

# CSCI 567: Machine Learning

Vatsal Sharan  
Fall 2022

Lecture 1, Aug 25

# Logistics

Course website: <https://vatsalsharan.github.io/fall22.html>

- Logistics, slides, homework etc.

Ed Discussion: <https://edstem.org/>

- Main forum for communication

DEN: <https://courses.uscden.net/d2l/home/23403>

- Recordings, homework submission

## Prerequisites

This is a mathematically advanced class: that makes it more interesting!

- (1) Undergraduate level training or coursework on linear algebra, (multivariate) calculus, and basic probability and statistics;
- (2) Basic skills in programming with Python;
- (3) Undergraduate level training in the analysis of algorithms (e.g. runtime analysis).

Overview of logistics, **go through course website** for details:

**Homeworks:** 4 homeworks (groups of 2), 1 late day per student (max 1 per HW)

**Quizzes:** **10/6** and **12/1** during lecture hours (5pm-7:20pm)

**Project:** Kaggle competition (groups of 4, more details later)

**Note:** Plagiarism and other unacceptable violations

- Neither ethical nor in your self-interest
- Zero-tolerance
- Read collaboration policy on course website



Machine Learning

MACHINE LEARNING

MACHINE LEARNING EVERYWHERE

# What is ML?

*"Humans appear to be able to learn new concepts without needing to be programmed explicitly in any conventional sense. In this paper we regard learning as the phenomenon of knowledge acquisition in the absence of explicit programming."*

--- *A Theory of the Learnable*, 1984, Leslie Valiant



# What is ML?

*"Humans appear to be able to learn new concepts without needing to be programmed explicitly in any conventional sense. In this paper we regard **learning** as the phenomenon of knowledge acquisition in the absence of explicit programming."*

--- *A Theory of the Learnable*, 1984, Leslie Valiant



*"A computer program is said to **learn** from experience  $E$  with respect to some class of tasks  $T$  and performance measure  $P$ , if its performance at tasks in  $T$ , as measured by  $P$ , improves with experience  $E$ ."*

--- *Machine Learning*, 1998, Tom Mitchell



# Enormous advances in recent years

The New York Times

THE SHIFT

## We Need to Talk About How Good A.I. Is Getting

We're in a golden age of progress in artificial intelligence. It's time to start taking its potential and risks seriously.

 Give this article    608



*DALL-E 2's output when given input "infinite joy"*

New York Times, August 24, 2022

# Enormous advances in recent years

The New York Times

THE SHIFT

## We Need to Talk About How Good A.I. Is Getting

We're in a golden age of progress in artificial intelligence. It's time to start taking its potential and risks seriously.

 Give this article    608



New York Times, August 24, 2022

DALL-E 2's output when given input "infinite joy"

Look at some examples they mentioned in the article..

# Image generation: Dall-E 2

The New York Times

## Meet DALL-E, the A.I. That Draws Anything at Your Command

New technology that blends language and images could serve graphic artists — and speed disinformation campaigns.

Give this article     145



I gave the prompt:

***A digital art image of a lecture on statistical machine learning. 200 students are sitting in a classroom, hearing about linear regression.***



# Text generation: GPT-3

The New York Times

Account ▾



## ***Meet GPT-3. It Has Learned to Code (and Blog and Argue).***

The latest natural-language system generates tweets, pens poetry, summarizes emails, answers trivia questions, translates languages and even writes its own computer programs.

I gave the following prompt to a much simpler model:  
***Today 250 students at the University of Southern California begin their machine learning class. What all will they learn??***

Today 250 students at the University of Southern California begin their machine learning class.  
What all will they learn??

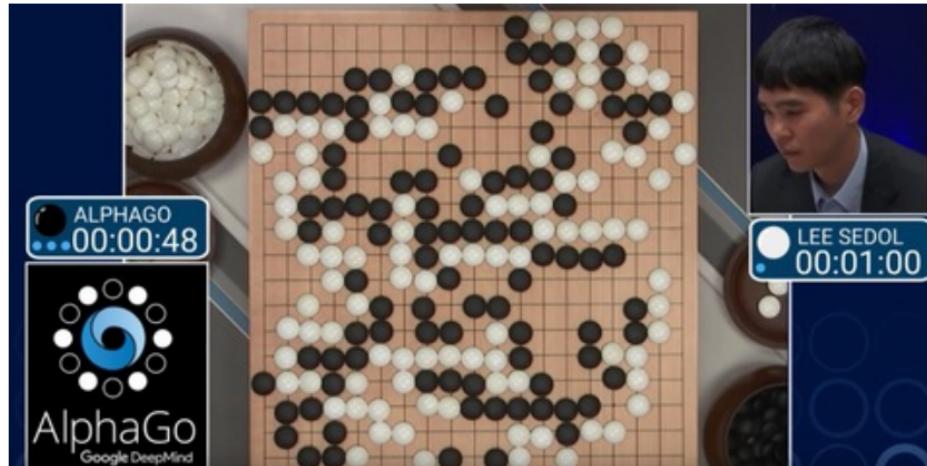
Machine learning is a fascinating subject that has been at the forefront of computer science for quite some time now. We're also seeing some new interest from those within the field. Here is one of our students here who has been working on machine learning for a number of years.

This is Dr. William Smith. He works on machine learning at the University of Southern California. He is a professor in computer science and the founder of the University of Southern California. It's a fantastic position to be part of the research community.

At the University of Southern California, you have a wide range of research interests. So, what's the most interesting research area to you?

GPT-2-simple response using <https://watt-ai.github.io/>

# Game playing: AlphaGo



# Protein folding: AlphaFold

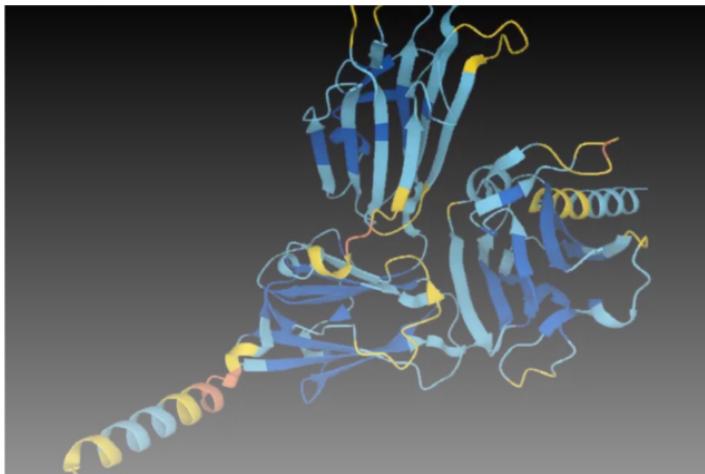
## DeepMind's protein-folding AI cracks biology's biggest problem

Artificial intelligence firm DeepMind has transformed biology by predicting the structure of nearly all proteins known to science in just 18 months, a breakthrough that will speed drug development and revolutionise basic science



TECHNOLOGY 28 July 2022

By [Matthew Sparkes](#)



Predicting the structure of proteins is one of the grand challenges of biology  
DeepMind

# Exciting time, but a lot needs to be done..

- Require significant computational resources
- Lack of understanding
- Fairness
- Robustness
- Interpretability
- Privacy
- Alignment
- ...

