# Midterm Review Session

**DSC 40A, Fall 2021 @ UC San Diego**

Suraj Rampure, with help from **many others**

# Agenda

- Bird’s eye view of the course.
- Select past exam problems.
- Review of select problems from Homeworks 1, 2, and 3.

# Bird’s eye view of the course

# What is this course about?

- So far, this course has really been about one thing: **learning from data**.
- The recipe:
  1. Choose a prediction rule.
  2. Choose a loss function.
  3. Minimize empirical risk, i.e. average loss, on your dataset to find the best predictions/parameters.
- Let's look at all of this a little more deeply.

# Choosing a prediction rule

In lecture, we've studied two prediction rules in depth:

- The **constant hypothesis**, $h$ (Lectures 1-5).
  - We didn't call it a "prediction rule" at the time, but it is one.
  - Equivalent to saying $H(x) = h$, i.e. we predict the same output for everyone.
- The **simple linear** prediction rule, $H(x) = w_0 + w_1 x$ (Lectures 6-7).
  - Now, predictions vary depending on $x$ ($x$ is called a **feature**).
- You also looked at $H(x) = w_1 x$ in Homework 3, 1f.

# Some questions...

- Suppose I've chosen to use a constant prediction rule $h$. Which $h$ do I use?
- Suppose I've chosen to use a simple linear prediction rule $H(x) = w_0 + w_1 x$. What should $w_0$ and $w_1$ be?
- **Answer:** Loss functions can help us.

# Loss functions

- A **loss function** $L(h, y)$ measures how a prediction $h$ is from the truth $y$.

- We've seen several loss functions so far:
  - Absolute loss: $L(h, y) = |y - h|$.
  - Squared loss: $L(h, y) = (y - h)^2$.
  - UCSD loss: $L(h, y) = 1 - e^{-(y-h)^2/\sigma^2}$.
  - 0-1 loss: $L(h, y) = 0$ if h = y, otherwise 1.

|actual−predicted|

absolute error

squared error

- Different loss functions have different properties, the key ones being their ease of minimization and their robustness to outliers.

# Empirical risk

- Loss functions are great — but they only measure the quality of a single prediction for a single true value.
- In order to get a sense of the quality of a prediction on our entire dataset, we must take the average of our chosen loss function over our entire dataset.
- The result is called empirical risk:

$$R(h) = \frac{1}{n} \sum_{i=1}^{n} L(h, y_i)$$

function of $h$!

  - If using absolute loss, $R$ is called **mean absolute error**.
  - If using squared loss, $R$ is called **mean squared error**.

# The constant hypothesis, $h$

To find the best $h$ (denoted as $h^*$) to make constant predictions with, we need to choose a loss function.

- If we choose absolute loss, the resulting empirical risk (i.e. mean absolute error) is

$$R_{abs}(h) = \frac{1}{n} \sum_{i=1}^{n} |y_i - h|$$

- If we choose squared loss, the resulting empirical risk (i.e. mean squared error) is

$$R_{sq}(h) = \frac{1}{n} \sum_{i=1}^{n} (y_i - h)^2$$

- We also looked at the resulting empirical risk when we choose 0-1 loss and UCSD loss.

# The simple linear prediction rule, $H(x) = w_0 + w_1 x$


$w_0$ and $w_1$ are called **parameters**. To find the **optimal parameters** (denoted as $w_0^*$ and $w_1^*$), we again need a loss function

- We could choose absolute loss — see Homework 3, Q3.
  - The resulting problem is called "least absolute deviations regression."
- The more common choice, though, is squared loss:

$$R_{sq}(w_0, w_1) = \frac{1}{n} \sum_{i=1}^{n} (y_i - (w_0 + w_1 x_i))^2$$

  - This problem is called **least squares regression**.

# Minimizing empirical risk

After choosing a prediction rule and loss function, and writing out the corresponding empirical risk, we need to **minimize** the empirical risk to find the best predictions/parameters.
Some ways we've minimized empirical risk:

- Calculus.
- Other algebraic arguments.
- Gradient descent.

# Minimizing empirical risk with calculus

- Strategy: take derivative(s), set it to 0, and solve.
- Constant hypothesis, squared loss:

$$R_{sq}(h) = \frac{1}{n}\sum_{i=1}^{n}(y_i - h)^2 \implies h^* = \bar{x}$$

$L(h,y) = (y-h)^2$ → $h^* = \frac{n}{\sum_{i=1}^{n}\frac{1}{y_i}}$

- Simple linear prediction rule, squared loss:

$H(x) = \beta x$

$$R_{sq}(w_0, w_1) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (w_0 + w_1 x_i)\right)^2$$

$$\implies w_1^* = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = r\frac{\sigma_y}{\sigma_x} \qquad w_0^* = \bar{y} - w_1^*\bar{x}$$

- Several homework problems.
  - HW 2 Q3, HW 3 Q1f.

# Minimizing empirical risk with algebraic arguments

- Since absolute loss is not differentiable, the resulting empirical risk (mean absolute error) also isn't. We couldn't use calculus.

- For the constant hypothesis, $R_{abs}(h) = \frac{1}{n} \sum_{i=1}^{n} |y_i - h|$.

