Welcome to:

EE353 Intro to Data Science and Machine Learning

EE769 Intro to Machine Learning

Course Introduction

Amit Sethi, EE (and KCDH, CMINDS, DSSE), IITB

MeDAL Lab (1st flr, EE), 3528, 7483, asethi@iitb.ac.in

Instructor Introduction

Employment: Current: IIT Bombay

Previous: IIT Guwahati,

ZS Associates, Chicago

NEC Labs, Cupertino

Research: Computational pathology, medical image analysis

Deep learning, machine learning,

Image processing, signal processing

Education: IIT Delhi, B Tech in EE

UIUC, PhD in ECE

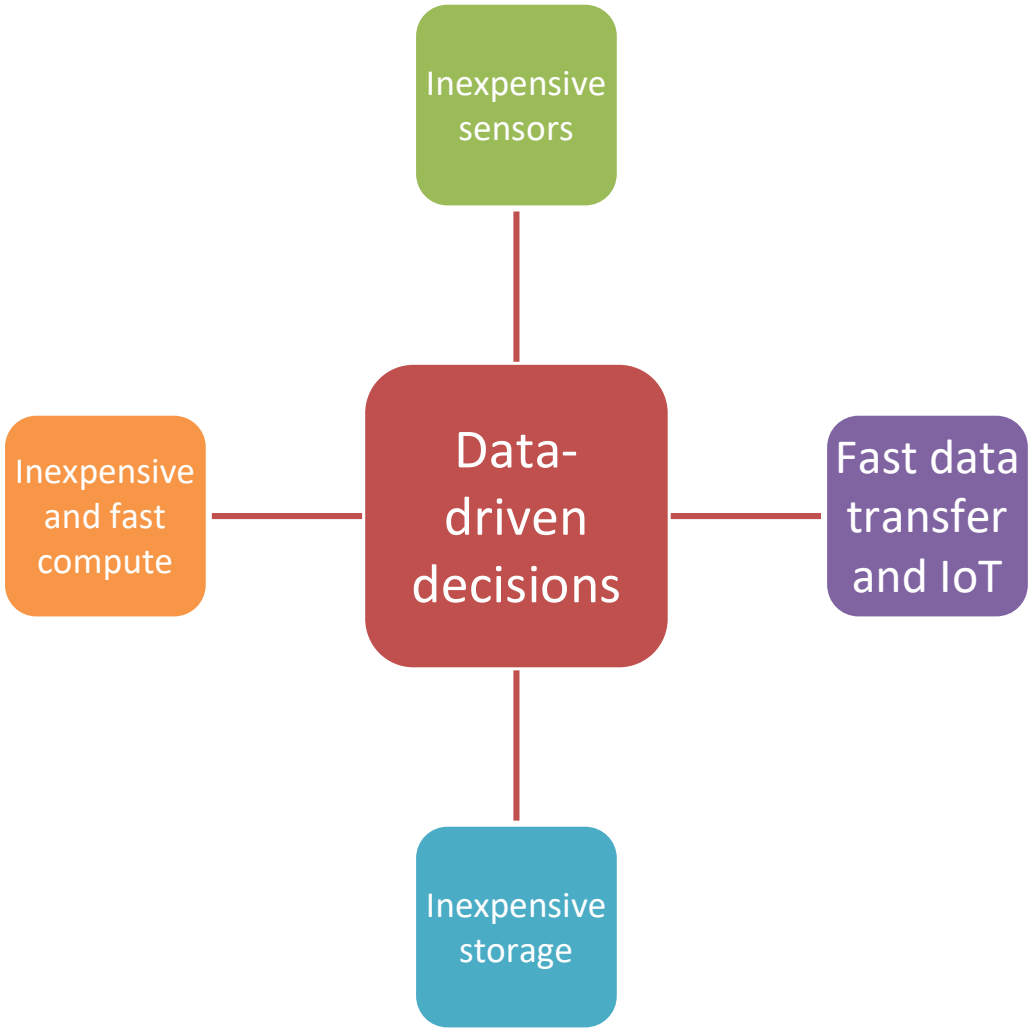
Sample research

Reverse Knowledge Distillation: Training a Large Model using a Small One for Retinal Image Matching on Limited Data SA Nasser, N Gupte, A Sethi Proceedings of the IEEE/CVF Winter Conference on Applications of Computer ...	1	2024
WaveMixSR: Resource-Efficient Neural Network for Image Super-Resolution P Jeevan, A Srinidhi, P Prathiba, A Sethi Proceedings of the IEEE/CVF Winter Conference on Applications of Computer ...		2024
Utilizing Radiomic Feature Analysis For Automated MRI Keypoint Detection: Enhancing Graph Applications SA Nasser, S Pathak, K Singhal, M Meena, N Gupte, A Chinmaya, P Garg, ... arXiv preprint arXiv:2311.18281		2023
Classification of Various Types of Damages in Honeycomb Composite Sandwich Structures using Guided Wave Structural Health Monitoring S Sawant, J Thalapil, S Tallur, S Banerjee, A Sethi arXiv preprint arXiv:2311.03765		2023
Utilizing Radiomic Feature Analysis For Automated MRI Keypoint Detection: Enhancing Graph Applications S Almahfouz Nasser, S Pathak, K Singhal, M Meena, N Gupte, ... arXiv e-prints, arXiv: 2311.18281		2023
Leveraging Segmentation to Improve Medical Image Registration SA Nasser, M Meena, G Sresth, A Sethi Authorea Preprints		2023
Combining Datasets with Different Label Sets for Improved Nucleus Segmentation and Classification A Parulekar, U Kanwat, RK Gupta, M Chippa, T Jacob, T Bameta, S Rane, ... arXiv preprint arXiv:2310.03346		2023
Domain-Adaptive Learning: Unsupervised Adaptation for Histology Images with Improved Loss Function Combination RK Gupta, S Das, A Sethi arXiv preprint arXiv:2309.17172		2023

Sample research

Transforming Breast Cancer Diagnosis: Towards Real-Time Ultrasound to Mammogram Conversion for Cost-Effective Diagnosis SA Nasser, A Sharma, A Saraf, AM Parulekar, P Haria, A Sethi arXiv preprint arXiv:2308.05449		2023
Heterogeneous graphs model spatial relationships between biological entities for breast cancer diagnosis A Krishna K, R Kant Gupta, N Cherian Kurian, P Jeevan, A Sethi arXiv e-prints, arXiv: 2307.08132		2023
WavePaint: Resource-efficient Token-mixer for Self-supervised Inpainting P Jeevan, DS Kumar, A Sethi arXiv preprint arXiv:2307.00407	1	2023
Leveraging Segmentation to Improve Medical Image Registration S Almahfouz Nasser, M Meena, G Sresth, A Sethi TechRxiv		2023
The ACROBAT 2022 Challenge: Automatic Registration Of Breast Cancer Tissue P Weitz, M Valkonen, L Solorzano, C Carr, K Kartasalo, C Boissin, ... arXiv preprint arXiv:2305.18033	2	2023
Multiscale deep learning framework captures systemic immune features in lymph nodes predictive of triple negative breast cancer outcome in large-scale studies G Verghese, M Li, F Liu, A Lohan, NC Kurian, S Meena, P Gazinska, ... The Journal of Pathology	3	2023
Quantification of subtype purity in Luminal A breast cancer predicts clinical characteristics and survival N Kumar, PH Gann, SM McGregor, A Sethi Breast Cancer Research and Treatment, 1-11	1	2023
EP178 ARTIFICIAL INTELLIGENCE BASED EOSINOPHIL COUNT IN GASTROINTESTINAL TRACT BIOPSY HC Shah, AD Amarpurkar, T Jacob, AM Parulekar, A Sethi Gastroenterology 164 (6), S-1229		2023
CHATTY: Coupled Holistic Adversarial Transport Terms with Yield for Unsupervised Domain Adaptation M Wagle, RK Gupta, A Sethi	1	2023

Why this course



- Lots of data generated and stored
- Data-driven decisions lead to better outcomes
- Same story across industries:
  - finance
  - healthcare
  - retail,
  - industrial automation,
  - environment and climate monitoring,
  - power,
  - automobiles,
  - …
- Handling data and programming are now a basic engineering skills

What is data science

- Extract useful insights from data that can be
  - Large in volume
  - Structured or unstructured
  - Captured and stored in different formats
- Using any of the following disciplines
  - Scientific method and statistics
  - Data mining and data visualization
  - Machine learning and deep learning
  - Algorithms, programming, and parallel processing

Types of data analysis

- Exploratory: what can be said about this data?
- Descriptive: does the data answer a question?
- Predictive: does the help predict something?
- Prescriptive: how can the data help us meet an objective?

Example 1: Increase company revenue

- Exploratory:
  - What data do we have on our customers?
  - Are there gaps in that data, e.g. certain seasons?
- Descriptive:
  - Do women really spend more than men?
  - Which age group spends the most?
- Predictive:
  - Can we predict when customers is ready to skip coming to our stores?
- Prescriptive:
  - Will a well-designed coupon campaign increase customer loyalty and sales?

Example 2: Increase car fuel efficiency

- Exploratory:
  - What data do we have about our car?
  - What data do we have about our customers and their driving habits?
- Descriptive:
  - What phase of the performance is crucial for sales?
- Predictive:
  - Can we predict among new fuel system design choices which ones will meet the performance and efficiency objectives?
- Prescriptive:
  - Which fuel system design choice will keep performance customers happy while fending criticism about our efficiency?

Types of data

- Structured:
  - Records with fields
  - Ordered on a grid
    - Time series
    - Images, videos, audio
    - Text
  - Relational
    - Graphs
- Unstructured

- Variables:
  - Nominal
  - Ordinal
  - Continuous
    - Floating point
    - Quantized continuous

Why make machines learn?

$Input_i \rightarrow$ **Model** $\rightarrow Output_i$

- We need models (functions, algorithms) to make predictions about inputs
- Many models are unknown and difficult to define
- Machine learning is the art and science of iteratively adjusting models based on inputs and some properties of the output

Some recent success stories

Autonomous driving

- Road recognition
- Automatic navigation

Speech recognition

- Speech to text
- Automated services over the phone

Face detection

- Facebook face tagging suggestions
- Camera autofocus for portraits

ML gives a model

- Elements of a model:
  - Input x_i
  - Function $f_\theta(x_i)$
- The model has to be useful:
  - Some notion of ideal output y_i
  - Loss $L(y_i, f_\theta(x_i), \theta)$

Good examples of ML problems

- Is a given face image of a male or female?
- Is there a Coke bottle in a given image?
- Is this image artistic?
- How can this text be improved or question answered?
- Is this customer likely to spend more later if we give her a discount now?
- Can I represent a 50-d data using 2-d?
- Can I divide my customers into logical groups?
- Can I generate music that sounds like Mozart?
- Inverse problem: source separation, super-resolution

Bad examples of ML problems

- Predict the next lunar eclipse
- Understand this text
- Should I pursue a PhD or not

Performance criteria and requirements

Performance criteria

- High accuracy
- Low risk
- More explanability
- Less memory
- Less compute

Requirements

- Good amount of data
- Clean and organized data
- Data labels and annotations
- Computational power

Learning outcomes for the course

- Translate real-world problem statements into different types of data analysis problems
- Define various data science tasks
- Demonstrate beginner-level proficiency in setting up all parts of a data analysis pipelines
- Define machine learning and cast ML problems
- Define and code popular ML algorithms
- Critique and compare ML algorithms and models

List of Topics

- Intro to random variables
- Intro to databases
- Exploratory data analysis
- Graphs and plots
- Statistical testing
- ML as a black box
- Linear regression
- Linear classification
- Regularization
- Kernelized classification
- Feature engineering
- Neural networks
- Deep neural networks
- Clustering
- Dimension reduction
- Density estimation

Prerequisites

- Basic linear algebra
  - Matrix-vector products, dot products, eigen vector definition, norms, …
- Intermediate probability
  - Continuous random variable, PDF, conditional distribution, marginalization
- Basic calculus
  - Derivatives, partial derivatives, critical points of a function
- Intermediate programming in python
  - Loops, functions, arrays, i/o, file i/o, graph plotting

Course eligibility

EE 353

- Core for EE juniors (BTech and DD)
- No other students allowed

EE 769

- Priority given to EE PGs
- Other students may be allowed, if their advisors recommend

Ineligible:

- First year and second year undergrads are not allowed
- CSE, IEOR, ME, CSRE students (except guide's reco)
- Any student who has done or is doing an equivalent (e.g. CS 419, 337, 725, 747, GNR 652, IE 506, 611, ME 781...) or advanced IITB course on ML (e.g. CS 726, 729, IE 643, 663, 712, GNR 638, EE 782...)

Tentative evaluation plan

Item	Wt.
Participation	10
Assignments (4)	40
Mid-sem exam	20
End-sem exam	30
Total	**100**

Min. marks	Grade
90	AA
80	AB
70	BB
60	BC
50	CC
40	CD
30	DD/AU

Evaluation items

- Precise understanding of concepts
- Express concepts mathematically
- Make basic mathematical derivations
- Program diligently
- Design experiments diligently
- Interpret results
- **Zero tolerance for academic malpractice**

Allowed vs. not-allowed

Allowed

- Exams
  - Notes
- Assignments
  - Discuss the assignment before starting a portion with friends
  - Consult code on the internet or LLMs
  - Disclosing sources of “inspiration,” indicating the exact lines copied and modified

Not allowed

- Exam
  - Open the internet
  - Communicate with others
- Assignments
  - Copy code from friend or internet and make trivial changes
  - Not disclosing sources of “inspiration” and the exact lines copied

How to get the most out of this course

- Attend lectures and take notes
- Read books and internet resources
- Attempt problems
- Discuss with classmates offline and online
- Become comfortable with programming
- Ask TAs and instructors any remaining doubts

Resources

- People:
  - Instructor: Amit Sethi (Chat on MS Teams, tag in channel)
  - TAs: TBA
- Office hours (Wednesdays 6pm to 7pm):
  - Amit Sethi's office, easternmost on second floor, EE building (7483)
  - MeDAL lab, first floor, EE building (3528)
- LMS
  - MS Teams code **yggbq3k** (Official channel for announcements and material, General channel for discussions and queries)
  - Moodle for assignment submissions and grades
- Books
  - Pattern Recognition and Machine Learning, by Christopher Bishop

Emergency absence policy

- Doctor’s note or documentation essential
- Assignments: extension of deadline
- Exams: Extrapolation of your performance based on your percentile (not percent)

Resume verification

- Sorry, I cannot verify resumes for such a large class
- Make a GitHub Repo, and upload your assignment and project reports on ArXiv for the recruiters to see

Immediate tasks

- **Python:**
  - https://www.learnpython.org/
- **Numpy and Google CoLab: (Basic python, numpy, pandas, matplotlib, scikit-learn)**
  - https://cs231n.github.io/python-numpy-tutorial/
- **Linear algebra and calculus:**
  - https://stanford.edu/~shervine/teaching/cs-229/refresher-algebra-calculus
- **Probability:**
  - https://stanford.edu/~shervine/teaching/cs-229/refresher-probabilities-statistics

EE769 Introduction to ML
Basic Mathematics for ML

Amit Sethi, EE, IITB

asethi, 7483

Objective

- Revise and gain comfort with the following:
  - Vectors and matrices
  - Calculus and convex optimization
  - Probability and statistics

Scalar-vector operations

$a \in \mathbb{R}$

$\bar{x} \in \mathbb{R}^{N \times 1}$

$$\bar{x} = \begin{bmatrix} 100 \\ -100 \\ 22 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \in \mathbb{R}^{3 \times 1}$$

$\mathbb{R}^3$

$$a\bar{x} = \begin{bmatrix} a x_1 \\ a x_2 \\ a x_3 \end{bmatrix}$$

$$\bar{x} + a = \begin{bmatrix} x_1 + a \\ x_2 + a \\ x_3 + a \end{bmatrix}$$

$$\bar{x} + a\mathbb{1} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + a \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

$N \times 1$ $N \times 1$

Scalar-vector operations

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \qquad x_1 \text{ scalar}$$

$$\mathbf{x} \in \mathbb{R}^3$$

$$a\mathbf{x} = \begin{bmatrix} a x_1 \\ a x_2 \\ a x_3 \end{bmatrix}$$

$$a \in \mathbb{R}$$

$\mathbf{x} + a$ — Not a math function

$$\equiv \begin{bmatrix} x_1 + a \\ x_2 + a \\ x_3 + a \end{bmatrix} = \mathbf{x} + a \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

$$= \mathbf{x} + a\mathbb{1}$$

($\mathbf{x}$: vector, a: scalar, $\mathbb{1}$: vector)

Vector-vector operations

$$\underset{\in \mathbb{R}^N}{\bar{x}} + \underset{\in \mathbb{R}^N}{\bar{y}} = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \end{bmatrix}$$

$$\bar{x} \cdot \bar{y} = \bar{x}^T \bar{y} = \langle x, y \rangle = \underbrace{x_1 y_1 + x_2 y_2}_{\text{Scalar}}$$

$$\underset{\in \mathbb{R}^{1\times 2}}{\begin{bmatrix} x_1 & x_2 \end{bmatrix}} \quad \underset{\in \mathbb{R}^{2\times 1}}{\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}}$$

$$\underset{N\times 1}{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}} \odot \underset{N\times 1}{\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}} = \underset{N\times 1}{\begin{bmatrix} x_1 y_1 \\ x_2 y_2 \end{bmatrix}}$$

Vector-vector operations

$$x + y = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ x_3 + y_3 \end{bmatrix}$$

x & y have the same dimension

$$x \cdot y = x_1 y_1 + x_2 y_2 + x_3 y_3 = x^T y$$

$$= \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \langle x, y \rangle$$

$$x \odot y = \begin{bmatrix} x_1 y_1 \\ x_2 y_2 \\ x_3 y_3 \end{bmatrix} \in \mathbb{R}^3$$

