CS 383: Machine Learning

Prof Adam Poliak

Fall 2024

09/25/2024

Lecture 09

Announcements

HW03 is due Tuesday night

- **Reading quiz: Thursday**
    - Duame 9.3 (2 pages)

Proposed updated schedule

Midterm 1 was Thursday October 3rd

3 lectures this week

lecture on Wednesday 10/02 but no lecture on Thursday 10/03 (was supposed to be midterm 1)

No lecture Wednesday 10/09

HW02 decision trees due tonight, HW03 polynomial regression due next Tuesday 10/01, HW04 naive Bayes due Tuesday 10/08 (it’ll be a shorter assignment)

Midterm 1 on Thursday 10/10

Outline

Normal equations vs SGD

Regularization

Probability

Naive Bayes

Pros and Cons

Gradient Descent

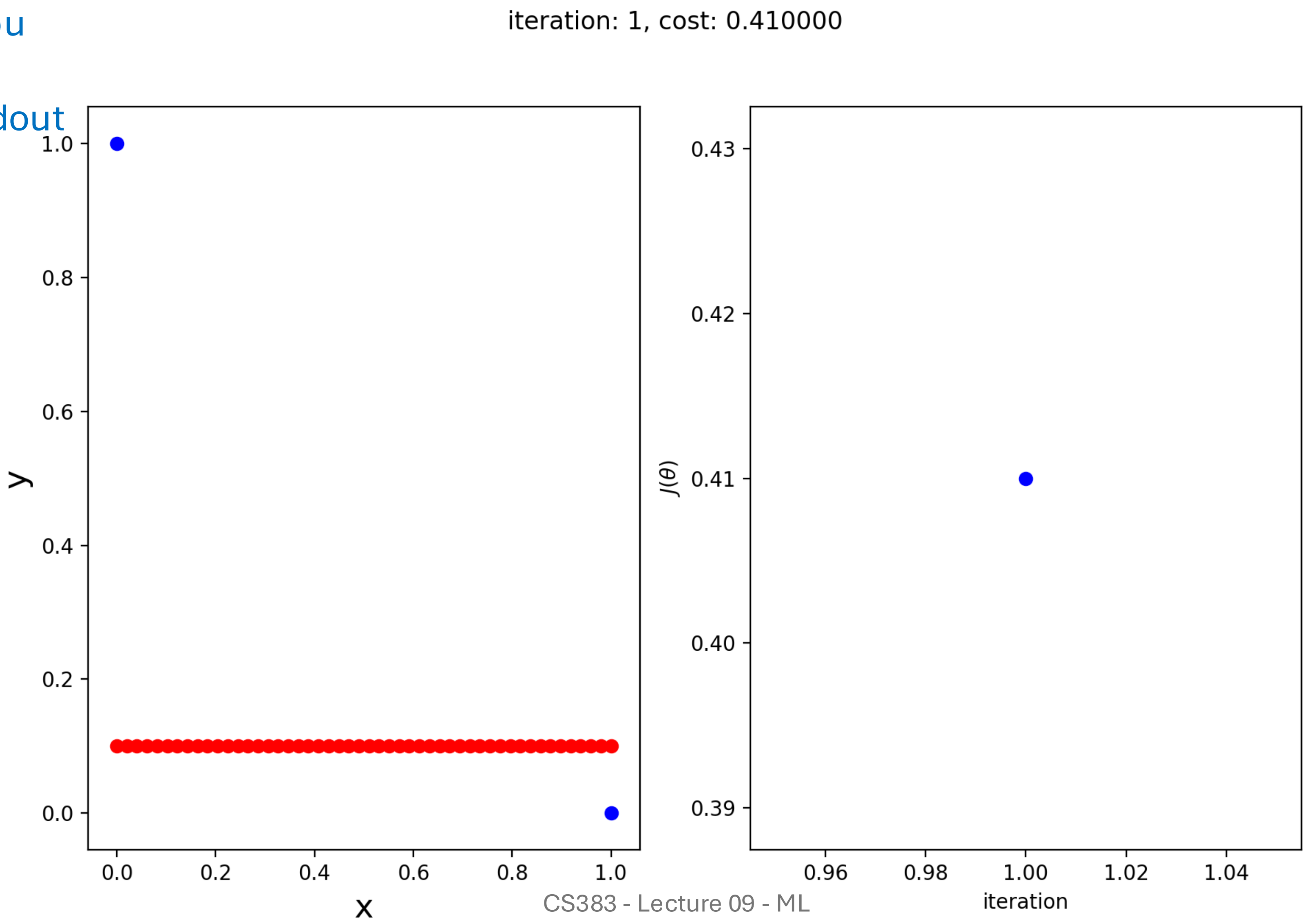
- Requires multiple iterations
- Need to choose η
- Works well when *n* is large
- Can support online learning