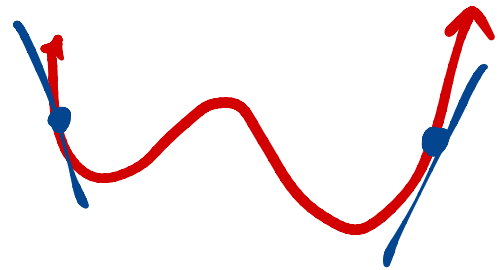
- We instead minimized $R_{abs}$ by finding a formula for the slope of $R$ at any $h$ (that isn't one of the $y_i$):

$$\text{slope of R at h} = \frac{1}{n} \left( \#(y_i < h) - \#(y_i > h) \right)$$

- The **median** is where the slope of $R$ goes from - to +; it minimizes $R_{abs}(h)$.

# Minimizing empirical risk using gradient descent

- Sometimes, even when our empirical risk is differentiable, there is no **closed-form solution** for the minimizing input.
  - Example: Empirical risk for $L_{ucsd}$.



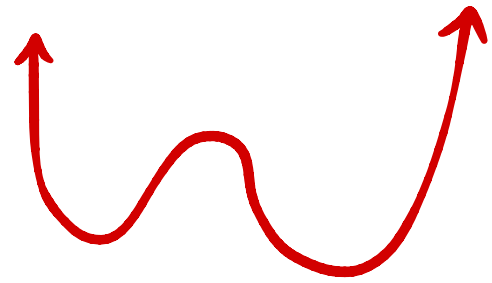
- Solution: **gradient descent**.

- Gradient descent tries to minimize a function $R(h)$ through an iterative process.
  - **Key idea**: Move opposite the direction of the slope.

  - Given an initial guess, $h_0$, for the minimizer and a step-size/learning rate $\alpha$, gradient descent updates are made with the update equation

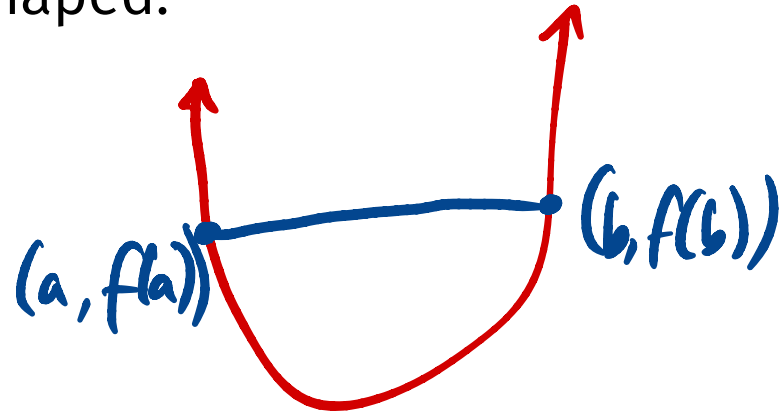
$$h_i = h_{i-1} - \alpha \cdot \frac{d}{dh} R(h_{i-1})$$

# Gradient descent



- **Key theorem:** Gradient descent is guaranteed to find the global minimum of a function if that function is **convex** and **differentiable**, given an appropriate step size.

- A function $f$ is **convex** if it is true that given any two inputs $a, b$, the line segment joining $(a, f(a))$ and $(b, f(b))$ does not go below the graph of $f$.
  - Convex functions are "bowl" shaped.
  - Second derivative test.

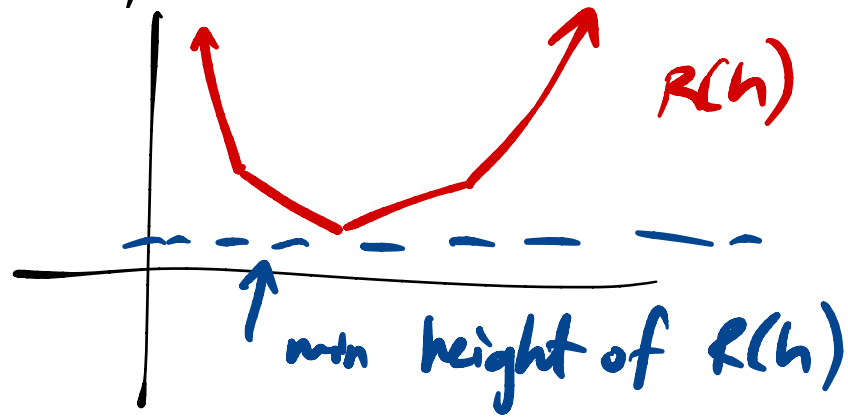


# Summary of key results

| Prediction rule | Loss function | Best prediction parameters |
|---|---|---|
| constant, $h$ | $\lvert y-h\rvert$ | $h^* = \text{median}$ |
| constant, $h$ | $(y-h)^2$ | $h^* = \text{mean}$ |
| constant, $h$ | $L(h,y) = \begin{cases} 0, & h=y \\ 1, & h \neq y \end{cases}$ | $h^* = \text{mode}$ |
| linear, $H(x) = w_0 + w_1 x$ | $(y-h)^2$ | $w_1^* = r\frac{\sigma_y}{\sigma_x}$<br>$w_0^* = \bar{y} - w_1^* \bar{x}$ |

# Other concepts — spread

- Different loss functions lead to empirical risk functions that, for the constant prediction rule, are minimized at various measures of **center**.
  - Absolute loss: **median**.
  - Squared loss: **mean**.
  - 0-1 loss: **mode**.




