

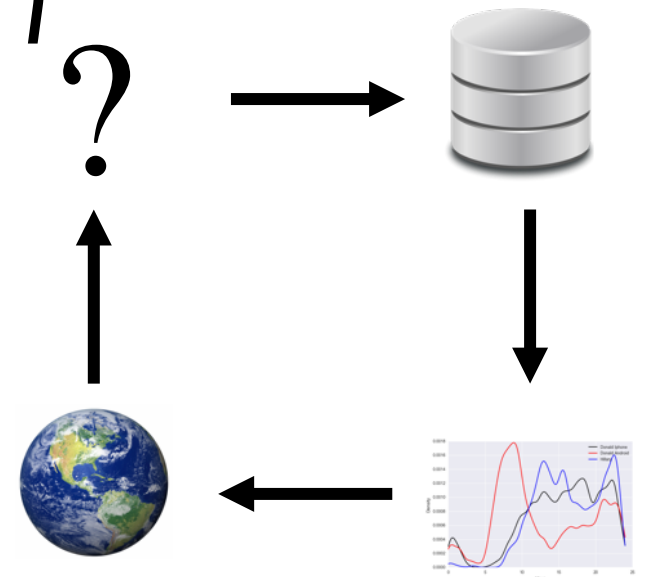
# Data Science 100

## Lecture 13: Modeling and Estimation

Slides by:

Joseph E. Gonzalez, [jegonzal@berkeley.edu](mailto:jegonzal@berkeley.edu)

2018 updates - Fernando Perez, [fernando.perez@berkeley.edu](mailto:fernando.perez@berkeley.edu)



# Recap ... so far we have covered

- **Data collection:** Surveys, sampling, administrative data
- **Data cleaning and manipulation:** Pandas, text & regexes.
- **Exploratory Data Analysis**
  - Joining and grouping data
  - Structure, Granularity, Temporality, Faithfulness and Scope
  - Basic exploratory data visualization
- **Data Visualization:**
  - Kinds of visualizations and the use of size, area, and color
  - Data transformations using Tukey Mosteller bulge diagram
- **An introduction to database systems and SQL**

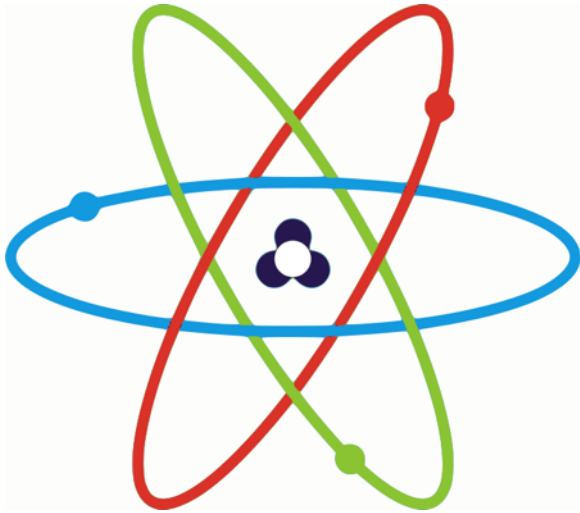
# Today – Models & Estimation

What is a model?

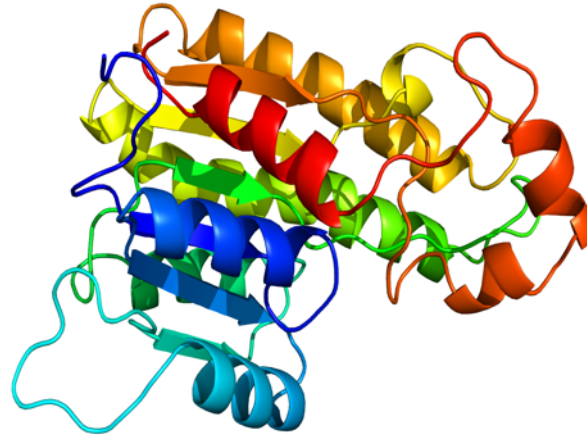


# What is a model?

A model is an an **idealized** representation of a system



Atoms don't actually work like this...



Proteins are far more complex



We haven't really seen one of these.

A black and white portrait of George Box, an elderly man with glasses, resting his chin on his hand. A speech bubble points from him towards the quote.

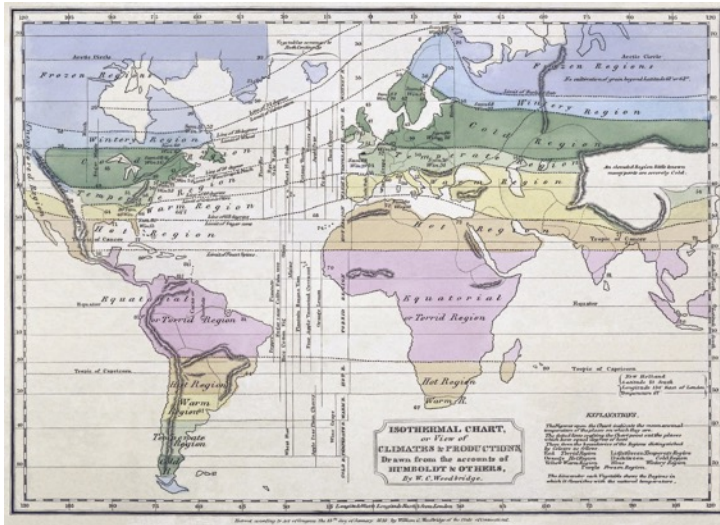
*“Essentially,  
all models are wrong,  
but some are useful.”*

George Box  
Statistician  
1919-2013

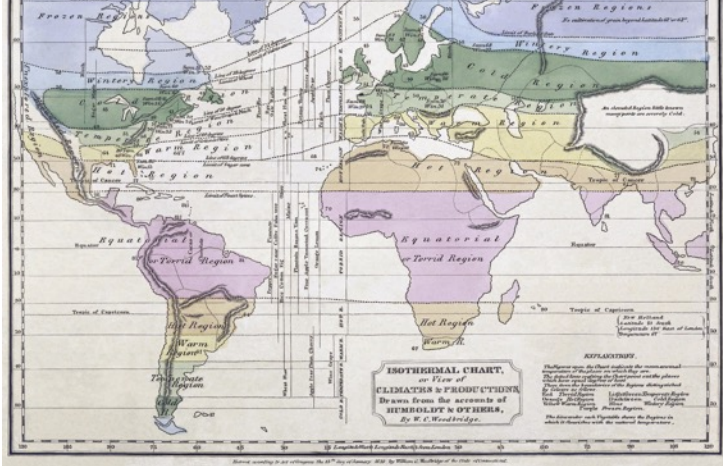
Why do we build models?

# Why do we build models?

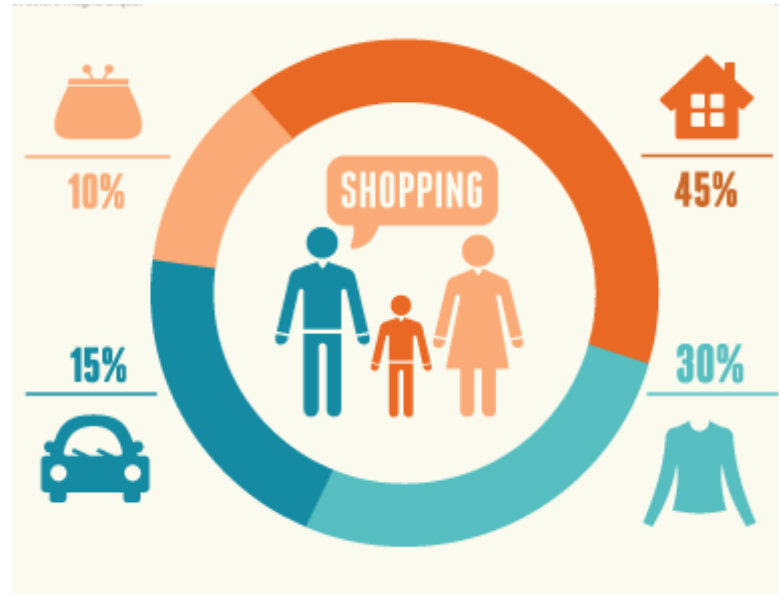
- Models enable us to make **accurate predictions**







➤ **Provide insight** into complex phenomena





# A few types of models: “physical” or “mechanistic”

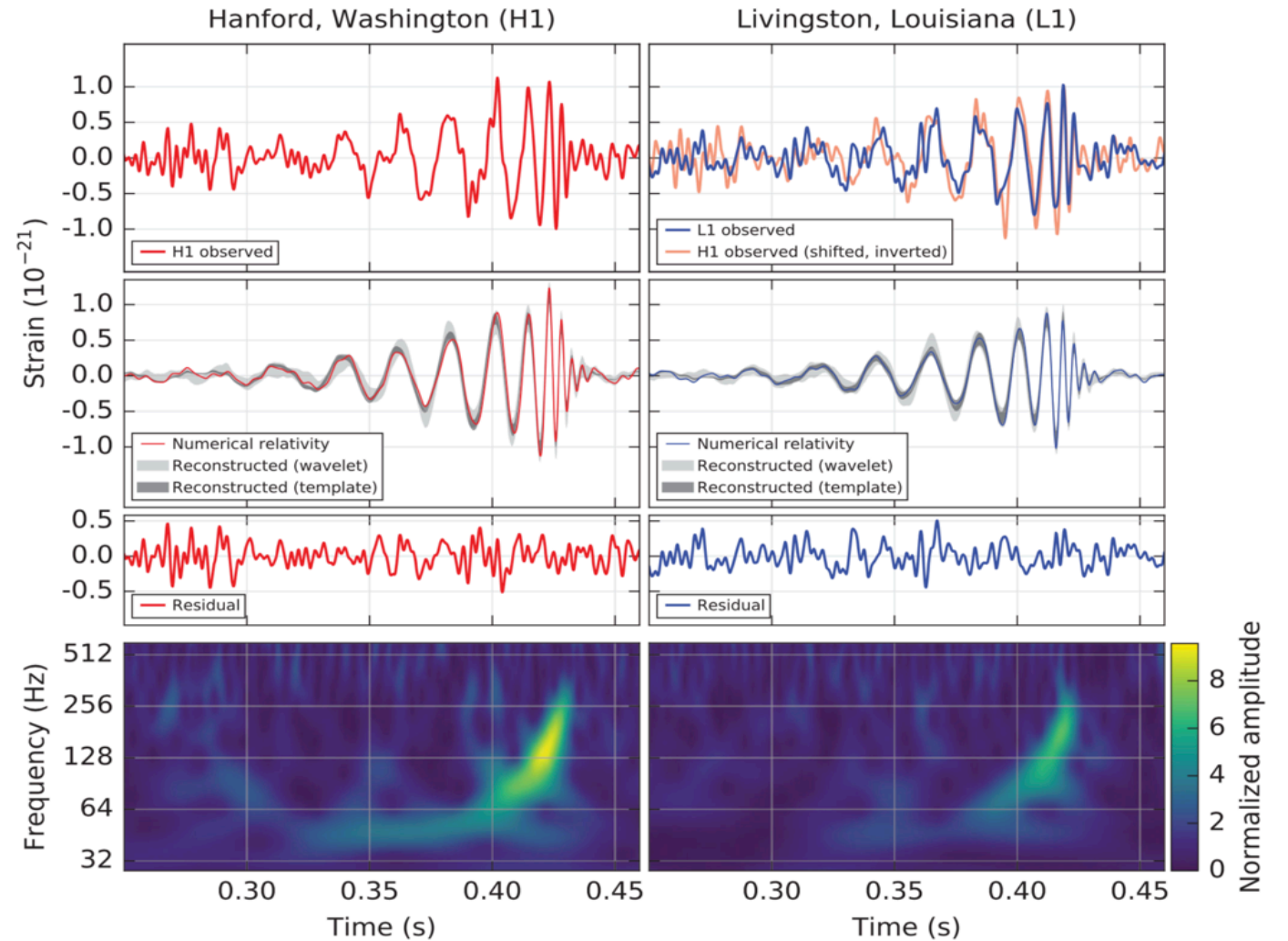
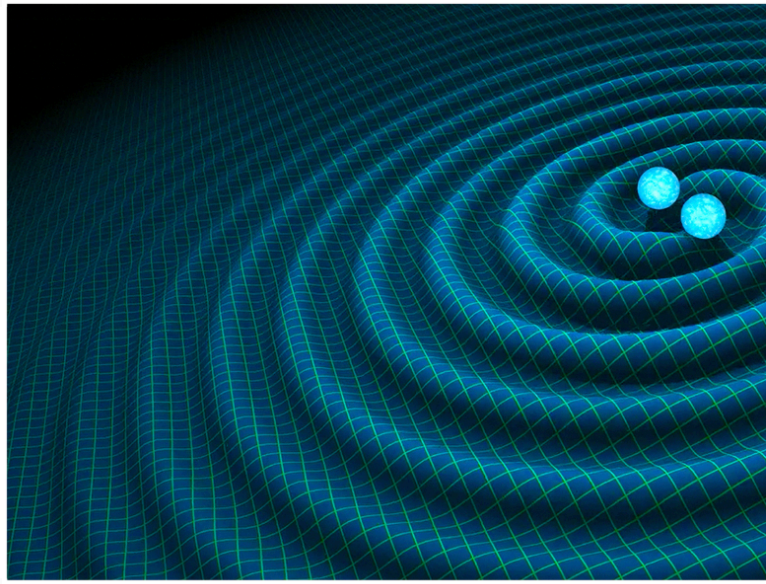
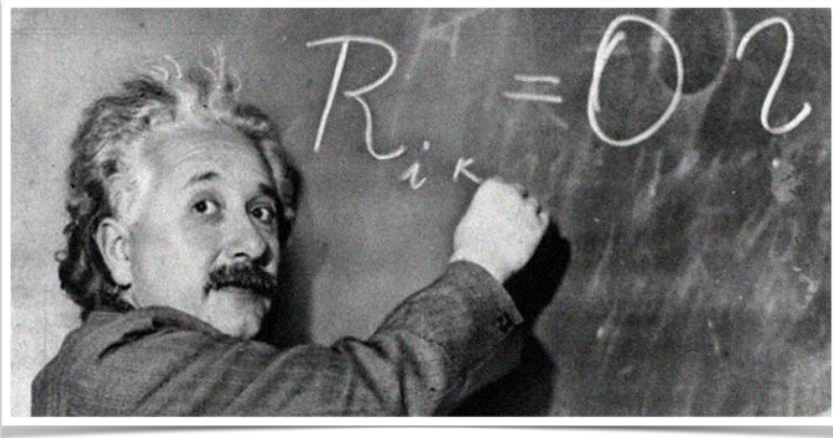
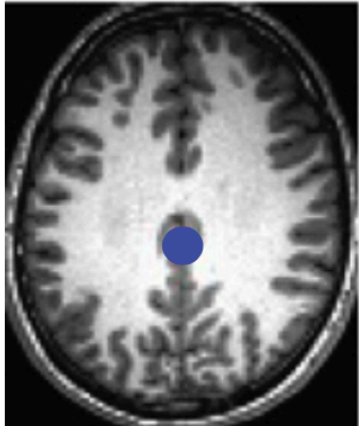
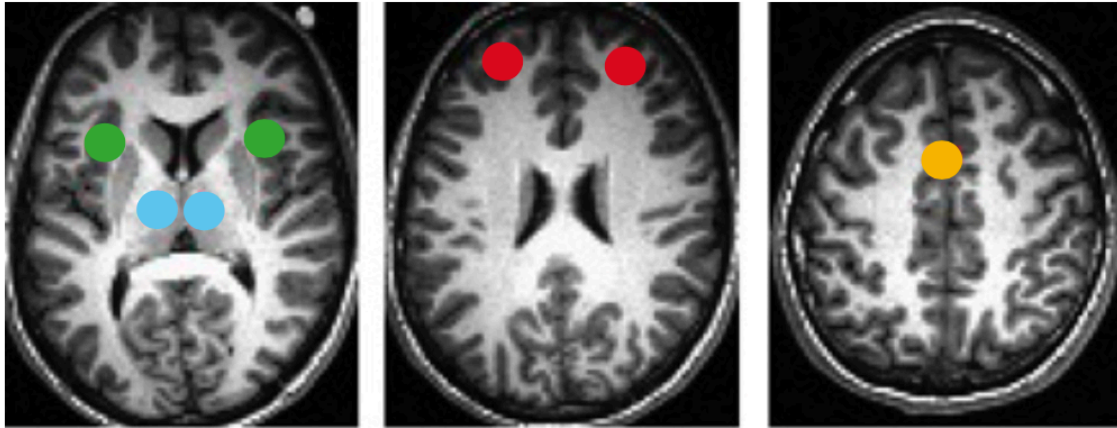


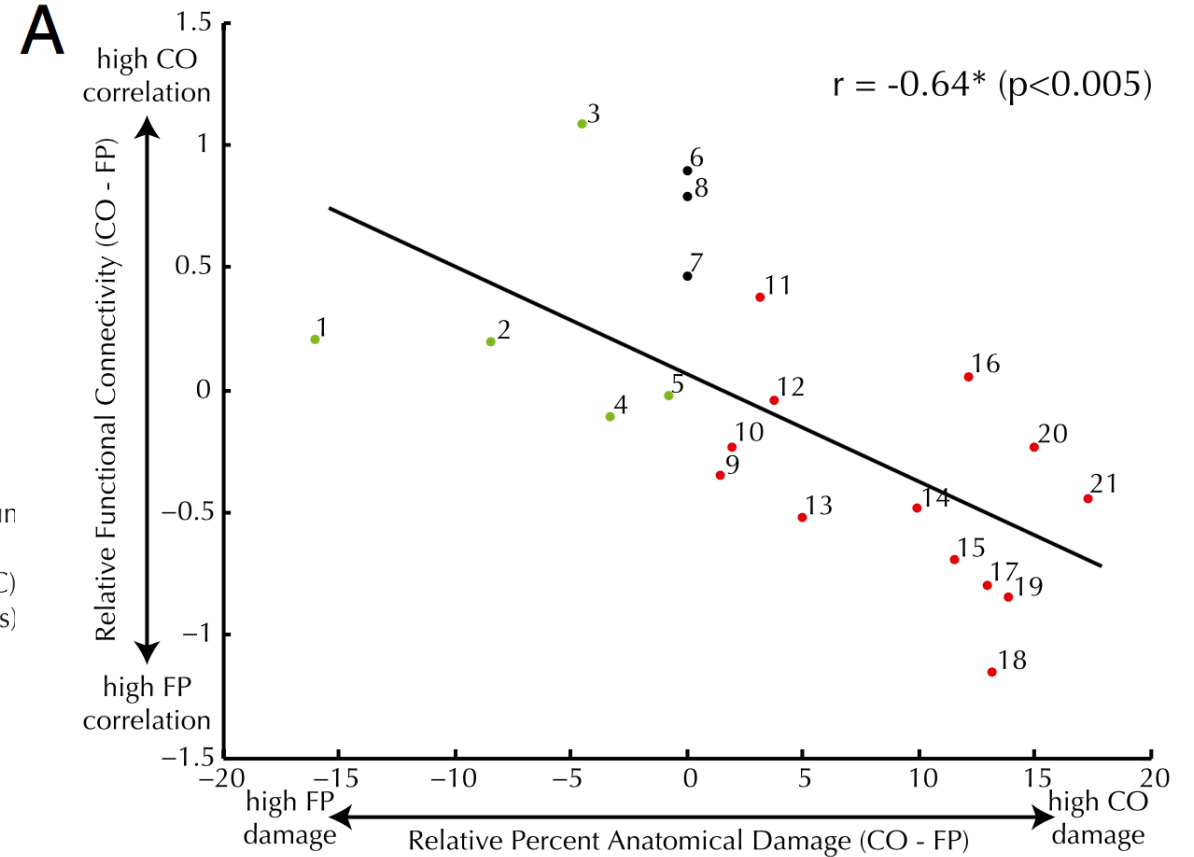
FIG. 1. The gravitational-wave event GW150914 observed by the LIGO Hanford (H1, left column panels) and Livingston (L1, right column panels) detectors. Times are shown relative to September 14, 2015 at 09:50:45 UTC. For visualization, all time series are filtered with a 35–350 Hz bandpass filter to suppress large fluctuations outside the detectors’ most sensitive frequency band, and band-reject

# Models: Statistical correlations (A)

**A** ROI coordinates from Dosenbach et al, 2007



- Cingulo-opercular (CO)**
- [1,2] anterior insula/frontal operculum
  - [3] dorsal anterior cingulate (dACC)
  - [4,5] anterior prefrontal cortex (aPFC)
  - [6,7] anterior thalamus (ant thalamus)
- Frontal-parietal (FP)**
- [1,2] intraparietal sulcus (IPS)
  - [3,4] frontal cortex
  - [5,6] precuneus
  - [7,8] intraparietal lobule (IPL)
  - [9,10] dorsolateral prefrontal cortex (dlPFC)
  - [11] midcingulate





# Models: statistical correlations (B)



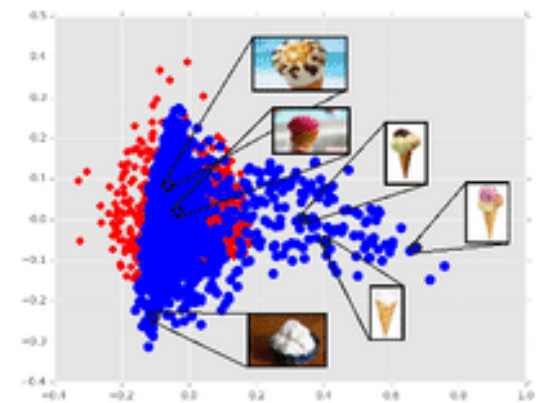
Icecream - ILSVRC12 (IN)



Icecream - WIN



Icecream - WINC



ILSVR red, WINC blue



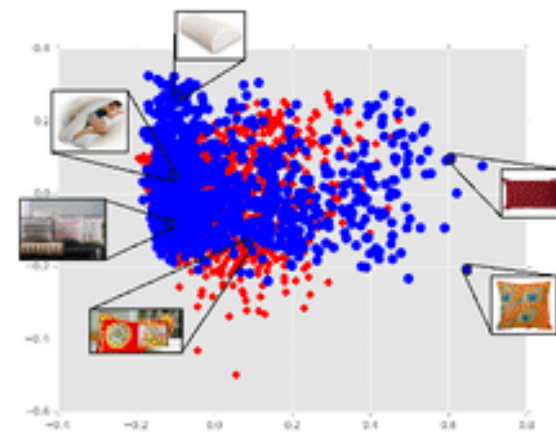
Pillow - ILSVRC12 (IN)



