

Lecture 4 – Spread, Other Loss Functions, Gradient Descent



DSC 40A, Fall 2021 @ UC San Diego
Suraj Rampure, with help from **many others**

Announcements

- ▶ **Make sure you submit Survey 1!**
- ▶ Homework 2 will be released today, due **Monday 10/11 at 11:59pm.**
- ▶ Groupwork 2 will be released today, due **Thursday 10/7 at 11:59pm. Must** submit in groups of 2-4.
- ▶ Discussion section is on Wednesday. Remote again.
 - ▶ Later today we'll send out a signup sheet where you can specify the breakout rooms you want.
 - ▶ If you have a group you want to meet with outside of discussion, go for it.
- ▶ Videos for Lecture 3 are posted on Campuswire.

Agenda

- ▶ Recap of empirical risk minimization.
- ▶ Center and spread.
- ▶ A new loss function.
- ▶ Gradient descent.

Recap of empirical risk minimization

Empirical risk minimization

- ▶ **Goal:** Given a dataset y_1, y_2, \dots, y_n , determine the best prediction h^* . *⑤ → best*
- ▶ Strategy:
 1. Choose a **loss function**, $L(h, y)$, that measures how far any particular prediction h is from the “right answer” y .
 2. Minimize **empirical risk** (also known as average loss) over the entire dataset. The value(s) of h that minimize empirical risk are the resulting “best predictions”.

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

Absolute loss and squared loss

- ▶ General form of empirical risk:

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

"absolute error"

- ▶ **Absolute loss:** $L_{\text{abs}}(h, y) = |y - h|$.

▶ Empirical risk: $R_{\text{abs}}(h) = \frac{1}{n} \sum_{i=1}^n |y_i - h|$. Also called "mean absolute error".

▶ Minimized by $h^* = \text{Median}(y_1, y_2, \dots, y_n)$.

- ▶ **Squared loss:** $L_{\text{sq}}(h, y) = (y - h)^2$.

"squared error"

▶ Empirical risk: $R_{\text{sq}}(h) = \frac{1}{n} \sum_{i=1}^n (y_i - h)^2$. Also called "mean squared error".

▶ Minimized by $h^* = \text{Mean}(y_1, y_2, \dots, y_n)$.

Discussion Question

Consider a dataset y_1, y_2, \dots, y_n .
Recall,

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

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

$$y_1 = 1, y_2 = 5 \\ h = 4$$

$$R_{abs}(4) = \frac{1}{2} \cdot [3 + 1] = \frac{1}{2} \cdot 4 = 2$$

Is it true that, for any h , $[R_{abs}(h)]^2 = R_{sq}(h)$?

- a) True
- b) False

$$R_{sq}(4) = \frac{1}{2} [9 + 1] = 5$$

To answer, go to menti.com and enter the code 1250 9212.

~~$$2^2 = 5$$~~

Center and spread

What does it mean?

- ▶ General form of empirical risk:

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

- ▶ The input h^* that minimizes $R(h)$ is some measure of the **center** of the data set.
 - ▶ e.g. median, mean, mode.
- ▶ The minimum output $R(h^*)$ represents some measure of the **spread**, or variation, in the data set.

Absolute loss

- The empirical risk for the absolute loss is

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

- $R_{abs}(h)$ is minimized at $h^* = \text{Median}(y_1, y_2, \dots, y_n)$.
- Therefore, the minimum value of $R_{abs}(h)$ is

$$R_{abs}(h^*) = R_{abs}(\text{Median}(y_1, y_2, \dots, y_n))$$

$$= \frac{1}{n} \sum_{i=1}^n |y_i - \text{Median}(y_1, y_2, \dots, y_n)|.$$

Mean absolute deviation from the median

- The minimum value of $R_{abs}(h)$ is the **mean absolute deviation from the median.**

$$\frac{1}{n} \sum_{i=1}^n |y_i - \text{Median}(y_1, y_2, \dots, y_n)|$$

- It measures how far each data point is from the median, on average.

Median : 3

Discussion Question

$$1001 \rightarrow \frac{1}{4}[1+0+0+1] = \frac{1}{2}$$

For the data set 2, 3, 3, 4, what is the mean absolute deviation from the median?

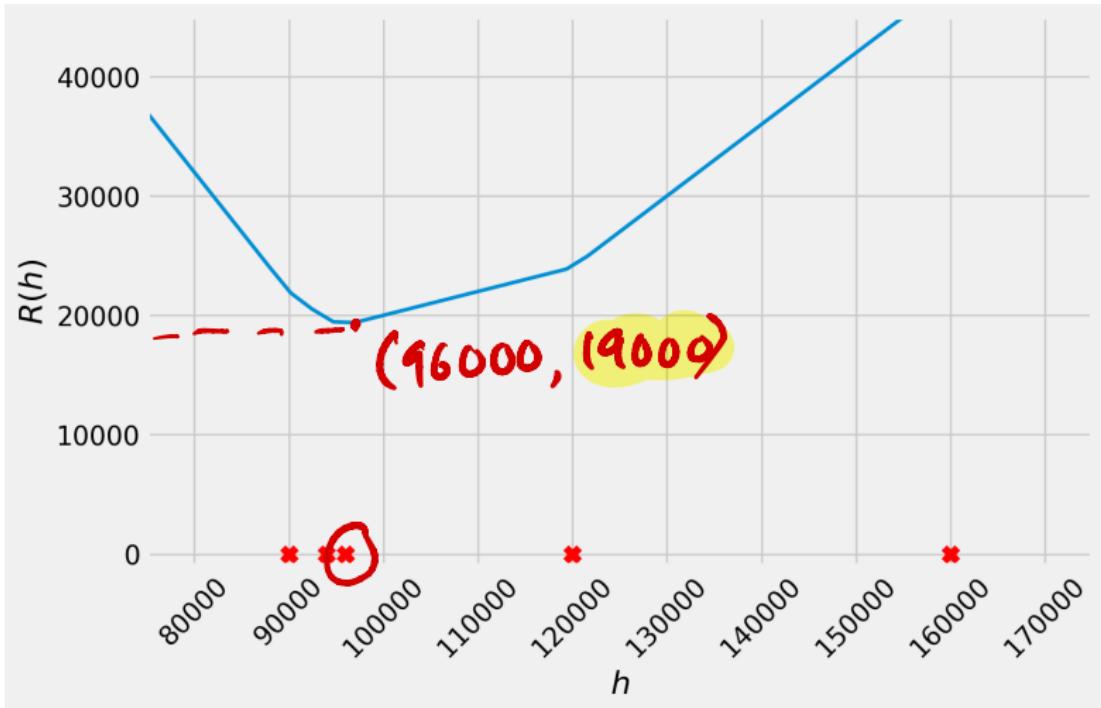
a) 0

b) $\frac{1}{2}$

c) 1

d) 2

Mean absolute deviation from the median



Squared loss

- The empirical risk for the squared loss is

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

- $R_{\text{sq}}(h)$ is minimized at $h^* = \text{Mean}(y_1, y_2, \dots, y_n)$.
- Therefore, the minimum value of $R_{\text{sq}}(h)$ is