- The minimum value of these empirical risks (i.e. the lowest height on the graph of $R$) is a measure of the **spread** of the data.
  - Absolute loss: **mean absolute deviation from the median**.
  - Squared loss: **variance**.
  - 0-1 loss: **proportion of values not equal to the mode**.

$$R_{abs}(h) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}}_{\text{mean}} |y_i - h|$$

$h^* = \text{median}$

$$R_{abs}(h^*) = R_{abs}(\text{median}) = \overbrace{\frac{1}{n}\sum_{i=1}^{n}}^{\text{mean}} |y_i - \text{median}|$$

mean absolute deviation from the median

2, 4, 9

$$\frac{1}{3}\left[|2-4| + |4-4| + |9-4|\right]$$

$$R_{sq}(h) = \frac{1}{n}\sum_{i=1}^{n}(y_i - h)^2 \Rightarrow h^* = \text{mean}$$

$$R_{sq}(h^*) = R_{sq}(\text{mean}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \text{mean})^2 = \text{variance}$$

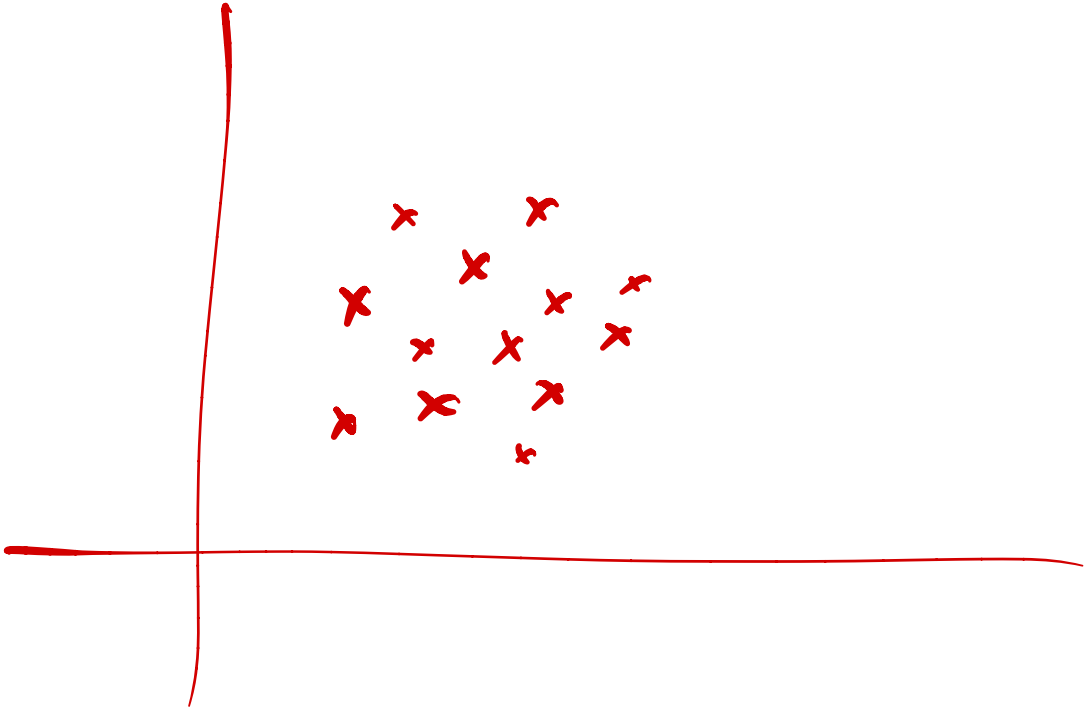
# Other concepts — correlation

- The **correlation coefficient**, $r$, is a measure of the **linear association between two variables**.

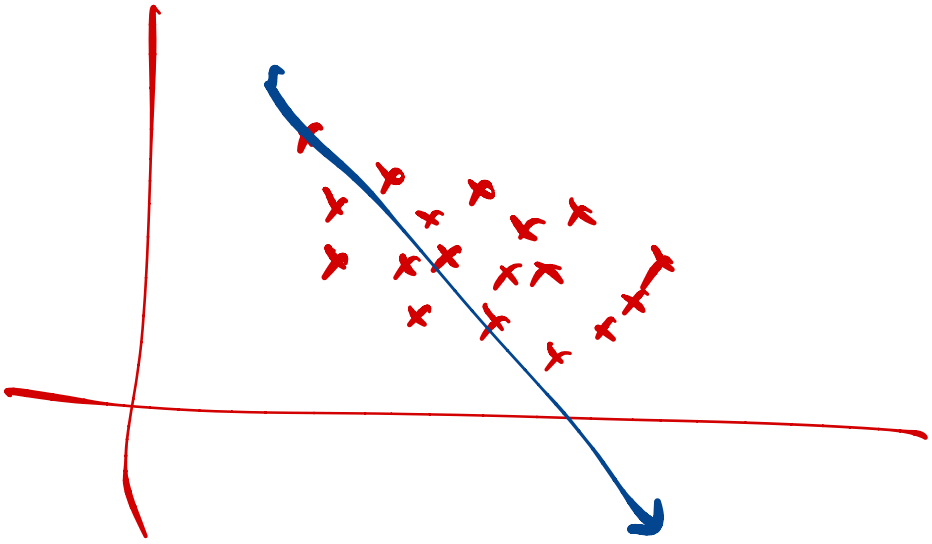
- It ranges between -1 and 1.
  - $r$ = 1 indicates a perfect positive linear association (x and y lie exactly on a straight line that is sloped upwards).
  - $r$ = –1 indicates a perfect negative linear association between $x$ and $y$.
  - The closer $r$ is to 0, the weaker the linear association between $x$ and $y$ is.

