

Lecture 4 – Spread, Other Loss Functions



DSC 40A, Fall 2022 @ UC San Diego

Mahdi Soleymani, with help from [many others](#)

Announcements

- ▶ **Groupwork 1 is out, due Monday, Oct. 3 at 11:59pm. Must submit in groups of 2-4.**
- ▶ HW 1 will be released today, **due Friday Oct. 7 at 2:00pm.**
- ▶ First Discussion session: Monday, Oct. 3.
- ▶ If you have a group you want to meet with outside of discussion, go for it.

Agenda

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

Recap of empirical risk minimization

Empirical risk minimization

- ▶ **Goal:** Given a dataset y_1, y_2, \dots, 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”.

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

Absolute loss and squared loss

- ▶ General form of empirical risk:

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

Average \leftarrow loss function

- ▶ **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 $h^* = \mathbf{Median}(y_1, y_2, \dots, y_n)$.
- ▶ **Squared loss:** $L_{\text{sq}}(h, y) = (y - h)^2$.
 - ▶ Empirical risk: $R_{\text{sq}}(h) = \frac{1}{n} \sum_{i=1}^n (y_i - h)^2$. Also called “**mean squared error**”.
 - ▶ Minimized by $h^* = \mathbf{Mean}(y_1, y_2, \dots, y_n)$.

Discussion Question

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

Recall,

$$(a+b)^2 = (a^2 + b^2) + 2ab$$

Dataset $\{1, 5\}$

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

$$R(4) \quad h=4$$

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

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

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

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

a) True

b) False

To answer, go to [menti.com](https://www.menti.com) and enter the code 7933 4859.

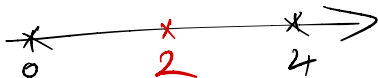
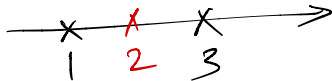
Center and spread

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

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

notion for spread

Mean absolute deviation from the median

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

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

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

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

Discussion Question

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

$$\text{Median} = 3$$

a) 0

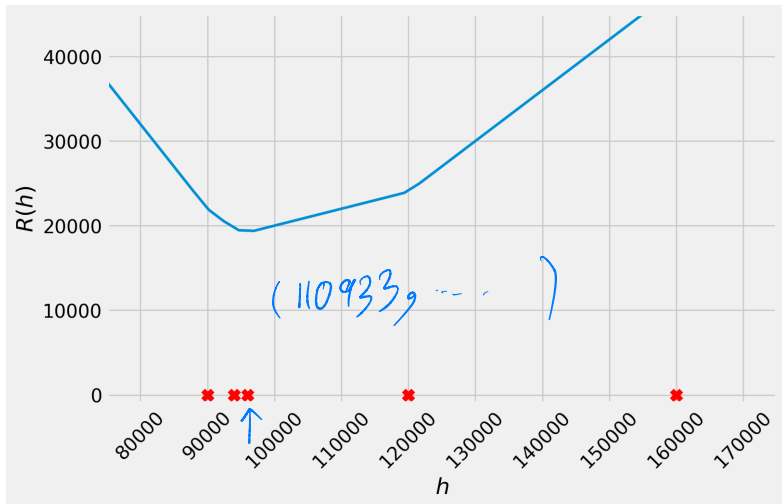
b) $\frac{1}{2}$

c) 1

d) 2

To answer, go to [menti.com](https://www.menti.com) and enter the code 7933 4859.

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

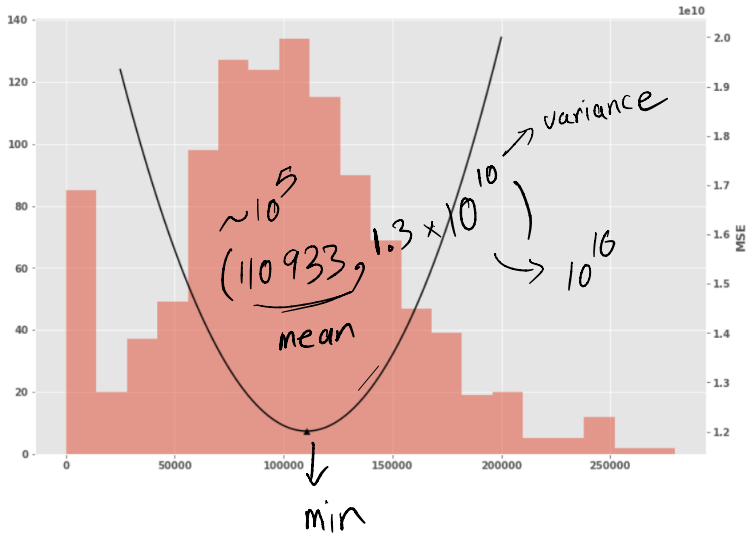
$$\frac{1}{n} \sum_{i=1}^n (y_i - \text{Mean}(y_1, y_2, \dots, y_n))^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2$$