Normal Equation

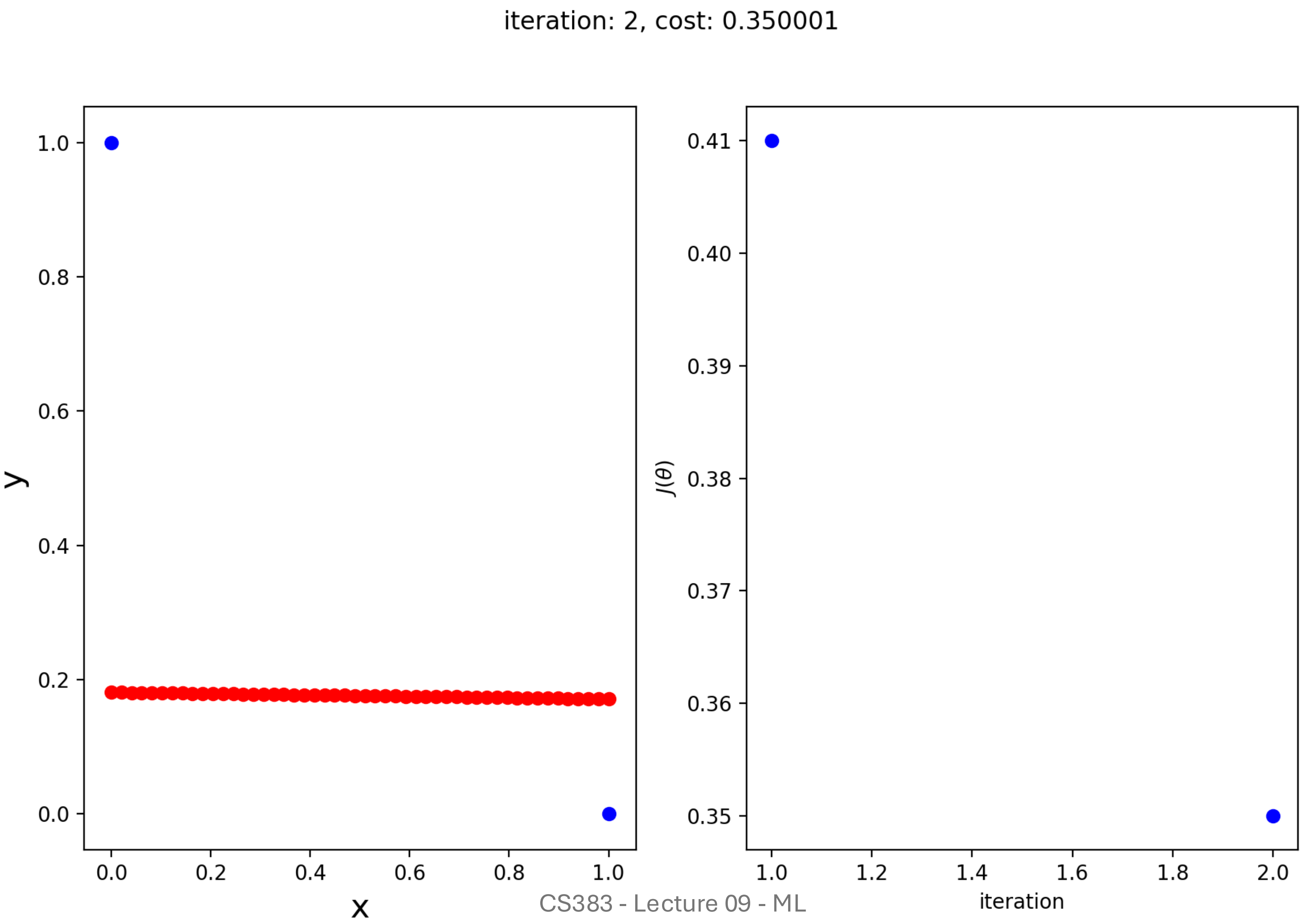
- Non-iterative
- No need to choose η
- Slow if *p* is large
  - Matrix inversion is $O(p^3)$

