CS 360: Machine Learning

Prof. Sara Mathieson

Fall 2019

HAVERFORD COLLEGE

Admin

- Office hours **Friday**! 3-5pm (Hilles 110)
- Start Lab 3 in lab today (I will do some instruction during lab)
- Lab 3 due **Tuesday night**

Outline for September 19

- Reading quiz #3
- Simple linear regression
- SGD (Stochastic Gradient Descent)
- Normal equations solution
- Week 3 feedback

Lab 1 feedback

- Lab 1 grades posted (10pts for knn, 3 for math)
- Keep large data out of repos (unless it is given as part of the starter)

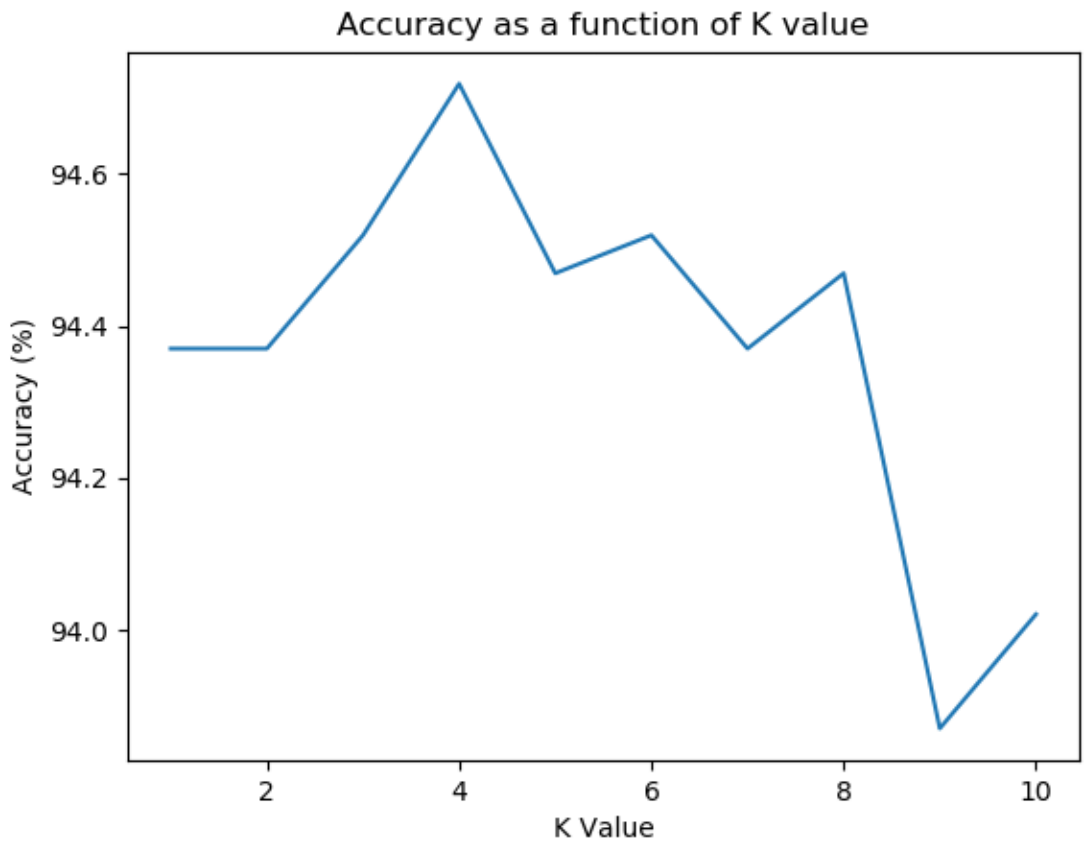
- Make sure code is in functions and function decomposition makes sense
- Keep lines less than 80 characters
- Official python style for methods/functions
  - my_function
  - ~~MyFunction~~ (uppercase is reserved for classes!)

- Some people are new to Python: I am happy to take some time during lab to go over syntax/common issues

Lab 1 Extensions

- Ahmed: implemented a weighting voting scheme for multi-class classification
- Amazing style: Fiona, Jason
- One of the fastest solutions: Emile
- Jiaping: command line options and automatic plot creation

Lab 1 Extensions: multi-class



Brian and Gareth created similar plots

Lab 1: how to make KNN faster?

- Runtime: exercise!

- Don't need to sort all distances – for small K, we can find the top K neighbors in linear time

- Save matrix of pair-wise distances across K
- Use less of the training data

- Put each training example in a "zone" or "cluster". For each test example, identify cluster and only consider neighbors within that cluster

Outline for September 19

- Reading quiz #3
- Simple linear regression
- SGD (Stochastic Gradient Descent)
- Normal equations solution
- Week 3 feedback

Reading Quiz #3

1. Name one goal of linear regression. Why would we want to fit a linear model?

Reading Quiz #3

1. Name one goal of linear regression. Why would we want to fit a linear model?

   Learn which features contribute to the response
   Predict future responses based on new features

2. In *simple* linear regression, how many features/predictors does each example have? How many parameters does our linear model have in this case?

   - # features (p) =
   - # model params =

Reading Quiz #3

1. Name one goal of linear regression. Why would we want to fit a linear model?

   Learn which features contribute to the response
   Predict future responses based on new features

2. In *simple* linear regression, how many features/predictors does each example have? How many parameters does our linear model have in this case?

   - # features (p) = 1
   - # model params = 2

3. What loss function are we trying to minimize in classical linear regression?

Reading Quiz #3

1. Name one goal of linear regression. Why would we want to fit a linear model?

   Learn which features contribute to the response
   Predict future responses based on new features

2. In *simple* linear regression, how many features/predictors does each example have? How many parameters does our linear model have in this case?

   - # features (p) = 1
   - # model params = 2

3. What loss function are we trying to minimize in classical linear regression?

   squared error

4. Does a closed-form (analytic) solution exist for the parameters of a linear model?

Reading Quiz #3

1. Name one goal of linear regression. Why would we want to fit a linear model?

   Learn which features contribute to the response
   Predict future responses based on new features

2. In *simple* linear regression, how many features/predictors does each example have? How many parameters does our linear model have in this case?