Pillow - WIN



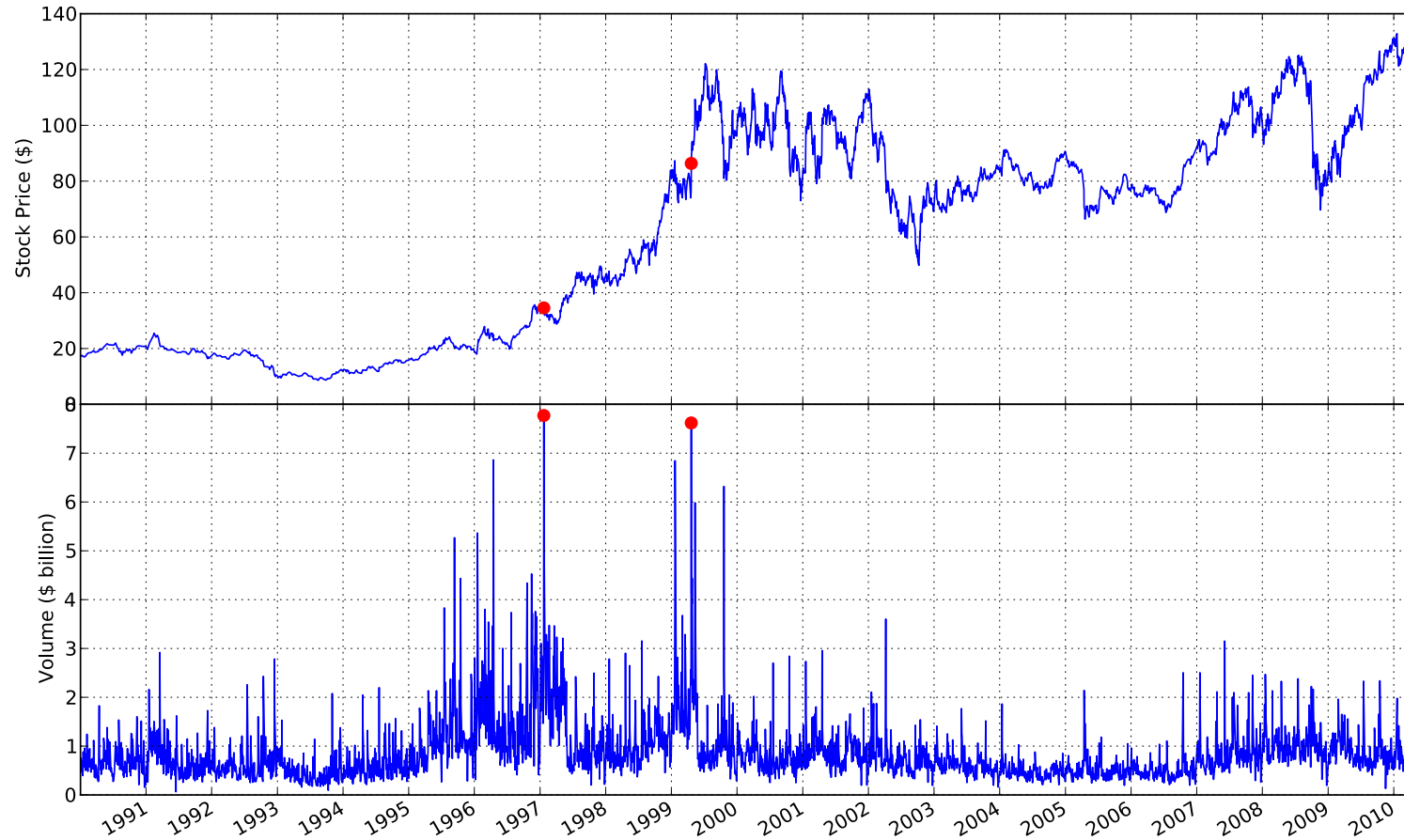
Pillow - WINC



ILSVR red, WINC blue



# Models: statistical correlations (C)



# Models and the World

- **Data Generation Process:** the real-world phenomena from which the data is collected
  - **Example:** *everyday there are some number of clouds and it rains or doesn't*
  - We don't know or can't compute this, could be stochastic or adversarial
- **Model:** a theory of the data generation process
  - **Example:** *if there are more than  $X$  clouds then it will rain*
  - How do we pick this model? EDA? Art?
  - May not reflect reality ... "all models are wrong ..."
- **Estimated Model:** an instantiation of the model
  - **Example:** *if there are more than 42 clouds then it will rain*
  - How do we estimate it?
  - What makes the estimate "good"?

# Example – Restaurant Tips

Follow along with the notebook ...

# Step 1: Understanding the Data (EDA)

```
data = sns.load_dataset("tips")
print("Number of Records:", len(data))
data.head()
```

Number of Records: 244

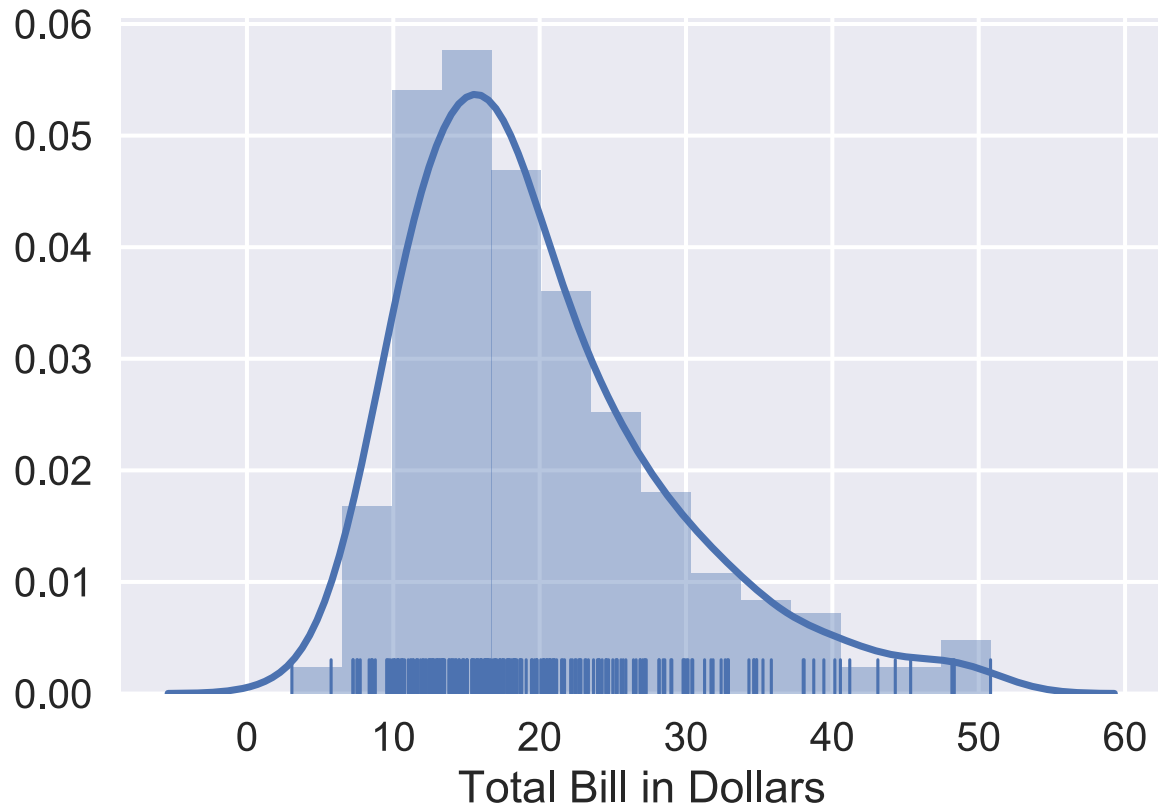
	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4

*Collected by a single waiter over a month*

Why?

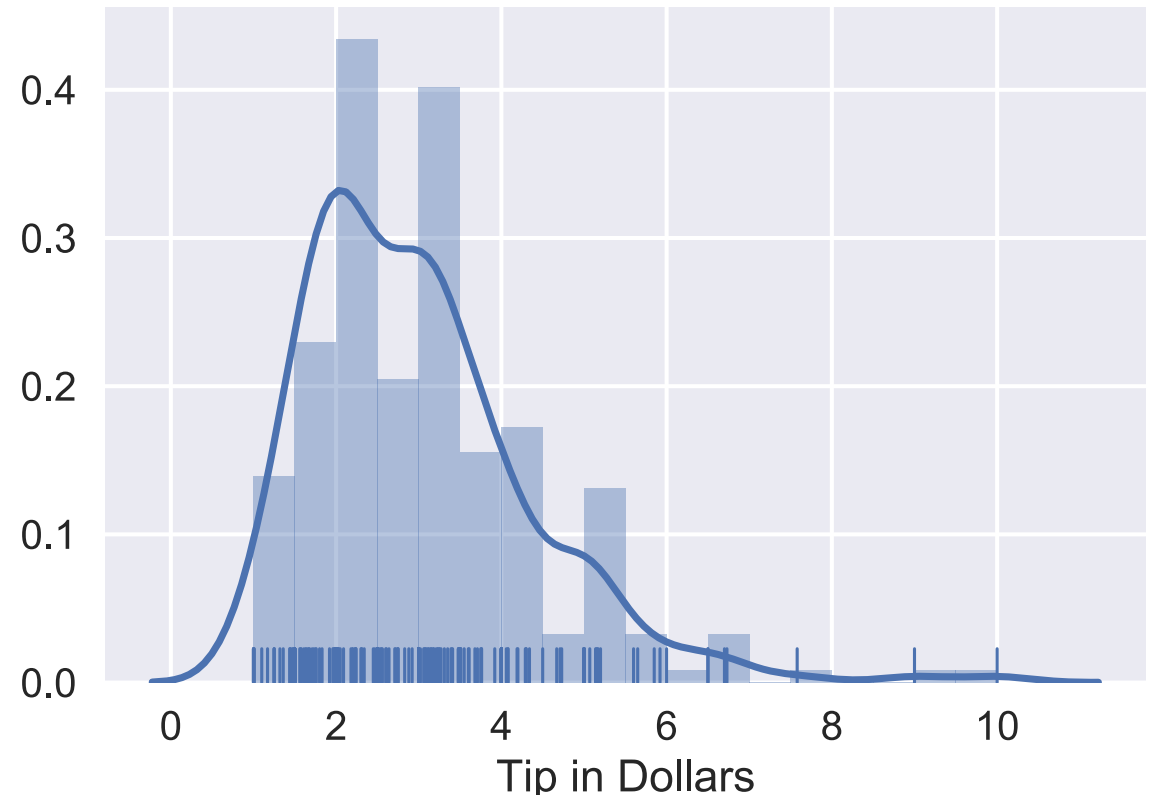
- **Predict** which tables will tip the highest
- **Understand** relationship between tables and tips

# Understanding the Tips



Observations:

- Right skewed
- Mode around \$15
- Mean around \$20
- No large bills



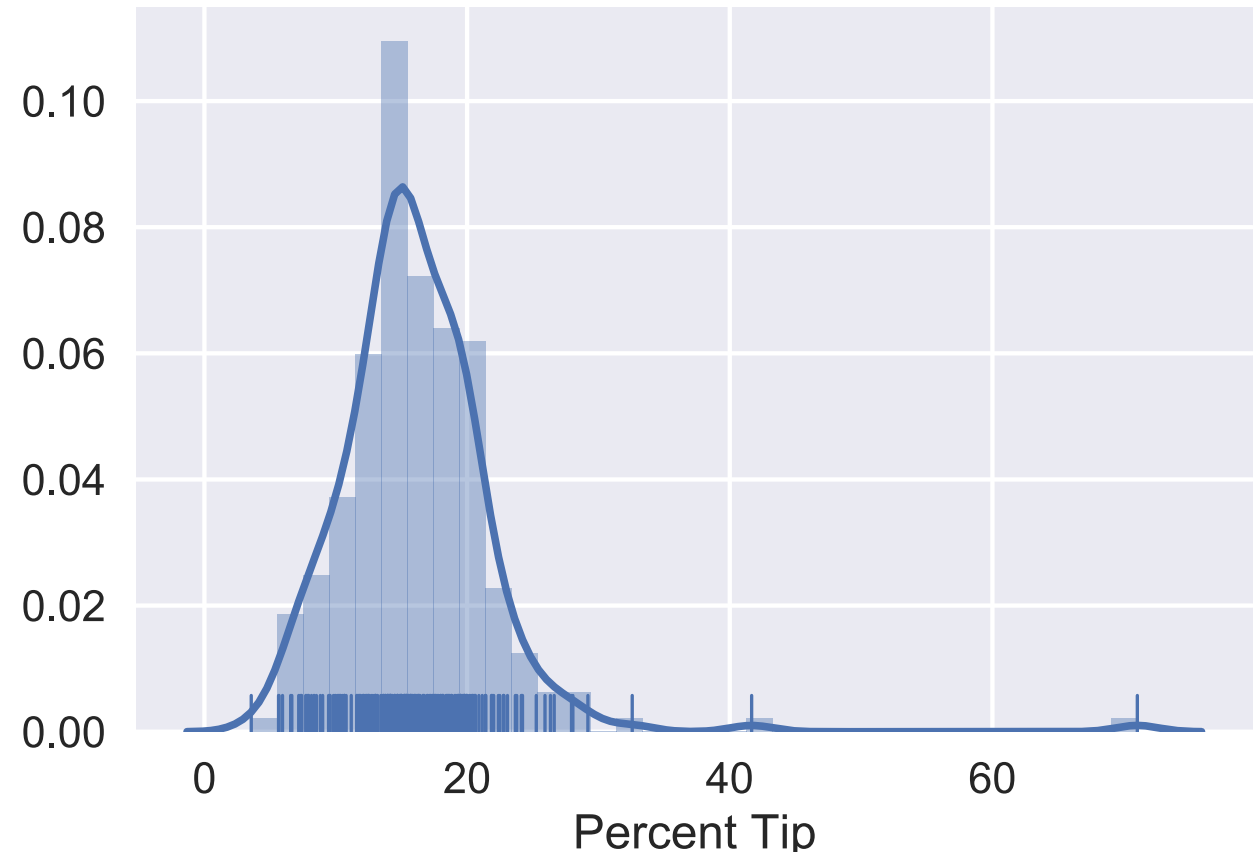
Observations:

- Right skewed
- Mean around 3
- Possibly bimodal? → Explanations?
- Large outliers → Explanations?

# Derived Variable: Percent Tip

$$\text{pct\_tip} = \frac{\text{tip}}{\text{total\_bill}} * 100$$

- Natural representation of tips
  - Why? Tradition in US is to tip %
- Issues in the plot?
  - Outliers
  - Explanation?
    - Small bills ... bad data?
  - Transformations?
    - Remove outliers



# Step 1: Define the Model

START SIMPLE!!

# Start with a **Simple Model**: Constant

$$\text{percentage tip} = \theta^*$$

\* Means true parameter determined by universe

- **Rationale:** There is a percent tip  $\theta^*$  that all customers pay
  - Correct?
    - *No! We have different percentage tips in our data*
    - *Why? Maybe people make mistakes calculating their bills?*
  - Useful?
    - Perhaps. A good estimate  $\theta^*$  could allow us to predict future tips ...
- The **parameter**  $\theta^*$  is determined by the universe
  - we generally don't get to see  $\theta^*$  ...
  - we will need to develop a procedure to **estimate  $\theta^*$  from the data**



# How do we estimate the parameter $\theta^*$

- Guess a number using **prior knowledge**: 15%
- **Use the data!** How?
- Estimate the value  $\theta^*$  as:
  - the percent tip from a **randomly selected** receipt
  - the **mode** of the distribution observed
  - the **mean** of the percent tips observed
  - the **median** of the percent tips observed
- Which is the best? How do I define best?
  - Depends on our goals ...

# Defining an the Objective (Goal)

- **Ideal Goal:** estimate a value for  $\theta^*$  such that the model makes good predictions about the future.
  - **Great goal!** Problem?
    - We don't know the future. How will we know if our estimate is good?
  - There is hope! ... we will return to this goal ... *in the future* 😊
- **Simpler Goal:** estimate a value for  $\theta^*$  such that the model “*fits*” the data
  - What does it mean to “*fit*” the data?
  - We can define a **loss function** that measures the error in our model on the data

# Step 2: Define the Loss

“Take the Loss”

# Loss Functions

- **Loss function:** a function that characterizes the cost, error, or loss resulting from a particular choice of model or model parameters.
- *Many definitions* of loss functions and the choice of loss function affects the **accuracy** and **computational cost of estimation**.
- The choice of loss function **depends on the estimation task**
  - quantitative (e.g., tip) or qualitative variable (e.g., political affiliation)
  - Do we care about the outliers?
  - Are all errors equally costly? (e.g., false negative on cancer test)

# Squared Loss

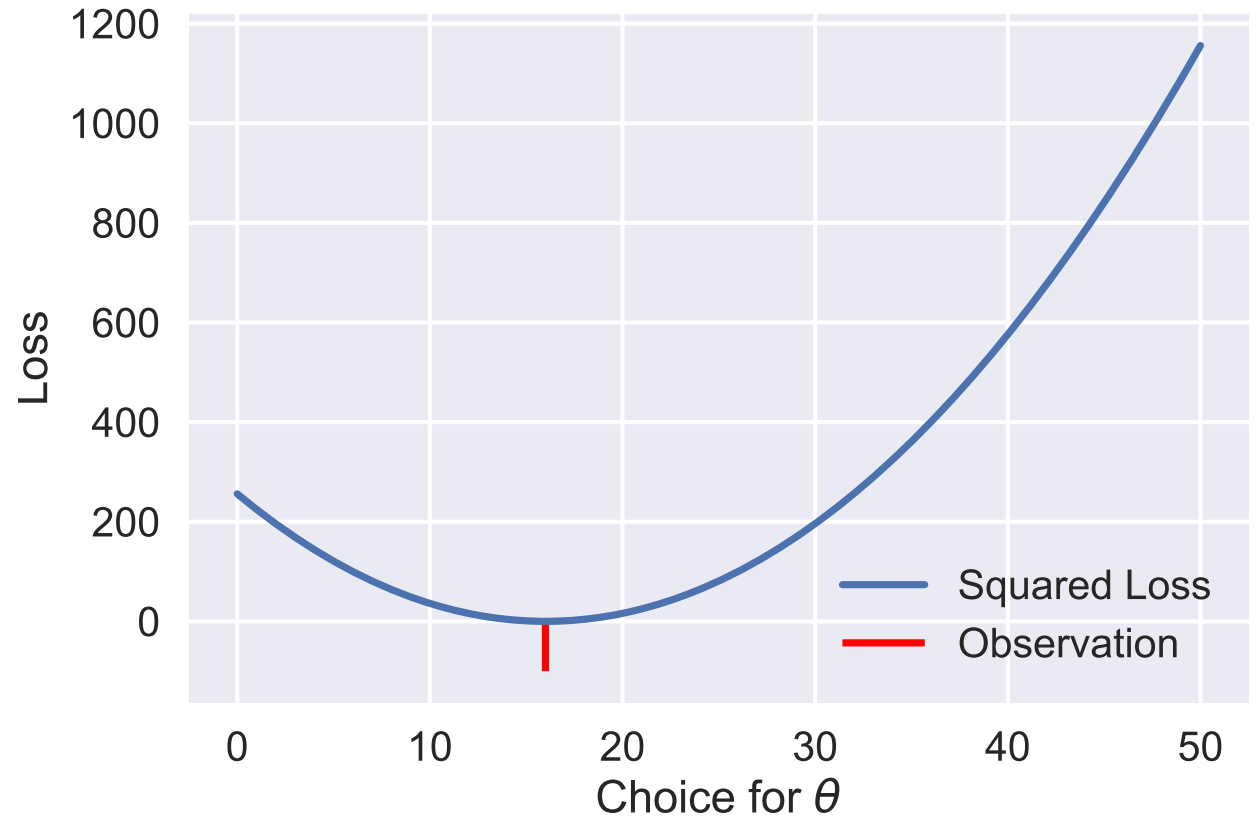
Widely used loss!

The predicted value

The “error” in our prediction

$$L(\theta, y) = (y - \theta)^2$$

An observed data point



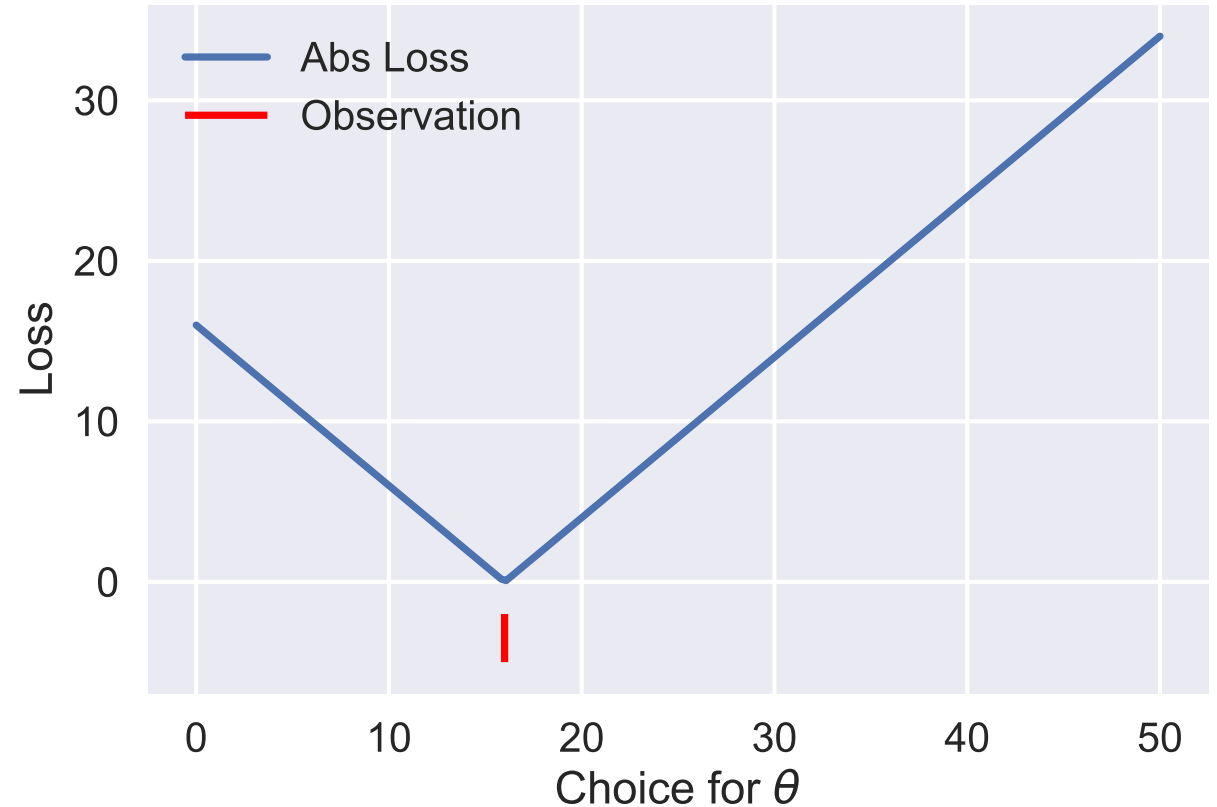
- Also known as the the  $L^2$  loss (pronounced “el two”)
- Reasonable?
  - $\theta = y \rightarrow$  good prediction  $\rightarrow$  good fit  $\rightarrow$  no loss!
  - $\theta$  far from  $y \rightarrow$  bad prediction  $\rightarrow$  bad fit  $\rightarrow$  lots of loss!