Sub-spaces spanned by vectors

$\forall \bar{v}_1, \bar{v}_2 \in V$

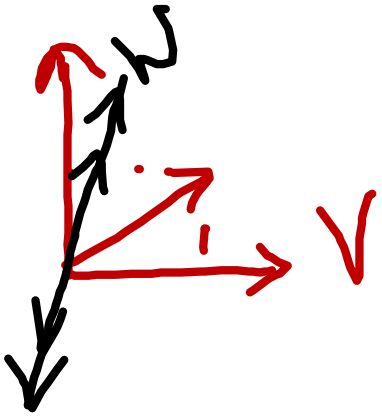
① $\bar{v}_1 + \bar{v}_2 \in V$

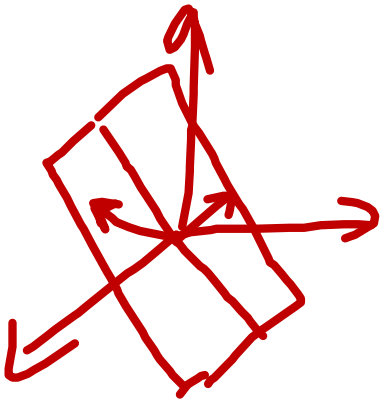
② $c\bar{v}_1 \in V$

$\forall \bar{w}_1, \bar{w}_2 \in W \subset V$

① $\bar{w}_1 + \bar{w}_2 \in W$

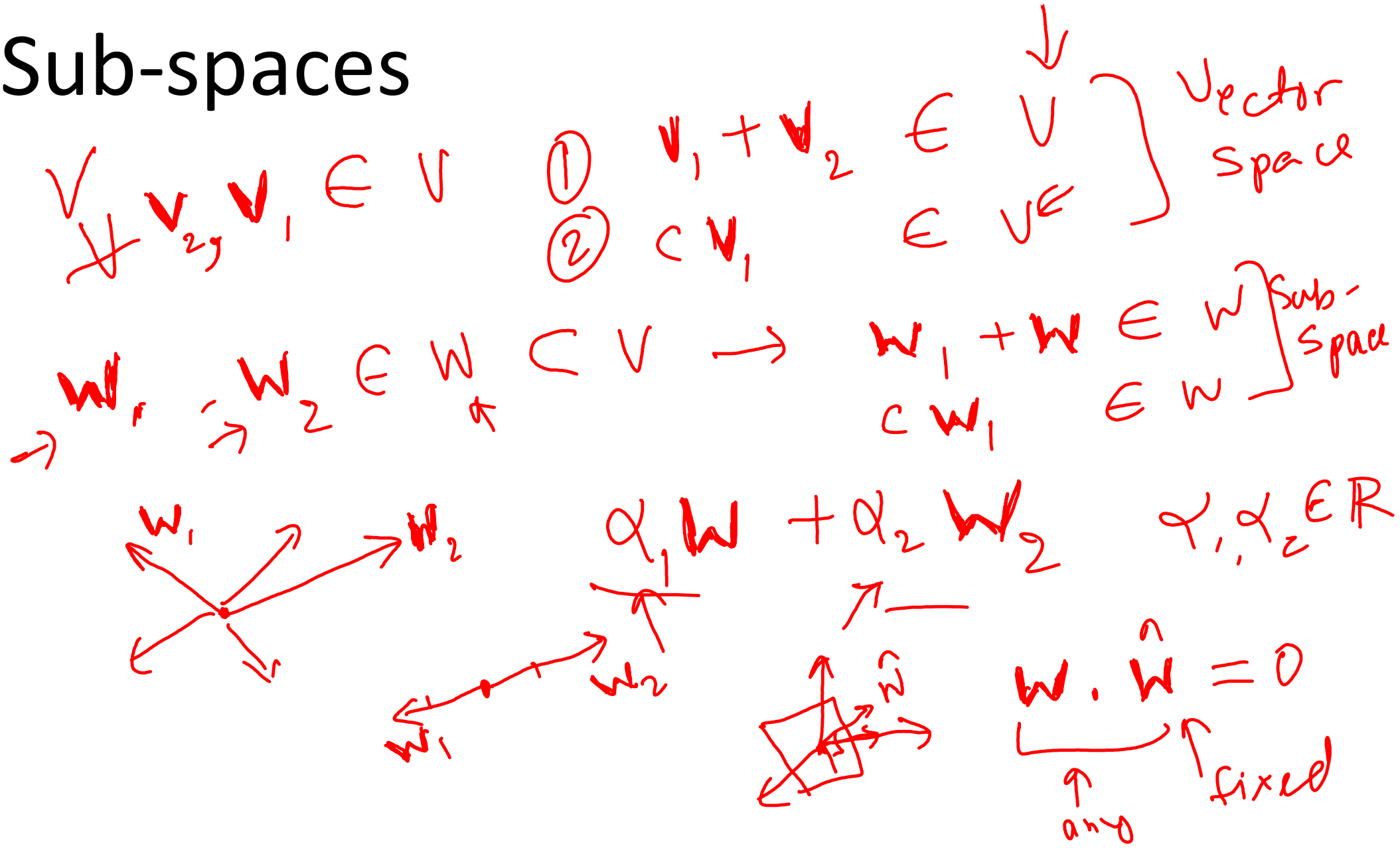
② $c\bar{w}_1 \in W$





$\alpha_1 \bar{w}_1 + \alpha_2 \bar{w}_2$

Sub-spaces


V
$\vec{v}_2, \vec{v}_1 \in V$
① $\vec{v}_1 + \vec{v}_2 \in V$
② $c\vec{v}_1 \in V$
Vector space
$\vec{w}_1, \vec{w}_2 \in W \subset V \rightarrow \vec{w}_1 + \vec{w} \in W$
$c\vec{w}_1 \in W$
sub-space
$\vec{w}_1$
$\vec{w}_2$
$\alpha_1 \vec{w} + \alpha_2 \vec{w}_2$
$\alpha_1, \alpha_2 \in \mathbb{R}$
$\hat{n}$
$\vec{w} \cdot \hat{n} = 0$
any
fixed

Matrix-matrix operations

$X \quad Y \in \mathbb{R}^{M \times N}$ ← Rows Col.s $\quad x \in \mathbb{R}^{N} \quad x \in \mathbb{R}^{N \times 1}$

$X \in \mathbb{R}^{M \times N}$

$$X + Y = \begin{bmatrix} (x_{11}+y_{11}) & \cdots & (x_{1N}+y_{1N}) \\ \vdots & & \\ (x_{M1}+y_{M1}) & \cdots & (x_{MN}+y_{MN}) \end{bmatrix}$$

$X \in \mathbb{R}^{M \times N}$, $Z \in \mathbb{R}^{N \times P}$

$$XZ = \begin{bmatrix} (x_{11}z_{11} + \cdots x_{1N}z_{N1}) & \cdots & \\ \cdots & & (x_{M1}z_{1P} + \cdots x_{MN}z_{NP}) \end{bmatrix}$$

$$X \odot Y \quad \begin{bmatrix} (x_{11}y_{11}) & \cdots (x_{MN}y_{MN}) \end{bmatrix}$$

Transpose, determinant, and inverse of a matrix

$$X^T = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{M1} \\ \vdots & \ddots & x_{ji} & \cdots \\ x_{1N} & \cdots & \cdots & x_{MN} \end{bmatrix} \in \mathbb{R}^{N \times M}$$

$$X \in \mathbb{R}^{2\times 2} \quad X = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \qquad \det(X) = ad - bc$$

$$X X^{-1} = I = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & 0 & \ddots & 1 \end{bmatrix} \qquad X^{-1} = \frac{1}{|X|} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

$$3 \times 3^{-1} = 1$$

$$\begin{bmatrix} 1 & -2 \\ 3 & 4 \end{bmatrix} \times \begin{bmatrix} 0.4 & 0.2 \\ -0.3 & 0.1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Rank of a matrix

$$\bar{X} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \in \mathbb{R}^{2\times 3}$$

$\text{Rank}(\bar{X}) \leq \min(M, N)$

$$\bar{Y} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 8 & 12 \end{bmatrix}$$

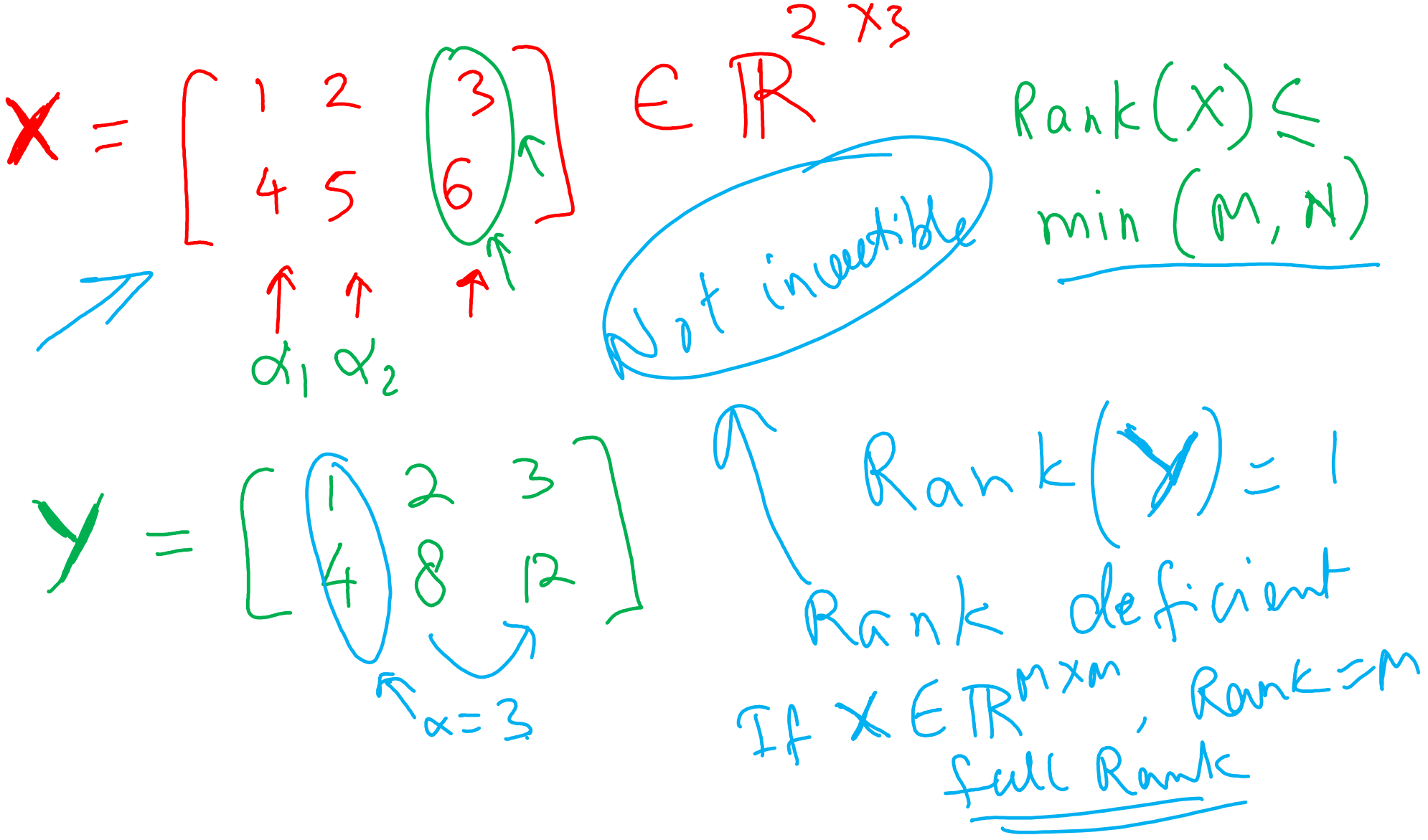
$\text{Rank}(\bar{Y}) = 1$

Rank deficient matrix

$$\bar{X} = \begin{bmatrix} 1 & 2 \\ 4 & 5 \end{bmatrix}$$ full Rank

$$\bar{Y} = \begin{bmatrix} 1 & 2 \\ 4 & 8 \end{bmatrix}$$ Rank deficient

Rank of a matrix



Pseudo-inverse of a matrix

$$\underset{M\times N}{\bar{\bar{X}}} \qquad \underset{N\times M}{\bar{\bar{X}}^{+}} = (\bar{\bar{X}}^{H}\bar{\bar{X}})^{-1}\bar{\bar{X}}^{H}$$

Conjugate transpose

$$\bar{\bar{X}}^{+}\bar{\bar{X}} = \left[(\bar{\bar{X}}^{H}\bar{\bar{X}})^{-1}\bar{\bar{X}}^{H}\right]\bar{\bar{X}} = I_{N\times N}$$

Pseudo-inverse of a matrix

X $\quad X^{+}$ $\quad (XX^{+})X = X$

$X^{+} = (X^{H}X)^{-1} X^{H}$

$M \times N$

$N \times M$

$N \times N$

Conjugate transpose

Eigen Decomposition

$\bar{A}\ \bar{v}_i = \lambda_i \bar{v}_i$ $\quad \bar{v}_i \rightarrow$ eigenvector of $\bar{A}$

$N \times N \quad N \times 1$ $\quad i \in \{1, \ldots, N\}$ $\quad \lambda_i \rightarrow$ eigenvalue of $\bar{A}$

$$\bar{A} = \bar{Q}\ \bar{\Lambda}\ Q^{-1}$$

$$\|v_i\|_2^2 = 1 \Leftarrow$$

$$\bar{Q} = [\bar{v}_1 \cdots \bar{v}_N]$$

$$\bar{\Lambda} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_N \end{bmatrix}$$

$$v_i^T v_j = \delta_{ij}$$

$$\delta_{ij} = 0, \quad i \neq j \Leftarrow$$

$$= 1, \quad i = j$$

Eigen Decomposition

$A v_i = \lambda_i v_i$

A: $N \times N$, v_i: $N \times 1$, result: $N \times 1$

λ_i: scalar, v_i: $N \times 1$

v is an eigen vector of A

λ " " " value of A

v_i is a unit vector

$A = Q \Lambda Q^{-1}$

$Q = [v_1 \; v_2 \; v_3 \cdots v_N]$

Λ → Diagonal matrix

$$\begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \cdots \lambda_N \end{bmatrix}$$

Tensors

$\underline{\alpha} \in \mathbb{R}$; $\underline{\bar{x}} \in \mathbb{R}^{N}$; $\underline{\bar{\bar{Y}}} \in \mathbb{R}^{M \times N}$

$\underline{\underline{T}} \in \mathbb{R}^{M \times N \times P \times \cdots Q}$

transpose(T, [3, 1, 2, 0])

0, 1, 2, 3

0, 1

1, 0

Tensors

$\alpha \in \mathbb{R}$

$\mathcal{X} \in \mathbb{R}^{N}$

$\underset{0}{M} \times \underset{1}{N} \times \underset{2}{P} \times \underset{3}{Q}$

$\mathbf{A} \in \mathbb{R}^{M \times N}$

$\mathcal{T} \in \mathbb{R}$

$R = \text{transpose}(T, [3,1,2,0])$

order of permutation of dimensions

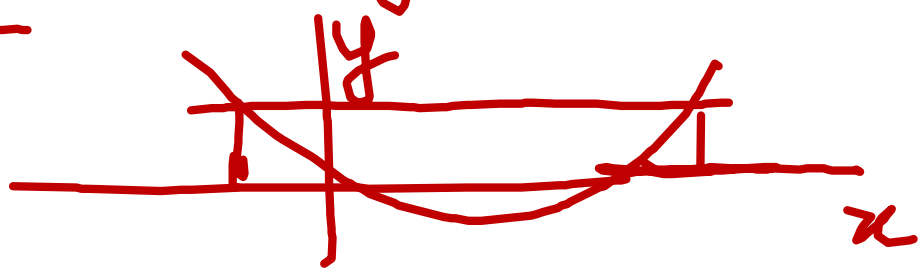
Functions

$f : X \to Y$ $\quad x \in X$ $\quad f(x) \in Y$

$y = f(x) \in Y$

many - to - one

$y = ax^2 + bx + c$



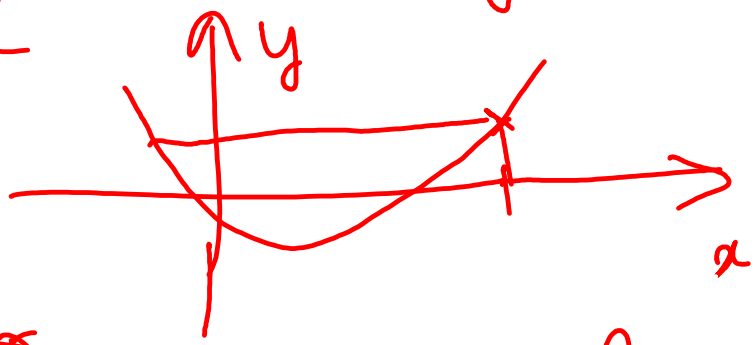
$x^2 + y^2 = 1 \Rightarrow y = \pm\sqrt{1-x^2}$ ✗ Not a function

