

# LECTURE 5

SPRING 2021

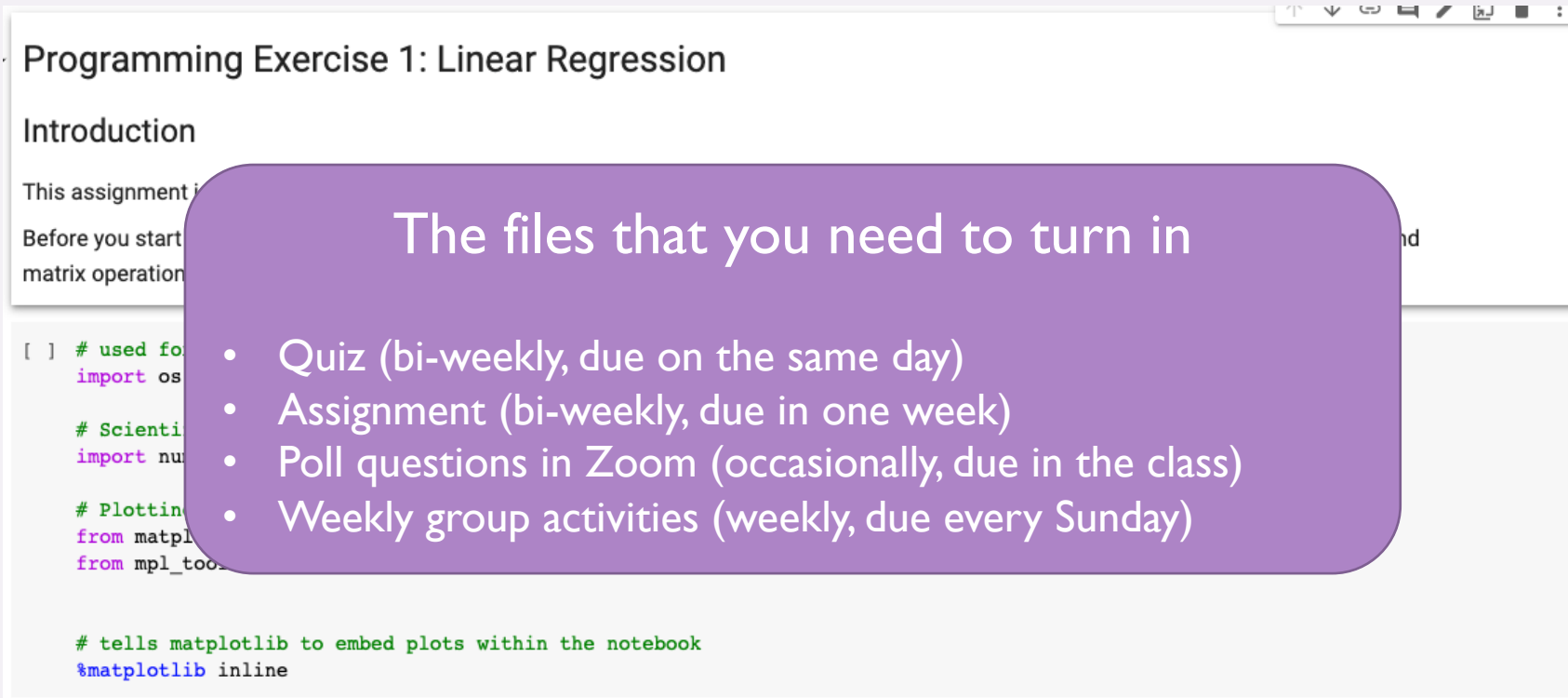
APPLIED MACHINE LEARNING

CIHANG XIE

# EXERCISES

- Google Colab Exercises (distributed during the lecture)

**NO NEED TO SUBMIT! JUST TO HELP YOU UNDERSTAND LECTURE**



The screenshot shows a Google Colab notebook interface. The title bar reads "Programming Exercise 1: Linear Regression". Below the title, the text "Introduction" is visible, followed by "This assignment" and "Before you start matrix operation". A code cell is partially visible, showing Python code for imports and plotting. A purple rounded rectangle is overlaid on the notebook, containing the text "The files that you need to turn in" and a bulleted list of activities.