   - # features (p) = 1
   - # model params = 2

3. What loss function are we trying to minimize in classical linear regression?

   squared error

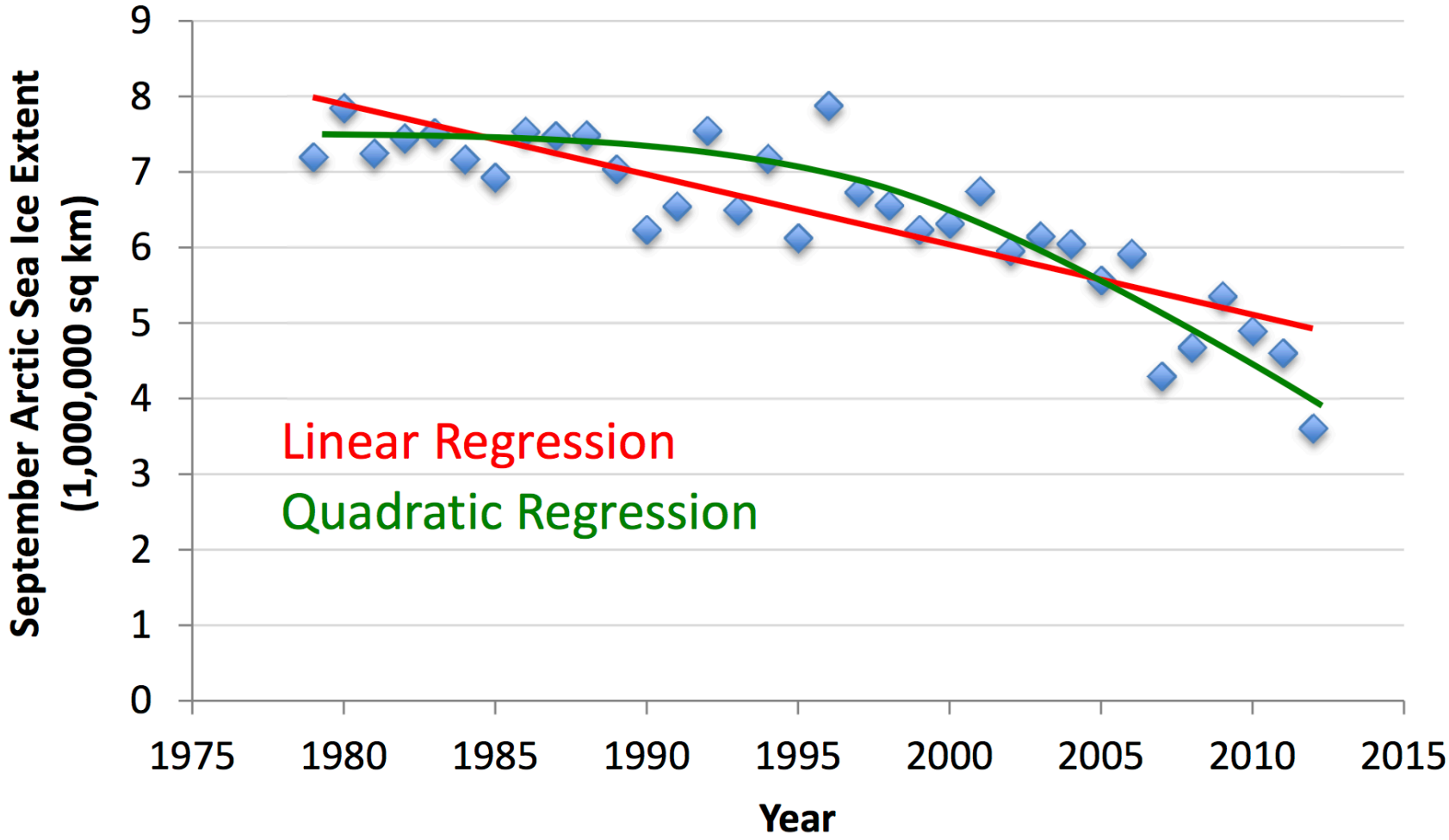
4. Does a closed-form (analytic) solution exist for the parameters of a linear model?

   Yes!

Outline for September 19

- Reading quiz #3
- Simple linear regression
- SGD (Stochastic Gradient Descent)
- Normal equations solution
- Week 3 feedback

Regression Example



Slide: modified from Jessica Wu

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

Goals of Inference

1) Which of the features/explanatory variables/ predictors (x) are associated with the response variable (y)?

2) What is the relationship between x and y?

3) Is a linear model enough?

4) Can we predict y given a new x?

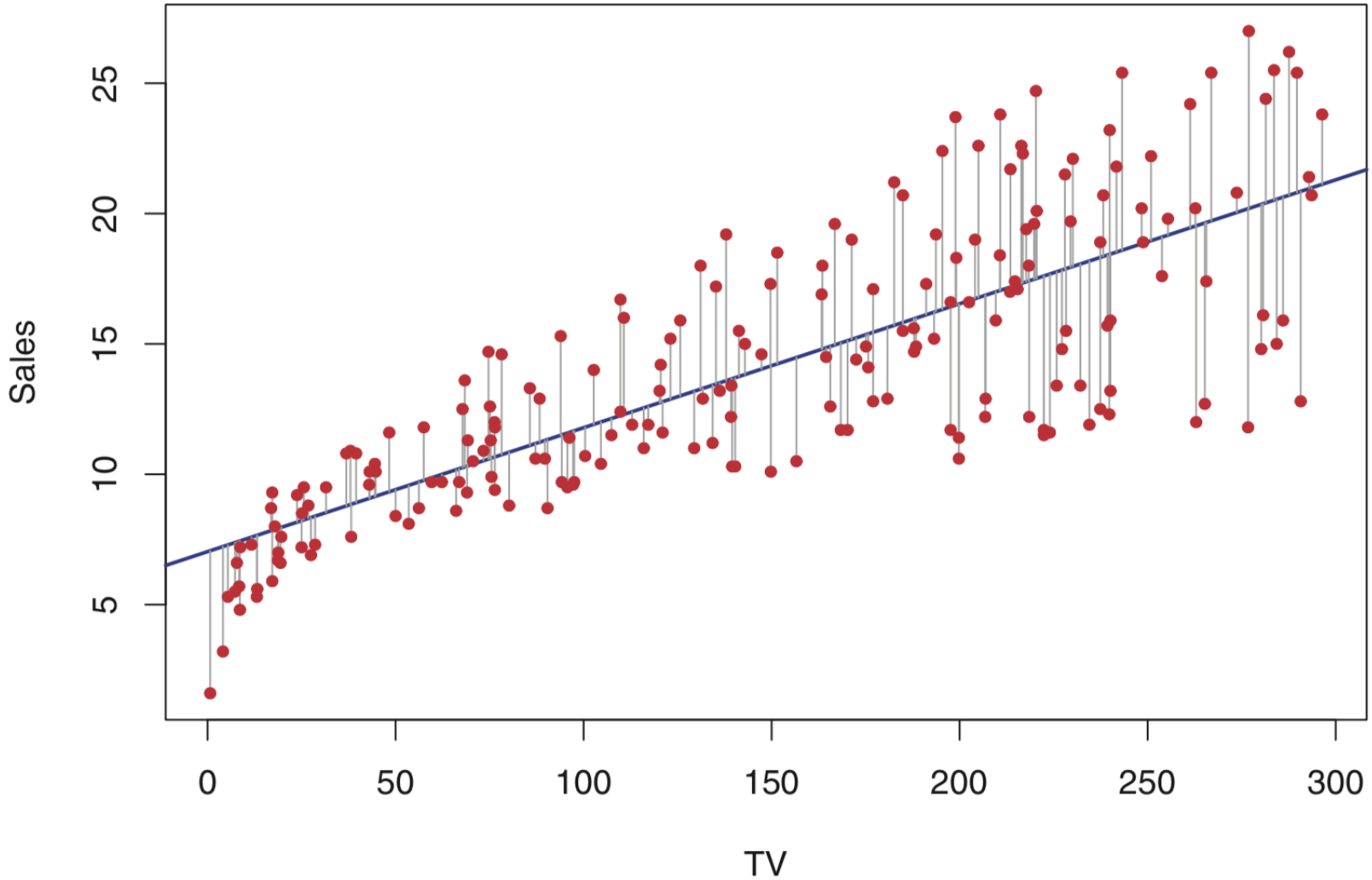
Linear Regression so far…

- Output (y) is continuous, not a discrete label

- <u>Learned model</u>: *linear function* mapping input to output (a *weight* for each feature + *bias*)

