Lecture 4 – ERM, Center and Spread

Winter 2024

DSC 40A, ~~Spring 2023~~

Last time: the mean minimizes mean squared error

- Our problem was: find h^* which minimizes the mean squared error, $R_{sq}(h) = \frac{1}{n}\sum_{i=1}^{n}(y_i - h)^2$.
  - The answer is: $\text{Mean}(y_1, \dots, y_n)$.
  - The **best prediction**, in terms of mean squared error, is the **mean**.
  - This answer is always unique!

Last time: the mean minimizes mean squared error

- Our problem was: find h^* which minimizes the mean squared error, $R_{sq}(h) = \frac{1}{n}\sum_{i=1}^{n}(y_i - h)^2$.
  - The answer is: $\cancel{\mathrm{Mean}(y_1, \ldots, y_n)}$.
    Median
  - The **best prediction**, in terms of mean squared error, is the ~~mean~~.
    median
  - This answer is always unique!
    not

Comparing the median and mean

Outliers

- Consider our original dataset of 5 salaries.

  90,000   94,000   96,000   120,000   160,000

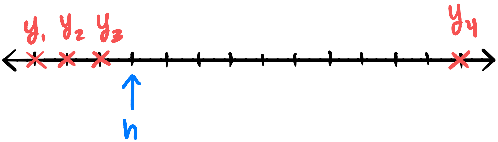
- As it stands, the **median is 96,000** and the **mean is 112,000**.
- What if we add 300,000 to the largest salary?

  90,000   94,000   96,000   120,000   460,000

- Now, the **median is still 96,000** but the **mean is 172,000**!
- **Key Idea:** The mean is quite **sensitive** to outliers.

Outliers

- The mean is quite **sensitive** to outliers.



- $|y_4 - h|$ is 10 times as big as $|y_3 - h|$.

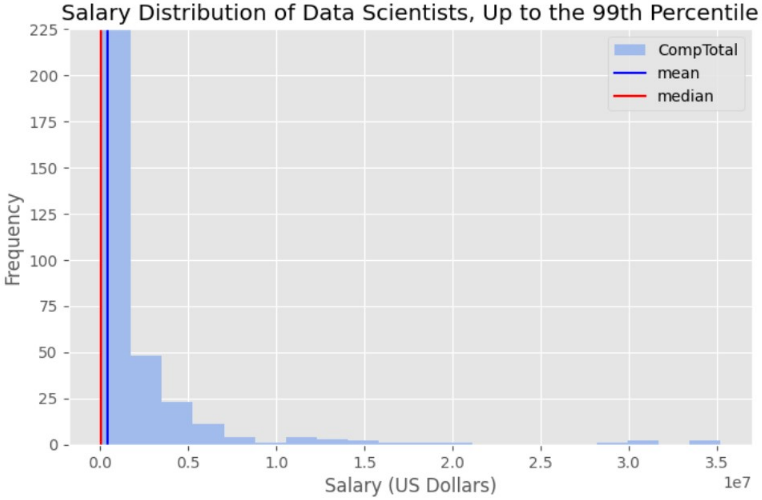
- But $(y_4 - h)^2$ is 100 times as big as $(y_3 - h)^2$.
  - This "pulls" h^* towards y_4.

- Squared error can be dominated by outliers.

Example: Data Scientist Salaries

- Dataset of 2,016 self-reported data science salaries in the United States from the 2022 StackOverflow survey.
- Median = $86,700.
- Mean = $501,425,531.
- Min = $20.
- Max = $1,000,000,000,000.
- 90th Percentile: $700,000.