Functions

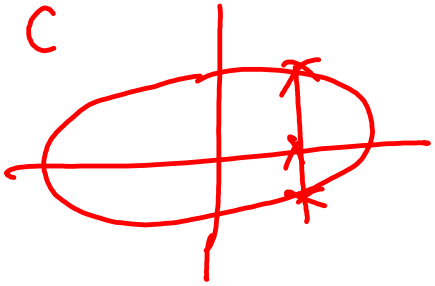
$f : X \to Y$ $\quad x \in X \quad f(x) \to y$

many to one

$f(x) = ax^2 + bx + c$



$ax^2 + by^2 = c^2$



Not a function

Continuity

$y = \sqrt{x}$ $y = \sin(x)$ $y = x^2$

x is continuous

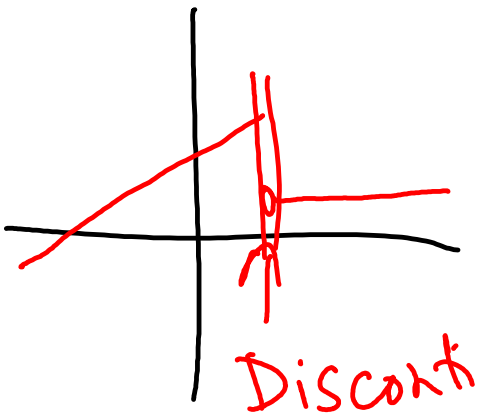
Lipschitz Continuity

If $f(x + \Delta x) = f(x)$

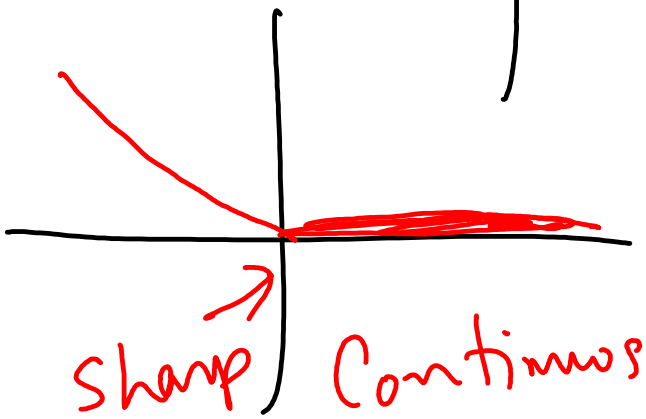
$\Delta x \to 0$

$|f(x_1) - f(x_2)| \leq K |x_1 - x_2|$

+K -K



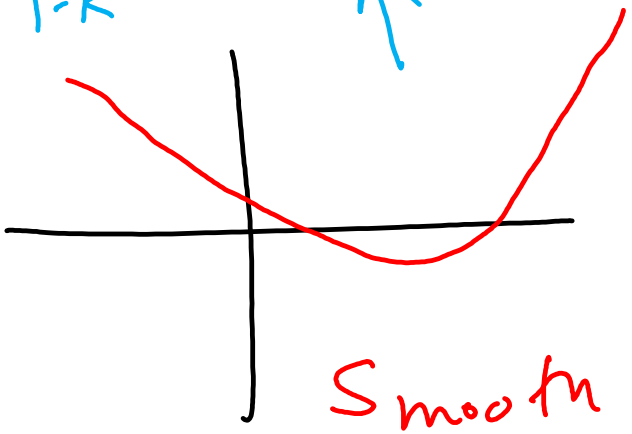
Discontinuous



Sharp corner

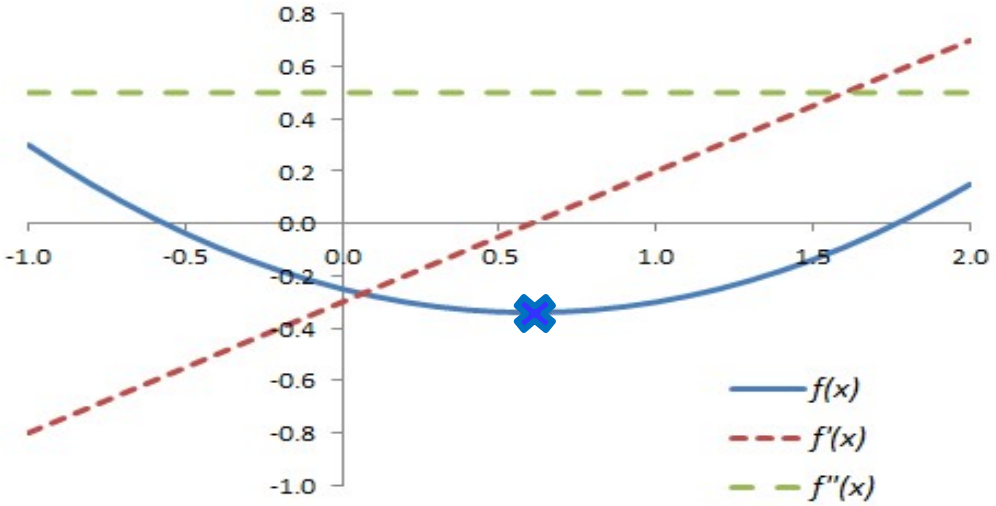
Continuous

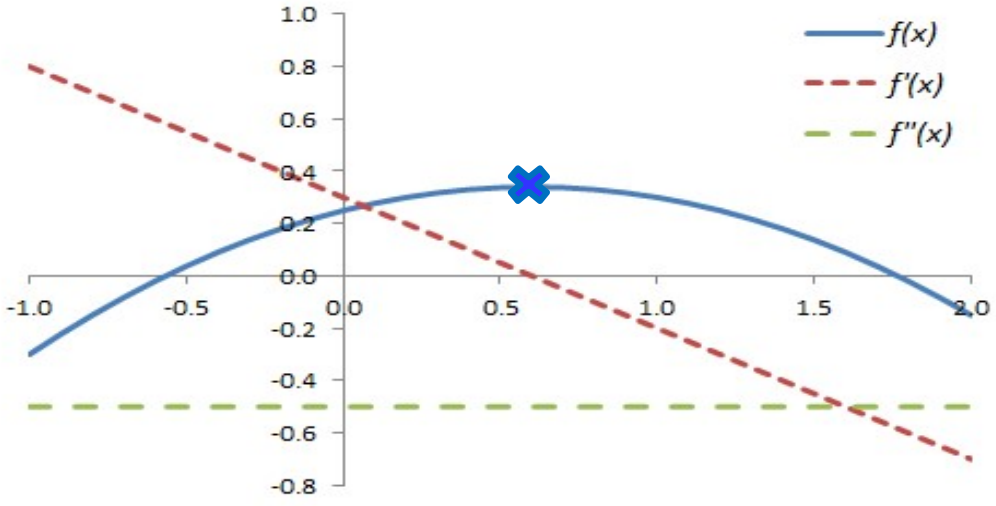
Not Smooth



Smooth & continuous

Derivative of a function of a scalar





E.g. $f(x) = ax^2 + bx + c, \quad f'(x) = 2ax + b, \quad f''(x) = 2a$

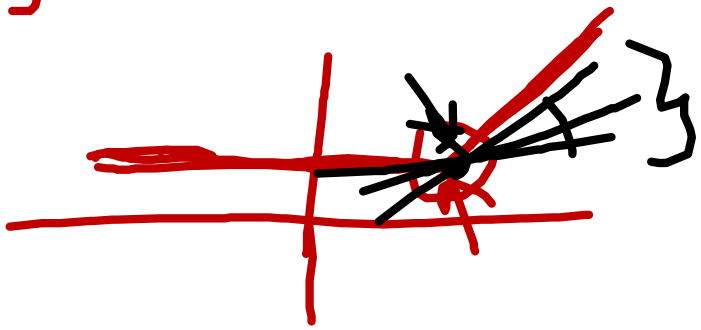
- Derivative $f'(x) = \frac{d\,f(x)}{d\,x}$ is the rate of change of $f(x)$ with x
- It is zero when then function is flat (horizontal), such as at the minimum or maximum of $f(x)$
- It is positive when $f(x)$ is sloping up, and negative when $f(x)$ is sloping down
- To move towards the maxima, taking a small step in a direction of the derivative

Derivative of a function

$$\lim_{\Delta x \to 0} f(x+\Delta x) = f(x) \quad \forall x$$

Continuous $f(x)$

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x}$$



Sub-tangent

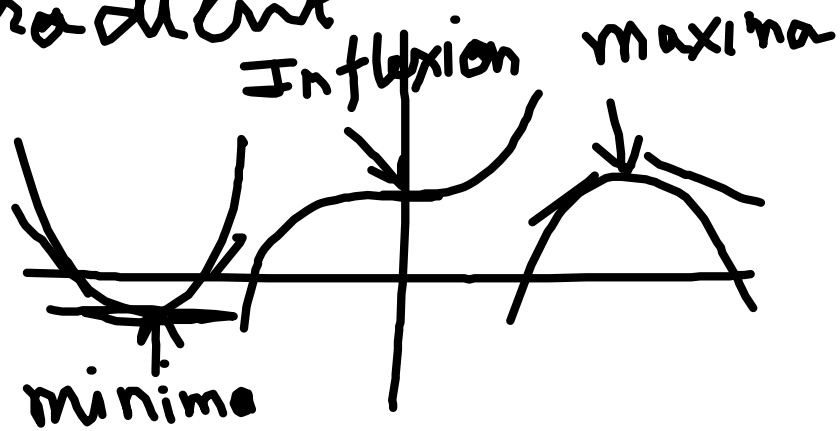
Sub-gradient

$f'(x) = 0 \rightarrow$ critical point

$f''(x) > 0 \rightarrow$ min

$< 0 \rightarrow$ max

$= 0 \rightarrow$ inflexion

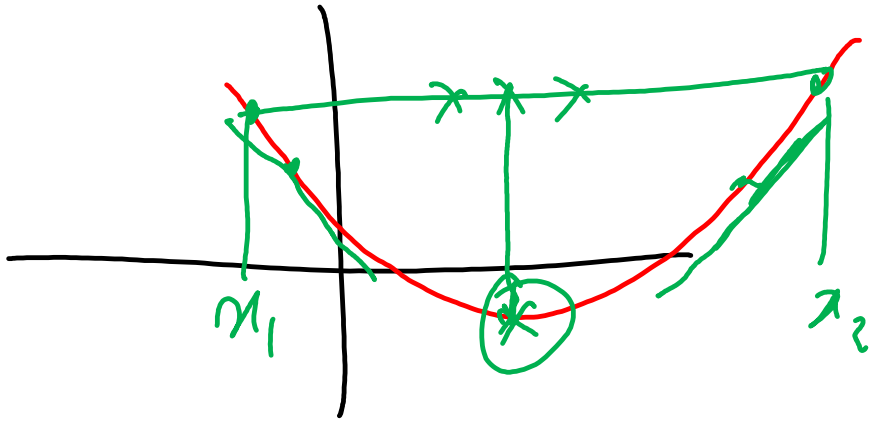


Derivative of a function

Gradient

$f(x)$

$f'(x) = 2ax + b$



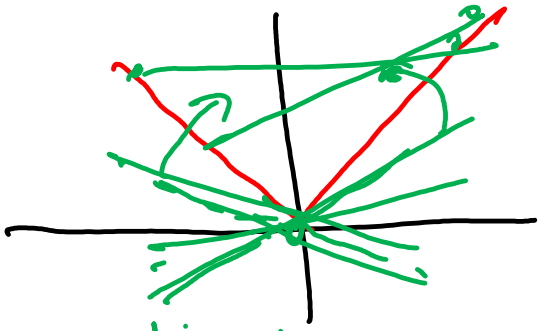
$\lambda f(x_1) + (1-\lambda) f(x_2)$

$\geq f(\lambda x_1 + (1-\lambda) x_2)$

$ax^2 + bx + c$

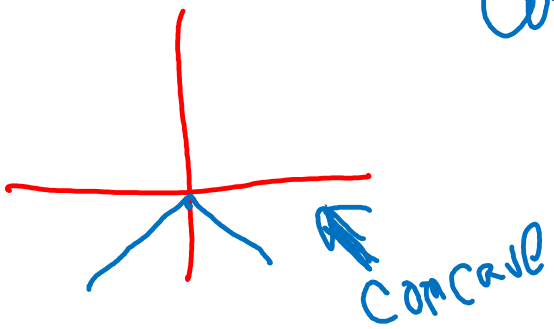
Convex function

Convex



$f(x) = |x|$

$$f'(x) = \begin{cases} -1, & \text{if } x < 0 \\ +1, & \text{if } x > 0 \end{cases}$$



Subgradient

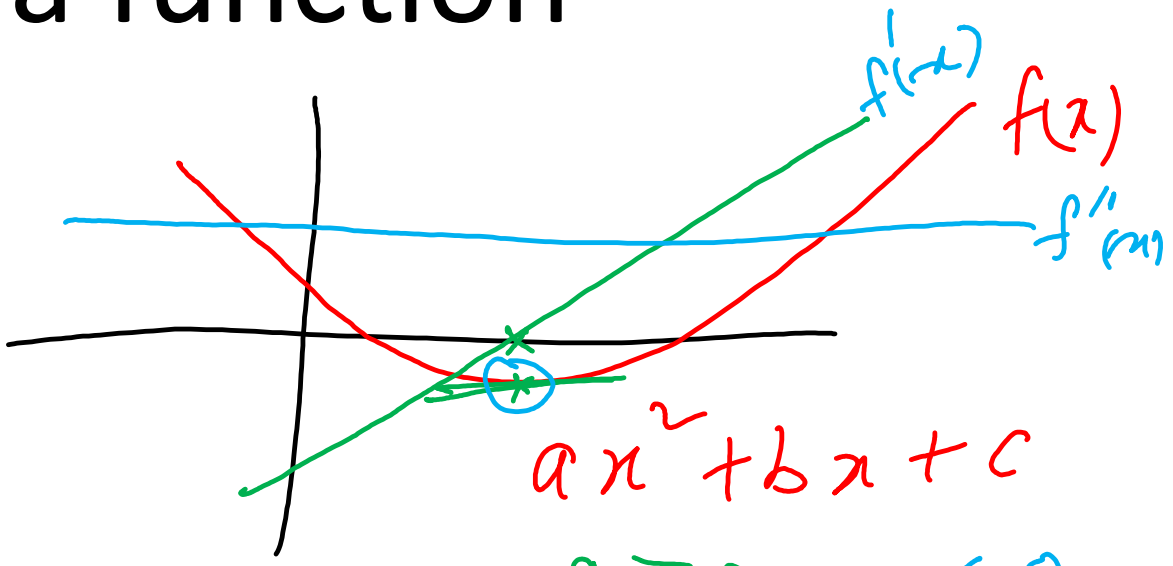
Sub-derivatives are slopes of subtangents

Critical points of a function

$f'(x) = 0$

$f'(x) = 2ax + b$

$x = \frac{-b}{2a}$



Maxima → if $a > 0$, $f''(x) < 0$

Minima → if $f''(x) > 0$

Inflection

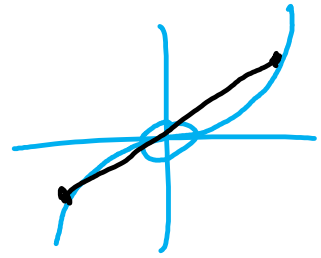
$f''(x) = 2a \geq 0$ ← Convex $\forall x$

Concave fn. $f''(x) \leq 0$

$f(x) = x^3$

$f'(x) = 3x^2$

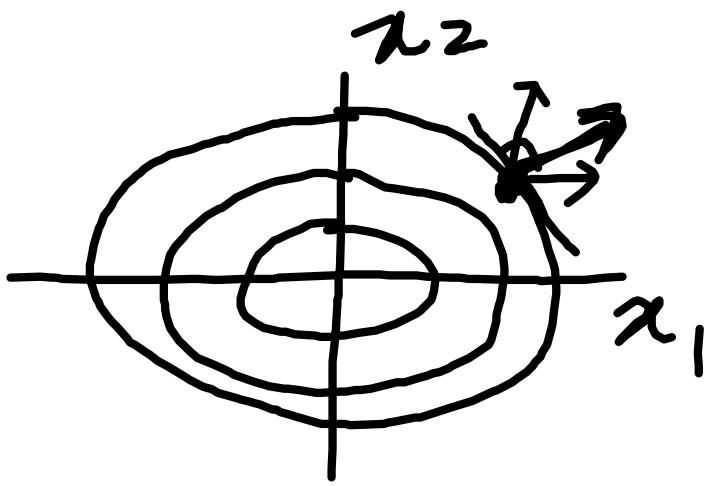
$f''(x) = 6x$



Multivariate functions

$$y = f(x_1, x_2) = a x_1^2 + b x_2^2 \Leftarrow$$

$$\nabla y = \begin{bmatrix} \partial y/\partial x_1 \\ \partial y/\partial x_2 \end{bmatrix} = \begin{bmatrix} 2a x_1 \\ 2b x_2 \end{bmatrix} \Leftarrow$$

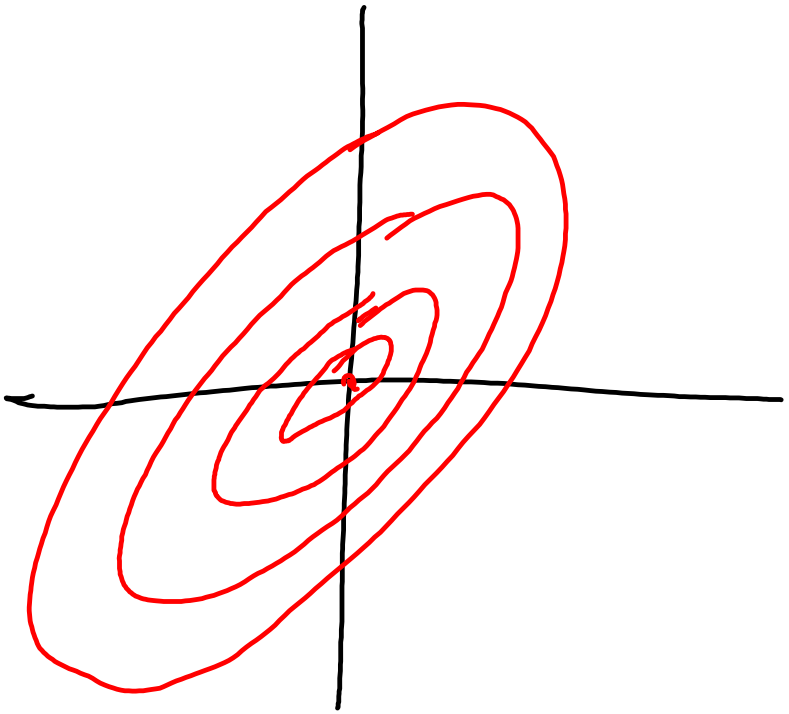


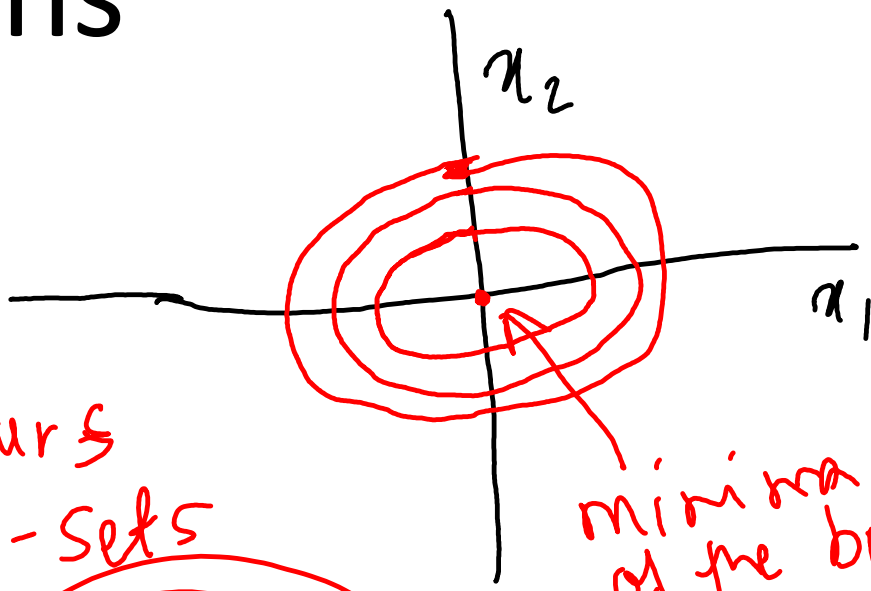
G. D. in 2-d

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \Leftarrow \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \eta \nabla y$$