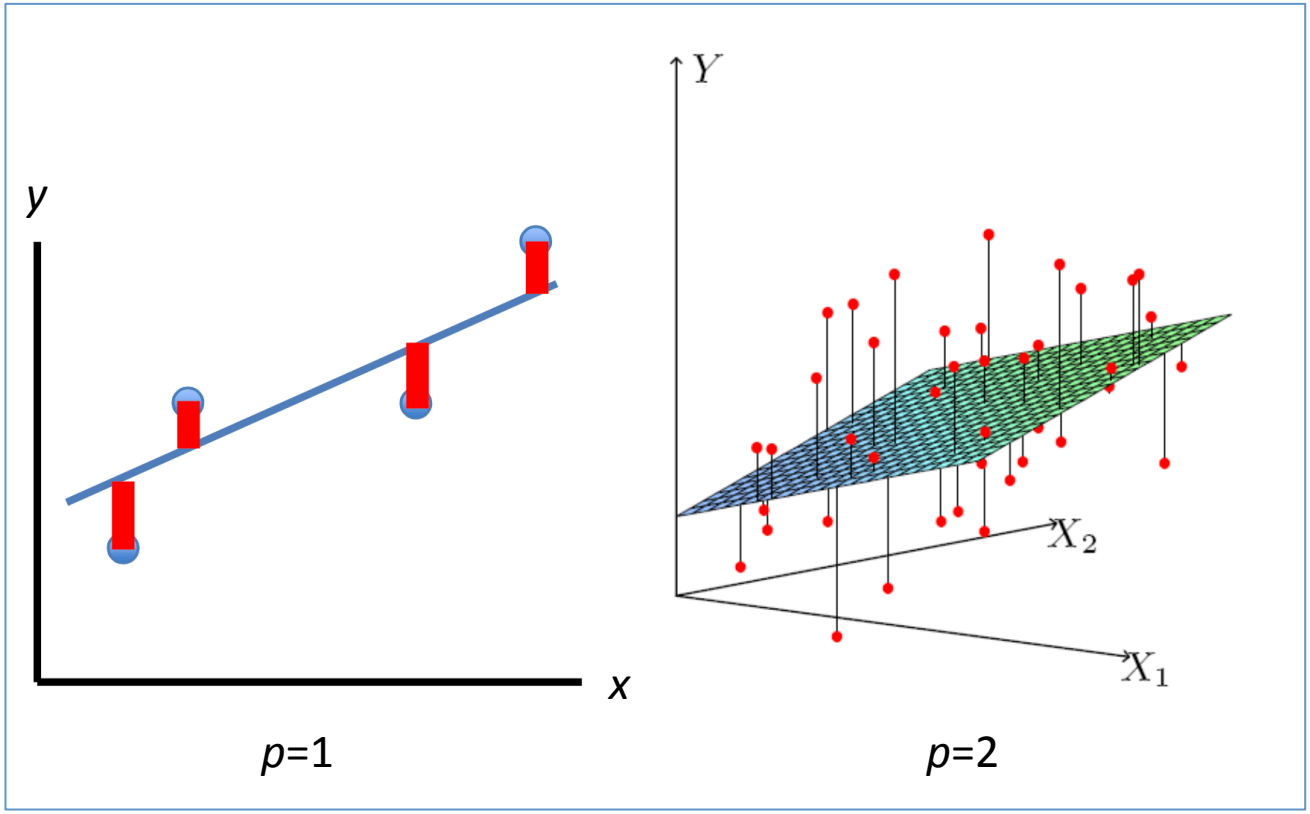
- <u>Goal</u>: minimize the *RSS* (residual sum of squared errors) or *SSE* (sum of squared errors)

Example: predict sales from TV advertising budget



ISL: Figure 3.1

Cost Function: sum of squared errors



Slide: modified from Jessica Wu
Original: Eric Eaton

"Simple" linear regression

$\boxed{p=1}$

"best fit" $\left\{ \hat{f}(x) = \hat{w}_0 + \hat{w}_1 x \right.$

$= \hat{y}$

$= h_{\vec{w}}(x)$

Goal: minimize

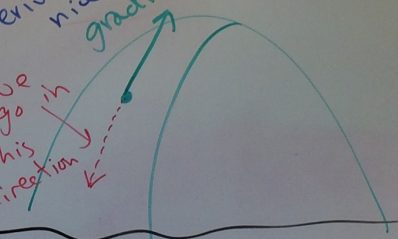
$$MSE = \frac{1}{n} \underbrace{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}_{\substack{RSS \\ SSE}}$$

Cost function:

$$J(\vec{w}) = \frac{1}{2} \sum_{i=1}^{n} (w_0 + w_1 x_i - y_i)^2$$

make derivative nice

gradient



gradient

$$\nabla J = \begin{bmatrix} \frac{\partial J}{\partial w_0} \\ \frac{\partial J}{\partial w_1} \end{bmatrix}$$

partial derivatives

$$\frac{\partial J}{\partial w_0} = \sum_{i=1}^{n} (w_0 + w_1 x_i - y_i) = 0$$

solve for w_0

$$\frac{n w_0}{n} = -w_1 \underbrace{\frac{1}{n} \sum_{i=1}^{n} x_i}_{\substack{\text{avg/mean} \\ \bar{x}}} + \frac{1}{n} \sum_{i=1}^{n} y_i$$

$\bar{y}$ (mean of y's)

$$\boxed{\hat{w}_0 = \bar{y} - w_1 \bar{x}}$$

$$\frac{\partial J}{\partial w_1} = \sum_{i=1}^{n} (w_0 + w_1 x_i - y_i)\, x_i$$

$$\Rightarrow \quad w_1 = \frac{\sum_{i=1}^{n} (x_i y_i - x_i \bar{y})}{\sum_{i=1}^{n} (x_i^2 - x_i \bar{x})}$$

<u>note</u>:

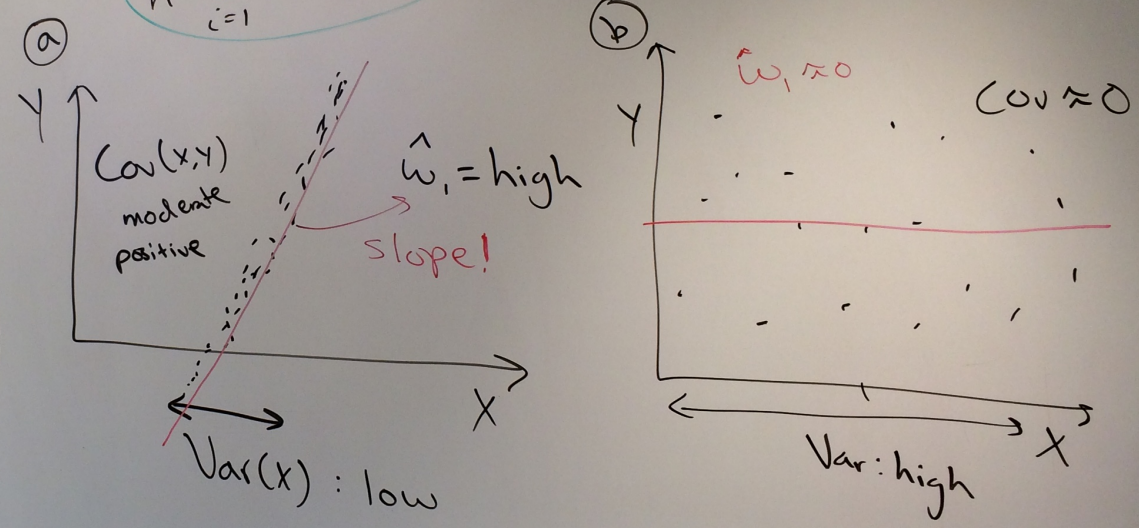
$$\sum_{i=1}^{n} (\bar{x}\bar{y} - y_i \bar{x}) = 0$$

$$\sum_{i=1}^{n} (\bar{x}^2 - x_i \bar{x}) = 0$$

$$\hat{w}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$Cov(X,Y)$

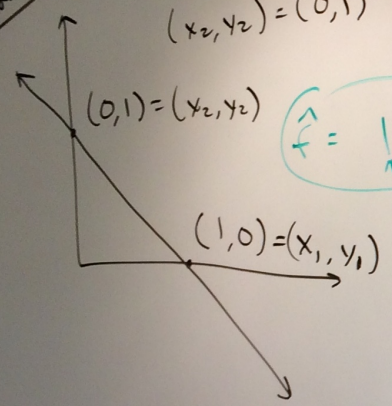
$Var(X)$



Handout 5

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

$(x_2, y_2) = (0, 1)$



$\hat{f} = \underset{w_0}{1} - \underset{w_1}{1}x$

$$\hat{w}_1 := \frac{Cov(X,Y)}{Var(X)}$$

$\bar{x} = \frac{1}{2}$

$\bar{y} = \frac{1}{2}$

$$= \frac{\frac{1}{2}\left[\left(1-\frac{1}{2}\right)\left(0-\frac{1}{2}\right) + \left(0-\frac{1}{2}\right)\left(1-\frac{1}{2}\right)\right]}{\frac{1}{2}\left[\left(1-\frac{1}{2}\right)^2 + \left(0-\frac{1}{2}\right)^2\right]}$$

$$\boxed{\hat{w}_1 = -1}$$

$$w_0 = \frac{1}{2} - (-1)\cdot\frac{1}{2}$$

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

Outline for September 19

- Reading quiz #3
- Simple linear regression
- SGD (Stochastic Gradient Descent)
- Normal equations solution
- Week 3 feedback

$\boxed{p>1}$

multiple linear regression

$x_0 = 1$ $\quad w_j x_j$

$$\hat{y} = \hat{f}(\vec{x}) = h_{\vec{w}}(\vec{x}) = w_0 \cdot 1 + w_1 x_1 + w_2 x_2 \cdots + w_p x_p$$

$$= \vec{w}^T \vec{x} = \vec{w} \cdot \vec{x}$$

dot product.

row x col.

$$\underbrace{\begin{bmatrix} w_0 & w_1 & \cdots & w_p \end{bmatrix}}_{\vec{w}^T} \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_p \end{bmatrix} = \text{Scalar}$$

$1 \times (p+1) \qquad (p+1) \times 1$

$$\vec{a} \cdot \vec{b} = \sum_{i=1}^{len} a_i b_i$$

Cost function (minimize!)