Toy example, iteration 1

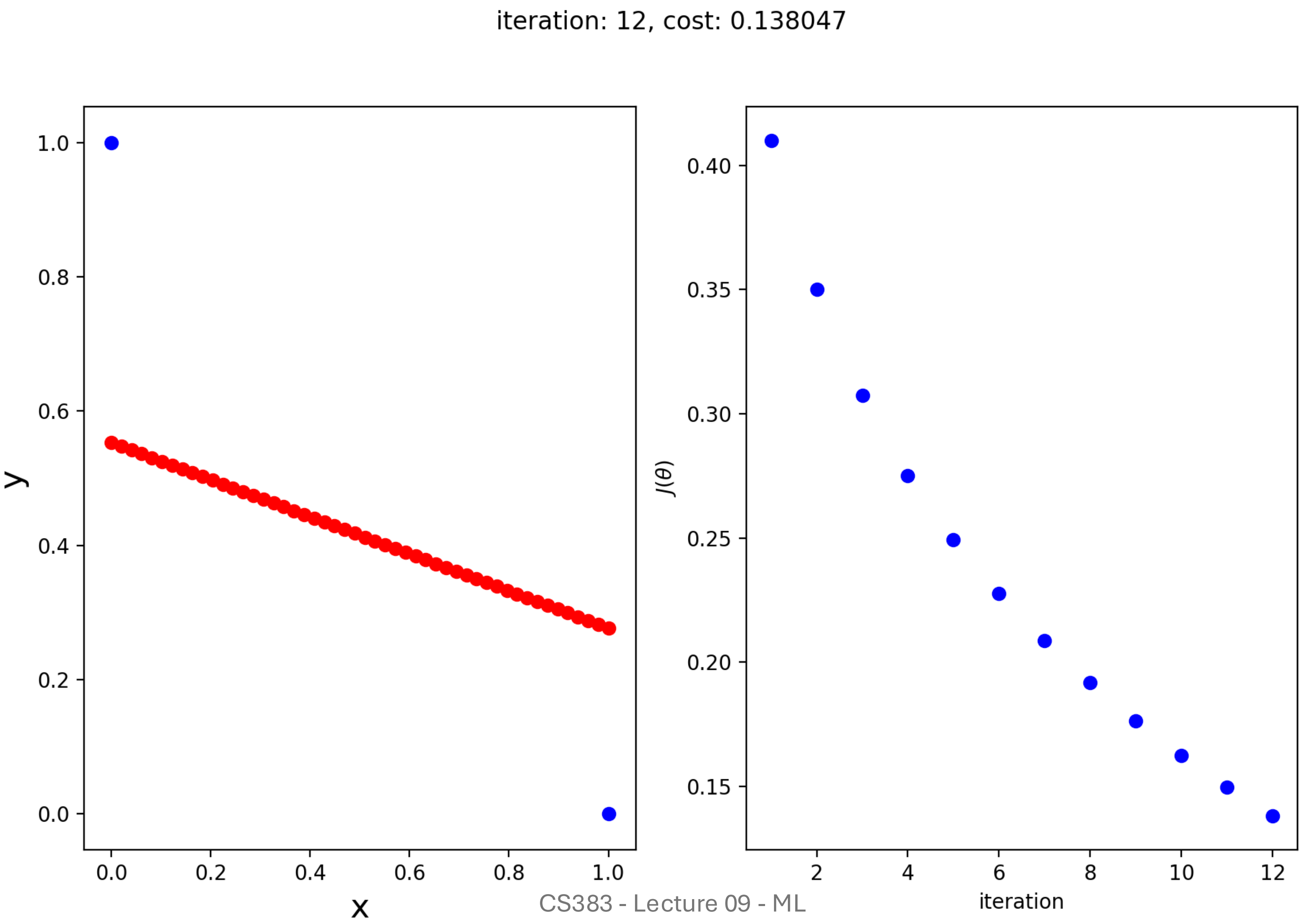
This is what you should have obtained in Handout 7!