- $w_1^*$ can be written in terms of $r$:

$$w_1^* = r\frac{\sigma_y}{\sigma_x}$$



$r = 0$



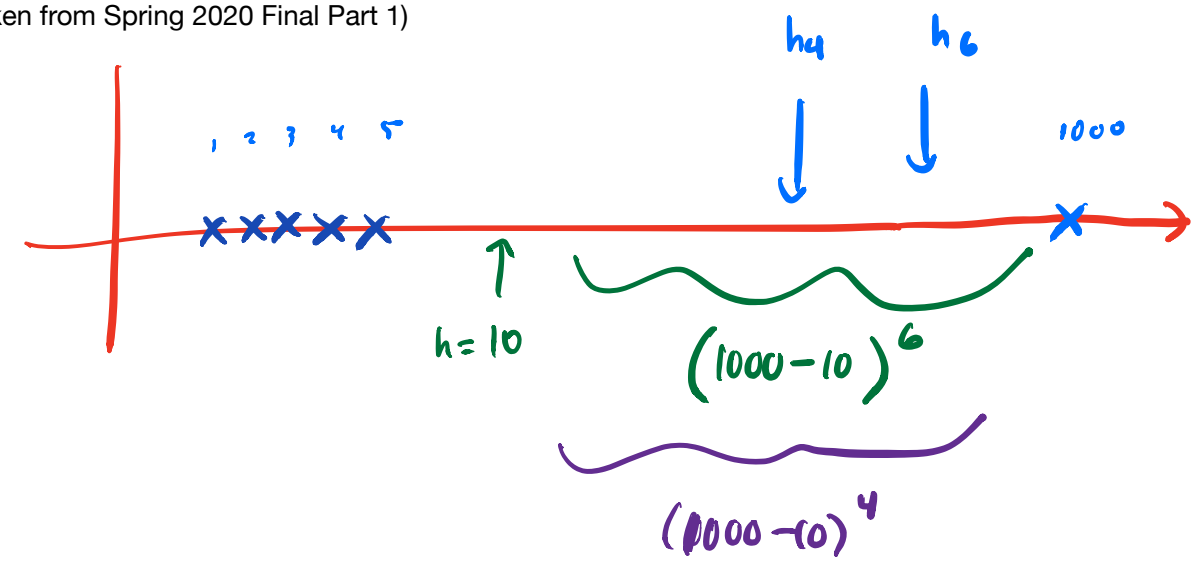
$r = -0.6$

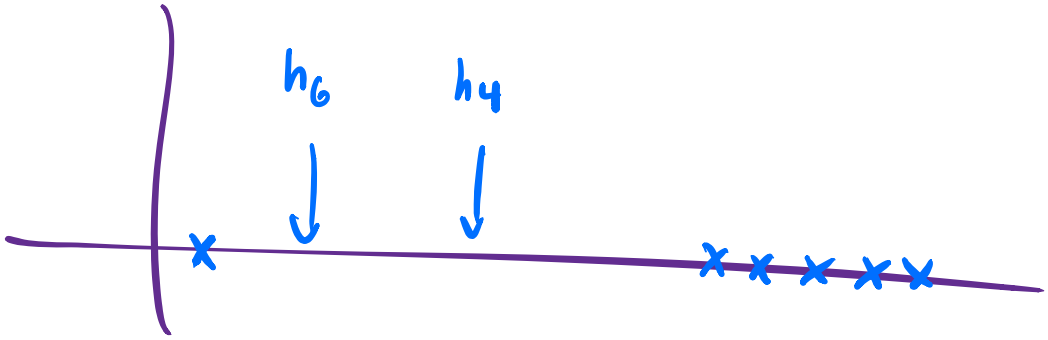
$0.5^6 < 0.5^4$

**Problem 2.** (10 points)

Suppose you have a data set $y_1, \ldots, y_n$ with one outlier whose value is significantly higher than the others. If we use empirical risk minimization to make a prediction, which choice of loss function would lead to a larger prediction, $L(h, y) = (h - y)^4$, or $L(h, y) = (h - y)^6$? Explain.

(taken from Spring 2020 Final Part 1)



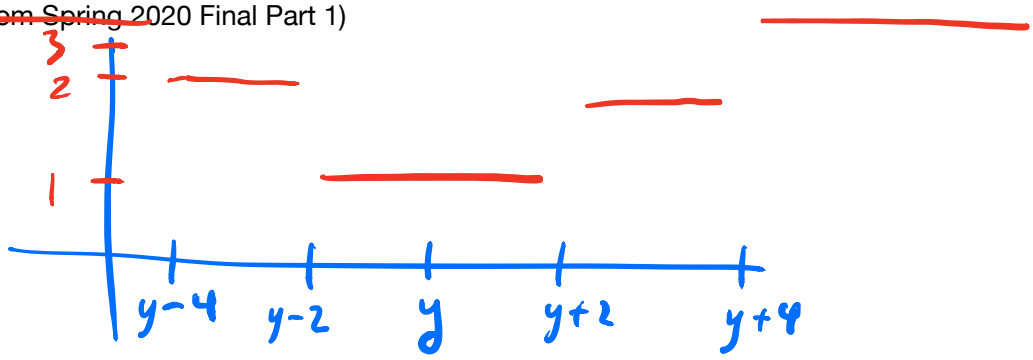
**Problem 3.** (10 points)