Example: Data Scientist Salaries



Example: Income Inequality

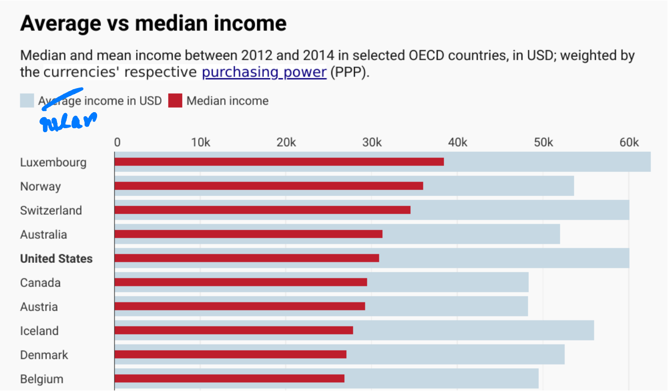
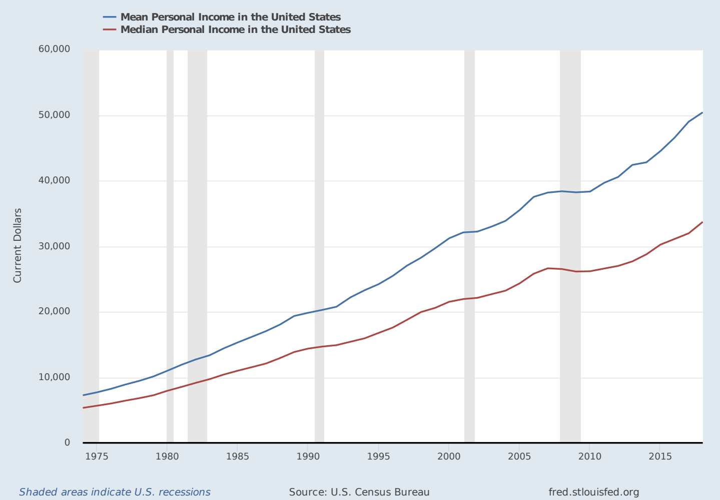


Chart: Lisa Charlotte Rost, Datawrapper

Example: Income Inequality



Empirical risk minimization

A general framework

- We started with the **mean absolute error**:

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

- Then we introduced the **mean squared error**:

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

- They have the same form: both are averages of some measurement that represents how different h is from the data.

A general framework

- Definition: A **loss function** $L(h, y)$ takes in a prediction h and a true value (i.e. a "right answer"), y, and outputs a number measuring how far h is from y (bigger = further).

- The **absolute loss**:

$$L_{\text{abs}}(h, y) = |y - h|$$

- The **squared loss**:

$$L_{\text{sq}}(h, y) = (y - h)^2$$

A general framework

- Suppose that $y_1, \ldots, y_n$ are some data points, h is a prediction, and L is a loss function. The **empirical risk** is the average loss on the data set:

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

- The goal of learning: find h that minimizes R_L. This is called **empirical risk minimization (ERM)**.

Empirical risk minimization (ERM)

- **Goal**: Given a dataset $y_1, y_2, ..., y_n$, determine the best prediction h^*.

- Strategy:
  1. Choose a **loss function**, $L(h, y)$, that measures how far any particular prediction h is from the "right answer" y.

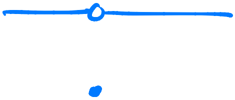
  2. Minimize **empirical risk** (also known as average loss) over the entire dataset. The value(s) of h that minimize empirical risk are the resulting "best predictions".

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

Key Idea

- The choice of loss function determines the properties of the result.
- **Different loss function = different minimizer = different prediction!**
  - Absolute loss yields the median.
  - Squared loss yields the mean.
  - The mean is easier to calculate but is more sensitive to outliers.
- ERM is a “recipe” that can be used to derive many machine learning algorithms.

Example: 0-1 Loss



1. Pick as our loss function the **0-1 loss**:

$$L_{0,1}(h, y) = \begin{cases} 0, & \text{if } h = y \\ 1, & \text{if } h \neq y \end{cases}$$

2. Minimize empirical risk:

$$R_{0,1}(h) = \frac{1}{n} \sum_{i=1}^{n} L_{0,1}(h, y_i)$$

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

Example: 0-1 Loss

1. Pick as our loss function the **0-1 loss**:

$$L_{0,1}(h, y) = \begin{cases} 0, & \text{if } h = y \\ 1, & \text{if } h \neq y \end{cases}$$

$\mathcal{L}(y_1, y_2) = 1$

2. Minimize empirical risk:

$$R_{0,1}(h) = \frac{1}{n} \sum_{i=1}^{n} L_{0,1}(h, y_i)$$

$= \frac{1}{n}\left(\overset{0}{\mathcal{L}(y_1, y_1)} + \overset{1}{\mathcal{L}(y_1, y_2)} + \dots + \overset{1}{\mathcal{L}(y_1, y_n)} \right)$