# Absolute Loss

It sounds worse than it is ...

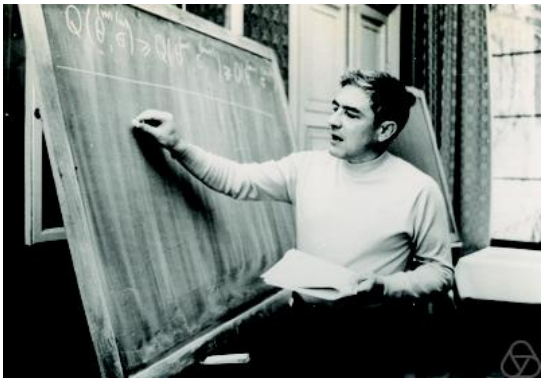
$$L(\theta, y) = |y - \theta|$$

Absolute value



- Also known as the the  $L^1$  loss (pronounced “el one”)
- Reasonable?
  - $\theta = y \rightarrow$  good prediction  $\rightarrow$  good fit  $\rightarrow$  no loss!
  - $\theta$  far from  $y \rightarrow$  bad prediction  $\rightarrow$  bad fit  $\rightarrow$  some loss

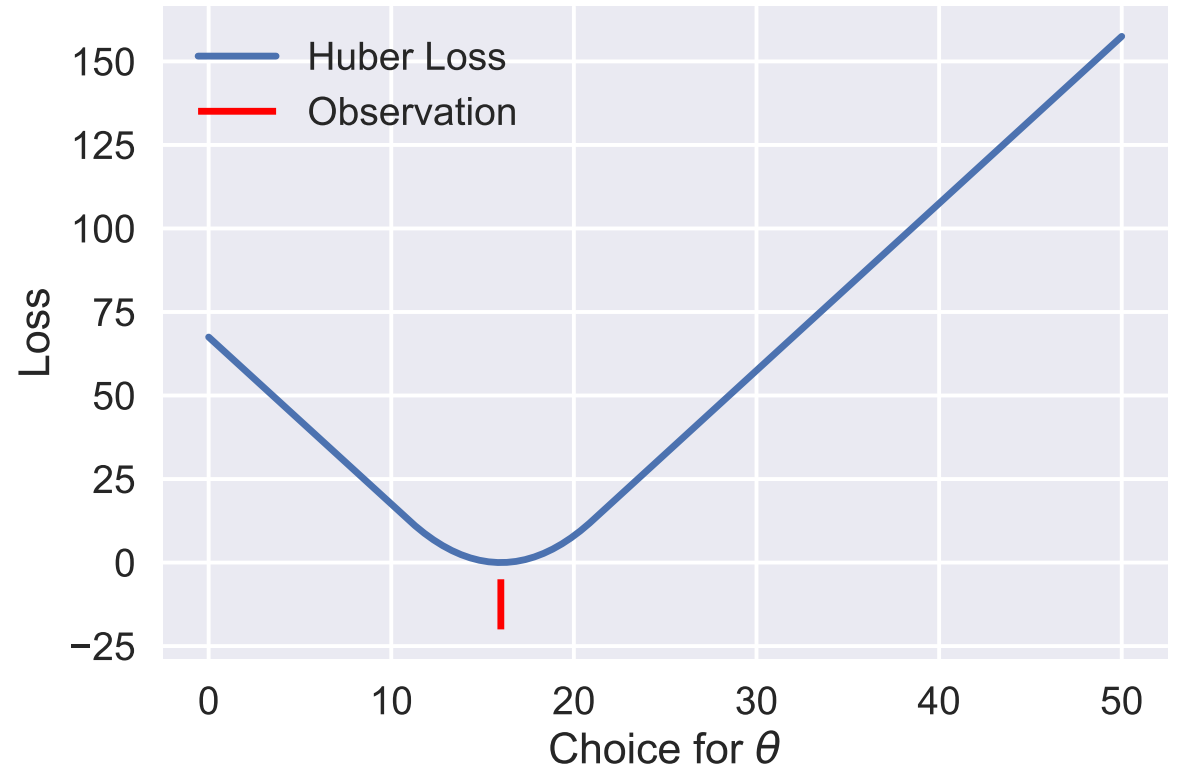
Can you think of another  
Loss Function?



$$L_{\alpha}(\theta, y) = \begin{cases} \frac{1}{2} (y - \theta)^2 & |y - \theta| < \alpha \\ \alpha (|y - \theta| - \frac{\alpha}{2}) & \text{otherwise} \end{cases}$$

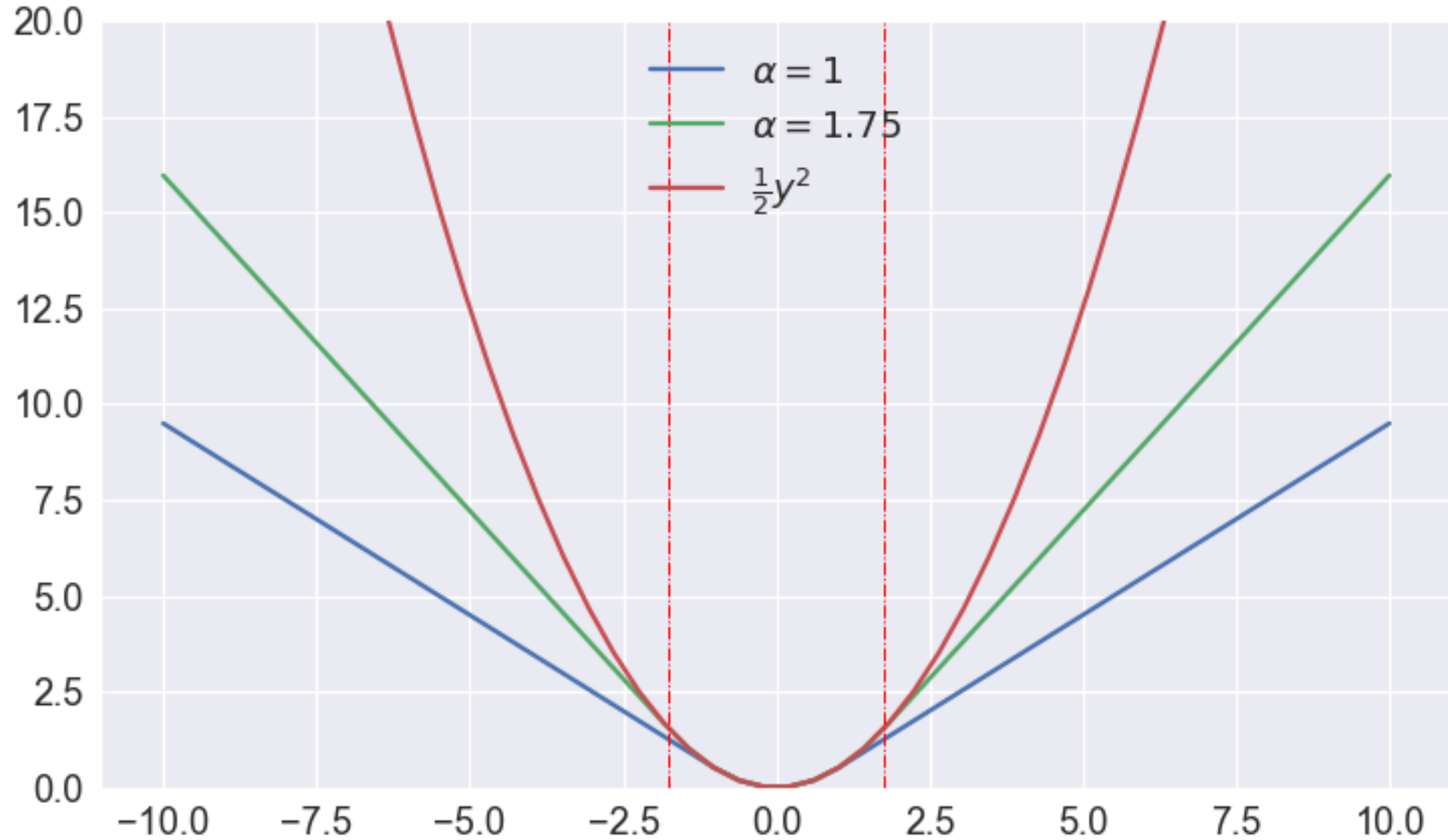
# Huber Loss

- Parameter  $\alpha$  that we need to choose.
- Reasonable?
  - $\theta = y \rightarrow$  good prediction  
 $\rightarrow$  good fit  $\rightarrow$  no loss!
  - $\theta$  far from  $y \rightarrow$  bad prediction  
 $\rightarrow$  bad fit  $\rightarrow$  some loss
- A hybrid of the L2 and L1 losses...



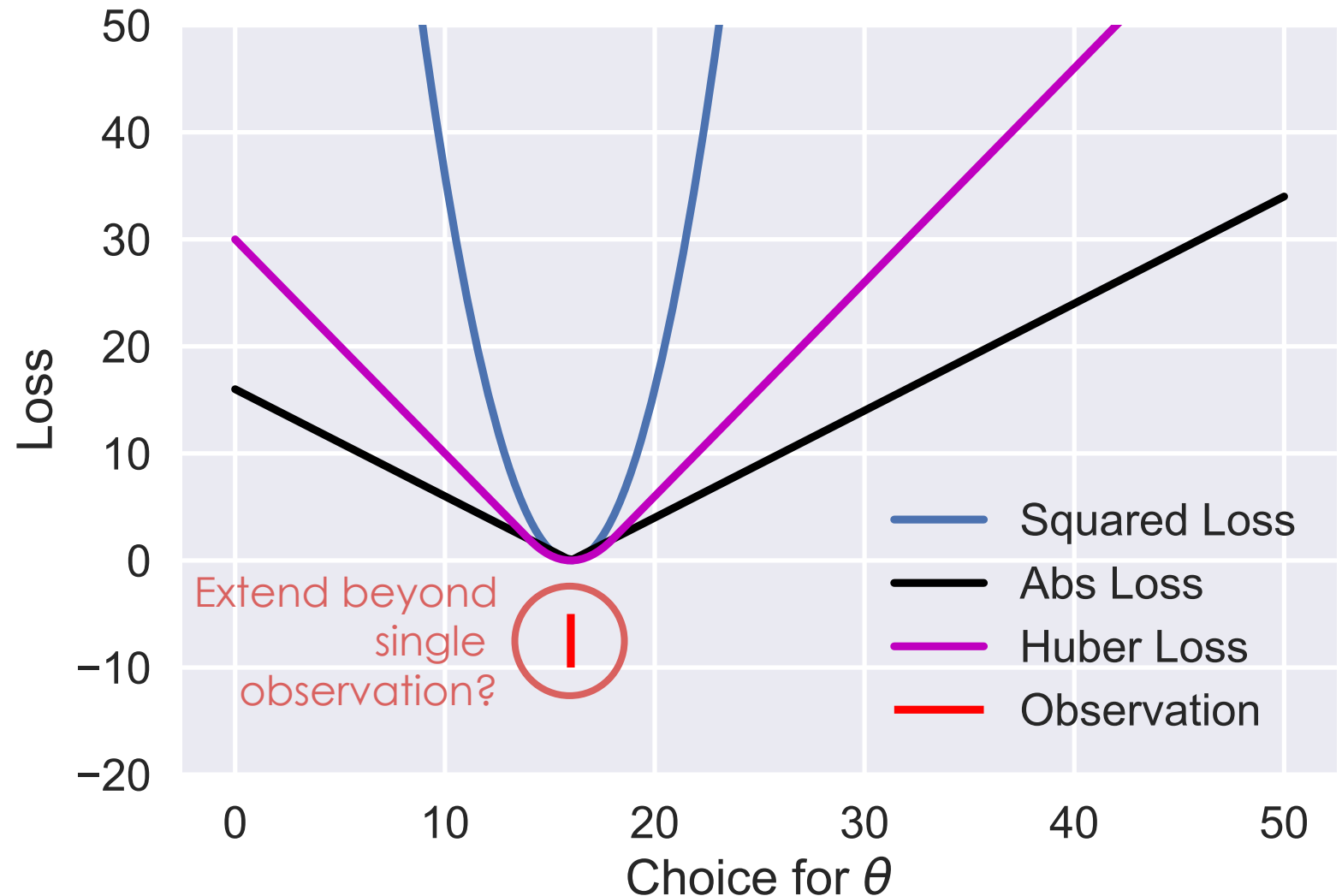


# The Huber loss function, interactively



# Comparing the Loss Functions

- All functions are zero when  $\theta = y$
- Different penalties for being far from observations
- Smooth vs. not smooth
- Which is the best?
  - Let's find out



# Average Loss

- A natural way to define the loss on our entire dataset is to compute the average of the loss on each record.

$$L(\theta, \mathcal{D}) = \frac{1}{n} \sum_{i=1}^n L(\theta, y_i)$$

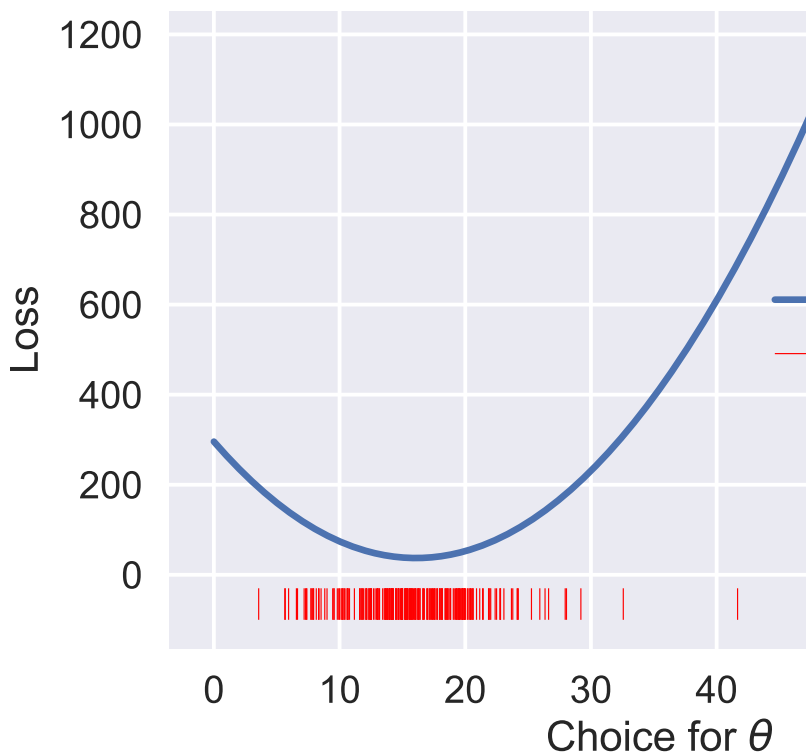
The set of  $n$  data points

- In some cases we might take a weighted average (when?)
  - Some records might be more important or reliable
- What does the average loss look like?

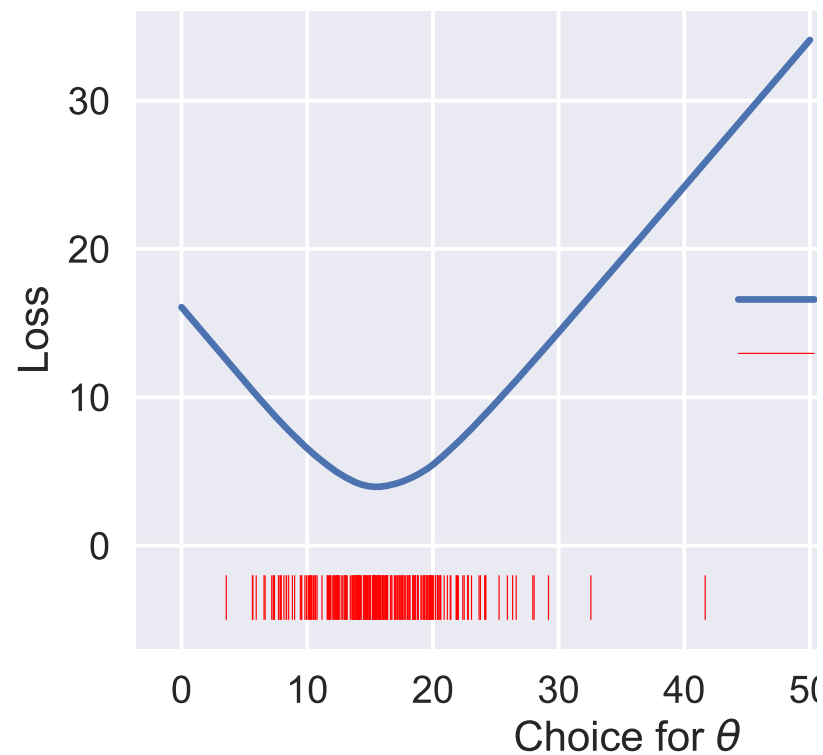
# Double Jeopardy

Name that Loss!

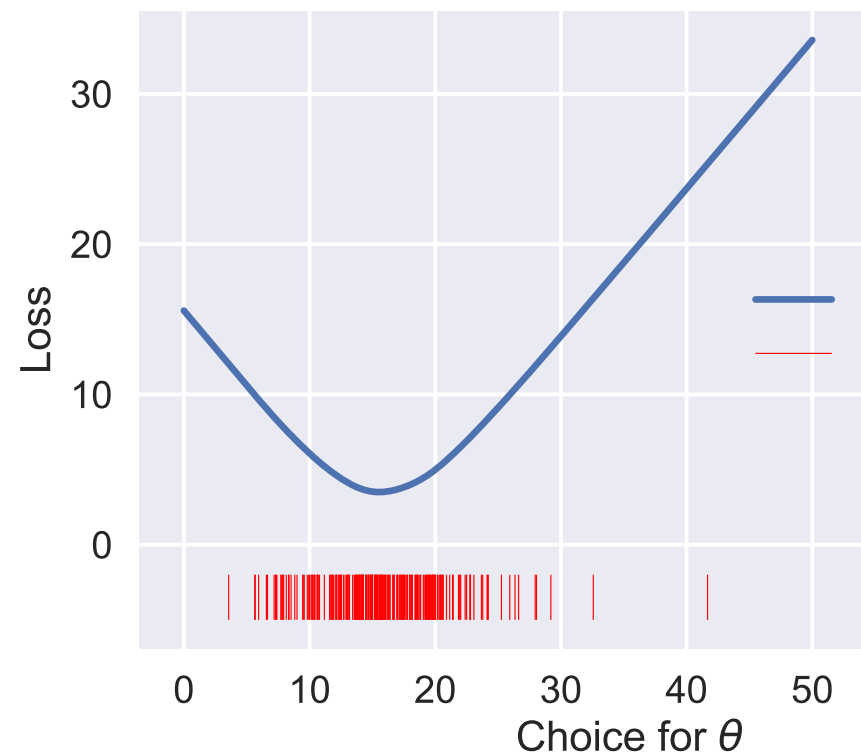
# Name that loss



(a)

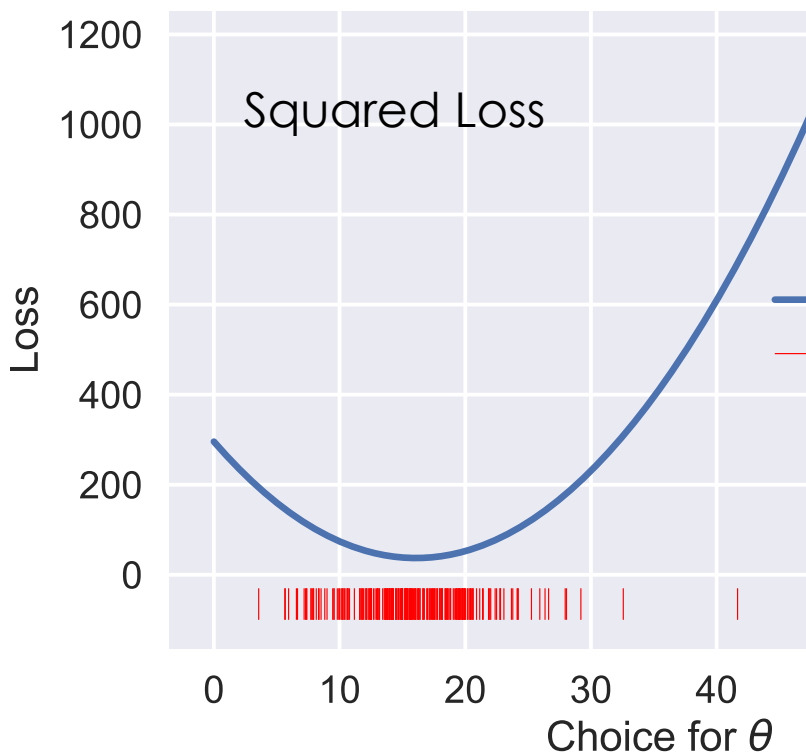


(b)

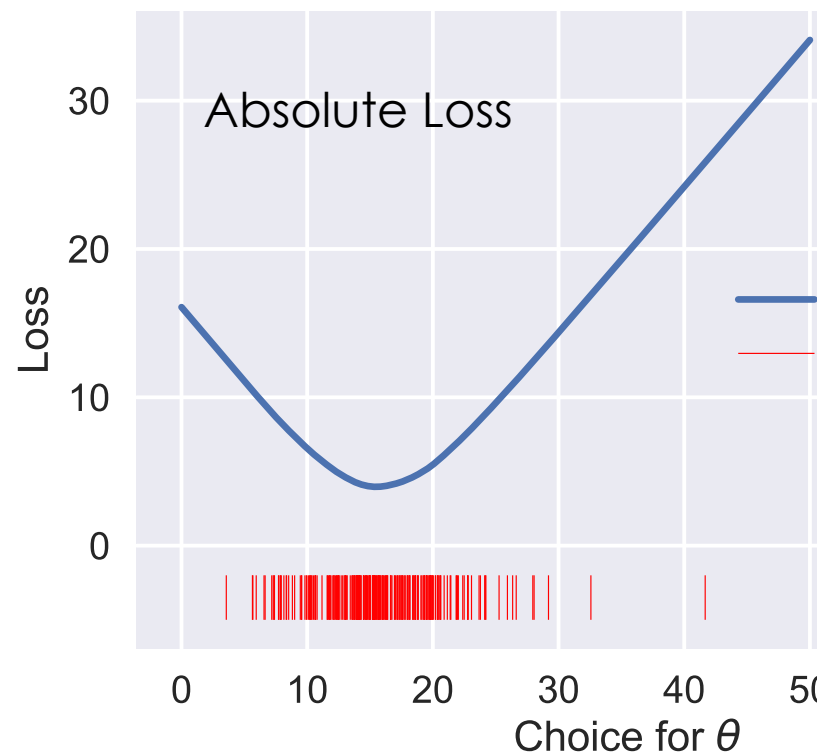


(c)