If we use empirical risk minimization to make predictions, which of the following loss functions would tend to generate the best predictions? Explain your choice.

$$L_A(h,y) = \begin{cases} 1, & 0 \le |h-y| \le 2 \\ 2, & 2 < |h-y| \le 4 \\ 3, & 4 < |h-y| \end{cases} \qquad L_B(h,y) = \begin{cases} 3, & 0 \le |h-y| \le 2 \\ 2, & 2 < |h-y| \le 4 \\ 1, & 4 < |h-y| \end{cases}$$

$$L_C(h,y) = \begin{cases} 2, & 0 \le |h-y| \le 2 \\ 3, & 2 < |h-y| \le 4 \\ 1, & 4 < |h-y| \end{cases} \qquad L_D(h,y) = \begin{cases} 2, & 0 \le |h-y| \le 2 \\ 1, & 2 < |h-y| \le 4 \\ 3, & 4 < |h-y| \end{cases}$$

(taken from Spring 2020 Final Part 1)



**1.** Given a data set of size $n = 8$ with $y_1 \leq y_2 \leq y_3 - 1 \leq y_3 \leq y_3 + 1 \leq y_4 \leq y_5 \leq y_6 \leq y_7 \leq y_8$ how does $R_{\text{abs}}(y_3 - 1)$ compare to $R_{\text{abs}}(y_3 + 1)$? Can you determine which is bigger, and by how much?

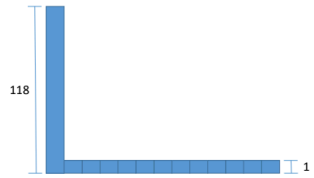
(taken from Spring 2021 Final Part 1)

Spring 20 MT Q4

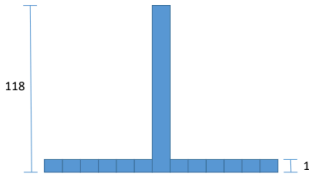
**Problem 5.**

Look at the three different data distributions shown below. Each has the same x-axis, where the markings are evenly spaced, splitting the data into thirteen bins of equal size. The frequency count for each bin is shown by the height of each bar.

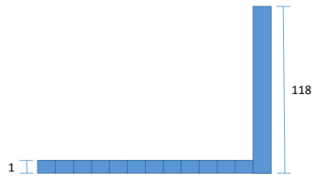
Distribution A:




Distribution B:




Distribution C:




**a)** Which of these three distributions has the smallest **mean absolute deviation from the median**? Justify your answer.

**b)** For the other two distributions, can you determine which has a larger mean absolute deviation from the median? Justify your answer.

(taken from Fall 2020 Final Part 1)

**Problem 2.**

Consider the following data set in which each point $x_i$ has an associated weight $\omega_i$:

| $i$ | $x_i$ | $\omega_i$ |
|---|---|---|
| 1 | 1 | 2 |
| 2 | 2 | 2 |
| 3 | 4 | 4 |
| 4 | 10 | 1 |

Define $R(h)$ as follows:

$$R(h) = \frac{1}{4}\sum_{i=1}^{4} \omega_i (x_i - h)^2.$$

That is, $R$ is the mean weighted square loss.

Run one iteration of gradient descent on $R$ using the data above, a learning rate of $\alpha = 1/8$, and an initial prediction of $h = 0$. Show your work.

$h_1 = h_0 - \alpha \frac{d}{dh} R(h_0)$

(taken from Winter 2020 Final)

$$R(h) = \frac{1}{4}\left[2(1-h)^2 + 2(2-h)^2 + 4(4-h)^2 + 1(10-h)^2\right]$$

$$\frac{d}{dh}R(h) = \frac{1}{4}\left[2(2)(1-h)(-1) + 2(2)(2-h)(-1) + 4(2)(4-h)(-1) + (1)(2)(10-h)(-1)\right]$$

$$\frac{d}{dh}R(h_0) = \frac{d}{dh}R(0) = \frac{1}{4}\left[2(2)(1-0)(-1) + (2)(2)(2-0)(-1) + 4(2)(4-0)(-1) + (1)(2)(10-0)(-1)\right]$$

$$= \frac{1}{4}\left[-4-8-32-20\right] = -1-2-8-5 = -16$$

$$h_1 = h_0 - \alpha \frac{d}{dh} R(h_0) = 0 - \frac{1}{8}(-16) = \boxed{2}$$

**Problem 5.** (12 points)

For each of the following statements, decide whether it is **true or false**, and justify your answer.

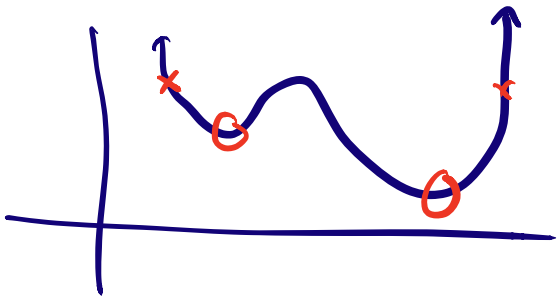
**a)** [6 points] In gradient descent, a larger learning rate sometimes allows you to find the minimum at a faster pace.