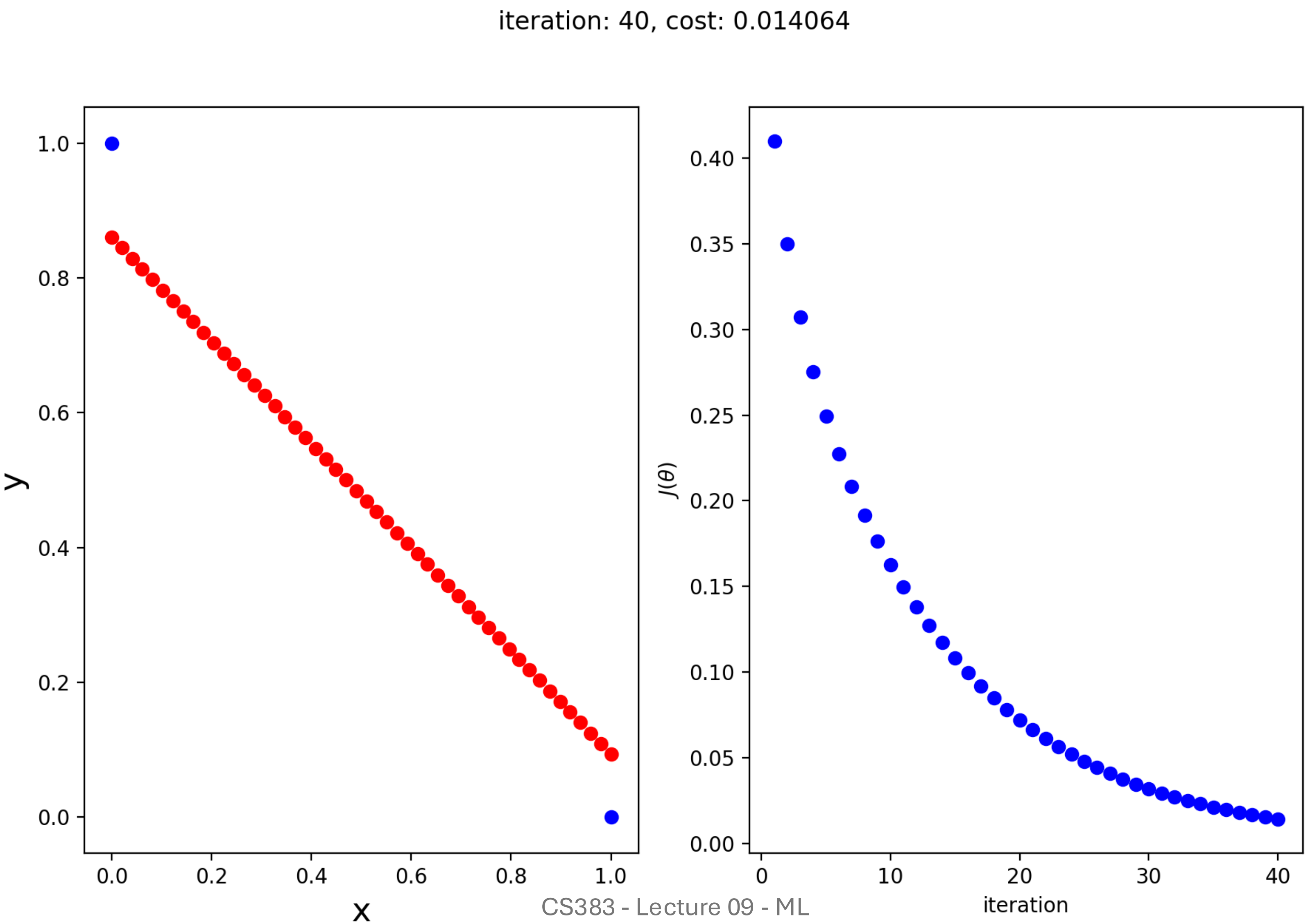
Toy example, iteration 2



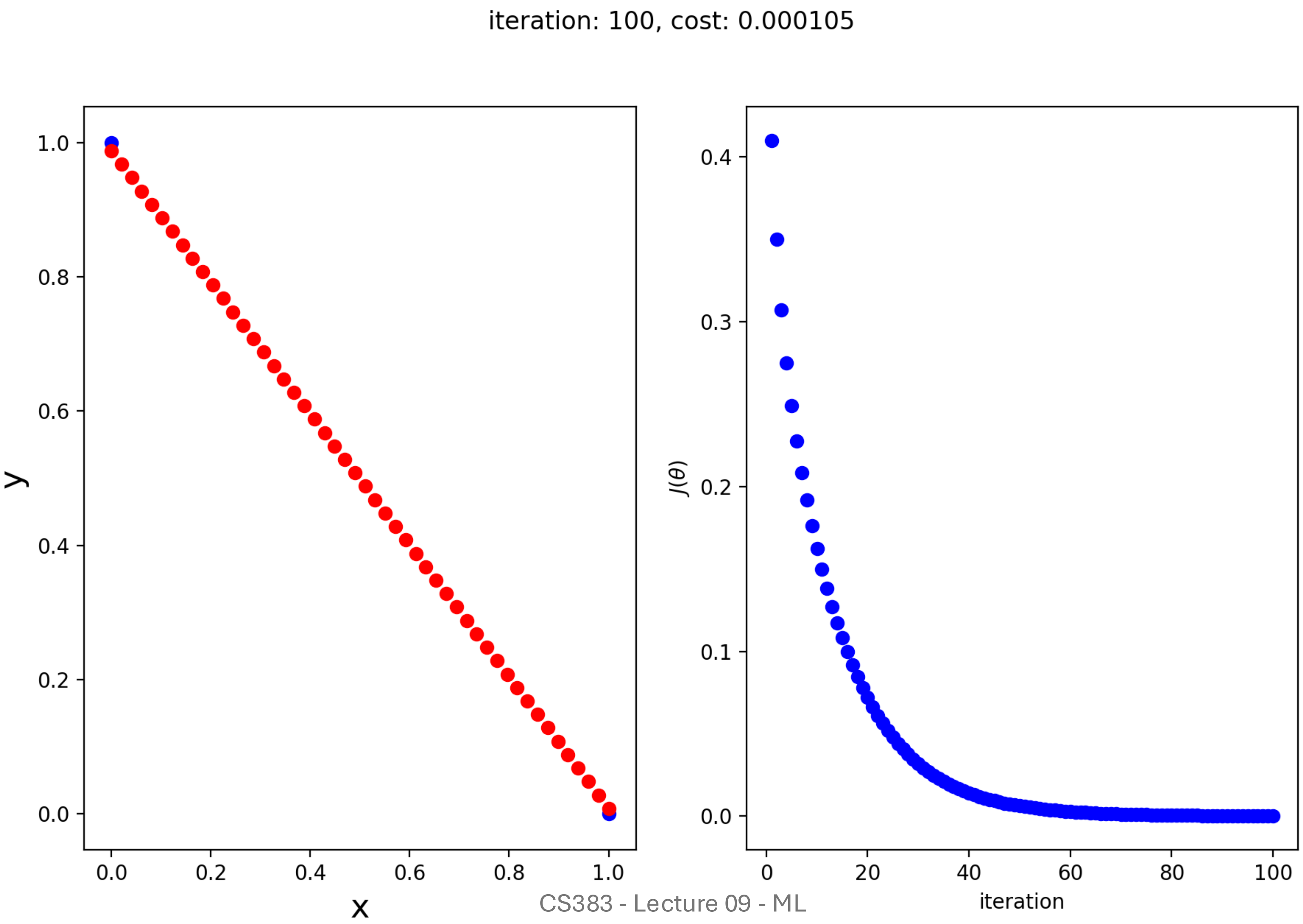
Toy example, iteration 12



Toy example, iteration 40