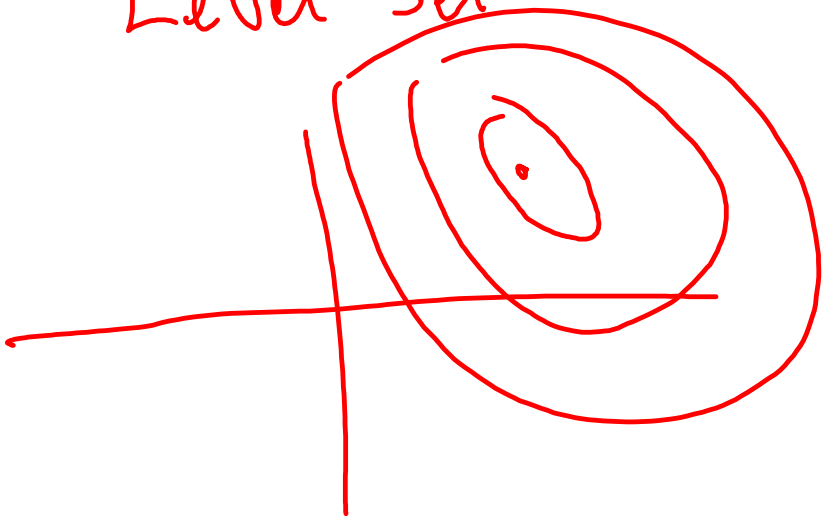
Multivariate functions

$$y = a x_1^2 + b x_2^2$$

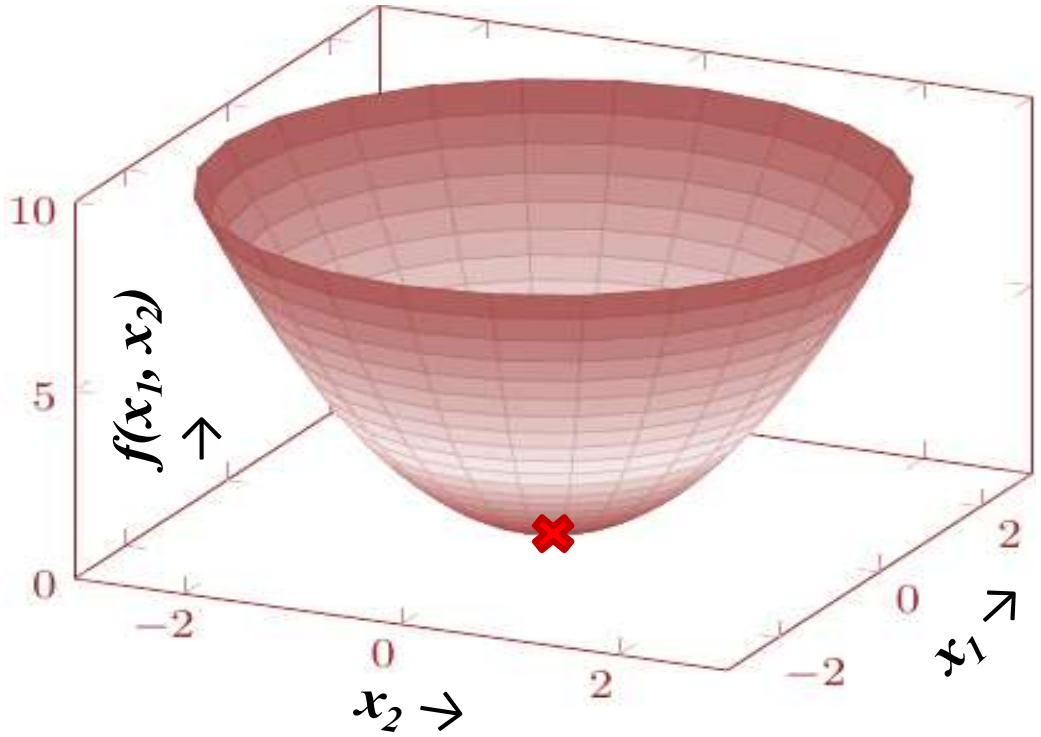




Contours
Level-sets



Gradient of a function of a vector

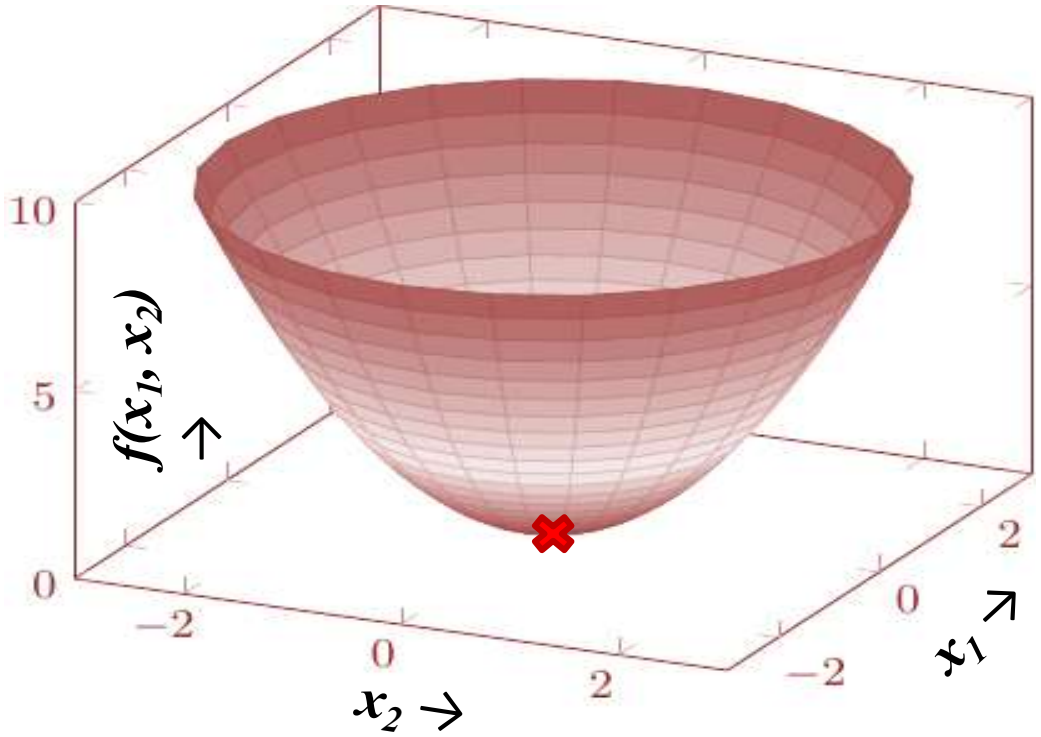


- Derivative with respect to each dimension, holding other dimensions constant
- $\nabla f(\boldsymbol{x}) = \nabla f(x_1, x_2) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix}$
- At a minima or a maxima the gradient is a zero vector

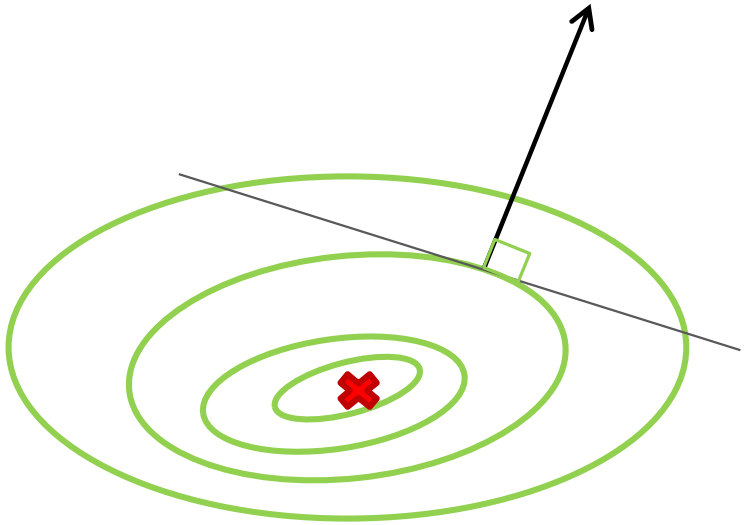
  The function is flat in every direction
- At a minima or a maxima the gradient is a zero vector

Original image source unknown

Gradient of a function of a vector



- Gradient gives a direction for moving towards the minima
- Take a small step towards negative of the gradient



Original image source unknown

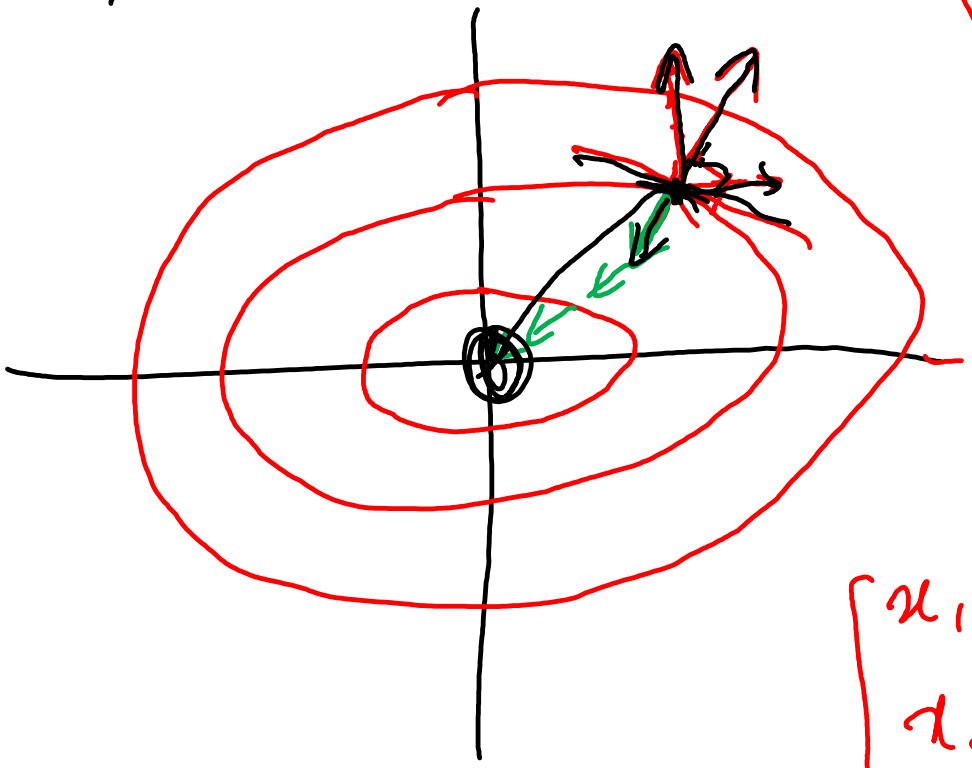
Example of gradient

- Let $f(\boldsymbol{x}) = f(x_1, x_2) = 5{x_1}^2 + 3{x_2}^2$
- Then $\nabla f(\boldsymbol{x}) = \nabla f(x_1, x_2) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 10x_1 \\ 6x_2 \end{bmatrix}$
- At a location $(2,1)$ a step in $\begin{bmatrix} 20 \\ 6 \end{bmatrix}$ or $\begin{bmatrix} 0.958 \\ 0.287 \end{bmatrix}$ direction will lead to maximal increase in the function

Gradients

$f(x_1, x_2) = y = a x_1^2 + b x_2^2$

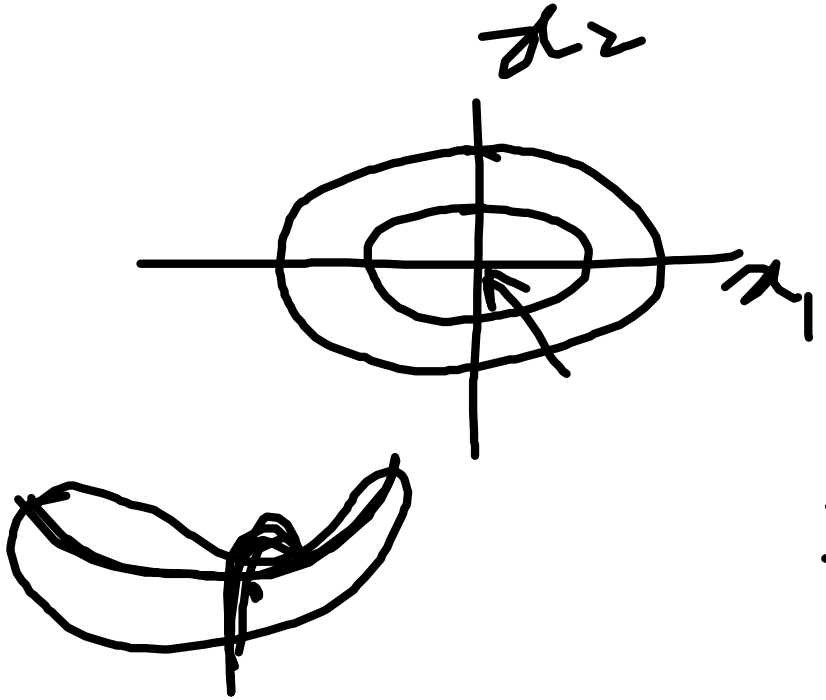
$$\nabla y = \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 2 a x_1 \\ 2 b x_2 \end{bmatrix}$$



Direction of steepest ascent

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \leftarrow \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \eta \begin{bmatrix} \partial y / \partial x_1 \\ \partial y / \partial x_2 \end{bmatrix}$$

Maxima and minima for multivariate functions



$$y = a x_1^2 + b x_2^2$$

$$\nabla y = 0 = \begin{bmatrix} 2ax_1 \\ 2bx_2 \end{bmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

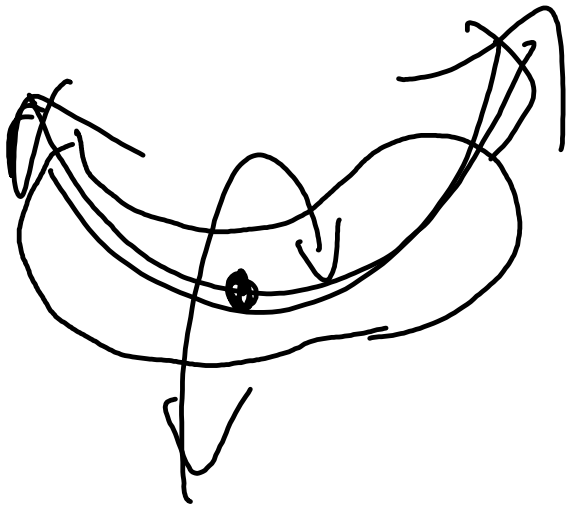
If $a > 0$; $b < 0$

$\Rightarrow (0,0)$ is max for x_2
min for x_1

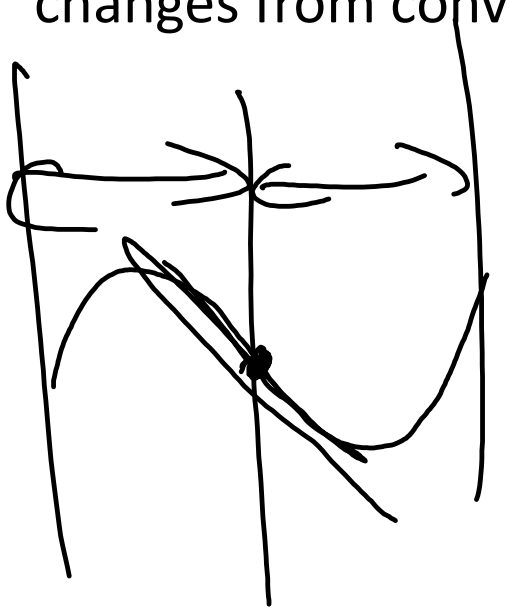
Saddle point

When functions are neither convex nor concave

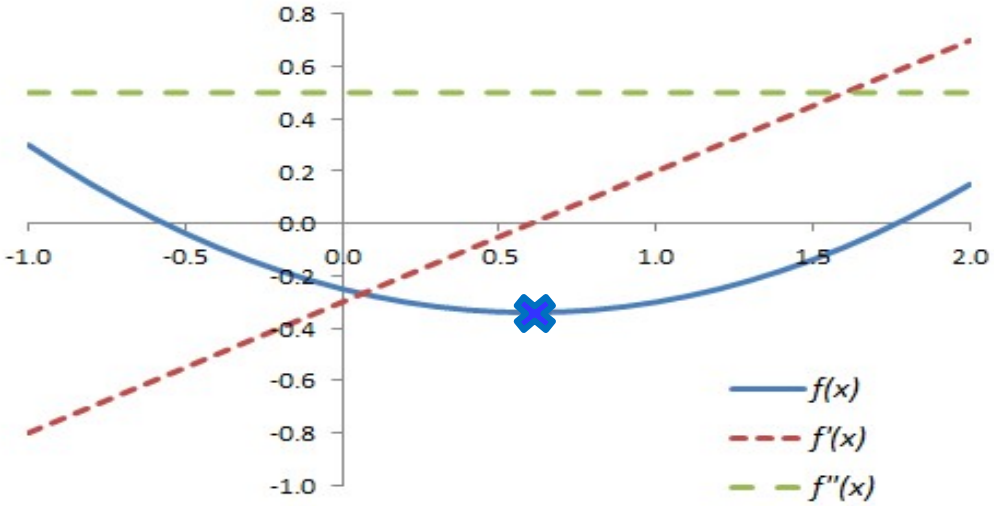
Saddle point
where derivative is zero
and one direction is convex,
and the other in concave

