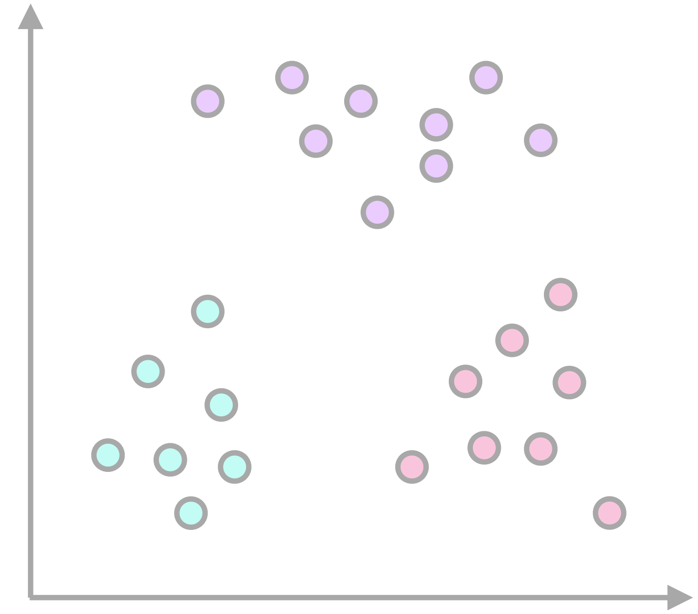
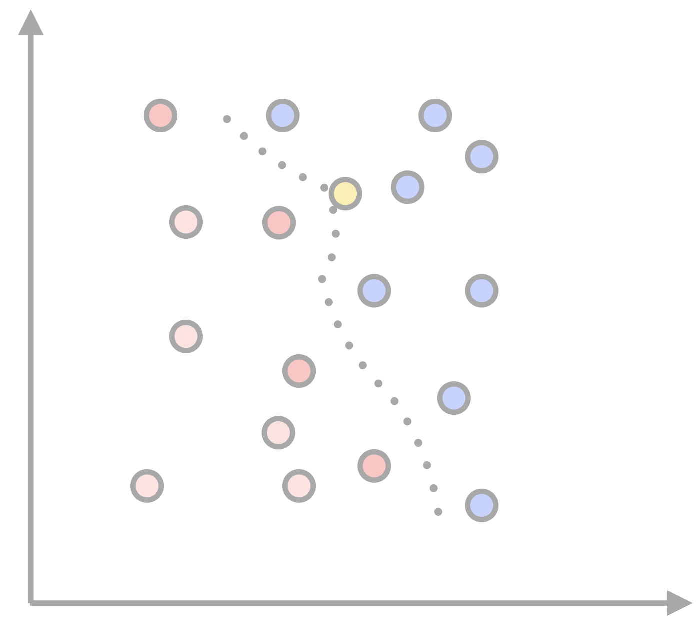
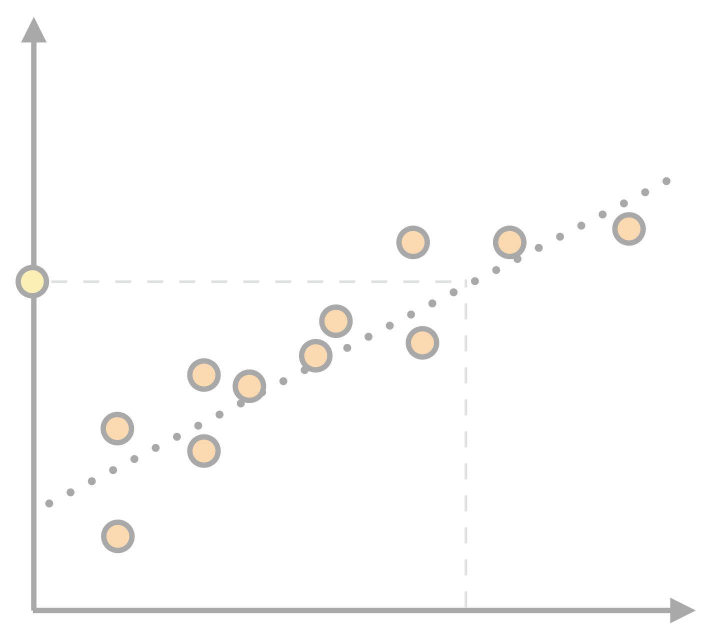
Machine Learning

1. Motivation + Theorie







Siegfried Gessulat

SAP Health Potsdam

Technical University of Munich
Chair of Proteomics and and Bioanalytics

s.gessulat@tum.de

FH Ludwigshafen
2018-10-09

Course Outline

Block I Foundations

Oct 08: Introduction
- Overview machine learning
- Theory: Linear Algebra
- Algorithms: Knn, K-means

Oct 09: Basics
- Theory: linear regression, logistic regression
- Algorithms: gradient descent

Block II Best practices

Oct 29: Neural Networks
- Data cleaning
- Algorithm: Neural Networks

Oct 30: Best practices
- Theory: Cross validation
- Theory: Regularization

Course Outline

Block III Dark Arts

<u>Nov 19: Tricks of the Trade</u>

- Ensembles
- Hyperparameter Search
- Deep Learning Black Magic

<u>Nov 20: Outlook</u>

- Theory: Dimensionality Reduction

Outline Today

1. Dataset: Boston
2. Linear Regression
3. Gradient Descent
4. Logistic Regression

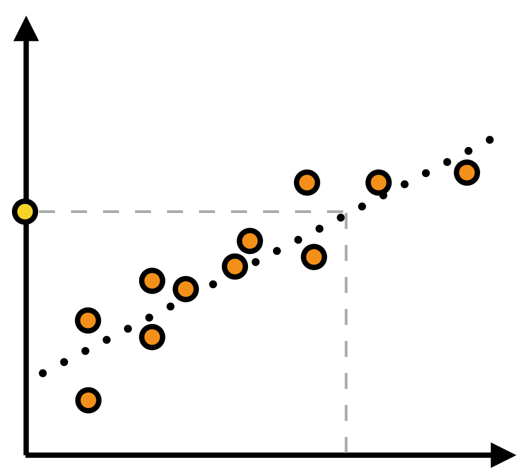
Machine Learning Overview

Do you have labeled data?

✓ **supervised**
what kind of label?

✗ **unsupervised**
what to generate?