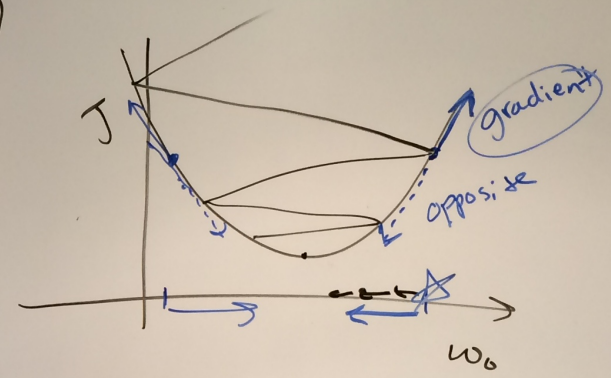
$$\boxed{J(\vec{w}) = \frac{1}{2} \sum_{i=1}^{n} \left(\vec{w}^T \vec{x}_i - y_i \right)^2}$$

$\boxed{\text{method 1}}$ **Stocastic Gradient Descent (SGD)**

$$\frac{\partial J}{\partial w_j} = \sum_{i=1}^{n} (\vec{w}^T \vec{x}_i - y_i) x_{ij}$$

slow!! when n is large.

$$\nabla J = \begin{bmatrix} \frac{\partial J}{\partial w_0} \\ \vdots \\ \frac{\partial J}{\partial w_p} \end{bmatrix}$$



algorithm

while not converged:

shuffle all data points

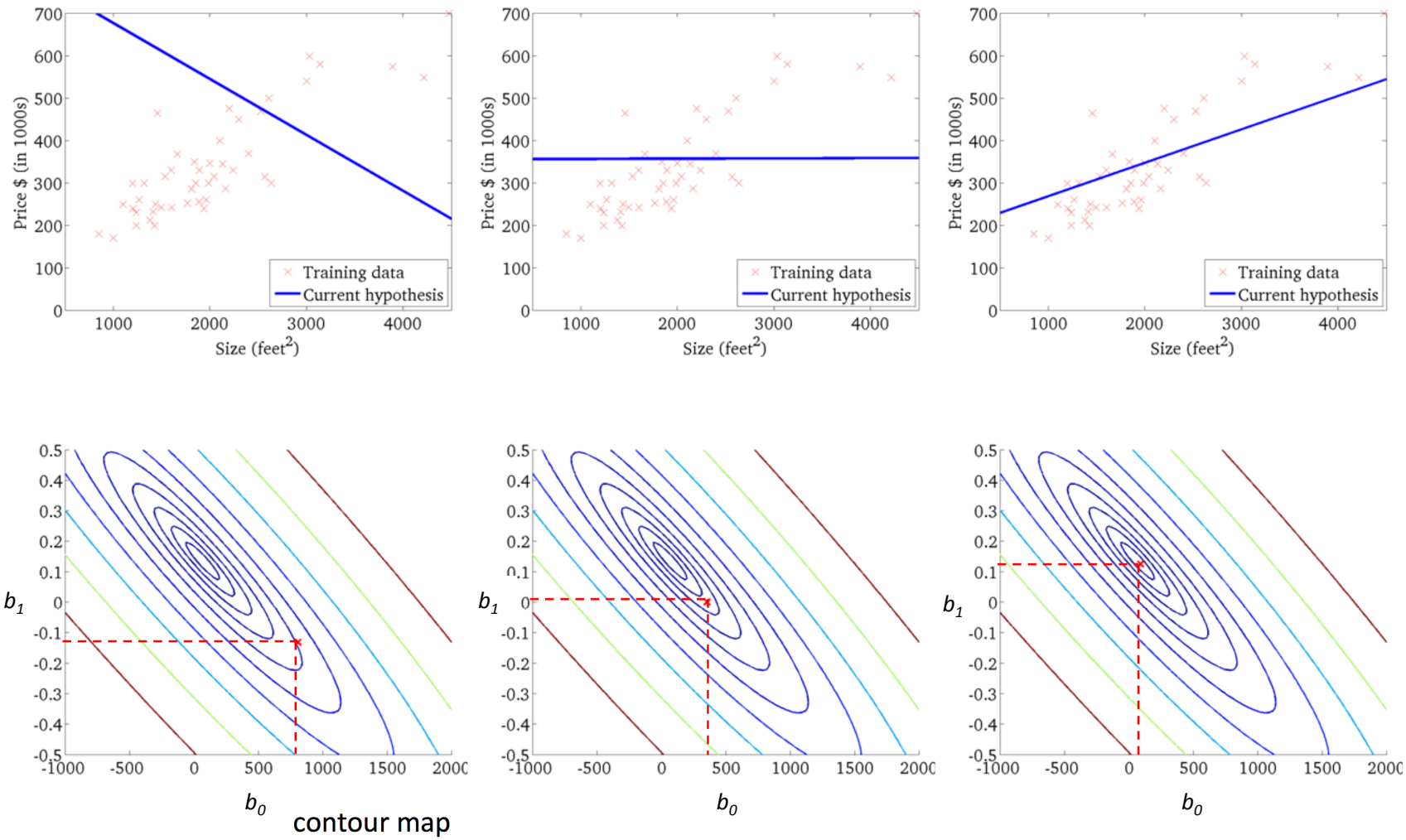
for $i = 1 \cdots n$ # go through one at a time!

$$\boxed{\vec{w} \leftarrow \vec{w} - \alpha \underbrace{(\vec{w}^T \vec{x}_i - y_i) \vec{x}_i}_{\text{gradient of } J \text{ wrt one data point.}}}$$ all weights!

opposite

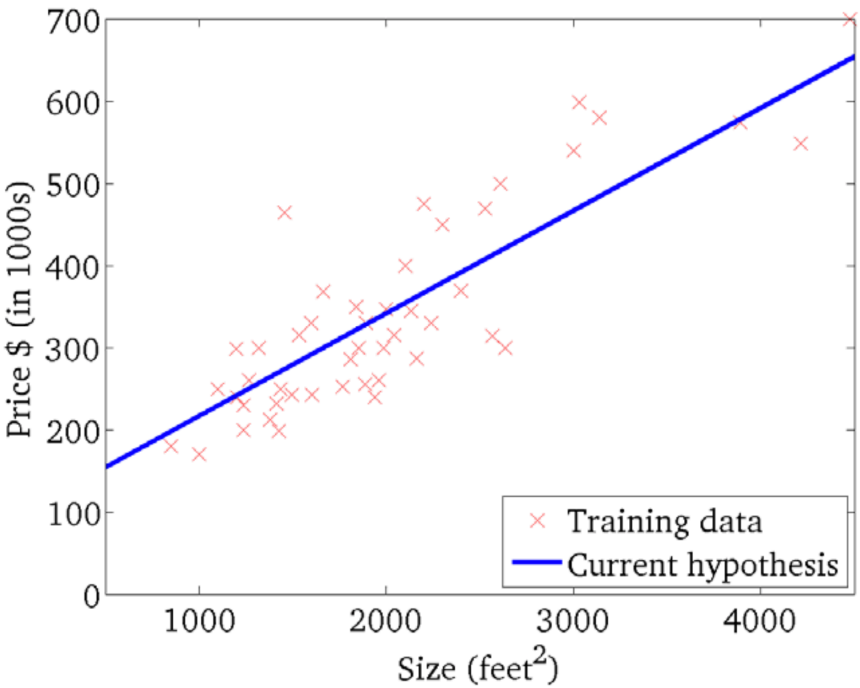
step size.