# This class:

- Understand the fundamentals
- Understand when ML works, its limitations, think critically

# This class:

- Understand the fundamentals
- Understand when ML works, its limitations, think critically

In particular,

- Study fundamental statistical ML methods (supervised learning, unsupervised learning, etc.)
- Solidify your knowledge with hand-on programming tasks
- Prepare you for studying advanced machine learning techniques

# A simplistic taxonomy of ML

## Supervised learning:

Aim to predict outputs of future datapoints

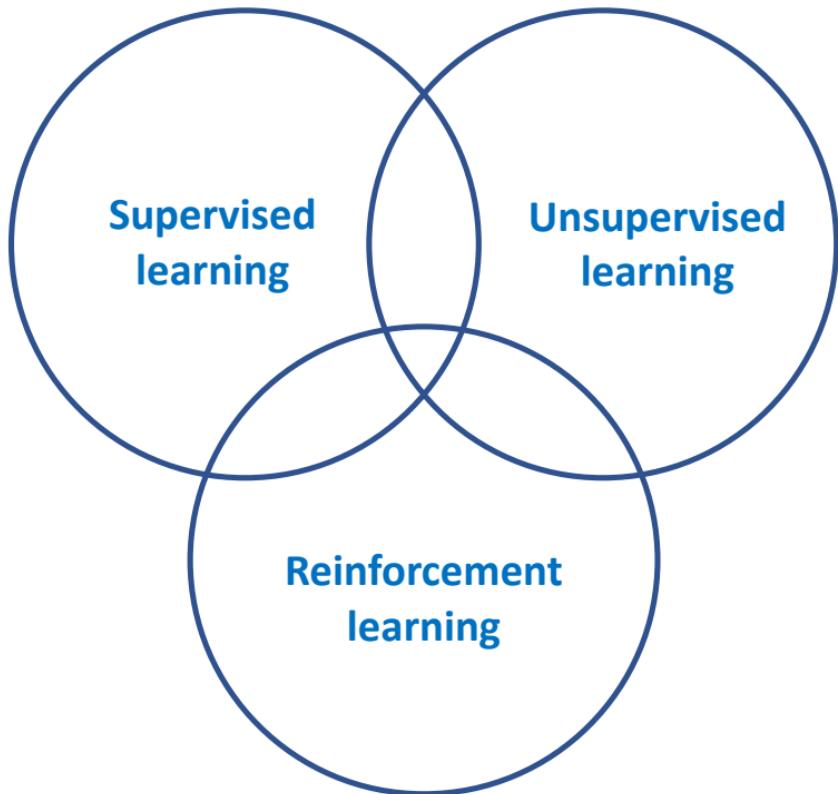
## Unsupervised learning:

Aim to discover hidden patterns and explore data

## Reinforcement learning:

Aim to make sequential decisions

# A simplistic taxonomy of ML





# Supervised Machine Learning

# Supervised ML: Predict future outcomes using past outcomes

true class = 7



true class = 2



true class = 1



true class = 0



true class = 4



true class = 1



true class = 4



true class = 9



true class = 5



Image classification

English - detected

Hindi

Welcome to our  
machine learning  
class!

हमारे मशीन लर्निंग क्लास में  
आपका स्वागत है!

hamaare masheen larning klaas mein  
apaka svaagat hai!



[Open in Google Translate](#) • [Feedback](#)

Machine translation

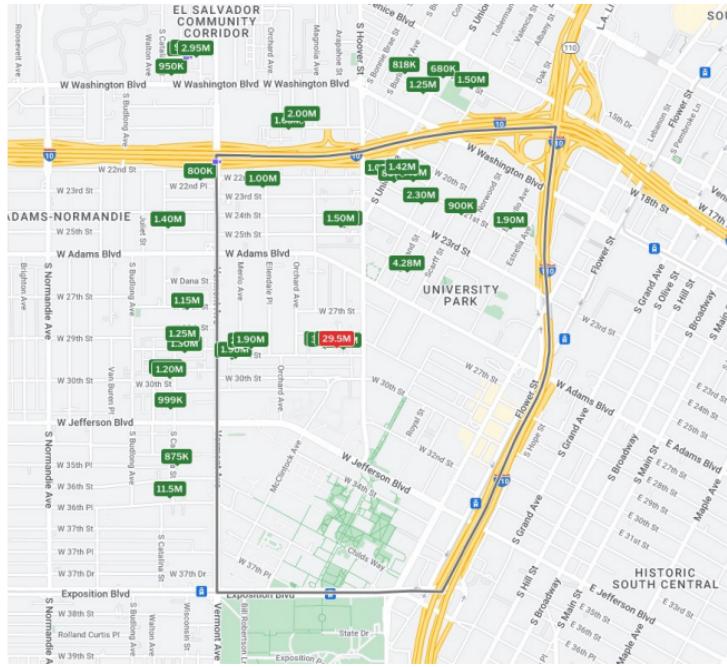
# Supervised ML: Predict future outcomes using past outcomes

The image shows a Zillow real estate listing for a house at 2640 Monmouth Ave, Los Angeles, CA 90007. The listing price is \$788,800, featuring 5 bedrooms, 2 bathrooms, and 1,944 square feet. It is marked as 'For sale by owner' with a Zestimate of \$888,500. The estimated payment is \$4,270. The page includes sections for Overview, Facts and Photos, and a prominent 'Get pre-approved' button. A red box highlights a 'What's a Zestimate?' pop-up window. This window defines the Zestimate as Zillow's best estimate of the home's market value, noting it is not an appraisal and should be used as a starting point. It also mentions that if facts are wrong, the Zestimate may be incorrect and provides a link to update them. The Zestimate is described as incorporating multiple data models and responding to factors like location. Below the listing is a map of the area, showing the house's location in University Park, Los Angeles, with nearby streets like W 23rd St, Flower St, S Grand Ave, S Highland Ave, S Main St, and W Adams Blvd.

Predicting sale price of a house

# Simplistic version: Predicting sale price of a house

Retrieve historical sales records (training data):



# Simplistic version: Predicting sale price of a house

## Features used to predict:

**3620 South BUDLONG**  
Los Angeles, CA 90007  
Status: Closed

[Overview](#) [Property Details](#) [Tour Insights](#) [Property History](#) [Public Records](#) [Activity](#) [Schools](#)



1 of 12 

Five unit apartment complex within 2 blocks of USC campus. Gate #6. Great for students (most student lessees have parents as guarantors). Most USC students live off campus, so housing units like this are always fully leased. Situated on a gated, corner lot, and across from an elementary school, this complex was recently renovated, and has in-unit laundry hook ups, wall-unit AC, and 12 parking spaces. It is within a DPS (Department of Public Safety) and Campus Cruiser patrolled area. This is a great income-generating property, not to be missed!