continuous
regression

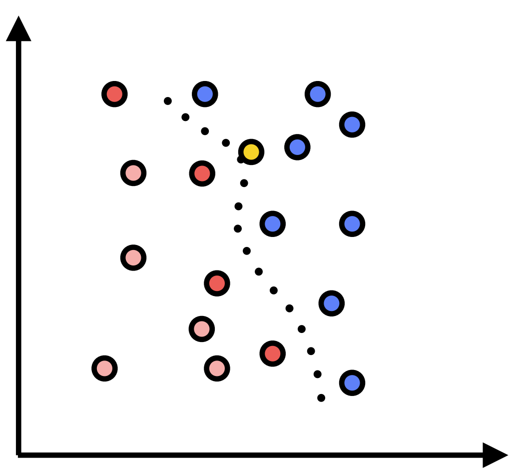


- K-nearest-neighbours

Oct 09

- Linear Regression

discrete
classification

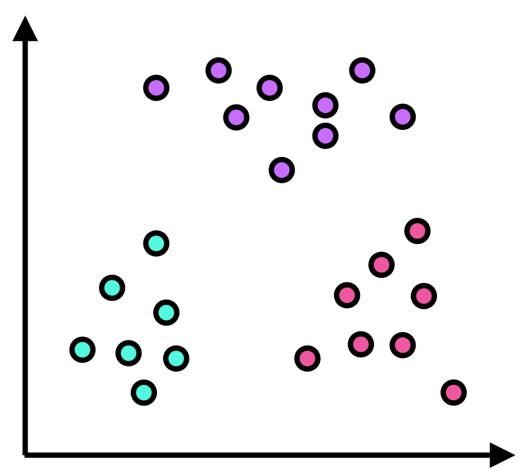


- Logistic Regression
- Support Vector Machines

Oct 09

- (Gaussian) Mixture Model

sample → label
clustering



- K-means
- (Gaussian) Mixture Models
- Self-organising maps*

sample → sample
generative

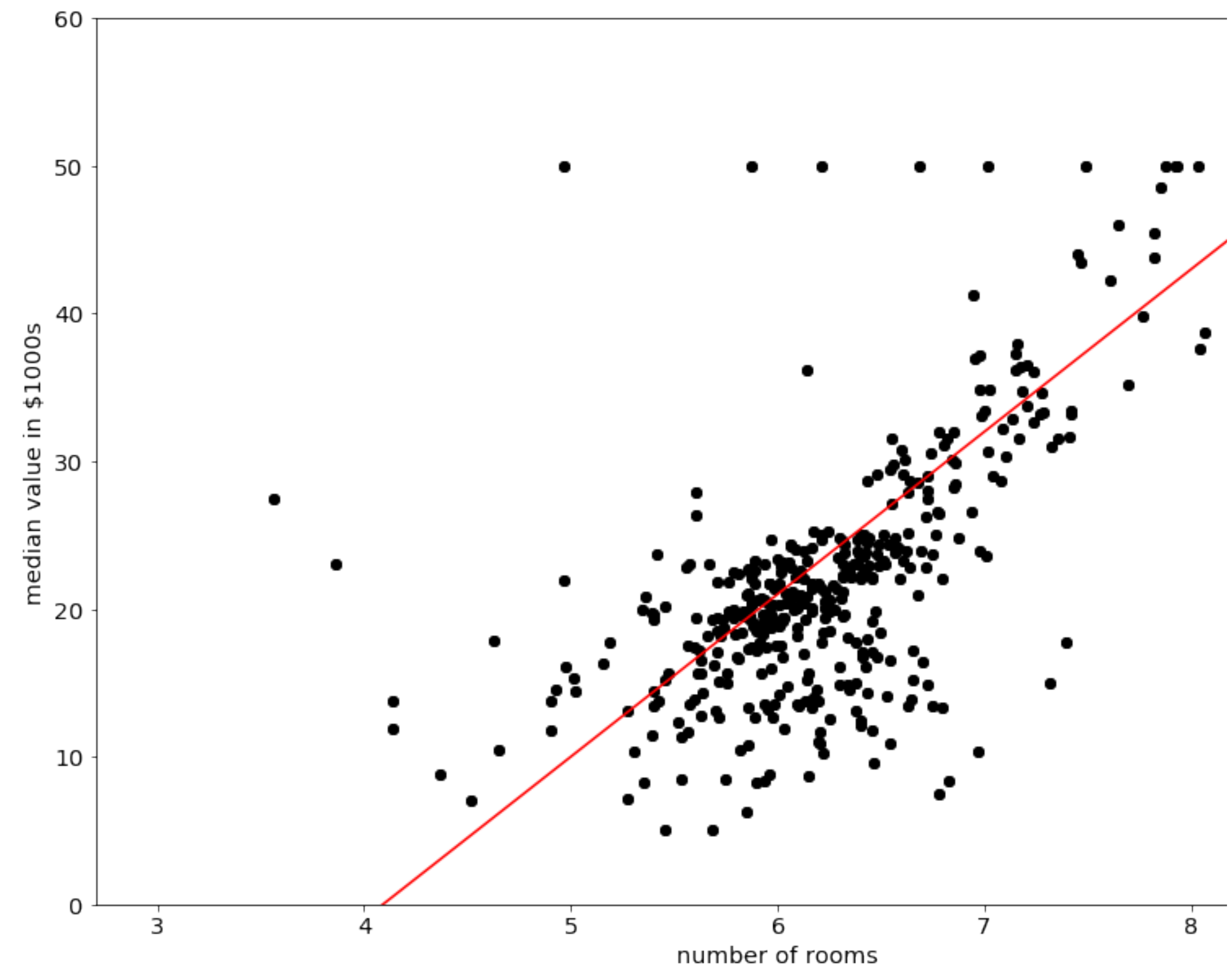
- Markov Chains
- Autoencoder*
- Generative Adversarial Networks*

* neural network

Dataset: Boston Housing Market

lib.stat.cmu.edu/datasets/boston

Harrison, D **Rubinfeld, D.L**



506 samples
14 attributes of real estate properties including its price.

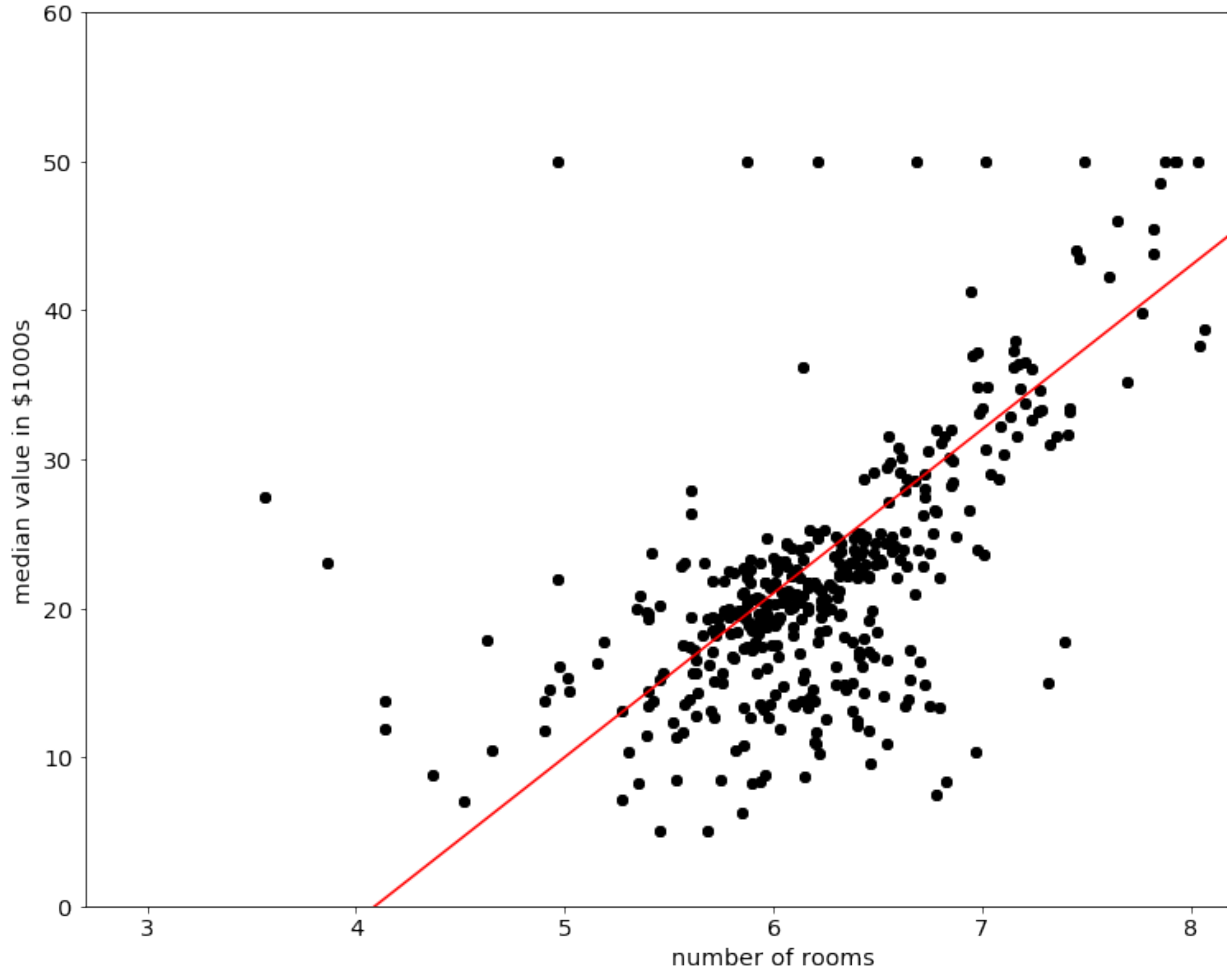
- **crim** - per capita crime rate by town.
- **indus** - proportion of non-retail business acres per town.
- **rm** - average number of rooms per dwelling.
- **age** - proportion of owner-occupied units built prior to 1940.
- **dis** - weighted mean of distances to five Boston employment centres.
- **tax** - full-value property-tax rate per $10,000.
- **ptratio** - pupil-teacher ratio by town.
- **medv** - median value of owner-occupied homes in $1000s.

Dataset: Boston Housing Market

lib.stat.cmu.edu/datasets/boston

Harrison, D **Rubinfeld, D.L**

506 samples
14 attributes of real estate properties including its price.