Programming Exercise 1: Linear Regression

Introduction

This assignment

Before you start matrix operation

[ ] # used for  
import os

# Scientific  
import numpy

# Plotting  
from matplotlib  
from mpl\_toolkits

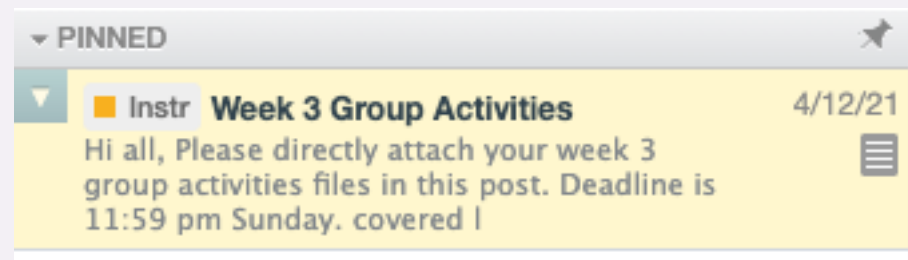
# tells matplotlib to embed plots within the notebook  
%matplotlib inline

The files that you need to turn in

- Quiz (bi-weekly, due on the same day)
- Assignment (bi-weekly, due in one week)
- Poll questions in Zoom (occasionally, due in the class)
- Weekly group activities (weekly, due every Sunday)

# GROUP ACTIVITIES

- ALL groups have submitted the files  
<https://piazza.com/class/kmmw9butjod4ay?cid=13>
- Please read notes & exercises from other groups
- Week 3 (lecture 4-5)



# TODAY

- Review of Gradient Descent Algorithm
- Choosing Learning Rate  $\alpha$
- Basis Functions

# LINEAR REGRESSION

# GRADIENT DESCENT

- Initialize  $\theta$
- Repeat until convergence

$$\theta_j \leftarrow \theta_j - \alpha \frac{\partial \text{Cost}(\theta)}{\partial \theta_j} \quad (\text{simultaneous update for } \theta_0, \theta_1, \dots, \theta_d)$$

- For linear regression:

$$\frac{\partial \text{Cost}(\theta)}{\partial \theta_j} = \frac{\partial}{\partial \theta_j} \frac{1}{2n} \sum_{i=1}^n (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

With  $h_{\theta}(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_d x_d = \sum_{j=0}^d \theta_j x_j$

# GRADIENT DESCENT

$$\frac{\partial Cost(\theta)}{\partial \theta_j} = \frac{\partial}{\partial \theta_j} \frac{1}{2n} \sum_{i=1}^n (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

$$= \frac{1}{2n} \frac{\partial}{\partial \theta_j} \sum_{i=1}^n (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

Scalar multiple rule

$$= \frac{1}{2n} \sum_{i=1}^n \frac{\partial}{\partial \theta_j} (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

Sum rule

$$= \frac{1}{2n} \sum_{i=1}^n 2(h_{\theta}(x^{(i)}) - y^{(i)}) \frac{\partial}{\partial \theta_j} (h_{\theta}(x^{(i)}) - y^{(i)})$$

Power rule

$$= \frac{1}{n} \sum_{i=1}^n (\sum_{k=0}^d \theta_k x_k^{(i)} - y^{(i)}) \frac{\partial}{\partial \theta_j} (\sum_{k=0}^d \theta_k x_k^{(i)} - y^{(i)})$$

$$= \frac{1}{n} \sum_{i=1}^n (\sum_{k=0}^d \theta_k x_k^{(i)} - y^{(i)}) x_j^{(i)}$$