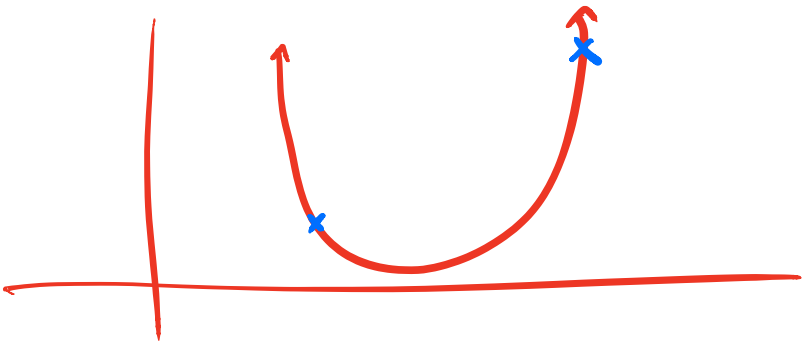
**b)** [6 points] If you start at two different initial predictions, gradient descent will find the minimum in either case, but it may take a different amount of time.

(Taken from Spring 2020 Final Part 1)

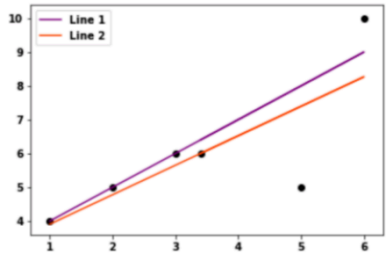
a) True!

b) False!





**4.** The picture below shows a data set and two lines. One of the lines is the least squares regression line and the other is the least absolute deviation regression line. Identify which line is which and explain your answer.




(taken from Spring 2021 Final Part 1)

**Problem 6.** (10 points)

In general, if we fit a regression line to a set of data points, some of which are duplicates, do we get the same or different regression line when we fit a regression line to the set of unique data points, with duplicates discarded? Justify your answer.

(taken from Spring 2020 Final Part 1)

# Homework 1

$P(d_{x_i})$

$L(d_{x_i})$

$$L(d, x_i) + P(d, x_i) = x_i$$

$$L_{sca}(h, y_i) = \begin{cases} a(h - y) & y \le h \\ b(y - h) & y > h \end{cases}$$

$$L'_{sca}(h, y_i) = \begin{cases} a & y \le h \\ -b & y > h \end{cases}$$

$$R_{sca}(h) = \frac{1}{n} \sum_{i=1}^{n} L_{sca}(h, y_i)$$

$$R'_{sca}(h) =$$

$$\frac{1}{n}\left[\sum_{i=1} L'_{sca}(h, y_i)\right]$$

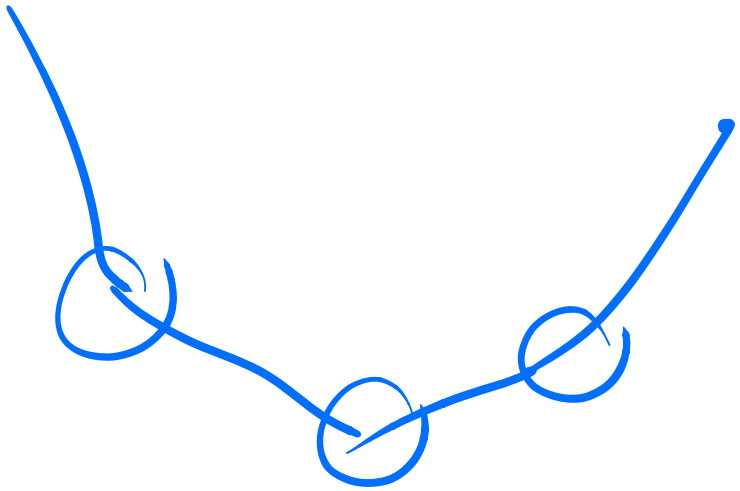
$$R'_{sca}(h) = \frac{1}{n}\left[a(\#y_i < h) - b(\#y_i > h)\right]$$