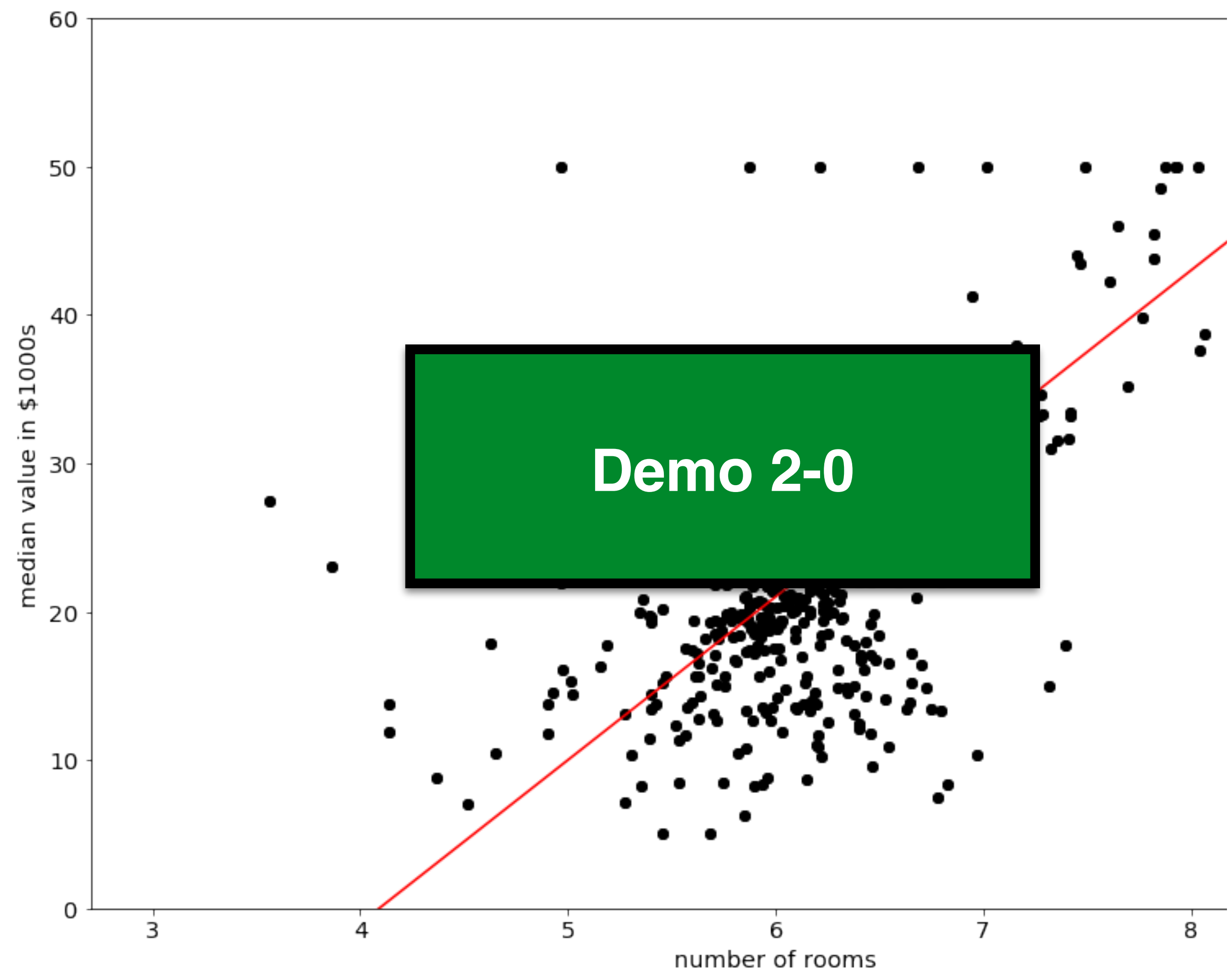


	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
0	1.23247	0.0	8.14	0.0	0.538	6.142	91.7	3.9769	4.0	307.0	21.0	396.90	18.72	15.2
1	0.02177	82.5	2.03	0.0	0.415	7.610	15.7	6.2700	2.0	348.0	14.7	395.38	3.11	42.3
2	4.89822	0.0	18.10	0.0	0.631	4.970	100.0	1.3325	24.0	666.0	20.2	375.52	3.26	50.0
3	0.03961	0.0	5.19	0.0	0.515	6.037	34.5	5.9853	5.0	224.0	20.2	396.90	8.01	21.1
4	3.69311	0.0	18.10	0.0	0.713	6.376	88.4	2.5671	24.0	666.0	20.2	391.43	14.65	17.7

Dataset: Boston Housing Market

lib.stat.cmu.edu/datasets/boston

Harrison, D **Rubinfeld, D.L**



506 samples
14 attributes of real estate properties including its price.

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
0	1.23247	0.0	8.14	0.0	0.538	6.142	91.7	3.9769	4.0	307.0	21.0	396.90	18.72	15.2
1	0.02177	82.5	2.03	0.0	0.415	7.610	15.7	6.2700	2.0	348.0	14.7	395.38	3.11	42.3
2	4.89822	0.0	18.10	0.0	0.631	4.970	100.0	1.3325	24.0	666.0	20.2	375.52	3.26	50.0
3	0.03961	0.0	5.19	0.0	0.515	6.037	34.5	5.9853	5.0	224.0	20.2	396.90	8.01	21.1
4	3.69311	0.0	18.10	0.0	0.713	6.376	88.4	2.5671	24.0	666.0	20.2	391.43	14.65	17.7

Linear Regression

A machine learning model is a function mapping X to Y.
Θ is what the model "learned".

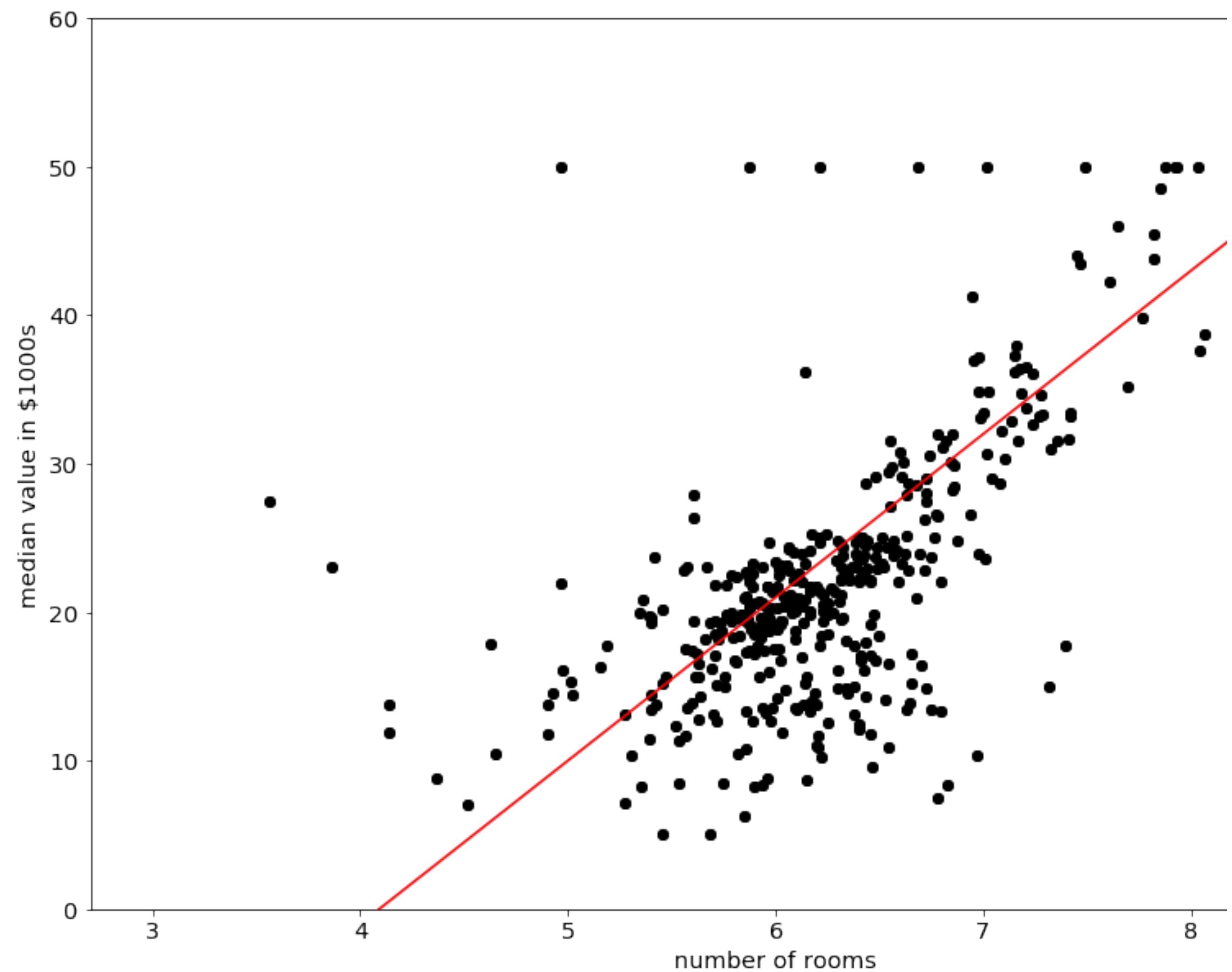
$$f_\theta : X \rightarrow Y$$

Hypothesis: prices follow a linear function of the number or rooms

$$f_\theta(x) = \theta_0 + \theta_1 x$$

θ_0 base value (no rooms)