# GRADIENT DESCENT

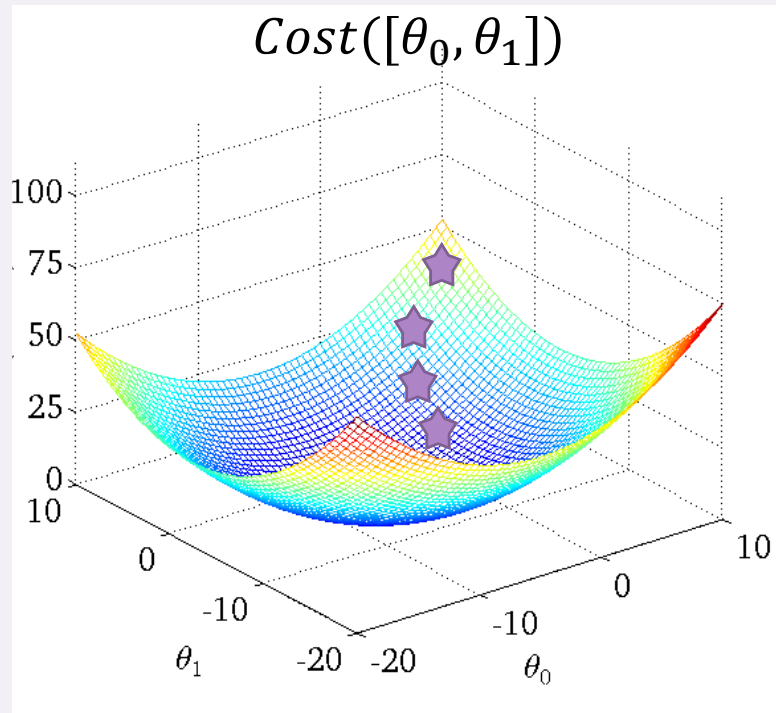
- Initialize  $\theta$

- Repeat until convergence

$$\theta_j \leftarrow \theta_j - \alpha \frac{\partial \text{Cost}(\theta)}{\partial \theta_j} \quad (\text{simultaneous update for } \theta_0, \theta_1, \dots, \theta_d)$$



# IN-CLASS QUIZ - WHICH ONE IS CORRECT?



Left:

$$\text{Temp0} \leftarrow \theta_0 - \alpha \frac{\partial Cost(\theta_0, \theta_1)}{\partial \theta_0}$$

$$\text{Temp1} \leftarrow \theta_1 - \alpha \frac{\partial Cost(\theta_0, \theta_1)}{\partial \theta_1}$$

$$\theta_0 \leftarrow \text{Temp0}$$

$$\theta_1 \leftarrow \text{Temp1}$$

Right:

$$\text{Temp0} \leftarrow \theta_0 - \alpha \frac{\partial Cost(\theta_0, \theta_1)}{\partial \theta_0}$$

$$\theta_0 \leftarrow \text{Temp0}$$

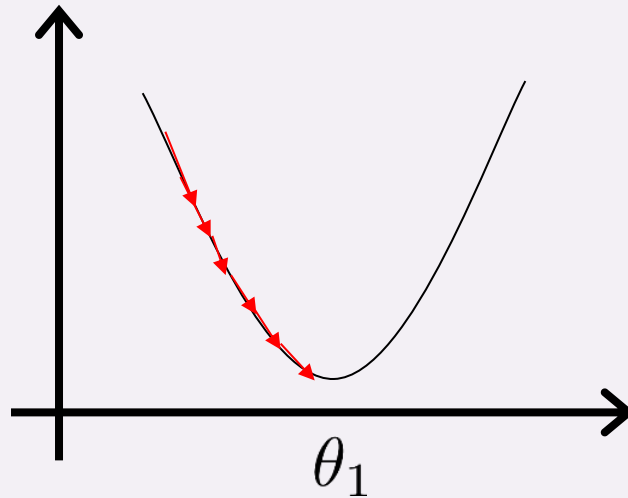
$$\text{Temp1} \leftarrow \theta_1 - \alpha \frac{\partial Cost(\theta_0, \theta_1)}{\partial \theta_1}$$

$$\theta_1 \leftarrow \text{Temp1}$$

# CHOOSE $\alpha$

$$\theta_j \leftarrow \theta_j - \alpha \frac{\partial \text{Cost}(\theta)}{\partial \theta_j} \quad (\text{simultaneous update for } \theta_0, \theta_1, \dots, \theta_d)$$

Learning rate

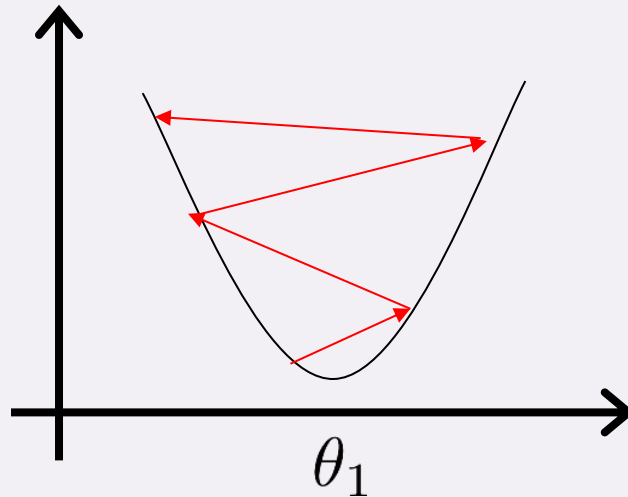


If  $\alpha$  is too small, gradient descent can be very slow.