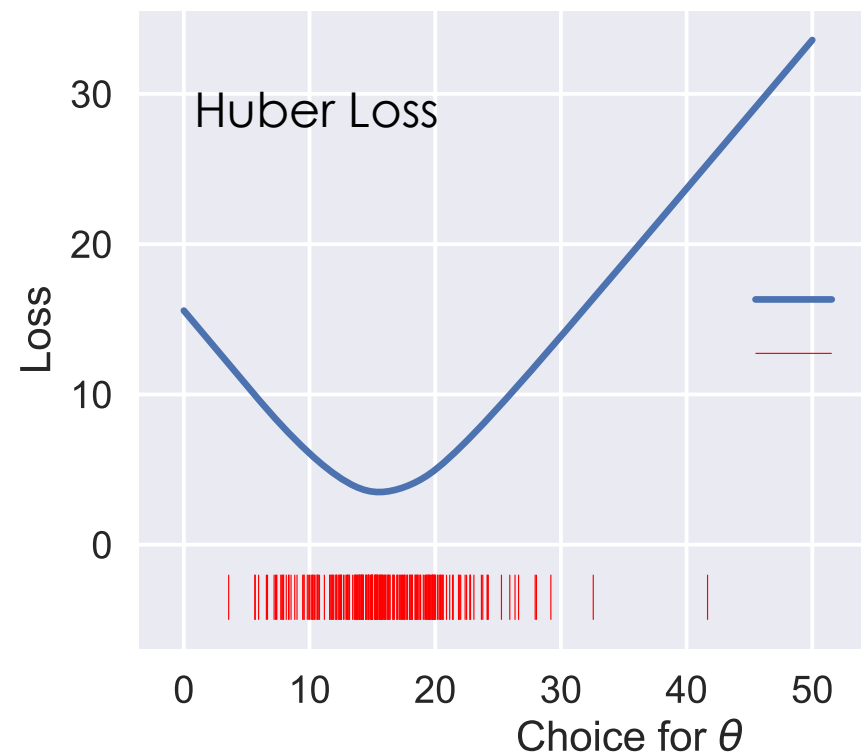
# Name that loss



(a)

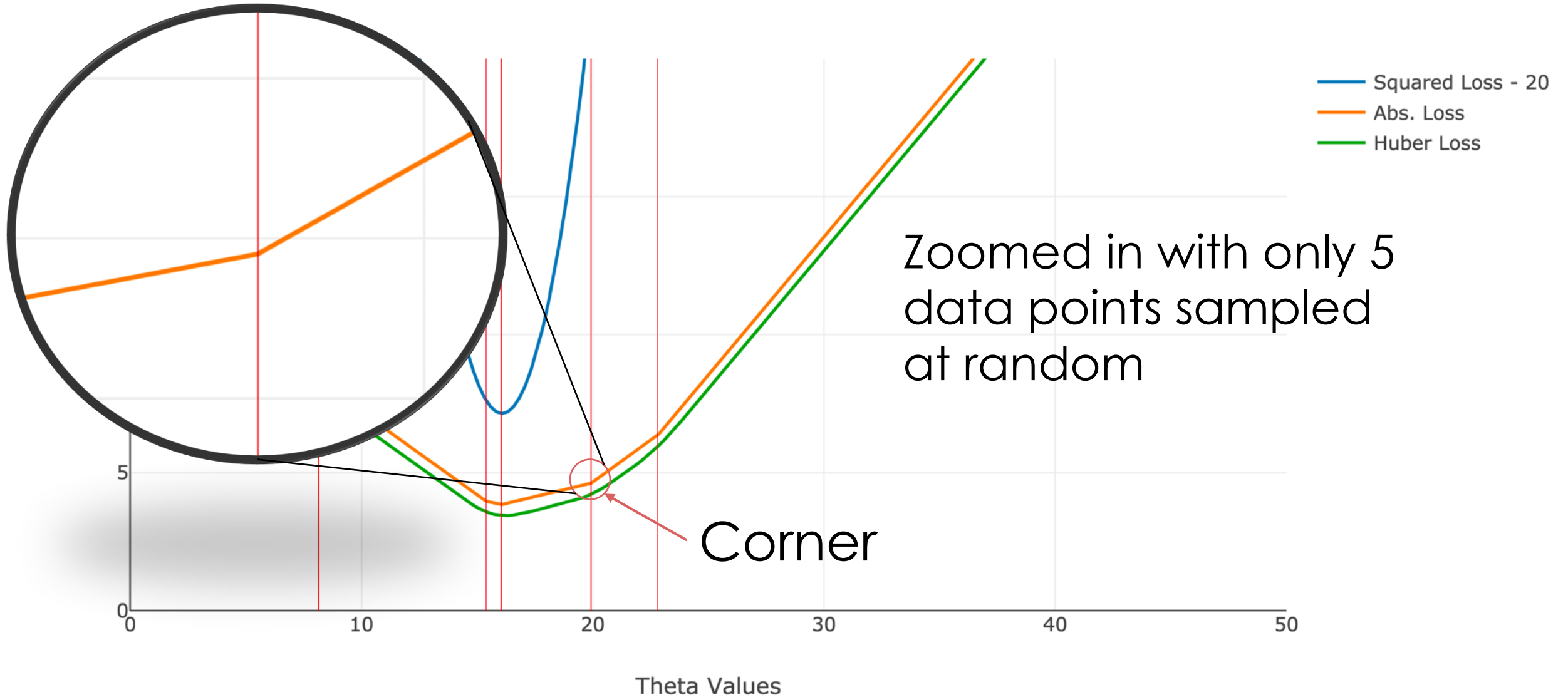


(b)

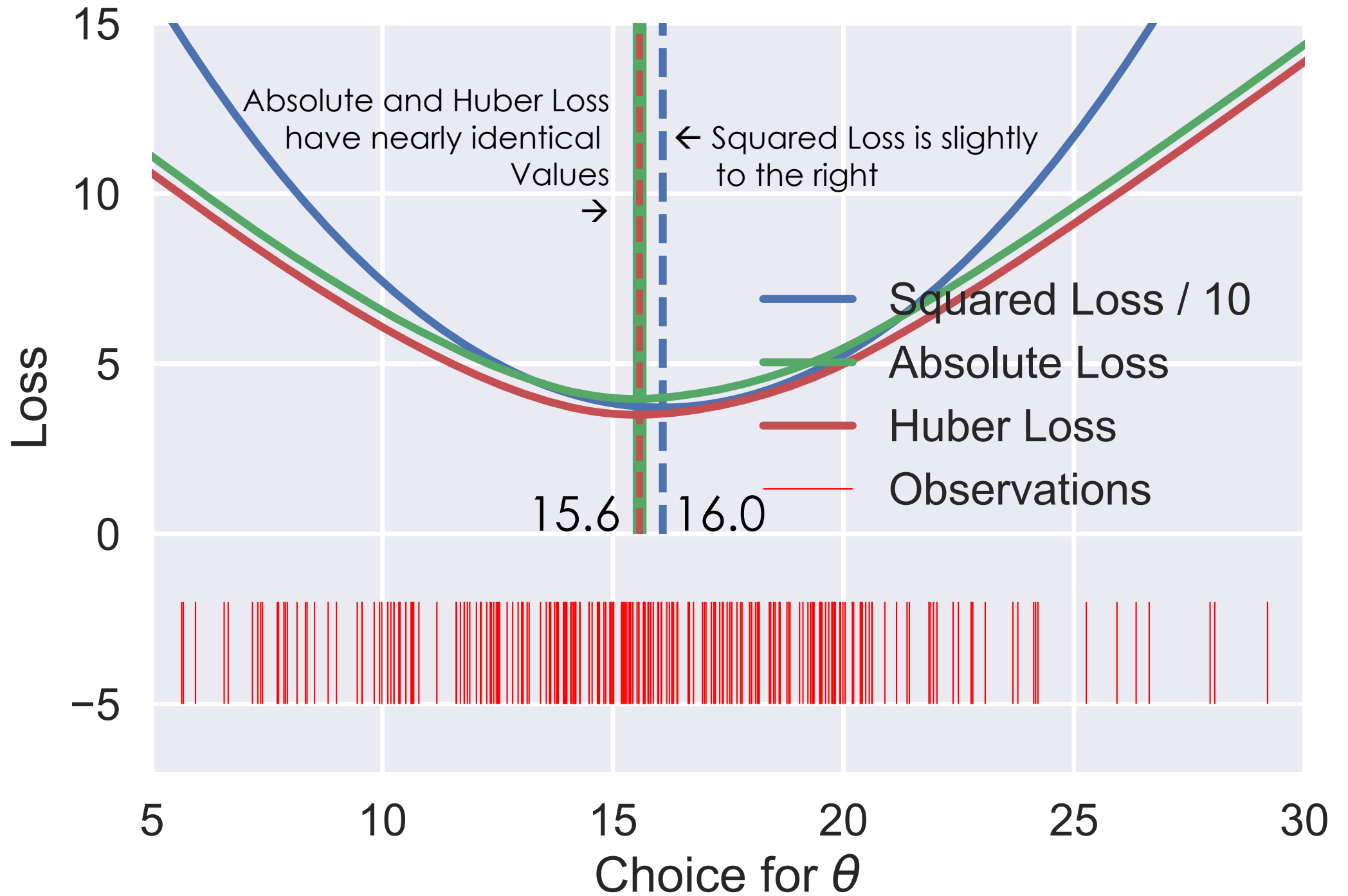


(c)

# Difference between Huber and L1

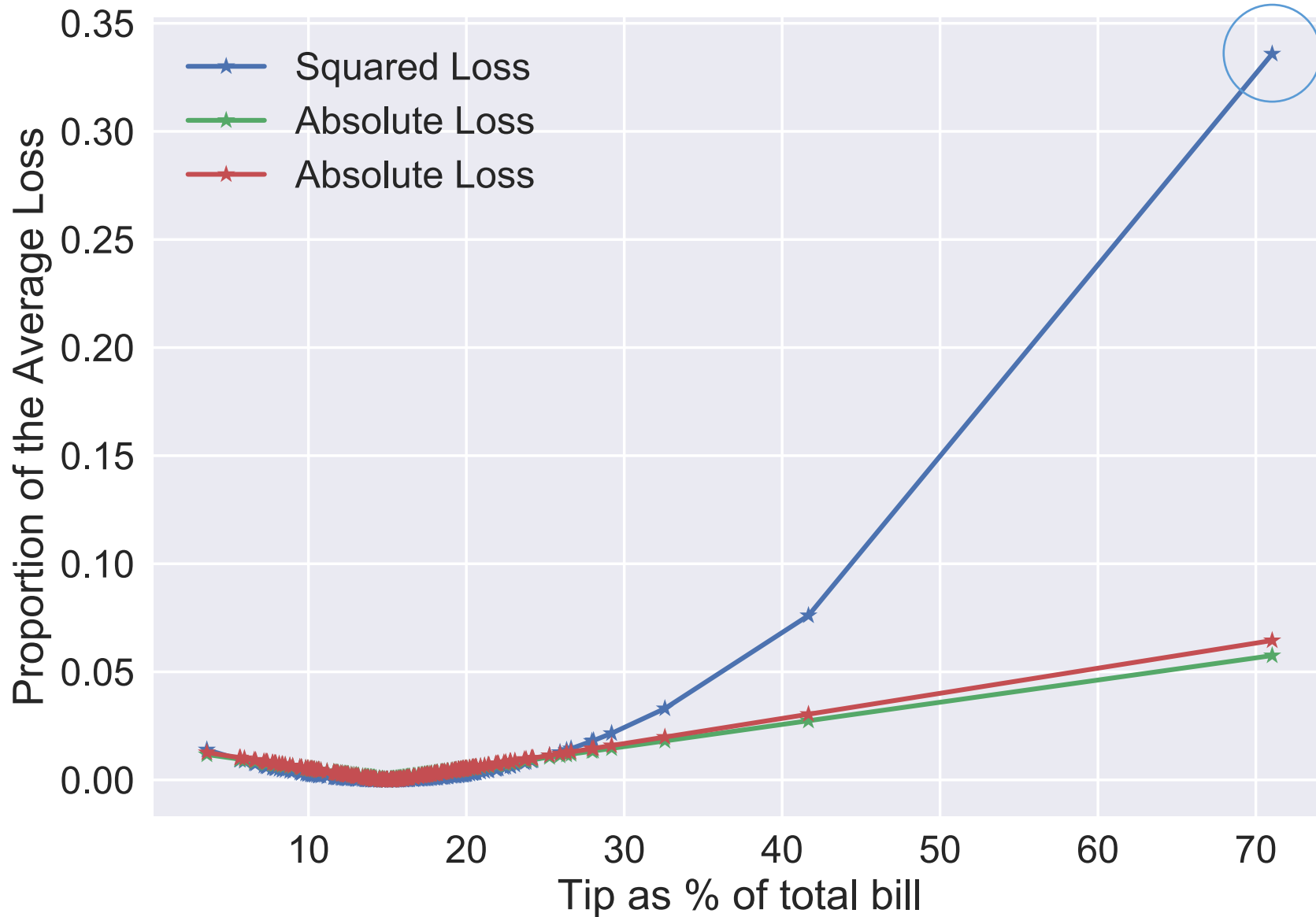


# Different Minimizers





# Sensitivity to Outliers



34% of loss due to a **single point**

Small fraction of loss on outliers...

# Recap on Loss Functions

- **Loss functions:** *a mechanism to measure how well a particular instance of a model fits a given dataset*
- **Squared Loss:** *sensitive to outliers but a smooth function*
- **Absolute Loss:** *less sensitive to outliers but not smooth*
- **Huber Loss:** *less sensitive to outliers and smooth but has an extra parameter to deal with*
- *Why is smoothness an issue → Optimization! ...*

# Summary of Model Estimation (so far...)

- 1. Define the Model:** simplified representation of the world
  - Use domain knowledge but ... **keep it simple!**
  - Introduce **parameters** for the unknown quantities
- 2. Define the Loss Function:** measures how well a particular instance of the model “fits” the data
  - We introduced  $L^2$ ,  $L^1$ , and Huber losses for each record
  - Take the average loss over the entire dataset
- 3. Minimize the Loss Function:** find the parameter values that minimize the loss on the data
  - So far we have done this graphically
  - Now we will **minimize the loss analytically**

Step 3: Minimize the Loss

# A Brief Review of Calculus

# Minimizing a Function

- Suppose we want to minimize:

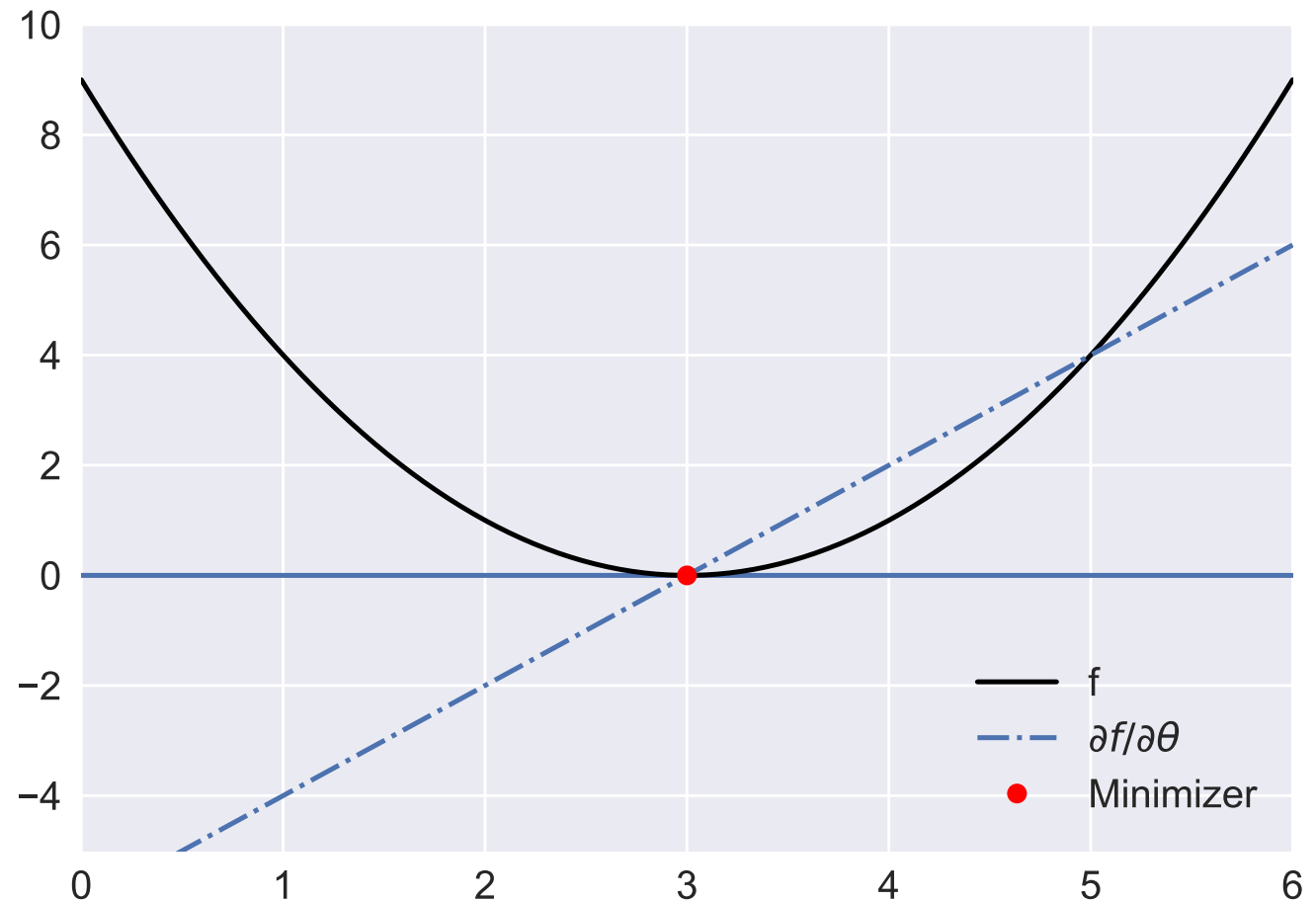
$$f(\theta) = (\theta - 3)^2$$

- Solve for derivative = 0:

$$\frac{\partial}{\partial \theta} f(\theta) = 2(\theta - 3) = 0$$

- Procedure:

1. take derivative
2. Set equal to zero
3. Solve for parameters



# Quick Review of the Chain Rule

- How do I compute the derivative of composed functions?

$$\begin{aligned}\frac{\partial}{\partial \theta} h(\theta) &= \frac{\partial}{\partial \theta} f(g(\theta)) \\ &= \left( \frac{\partial}{\partial u} f(u) \Big|_{u=g(\theta)} \right) \frac{\partial}{\partial \theta} g(\theta)\end{aligned}$$

Derivative of  $f$   
evaluated  
at  $g(\theta)$

Derivative  
of  $g(\theta)$

# Using the Chain Rule

$$\frac{\partial}{\partial \theta} \exp(\sin(\theta^2)) = \left( \frac{\partial}{\partial u} \exp(u) \Big|_{u=\sin(\theta^2)} \right) \frac{\partial}{\partial \theta} \sin(\theta^2)$$

First application of chain rule

$$\text{Derivative of exponent} = \left( \exp(u) \Big|_{u=\sin(\theta^2)} \right) \frac{\partial}{\partial \theta} \sin(\theta^2)$$

$$\text{Substituting } u = \exp(\sin(\theta^2)) \frac{\partial}{\partial \theta} \sin(\theta^2)$$

$$\text{Second application of the chain rule} = \exp(\sin(\theta^2)) \left( \frac{\partial}{\partial u} \sin(u) \Big|_{u=\theta^2} \right) \frac{\partial}{\partial \theta} \theta^2$$

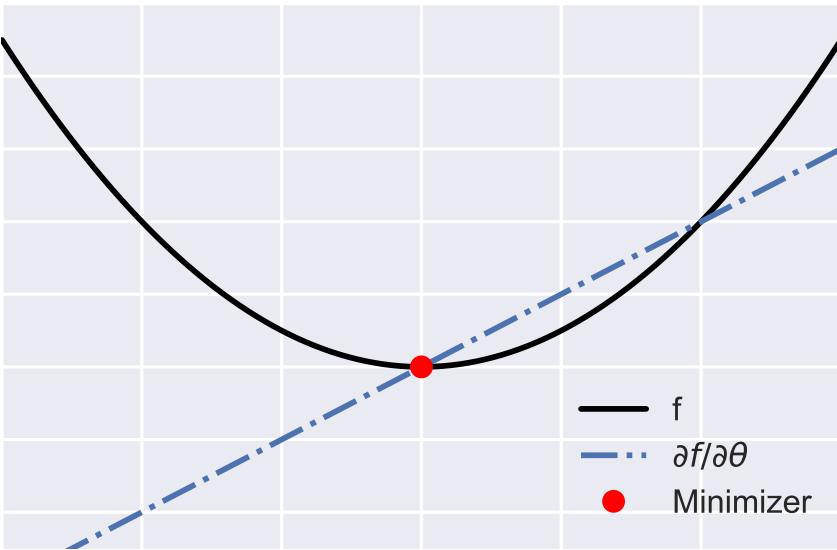
$$\text{Derivative of sine function} = \exp(\sin(\theta^2)) \left( \cos(u) \Big|_{u=\theta^2} \right) \frac{\partial}{\partial \theta} \theta^2$$

$$\text{Computing the remaining derivative} = \exp(\sin(\theta^2)) \cos(\theta^2) \frac{\partial}{\partial \theta} \theta^2$$