↓
mean

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

$$\sqrt{\text{Variance}} = \text{Standard Deviation}$$

Variance



0-1 loss

- ▶ The empirical risk for the 0-1 loss is

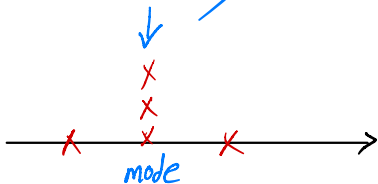
$$R_{0,1}(h) = \frac{1}{n} \sum_{i=1}^n \begin{cases} 0, & \text{if } h = y_i \\ 1, & \text{if } h \neq y_i \end{cases}$$

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

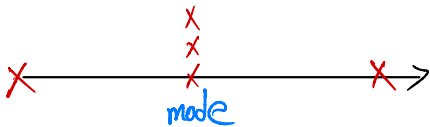
A poor way to measure spread

- ▶ 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.

$$R_{0,1} = \frac{2}{5}$$



$$R_{0,1} = \frac{2}{5}$$



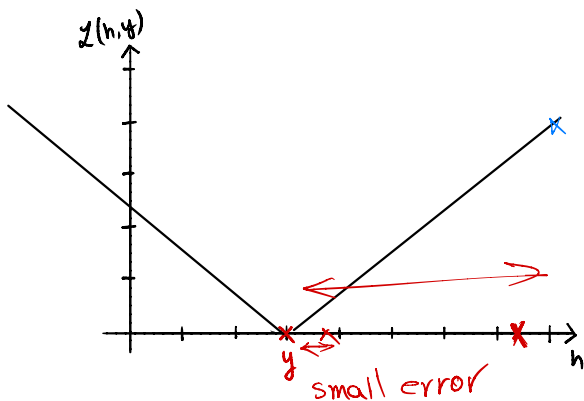
Summary of center and spread

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

A new loss function

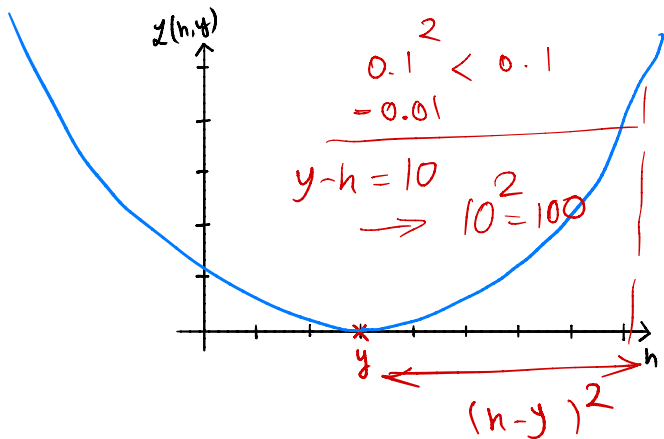
Plotting a loss function

- ▶ 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 $L_{sq}(h, y) = (y - h)^2$:

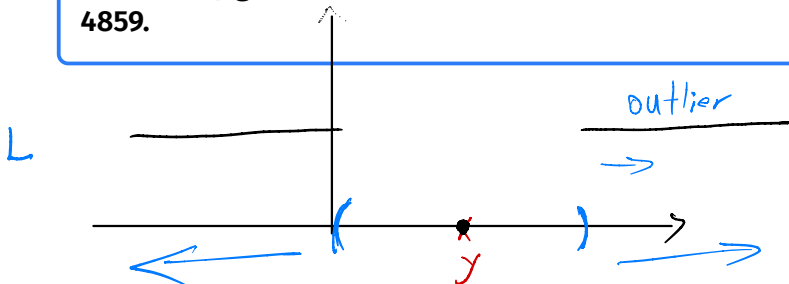


Discussion Question

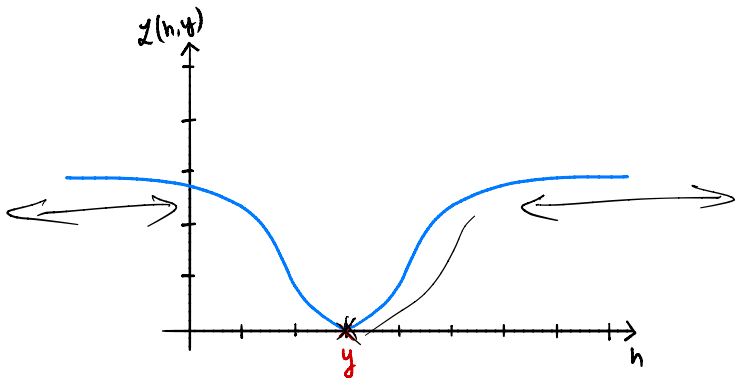
Suppose L considers all outliers to be equally as bad. What would it look like far away from y ?