# CHOOSE $\alpha$

$$\theta_j \leftarrow \theta_j - \alpha \frac{\partial \text{Cost}(\theta)}{\partial \theta_j} \quad (\text{simultaneous update for } \theta_0, \theta_1, \dots, \theta_d)$$

Learning rate



If  $\alpha$  is too large, gradient descent can overshoot the minimum. It may fail to converge, or even diverge.

# CHOOSE $\alpha$

$$\theta_j \leftarrow \theta_j - \alpha \frac{\partial \text{Cost}(\theta)}{\partial \theta_j} \quad (\text{simultaneous update for } \theta_0, \theta_1, \dots, \theta_d)$$

Learning rate

For certain functions  $\text{Cost}(\theta)$ , we can theoretically guarantee the convergence of gradient descent by choosing an appropriate  $\alpha$

If interested, please read machine learning course at UBC: lecture 4, starting from page 9  
<https://www.cs.ubc.ca/~schmidtm/Courses/540-WI18/L4.pdf>

# EXTENDING LINEAR REGRESSION TO MORE COMPLEX MODELS

- The inputs  $X$  for linear regression can be:
  - Original quantitative inputs
  - Transformation of quantitative inputs (log, exp, square, etc.)
  - Polynomial transformation (example:  $y = \beta_0 + \beta_1x + \beta_2x^2 + \beta_3x^3$ )
  - Interactions between variables (example:  $x_3 = x_1 \times x_2$ )
- This allows use of linear regression techniques to fit non-linear datasets.

# LINEAR BASIS FUNCTION MODEL

- Generally,

$$h_{\theta}(x) = \sum_{j=0}^d \theta_j \phi_j(x)$$

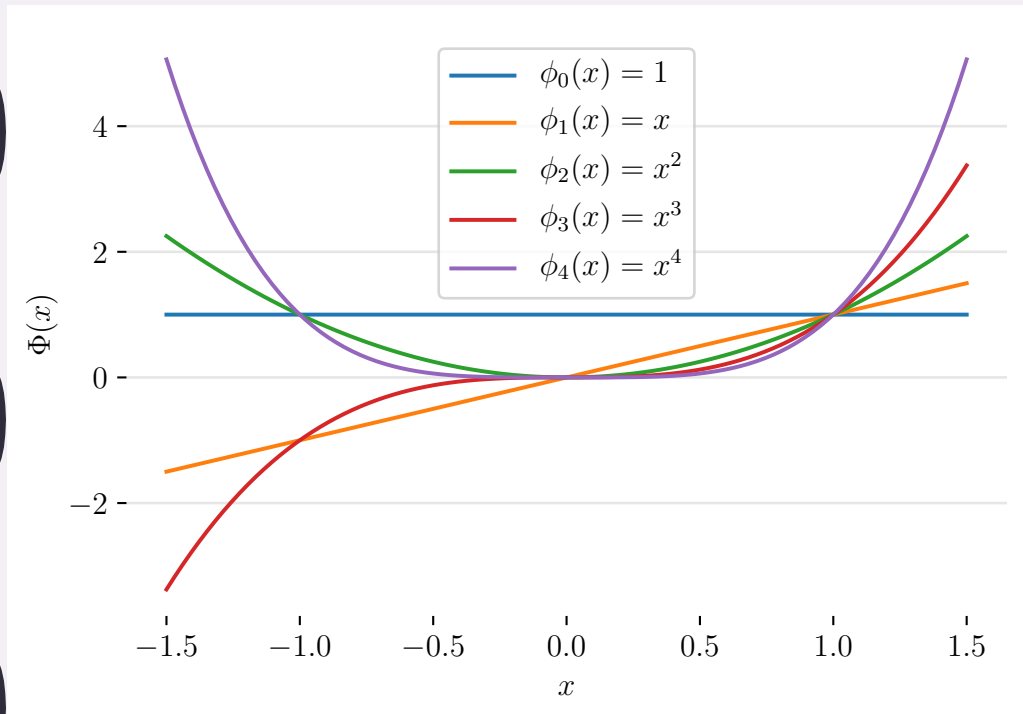
← Basis Function

- Typically,  $\phi_0(x) = 1$  so that  $\theta_0$  acts as a bias.
- In the simplest case, we can use linear basis function:

$$\phi_j(x) = x_j$$

- Polynomial basis function:  $\phi_j(x) = x^j$
- Gaussian basis function:  $\phi_j(x) = e^{-\frac{(x-\mu_j)^2}{2s^2}}$