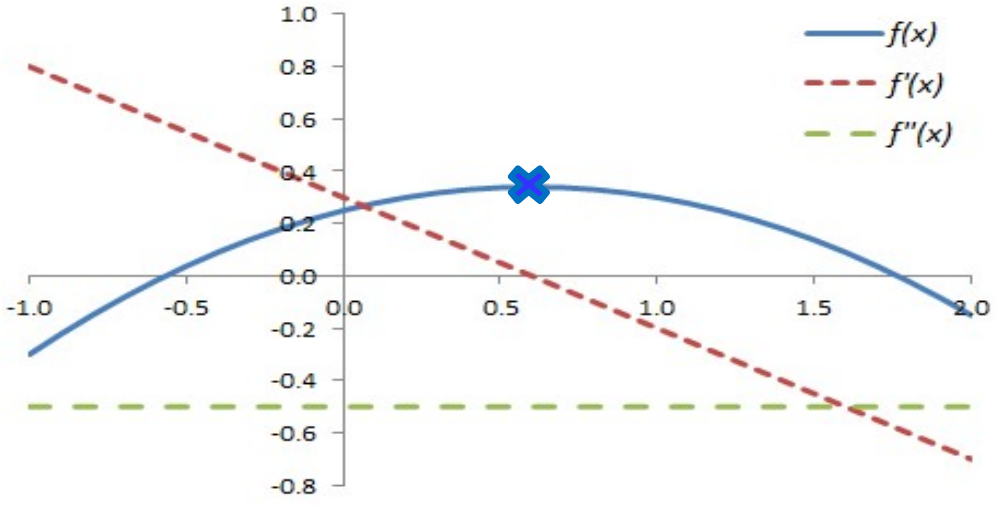


Inflection point
where second derivative
is zero and the function
changes from convex to concave



Double derivative

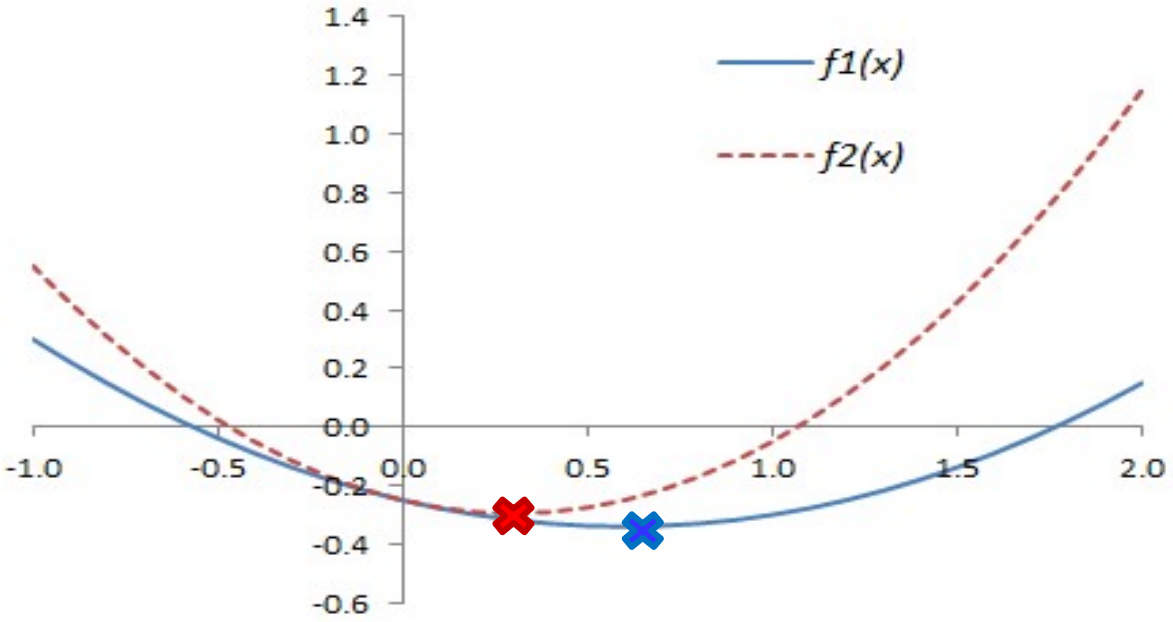




E.g. $f(x) = ax^2 + bx + c, \quad f'(x) = 2ax + b, \quad f''(x) = 2a$

- Double derivative $f''(x) = \frac{d^2 f(x)}{d x^2}$ is the derivative of derivative of $f(x)$
- Double derivative is positive for convex functions (have a single minima), and negative for concave functions (have a single maxima)

Double derivative



$$f(x) = ax^2 + bx + c,$$
$$f'(x) = 2ax + b,$$
$$f''(x) = 2a$$

- Double derivative tells how far the minima might be from a given point.
- From $x = 0$ the minima is closer for the red dashed curve than for the blue solid curve, because the former has a larger second derivative (its slope reverses faster)

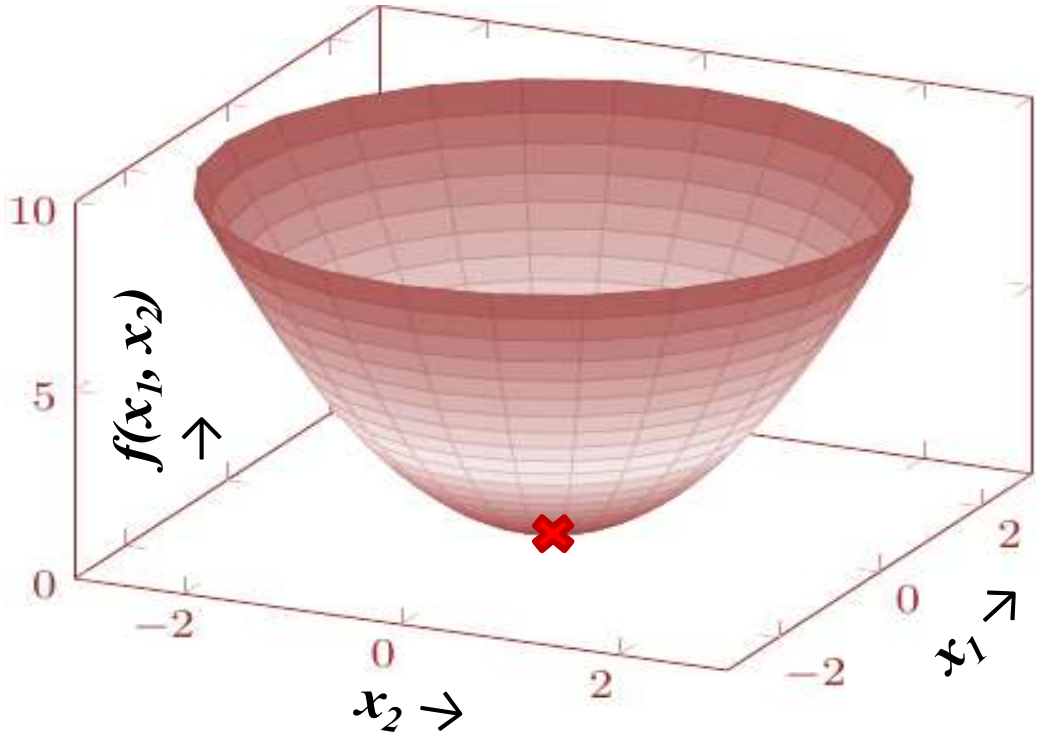
Perfect step size for a paraboloid

- Let $f(x) = ax^2 + bx + c$
- Assuming $a < 0$
- Minima is at: $x^* = -\frac{b}{2a}$
- For any x the perfect step would be:

$$-\frac{b}{2a} - x = -\frac{2ax+b}{2a} = -\frac{f'(x)}{f''(x)}$$

- So, the perfect learning rate is: $\eta^* = \frac{1}{f''(x)}$
- In multiple dimensions, $\boldsymbol{x} \leftarrow \boldsymbol{x} - H\big(f(\boldsymbol{x})\big)^{-1} \nabla(f(\boldsymbol{x}))$
- Practically, we do not want to compute the inverse of a Hessian matrix, so we approximate Hessian inverse

Hessian of a function of a vector



- Double derivative with respect to a pair of dimensions forms the Hessian matrix:
- If all eigenvalues of a Hessian matrix are positive, then the function is convex

Original image source unknown

Example of Hessian

- Let $f(\boldsymbol{x}) = f(x_1, x_2) = 5{x_1}^2 + 3{x_2}^2 + 4x_1x_2$
- Then $\nabla f(\boldsymbol{x}) = \nabla f(x_1, x_2) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 10x_1 + 4x_2 \\ 6x_2 + 4x_1 \end{bmatrix}$
- And, $H(f(\boldsymbol{x})) = \begin{bmatrix} \frac{\partial^2 f}{\partial {x_1}^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial {x_2}^2} \end{bmatrix} = \begin{bmatrix} 10 & 4 \\ 4 & 6 \end{bmatrix}$

Saddle points, Hessian and long local furrows

- Some variables may have reached a local minima while others have not
- Some weights may have almost zero gradient
- At least some eigenvalues may not be negative

