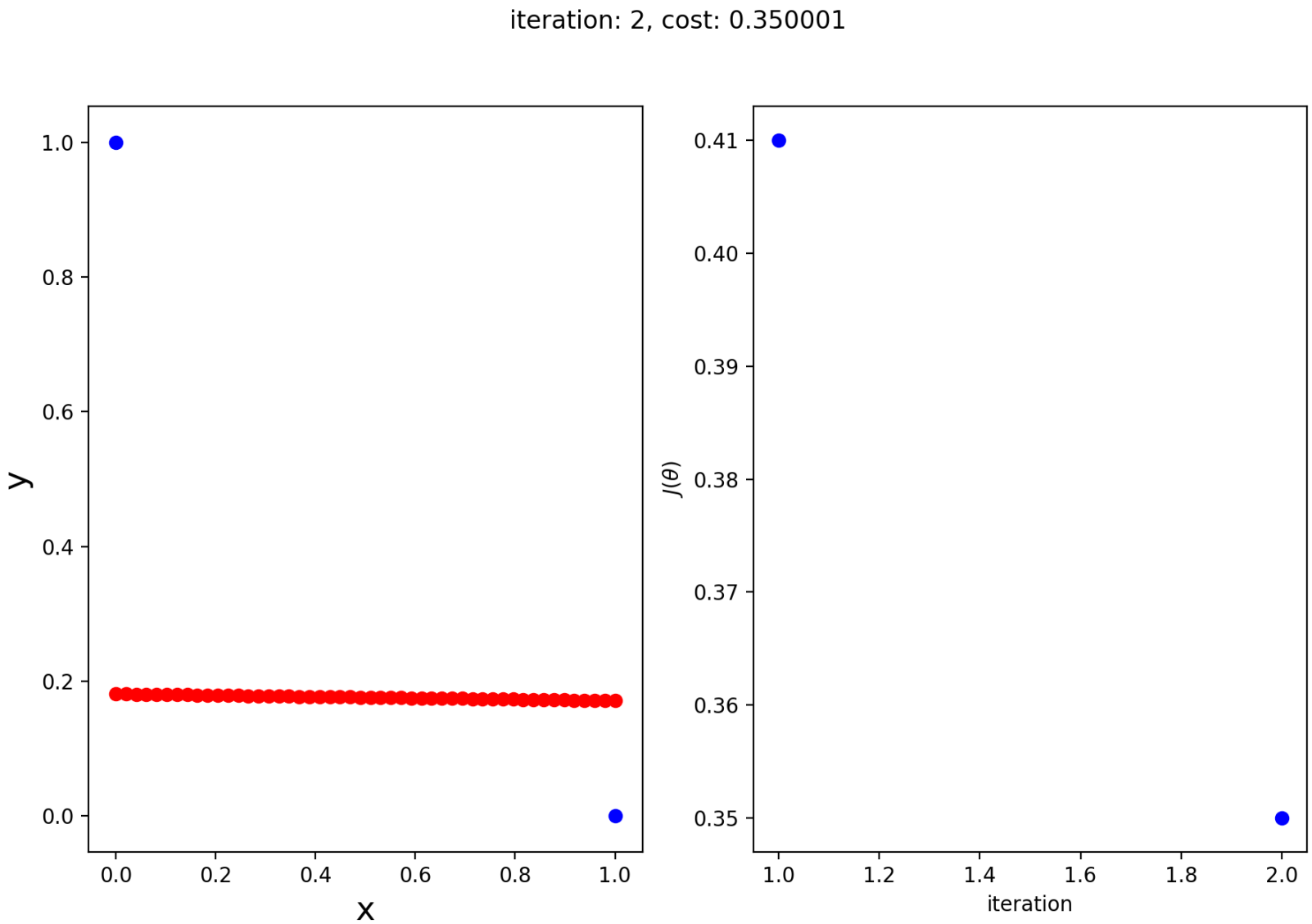
Handout 2, Question 2(a)