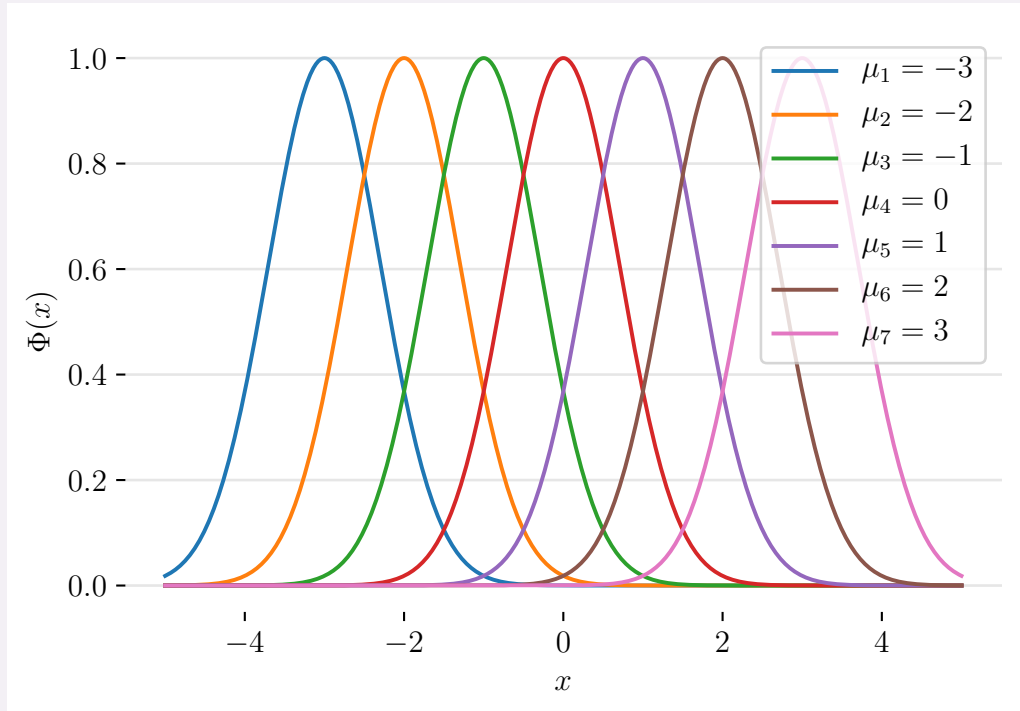
# EXAMPLE - POLYNOMIAL BASIS FUNCTION



(a) Polynomial basis out to degree 4.

$$y = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4 = \sum_{j=0}^4 \theta_j x^j$$

# EXAMPLE - GAUSSIAN BASIS FUNCTION



(a) Examples of Gaussian-type radial basis functions.

$$y = \theta_0 + \theta_1 e^{-\frac{(x-\mu_1)^2}{2s^2}} + \dots + \theta_7 e^{-\frac{(x-\mu_7)^2}{2s^2}}$$





# EXERCISE

[HTTPS://COLAB.RESEARCH.GOOGLE.COM/DRIVE/IVIFN\\_VBCFAXXAPHZW-WUGJRRGXGLFVFQ?USP=SHARING](https://colab.research.google.com/drive/IVIFN_VBCFAXXAPHZW-WUGJRRGXGLFVFQ?USP=SHARING)

A decorative graphic on the left side of the slide consisting of two parallel, wavy vertical lines. The inner line is a light purple color, and the outer line is a slightly darker shade of purple. They extend from the top to the bottom of the slide.

# QUESTIONS?

# HW1 (DUE 4/20)



**HW 1**

Due Apr 20 at 11:59pm | 60 pts