Property Type: Multi-Family      Style: Two Level, Low Rise  
Community: Downtown Los Angeles      County: Los Angeles  
MLS# 22176741

### Property Details for 3620 South BUDLONG, Los Angeles, CA 90007

Details provided by i-Tech MLS and may not match the public record. [Learn More](#)

#### Interior Features

##### Kitchen Information

- Remodeled
- Oven, Range

##### Laundry Information

- Inside Laundry

##### Heating & Cooling

- Wall Cooling Unit(s)

#### Multi-Unit Information

##### Community Features

- Units in Complex (Total): 5
- Multi-Family Information
  - # Leased: 5
  - # of Buildings: 1
  - Owner Pays Water
  - Tenant Pays Electricity, Tenant Pays Gas

##### Unit 2 Information

- # of Beds: 3
- # of Baths: 1
- Unfurnished
- Monthly Rent: \$2,250

##### Unit 5 Information

- # of Beds: 3
- # of Baths: 2
- Unfurnished
- Monthly Rent: \$2,325

##### Unit 1 Information

- # of Beds: 2
- # of Baths: 1
- Unfurnished
- Monthly Rent: \$1,700

##### Unit 3 Information

- Unfurnished
- # of Beds: 3
- # of Baths: 1
- Unfurnished

##### Unit 6 Information

- # of Beds: 3
- # of Baths: 1
- Monthly Rent: \$2,250

#### Property / Lot Details

##### Property Features

- Automatic Gate, Card/Code Access

##### Automatic Gate, Lawn, Sidewalks

- Tax Parcel Number: 50400017019

##### Lot Information

- Lot Size (Sq. Ft.): 9,849
- Lot Size (Acres): 0.2215
- Lot Size Source: Public Records

##### Corner Lot, Near Public Transit

##### Property Information

- Updated/Renovated
- Square Footage Source: Public Records

#### Parking / Garage, Exterior Features, Utilities & Financing

##### Parking Information

- # of Parking Spaces (Total): 12
- Parking Space
- Gated

##### Utility Information

- Green Certification Rating: 0.00
- Green Location: Transportation, Walkability
- Green Walk Score: 0
- Green Year Certified: 0

##### Financial Information

- Capitalization Rate (%): 6.25
- Actual Annual Gross Rent: \$128,331
- Gross Rent Multiplier: 11.29

#### Location Details, Misc. Information & Listing Information

##### Location Information

- Cross Streets: W 36th Pl

##### Expense Information

- Operating: \$37,864

##### Listing Information

- Listing Terms: Cash, Cash To Existing Loan
- Buyer Financing: Cash

# Simplistic version: Predicting sale price of a house

## Features used to predict:

**3620 South BUDLONG**  
Los Angeles, CA 90007  
Status: Closed

**\$1,510,000** | **14** Beds | **6** Baths  
Last Sold Price | Built: 1958 | Lot Size: 9,648 Sq. Ft. | Sold On: Jul 26, 2013

**4,418 Sq. Ft.**  
\$342 / Sq. Ft.  
Schools

**OVERVIEW** **PROPERTY DETAILS** **TOUR INSIGHTS** **PROPERTY HISTORY** **PUBLIC RECORDS** **ACTIVITY**

1 of 12

Property Type: Multi-Family  
Community: Downtown Los Angeles  
MLS# 22176741  
Style: Two Level, Low Rise  
County: Los Angeles

Five unit apartment complex within 2 blocks of USC campus. Gate #6. Great for students (most student lessees have parents as guarantors). Most USC students live off campus, so housing units like this are always fully leased. Situated on a gated, corner lot, and across from an elementary school, this complex was recently renovated, has in-unit laundry hook ups, wall-unit AC, and 12 parking spaces. It is within a DPS (Department of Public Safety) and Campus Cruiser patrolled area. This is a great income-generating property, not to be missed!

### Numeric data

#### Property Details for 3620 South BUDLONG, Los Angeles, CA 90007

Details provided by i-Tech MLS and may not match the public record. [Learn More](#)

##### INTERIOR FEATURES

###### Kitchen Information

- Remodeled
- Oven, Range

###### MULTI-UNIT INFORMATION

###### COMMUNITY FEATURES

- Units in Complex (Total): 5
- Multi-Family Information
  - # of Beds: 2
  - # of Leased: 5
  - # of Buildings: 1

###### TENANT PAY INFORMATION

- Tenant Pays Electricity, Tenant Pays Gas
- Unit 1 Information
  - # of Beds: 2
  - # of Baths: 1
  - Unfurnished
  - Monthly Rent: \$1,700

###### PROPERTY FEATURES

- Automatic Gate, Card/Code Access

###### LOT INFORMATION

- Lot Size (Sq. Ft.): 9,648
- Lot Size (Acres): 0.2215
- Lot Size Source: Public Records

###### PARKING / GARAGE, EXTERIOR FEATURES, UTILITIES & FINANCING

###### PARKING INFORMATION

- # of Parking Spaces (Total): 12
- Parking Space
- Gated

###### BUILDING INFORMATION

- Total Floors: 2

###### LOCATION DETAILS, MISC. INFORMATION & LISTING INFORMATION

###### LOCATION INFORMATION

- Cross Streets: W 36th Pl

### Free-form text

###### LAUNDRY INFORMATION

- Inside Laundry

###### HEATING & COOLING

- Wall Cooling Unit(s)

###### UNIT 2 INFORMATION

- # of Beds: 3
- # of Baths: 1
- Unfurnished
- Monthly Rent: \$2,250

###### UNIT 3 INFORMATION

- Unfurnished
- Unfurnished
- Unfurnished
- Unfurnished

###### UNIT 4 INFORMATION

- # of Beds: 3

- # of Baths: 1

- Unfurnished

###### UNIT 5 INFORMATION

- # of Beds: 3
- # of Baths: 2
- Unfurnished
- Monthly Rent: \$2,325

###### UNIT 6 INFORMATION

- # of Beds: 3
- # of Baths: 1
- Monthly Rent: \$2,250

###### TAX PARCEL NUMBER

- Tax Parcel Number: 50400017019

###### PROPERTY INFORMATION

- Updated/Renovated

- Square Footage Source: Public Records

###### FINANCIAL INFORMATION

- Capitalization Rate (%): 6.25

- Actual Annual Gross Rent: \$128,331

- Gross Rent Multiplier: 11.29

###### LISTING INFORMATION

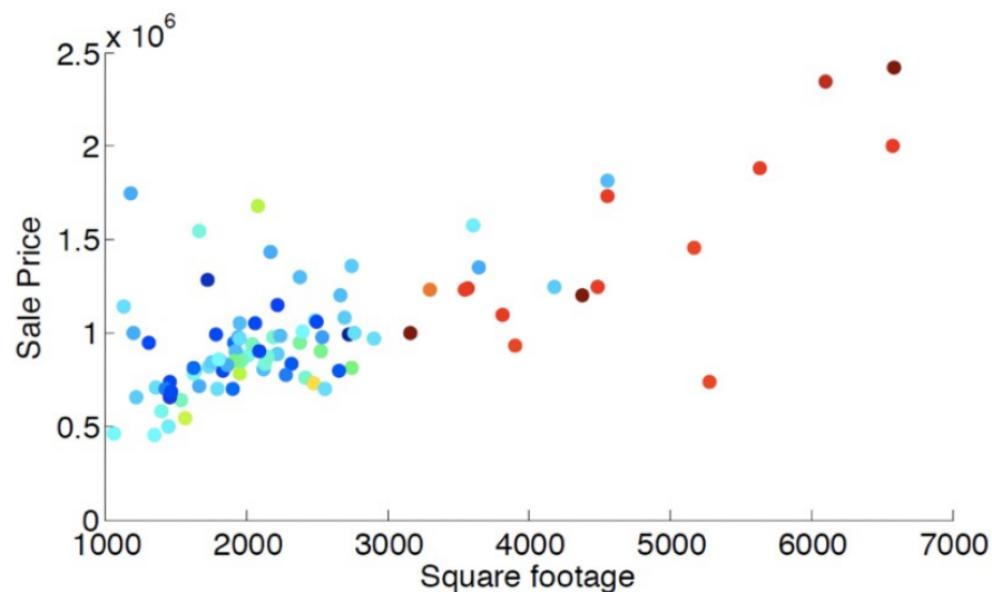
- Listing Terms: Cash, Cash To Existing Loan