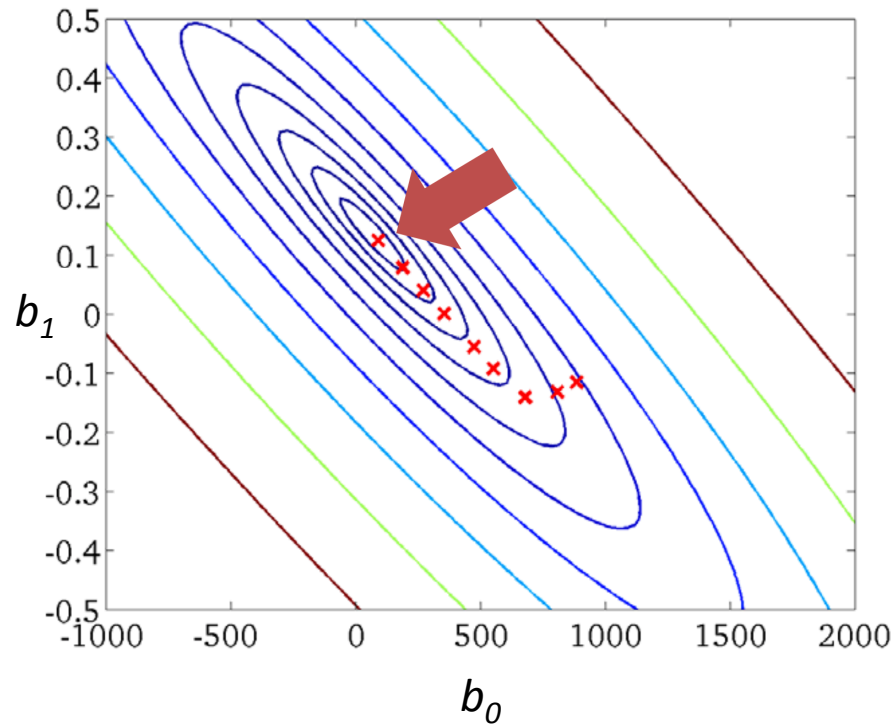
Cost Function



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

Gradient Descent: walking toward the minimum



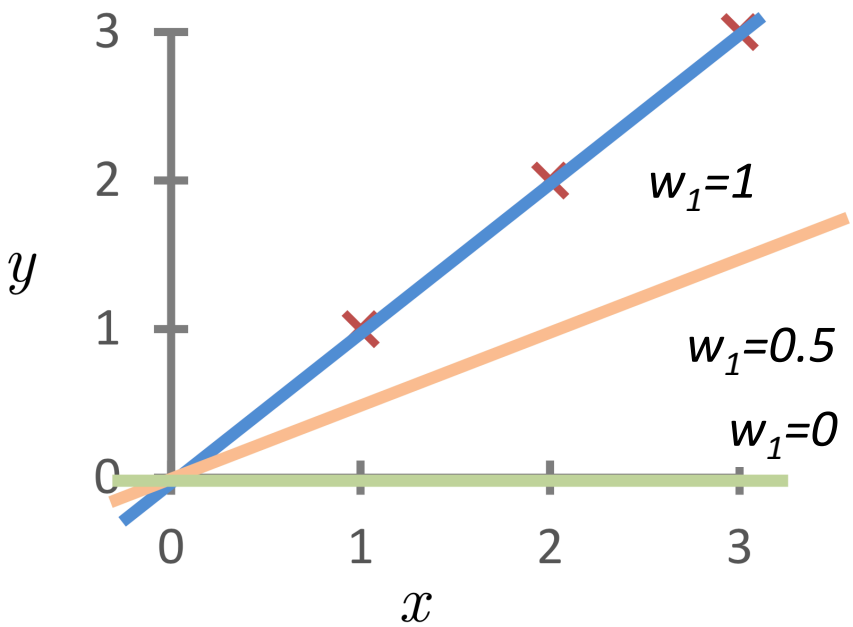


Slide: modified from Jessica Wu
Based on slide by Andrew Ng

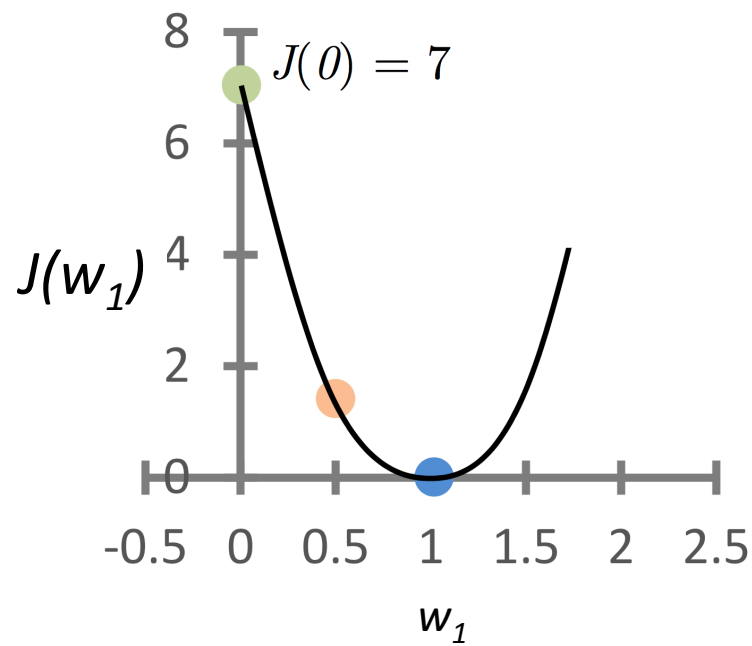
Cost Function (extra practice)

$h_w(x)$

(assume $w_0=0$ for this example)



$J(w_1)$

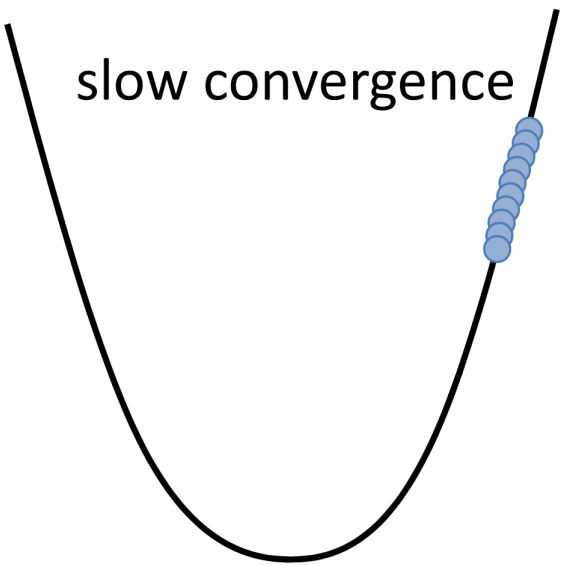


$$J(0.5) = \frac{1}{2}\left[(0.5-1)^2 + (1-2)^2 + (1.5-3)^2\right] = 1.75$$

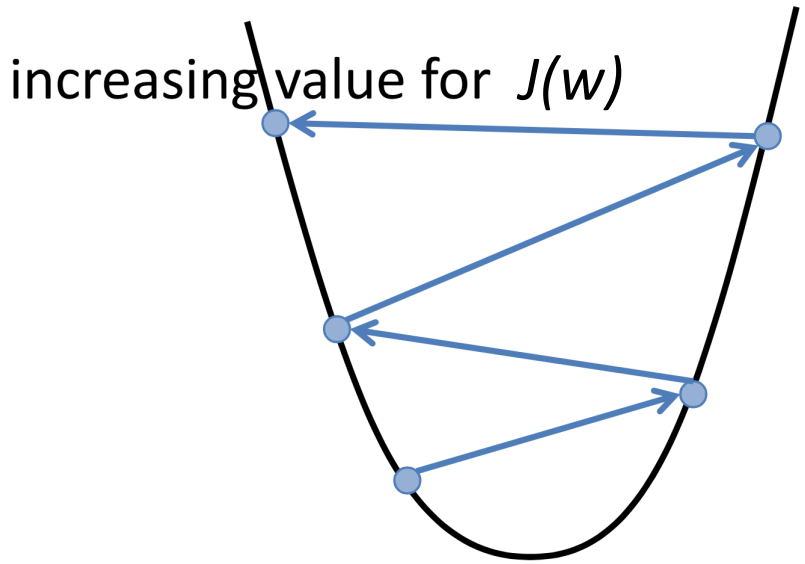
Slide: modified from Jessica Wu
Based on slide by Eric Eaton
[Example by Andrew Ng]