$$= 0$$

$$a(\#y_i < h) - b(\#y_i > h) = 0$$

$$\frac{a}{b} = \frac{(\#\, y_i > h)}{(\#\, y_i < h)}$$

$$\frac{1}{3} = \frac{(\#\, y_i > h)}{(\#\, y_i < h)}$$



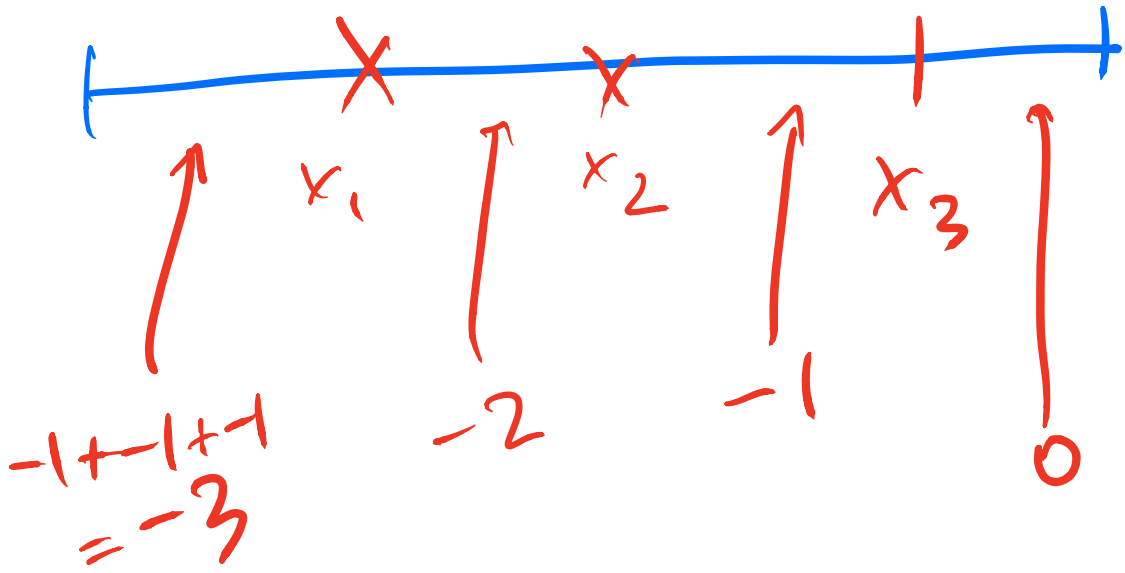
# Homework 2

slope = −1

$$L(d; x_i) = \begin{cases} x_i - d & d \le x_i \\ x_i & \text{else} \end{cases}$$

slope = 0

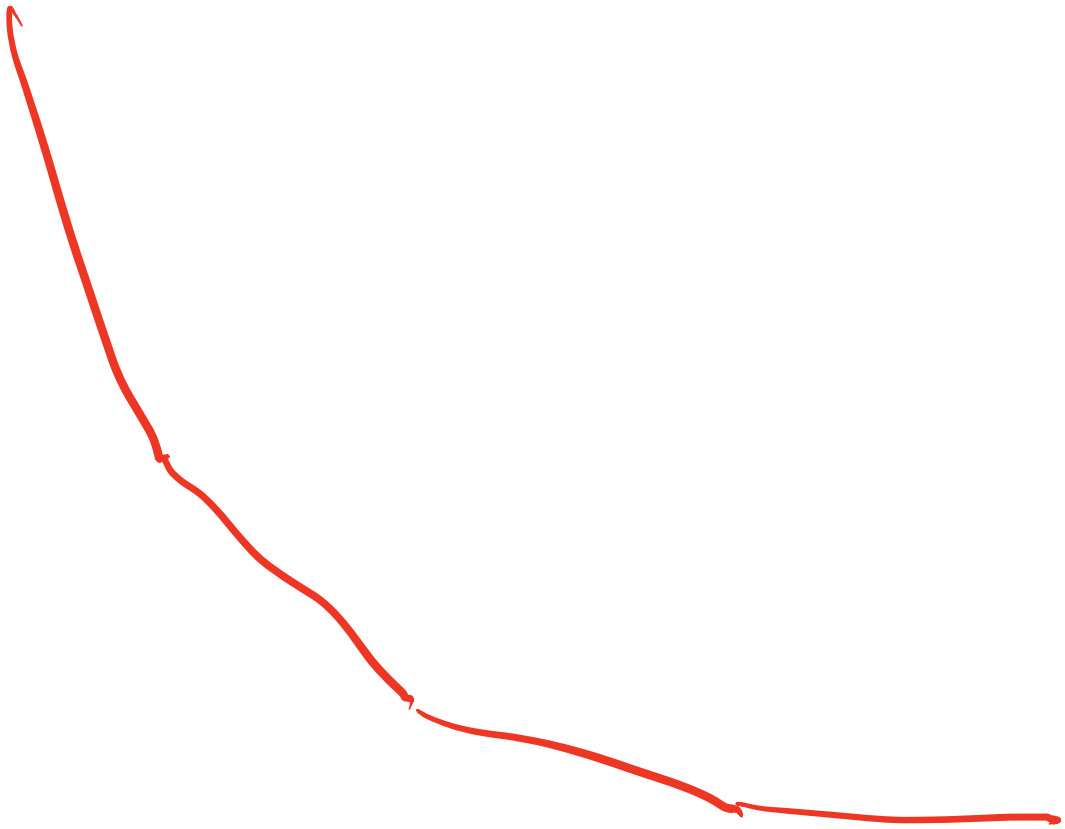
$$T(d) = \sum_{i=1}^{\hat{}} L(d; x_i)$$