$$R_{\text{sq}}(h^*) = R_{\text{sq}}(\text{Mean}(y_1, y_2, \dots, y_n))$$

$$= \frac{1}{n} \sum_{i=1}^n (y_i - \text{Mean}(y_1, y_2, \dots, y_n))^2.$$

Variance

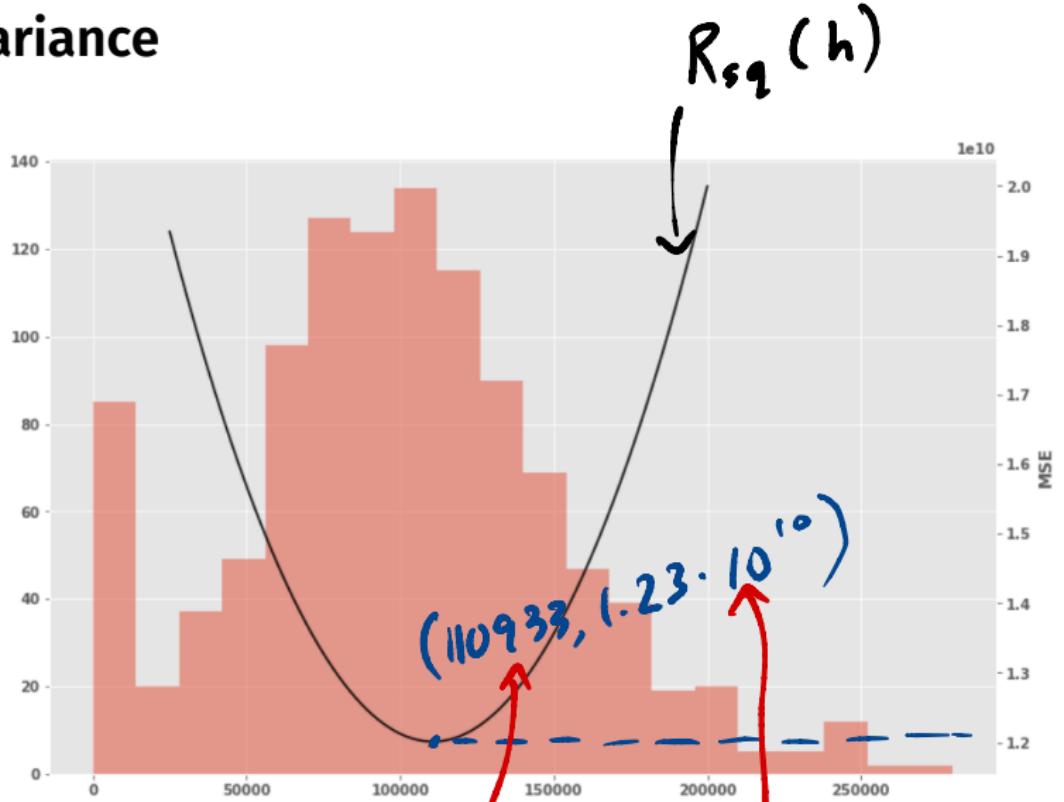
- ▶ The minimum value of $R_{\text{sq}}(h)$ is the mean squared deviation from the mean, more commonly known as the **variance**.

$$\frac{1}{n} \sum_{i=1}^n (y_i - \text{Mean}(y_1, y_2, \dots, y_n))^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2$$

σ_y^2

- ▶ It measures the squared distance of each data point from the mean, on average.
- ▶ Its square root is called the **standard deviation**.

Variance



mean of
dataset

variance of
dataset

0-1 loss

$$R_{0,1}(4) = \text{fraction of pts not equal to 4}$$

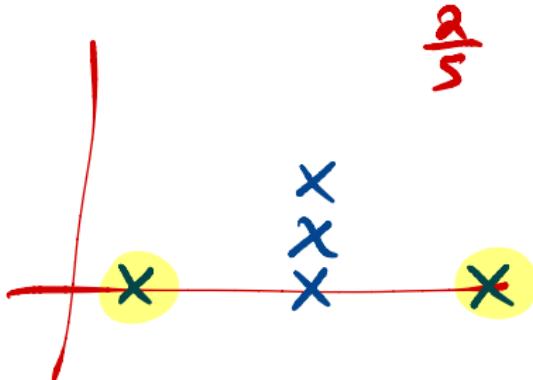
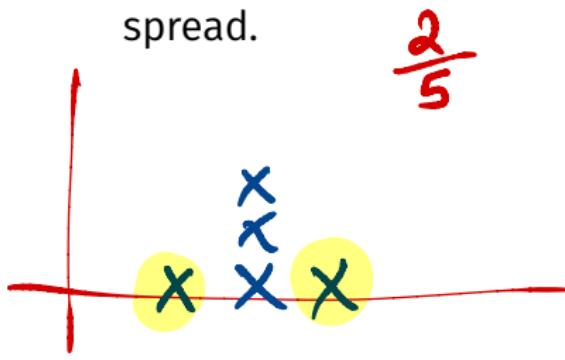
- The empirical risk for the 0-1 loss is

$$R_{0,1}(h) = \frac{1}{n} \sum_{i=1}^n \begin{cases} 0, & \text{if } h = y_i \\ 1, & \text{if } h \neq y_i \end{cases}$$

- This is the proportion (between 0 and 1) of data points not equal to h .
-fraction
- $R_{0,1}(h)$ is minimized at $h^* = \text{Mode}(y_1, y_2, \dots, y_n)$.
- Therefore, $R_{0,1}(h^*)$ is the proportion of data points not equal to the mode.