Toy example, iteration 100



Outline

Normal equations vs SGD

Regularization

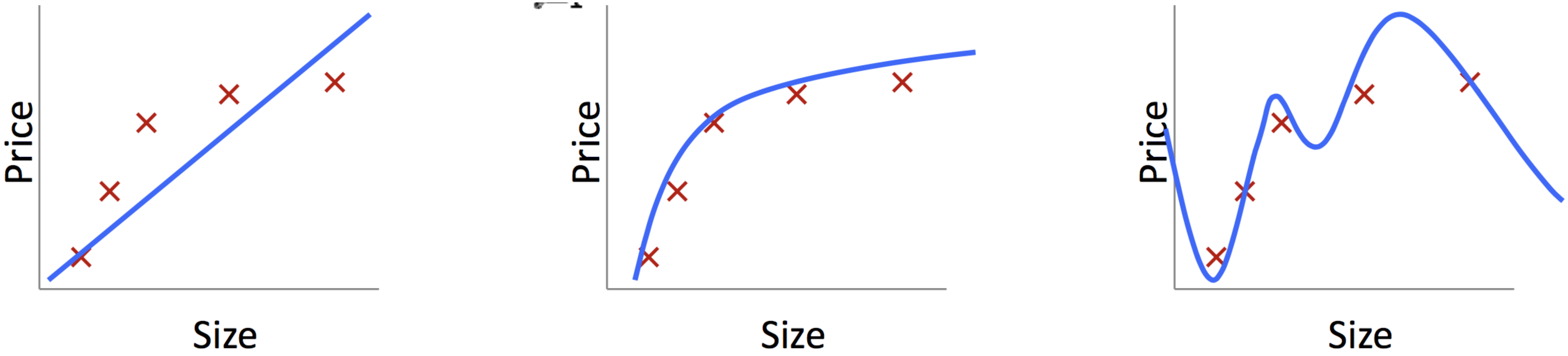
Probability

Naive Bayes

Adapted from slide by Jessica Wu
[example by Andrew Ng]

Generalization Error

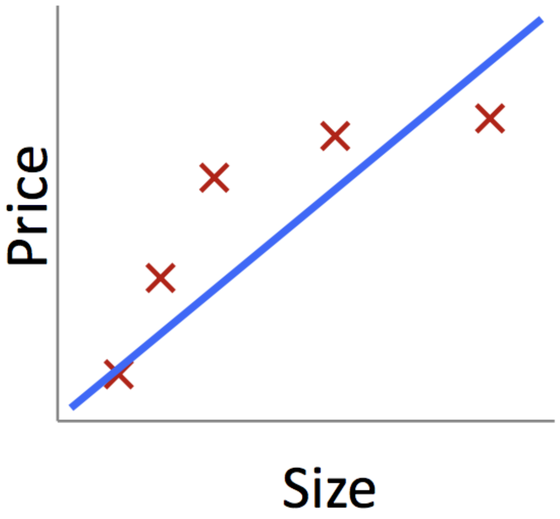
Example: price vs. size (i.e. of a house or car)