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

To answer, go to [menti.com](https://www.menti.com) and enter the code 7933 4859.



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

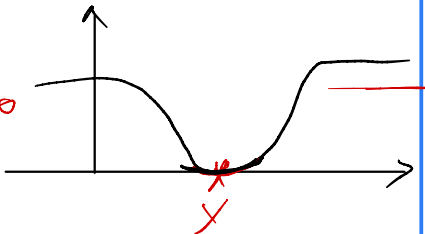
$x = h$

✓ b) $1 - e^{-(y-h)^2}$

$e^{-\infty} = 0$

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

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



To answer, go to [menti.com](https://www.menti.com) and enter the code 7933 4859.

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 :

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

Adding a scale parameter

Empirical risk minimization

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

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

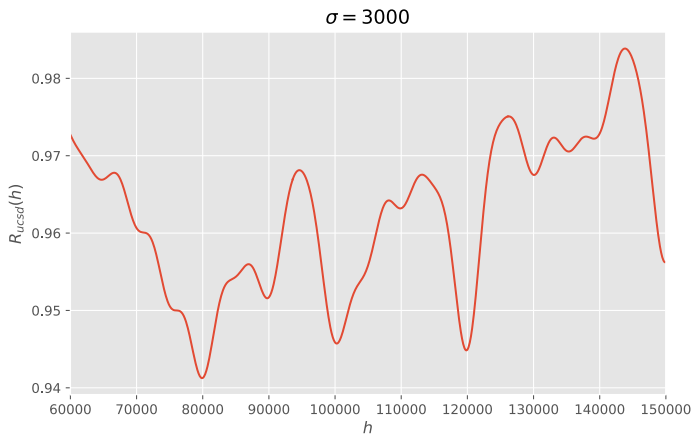
Let's plot R_{ucsd}

- ▶ Recall:

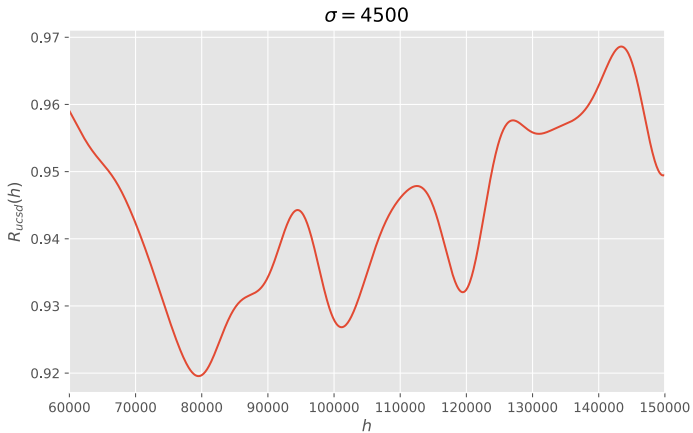
$$R_{ucsd}(h) = \frac{1}{n} \sum_{i=1}^n \left[1 - e^{-(y_i-h)^2/\sigma^2} \right]$$

- ▶ 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 .