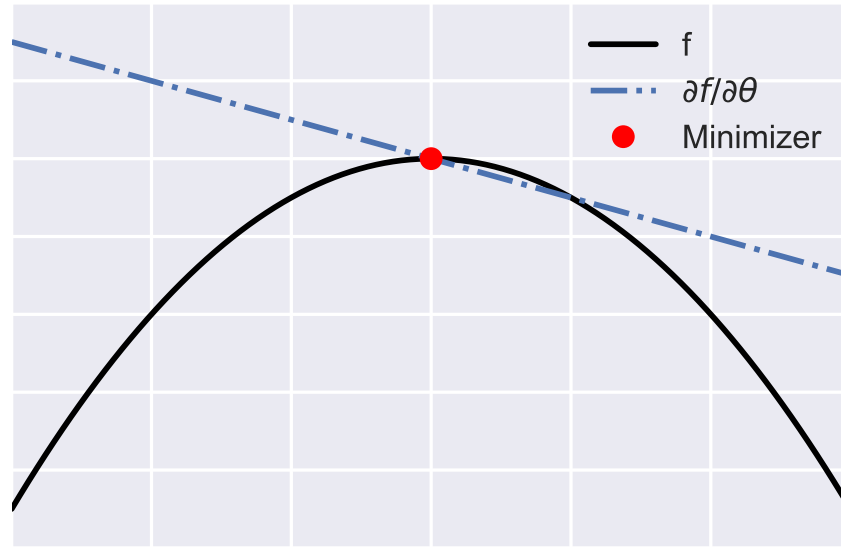
$$= \exp(\sin(\theta^2)) \cos(\theta^2) 2\theta$$



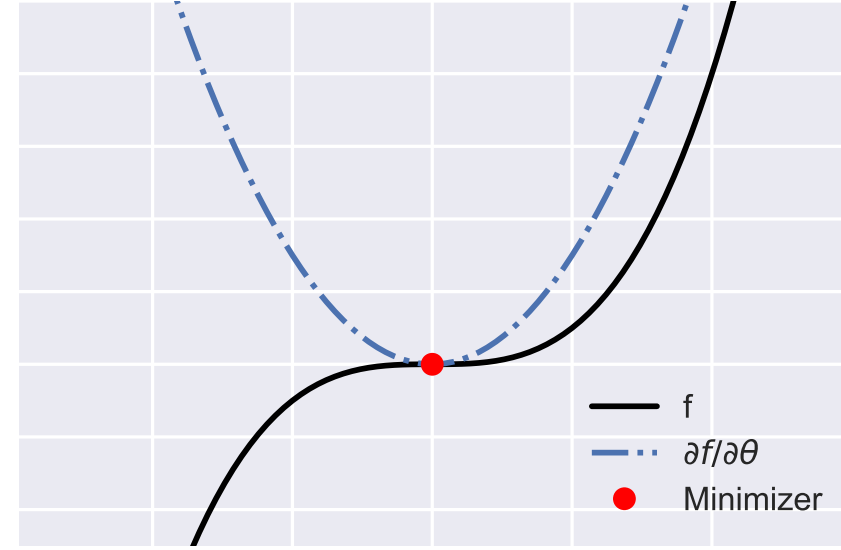
$$f(\theta) = (\theta - 3)^2$$



$$f(\theta) = -(\theta - 3)^2$$



$$f(\theta) = (\theta - 3)^3$$



$$\frac{\partial}{\partial \theta} f(\theta) = 2(\theta - 3)$$

$$\frac{\partial}{\partial \theta} f(\theta) = -2(\theta - 3)$$

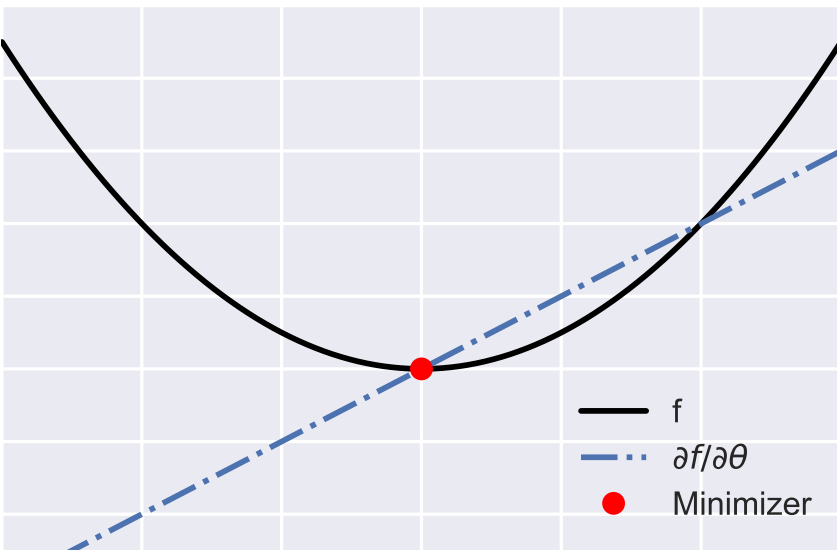
$$\frac{\partial}{\partial \theta} f(\theta) = 3(\theta - 3)^2$$

All of the above functions have zero derivatives at  $\theta = 3$   
 $\rightarrow$  is  $\theta=3$  minimizer for all the above functions?

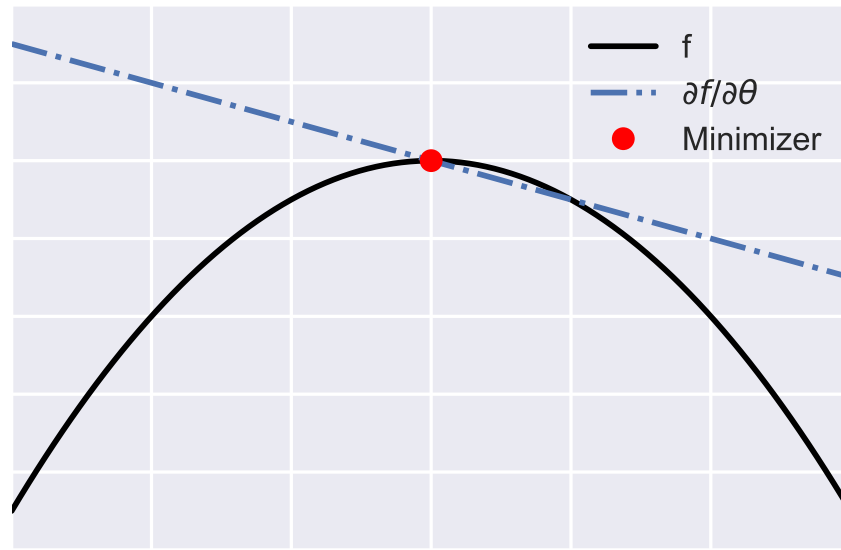
No!

Need to check second derivative is positive...

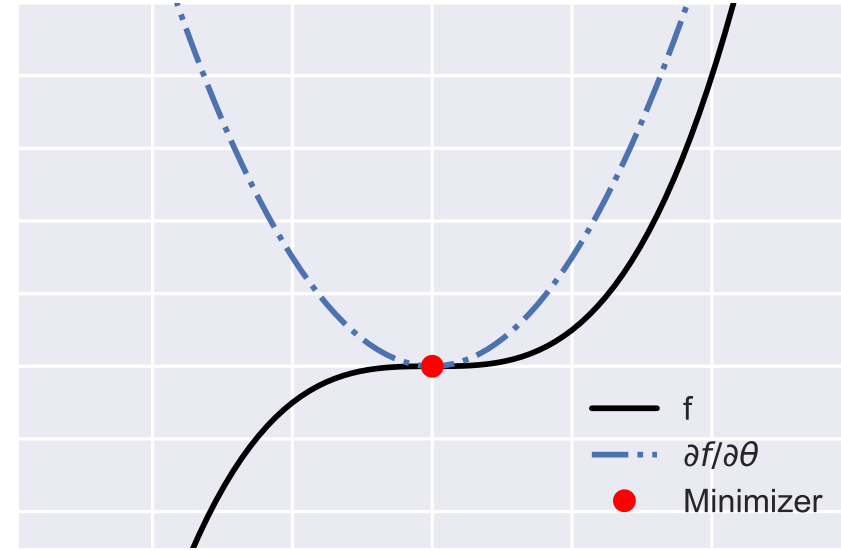
$$\frac{\partial}{\partial \theta} f(\theta) = 2(\theta - 3)$$



$$\frac{\partial}{\partial \theta} f(\theta) = -2(\theta - 3)$$



$$\frac{\partial}{\partial \theta} f(\theta) = 3(\theta - 3)^2$$



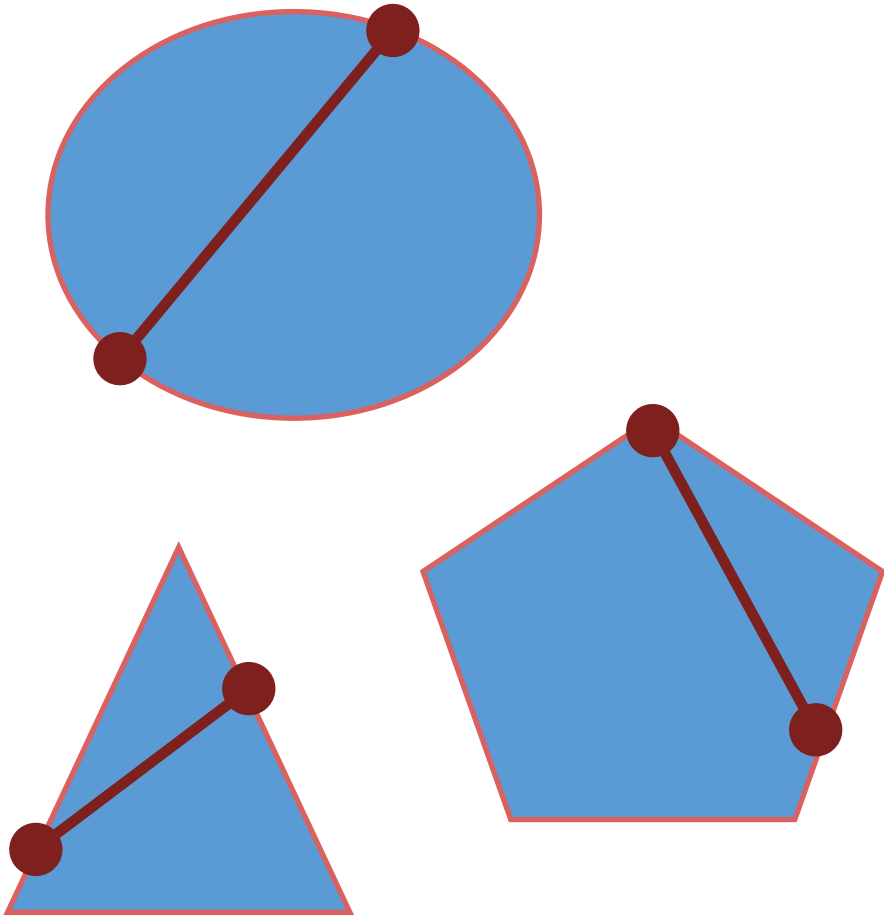
$$\frac{\partial^2}{\partial \theta^2} f(\theta) = 2$$

$$\frac{\partial^2}{\partial \theta^2} f(\theta) = -2$$

$$\frac{\partial^2}{\partial \theta^2} f(\theta) = 6(\theta - 3)$$

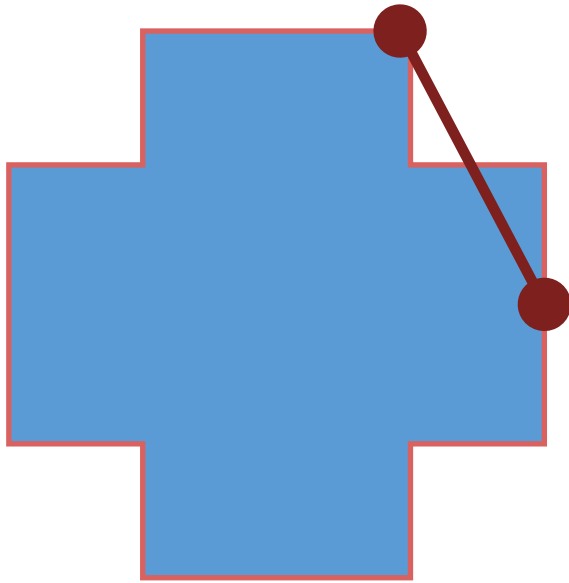
Generally we are interested in **convex functions** with respect to the parameters  $\theta$ .

# Convex sets and polygons

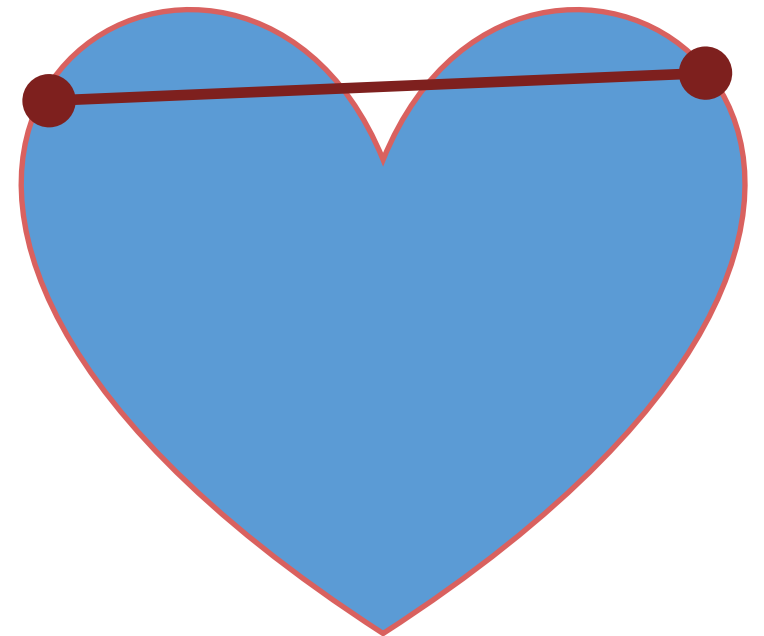
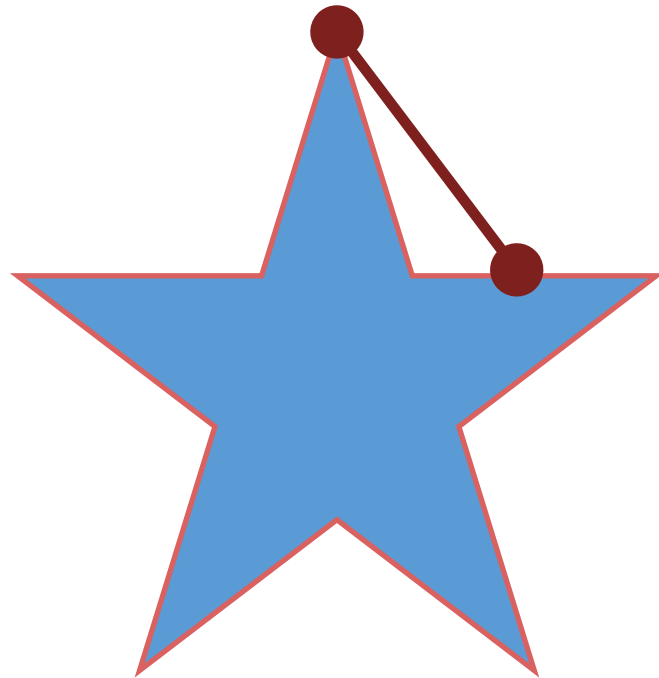


- No line segment between any two points on the boundary ever leaves the polygon.
- Equivalently, all angles are  $\leq 180^\circ$ .
- The interior is a convex set.

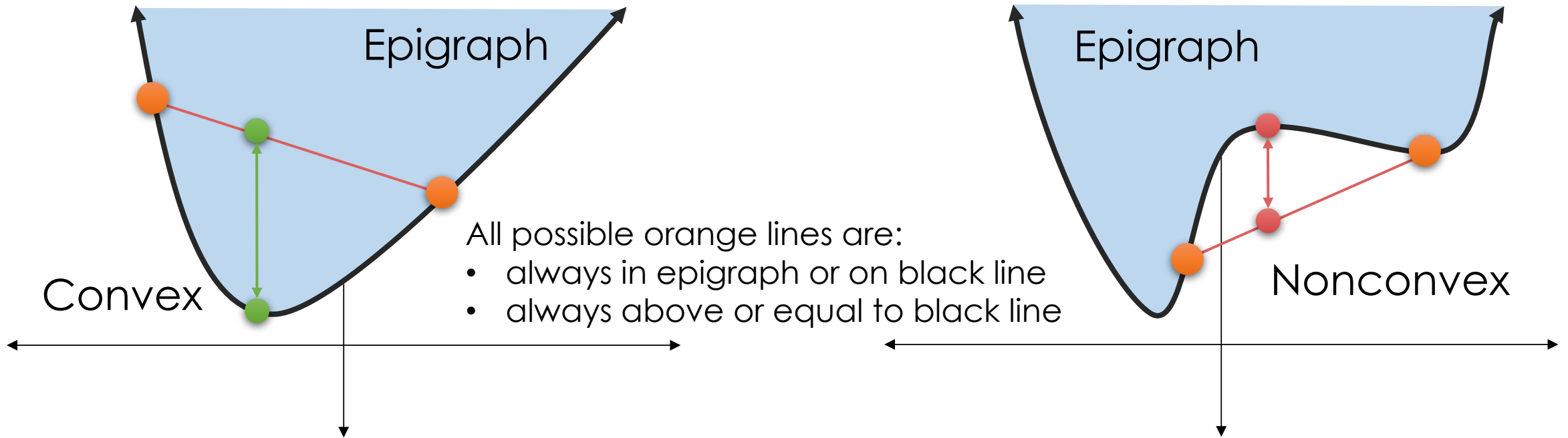
# Non-Convex sets and polygons



- There is at least one line segment between two points on the boundary that leaves the set.



# Formal Definition of Convex Functions

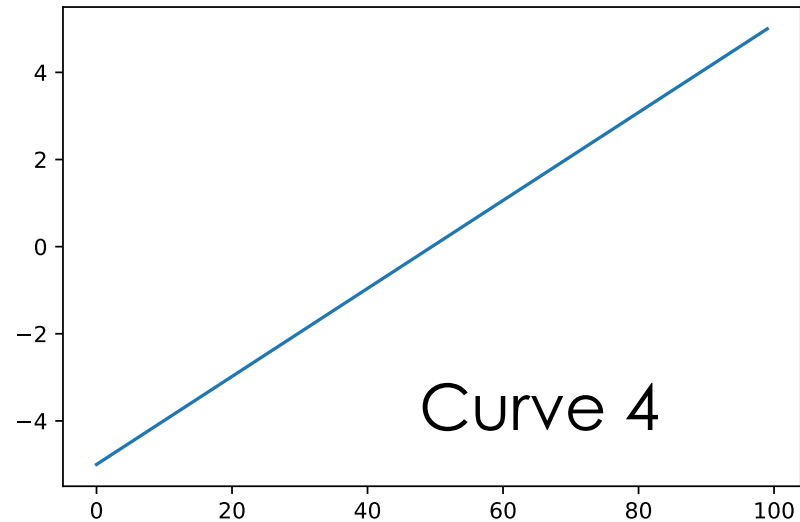
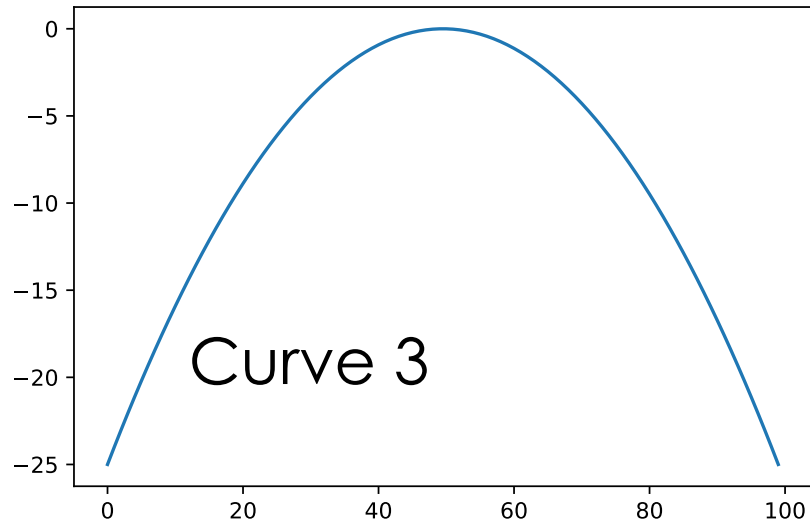
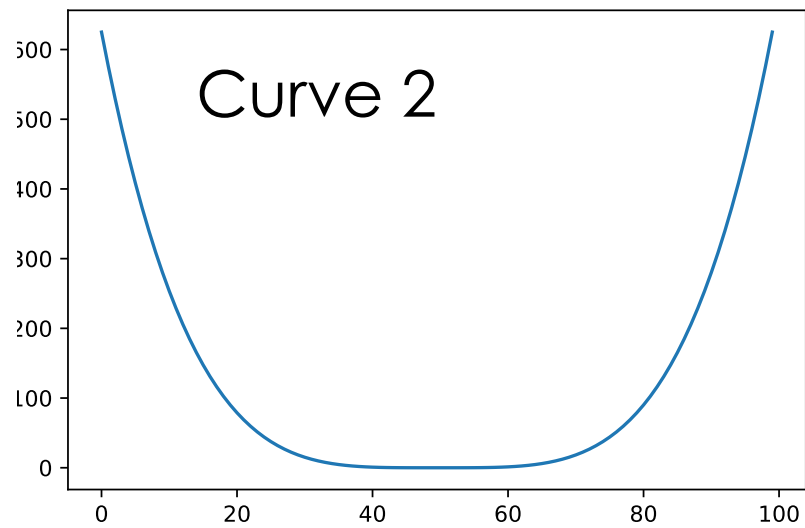
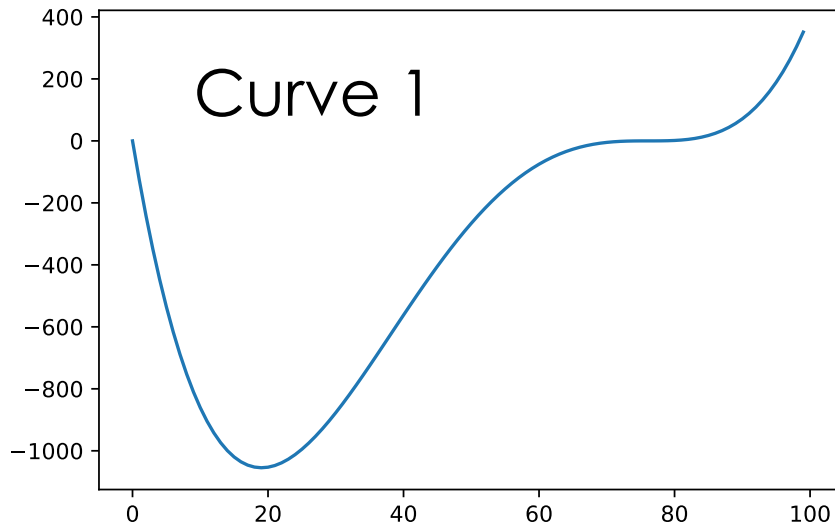