Choosing 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 September 19

- Reading quiz #3
- Simple linear regression
- SGD (Stochastic Gradient Descent)
- Normal equations solution
- Week 3 feedback

Normal Equations

$$J(\vec{\omega}) = \frac{1}{2}\sum_{i=1}^{n}\Big(\underbrace{\vec{\omega}^T\vec{x}_i}_{\text{pred}} - \underbrace{y_i}_{\text{truth}}\Big)^2$$

cost function

scalar

$$= \boxed{\frac{1}{2}\underbrace{\left(X\vec{\omega} - \vec{y}\right)^T}_{(1\times n)}\underbrace{\left(X\vec{\omega} - \vec{y}\right)}_{(n\times 1)}}$$

technically!

$$\underbrace{\begin{bmatrix} 1 & \text{---} & \vec{x}_1^{\,T} & \text{---} \\ 1 & \text{---} & \vec{x}_2 & \text{---} \\ \vdots & & \vdots & \\ 1 & \text{---} & \vec{x}_n & \text{---} \end{bmatrix}}_{X}_{n\times(p+1)} \underbrace{\begin{bmatrix} \omega_0 \\ \omega_1 \\ \vdots \\ \omega_p \end{bmatrix}}_{(p+1)\times 1} = \underbrace{\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix}}_{n\times 1}$$

$$1\cdot\omega_0 + x_{1}\cdot\omega_1 + x_{2}\cdot\omega_2 \cdots x_{p}\cdot\omega_p$$

take derivative
$\Rightarrow$ set equal to 0

$$\boxed{\hat{\vec{\omega}} = \underbrace{\left(X^TX\right)^{-1}}_{O(p^3)\ \Rightarrow\ \text{slow.}} X^T\vec{y}}$$

normal equation

closed form solution.

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{\underbrace{ad-bc}_{\text{det.}}}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

$$\begin{bmatrix} 1 & 3 \\ 1 & 3 \end{bmatrix}^{-1} = \frac{1}{3-3}\left(\quad\right)$$

AHH!!!

Computer Science

polynomial regression

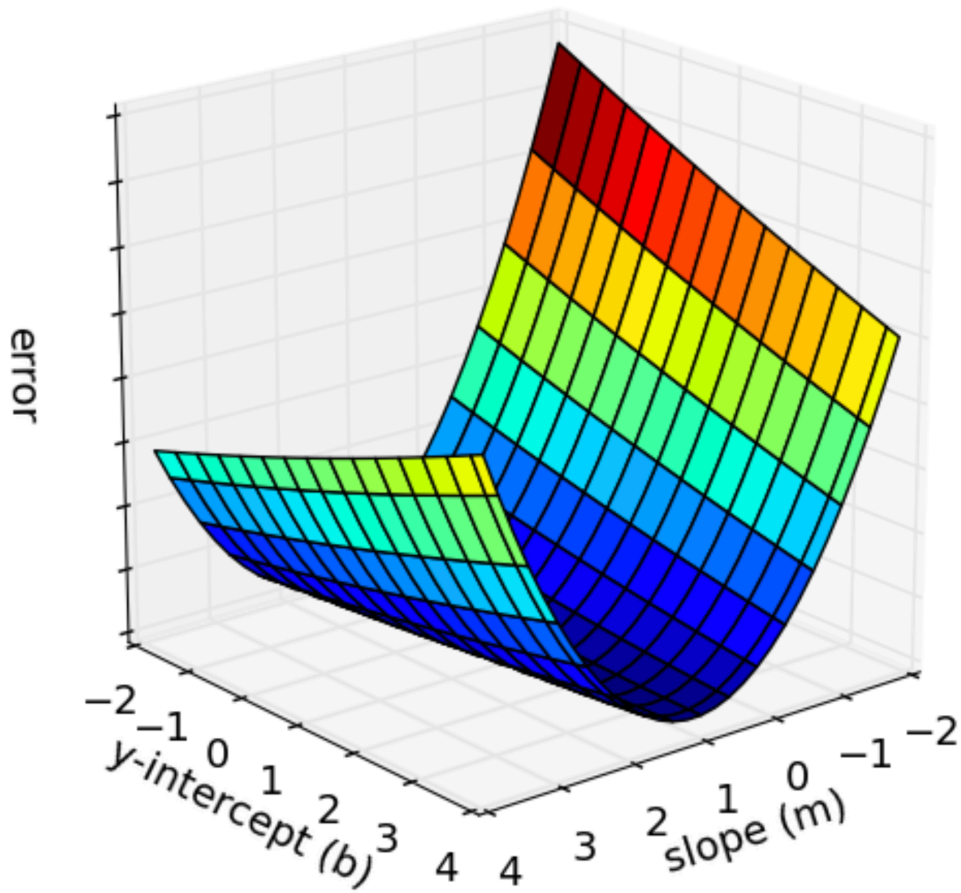
d degree polynomial.

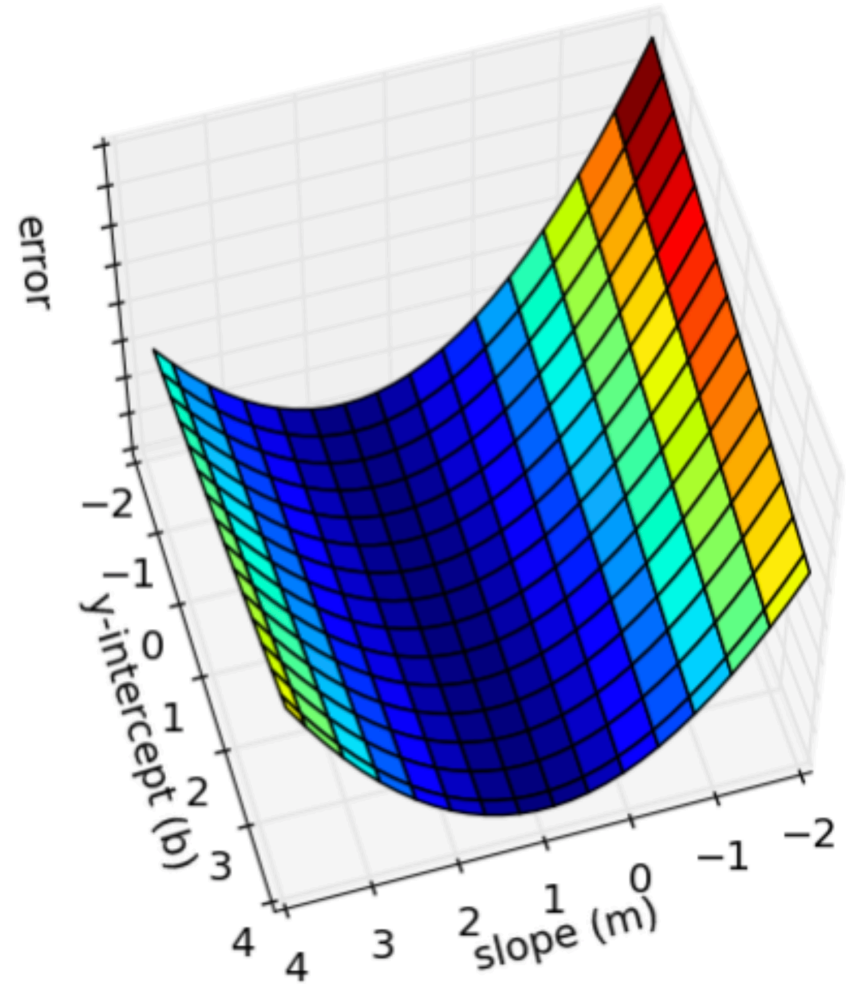
$\boxed{p=1}$ $h_{\vec{w}}(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + \cdots w_d x^d$