Discussion Question

Suppose $y_1, \ldots, y_n$ are all distinct. Find $R_{0,1}(y_1)$. $h = y_1$

a) 0 b) $\frac{1}{n}$ c) $\frac{n-1}{n}$ d) 1

Minimizing empirical risk

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

$$= \frac{1}{n}\left(\#\text{ of } y_i\text{'s that are} \neq h\right)$$

Minimizer: pick h to be the "most common" data point,

i.e, the <u>mode</u>

Different loss functions lead to different predictions

Loss	Minimizer	Outliers	Differentiable
L_{abs}	median	**insensitive**	**no**
L_{sq}	mean	**sensitive**	**yes**
$L_{0,1}$	mode	**insensitive**	**no**

- The optimal predictions are all **summary statistics** that measure the **center** of the data set in different ways.

Summary

Summary

- $h^* = \text{Mean}(y_1, \dots, y_n)$ minimizes $R_{sq}(h) = \frac{1}{n} \sum_{i=1}^{n} (y_i - h)^2$, i.e. the mean minimizes mean squared error.

- The mean absolute error and the mean squared error fit into a general framework called **empirical risk minimization**.
  - Pick a loss function. We've seen absolute loss, $|y - h|^2$, squared loss, $(y - h)^2$, and 0-1 loss.
  - Pick a way to minimize the average loss (i.e. empirical risk) on the data.

- By changing the loss function, we change which prediction is considered the best.

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

$$\begin{aligned} R_{abs}(h^*) &= R_{abs}(\text{Median}(y_1, y_2, \dots, y_n)) \\ &= \frac{1}{n}\sum_{i=1}^{n} |y_i - \text{Median}(y_1, y_2, \dots, y_n)|. \end{aligned}$$

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.

3, 3, 3, 3

Discussion Question

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

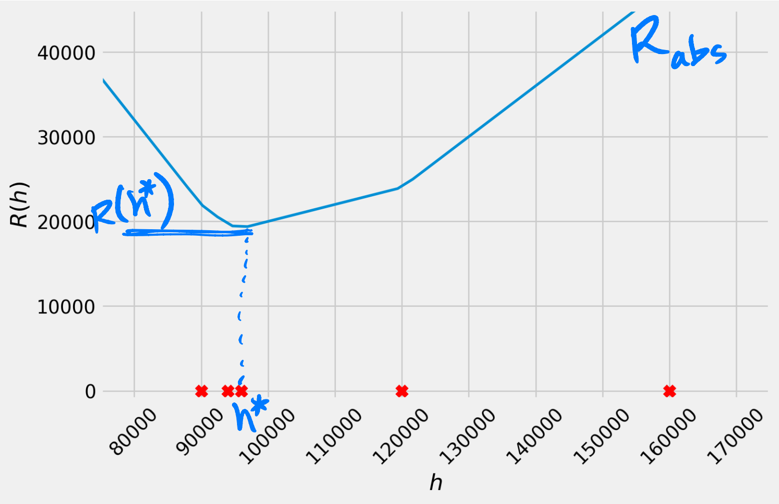
Median = 3

MADM $= \frac{1}{4}(|2-3| + |3-3| + |3-3| + |4-3|)$

$= 1/2$

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

$$\begin{aligned} R_{\text{sq}}(h^*) &= R_{\text{sq}}(\text{Mean}(y_1, y_2, \dots, y_n)) \\ &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \text{Mean}(y_1, y_2, \dots, y_n))^2. \end{aligned}$$

Variance

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

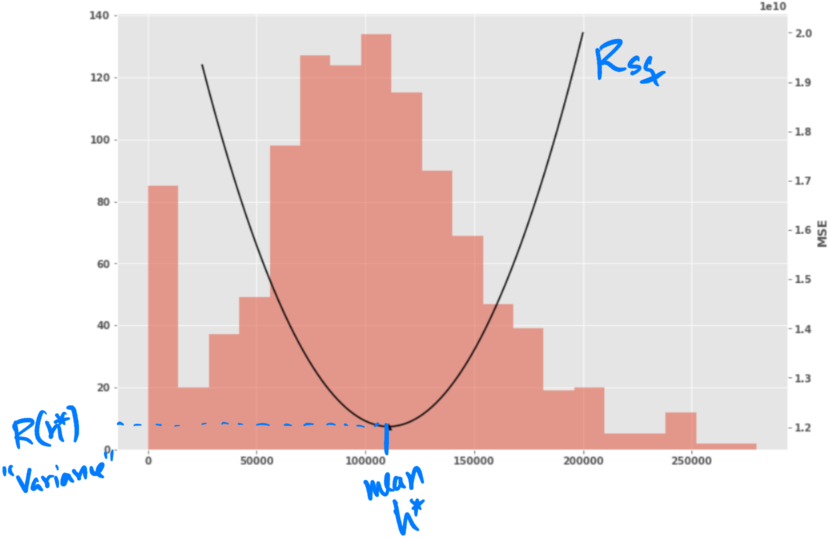
$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \text{Mean}(y_1, y_2, \dots, y_n))^2$$