➤ A function  $f$  is convex if and only if:

$$tf(a) + (1 - t)f(b) \geq f(ta + (1 - t)b)$$

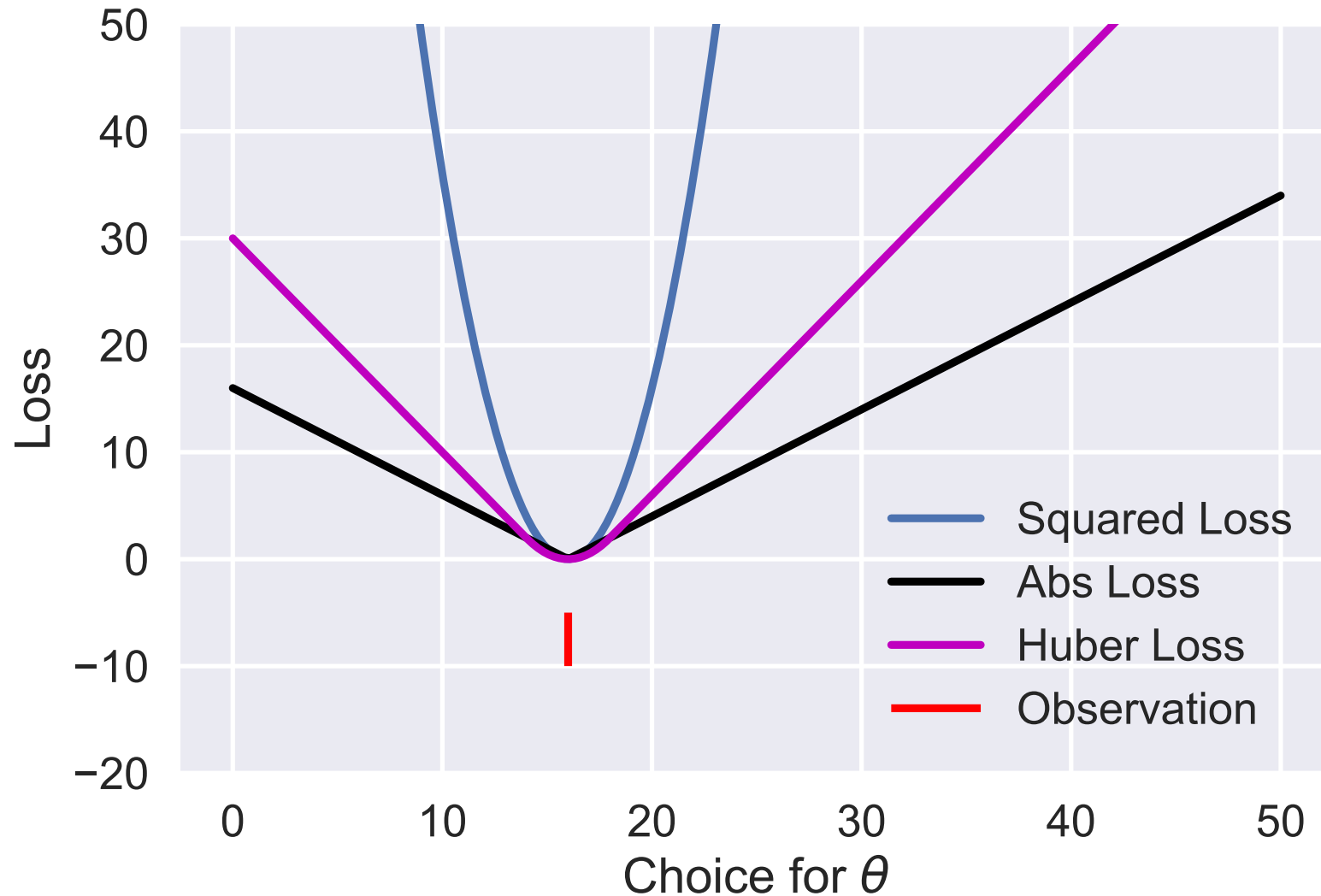
$$\forall a, \forall b, t \in [0, 1]$$

<http://bit.ly/ds100-sp18-cvx>



Convex or  
Not Convex

# Are our previous loss functions convex?



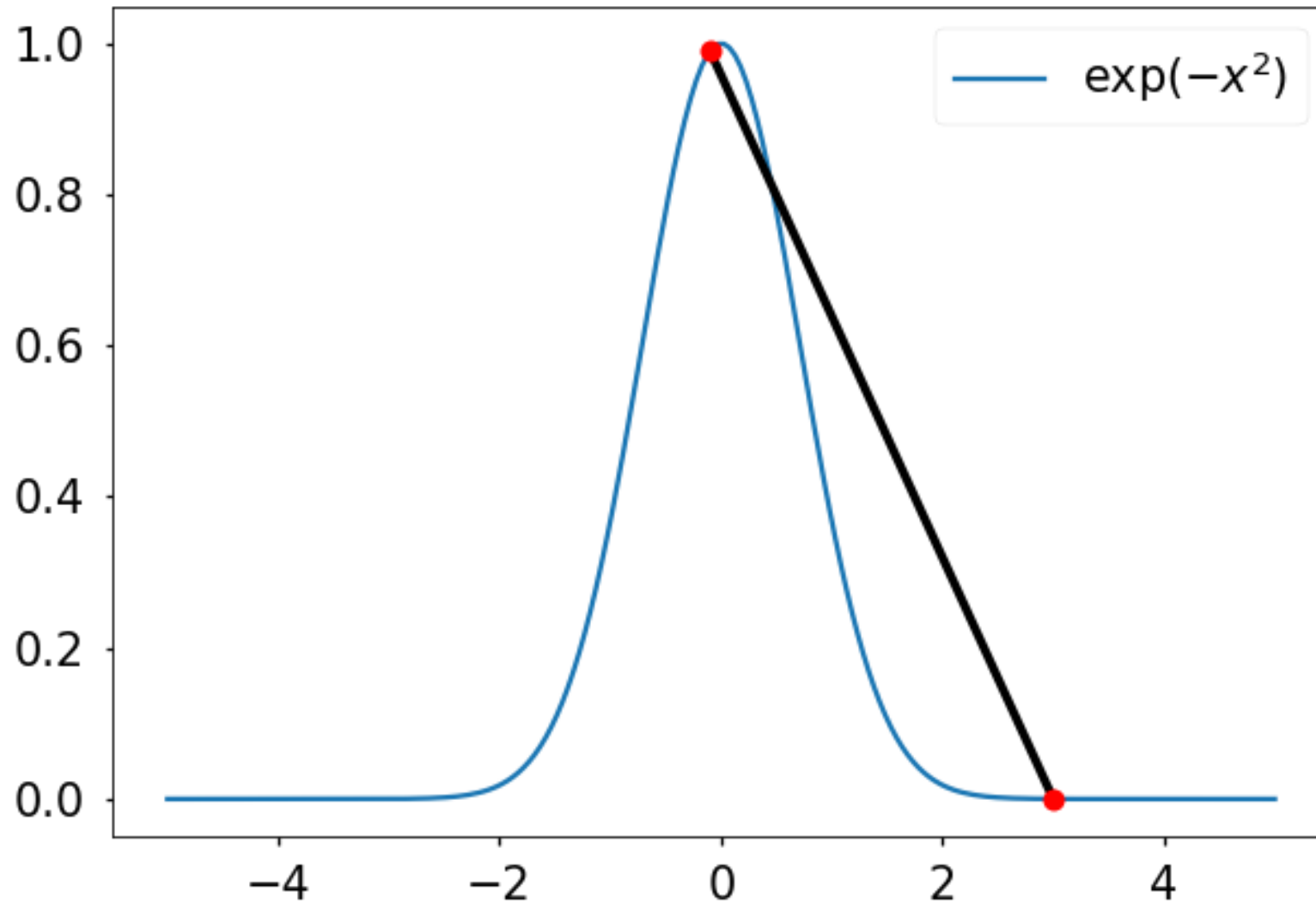
Yes!

Average Loss?

Yes!

(Sum of convex functions is convex)

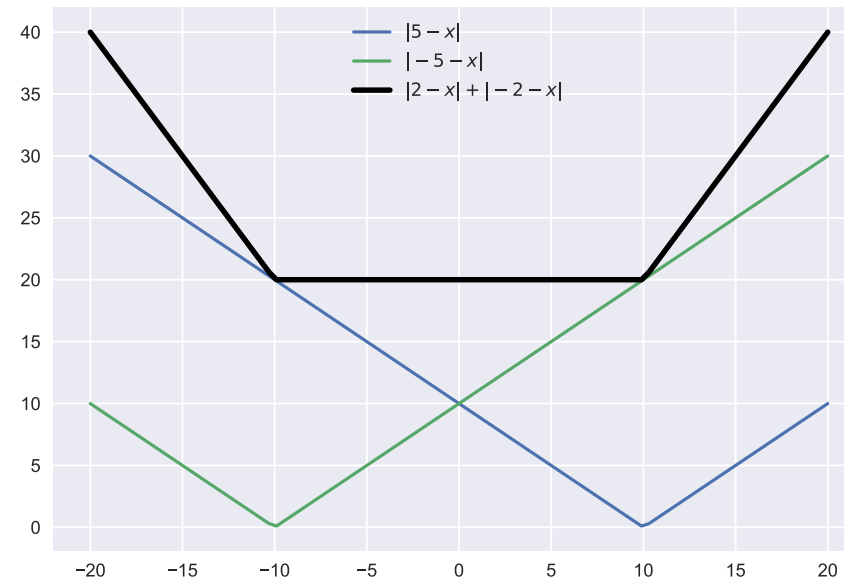
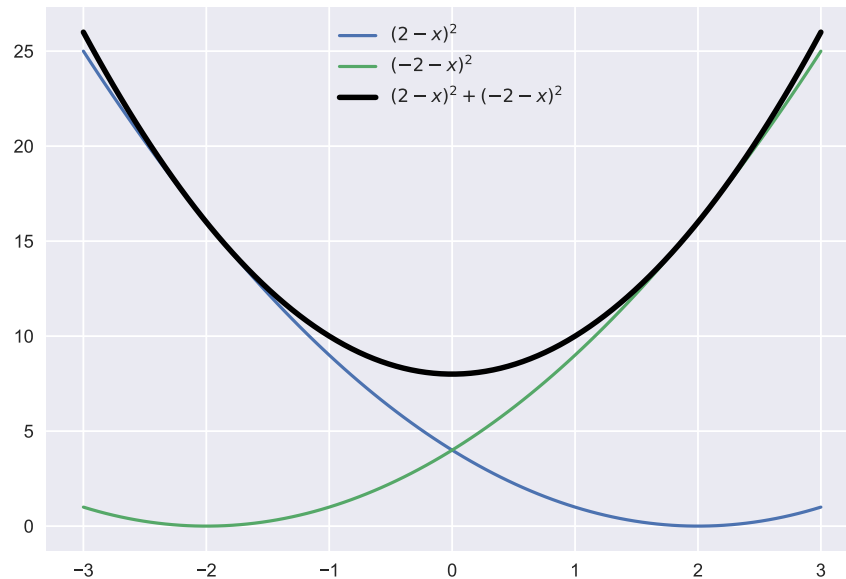
# Is a Gaussian convex?





# Sum of Convex Functions is Convex

- In class professor Gonzalez was asked: “Are you sure that the sum of convex functions is convex?”
  - The answer is yes! Always!
    - Professor Gonzalez should have had a proof ready! ☹ It's Easy!
- Proposed counter examples ... (not entirely obvious)



# Formal Proof

- Suppose you have two convex functions  $f$  and  $g$ :

$$tf(a) + (1 - t)f(b) \geq f(ta - (1 - t)a)$$

$$tg(a) + (1 - t)g(b) \geq g(ta - (1 - t)a)$$

$$\forall a, \forall b, t \in [0, 1]$$

- We would like to show:

$$th(a) + (1 - t)h(b) \geq h(ta - (1 - t)a)$$

- Where:  $h(x) = f(x) + g(x)$

➤ We would like to show:

$$th(a) + (1 - t)h(b) \geq h(ta + (1 - t)b)$$

➤ Where:  $h(x) = f(x) + g(x)$

➤ Starting on the left side

Substituting definition of h:

$$th(a) + (1 - t)h(b) = t(f(a) + g(a)) + (1 - t)(f(b) + g(b))$$

Re-arranging terms:  $= [tf(a) + (1 - t)f(b)] + [tg(a) + (1 - t)g(b)]$

Convexity in  $f \geq f(ta + (1 - t)b) + [tg(a) + (1 - t)g(b)]$

Convexity in  $g \geq f(ta + (1 - t)b) + g(ta + (1 - t)b)$

Definition of h  $= h(ta + (1 - t)b)$



# Minimizing the Average Squared Loss

# Minimizing the Average Squared Loss

$$L_{\mathcal{D}}(\theta) = \frac{1}{n} \sum_{i=1}^n (y_i - \theta)^2 \rightarrow \frac{\partial}{\partial \theta} L_{\mathcal{D}}(\theta) = \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta} (y_i - \theta)^2$$

➤ Take the derivative

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

# Minimizing the Average Squared Loss

$$L_{\mathcal{D}}(\theta) = \frac{1}{n} \sum_{i=1}^n (y_i - \theta)^2 \rightarrow \frac{\partial}{\partial \theta} L_{\mathcal{D}}(\theta) = \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta} (y_i - \theta)^2$$

- Take the derivative
- Set the derivative equal to zero

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

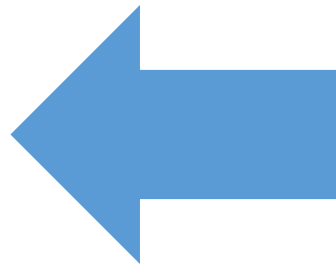
$$0 = -\frac{2}{n} \sum_{i=1}^n (y_i - \theta)$$

# Minimizing the Average Squared Loss

- Take the derivative
- Set the derivative equal to zero
- Solve for parameters

Hat  
(Estimator)

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n y_i$$




$$0 = -\frac{2}{n} \sum_{i=1}^n (y_i - \theta)$$

$$0 = \sum_{i=1}^n (y_i - \theta)$$

$$0 = \left( \sum_{i=1}^n y_i \right) - \left( \sum_{i=1}^n \theta \right)$$

$$0 = \left( \sum_{i=1}^n y_i \right) - n\theta$$

# Minimizing the Average Squared Loss

Hat  
(Estimator) 

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n y_i$$

**Mean  
(Average)!**

- The estimate for percent tip that minimizes the squared loss is the mean (average) of the percent tips
  - We guessed that already!



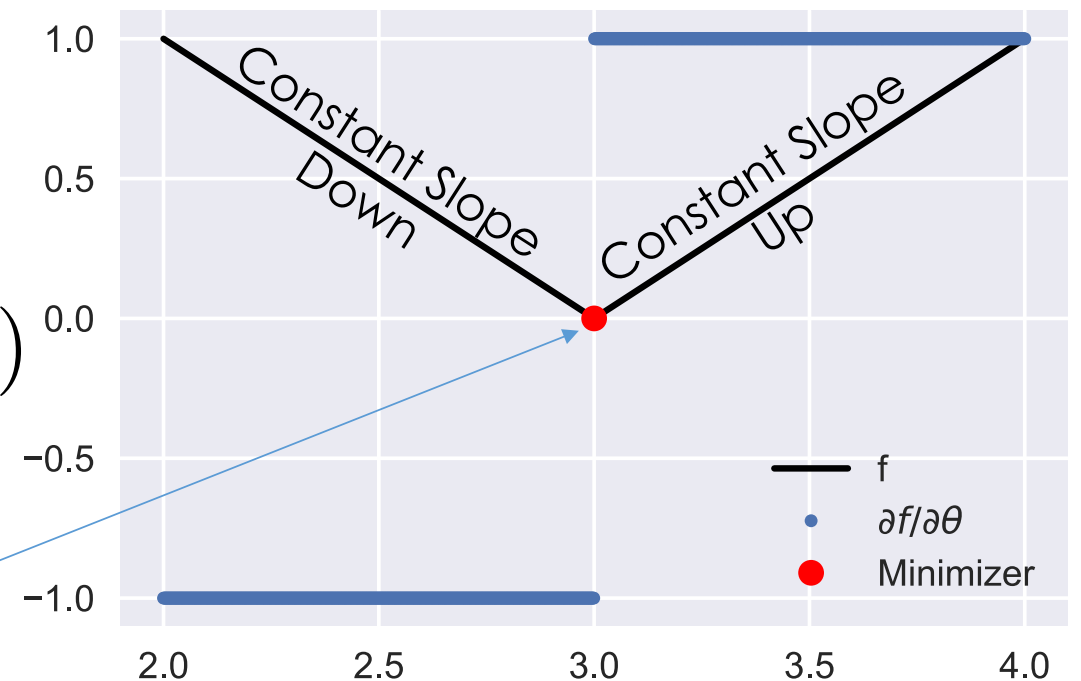
# Minimizing the Average Absolute Loss

$$L_{\mathcal{D}}(\theta) = \frac{1}{n} \sum_{i=1}^n |y_i - \theta| \quad \rightarrow \quad \frac{\partial}{\partial \theta} L_{\mathcal{D}}(\theta) = \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta} |y_i - \theta|$$

- Take the derivative
  - How?

$$\frac{\partial}{\partial \theta} L_{\mathcal{D}}(\theta) = -\frac{1}{n} \sum_{i=1}^n \mathbf{sign}(y_i - \theta)$$

What is  $\mathbf{sign}(0)$  ?



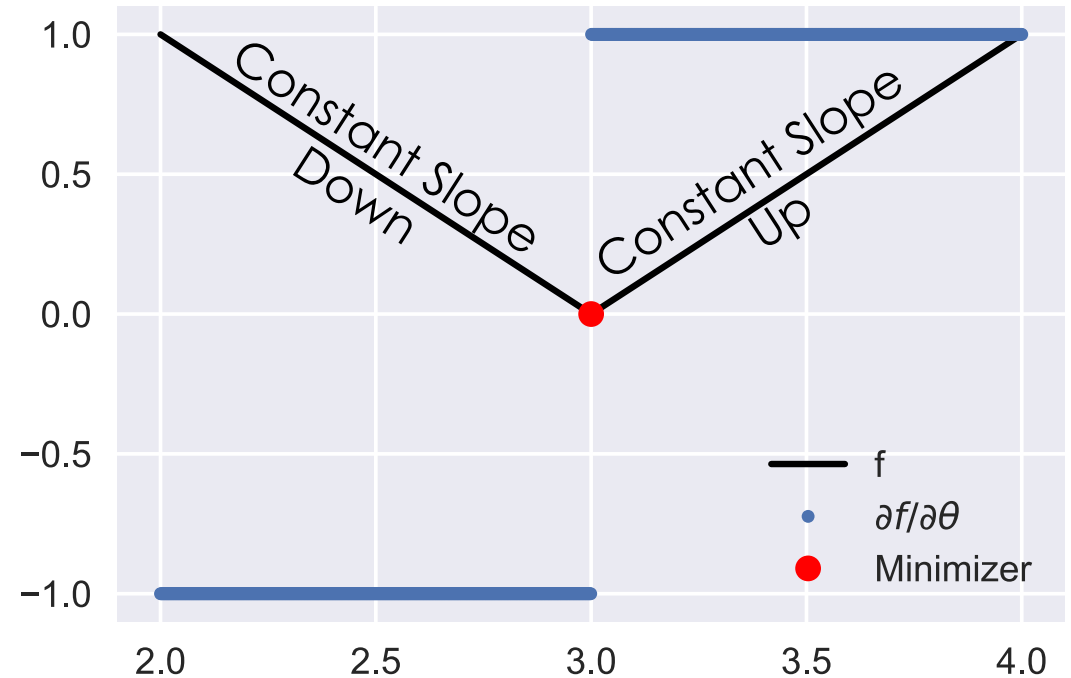
# Minimizing the Average Absolute Loss

- Take the derivative
  - How?

$$\frac{\partial}{\partial \theta} L_{\mathcal{D}}(\theta) = -\frac{1}{n} \sum_{i=1}^n \mathbf{sign}(y_i - \theta)$$

- Derivative at the corner?
  - What is the sign of 0?
- Convention:

$$\mathbf{sign}(0) = 0$$



# Minimizing the Average Absolute Loss

- Take the derivative
- Set derivative to zero and solve for parameters

$$\frac{\partial}{\partial \theta} L_{\mathcal{D}}(\theta) = -\frac{1}{n} \sum_{i=1}^n \mathbf{sign}(y_i - \theta)$$

$$= -\frac{1}{n} \left( \sum_{y_i < \theta} -1 + \sum_{y_i > \theta} +1 \right)$$

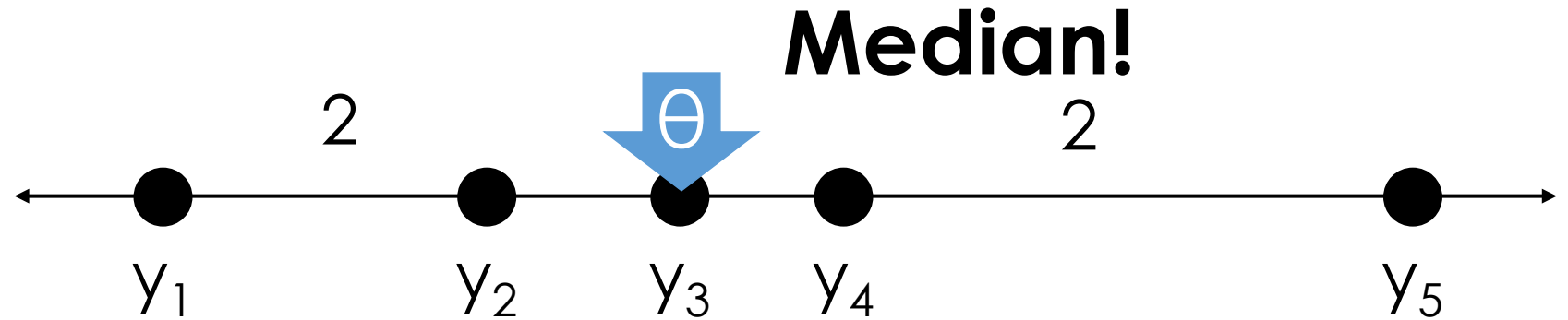
$$\left( \sum_{y_i < \theta}^n 1 \right) = \left( \sum_{y_i > \theta} 1 \right) \quad \leftarrow \quad 0 = \left( \sum_{y_i < \theta}^n -1 \right) + \left( \sum_{y_i > \theta} +1 \right)$$