- Buyer Financing: Cash

### Categorical data

## Simplistic version: Predicting sale price of a house

Correlation between square footage and sale price:



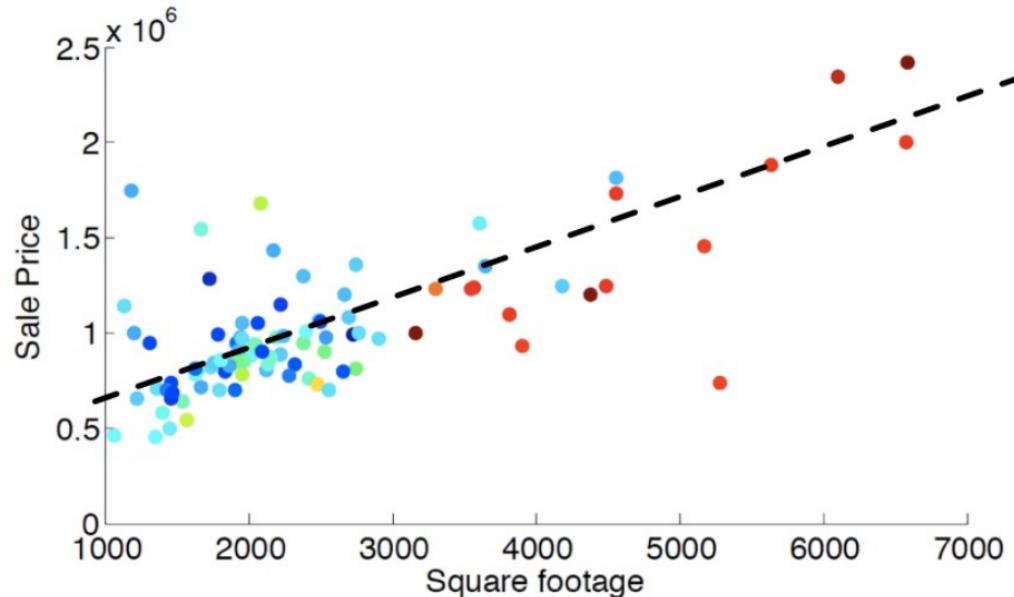
## Simplistic version: Predicting sale price of a house

Possibly linear relationship:

Sale price  $\approx$  **price per sqft**  $\times$  square footage + **fixed expense**

(*slope*)

(*intercept*)



## General framework for supervised learning

→ An **input space** :  $X \subset \mathbb{R}^d$

- \* Data points in  $d$  dimensions

- \* In previous e.g.,  $d=1$  (e.g. footage)

} Feature engineering !

→ An **output space** :  $Y$

- \*  $Y \in \mathbb{R}$  for sale price prediction

- \*  $Y \in \{\pm 1\}$  for binary classification

Goal: Learn a predictor  $f(x) : X \rightarrow Y$

Loss function :  $l(f(x), y)$ . Depends on the task.

e.g. Squared loss for  $y \in \mathbb{R}$  :  $l(f(x), y) = (f(x) - y)^2$

What to minimize over?

Minimize loss over some distribution  $D$  over instances  $(x, y)$

Definition : Risk of predictor  $f(x)$  is :

$$\begin{aligned} R(f) &= \mathbb{E}_{(x,y) \sim D} [l(f(x), y)] \\ &= \sum_{x'} \text{Prob}_D(x=x', y=y') l(f(x'), y') \end{aligned}$$

Challenge : Don't know D

\* i.i.d. assumption: We assume that we have a set of labelled instances drawn independently & identically (i.i.d.) from distribution D.

\* theoretical abstraction, often useful.

Pay attention to whether this is valid! (need "stationarity")

Definition: Given a set of labelled datapoints

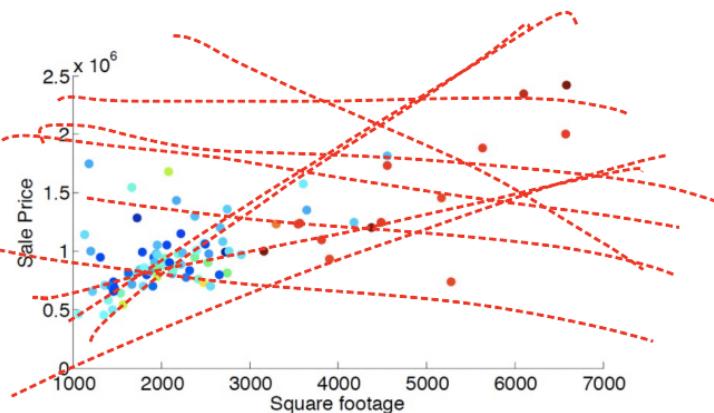
$S = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$  the empirical risk of any  $f: X \rightarrow Y$  w.r.t. S is  $\hat{R}_S(f) = (1/n) \sum_{i=1}^n l(f(x_i), y_i)$ .

## Function class

Def: A **function class** is a collection of functions  $f: X \rightarrow Y$ .

E.g.  $X = \mathbb{R}$ ,  $Y = \mathbb{R}$ ,  $f = \{f: y = w_0x + c\}$

Each of  
these is a  
linear function



The class of  
all linear functions  
is a function class.

## Empirical risk minimizer (ERM)

Def: Given a function class  $\mathcal{F} = \{f: X \rightarrow \mathbb{Y}\}$  & set of labelled datapoints  $S$ , ERM corresponds to

$$\min_{f \in \mathcal{F}} \hat{R}_S(f) = \frac{1}{n} \sum_{i=1}^n l(f(x_i), y_i)$$

## Generalization

A tautology :

$$R(f) = \hat{R}_s(f) + (R(f) - \hat{R}_s(f))$$

To Minimize  $R(f)$

- \* First try to minimize  $\hat{R}_s(f)$
- \* What's left is  $R(f) - \hat{R}_s(f)$ . This is known as the **generalization gap**.

Generalization: How well does our predictor "generalize" to new samples?

## Measuring generalization: **Training/Test paradigm**

In theory: Generalization bounds (based on "complexity" of the model)

In practice: empirical evaluation

Divide data into

**training set** - a subset of data to train model.

**test set** - a " " to test model.

Ideally: only use test set **once** (or a few times)

## Supervised learning in one slide

### Loss function:

What is the right loss function for the task?