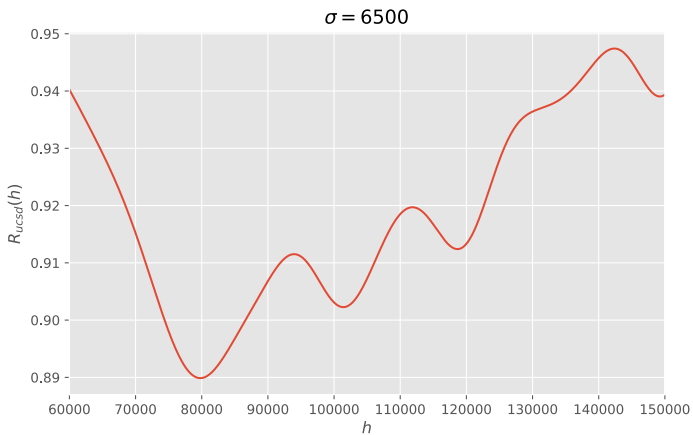
Plot of $R_{ucsd}(h)$



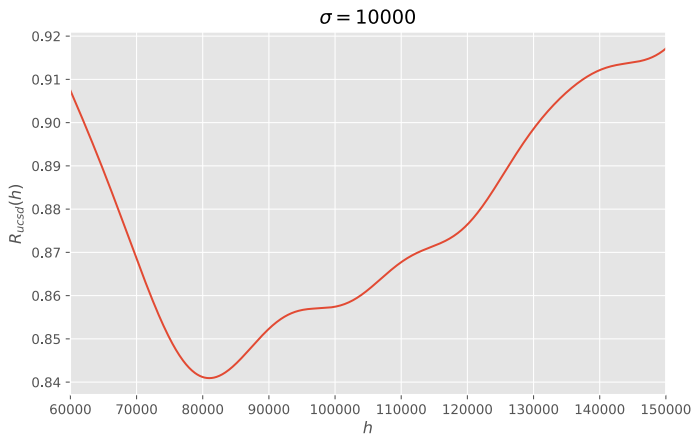
Plot of $R_{ucsd}(h)$



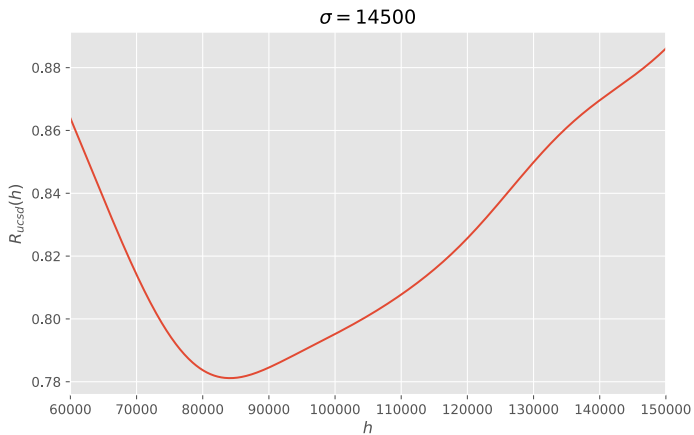
Plot of $R_{ucsd}(h)$



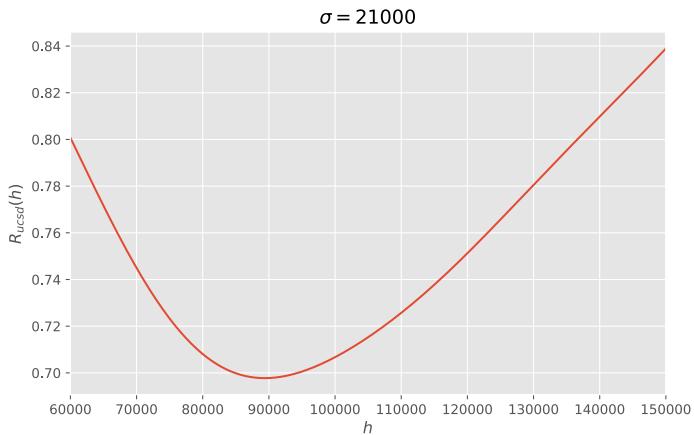
Plot of $R_{ucsd}(h)$



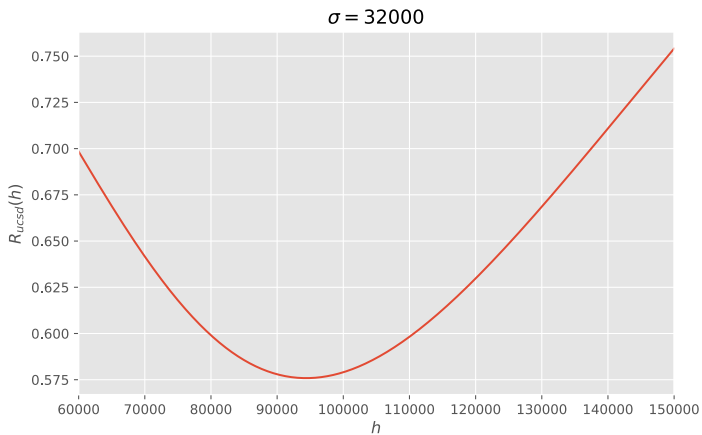
Plot of $R_{ucsd}(h)$



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.

Step 1: Taking the derivative

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

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???

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.