# Minimizing the Average Absolute Loss

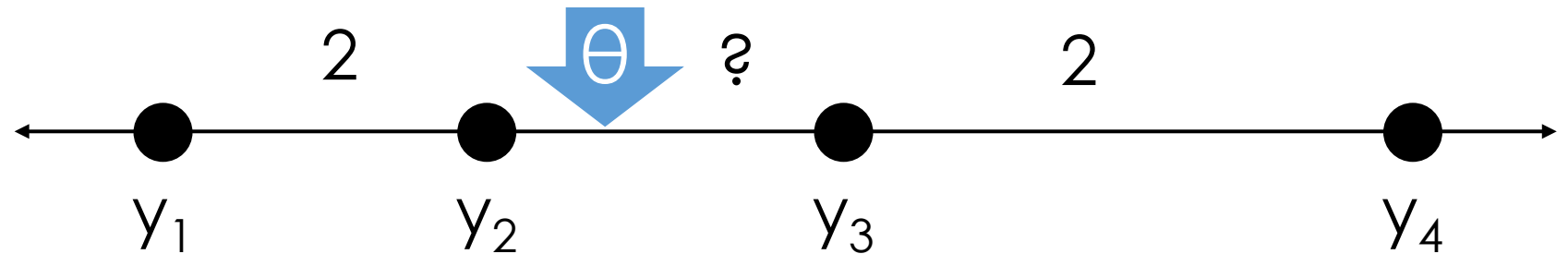
- Take the derivative
- Set derivative to zero and solve for parameters

$$\left( \sum_{y_i < \theta}^n 1 \right) = \left( \sum_{y_i > \theta} 1 \right)$$

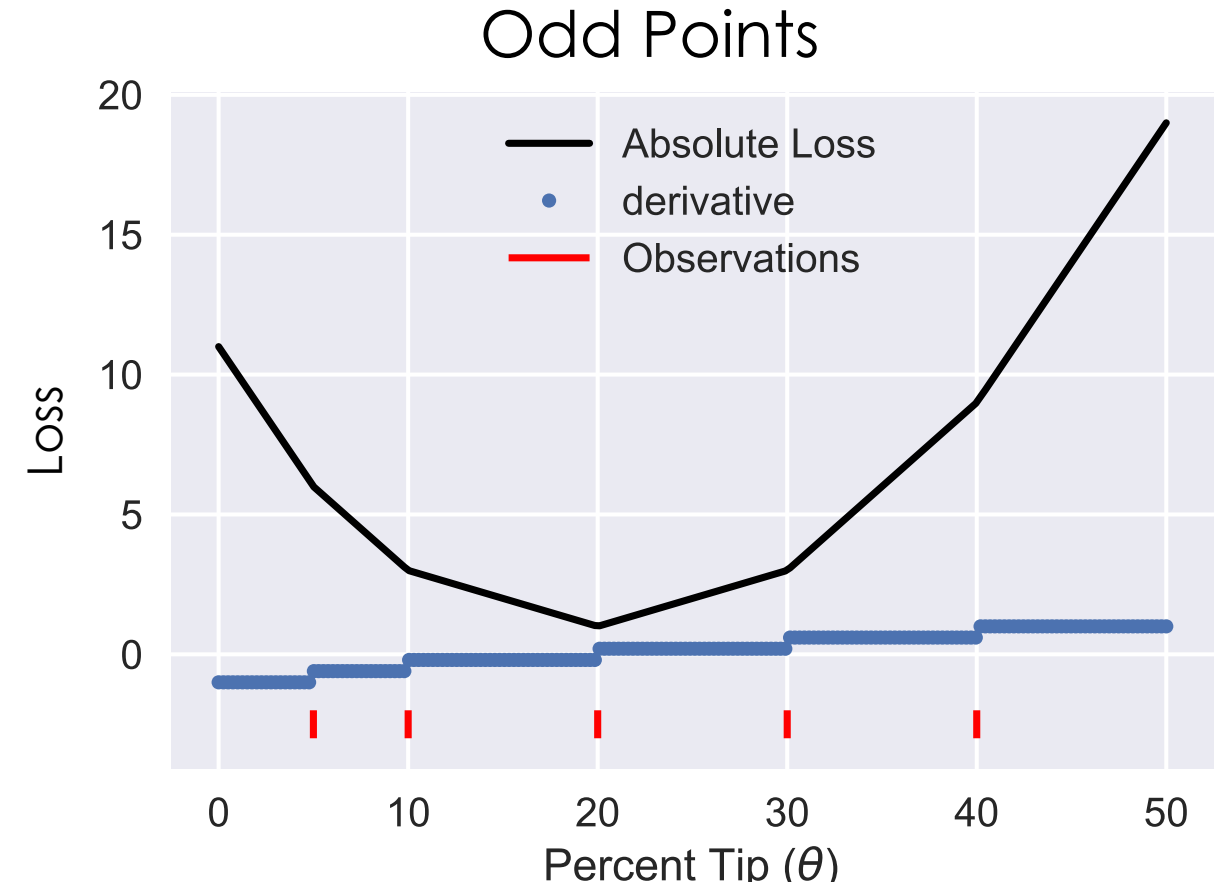
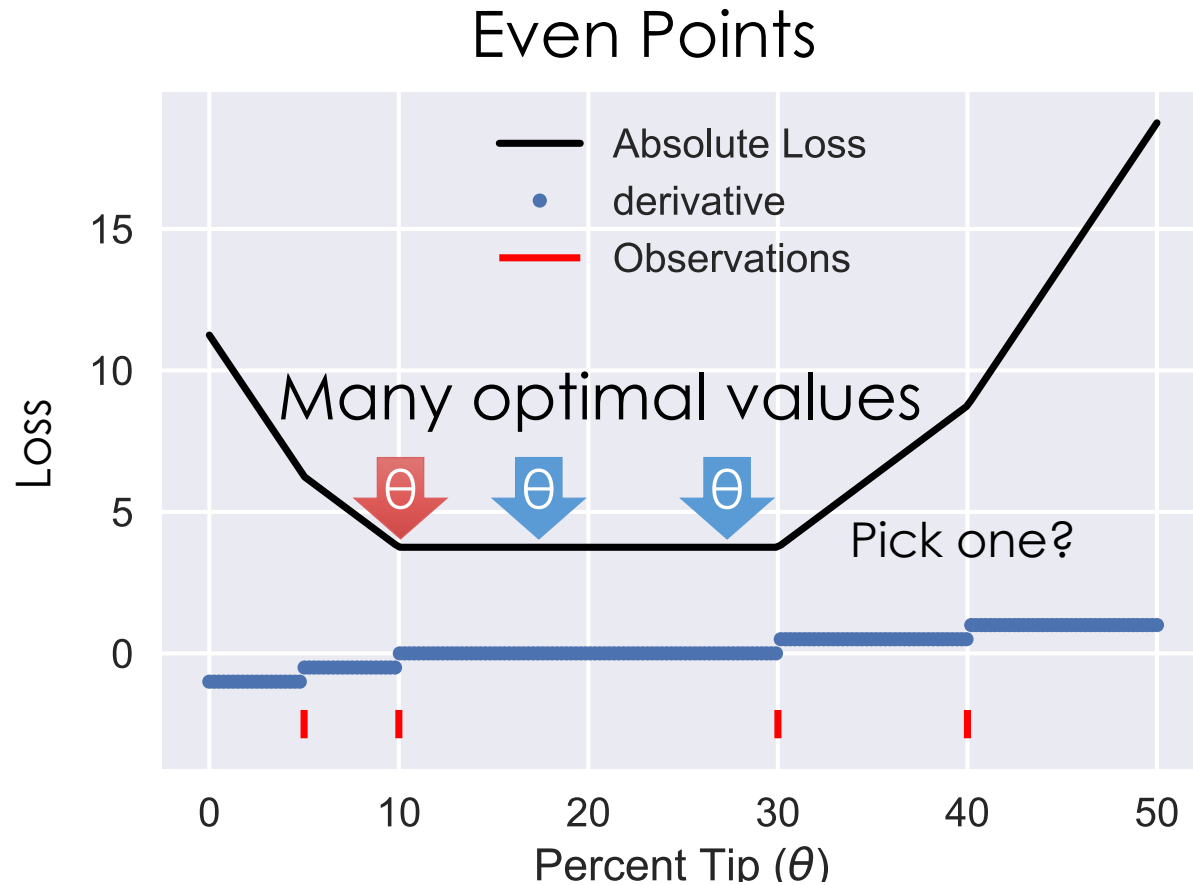
Percent Tips in sorted order



Percent Tips in sorted order



# Absolute Loss Even and Odd Data



The **median** minimizes the absolute loss → Robust!  
not sensitive to outliers

# Calculus for Loss Minimization

- General Procedure:
  - Verify that function is convex (we often will assume this...)
  - Compute the derivative
  - Set derivative equal to zero and solve for the parameters
- Using this procedure we discovered:

$$\hat{\theta}_{L^2} = \frac{1}{n} \sum_{I=1}^n y_i = \mathbf{mean}(\mathcal{D}) \quad \hat{\theta}_{L^1} = \mathbf{median}(\mathcal{D})$$

$$\hat{\theta}_{\text{Huber}} = ?$$

# Minimizing the Average Huber Loss

$$L_{\alpha}(\theta, y) = \begin{cases} \frac{1}{2} (y - \theta)^2 & |y - \theta| < \alpha \\ \alpha \left( |y - \theta| - \frac{\alpha}{2} \right) & \text{otherwise} \end{cases}$$

➤ Take the derivative of the average Huber Loss

$$\frac{\partial}{\partial \theta} L_{\mathcal{D}}(\theta) = \frac{1}{n} \sum_{i=1}^n \begin{cases} -(y_i - \theta) & |y_i - \theta| < \alpha \\ -\alpha \operatorname{sign}(y_i - \theta) & \text{otherwise} \end{cases}$$

# Minimizing the Average Huber Loss

$$L_{\alpha}(\theta, y) = \begin{cases} \frac{1}{2} (y - \theta)^2 & |y - \theta| < \alpha \\ \alpha \left( |y - \theta| - \frac{\alpha}{2} \right) & \text{otherwise} \end{cases}$$

➤ Take the derivative of the average Huber Loss

$$\frac{\partial}{\partial \theta} L_{\mathcal{D}}(\theta) = \frac{1}{n} \sum_{i=1}^n \begin{cases} -(y_i - \theta) & |y_i - \theta| < \alpha \\ -\alpha \operatorname{sign}(y_i - \theta) & \text{otherwise} \end{cases}$$



# Minimizing the Average Huber Loss

$$L_{\alpha}(\theta, y) = \begin{cases} \frac{1}{2} (y - \theta)^2 & |y - \theta| < \alpha \\ \alpha \left( |y - \theta| - \frac{\alpha}{2} \right) & \text{otherwise} \end{cases}$$

- Take the derivative of the average Huber Loss

$$\frac{\partial}{\partial \theta} L_{\mathcal{D}}(\theta) = \frac{1}{n} \sum_{i=1}^n \begin{cases} -(y_i - \theta) & |y_i - \theta| < \alpha \\ -\alpha \operatorname{sign}(y_i - \theta) & \text{otherwise} \end{cases}$$

# Minimizing the Average Huber Loss

$$L_{\alpha}(\theta, y) = \begin{cases} \frac{1}{2} (y - \theta)^2 & |y - \theta| < \alpha \\ \alpha \left( |y - \theta| - \frac{\alpha}{2} \right) & \text{otherwise} \end{cases}$$

➤ Take the derivative of the average Huber Loss

$$\frac{\partial}{\partial \theta} L_{\mathcal{D}}(\theta) = \frac{1}{n} \sum_{i=1}^n \begin{cases} -(y_i - \theta) & |y_i - \theta| < \alpha \\ -\alpha \mathbf{sign}(y_i - \theta) & \text{otherwise} \end{cases}$$

$$\frac{\partial}{\partial \theta} L_{\mathcal{D}}(\theta) = \frac{1}{n} \sum_{i=1}^n \begin{cases} -(y_i - \theta) & |y_i - \theta| < \alpha \\ -\alpha \mathbf{sign}(y_i - \theta) & \text{otherwise} \end{cases}$$

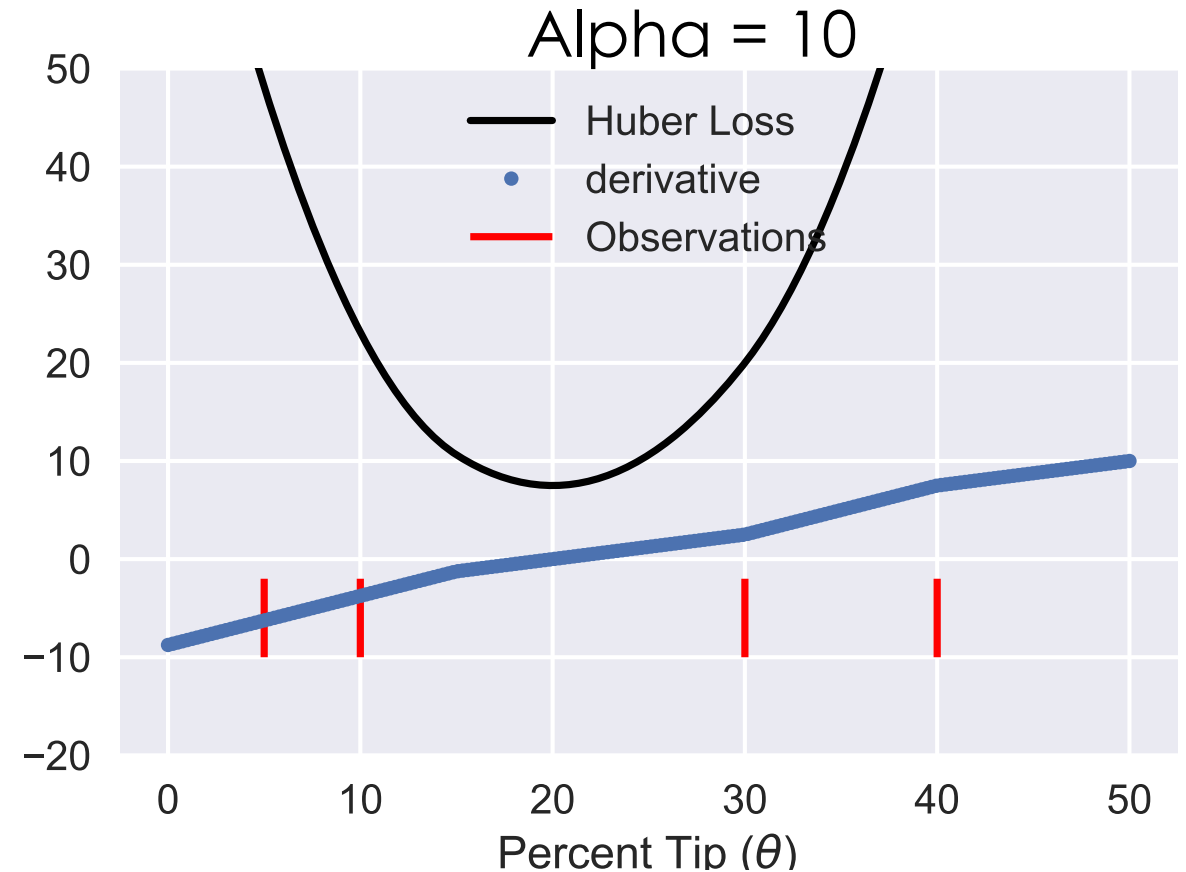
- Set derivative equal to zero:

$$\left( \sum_{\theta \geq y_i + \alpha} \alpha \right) - \left( \sum_{\theta \leq y_i - \alpha} \alpha \right) - \left( \sum_{|y_i - \theta| < \alpha} (y_i - \theta) \right) = 0$$

- Solution?
- No simple analytic solution ...
  - We can still plot the derivative

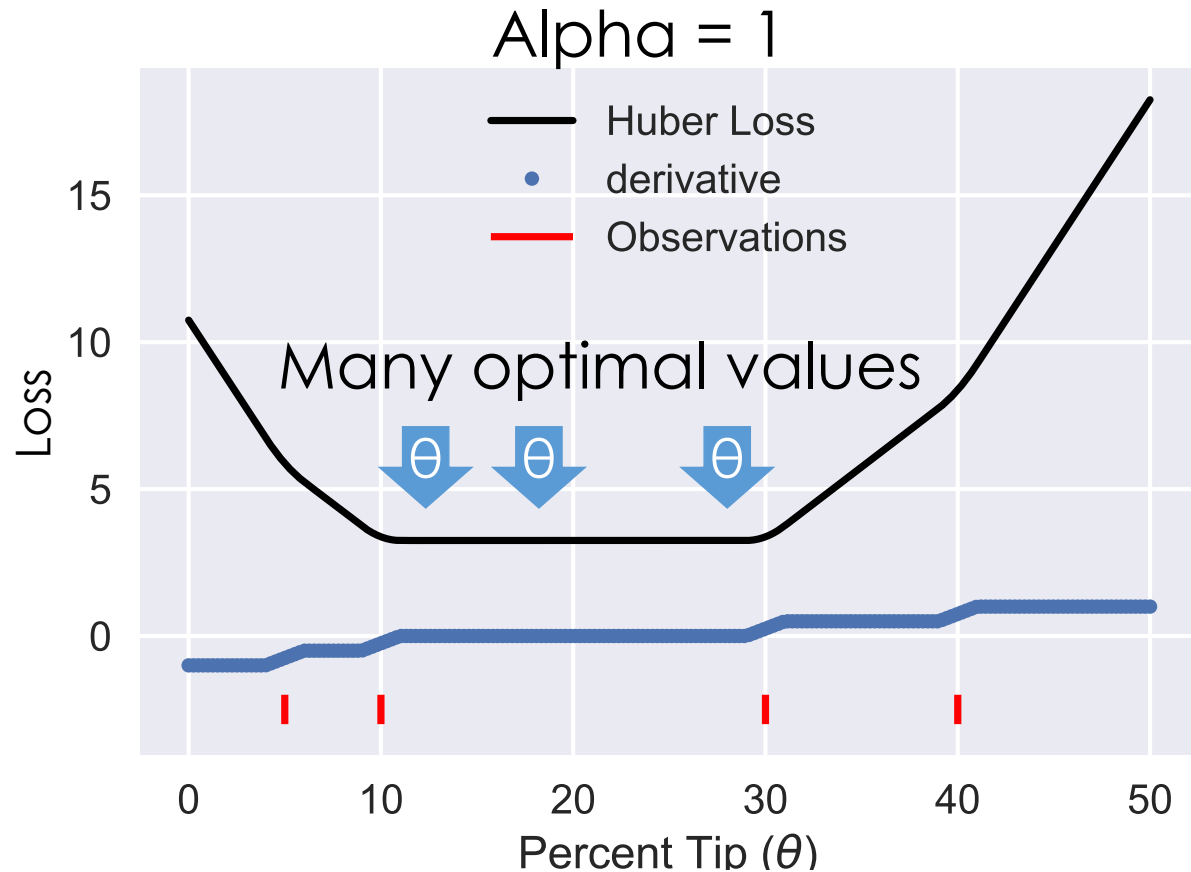
# Visualizing the Derivative of the Huber Loss

$$L_{\alpha}(\theta, y) = \begin{cases} \frac{1}{2} (y - \theta)^2 & |y - \theta| < \alpha \\ \alpha (|y - \theta| - \frac{\alpha}{2}) & \text{otherwise} \end{cases}$$

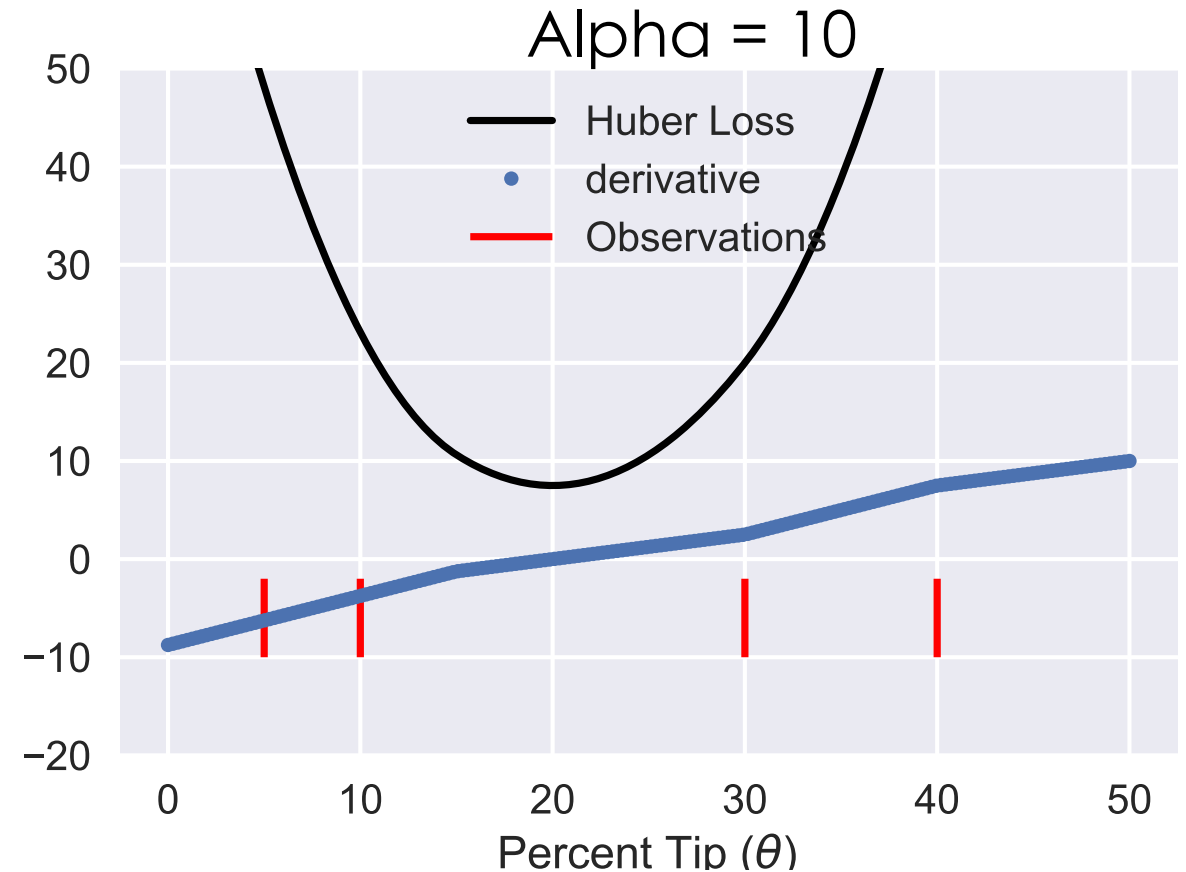


➤ Large  $\alpha \rightarrow$  unique optimum like squared loss

# Visualizing the Derivative of the Huber Loss



- Derivative is continuous
- Small  $\alpha \rightarrow$  many optima



- Large  $\alpha \rightarrow$  unique optimum like squared loss

# Numerical Optimization

# Minimizing the Huber Loss Numerically

Often we will use  
numerical optimization  
methods

The following are **helpful properties** when using numerical optimization methods:

- **convex** loss function
- **smooth** loss function
- analytic **derivative**

```
from scipy.optimize import minimize

def huber_loss_derivative(est, y_obs, alpha=1):
    d = abs_loss(est, y_obs)
    return np.where(d < alpha,
                    -(y_obs - est),
                    -alpha * np.sign(y_obs - est))

f = lambda theta: data['pcttip'].apply(
    lambda y: huber_loss(theta, y)).mean()
df = lambda theta: data['pcttip'].apply(
    lambda y: huber_loss_derivative(theta, y)).mean()
minimize(f, x0=0.0, jac=df)
```

```
      fun: 3.4999248461189802
 hess_inv: array([[ 5.08333333]])
       jac: array([ 4.36809059e-17])
 message: 'Optimization terminated successfully.'
      nfev: 10
       nit: 7
      njev: 10
   status: 0
  success: True
         x: array([ 15.53063381])
```

# Summary of Model Estimation

- 1. Define the Model:** simplified representation of the world
  - Use domain knowledge but ... **keep it simple!**
  - Introduce **parameters** for the unknown quantities
- 2. Define the Loss Function:** measures how well a particular instance of the model “fits” the data
  - We introduced  $L^2$ ,  $L^1$ , and Huber losses for each record
  - Take the average loss over the entire dataset
- 3. Minimize the Loss Function:** find the parameter values that minimize the loss on the data
  - We did this graphically
  - **Minimize the loss analytically using calculus**
  - **Minimize the loss numerically**



Improving the Model

# Going beyond the simple model

$$\text{percentage tip} = \theta^*$$

- How could we improve upon this model?
- Things to consider when improving the model
  - **Related factors** to the quantity of interest
    - Examples: quality of service, table size, time of day, total bill
    - Do we have data for these factors?
  - The **form of the relationship** to the quantity of interest
    - Linear relationships, step functions, etc ...
  - Goals for improving the model
    - Improve **prediction accuracy** → more complex models
    - Provide **understanding** → simpler models
  - Is my model “identifiable” (is it possible to estimate the parameters?)
    - $\text{percent tip} = \theta_1^* + \theta_2^*$  ← many identical parameterizations

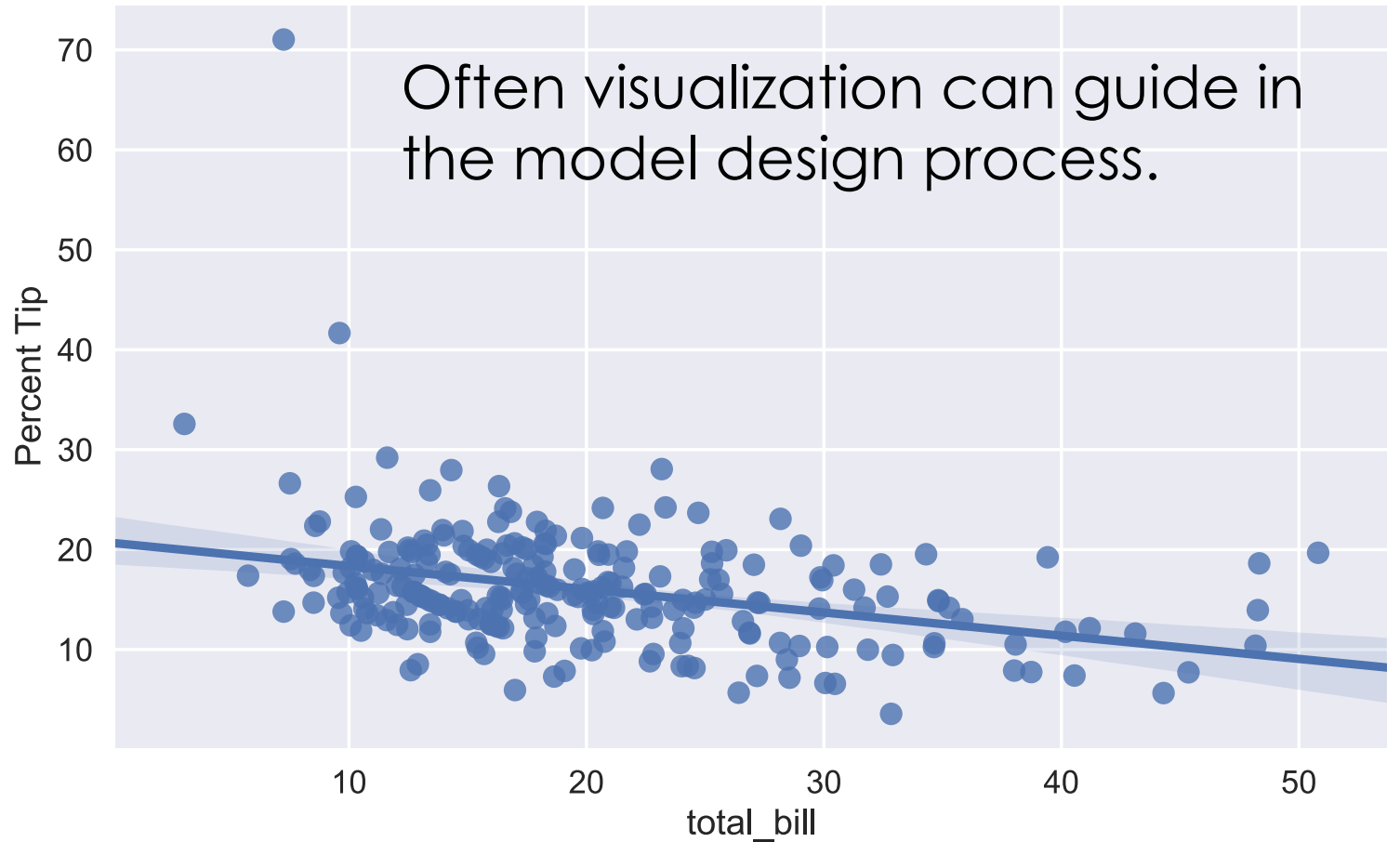
$$\text{percentage tip} = \theta_1^* + \theta_2^* * \text{total bill}$$

### Rationale:

Larger bills result in larger tips and people tend to be more careful or stingy on big tips.

### Parameter Interpretation:

- $\theta_1$ : Base tip percentage
- $\theta_2$ : Reduction/increase in tip for an increase in total bill.



# Estimating the model parameters:

$$\text{percentage tip} = \theta_1^* + \theta_2^* * \text{total bill}$$

- Write the loss (e.g., average squared loss)

$$L_{\mathcal{D}}(\theta_1, \theta_2) = \frac{1}{n} \sum_{i=1}^n (y_i - (\theta_1 + \theta_2 x_i))^2$$

$n$  {

	$x_i$ (Total Bill)	$y_i$ (% Tip)
0	16.99	5.944673
1	10.34	16.054159
2	21.01	16.658734
3	23.68	13.978041
4	24.59	14.680765

% Tip

Total Bill

$$L_{\mathcal{D}}(\theta_1, \theta_2) = \frac{1}{n} \sum_{i=1}^n (y_i - (\theta_1 + \theta_2 x_i))^2$$

➤ Take the derivative(s):

$$\begin{aligned} \frac{\partial}{\partial \theta_1} L_{\mathcal{D}}(\theta_1, \theta_2) &= \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta_1} (y_i - (\theta_1 + \theta_2 x_i))^2 \\ &= -\frac{2}{n} \sum_{i=1}^n (y_i - (\theta_1 + \theta_2 x_i)) \end{aligned}$$

$$L_{\mathcal{D}}(\theta_1, \theta_2) = \frac{1}{n} \sum_{i=1}^n (y_i - (\theta_1 + \theta_2 x_i))^2$$

➤ Take the derivative(s):

$$\frac{\partial}{\partial \theta_1} L_{\mathcal{D}}(\theta_1, \theta_2) = -\frac{2}{n} \sum_{i=1}^n (y_i - (\theta_1 + \theta_2 x_i))$$

$$\frac{\partial}{\partial \theta_2} L_{\mathcal{D}}(\theta_1, \theta_2) = -\frac{2}{n} \sum_{i=1}^n (y_i - (\theta_1 + \theta_2 x_i)) \frac{\partial}{\partial \theta_2} \theta_2 x_i$$

$$= -\frac{2}{n} \sum_{i=1}^n (y_i - (\theta_1 + \theta_2 x_i)) x_i$$

$$L_{\mathcal{D}}(\theta_1, \theta_2) = \frac{1}{n} \sum_{i=1}^n (y_i - (\theta_1 + \theta_2 x_i))^2$$

➤ Take the derivative(s):

$$\frac{\partial}{\partial \theta_1} L_{\mathcal{D}}(\theta_1, \theta_2) = -\frac{2}{n} \sum_{i=1}^n (y_i - (\theta_1 + \theta_2 x_i))$$

$$\frac{\partial}{\partial \theta_2} L_{\mathcal{D}}(\theta_1, \theta_2) = -\frac{2}{n} \sum_{i=1}^n (y_i - (\theta_1 + \theta_2 x_i)) x_i$$

➤ Set derivatives equal to zero and solve for parameters

# Solving for $\theta_1$

$$0 = -\frac{2}{n} \sum_{i=1}^n (y_i - (\theta_1 + \theta_2 x_i))$$

Breaking apart the sum

$$= -\frac{2}{n} \left( \left( \sum_{i=1}^n y_i \right) - n\theta_1 - \theta_2 \sum_{i=1}^n x_i \right)$$

Rearranging  
Terms



$$\sum_{i=1}^n y_i = n\theta_1 + \theta_2 \sum_{i=1}^n x_i$$



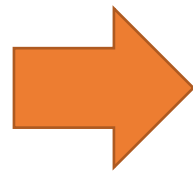
# Solving for $\theta_1$

$$\sum_{i=1}^n y_i = n\theta_1 + \theta_2 \sum_{i=1}^n x_i \quad \xrightarrow{\text{Divide by } n} \quad \frac{1}{n} \sum_{i=1}^n y_i = \theta_1 + \theta_2 \frac{1}{n} \sum_{i=1}^n x_i$$

➤ Define the average of x and y:

$$\bar{x} := \frac{1}{n} \sum_{i=1}^n x_i$$

$$\bar{y} := \frac{1}{n} \sum_{i=1}^n y_i$$



$$\bar{y} = \theta_1 + \theta_2 \bar{x}$$



$$\theta_1 = \bar{y} - \theta_2 \bar{x}$$

# Solving for $\theta_2$

Scratch

$$\theta_1 = \bar{y} - \theta_2 \bar{x}$$

$$\frac{\partial}{\partial \theta_2} L_{\mathcal{D}}(\theta_1, \theta_2) = -\frac{2}{n} \sum_{i=1}^n (y_i - (\theta_1 + \theta_2 x_i)) x_i$$

Distributing the  $x_i$  term

$$= -\frac{2}{n} \sum_{i=1}^n (y_i x_i - \theta_1 x_i - \theta_2 x_i^2)$$

Breaking apart the sum

$$= -\frac{2}{n} \left( \left( \sum_{i=1}^n y_i x_i \right) - \left( \theta_1 \sum_{i=1}^n x_i \right) - \theta_2 \sum_{i=1}^n x_i^2 \right)$$

# Solving for $\theta_2$

Scratch

$$\theta_1 = \bar{y} - \theta_2 \bar{x}$$

$$0 = -\frac{2}{n} \left( \left( \sum_{i=1}^n y_i x_i \right) - \left( \theta_1 \sum_{i=1}^n x_i \right) - \theta_2 \sum_{i=1}^n x_i^2 \right)$$

Rearranging  
Terms 

$$\sum_{i=1}^n y_i x_i = \theta_1 \sum_{i=1}^n x_i + \theta_2 \sum_{i=1}^n x_i^2$$

Divide by n 

$$\frac{1}{n} \sum_{i=1}^n y_i x_i = \theta_1 \frac{1}{n} \sum_{i=1}^n x_i + \theta_2 \frac{1}{n} \sum_{i=1}^n x_i^2$$

# Solving for $\theta_2$

Scratch

$$\theta_1 = \bar{y} - \theta_2 \bar{x}$$

$$\frac{1}{n} \sum_{i=1}^n y_i x_i = \theta_1 \frac{1}{n} \sum_{i=1}^n x_i + \theta_2 \frac{1}{n} \sum_{i=1}^n x_i^2$$

$$\bar{x} := \frac{1}{n} \sum_{i=1}^n x_i$$



$$\overline{xy} = \theta_1 \bar{x} + \theta_2 \overline{x^2}$$

$$\overline{xy} := \frac{1}{n} \sum_{i=1}^n x_i y_i$$

$$\overline{x^2} := \frac{1}{n} \sum_{i=1}^n x_i^2$$

# System of Linear Equations

- Substituting  $\theta_1$  and solving for  $\theta_2$

$$\overline{xy} = (\bar{y} - \theta_2 \bar{x}) \bar{x} + \theta_2 \overline{x^2}$$

$$\theta_1 = \bar{y} - \theta_2 \bar{x}$$

$$= \bar{y}\bar{x} - \theta_2 \bar{x}^2 + \theta_2 \overline{x^2}$$

$$\overline{xy} = \theta_1 \bar{x} + \theta_2 \overline{x^2}$$

$$= \bar{y}\bar{x} + \theta_2 (\overline{x^2} - \bar{x}^2)$$

 solving for  $\theta_2$

$$\theta_2 = \frac{\overline{xy} - \bar{y}\bar{x}}{\overline{x^2} - \bar{x}^2} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n} \sum_{I=1}^n (x_i - \bar{x})^2}$$

Algebra...

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

➤ Completing the squares:

# Denominator Derivation

Skipped in Class

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

$$= \sum_{i=1}^n \left( (x_i - \bar{x})^2 - \bar{x}^2 + \bar{x}x_i \right)$$

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

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

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

➤ Completing the squares:

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

# Numerator Derivation

Skipped in Class

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

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

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

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

# Summary so far ...

- **Step 1:** Define the model with unknown parameters

$$\text{percentage tip} = \theta_1^* + \theta_2^* * \text{total bill}$$

- **Step 2:** Write the loss (we selected an average squared loss)

$$L_{\mathcal{D}}(\theta_1, \theta_2) = \frac{1}{n} \sum_{i=1}^n (y_i - (\theta_1 + \theta_2 x_i))^2$$

- **Step 3:** Minimize the loss
  - Analytically (using calculus)
  - Numerically (using optimization algorithms)



$$L_{\mathcal{D}}(\theta_1, \theta_2) = \frac{1}{n} \sum_{i=1}^n (y_i - (\theta_1 + \theta_2 x_i))^2$$

➤ **Step3:** Minimize the loss

- Analytically (using calculus)
- Numerically (using optimization algorithms)

$$\frac{\partial}{\partial \theta_1} L_{\mathcal{D}}(\theta_1, \theta_2) = -\frac{2}{n} \sum_{i=1}^n (y_i - (\theta_1 + \theta_2 x_i))$$

$$\frac{\partial}{\partial \theta_2} L_{\mathcal{D}}(\theta_1, \theta_2) = -\frac{2}{n} \sum_{i=1}^n (y_i - (\theta_1 + \theta_2 x_i)) x_i$$

- Set derivatives equal to zero and solve for parameter values

$$\theta_1 = \bar{y} - \theta_2 \bar{x}$$

$$\theta_2 = \frac{\overline{xy} - \bar{y}\bar{x}}{\overline{x^2} - \bar{x}^2} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

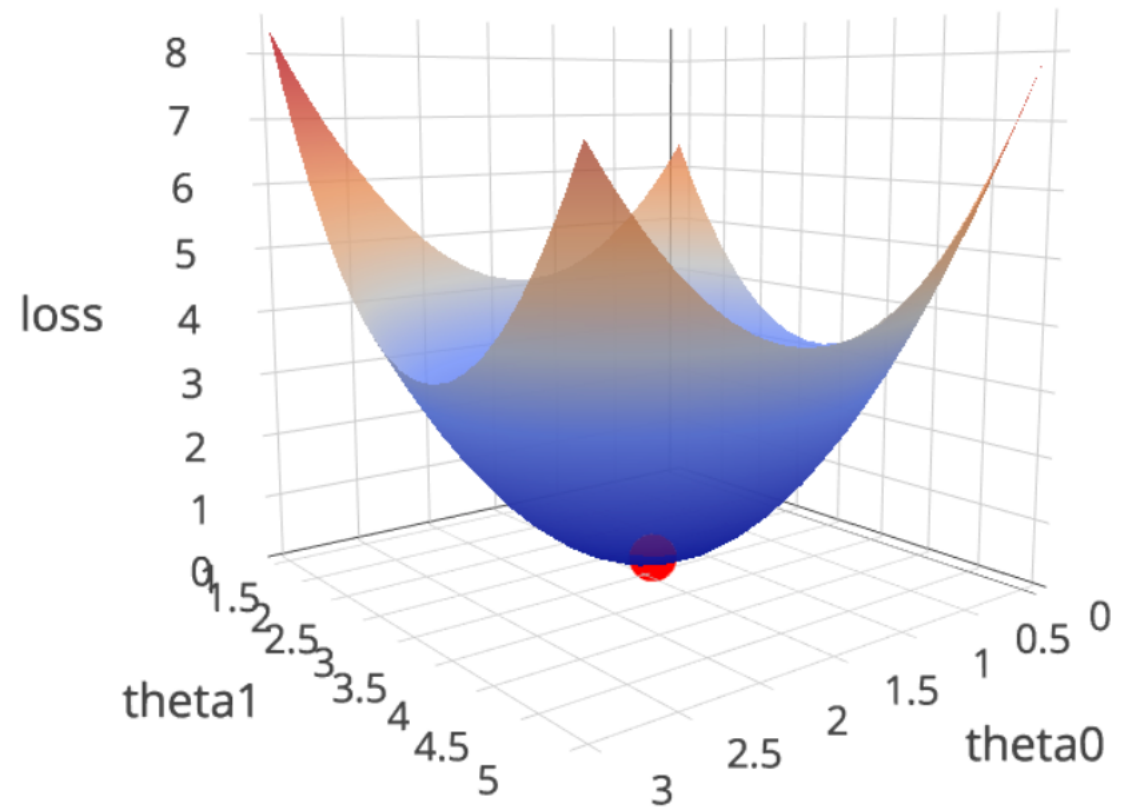
- Is this a local minimum?

$$\frac{\partial^2}{\partial \theta_1^2} L_{\mathcal{D}}(\theta_1, \theta_2) = -\frac{2}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta_1} (y_i - (\theta_1 + \theta_2 x_i)) = -\frac{2}{n} \sum_{i=1}^n -1 = 2$$

$$\frac{\partial^2}{\partial \theta_2^2} L_{\mathcal{D}}(\theta_1, \theta_2) = -\frac{2}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta_2} (y_i - (\theta_1 + \theta_2 x_i)) = \frac{2}{n} \sum_{i=1}^n x_i^2 > 0$$

# Visualizing the **Higher Dimensional Loss**

- What does the loss look like?
- Go to notebook ...



“Improving” the Model  
(more...)

$$\text{percentage tip} = \theta_1^* + \theta_2^* * \text{is Male} \\ + \theta_3^* * \text{is Smoker} + \theta_4^* * \text{table size}$$

### **Rational:**

Each term encodes a potential factor that could affect the percentage tip.

### **Possible Parameter Interpretation:**

- $\theta_1$ : base tip percentage paid by female non-smokers without accounting for table size.
- $\theta_2$ : tip change associated with male patrons ...

Maybe difficult to estimate ... what if all smokers are male?

Difficult  
to  
Plot

[Go to Notebook](#)

# Define the model

- Use python to define the function

```
def f(theta, data):  
    return (  
        theta[0] +  
        theta[1] * (data['sex'] == 'Male') +  
        theta[2] * (data['smoker'] == "Yes") +  
        theta[3] * data['size']  
    )
```

# Define and Minimize the Loss

```
def l2(theta):  
    return np.mean(squared_loss(f(theta, data), data['pcttip']).values)  
  
minimize(l2, x0=np.zeros(4))
```

```
fun: 36.25888793122608  
hess_inv: array([[ 5.00852276, -1.03468734, -1.13297213, -1.36869473],  
                [-1.03468734,  2.06166674,  0.00679159, -0.11857307],  
                [-1.13297213,  0.00679159,  2.08462848,  0.14029876],  
                [-1.36869473, -0.11857307,  0.14029876,  0.55080528]])  
jac: array([ 3.81469727e-06,  3.33786011e-06,  4.76837158e-07,  
            8.10623169e-06])  
message: 'Optimization terminated successfully.'  
nfev: 84  
nit: 13  
njev: 14  
status: 0  
→ success: True  
→ x: array([ 18.73866929, -0.73513124,  0.16122391, -0.87437012])
```

# Define and Minimize the Loss

```
def l1(theta):  
    return np.mean(abs_loss(f(theta, data), data['pcttip']).values)  
  
minimize(l1, np.zeros(4))
```

```
fun: 3.90957158852356  
hess_inv: array([[ 443.57329609, -215.55179077, -211.52560242, -109.7383045 ],  
                [-215.55179077,  104.77953797,  102.80962477,   53.31466531],  
                [-211.52560242,  102.80962477,  100.96345597,   52.31890909],  
                [-109.7383045 ,   53.31466531,   52.31890909,   27.15457305]])  
jac: array([ 0.00750431,  0.00340596,  0.00340596,  0.01979941])  
message: 'Desired error not necessarily achieved due to precision loss.'  
nfev: 1104  
nit: 31  
njev: 182  
status: 2  
→ success: False  
→ x: array([ 18.02471408, -0.72038142, -0.9579457 , -0.77126898])
```

**Why? Function is not smooth → Difficult to optimize**



# Define and Minimize the Loss

```
def huber(theta):  
    return np.mean(huber_loss(f(theta, data), data['pcttip']))  
  
minimize(huber, np.zeros(4))
```

```
fun: 3.4476306812527757  
hess_inv: array([[ 77.24012512, -19.71060902, -26.073196  , -20.40690306],  
                [-19.71060902,  20.85365616,  4.85116291,  2.01663757],  
                [-26.073196  ,  4.85116291, 28.8990574  ,  5.65213441],  
                [-20.40690306,  2.01663757,  5.65213441,  6.76874477]])  
jac: array([-1.19209290e-07, -8.94069672e-08, -1.19209290e-07,  
            -1.78813934e-07])  
message: 'Optimization terminated successfully.'  
nfev: 150  
nit: 21  
njev: 25  
status: 0  
→ success: True  
→ x: array([ 18.53021329, -0.90174037, -0.87843472, -0.84144212])
```