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

$J = 0.41$

Error as a function of slope & y-intercept





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

Image: Bryce Wiedenbeck

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

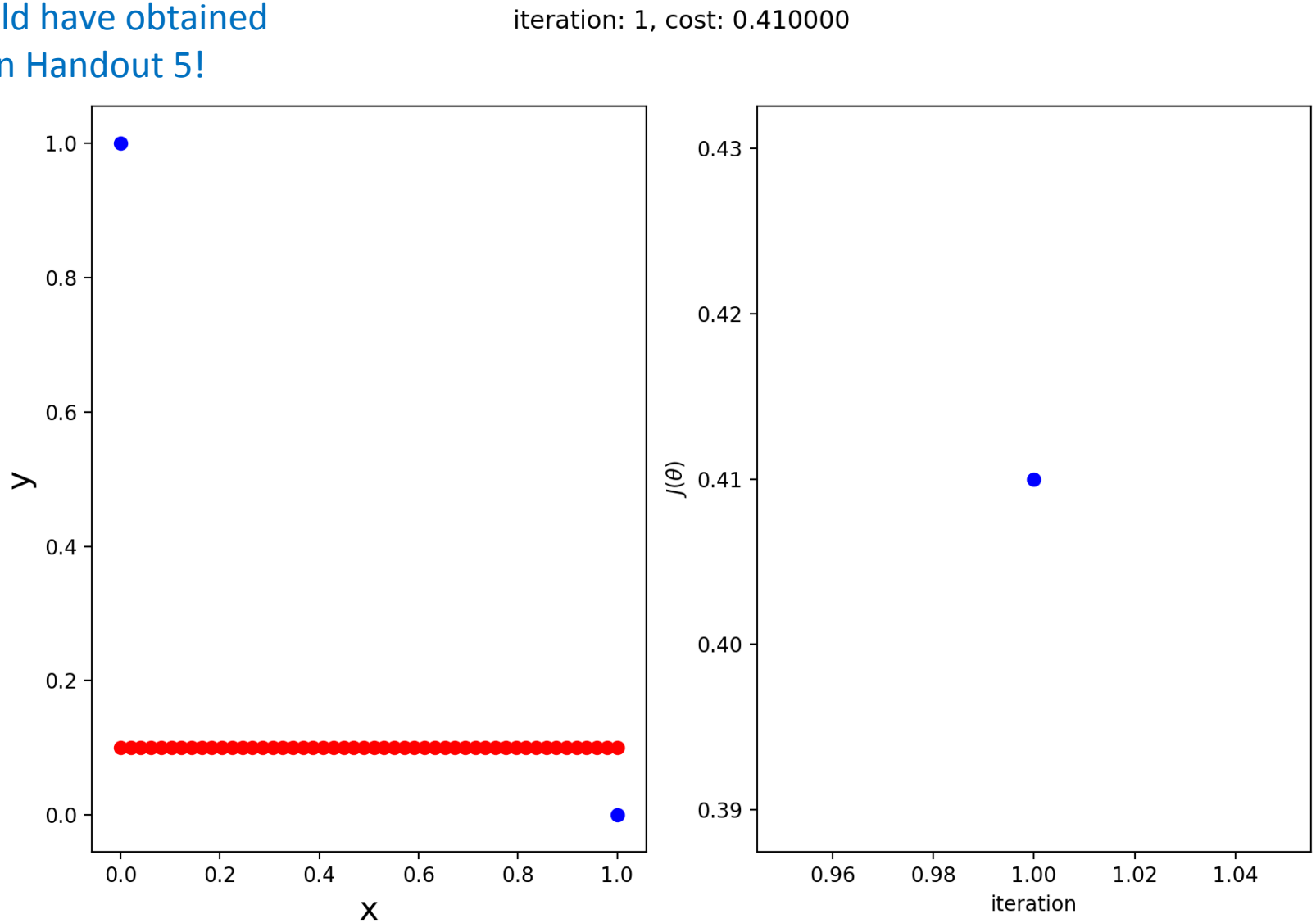
Outline for September 19

- Reading quiz #3
- Simple linear regression
- SGD (Stochastic Gradient Descent)
- Normal equations solution
- Week 3 feedback