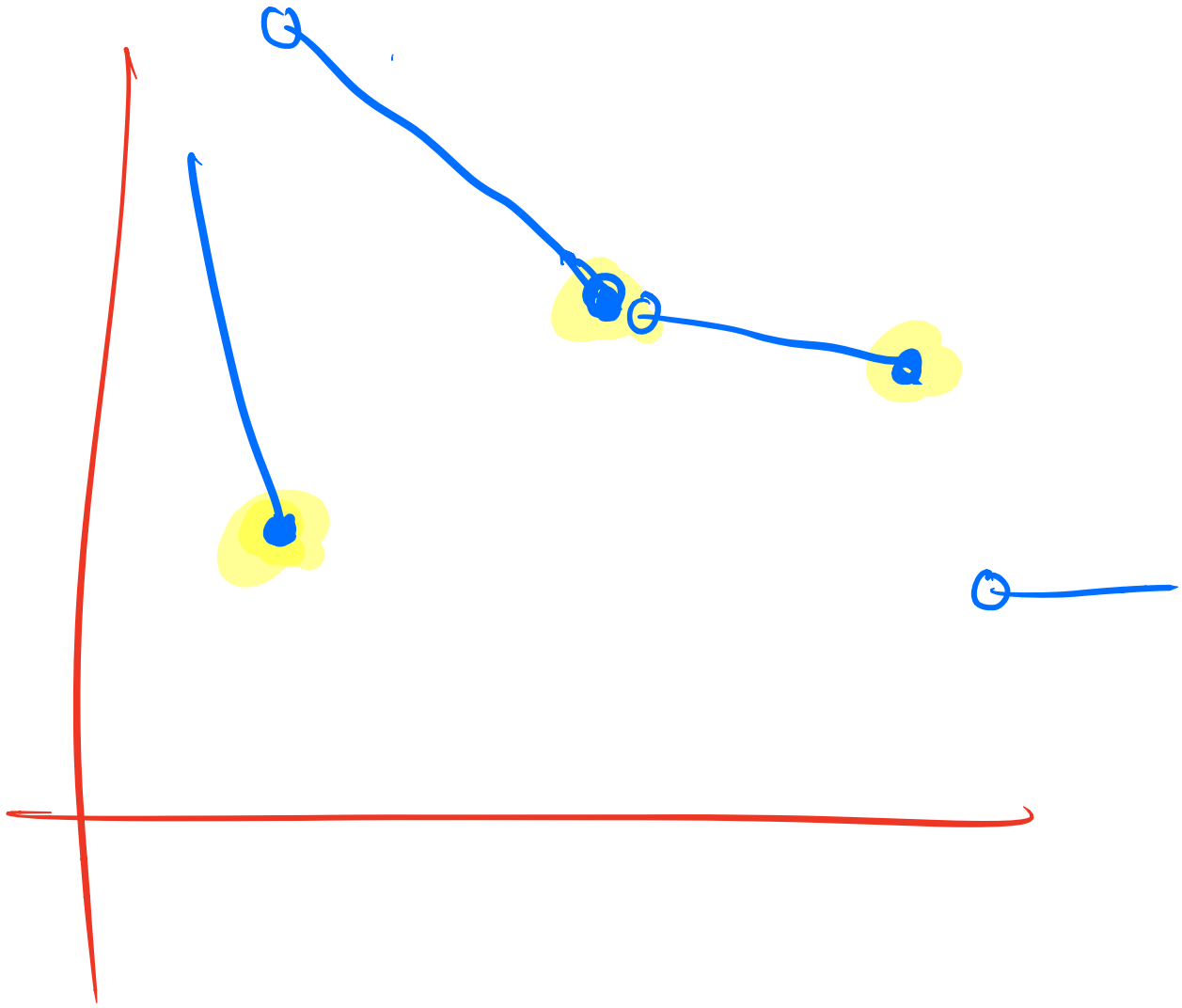


c) slope of $T(d)$

$$= -\left(\#\, x_i > d\right)$$

d)





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

$$\bar{y} - w_1 \bar{x}$$

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

$$w_1^* = \frac{\sum_{i \neq j} (x_i - \bar{x}) y_i + (0 - \bar{x}) 120}{n s^2}$$

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

$$w' = \frac{\sum_{i \neq j} (x_i - \bar{x})^2}{\quad}$$

$$\frac{+ (10-\bar{x}) \cdot 320}{n\xi^2}$$

$$W' - W_1^* = (10-\bar{x})320 - \frac{(10-\bar{x})120}{n\xi^2}$$

$$= \frac{(10-90)(320-120)}{200 \times 20^2}$$

$$= -\frac{1}{5}$$

C).

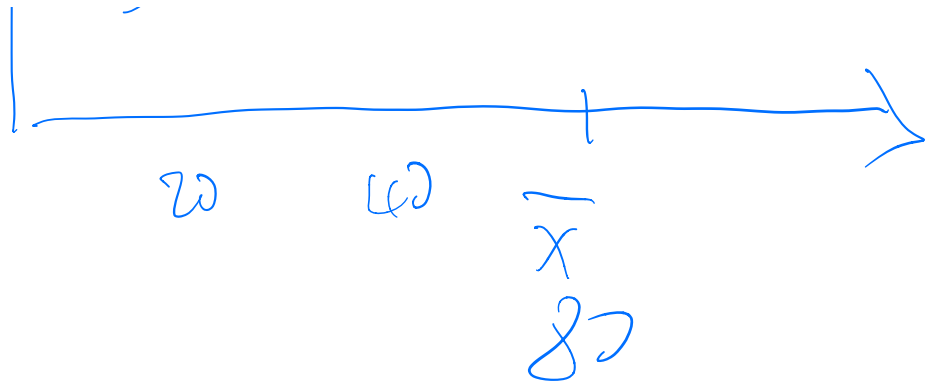
A. $H'(20) - H^*(20)$

✓

B. $H'(40) - H^*(40)$

2021 > 2020

$H'$ $H'(\bar{x}) > H^*(\bar{x})$




$$k = w_1' - w_1^* = \frac{(\bar{x} - \xi_0)\delta w}{n\delta^2} \geq 0$$

$$\xi_0 - w = 0.$$

$$R(\xi_0) = 0.$$

$$R(-\xi_0) > 0.$$