*Depends on the problem that one is trying to solve, and on the rest...*

## Supervised learning in one slide

**Loss function:** What is the right loss function for the task?

**Representation:** What class of functions should we use?

*Also known as the “inductive bias”.*

*No-free lunch theorem from learning theory tells us that  
**no model can do well on every task***

*“All models are wrong, but some are useful”, George Box*

# Supervised learning in one slide

- Loss function:** What is the right loss function for the task?
- Representation:** What class of functions should we use?
- Optimization:** How can we efficiently solve the empirical risk minimization problem?

*Depends on all the above and also...*

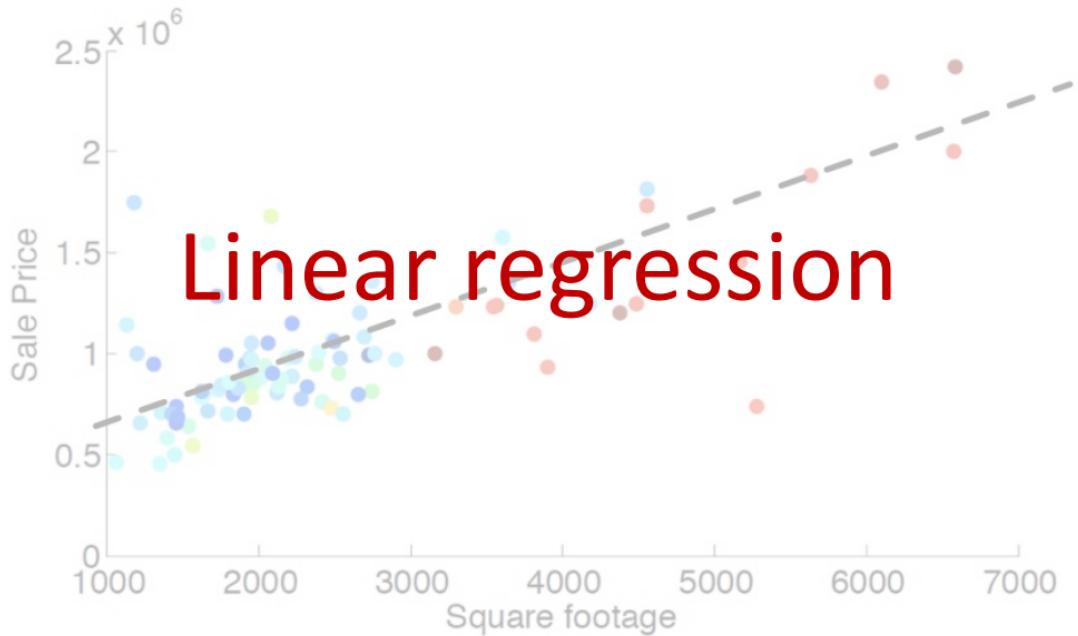
# Supervised learning in one slide

- Loss function:** What is the right loss function for the task?
- Representation:** What class of functions should we use?
- Optimization:** How can we efficiently solve the empirical risk minimization problem?
- Generalization:** Will the predictions of our model transfer gracefully to unseen examples?

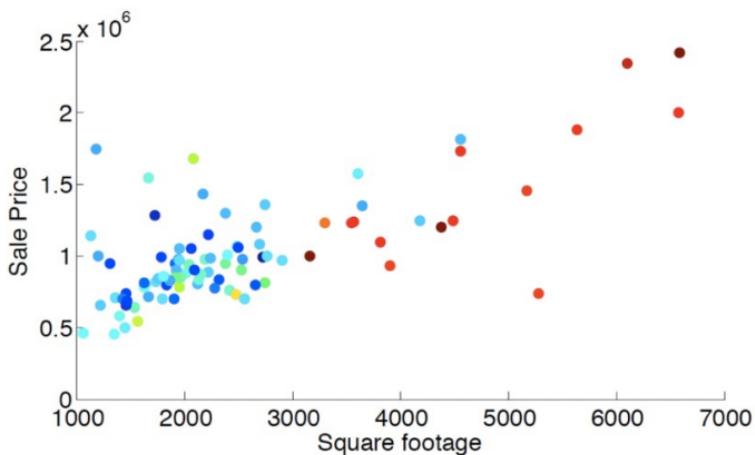
# Supervised learning in one slide

- Loss function:** What is the right loss function for the task?
- Representation:** What class of functions should we use?
- Optimization:** How can we efficiently solve the empirical risk minimization problem?
- Generalization:** Will the predictions of our model transfer gracefully to unseen examples?

*All related! And the fuel which powers everything is **data**.*



# House price prediction: the loss function



We're looking at  
real-valued outputs

\* **Squared error:**

$$(\text{prediction} - \text{sale price})^2$$
 (most common)

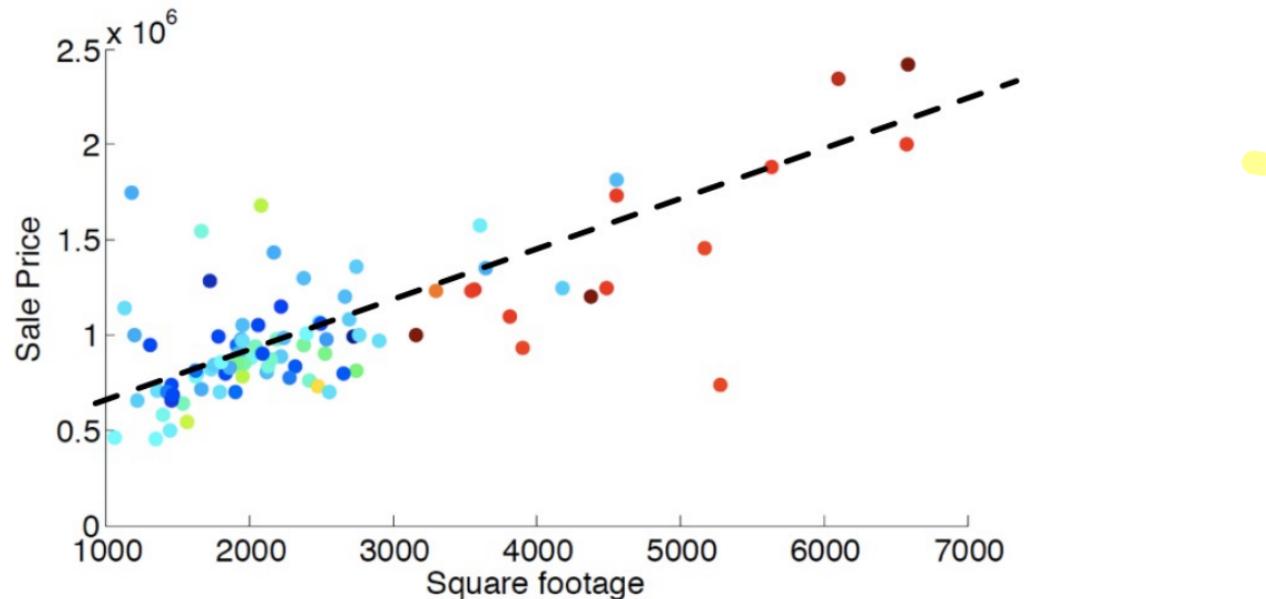
\* **Absolute error:**

$$| \text{prediction} - \text{sale price} |$$

## House price prediction: **the function class**

Possibly linear relationship:

Sale price  $\approx$  **price per sqft**  $\times$  square footage + **fixed expense**



## Linear regression

Predicted sale price = **price\_per\_sqft** × square footage + **fixed\_expense**

one model:  $\text{price\_per\_sqft} = 0.3K$ ,  $\text{fixed\_expense} = 210K$

sqft	sale price (K)	prediction (K)	squared error
2000	810	810	0
2100	907	840	$67^2$
1100	312	540	$228^2$
5500	2,600	1,860	$740^2$
...	...	...	...
Total			$0 + 67^2 + 228^2 + 740^2 + \dots$

Adjust  $\text{price\_per\_sqft}$  and  $\text{fixed\_expense}$  such that the total squared error is minimized.