$$\sqrt{\frac{1}{n}\sum_i (y_i - \mu)^2}$$

- It measures the squared distance of each data point from the mean, on average.

- Its square root is called the **standard deviation.**

Variance


1e10
MSE
0
20
40
60
80
100
120
140
0
50000
100000
150000
200000
250000
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
2.0
Rsq
R(h*)
"Variance"
mean
h*

0-1 loss

- 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.

- $R_{0,1}(h)$ is minimized at $h^* = \text{Mode}(y_1, y_2, \ldots, y_n)$.

- Therefore, $R_{0,1}(h^*)$ is the proportion of data points not equal to the mode.

A poor way to measure spread

- The minimium 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.

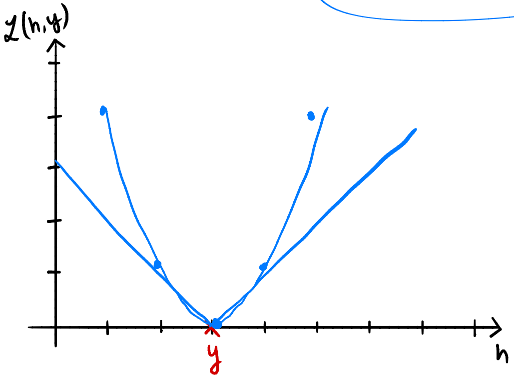
Summary of center and spread

- 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**.

A new loss function

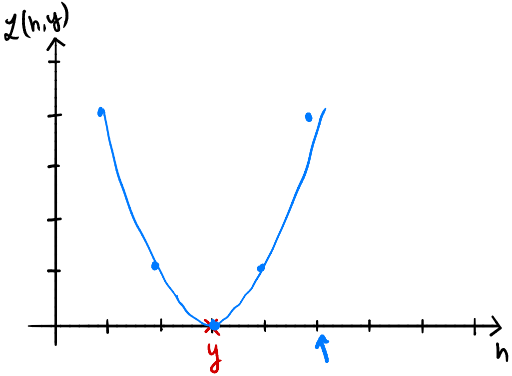
Plotting a loss function

- 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 $L_{sq}(h, y) = (y - h)^2$:



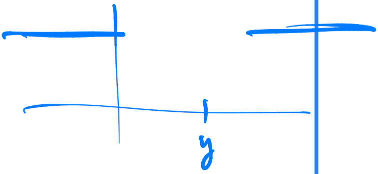
Discussion Question

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

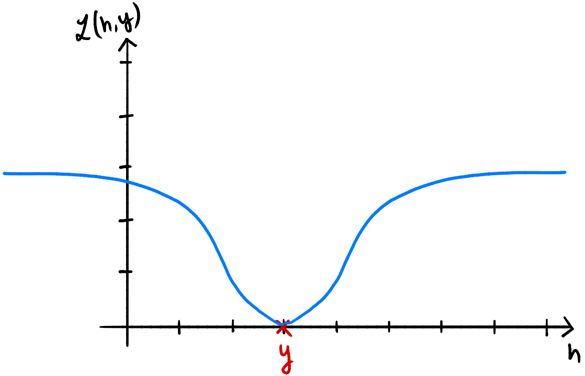
a) flat

b) rapidly decreasing

c) rapidly increasing



A very insensitive loss



- We'll call this loss L_{ucsd} because we made it up at UCSD.