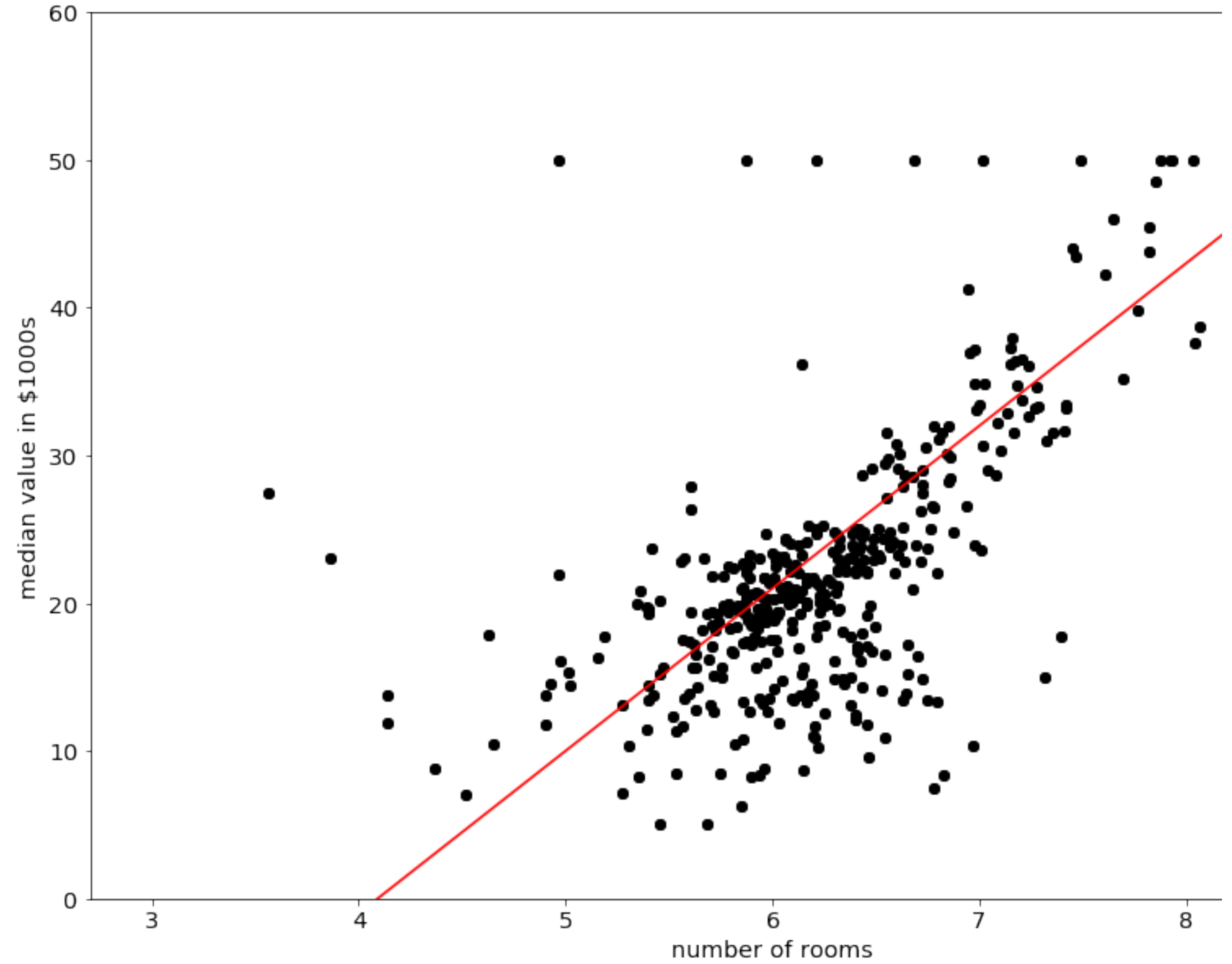
θ_1 price increase per room



	rm	medv
0	6.142	15.2
1	7.610	42.3
2	4.970	50.0
3	6.037	21.1
4	6.376	17.7

Linear Regression

Cost Function



We define a loss function to evaluate different hypotheses

$$L_{f,m}(\theta) = m(f_\theta(x), y)$$

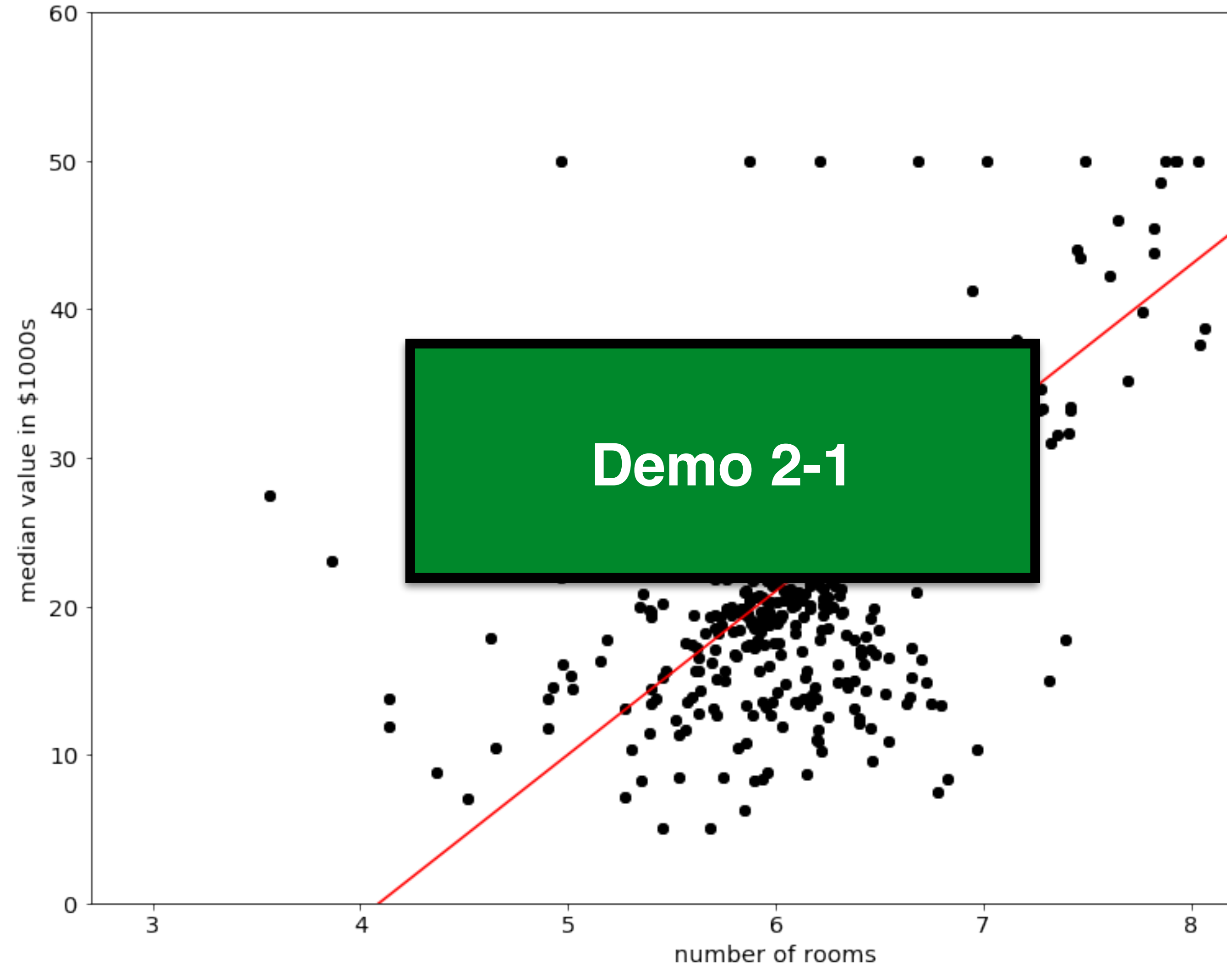
Choosing **mean squared error** as an error metric:

$$L(\theta) = \frac{1}{2n} \sum_{i=1}^{n} \left(f_\theta(x_i) - y_i\right)^2$$

Minimize L by evaluating different Θ.

Linear Regression

Cost Function



We define a loss function to evaluate different hypotheses

$$L_{f,m}(\theta) = m(f_\theta(x), y)$$

Choosing **mean squared error** as an error metric:

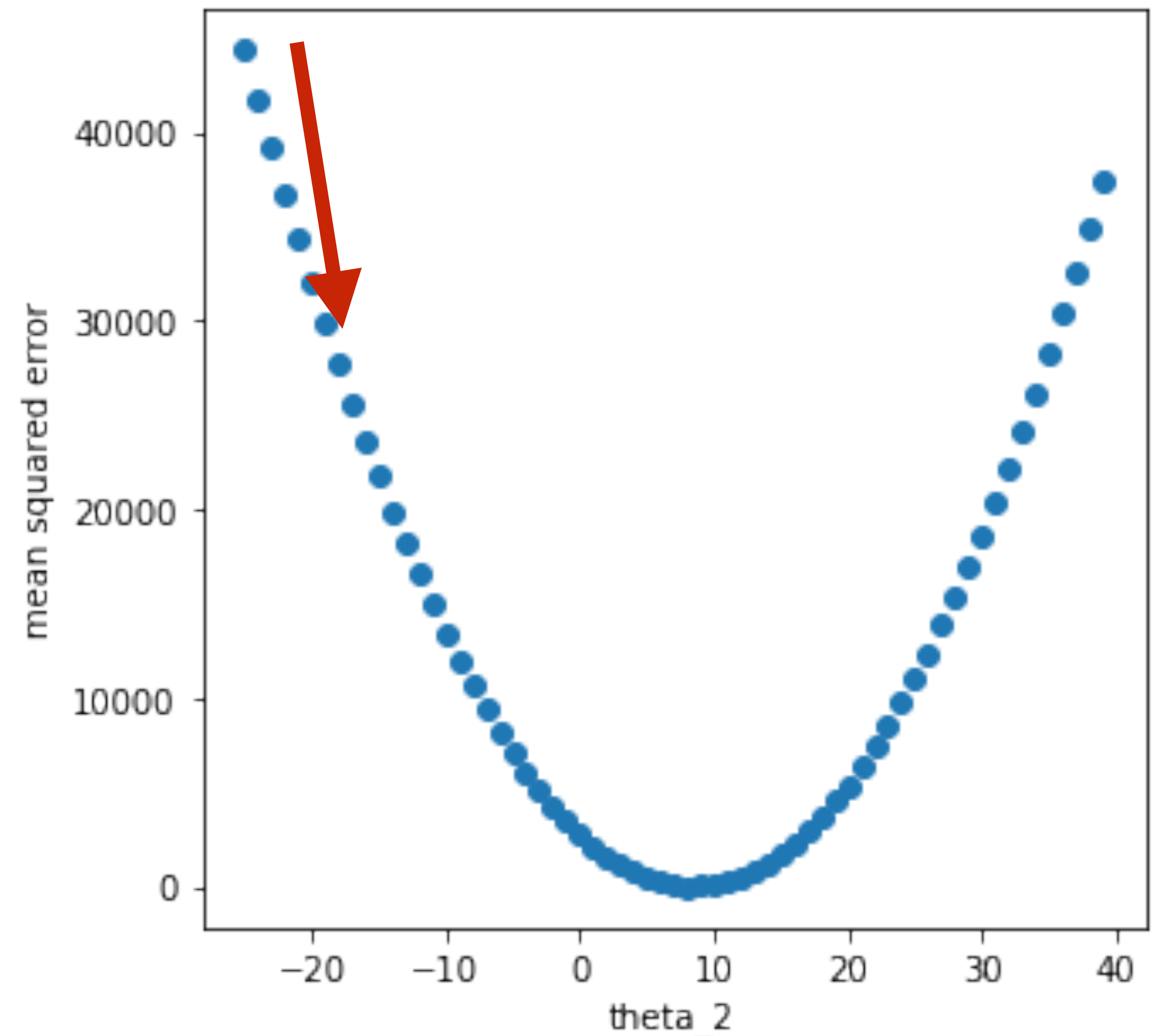
$$L(\theta) = \frac{1}{2n} \sum_{i=1}^{n} \left(f_\theta(x_i) - y_i\right)^2$$

Minimize L by evaluating different Θ.

Gradient Descent

Intuition

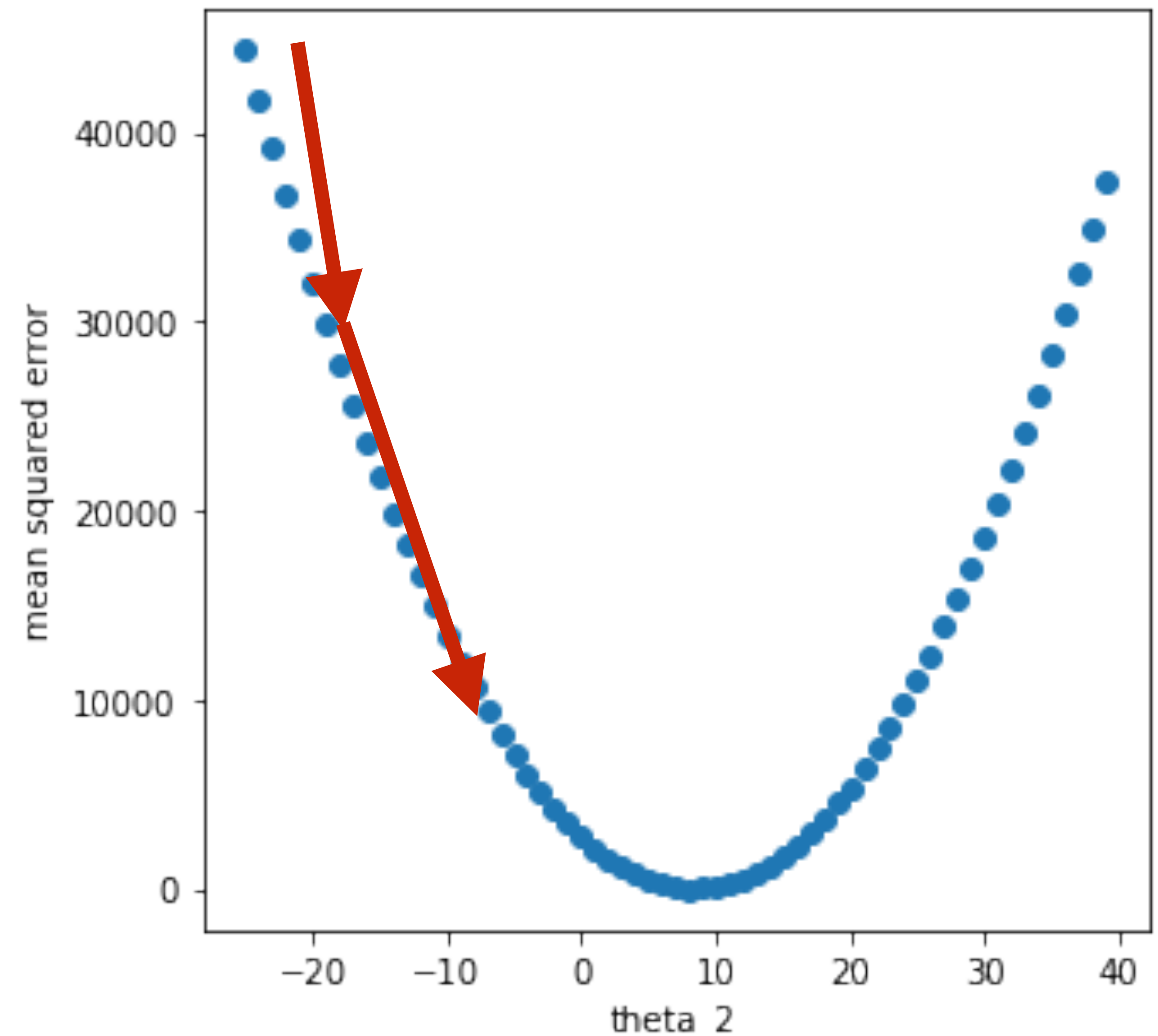
0. Start with random **θ**
1. Change **θ** to reduce **L**
2. Repeat till we reach minimum



Gradient Descent

Intuition

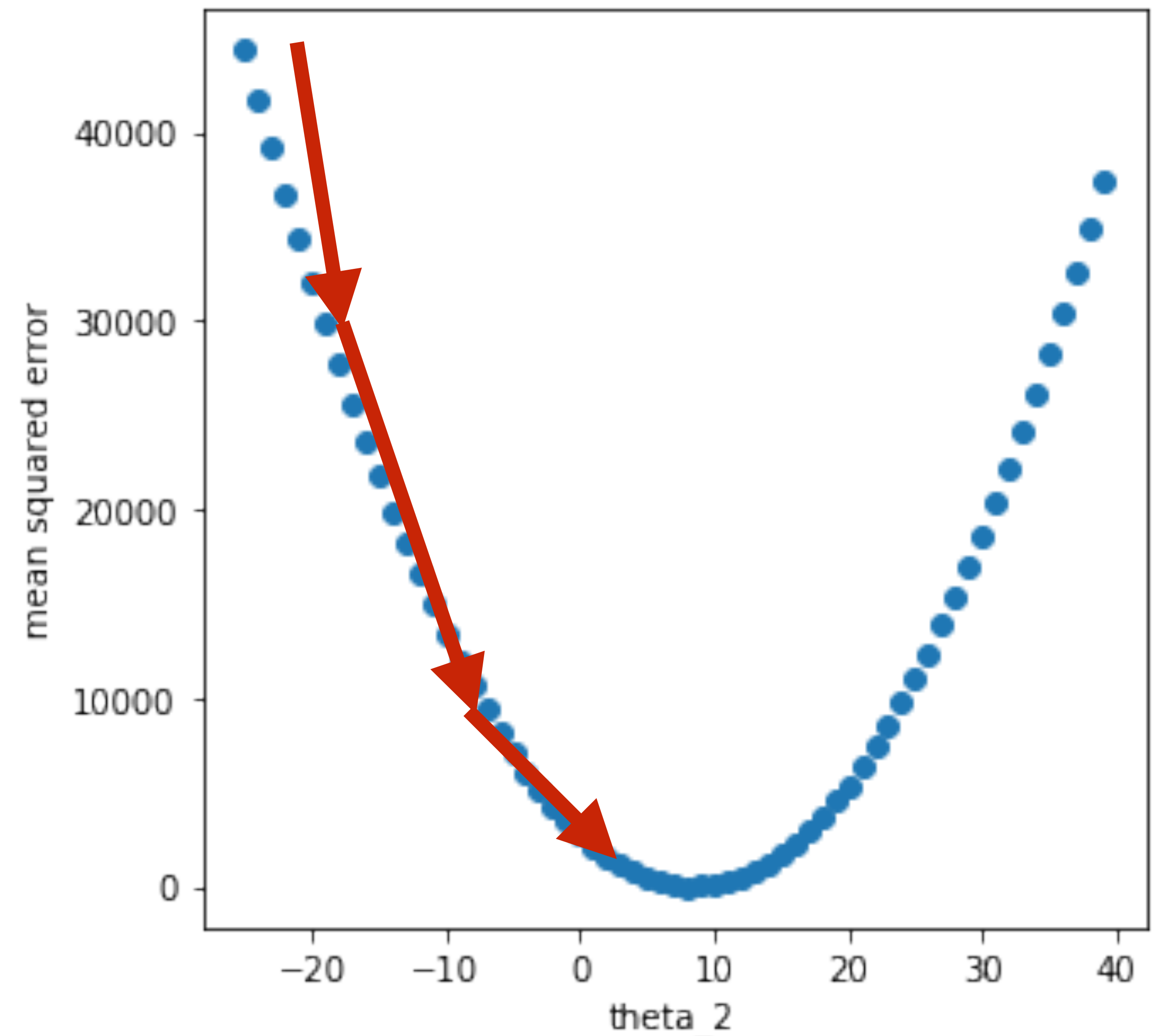
0. Start with random **θ**
1. Change **θ** to reduce **L**
2. Repeat till we reach minimum