Toy example, iteration 1

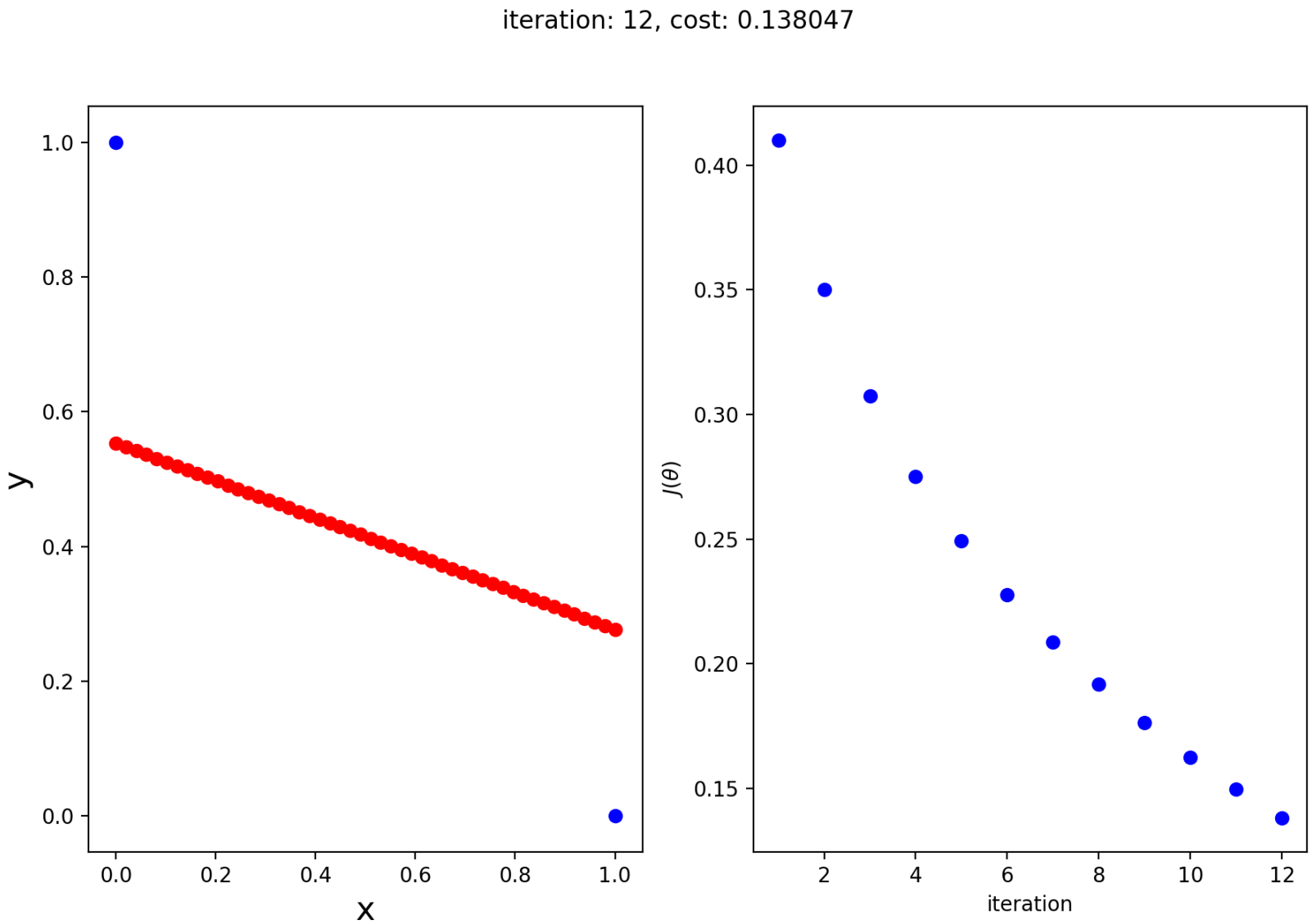
This is what you should have obtained in Handout 2!