A poor way to measure spread

- ▶ The minimum value of $R_{0,1}(h)$ is the proportion of data points not equal to the mode.
- ▶ A higher value means less of the data is clustered at the mode.
- ▶ Just as the mode is a very simplistic way to measure the center of the data, this is a very crude way to measure spread.



Summary of center and spread

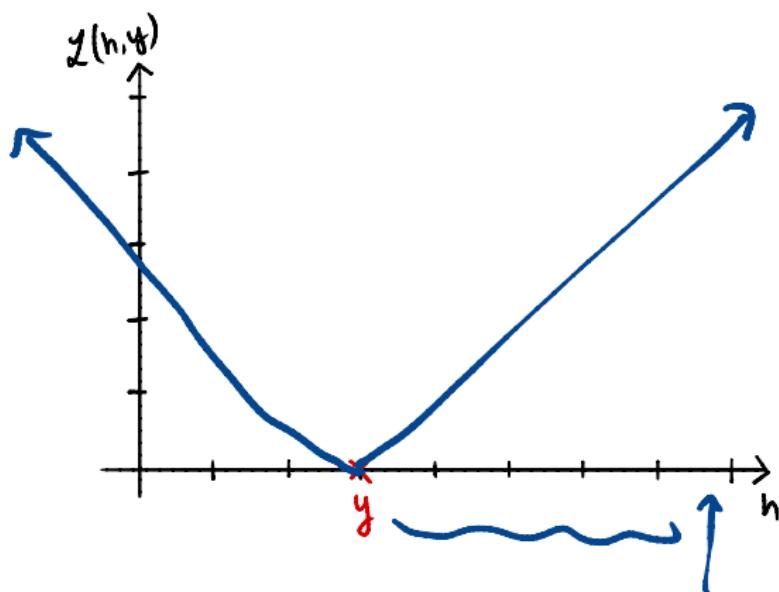
- ▶ Different loss functions lead to empirical risk functions that are minimized at various measures of **center**.
- ▶ The minimum values of these risk functions are various measures of **spread**.
- ▶ There are many different ways to measure both center and spread. These are sometimes called **descriptive statistics**.

summary statistics

A new loss function

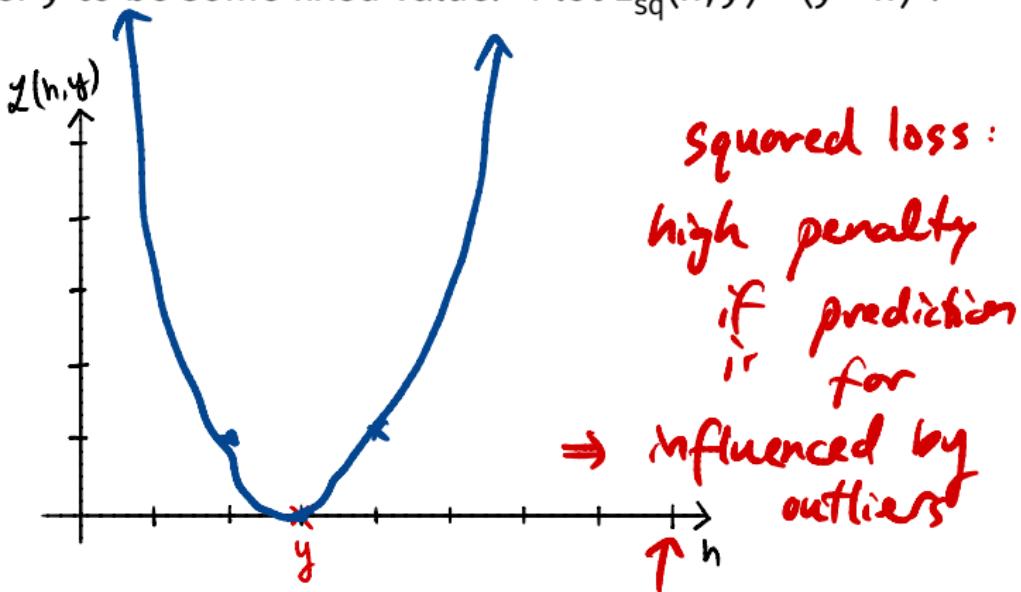
Plotting a loss function

- ▶ The plot of a loss function tells us how it treats outliers.
- ▶ Consider y to be some fixed value. Plot $L_{\text{abs}}(h, y) = |y - h|$:



Plotting a loss function

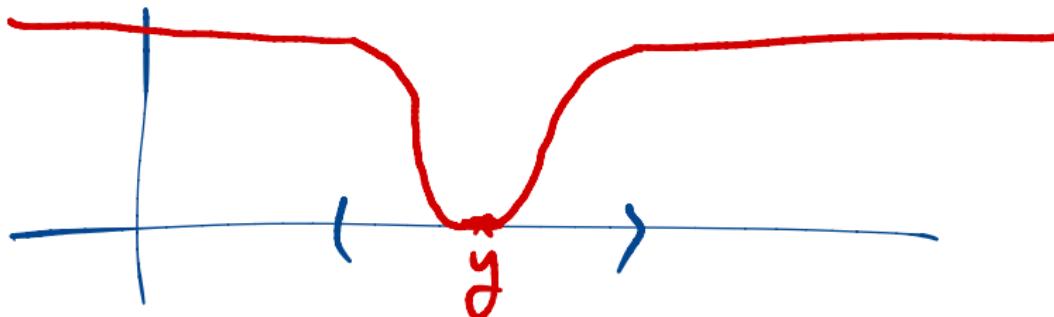
- The plot of a loss function tells us how it treats outliers.
- Consider y to be some fixed value. Plot $L_{\text{sq}}(h, y) = (y - h)^2$:



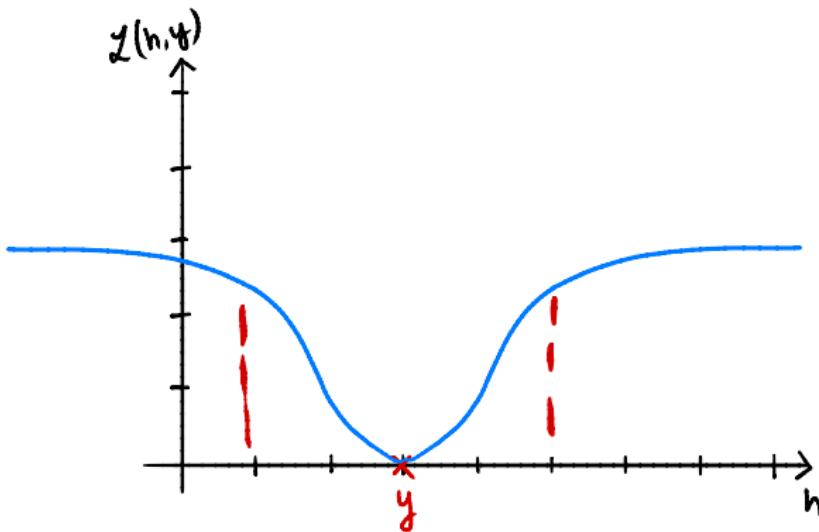
Discussion Question

Suppose L considers all outliers to be equally as bad.
What would it look like far away from y ?

- a) flat
- b) rapidly decreasing
- c) rapidly increasing



A very insensitive loss



- We'll call this loss L_{ucsd} because it doesn't have a name.

Discussion Question

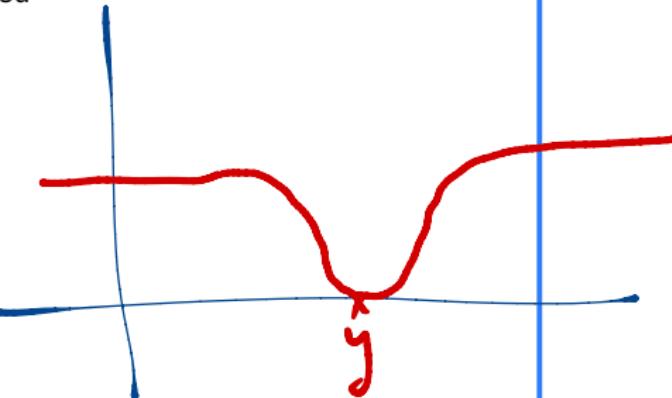
Which of these could be $L_{ucsd}(h, y)$?

a) ~~$e^{(y-h)^2}$~~

b) $1 - e^{-(y-h)^2}$

c) ~~$1 - (y - h)^2$~~

d) $1 - e^{-|y-h|}$



To answer, go to menti.com and enter the code 1250 9212.

Adding a scale parameter

- ▶ Problem: L_{ucsd} has a fixed scale. This won't work for all datasets.
 - ▶ If we're predicting temperature, and we're off by 100 degrees, that's bad.
 - ▶ If we're predicting salaries, and we're off by 100 dollars, that's pretty good.
- ▶ What we consider to be an outlier depends on the scale of the data.
- ▶ Fix: add a **scale parameter**, σ :

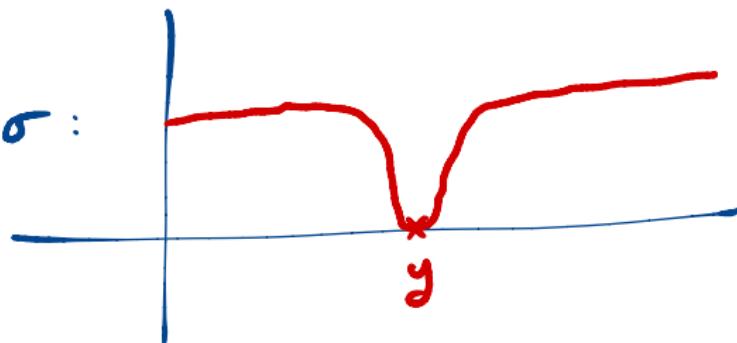
You get to choose! $L_{ucsd}(h, y) = 1 - e^{-(y-h)^2/\sigma^2}$
NOT std deviation!!!

similar to bell curve

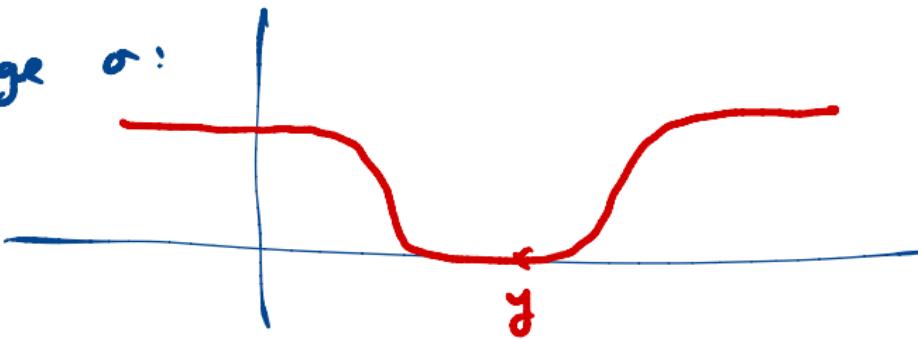
Adding a scale parameter

$$L_{csd} = 1 - e^{-\frac{(y-h)^2}{\sigma^2}}$$

small σ :

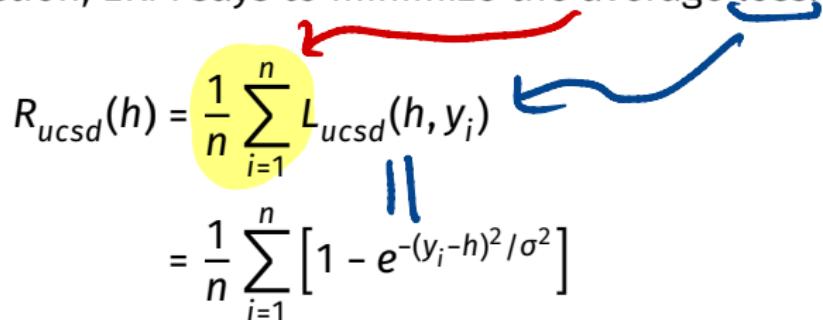


large σ :



Empirical risk minimization

- ▶ We have salaries y_1, y_2, \dots, y_n .
- ▶ To find prediction, ERM says to minimize the average loss;

$$\begin{aligned} R_{ucsd}(h) &= \frac{1}{n} \sum_{i=1}^n L_{ucsd}(h, y_i) \\ &= \frac{1}{n} \sum_{i=1}^n \left[1 - e^{-(y_i-h)^2/\sigma^2} \right] \end{aligned}$$