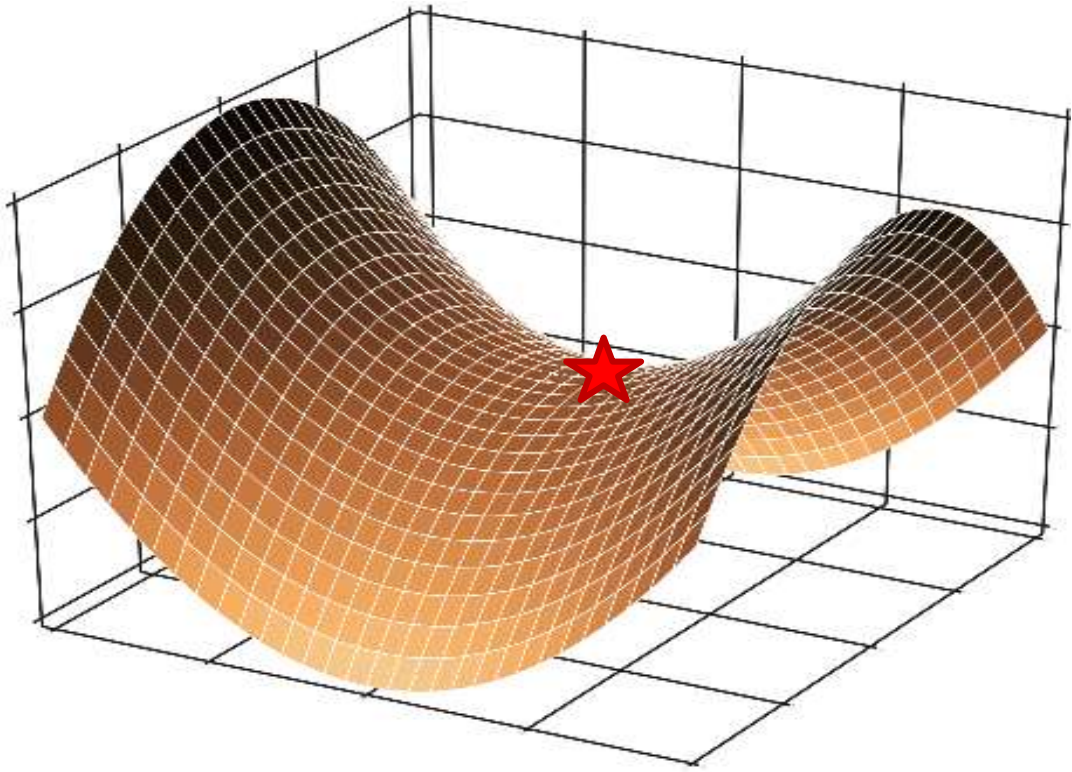


Image source: Wikipedia

A realistic picture

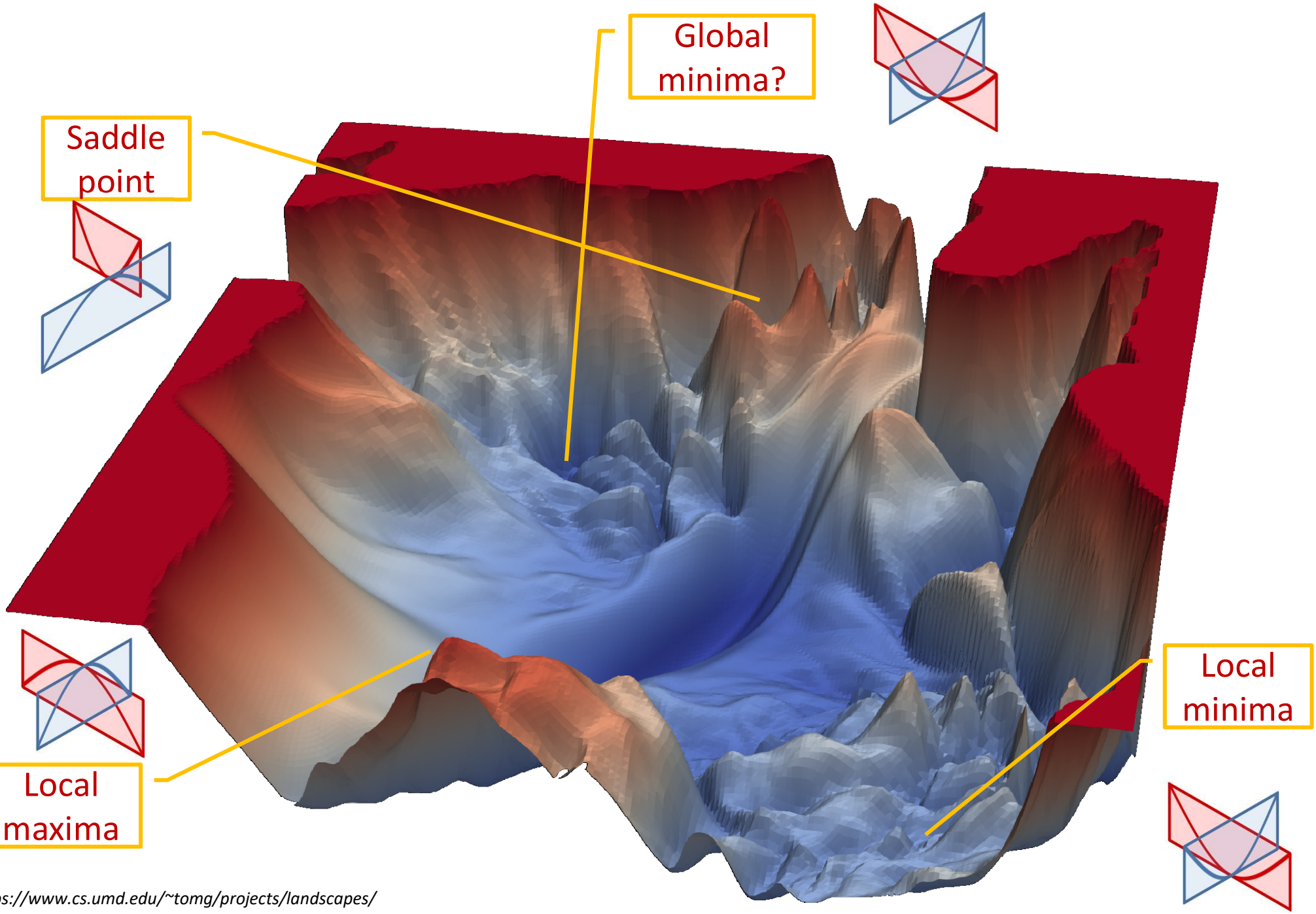
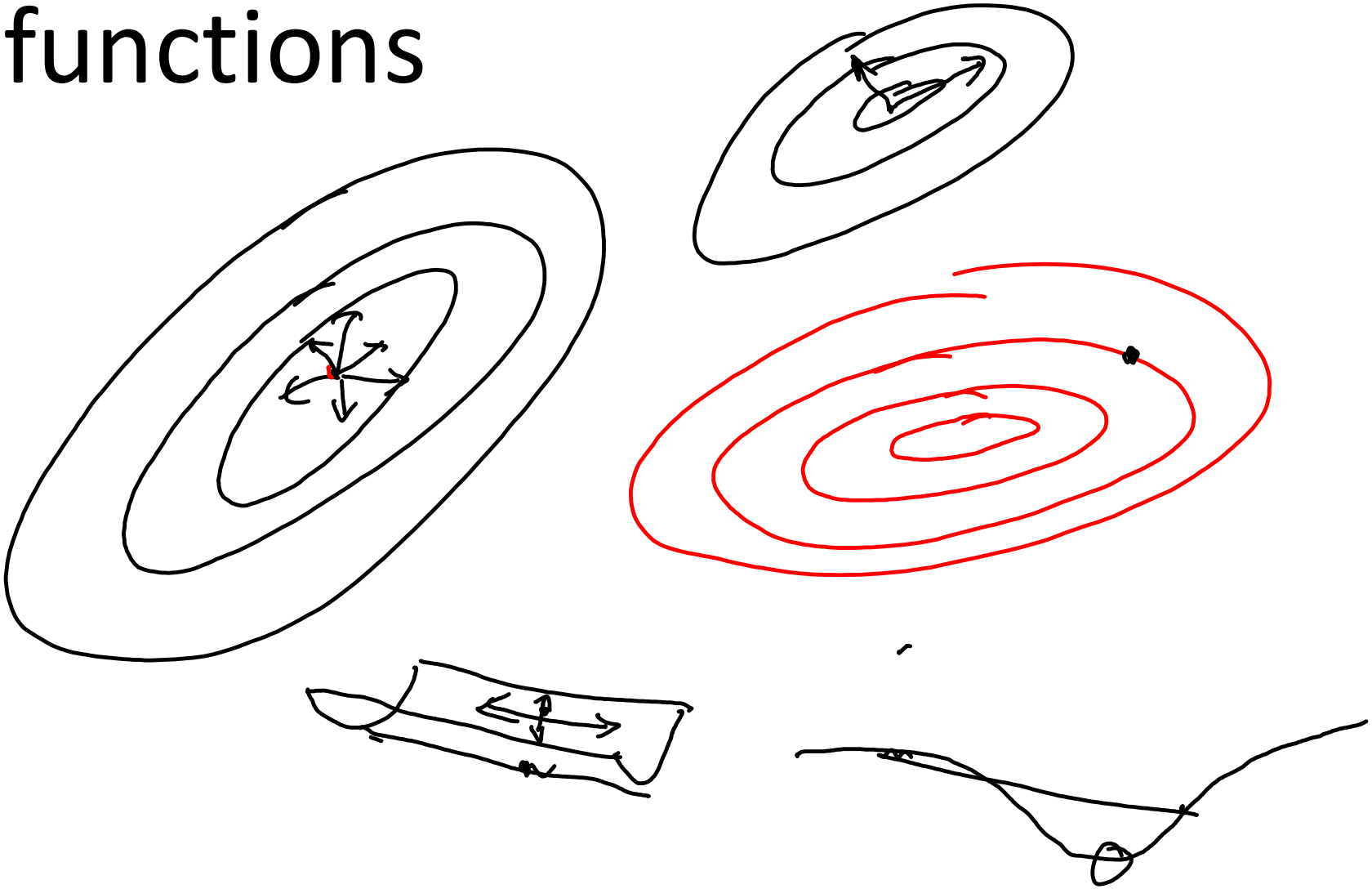
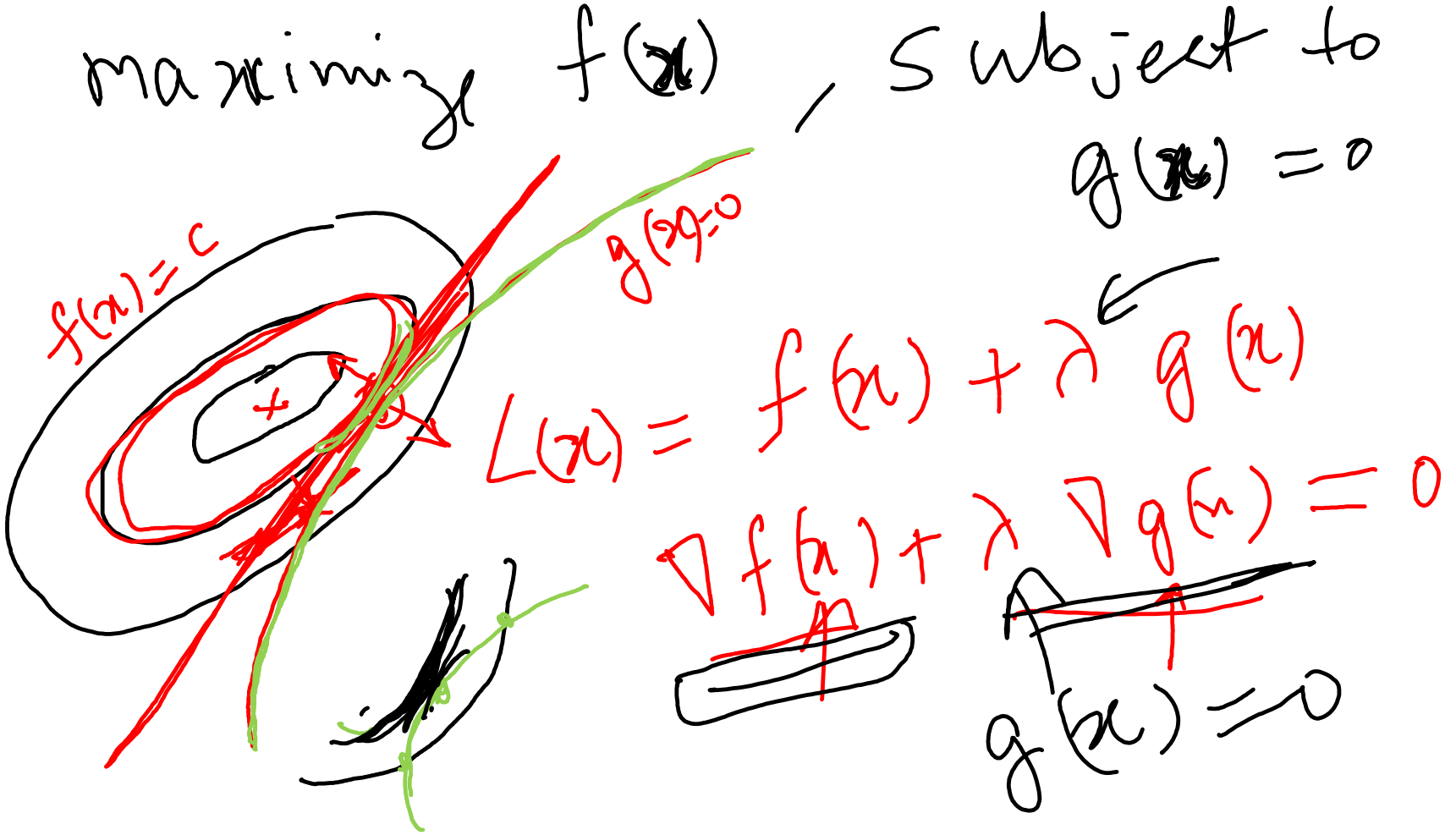


Image source: https://www.cs.umd.edu/~tomg/projects/landscapes/

Some functions



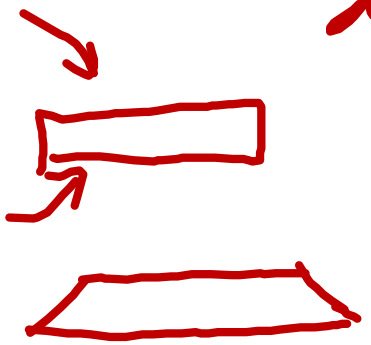
Constrained optimization using Lagrange multiplier



Random variable

$X = x$ $\{H, T\}$ M.E. C.E

$x \in \{0, 1\}$



Fair coin

$P(X=0) = 1 - P(X=1)$

$= 0.5$

$\forall x, P(X=x) \geq 0$

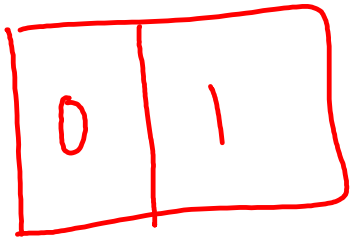
$\sum_x P(X=x) = 1$

Random variable

$X = x$ $\{H, T\}$ Event space

$X \in \{0, 1\}$ M.E.C.E.

Fair coin 



$P(X=0) = 1 - P(X=1) = 0.5$

Biased coin 

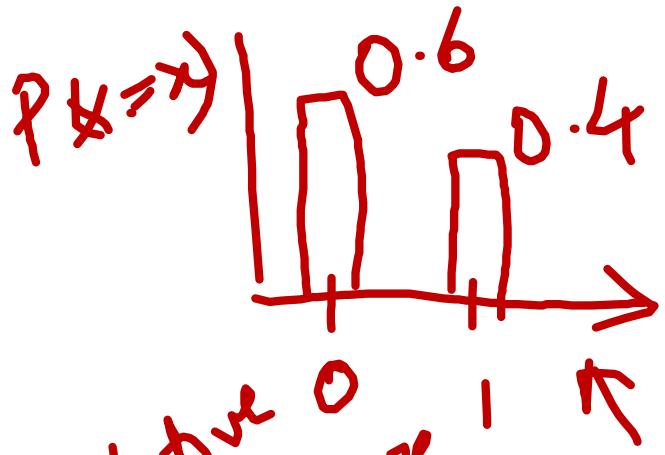
$P(X=0) = 0.6$

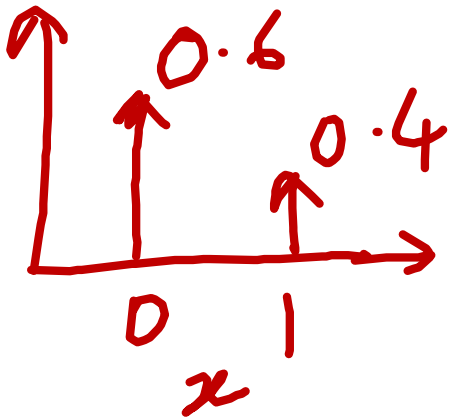
$P(X=1) = 0.4$

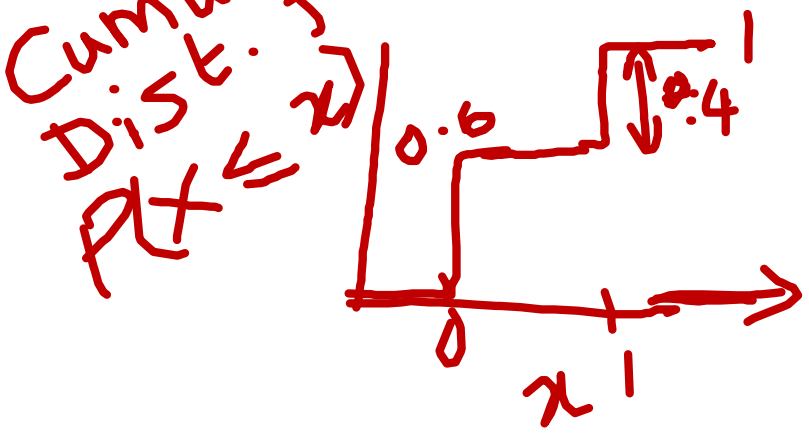
$\forall x, P(X=x) \geq 0$

$\sum_x P(X=x) = 1$

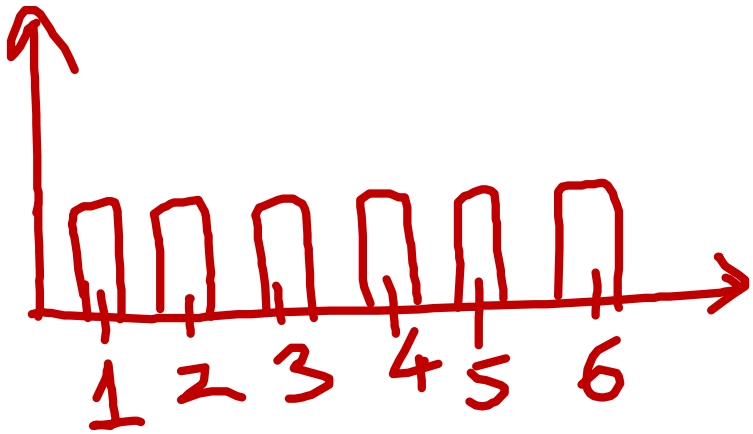
Probability mass function







6-faced die



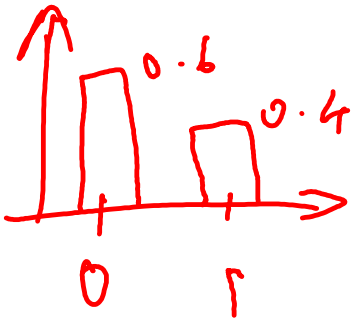
Probability mass function

$P(X = x)$

$P_X(x) : x \longrightarrow [0, 1]$

$\uparrow \quad \sum_x P_X(x) = 1$

Discreet R.V.



Some common PMFs

Bernoulli Dist.

$x \in \{0, 1\}$

$\mu = P(X = 1)$

$\text{Bern}(x \mid \mu) = \mu^x (1-\mu)^{1-x} = \begin{cases} \mu, & \text{if } x = 1 \\ 1-\mu, & \text{if } x = 0 \end{cases}$

$\mathbb{E}[x] = \sum_x x \, P(X = x) = \mu$

$\mathbb{E}[(x - \bar{x})^2] = \mu(1-\mu)$

Binomial Dist.

$\text{Bin}(y \mid \mu, N) = \binom{N}{y} \mu^y (1-\mu)^{N-y}$

$\mathbb{E}[f(x)] = \sum_x f(x) P(X = x)$

Some common PMFs

① Bernoulli $x \in \{0, 1\}$

$\mu = P(x=1)$ $\mu \in [0, 1]$

$\text{Bern}(x|\mu) = \mu^x (1-\mu)^{1-x}$

given $\begin{cases} \mu, \text{ if } x=1 \\ 1-\mu, \text{ if } x=0 \end{cases}$

$\mathbb{E}(x) = \sum_x x\, P(x) = 0 \cdot (1-\mu) + 1 \cdot \mu = \mu$

$\mathbb{E}f(x) = \sum_x f(x) P(x) = f(0)(1-\mu) + f(1)\mu$

$\mathbb{E}(x^2) = 0(1-\mu) + 1(\mu) = \mu$

Variance of x

$= \mathbb{E}(x - \mathbb{E}(x))^2$

$= \mathbb{E}(x-\mu)^2$

$= \mathbb{E}(x^2 - 2x\mu + \mu^2)$

$= \mathbb{E}(x^2) - 2\mu\mathbb{E}(x) + \mu^2$

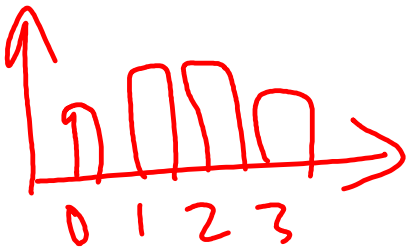
$= \mu(1-\mu)$

Binomial distribution

μ, N I.I.D.

$$Bin(y \mid \mu, N) = \binom{N}{y} \mu^{y} (1-\mu)^{N-y}$$

$$= {}^{N}C_{y}$$



0 0 0	0
0 0 1	1
0 1 0	1
0 1 1	2
1 0 0	1
1 0 1	2
1 1 0	2
1 1 1	3

$E(y) = N\mu$

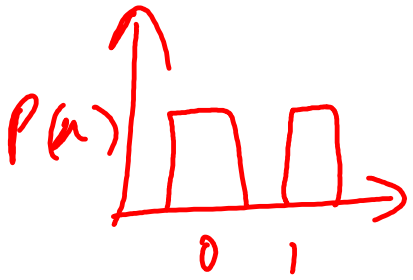
Entropy

Entropy of a random variable

$$H(x) = -E[\log(P(x))]$$

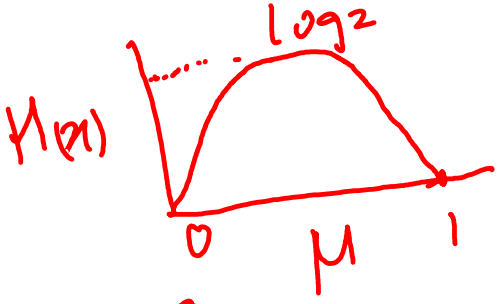
$$= -\sum_x \log[P(x)]\, P(x)$$

$$= -[(1-\mu)\log(1-\mu) + \mu \log \mu]$$



High $H(x)$

Low $H(x)$



$\mu = 0.5 \Rightarrow H(x) = -\log\left(\frac{1}{2}\right) = \log 2$

$= 0$

$$\text{Lt}_{\mu \to 0} \quad H(x) = -\underbrace{(1-\mu)}\log\underbrace{(1-\mu)} - \underbrace{\mu}\log\underbrace{\mu}$$

Joint, conditional, and marginal probability

Dice: $x \in \{1,2,3,4,5,6\}$

$y \in \{0,1\}$, $0 \Rightarrow$ even

$z \in \{0,1\}$, $x < 3$

Joint

$P(y,z) \neq 0$

	$z=0$	$z=1$	$p(y)$
$y=0$	$1/6$	$2/6$	$1/2$
$y=1$	$1/6$	$2/6$	$1/2$
$p(z)$	$1/3$	$2/3$	