Gradient Descent

Intuition

0. Start with random **θ**
1. Change **θ** to reduce **L**
2. Repeat till we reach minimum

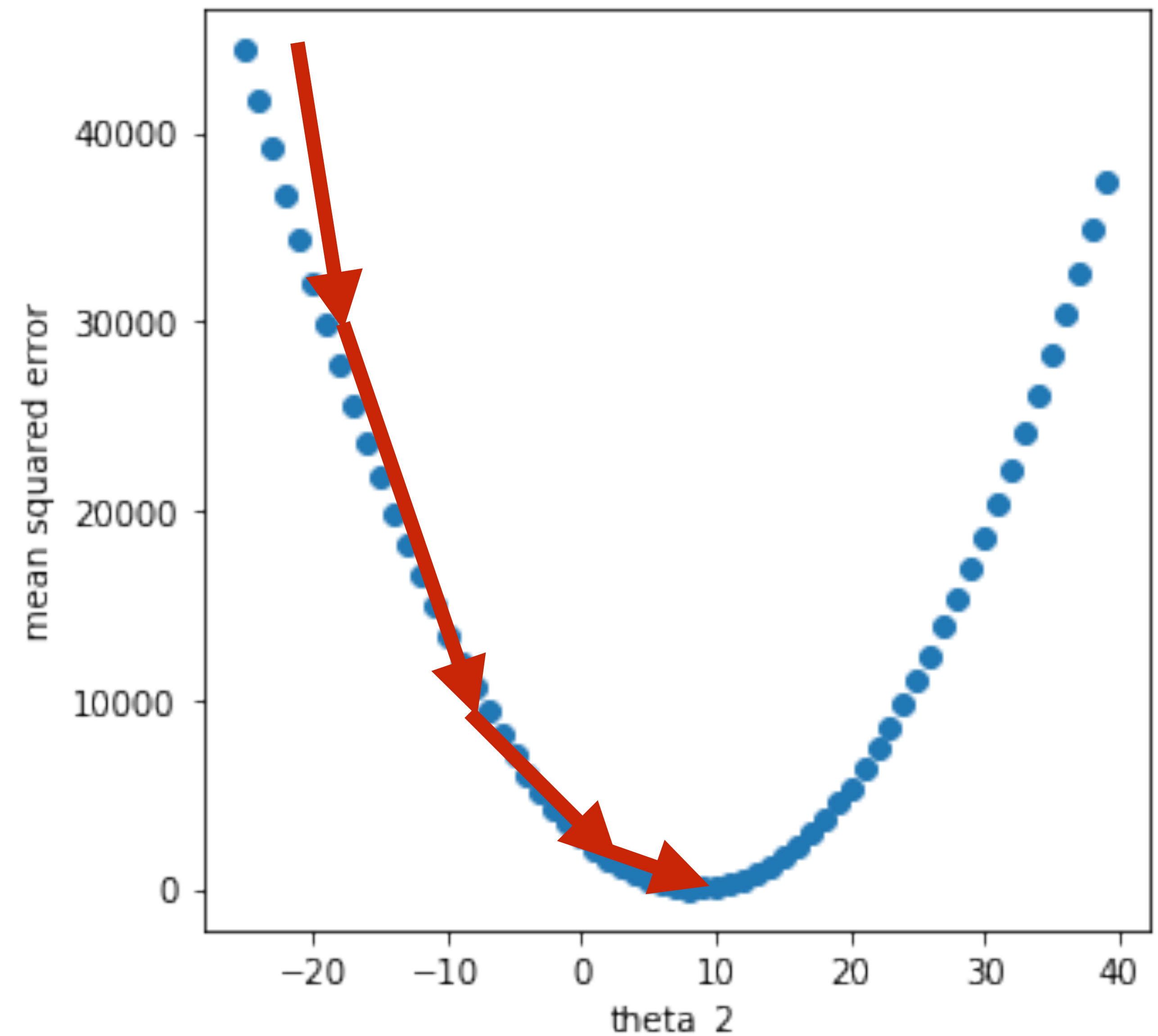


Gradient Descent

Intuition

0. Start with random **θ**
1. Change **θ** to reduce **L**
2. Repeat till we reach minimum

How to change **θ**?



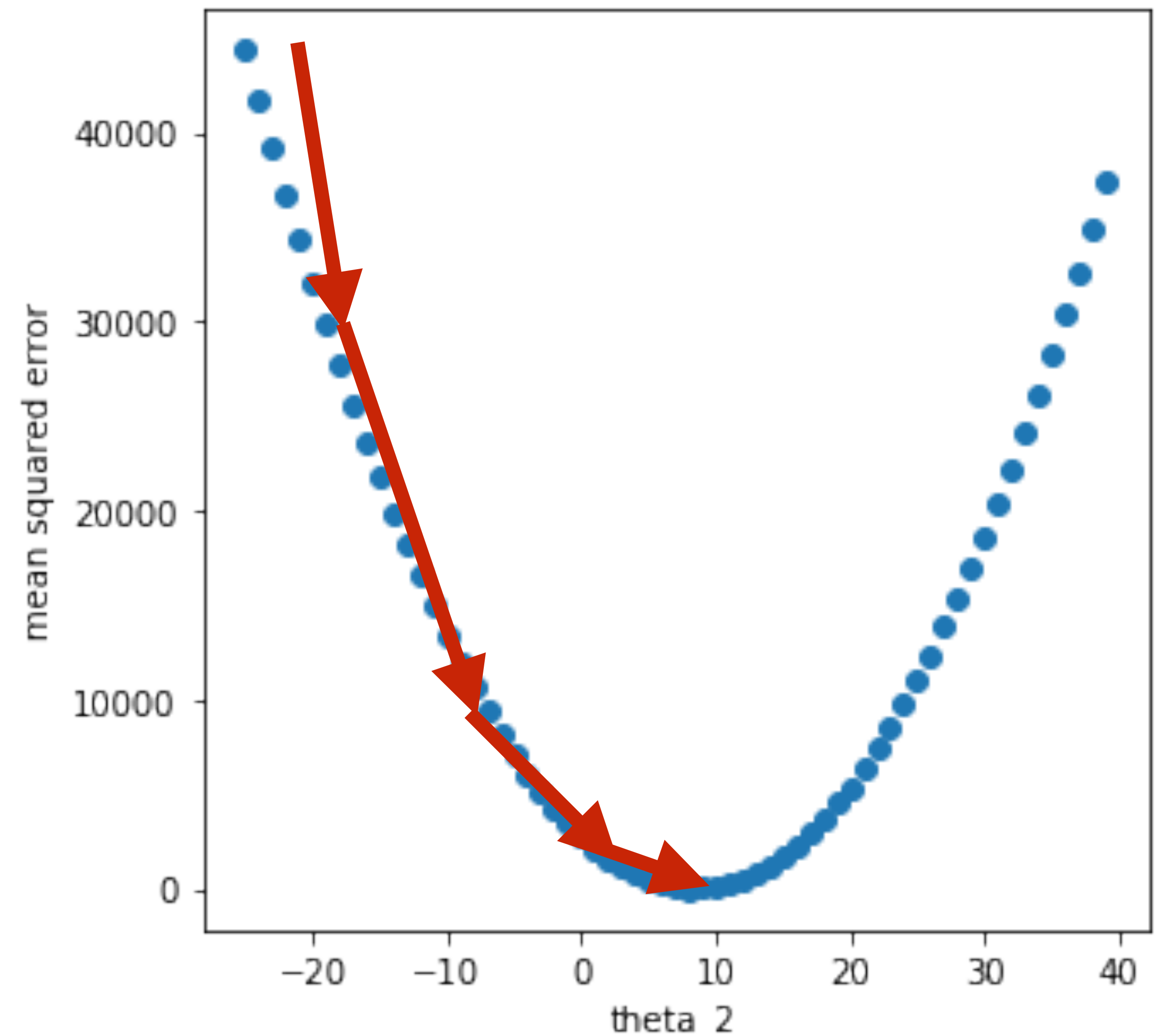
Gradient Descent

Intuition

0. Start with random **θ**
1. Change **θ** to reduce **L**
2. Repeat till we reach minimum

How to change **θ**?

$$\theta_i := \theta_i - \alpha \frac{\delta}{\delta\theta_j} J(\theta)$$



Gradient Descent

Algorithm

Input
- training dataset **X,Y** - matrices of samples (**x**, **y**)
- loss function **L**
- learning rate **α**

Output
- **θ** that minimizes **L** on **X, Y**

?