## Formal setup for linear regression

Input :  $x \in \mathbb{R}^d$

Output :  $y \in \mathbb{R}$

Training data  $S = \{(x_i, y_i), i=1, \dots, n\}$

Linear model :  $f: \mathbb{R}^d \rightarrow \mathbb{R}$  with  $f(x) = w_0 + \sum_{i=1}^d w_i x_i$   
 $= w_0 + w^T x$

- \*  $w = [w_1, w_2, \dots, w_d]^T$  (weights, weight vector)
- \* bias  $w_0$

## Note: For notational convenience

Append 1 to each  $\mathbf{x}$  as first feature:  $\tilde{\mathbf{x}} = [ 1 \ x_1 \ x_2 \ \dots \ x_d ]^T$

Let  $\tilde{\mathbf{w}} = [ w_0 \ w_1 \ w_2 \ \dots \ w_d ]^T$  represent all  $d + 1$  parameters

Model becomes  $f(\mathbf{x}) = \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}$

Sometimes, we'll use  $\mathbf{w}, \mathbf{x}, d$  for  $\tilde{\mathbf{w}}, \tilde{\mathbf{x}}, d + 1$

## Goal

Minimize total squared error

$$\hat{R}_S(\tilde{w}) = \frac{1}{n} \sum_i (f(x_i) - y_i)^2 = \frac{1}{n} \sum_i (\tilde{x}_i^\top \tilde{w} - y_i)^2$$

Define (Residual sum of squares) :

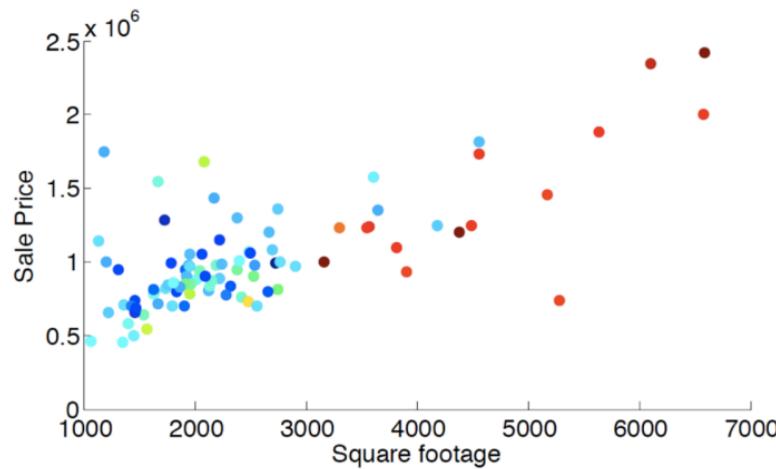
$$RSS(\tilde{w}) = n \hat{R}_S(\tilde{w}) = \sum_i (\tilde{x}_i^\top \tilde{w} - y_i)^2$$

ERM: find  $\tilde{w}^* = \underset{\tilde{w} \in \mathbb{R}^{d+1}}{\operatorname{argmin}} RSS(\tilde{w})$

Known as least squares solution

## Warmup: $d = 0$

Only one parameter  $w_0$ : constant prediction  $f(x) = w_0$



$f$  is a horizontal line, where should it be?

Warmup:  $d = 0$

$$RSS(w_0) = \sum_i (w_0 - y_i)^2$$

$$= n w_0^2 - 2 \left( \sum_i y_i \right) w_0 + \text{const}$$

$$= n \left( w_0 - \frac{1}{n} \sum_i y_i \right)^2 + \text{const}$$

(completion of squares)

$$w_0^* = \frac{1}{n} \sum_i y_i \quad (\text{the average})$$

Think about what should be the solution for  
absolute error ( $l(f(x), y) = |f(x) - y|$ )

Warmup:  $d = 1$

$$RSS(\tilde{w}) = \sum_i (w_0 + w_1 x_i - y_i)^2$$

General approach: find **stationary point**

(i.e. points with zero gradient)

$$\frac{\partial RSS(\tilde{w})}{\partial w_0} = 0 \Rightarrow \sum_i (w_0 + w_1 x_i - y_i) = 0$$

$$\frac{\partial RSS(\tilde{w})}{\partial w_1} = 0 \Rightarrow \sum_i (w_0 + w_1 x_i - y_i) x_i = 0$$

Warmup:  $d = 1$

$$\begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}$$

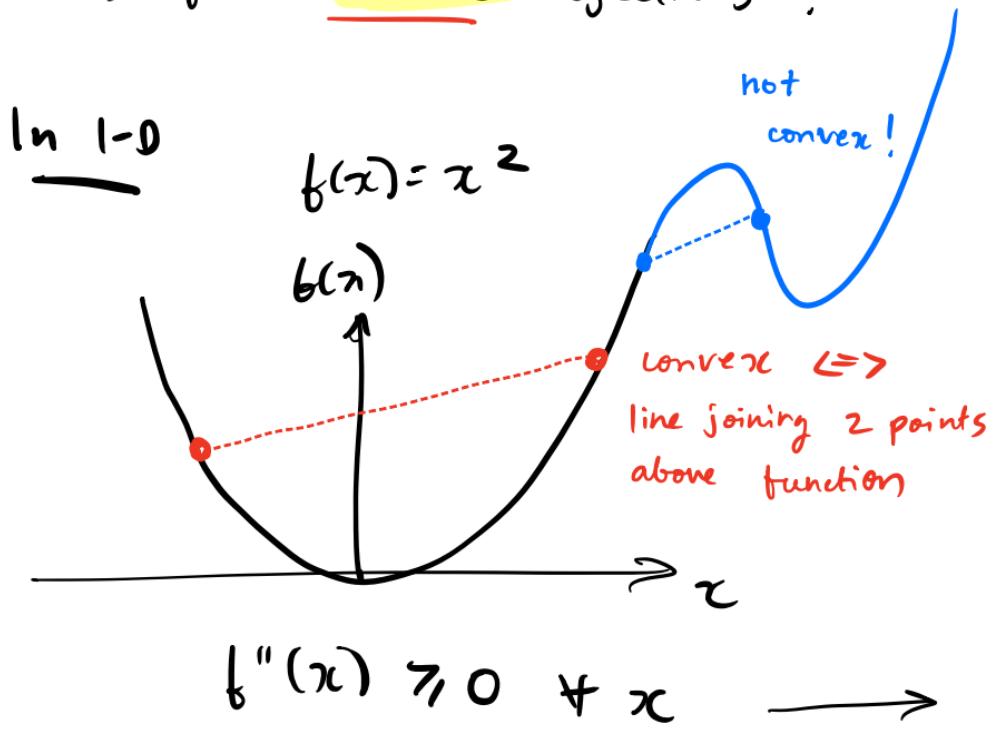
( a linear system)

$$\begin{pmatrix} w_0^* \\ w_1^* \end{pmatrix} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}$$