$\sum_y \sum_z p(y,z) = 1$

Marginal Dist

$p(z) = \sum_y p(y,z)$

$p(y) = \sum_z p(y,z)$

Conditional

$p(y|z) = \frac{p(y,z)}{p(z)}$

$p(y|z=0)$

$p(y=0|z=0) = \frac{1/6}{1/3} = \frac{1}{2}$

Joint, conditional, and marginal probability

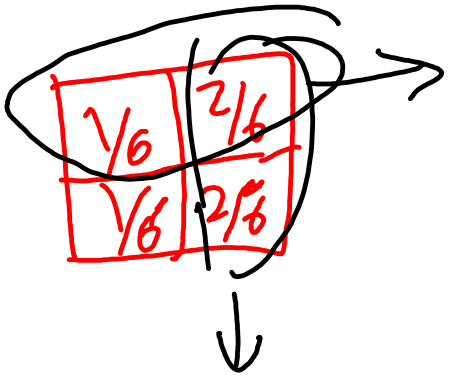
Dice $x \in \{1, 2, 3, 4, 5, 6\}$

$y \in \{0, 1\}$ 0 => Even, 1 means odd

$z \in \{0, 1\}$ $0 \Rightarrow x < 3$, $1 \Rightarrow x \not< 3$

$P(y, z)$

	$z=0$	$z=1$
$y=0$	$\{2\}$ 1/6	$\{4,6\}$ 2/6
$y=1$	$\{1\}$ 1/6	$\{3,5\}$ 2/6



$\sum_y \sum_z P(y, z) = 1$

$P(y, z) \in [0, 1]$

$P(y,z)$

	$z=0$	$z=1$
$y=0$	$1/6$	$2/6$
$y=1$	$1/6$	$2/6$

$P(y|z)$ ← given

Conditional

$P(x_i \text{ Even} \mid x \geq 3) = \frac{2/6}{2/6+2/6} = \frac{1}{2}$

$P(y|z) = \frac{P(y,z)}{P(z)} = \frac{2/6}{4/6} = \frac{1}{2}$

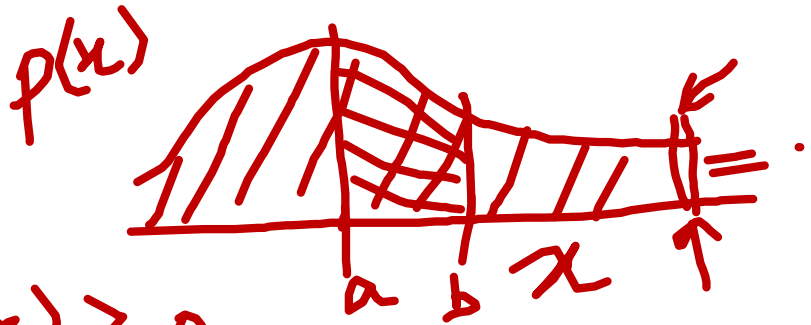
$P(y,z) = P(y|z)\,P(z)$

Marginal $P(z) = \sum_y P(y,z) = \frac{2}{6} + \frac{2}{6} = \frac{4}{6}$

Continuous random variable and PDF

$P(X = x) = 0$ $x = 170\,cm$

$p(x)$



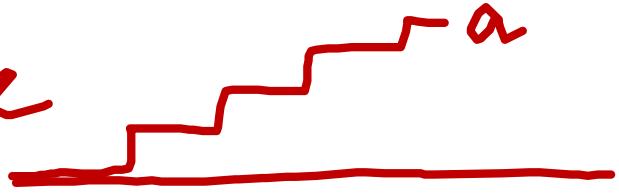
$$\int_{-\infty}^{\infty} p(x)\,dx = 1 \qquad p(x) \geq 0$$

$$P(a < x \leq b) = \int_{a}^{b} p(x)\,dx$$

Cumulative D.F.

$$\int_{-\infty}^{a} p(x)\,dx$$

Discrete



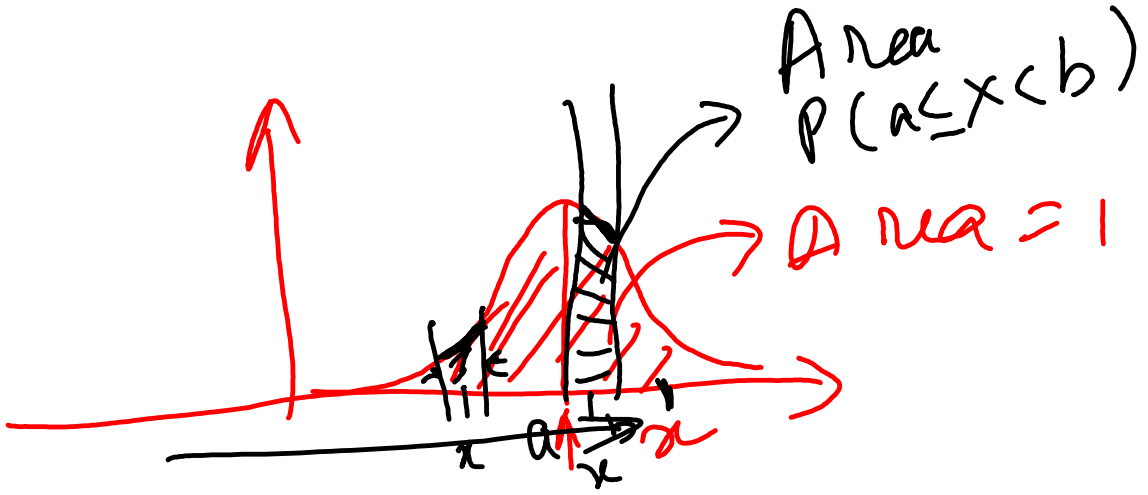
Continuous random variable and PDF

$P(x = x) = 0$

Prob. density fn.

$p(x) \geq 0$

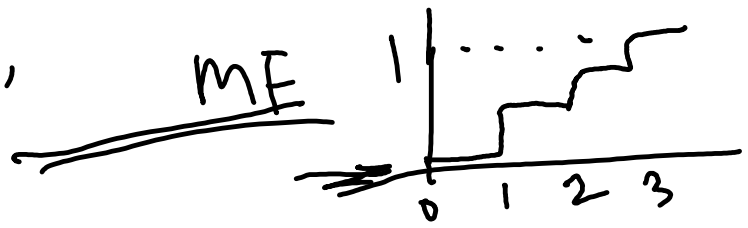
$\int_{-\infty}^{\infty} p(x) = 1$



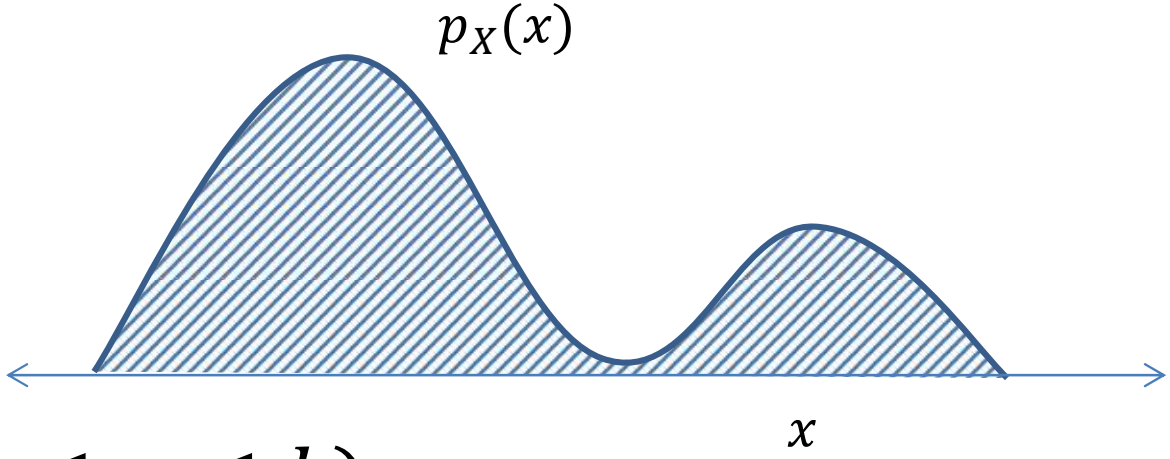
170 cm 169.99
170.01

Cont. R.V. Cumulative DF $= \int_{-\infty}^{x} P_x(y)\,dy$

Discrete " " MF



Probability density function (PDF)

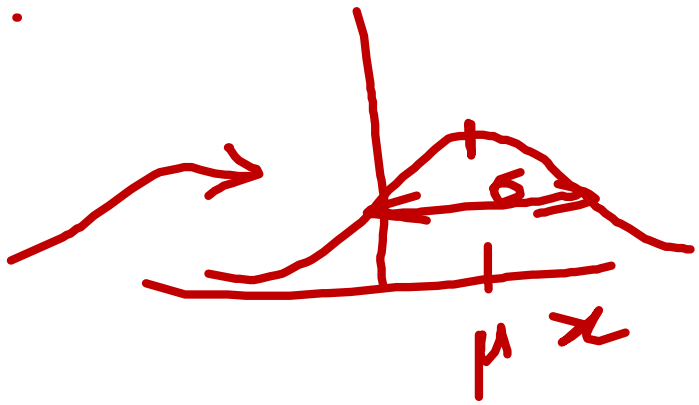


- PDF: $p_X(x) \geq 0$
- $\int_a^b p_X(\hat{x})d\hat{x} = Pr(a \leq x \leq b)$
- $\int_{-\infty}^{+\infty} p_X(\hat{x})d\hat{x} = 1$
- CDF: $P_X(x) = \int_{-\infty}^{x} p_X(\hat{x})d\hat{x}$
- $p_X(x) = \frac{dP_X(x)}{dx}$

Some common PDFs

Gaussian / Normal Dist.

$$\mathcal{N}(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



$$E(x) = \int_{-\infty}^{\infty} x\, p(x)\, dx = \mu$$

$$E[(x-\mu)^2] = \sigma^2$$

Unif$(x;a,b)$

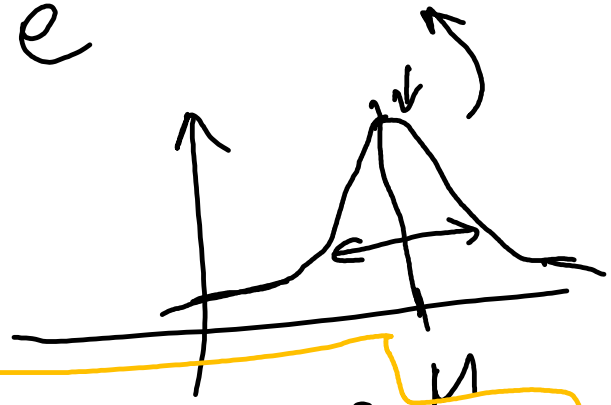
Beta dist. $x \in [0,1]$

$$\text{Beta}(x;a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$$

Some common PDFs

Gaussian / Normal

$$\mathcal{N}(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]}$$



$$H(x) = \int_{-\infty}^{\infty} [\log p(x)]\, p(x)\, dx$$

$$\mathbb{E}[f(x)] = \int_{-\infty}^{\infty} f(x)\, p(x)\, dx$$

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} x\, p(x)\, dx$$

Beta

Unif.

$\frac{1}{b-a}$

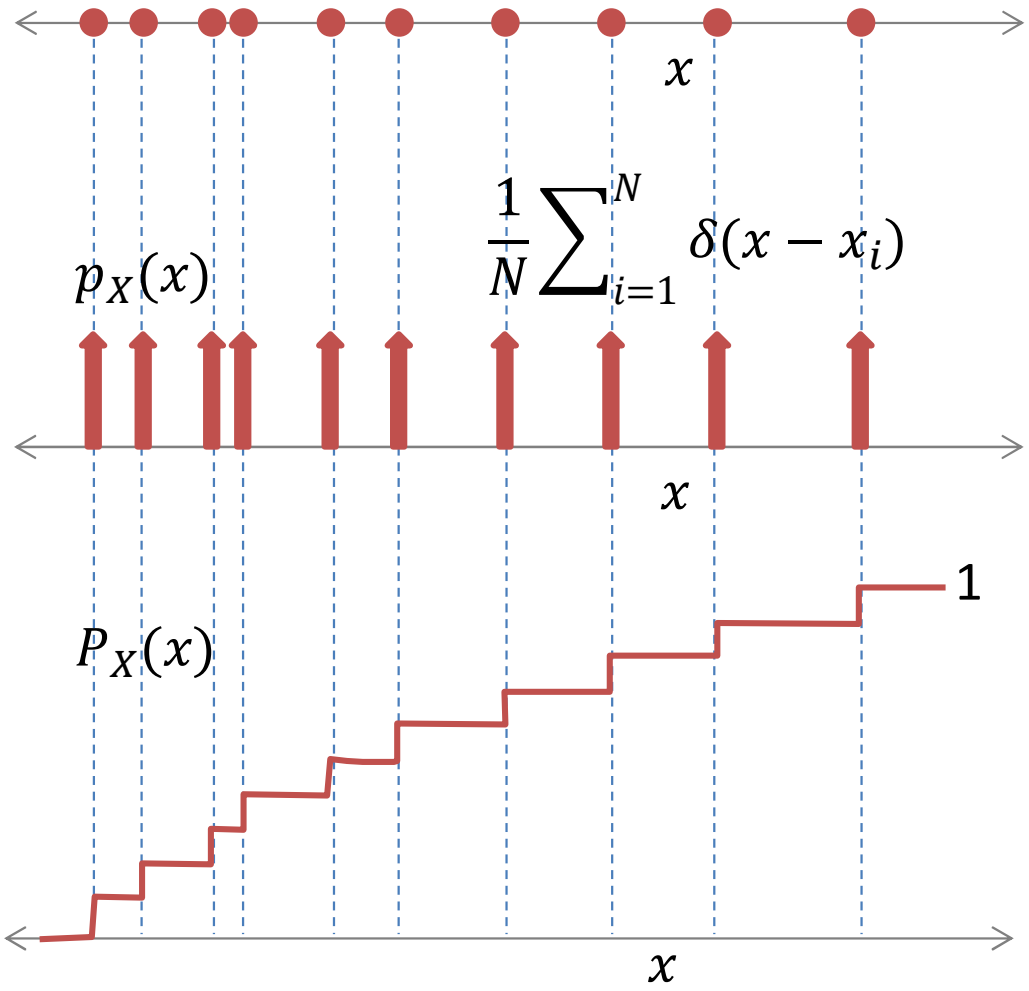
a, b $\quad \mathrm{Beta}(x|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$

for $0 \le x \le 1$

0 otherwise

Empirical distribution

- Given samples
- What is the “best” explanation of the data?
- Empirical PDF (train of Dirac delta functions scaled 1/N)
- CDF has N steps



Problem with empirical distribution

- Rote learning
- Does not generalize
- If we randomly split the sample into training and validation
- How likely is the validation subset based on the distribution learnt from the training subset?

Between two distributions, which is better?

- The one under which the data has higher likelihood
- If $p_A(x_i) > p_B(x_i)$ then p_A is a better distribution to explain the observation x_i
- This can be maximized by having a Dirac delta function $\delta(x - x_i)$
- What about for the entire data sample X?

The IID Assumption

- Independence of x_i and x_j for $i \neq j$.
- $p(x_i, x_j) = p(x_i)p(x_j)$

- All samples drawn from the same (identical) distributed
- $x_i, x_j \sim p_X$

Implication of the IID assumption

- Likelihood of the whole data factorizes:

- $p(X) = p(x_1, \ldots, x_N) = p_X(x_1) \times \ldots \times p_X(x_N)$
  $= \prod_{i=1}^{N} p_X(x_i)$

- $L(X) = \log p(X) = \sum_i \log p_X(x_i)$

MLE of parameterized distribution

- Between two distributions p_A is a better explanation than p_B of the entire data X if $\prod_i p_A(x_i) > \prod_i p_B(x_i)$

- By extension, if a family of distributions is parameterized by θ, then we are interested in

$$\begin{aligned} \arg\max_\theta \prod_i p_\theta(x_i) &= \arg\max_\theta \sum_i \log p_\theta(x_i) \\ &= \arg\max_\theta L_\theta(X) \end{aligned}$$

Example 1: Exponential distribution

- $p_\lambda(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0 & , \ x < 0 \end{cases}$
- Log likelihood $L_\lambda(X) = \sum_i \log p_\lambda(x_i)$
- Differntiating $\frac{\partial L_\lambda(X)}{\partial \lambda} = \sum_{x_{i \geq 0}} \left(\frac{\partial \log \lambda}{\partial \lambda} - \frac{\partial \lambda x_i}{\partial \lambda} \right) = 0$
- (assuming all samples are non-negative)
- Implies $\lambda = \frac{N}{\sum_i x_i}$ = inverse of the sample mean

Example 2: Uniform distribution

- $p_{a,b}(x) = \begin{cases} \frac{1}{b-a}, a \leq x \leq b \\ 0 \quad , \text{otherwise} \end{cases}$
- Log likelihood $L_{a,b}(X) = \sum_i \log p_{a,b}(x_i)$
- $= K \log(0) \; - (\mathrm{N} - \mathrm{K}) \log(b - a)$
- Reduce the K (terms outside $[a, b]$) to zero, and minimize $(b - a)$ by differentiating wrt a, b
- So, $a = \min x_i$, $b = \max x_i$

Example 3: Gaussian distribution

- $p_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
- MLE by differentiating log likelihood wrt μ, σ
- gives $\mu = \bar{x} = \frac{\sum_i x_i}{N}$; i.e., sample mean
- and, $\sigma = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{N}}$

Sufficient statistics

- Statistic is a function of the sample

$$T(X) = T(x_1, \ldots, x_N)$$

- For some distributions, computing a few statistics is sufficient for MLE estimate
- Gives complete information about the distribution
- Examples:
  - Sample mean and variance for Gaussian distribution
  - Sample mean for exponential distribution
  - Max and min for uniform distribution

Comparing two parametric distributions

- Let there be two candidate families of distributions $p_\theta(X)$ and $p_\phi(X)$ to explain the data
- Can we compare $\max L_\theta(X)$ and $\max L_\phi(X)$?
- Yes, we can, but we might overfit
- Narrow down the family of distributions based on domain knowledge (e.g. physical phenomenon)
- E.g. “Can the random variable take negative values?”