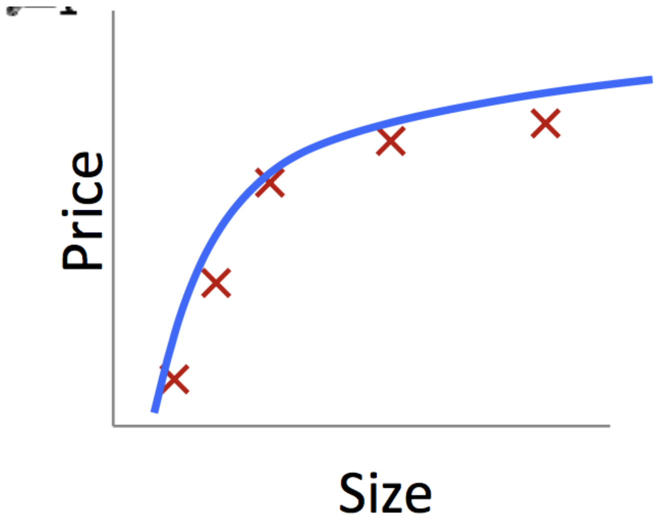
Adapted from slide by Jessica Wu
[example by Andrew Ng]

Generalization Error

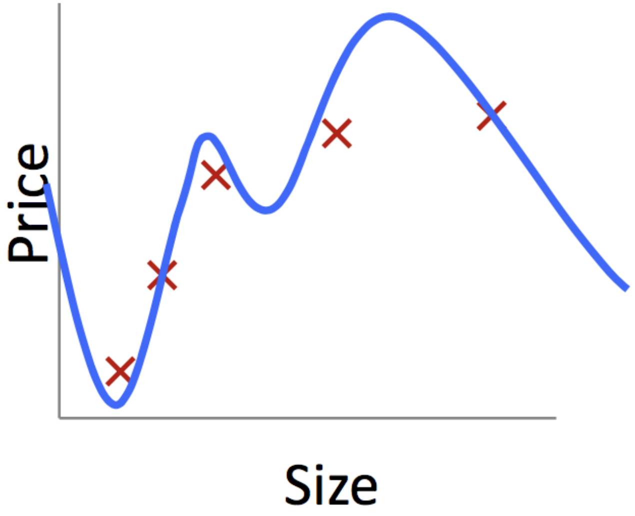
Example: price vs. size (i.e. of a house or car)



underfitting
(high bias)



correct fit



overfitting
(high variance)

Adapted from slide by Jessica Wu
[example by Andrew Ng]