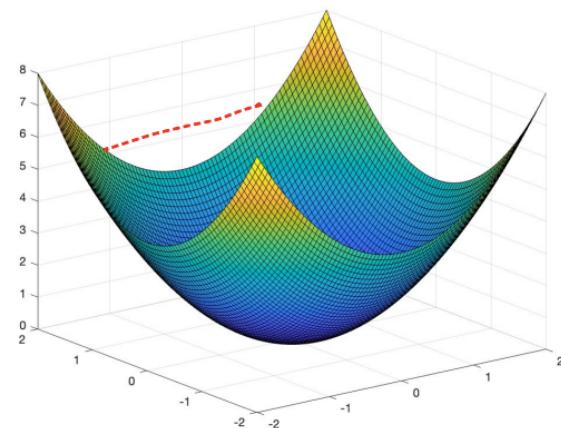
(assuming invertible)

Are stationary points minimizers?

Yes, for convex objectives !



In high dimensions:



$\nabla^2(f(x))$  is positive  
semi-definite (psd)

## General least square solution

$$RSS(\tilde{\omega}) = \sum_i (\tilde{x}_i^\top \tilde{\omega} - y_i)^2$$

Set  $\nabla RSS(\tilde{\omega}) = 0$

$$\nabla RSS(\tilde{\omega}) = 2 \sum_i \tilde{x}_i (\tilde{x}_i^\top \tilde{\omega} - y_i)$$

$$\propto \left( \sum_i \tilde{x}_i^\top \tilde{x}_i \right) \tilde{\omega} - \sum_i \tilde{x}_i y_i$$

$$= (\tilde{x}^\top \tilde{x}) \tilde{\omega} - \tilde{x}^\top y = 0$$

$$\tilde{x} = \begin{pmatrix} x_1^\top \\ x_2^\top \\ \vdots \\ x_n^\top \end{pmatrix} \in \mathbb{R}^{(n + (d+1))}$$

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n$$

$$\begin{pmatrix} \tilde{x}^\top \\ 1 \end{pmatrix} \tilde{w} = \tilde{x}^\top y \Rightarrow \tilde{w}^* = \underbrace{\begin{pmatrix} \tilde{x}^\top \\ 1 \end{pmatrix}}_{\text{invertible}}^{-1} \tilde{x}^\top y$$

## Covariance matrix and understanding LS

$$\tilde{x}^T \tilde{x} = \left( \begin{array}{cccc} | & | & | & | \\ \tilde{x}_1 & \tilde{x}_2 & \dots & \tilde{x}_n \\ | & | & & | \end{array} \right) \left( \begin{array}{c} \text{---} \\ \text{---} \\ \vdots \\ \text{---} \end{array} \right)$$

each row sums to 0

Suppose each feature is 0-mean

$\tilde{x}^T \tilde{x}$ : covariance matrix

$$\text{Suppose } \tilde{x}^T \tilde{x} = I, \quad \tilde{w}^T = \tilde{x}^T y$$

each weight is the covariance of the feature with label y. Highly correlated features have higher weights.

## Another approach

RSS is a **quadratic**, so let's complete the square:

$$\begin{aligned}\text{RSS}(\tilde{\boldsymbol{w}}) &= \sum_i (\tilde{\boldsymbol{w}}^T \tilde{\mathbf{x}}_i - y_i)^2 = \|\tilde{\mathbf{X}}\tilde{\boldsymbol{w}} - \mathbf{y}\|_2^2 \\ &= (\tilde{\mathbf{X}}\tilde{\boldsymbol{w}} - \mathbf{y})^T (\tilde{\mathbf{X}}\tilde{\boldsymbol{w}} - \mathbf{y}) \quad \text{completion of squares} \\ &= \tilde{\boldsymbol{w}}^T \tilde{\mathbf{X}}^T \tilde{\mathbf{X}} \tilde{\boldsymbol{w}} - \mathbf{y}^T \tilde{\mathbf{X}} \tilde{\boldsymbol{w}} - \tilde{\boldsymbol{w}}^T \tilde{\mathbf{X}}^T \mathbf{y} + \text{cnt.} \\ &= (\tilde{\boldsymbol{w}} - (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T \mathbf{y})^T (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}}) (\tilde{\boldsymbol{w}} - (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T \mathbf{y}) + \text{cnt.}\end{aligned}$$

**Note:**  $\mathbf{u}^T (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}}) \mathbf{u} = (\tilde{\mathbf{X}} \mathbf{u})^T \tilde{\mathbf{X}} \mathbf{u} = \|\tilde{\mathbf{X}} \mathbf{u}\|_2^2 \geq 0$  and is 0 if  $\mathbf{u} = 0$ .

So  $\tilde{\boldsymbol{w}}^* = (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T \mathbf{y}$  is the minimizer.

## Computational complexity

(running time)