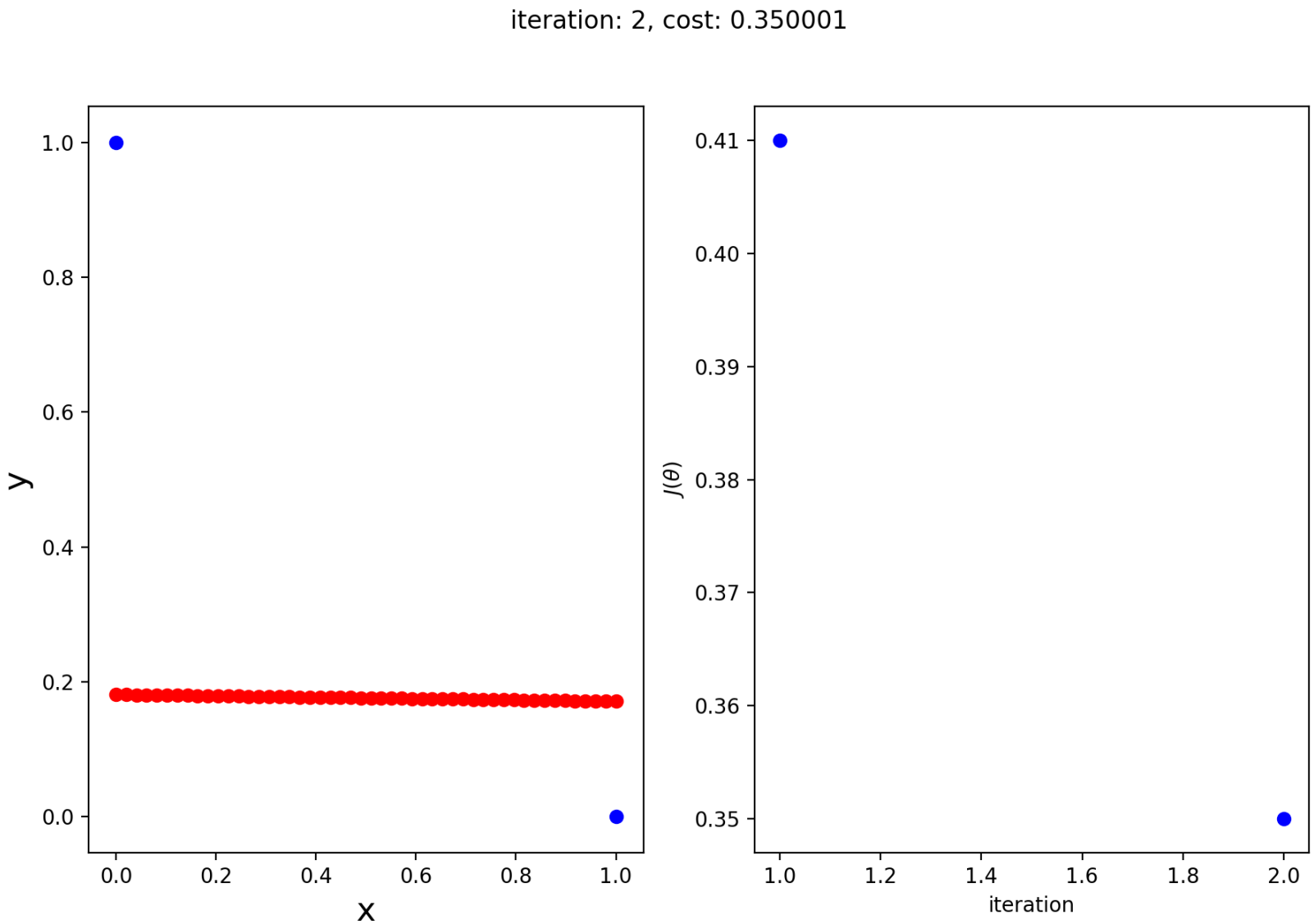
Lab 3 applied to Handout 5

Toy example, iteration 1

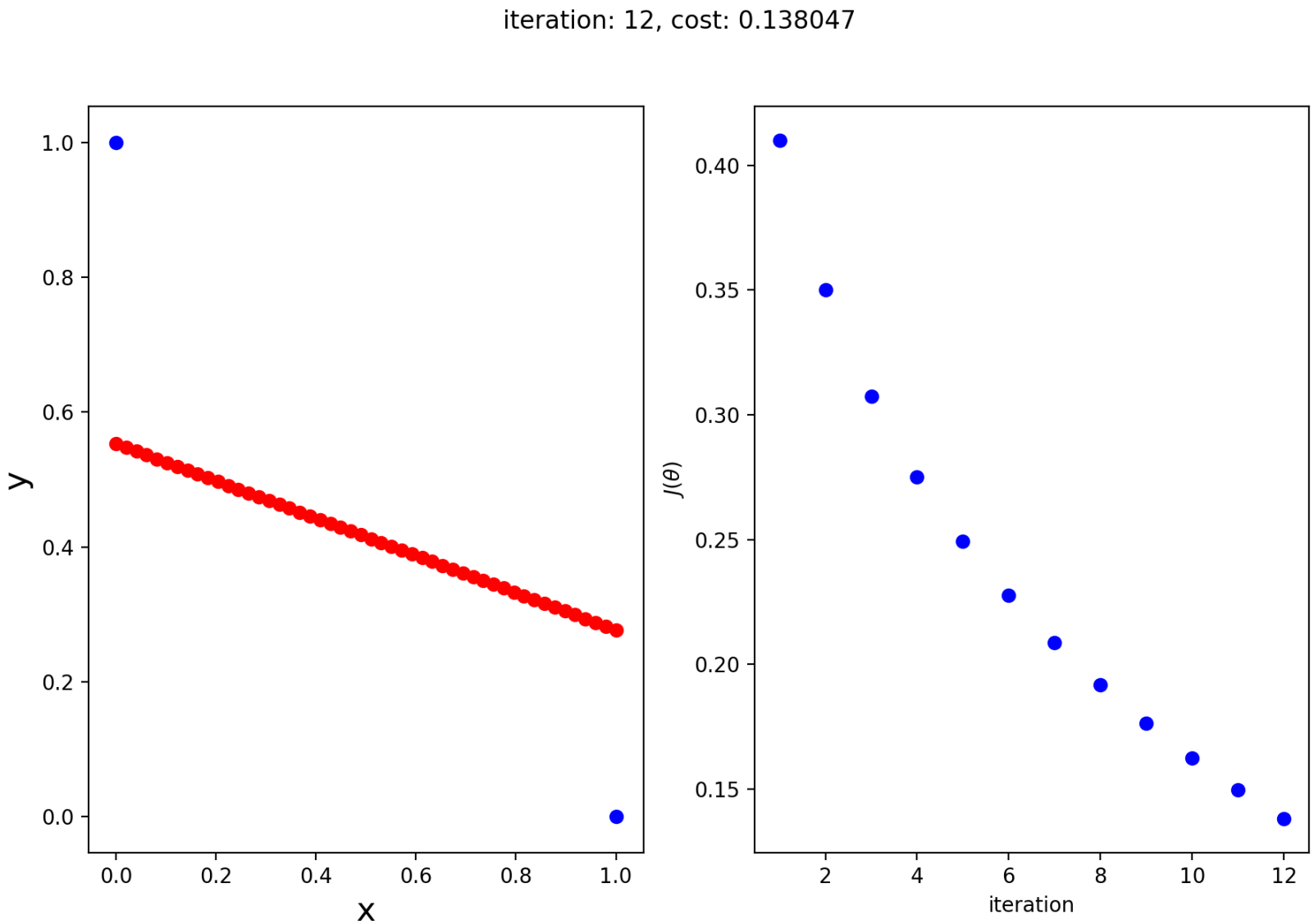
This is what you should have obtained in Handout 5!



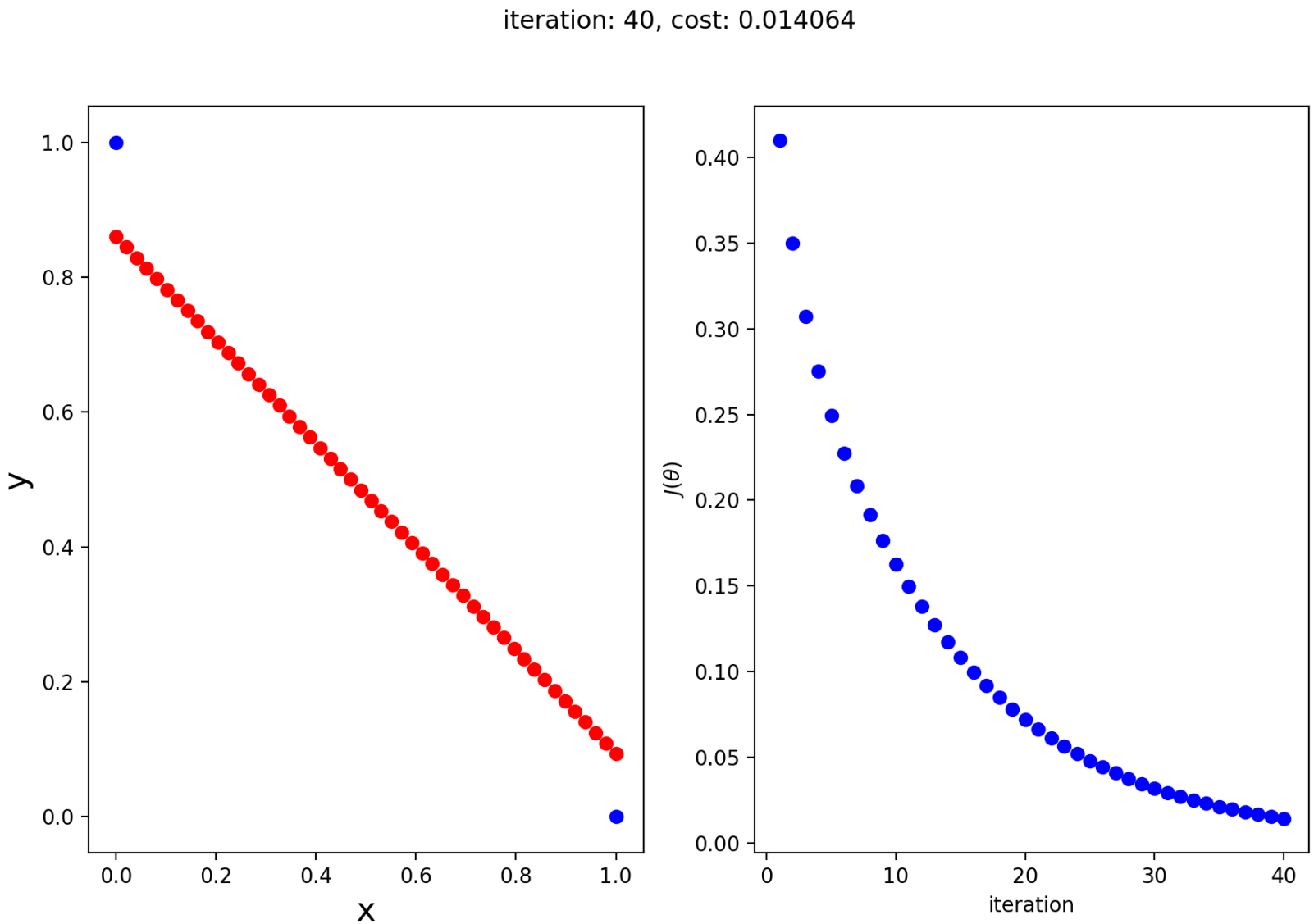
Toy example, iteration 2



Toy example, iteration 12



Toy example, iteration 40



Toy example, iteration 100