Linear Regression

hypothesis

$$f_\theta(x) = \theta_0 + \theta_1 x$$

cost function

$$L(\theta) = \frac{1}{2n} \sum_{i=1}^{n} \left(f_\theta(x_i) - y_i \right)^2$$

derivatives

$$\frac{\delta}{\delta\theta_0} L(\theta) = \frac{1}{n} \sum_{i=1}^{n} f_\theta(x_i) - y_i$$

$$\frac{\delta}{\delta\theta_1} L(\theta) = \frac{1}{n} \sum_{i=1}^{n} (f_\theta(x_i) - y_i) x_i$$

Gradient Descent

Algorithm

Input
- training dataset **X,Y** - matrices of samples (**x**, **y**)
- loss function **L**
- learning rate **α**

Output
- **θ** that minimizes **L** on **X, Y**

Repeat until convergence
for all *i:*

$$\theta_i := \theta_i - \alpha \frac{\delta}{\delta \theta_j} L(\theta)$$

update simultaneously!

Linear Regression

hypothesis

$$f_\theta(x) = \theta_0 + \theta_1 x$$

cost function

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

derivatives

$$\frac{\delta}{\delta \theta_0} L(\theta) = \frac{1}{n} \sum_{i=1}^{n} f_\theta(x_i) - y_i$$

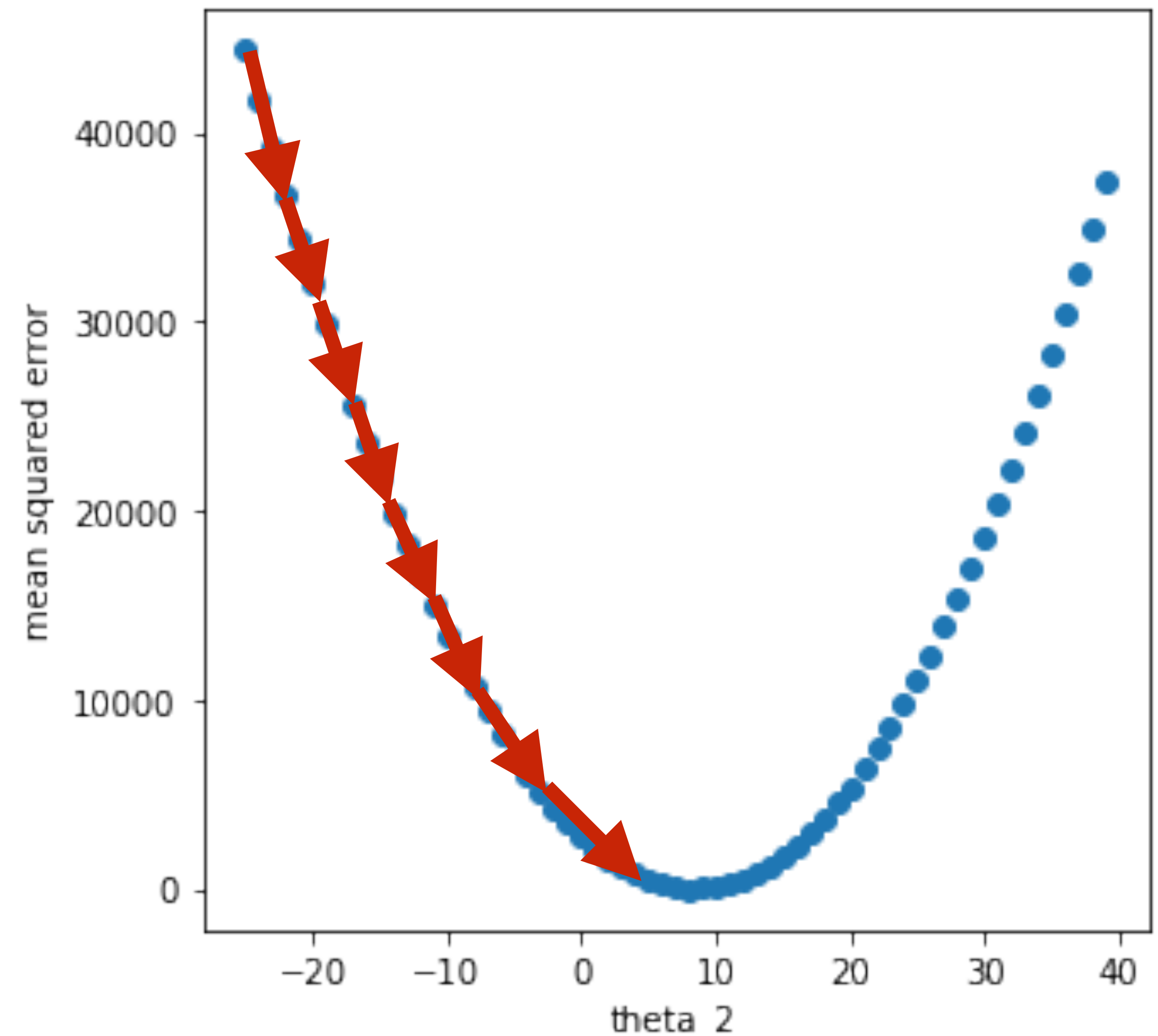
$$\frac{\delta}{\delta \theta_1} L(\theta) = \frac{1}{n} \sum_{i=1}^{n} (f_\theta(x_i) - y_i) x_i$$

Gradient Descent

How to change **θ**?

$$\theta_i := \theta_i - \alpha \frac{\delta}{\delta \theta_j} J(\theta)$$

small learning rate

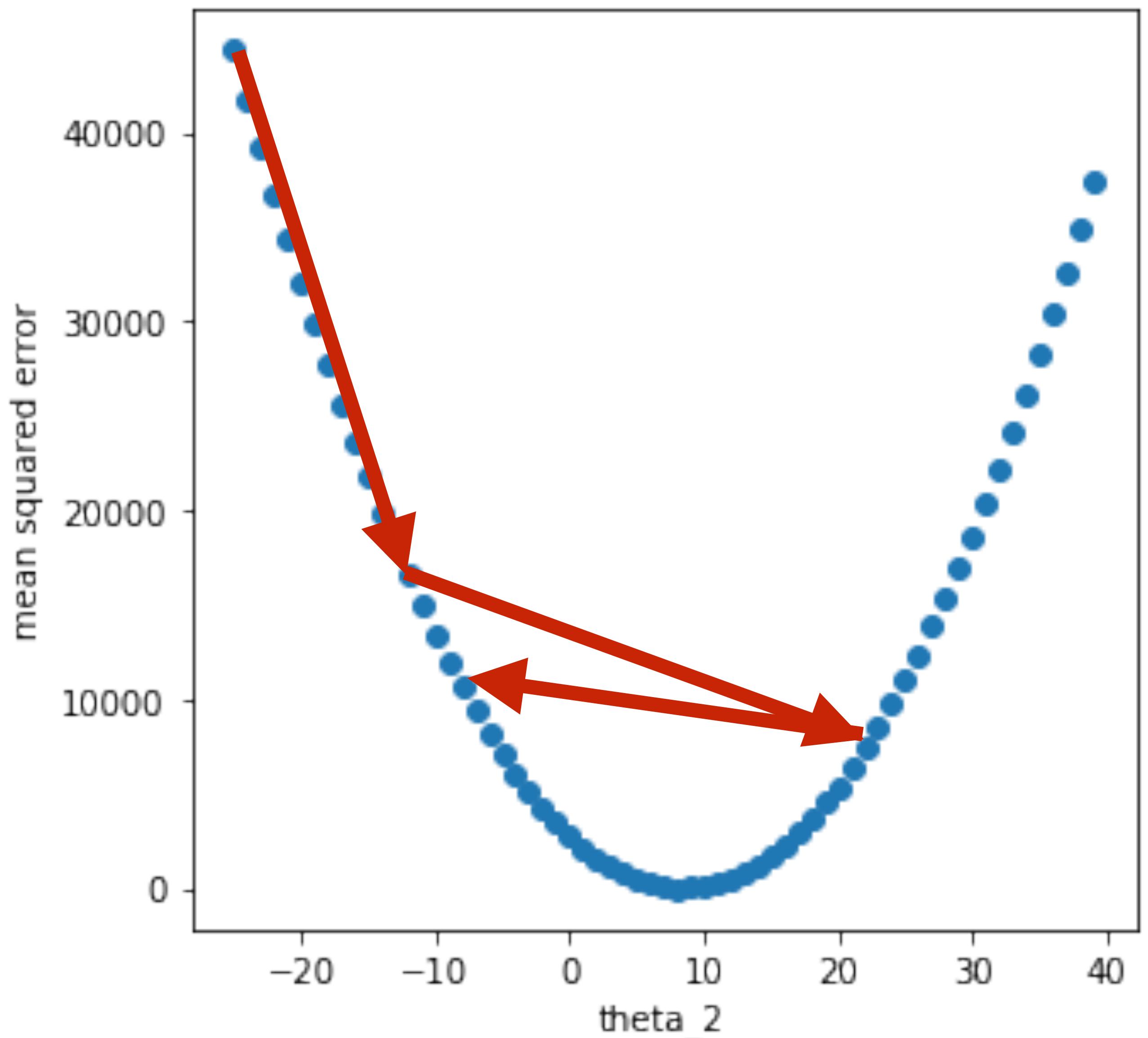


Gradient Descent

How to change **θ**?