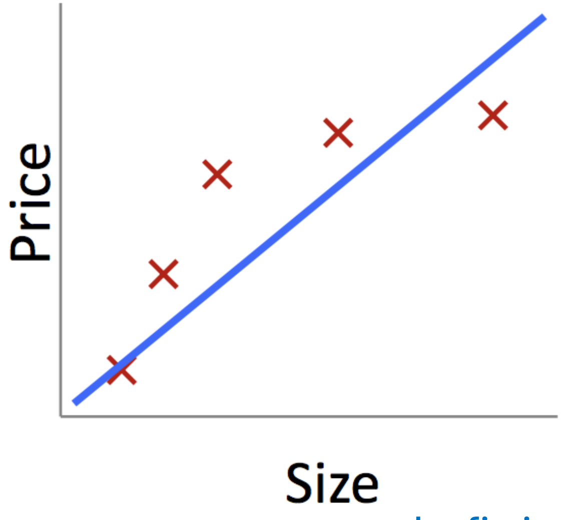
Generalization Error

Structural error:
Hypothesis space cannot model true relationship

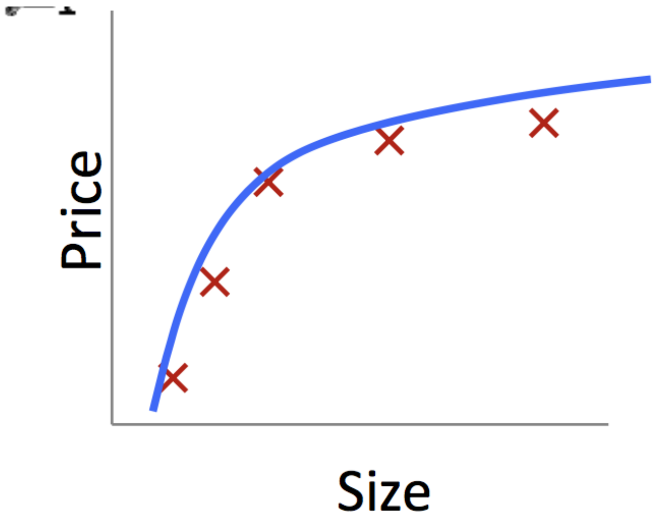
-More data doesn’t help
-Need a more flexible model

balance ⇔

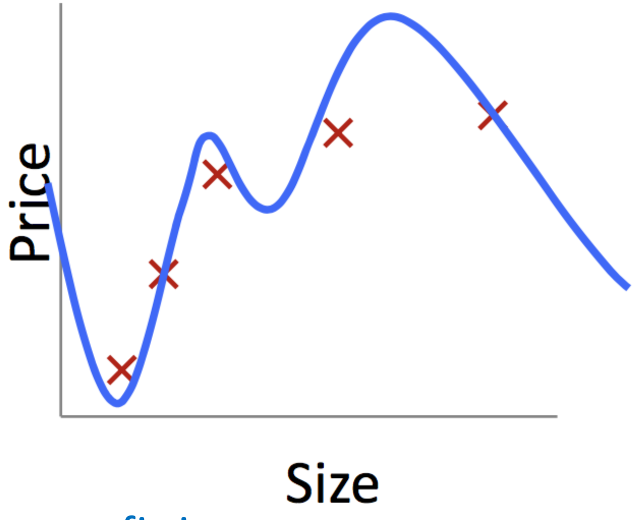
Estimation (approximation) error:
Hypothesis space *can* model true relationship, BUT hard to identify correct model due to large hypothesis space, small *n*, or noise
Reduce hypothesis space
Add more data



underfitting
(high bias)



correct fit



overfitting
(high variance)

Regularization

What if …

- we have a limited # of training examples (*n*<*p*), or
- we want to automatically control the complexity of the learned hypothesis?

Regularization

What if …

- we have a limited # of training examples (*n*<*p*), or
- we want to automatically control the complexity of the learned hypothesis?

Idea: penalize large values of w_j

Why prefer small weights?

- if large weights, small change in feature can result in large change in prediction
- prevent giving too much weight to any one feature
- might prefer zero weight for useless features

Common Regularizers

$$||\vec{w}||_0 = \sum_{j:w_j \neq 0} 1$$

L_0 **norm**

- Number of non-zero entries
- Minimizing L_0 norm is NP hard

$$||\vec{w}||_1 = \sum_{j=1}^{p} |w_j|$$

L_1 **norm**

- Sum of magnitude of weights
- Not differentiable

$$||\vec{w}||_2 = \sqrt{\sum_{j=1}^{p} w_j^2}$$

L_2 **norm**

- Sum of squared weights
- Differentiable

Slide adapted from Jessica Wu

Outline

Normal equations vs SGD

Regularization

Probability

Naive Bayes

Probability & Bayes Derivation

Bayes Rule

Conditional Probability

Marginal Probability

Bayes Rule

$$
\begin{aligned}
P(A, B) = & \\
&= P(A)P(A|B) \\
&= P(B)P(B|A)
\end{aligned}
$$

Hence:

$$P(A|B) = \frac{P(A|B)P(A)}{P(B)}$$

Joint & Conditional Probability

Joint Probability of Multiple Variables $P(A, B, C) =$

$$= P(C)P(A, B|\ C)$$

$$= P(C)P(B|C)P(A|B, C)$$

If A and B are independent:

$P(A, B) = P(A)P(B)$

If A and B are conditionally independent given C

$P(A, B|C) = P(A|C)P(B|C)$

If A, B, C are independent:

$P(A, B, C) = P(A)P(B)P(C)$