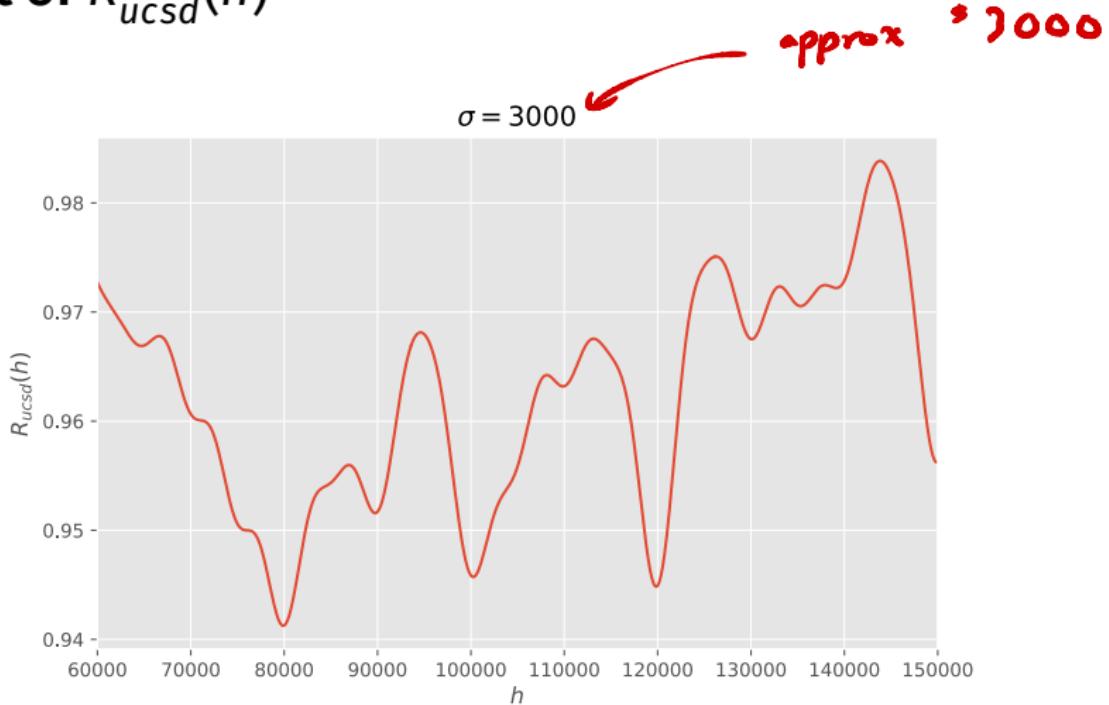
Let's plot R_{ucsd}

- ▶ Recall:

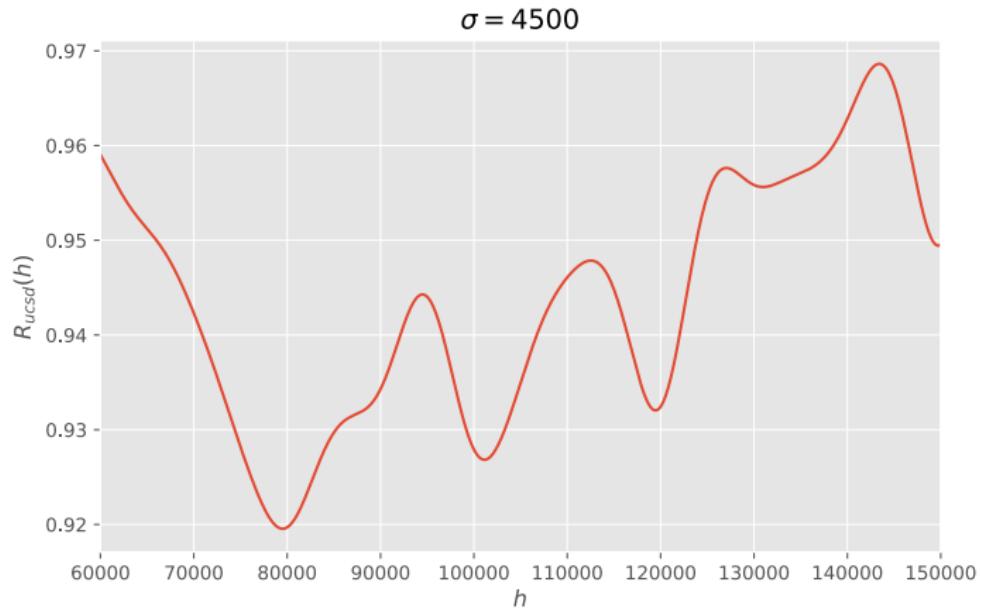
$$R_{ucsd}(h) = \frac{1}{n} \sum_{i=1}^n \left[1 - e^{-(y_i - h)^2 / \sigma^2} \right]$$

- ▶ Once we have data y_1, y_2, \dots, y_n and a scale σ , we can plot $R_{ucsd}(h)$.
- ▶ We'll use full the StackOverflow dataset ($n = 1121$).
- ▶ Let's try several scales, σ .