Discussion Question

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

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

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

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

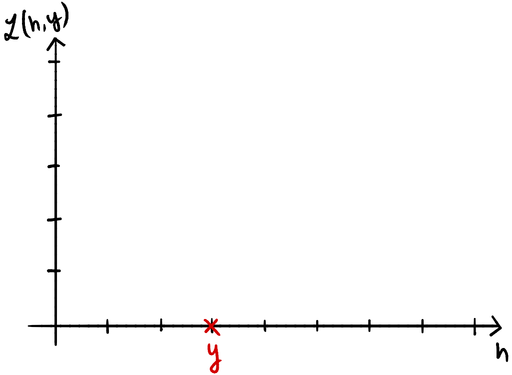
d) $1 - e^{-|y-h|}$

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:

$$L_{ucsd}(h, y) = 1 - e^{-(y-h)^2/\sigma^2}$$

Scale parameter controls width of bowl



Empirical risk minimization

- We have salaries $y_1, y_2, ..., y_n$.
- 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}

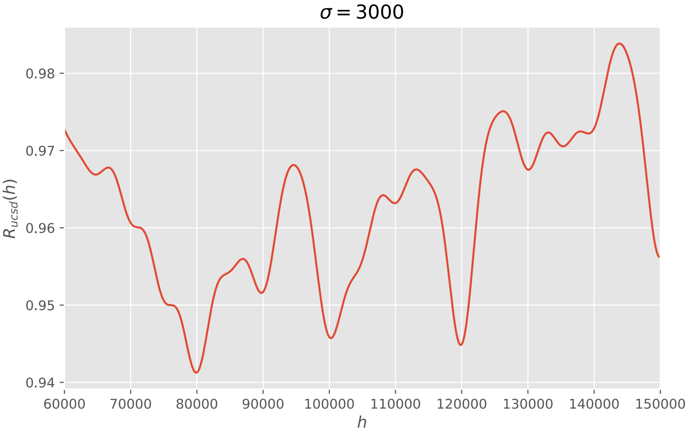
- Recall:

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

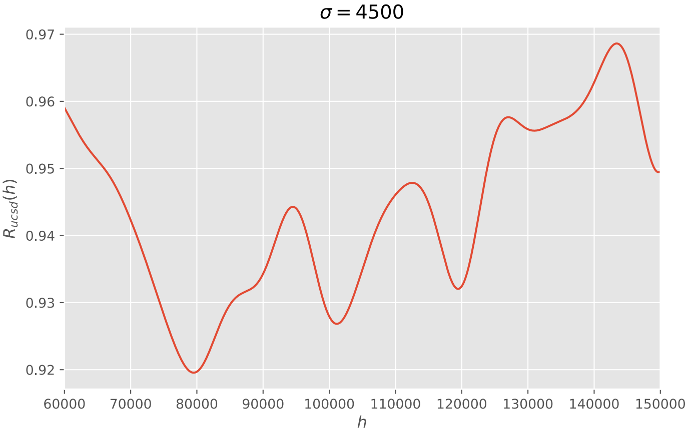
- Once we have data $y_1, y_2, \dots, y_n$ and a scale σ, we can plot $R_{ucsd}(h)$.

- Let's try several scales, σ, for the data scientist salary data.