$$\theta_i := \theta_i - \alpha \frac{\delta}{\delta \theta_j} J(\theta)$$

learning rate too big

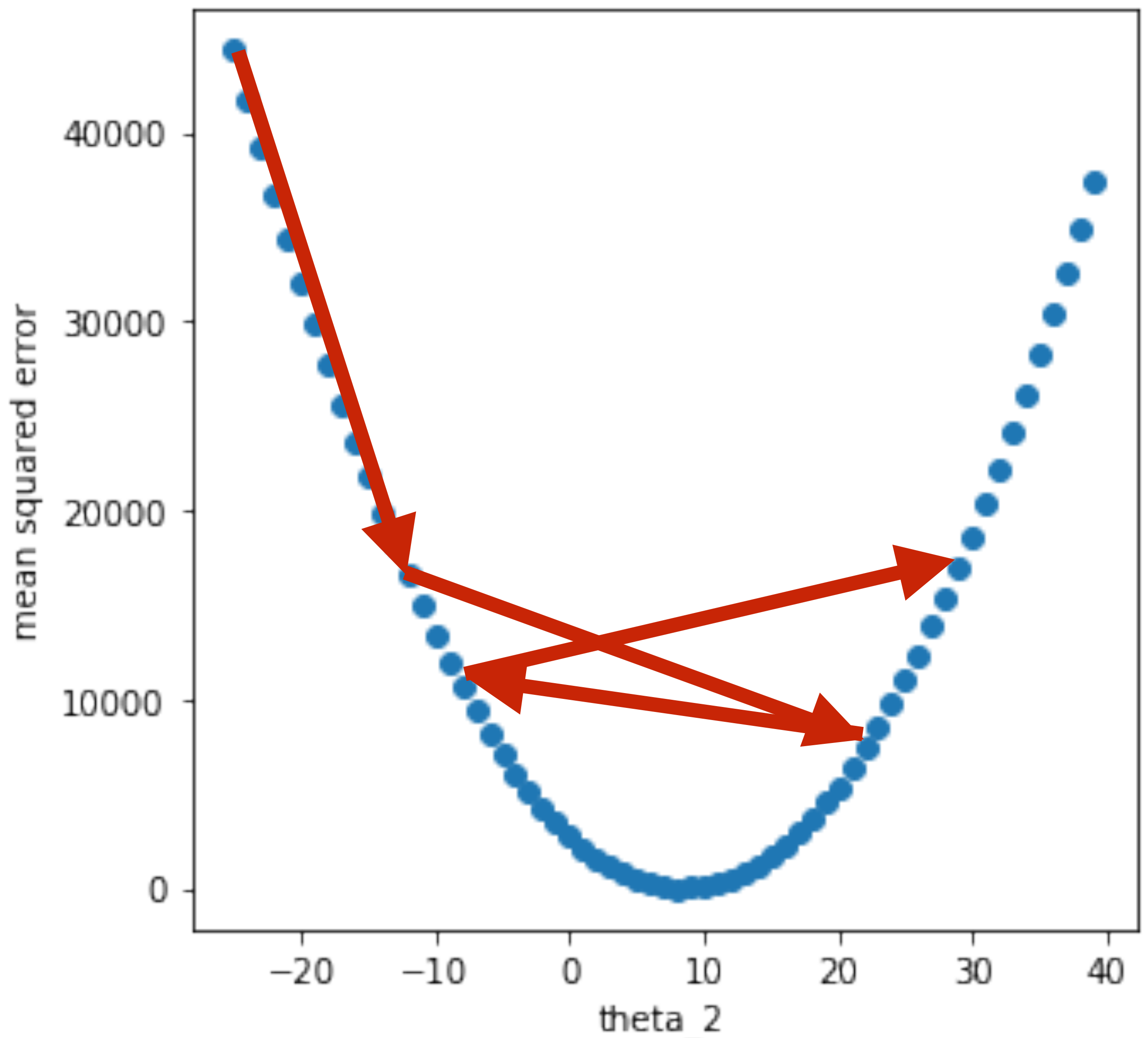


Gradient Descent

How to change **θ**?

$$\theta_i := \theta_i - \alpha \frac{\delta}{\delta \theta_j} J(\theta)$$

learning rate too big



Logistic Regression

Hypothesis

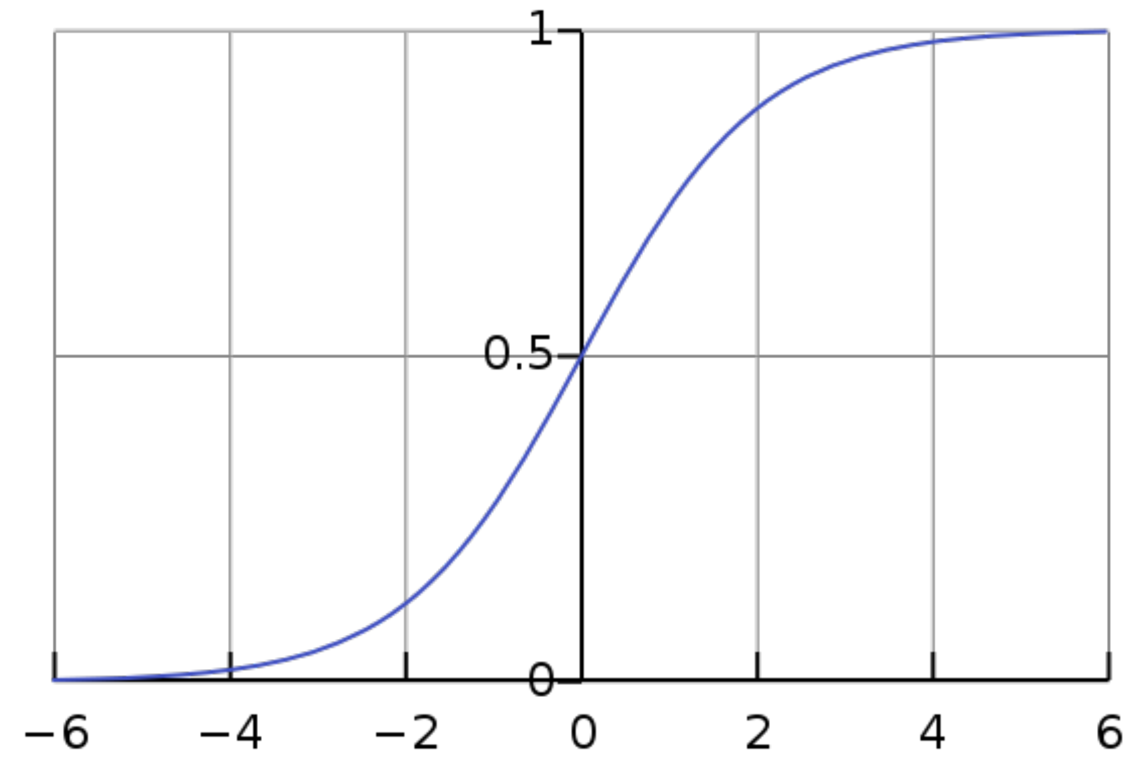
A machine learning model is a function mapping X to Y.
Θ is what the model “learned”.

$$f_\theta : X \rightarrow Y$$

Hypothesis: prices follow a linear function of the number or rooms

$$f_\theta(x) = sigmoid(\theta^T x)$$

$$sigmoid(x) = \frac{1}{1+e^{-x}}$$



Recap:
Linear Regression

$$f_\theta(x) = \theta_0 + \theta_1 x$$

$$f_\theta(x) = \theta^T x$$

Logistic Regression

Cost Function

We define a loss function to evaluate different hypotheses

$$L_{f,m}(\theta) = m(f_\theta(x), y)$$

Choosing **softmax** error as an error metric:

$$softmax(z)_i = \frac{e^{z}{}_{i}}{\sum_{j=1}^{m} e^{z_j}}$$

$$L_f(\theta) = softmax(f_\theta - y)$$

Minimize L by evaluating different Θ.