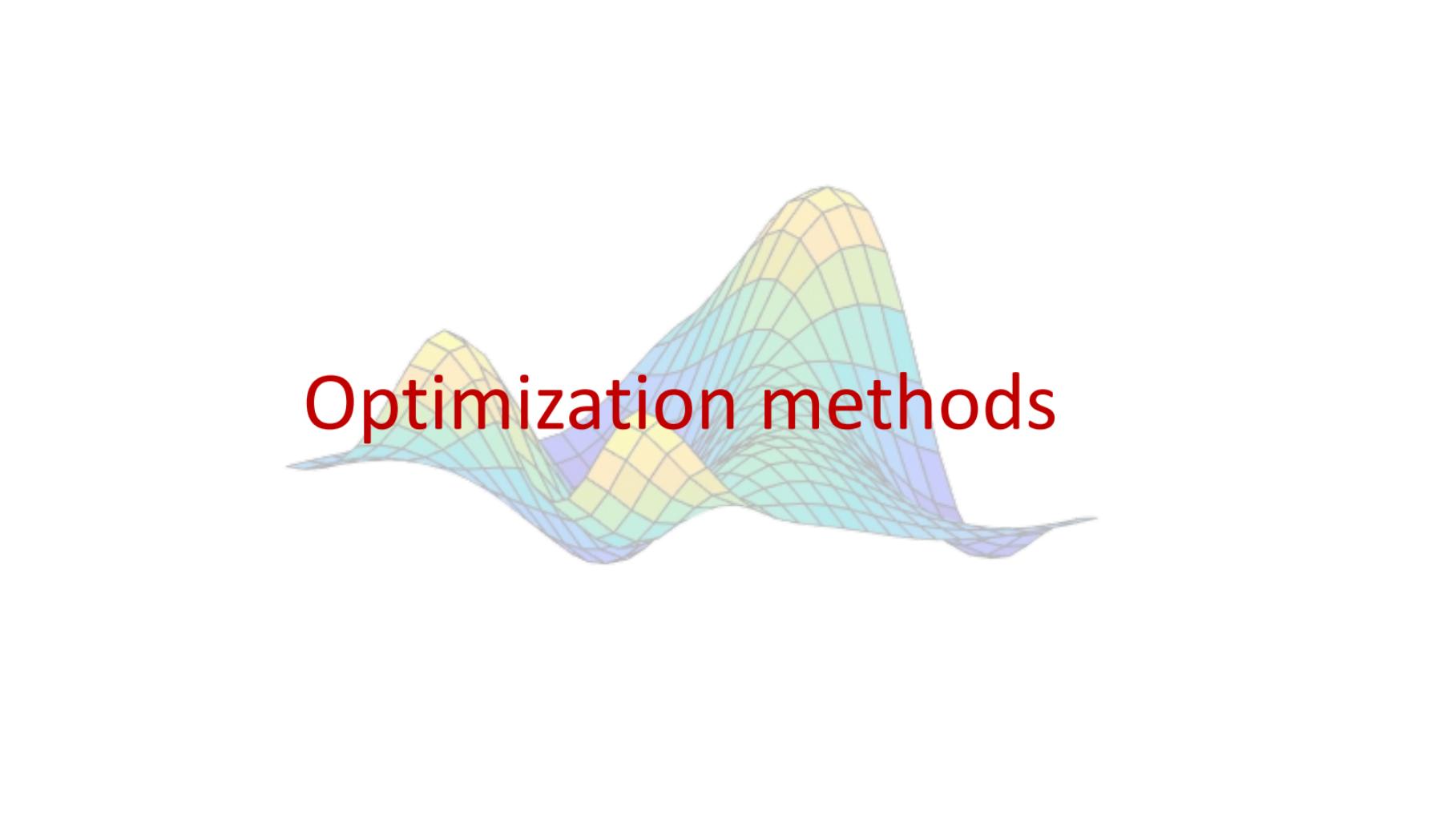
Bottleneck of computing

$$\tilde{\mathbf{w}}^* = (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T \mathbf{y}$$

is to invert the matrix  $\tilde{\mathbf{X}}^T \tilde{\mathbf{X}} \in \mathbb{R}^{(d+1) \times \mathbb{R}^{(d+1)}}$ .

Takes time  $O(d^3)$

This can be quite expensive in high dimensions



# Optimization methods

## Problem setup

Given: a function  $F(\mathbf{w})$

Goal: minimize  $F(\mathbf{w})$  (approximately)

Two simple yet extremely popular methods

**Gradient Descent (GD):** simple and fundamental

**Stochastic Gradient Descent (SGD):** faster, effective for large-scale problems

Gradient is the *first-order information* of a function.

Therefore, these methods are called *first-order methods*.

## Gradient descent

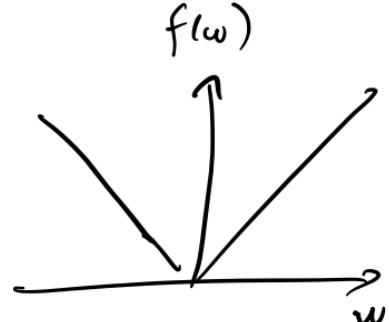
**GD:** keep moving in the *negative gradient direction*

Start with some  $w^{(0)}$ . For  $t = 0, 1, \dots, T$

$$w^{(t+1)} \leftarrow w^{(t)} - \eta \nabla F(w^{(t)})$$

where  $\eta > 0$  is called *step size / learning rate*

- in theory  $\eta$  should be set in terms of some parameters of  $f$
- in practice we just try several small values
- might need to be changing over iterations (think  $f(w) = |w|$ )
- adaptive and automatic step size tuning is an active research area



# An example

Example:  $F(\mathbf{w}) = 0.5(w_1^2 - w_2)^2 + 0.5(w_1 - 1)^2$ . Gradient is

$$\frac{\partial F}{\partial w_1} = 2(w_1^2 - w_2)w_1 + w_1 - 1 \quad \frac{\partial F}{\partial w_2} = -(w_1^2 - w_2)$$

GD:

- Initialize  $w_1^{(0)}$  and  $w_2^{(0)}$  (to be 0 or *randomly*),  $t = 0$
- do

$$w_1^{(t+1)} \leftarrow w_1^{(t)} - \eta \left[ 2(w_1^{(t)} - w_2^{(t)})w_1^{(t)} + w_1^{(t)} - 1 \right]$$

$$w_2^{(t+1)} \leftarrow w_2^{(t)} - \eta \left[ -(w_1^{(t)} - w_2^{(t)}) \right]$$

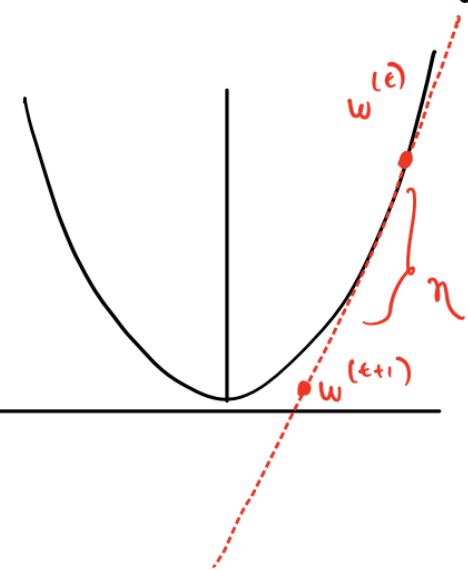
$$t \leftarrow t + 1$$

- until  $F(\mathbf{w}^{(t)})$  does not change much or  $t$  reaches a fixed number

## Why GD?

Intuition: First-order Taylor approximation

$$F(w) \approx F(w^{(t)}) + \nabla F(w^{(t)})^\top (w - w^{(t)})$$



$$\begin{aligned} F(w^{(t+1)}) &\approx F(w^{(t)}) - \eta \underbrace{\|\nabla F(w^{(t)})\|_2^2}_{>0} \\ \Rightarrow F(w^{(t+1)}) &\stackrel{\approx}{\leq} F(w^{(t)}) \end{aligned}$$

(this is only an approximation, and can be invalid if step size is too large)

# Switch to Colab

optimization.ipynb

File Edit View Insert Runtime Tools Help

+ Code + Text

▶

```
this_theta[1] = last_theta[1] - eta * grad1
theta.append(this_theta)
J.append(cost_func(*this_theta))

# Annotate the objective function plot with coloured points indicating the
# parameters chosen and red arrows indicating the steps down the gradient.
for j in range(1,N):
    ax.annotate('',
        xy=theta[j], xytext=theta[j-1],
        arrowprops={'arrowstyle': '->', 'color': 'orange', 'lw': 1},
        va='center', ha='center')
    ax.scatter(*zip(*theta), facecolors='none', edgecolors='r', lw=1.5)

# Labels, titles and a legend.
ax.set_xlabel(r'$w_1$')
ax.set_ylabel(r'$w_2$')
ax.set_title('objective function')

plt.show()
```

□\*

objective function

w<sub>1</sub>

w<sub>2</sub>