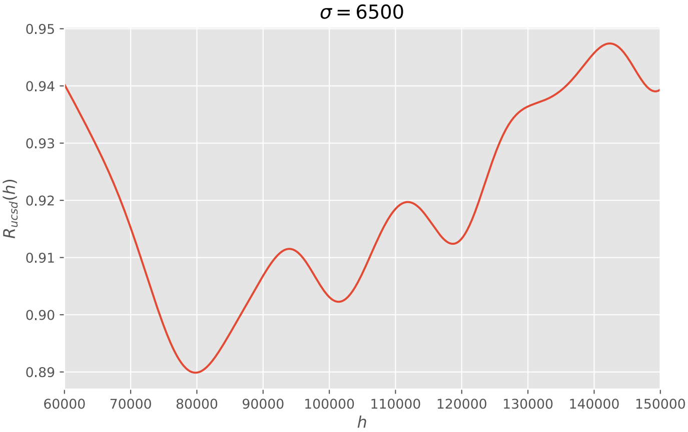
Plot of $R_{ucsd}(h)$



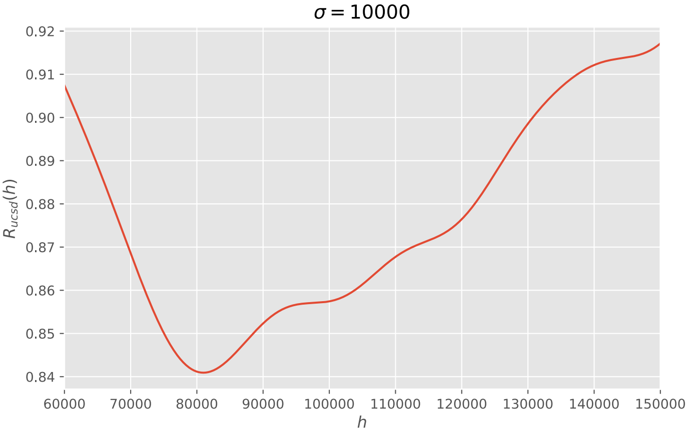
Plot of $R_{ucsd}(h)$



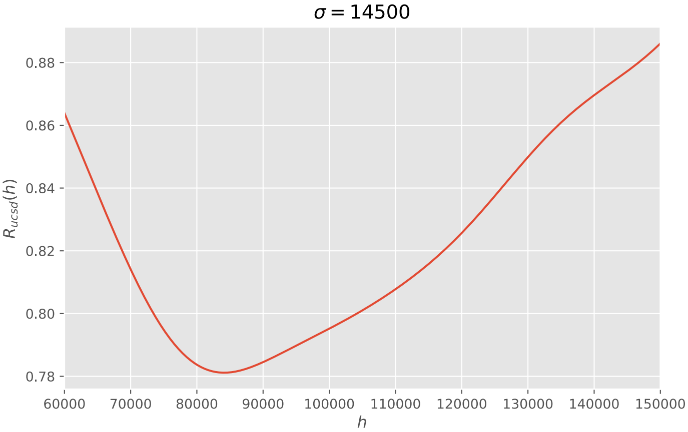
Plot of $R_{ucsd}(h)$



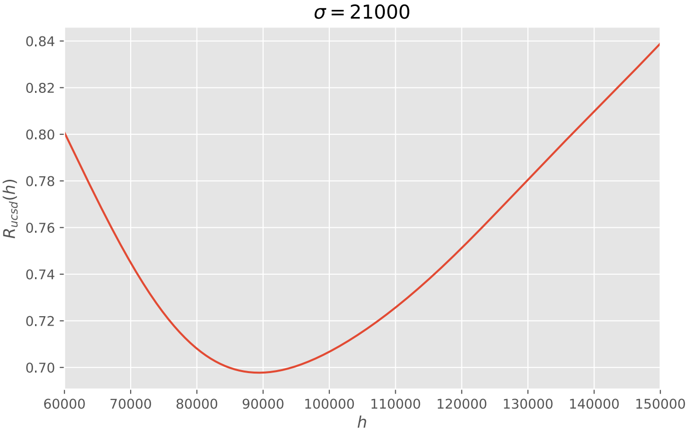
Plot of $R_{ucsd}(h)$



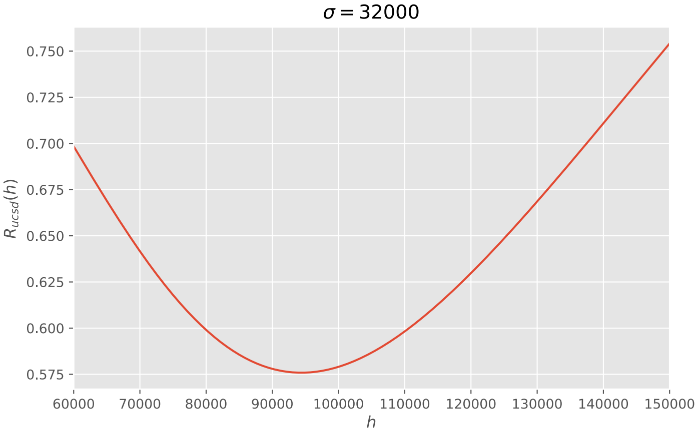
Plot of $R_{ucsd}(h)$



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.

Step 1: Taking the derivative

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

Step 2: Setting to zero and solving

- We found:

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

- 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.

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.
- **Next Time:** We'll learn a computational tool to approximate the minimizer of R_{ucsd}.