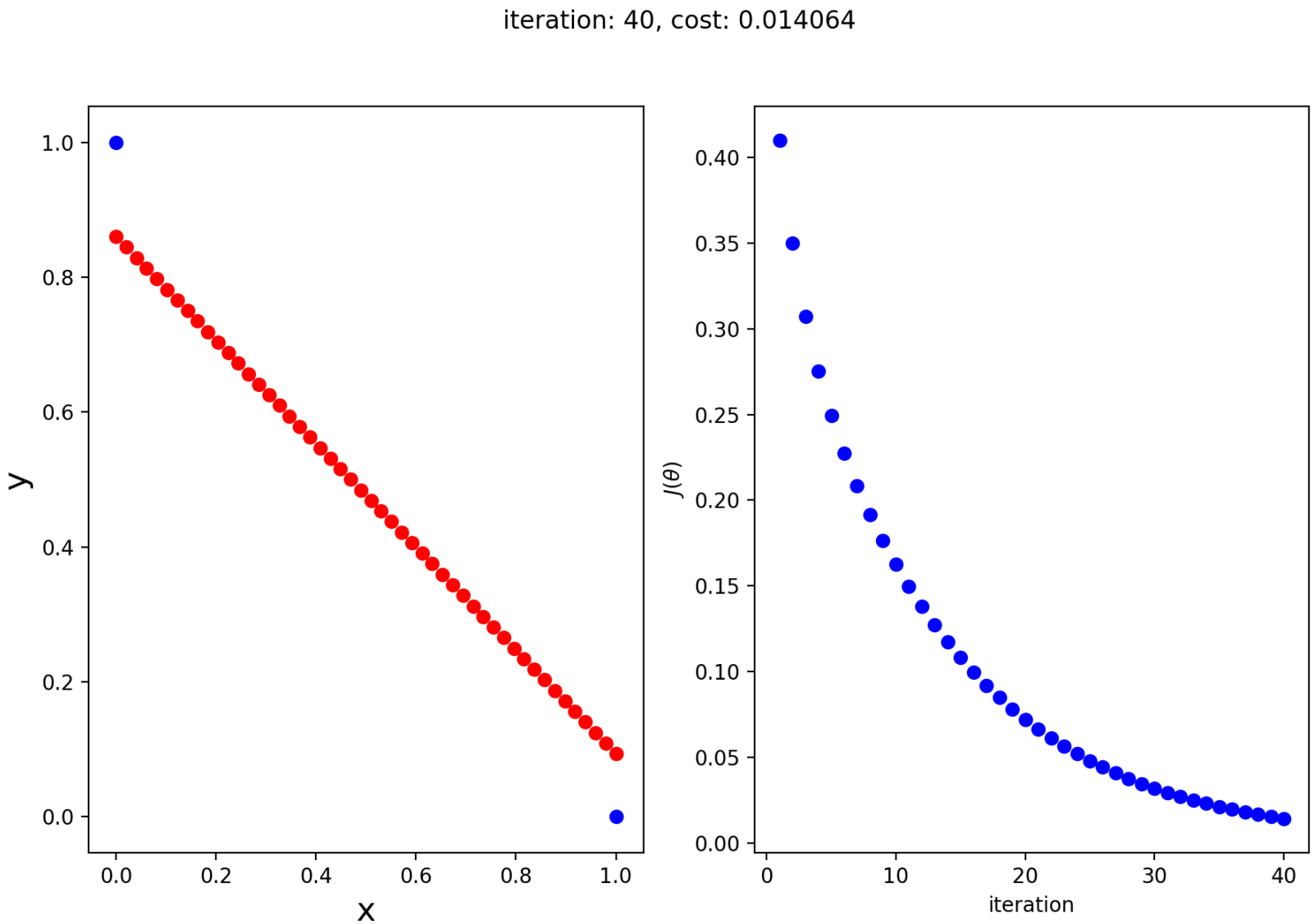
Toy example, iteration 2



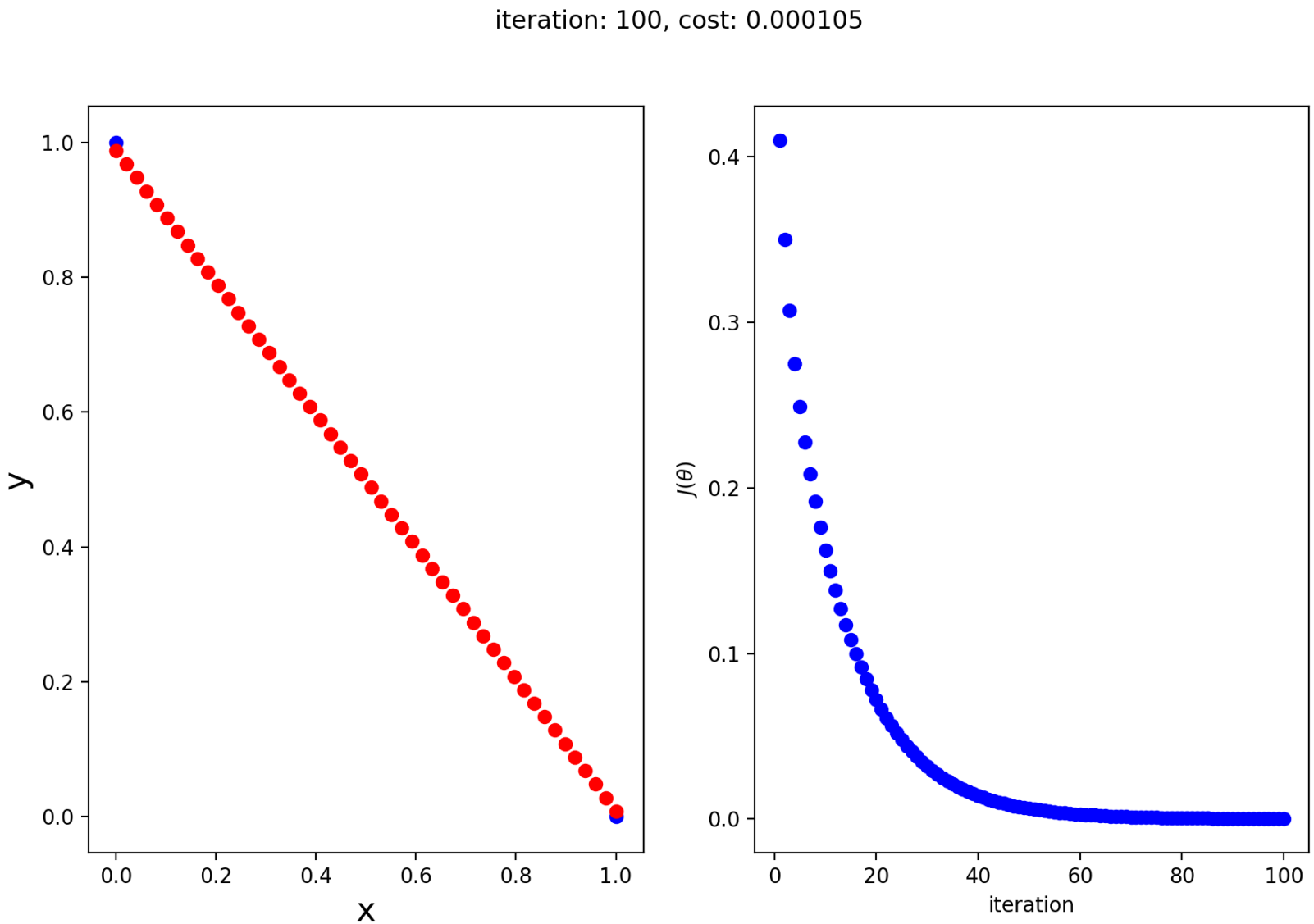
Toy example, iteration 12



Toy example, iteration 40

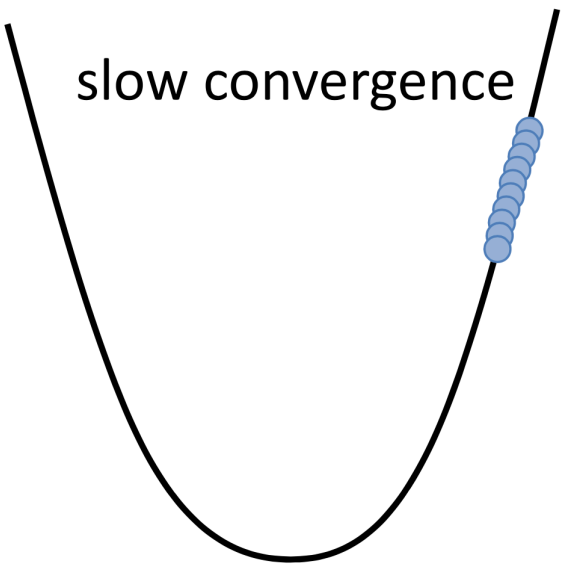


Toy example, iteration 100

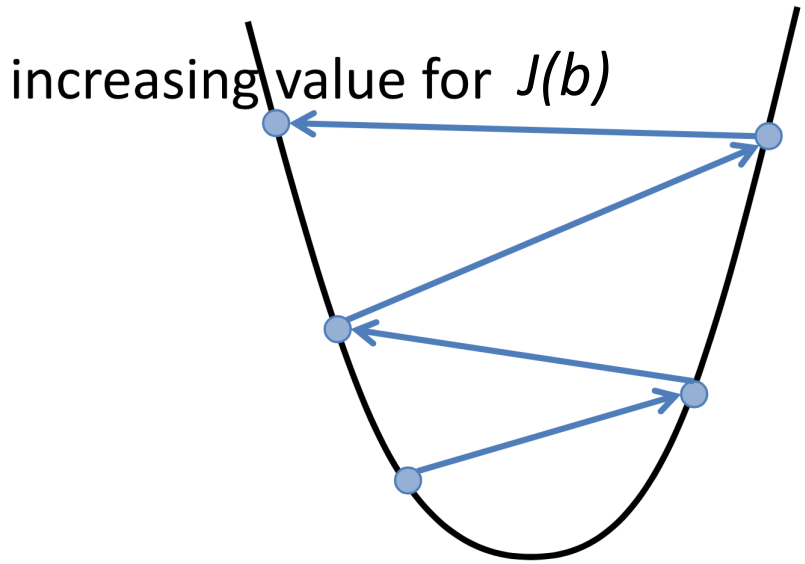


Choosing step size α

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]

Outline for February 8

- Where we are with regression so far
- Cost function and multiple linear regression
- SGD (Stochastic Gradient Descent)
- Normal equations solution

Pros and Cons

Gradient Descent

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

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