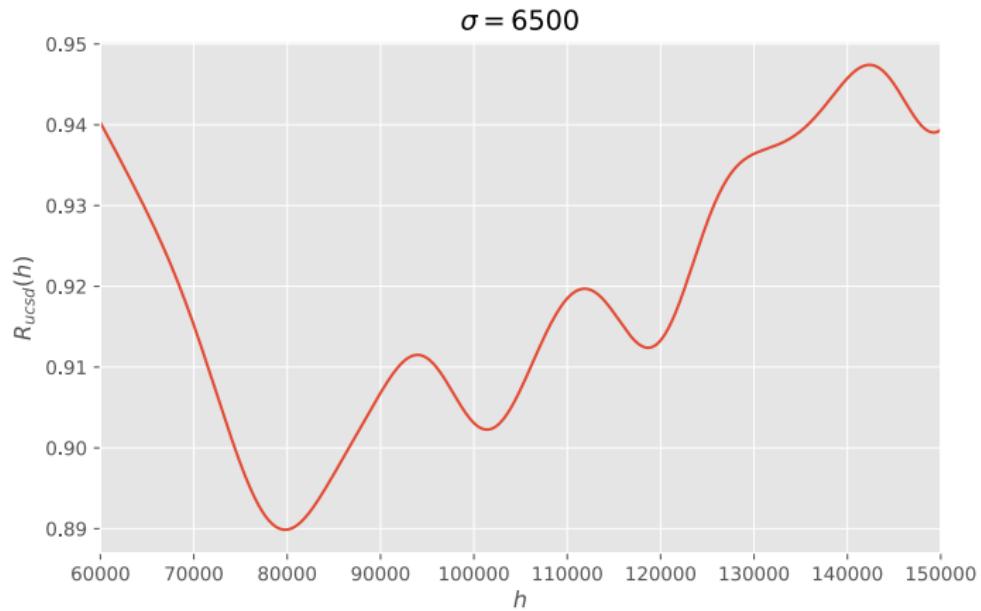
Plot of $R_{ucsd}(h)$



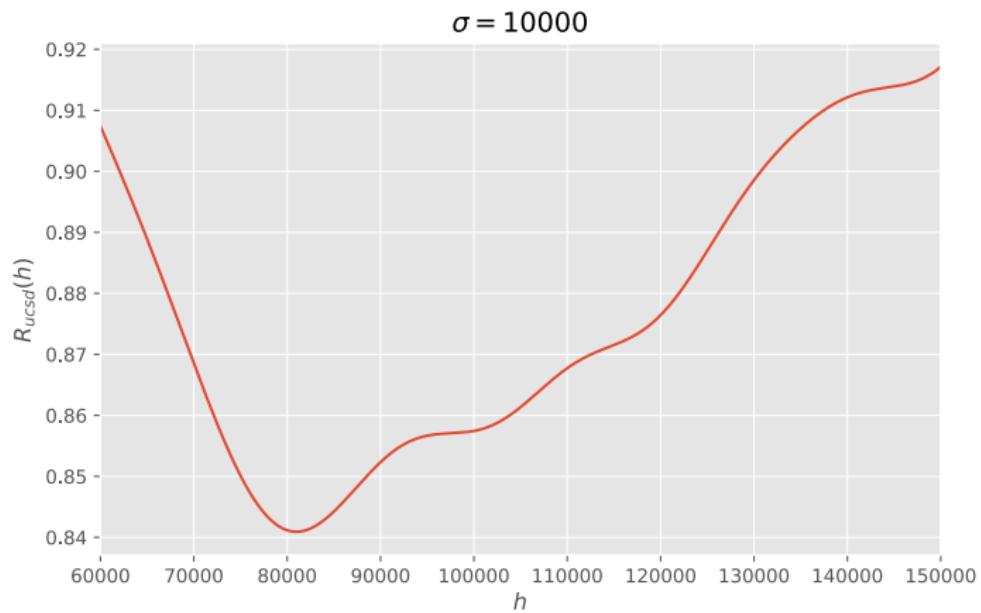
Plot of $R_{ucsd}(h)$



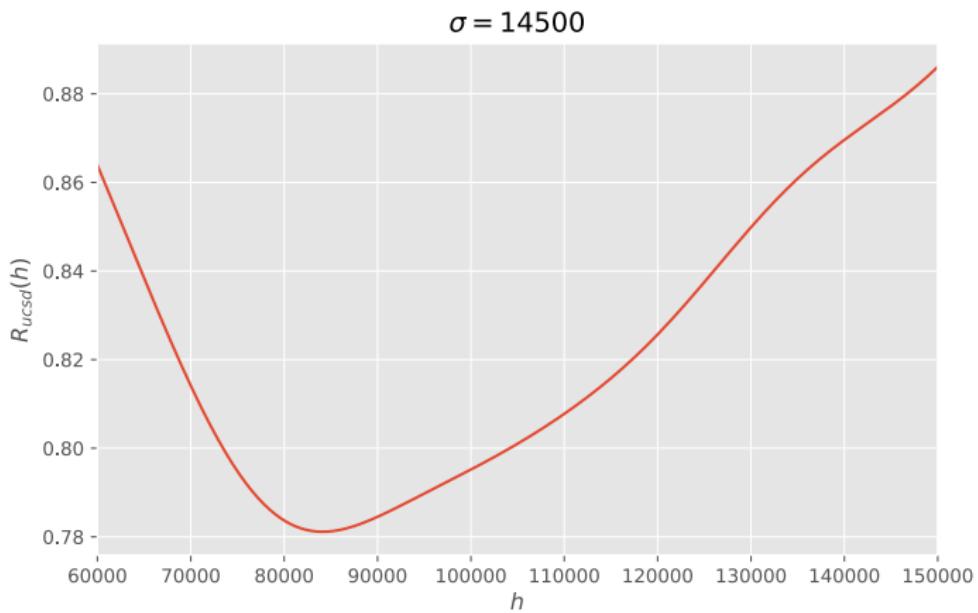
Plot of $R_{ucsd}(h)$



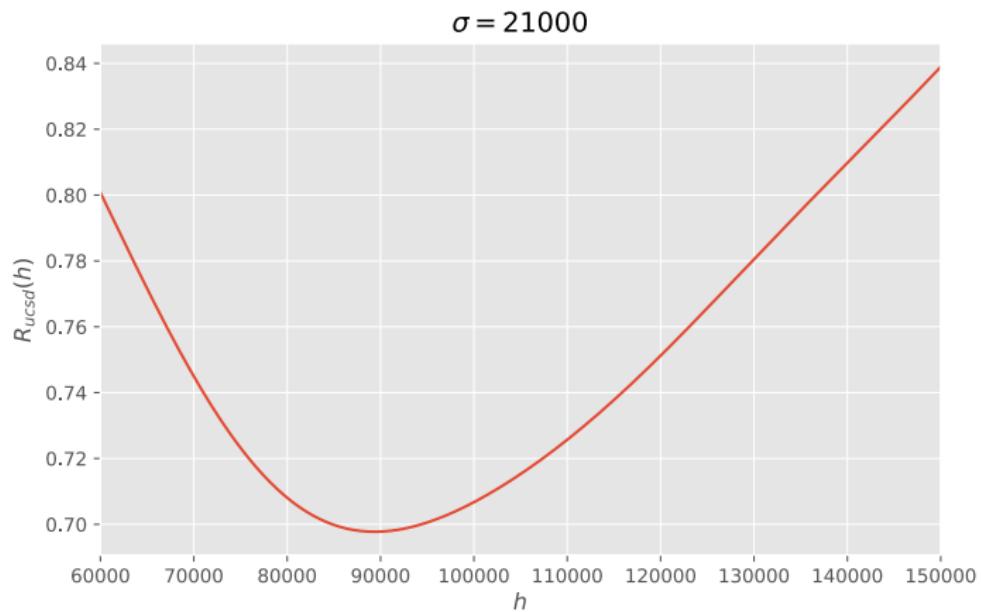
Plot of $R_{ucsd}(h)$



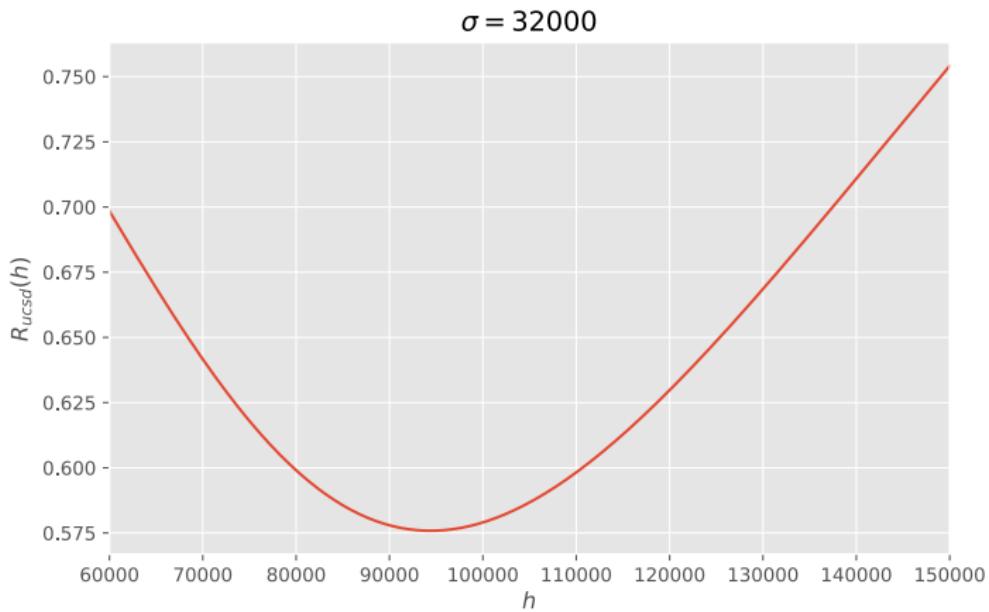
Plot of $R_{ucsd}(h)$



Plot of $R_{ucsd}(h)$



Plot of $R_{ucsd}(h)$



Minimizing R_{ucsd}

- ▶ To find the best prediction, we find h^* minimizing $R_{ucsd}(h)$.
- ▶ $R_{ucsd}(h)$ is **differentiable**.
- ▶ To minimize: take derivative, set to zero, solve.

$$R_{ucsd}(h) = \frac{1}{n} \sum_{i=1}^n \left[1 - e^{-\frac{(y_i - h)^2}{\sigma^2}} \right]$$

Step 1: Taking the derivative

$$\begin{aligned}\frac{dR_{ucsd}}{dh} &= \frac{d}{dh} \left(\frac{1}{n} \sum_{i=1}^n \left[1 - e^{-(y_i-h)^2/\sigma^2} \right] \right) \\&= \frac{1}{n} \sum_{i=1}^n \left[\frac{d}{dh} \left[1 - e^{-\frac{(y_i-h)^2}{\sigma^2}} \right] \right] \\&= \frac{1}{n} \sum_{i=1}^n \left[-e^{-\frac{(y_i-h)^2}{\sigma^2}} \cdot \frac{d}{dh} \left(-\frac{(y_i-h)^2}{\sigma^2} \right) \right] \\&= \frac{1}{n} \sum_{i=1}^n \left[-e^{-\frac{(y_i-h)^2}{\sigma^2}} \cdot \underline{\underline{(-1) \cdot \frac{2(y_i-h)}{\sigma^2} (-1)}} \right] \\&= \underline{\underline{\frac{2}{n\sigma^2} \sum_{i=1}^n \left[e^{-\frac{(y_i-h)^2}{\sigma^2}} \cdot (h-y_i) \right]}}$$

Step 2: Setting to zero and solving

- We found:

$$\frac{d}{dh}(h) = \frac{2}{n\sigma^2} \sum_{i=1}^n (h - y_i) \cdot e^{-(h-y_i)^2/\sigma^2}$$

- Now we just set to zero and solve for h :

$$0 = \frac{2}{n\sigma^2} \sum_{i=1}^n (h - y_i) \cdot e^{-(h-y_i)^2/\sigma^2}$$

- We **can** calculate derivative, but we **can't** solve for h ; we're stuck again.
- Now what???

Gradient descent

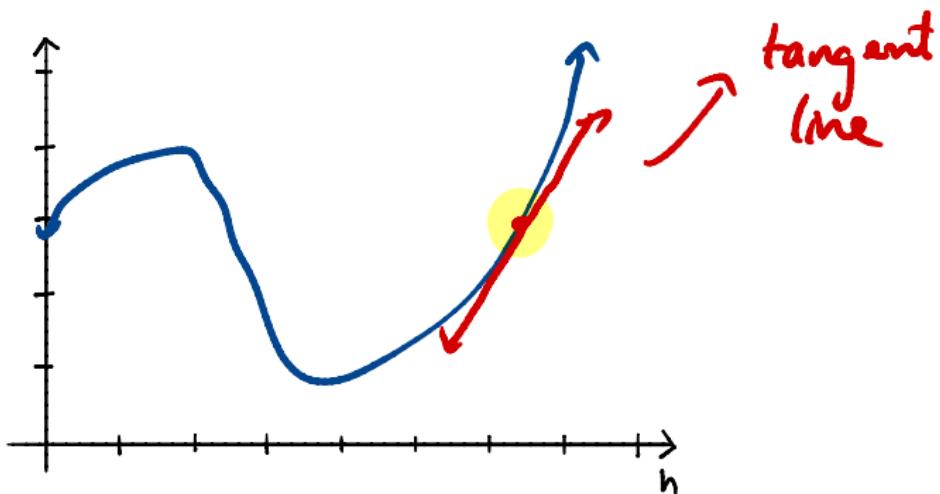
The general problem

- ▶ **Given:** a differentiable function $R(h)$.
- ▶ **Goal:** find the input h^* that minimizes $R(h)$.

Gradient descent works for any
differentiable function, not just
empirical risk!

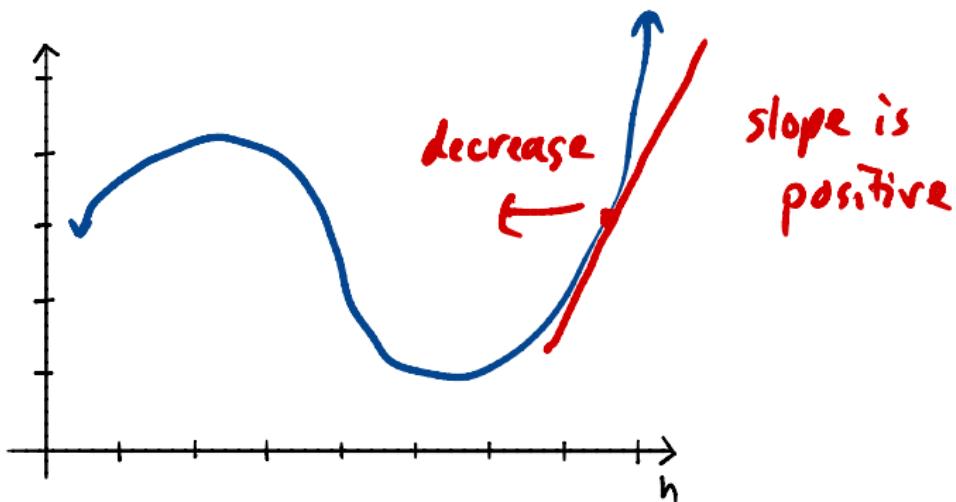
Meaning of the derivative

- ▶ We're trying to minimize a **differentiable** function $R(h)$. Is calculating the derivative helpful?
- ▶ $\frac{dR}{dh}(h)$ is a function; it gives the **slope** at h .



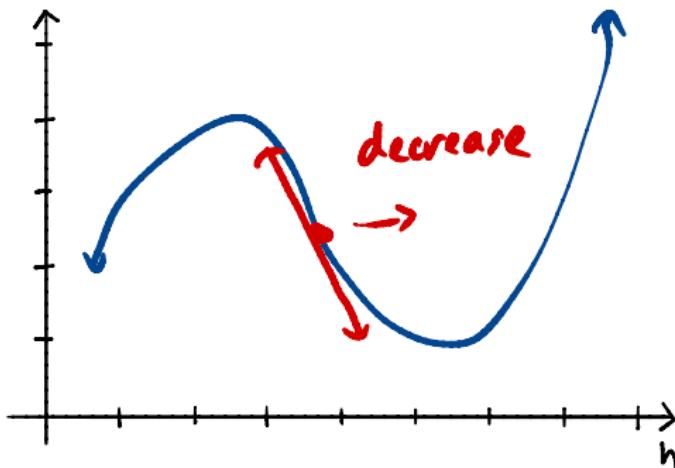
Key idea behind gradient descent

- ▶ If the slope of R at h is **positive** then moving to the **left** decreases the value of R .
- ▶ i.e., we should **decrease** h .



Key idea behind gradient descent

- ▶ If the slope of R at h is **negative** then moving to the **right** decreases the value of R .
- ▶ i.e., we should **increase** h .



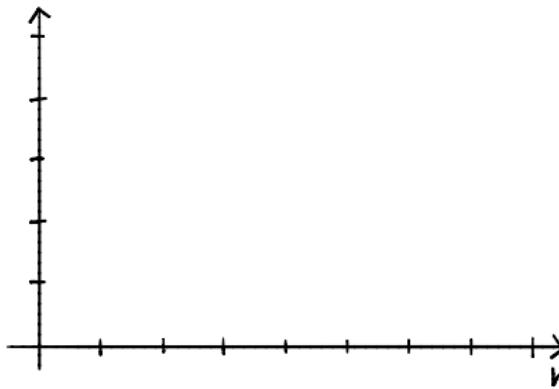
Key idea behind gradient descent

- ▶ Pick a starting place, h_0 . Where do we go next?
- ▶ Slope at h_0 negative? Then increase h_0 .
- ▶ Slope at h_0 positive? Then decrease h_0 .
- ▶ This will work:

$$h_1 = h_0 - \frac{dR}{dh}(h_0)$$

Gradient Descent

- ▶ Pick α to be a positive number. It is the **learning rate**, also known as the **step size**.
- ▶ Pick a starting prediction, h_0 .
- ▶ On step i , perform update $h_i = h_{i-1} - \alpha \cdot \frac{dR}{dh}(h_{i-1})$
- ▶ Repeat until convergence (when h doesn't change much).



You will not be responsible for implementing gradient descent in this class, but here's an implementation in Python if you're curious:

```
def gradient_descent(derivative, h, alpha, tol=1e-12):
    """Minimize using gradient descent."""
    while True:
        h_next = h - alpha * derivative(h)
        if abs(h_next - h) < tol:
            break
        h = h_next
    return h
```

Example: Minimizing mean squared error

- Recall the mean squared error and its derivative:

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

Discussion Question

Let $y_1 = -4$, $y_2 = -2$, $y_3 = 2$, $y_4 = 4$. Pick $h_0 = 4$ and $\alpha = 1/4$. What is h_1 ?

- a) -1
- b) 0
- c) 1
- d) 2

To answer, go to menti.com and enter the code 1250 9212.

Solution

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

Data values are $-4, -2, 2, 4$. Pick $h_0 = 4$ and $\alpha = 1/4$. Find h_1 .

Summary

Summary

- ▶ Different loss functions lead to empirical risk functions that are minimized at various measures of **center**.
- ▶ The minimum values of these empirical risk functions are various measures of **spread**.
- ▶ We came up with a more complicated loss function, L_{ucsd} , that treats all outliers equally.
 - ▶ We weren't able to minimize its empirical risk R_{ucsd} by hand.
- ▶ We invented **gradient descent**, which repeatedly updates our prediction by moving in the opposite direction of the derivative.
- ▶ **Next Time:** We'll look at gradient descent in action.