MLE vs. Bayesian estimate

- MLE finds θ that maximizes $L_\theta(X)$
- Bayesian estimate takes the expected value of θ w.r.t. $L_\theta(X)$
- Bayesian estimate: $\int \theta L_\theta(X) d\theta \;/\; \int L_\theta(X) d\theta$
- We can also incorporate a prior over θ
- $\int \theta p_\Theta(\theta) L_\theta(X) d\theta \;/\; \int p_\Theta(\theta) L_\theta(X) d\theta$

Multivariate PDF

$$p(x_1, x_2) \geq 0$$

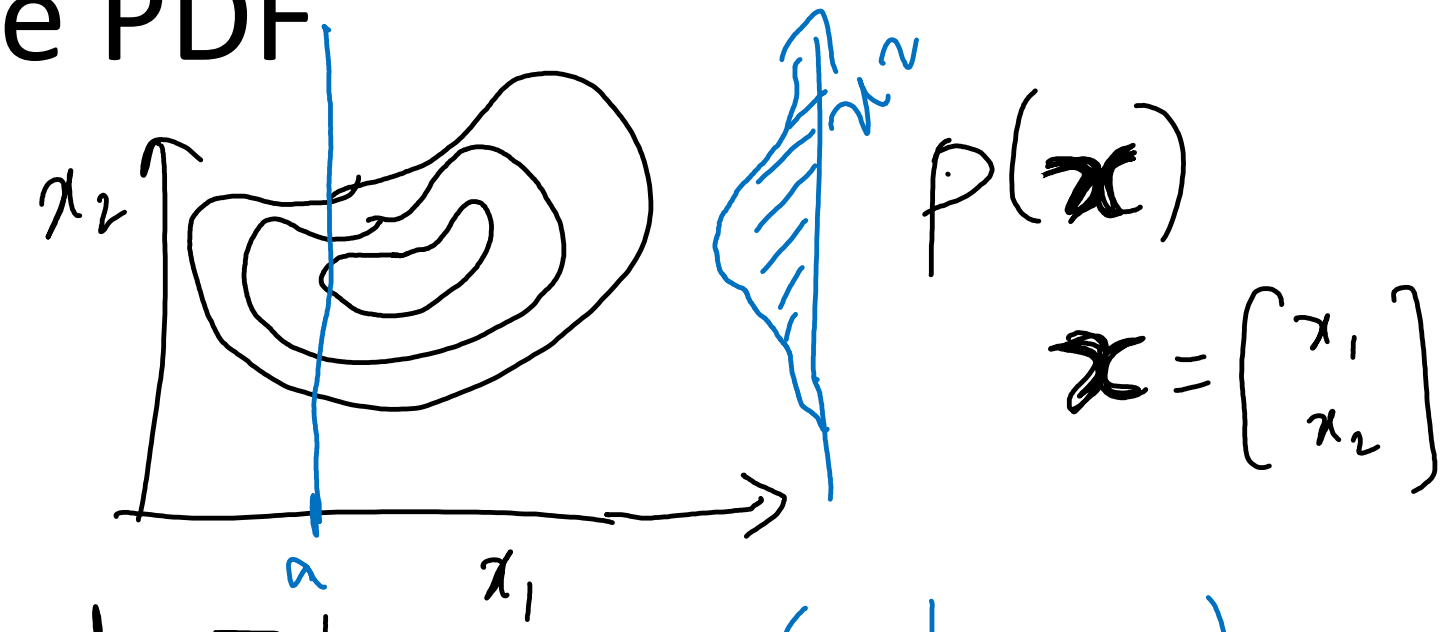
$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x_1, x_2)\, dx_1\, dx_2 = 1$$

$$\int_{-\infty}^{\infty} p(x_1, x_2)\, dx_1 = p(x_2)$$

$$p(x_2 \mid x_1 = a) = \frac{p(x_1 = a, x_2)}{p(x_1 = a)}$$

Multivariate PDF

$p(x_1, x_2)$



$p(x_1, x_2) \geq 0$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x_1, x_2)\, dx_1\, dx_2 = 1$$

$$p(x_1) = \int_{-\infty}^{\infty} p(x_1, x_2)\, dx_2$$

marginal

Conditional

$$p(x_2 \mid x_1 = a) = \frac{p(x_1 = a, x_2)}{p(x_1 = a)}$$

Multivariate Gaussian

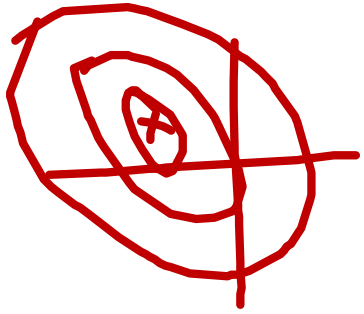
$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

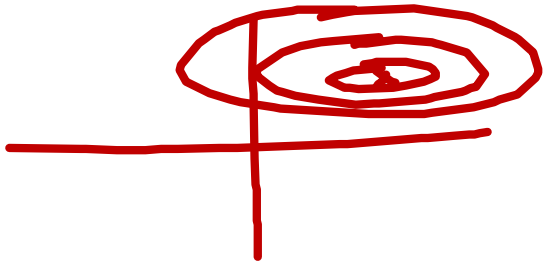
$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{21} & \sigma_{22}^2 \end{bmatrix}$$

$p(x_1, x_2)$

$$p(x) = \frac{1}{\sqrt{(2\pi)^k \, \det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

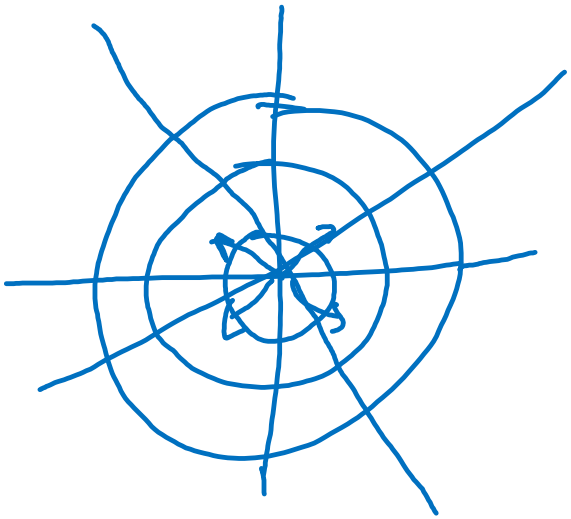




Multivariate Gaussian PDF

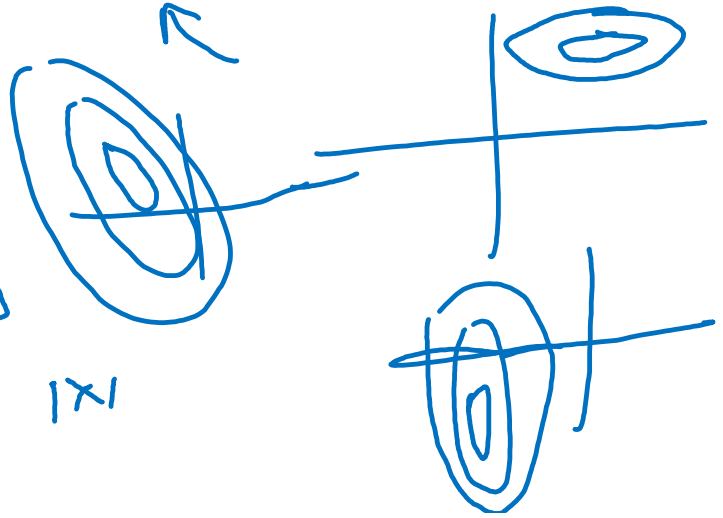
$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{21} & \sigma_{22}^2 \end{bmatrix}$$



$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^k \det(\Sigma)}} \times e^{\left[-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)\right]}$$

$1 \times k \quad k \times k \quad k \times 1 \quad 1 \times 1$



Basic Statistical Testing

Amit Sethi, EE, IITB

Learning Objectives

- Compare distributions that can explain a given sample
- Test if two samples are from two different distributions
- Test the strength of relation between two variables

The IID Assumption

- Independence of x_i and x_j for $i \neq j$.
- $p(x_i, x_j) = p(x_i)p(x_j)$

- All samples drawn from the same (identical) distributed
- $x_i, x_j \sim p_X$

Implication of the IID assumption

- Likelihood of the whole data factorizes:

- $p(X) = p(x_1, \ldots, x_N) = p_X(x_1) \times \ldots \times p_X(x_N)$
  $= \prod_{i=1}^{N} p_X(x_i)$

- $L(X) = \log p(X) = \sum_i \log p_X(x_i)$

MLE of parameterized distribution

- Between two distributions p_A is a better explanation than p_B of the entire data X if $\prod_i p_A(x_i) > \prod_i p_B(x_i)$

- By extension, if a family of distributions is parameterized by θ, then we are interested in

$$\begin{aligned} \arg\max_\theta \prod_i p_\theta(x_i) &= \arg\max_\theta \sum_i \log p_\theta(x_i) \\ &= \arg\max_\theta L_\theta(X) \end{aligned}$$

Example 1: Exponential distribution

- $p_\lambda(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0 & , \ x < 0 \end{cases}$
- Log likelihood $L_\lambda(X) = \sum_i \log p_\lambda(x_i)$
- Differntiating $\frac{\partial L_\lambda(X)}{\partial \lambda} = \sum_{x_{i \geq 0}} \left(\frac{\partial \log \lambda}{\partial \lambda} - \frac{\partial \lambda x_i}{\partial \lambda} \right) = 0$
- (assuming all samples are non-negative)
- Implies $\lambda = \frac{N}{\sum_i x_i}$ = inverse of the sample mean

Example 2: Uniform distribution

- $p_{a,b}(x) = \begin{cases} \frac{1}{b-a}, a \leq x \leq b \\ 0 \quad , \text{otherwise} \end{cases}$
- Log likelihood $L_{a,b}(X) = \sum_i \log p_{a,b}(x_i)$
- $= K \log(0) \; - (\mathrm{N} - \mathrm{K}) \log(b - a)$
- Reduce the K (terms outside $[a, b]$) to zero, and minimize $(b - a)$ by differentiating wrt a, b
- So, $a = \min x_i$, $b = \max x_i$

Example 3: Gaussian distribution

- $p_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
- MLE by differentiating log likelihood wrt μ, σ
- gives $\mu = \bar{x} = \frac{\sum_i x_i}{N}$; i.e., sample mean
- and, $\sigma = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{N}}$

Sufficient statistics

- Statistic is a function of the sample

$$T(X) = T(x_1, \ldots, x_N)$$

- For some distributions, computing a few statistics is sufficient for MLE estimate
- Gives complete information about the distribution
- Examples:
  - Sample mean and variance for Gaussian distribution
  - Sample mean for exponential distribution
  - Max and min for uniform distribution

Comparing two parametric distributions

- Let there be two candidate families of distributions $p_\theta(X)$ and $p_\phi(X)$ to explain the data
- Can we compare $\max L_\theta(X)$ and $\max L_\phi(X)$?
- Yes, we can, but we might overfit
- Narrow down the family of distributions based on domain knowledge (e.g. physical phenomenon)
- E.g. “Can the random variable take negative values?”

MLE vs. Bayesian estimate

- MLE finds θ that maximizes $L_\theta(X)$
- Bayesian estimate takes the expected value of θ w.r.t. $L_\theta(X)$
- Bayesian estimate: $\int \theta L_\theta(X) d\theta \ / \ \int L_\theta(X) d\theta$
- We can also incorporate a prior over θ
- $\int \theta p_\Theta(\theta) L_\theta(X) d\theta \ / \ \int p_\Theta(\theta) L_\theta(X) d\theta$

Recipe for statistical testing

1. **Explore** reasonable assumptions about the data, e.g. distribution type (including "cannot be assumed"), mean, variance, etc. and ask what do we want to verify
2. **Form null hypothesis H_0** that we want to reject, e.g. "The two means are NOT different"
3. **Form alternative hypothesis** that we hope is true, e.g. "The two means are different"
4. **Decide on a significance level** (1 − confidence) to reject the null hypothesis BEFORE performing a test, e.g. $p < 0.05$ or $p < 0.01$
5. **Perform the test** by performing the calculations
6. **Check if the result was significant** enough to reject the null hypothesis and accept the alternative hypothesis, i.e., the alternative hypothesis was not just a chance outcome, but we are 95% or 99% confident that it is more likely than the null hypothesis

Confidence interval

- Given sample $x_1, \ldots, x_N$ and sample mean $\bar{x}$
- What is the interval $\bar{x} \pm \varepsilon$ within which the true mean will lie with confidence $1 - \alpha$ (e.g. 95%)

$$Pr(|\bar{x} - \mu| > \varepsilon) < \alpha$$

- For Gaussian distribution

$$\varepsilon = z_{\alpha/2} \frac{\sigma}{\sqrt{N}}$$

- Standard Gaussian is used to define z
- Replace σ by $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N}}$

  for unknown σ

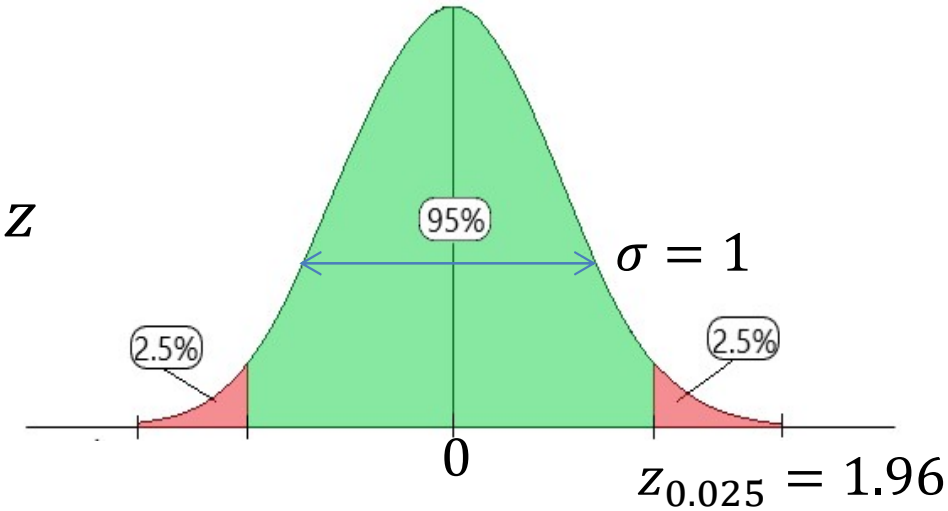


Image source: https://www.geeksforgeeks.org/confidence-interval/

Comparing means of two independent set of samples

- Given samples from two distributions, can we say with confidence that their means differ?
- $\mu_X = \frac{1}{n_X}\sum_{i=1}^{n_X} x_i, \qquad \sigma_X = \sqrt{\frac{1}{n_X}\sum_{i=1}^{n_X}(x_i - \mu_X)^2}$
- $\mu_Y = \frac{1}{n_Y}\sum_{i=1}^{n_Y} y_i, \qquad \sigma_Y = \sqrt{\frac{1}{n_Y}\sum_{i=1}^{n_Y}(y_i - \mu_Y)^2}$
- Welch's t-test: $t = \frac{\mu_X - \mu_Y}{\sqrt{\frac{{\sigma_X}^2}{n_X} + \frac{{\sigma_Y}^2}{n_Y}}}$ is matched to a table for the appropriate degrees of freedom (DoF): $\frac{\left(\frac{{\sigma_X}^2}{n_X} + \frac{{\sigma_Y}^2}{n_Y}\right)^2}{\frac{{\sigma_X}^4}{{n_X}^2(n_X-1)} + \frac{{\sigma_Y}^4}{{n_Y}^2(n_Y-1)}}$

Comparing means of paired-samples

Pre-yoga blood sugar	Post-yoga blood sugar	Δ
x_1	y_1	$d_1 = x_1 - y_1$
…	…	…
x_N	y_N	$d_N = x_N - y_N$

- Is there a post-event effect in a variable?
- E.g., “Does yoga lower blood sugar?”
- Mean of the difference $\overline{d} > 0$ with 95% confidence?
- Treat d as a random variable
- Is $\overline{d} - 1.8\frac{\sigma_d}{\sqrt{N}} > 0$, where $1.8 = z_{0.05}$?

Comparing paired variables without assuming a distribution

- Let there be two paired continuous variables
- We can compare their medians, if we do not want to assume a distribution, using Wilcoxon signed rank test
- Add all the ranks of positive Δ and negative Δ separately, and pick the smaller sum or ranks as test stat w_{test}
- Test stat should be smaller than $w_{critical}$ from a table for the given *N*

Pre-yoga blood sugar	Post-yoga blood sugar	Δ	|Δ|	Rank of |Δ|
x_1	y_1	$d_1 = x_1 - y_1$	$\|d_1\|$	r_1
…	…	…	…	…
x_N	y_N	$d_N = x_N - y_N$	$\|d_N\|$	r_N

Are two variables linearly related?

- Let there be two paired continuous variables
- Pearson's correlation coefficient
- $\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathrm{E}[X,Y] - \mathrm{E}[X]\mathrm{E}[Y]}{\sqrt{\mathrm{E}[X^2] - \mathrm{E}[X]^2}\sqrt{\mathrm{E}[Y^2] - \mathrm{E}[Y]^2}}$
- For a sample $r_{x,y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$
- Ranges from −1 to +1
- Does not imply causation, nor models nonlinear relations

Are two variables monotonically related?

- Find Spearman's correlation, which is Pearson's correlation between ranks of X and Y

$$r_S = \rho_{\mathrm{R}(X),\mathrm{R}(Y)} = \frac{\mathrm{Cov}(\mathrm{R}(X), \mathrm{R}(Y))}{\sigma_{\mathrm{R}(X)}\sigma_{\mathrm{R}(Y)}} = 1 - \frac{6\sum(\mathrm{R}(x_i) - \mathrm{R}(y_i))^2}{n(n^2-1)}$$

Some common statistical tests

Predictor	Outcome	Example	Parameteric test	Non-parameteric test
Categorical binary	Numerical unpaired	Do joggers have lower pulse rate than non-joggers	Independent t-test	Wilcoxon rank-sum test
Categorical binary	Numerical paired	Does blood sugar reduce after yoga	Paired t-test	Sign test, Wilcoxon signed-rank test
Numerical	Numerical	Are height and weight related	Pearson's correlation	Spearman's correlation
Categorical	Categorical	Does species predict color		Chi-square

How to choose a stat test

- Frame your problem
  - Predictor and outcome variable types
  - Decision to be expected
- Check if a widely accepted test is already available
- Check if the assumptions behind the test are applicable to your scenario
- Else, make your own test by using an existing test as a base for approach

Welcome to:

EE353 Intro to Data Science and Machine Learning

EE769 Intro to Machine Learning

Intro to RDBMS and SQL

Amit Sethi, EE (and KCDH, CMINDS, DSSE), IITB

MeDAL Lab (1st flr, EE), 3528, 7483, asethi@iitb.ac.in

Learning objectives

- List the key concepts of relational databases
- Define table, record, field, key
- Define structure of basic SQL queries

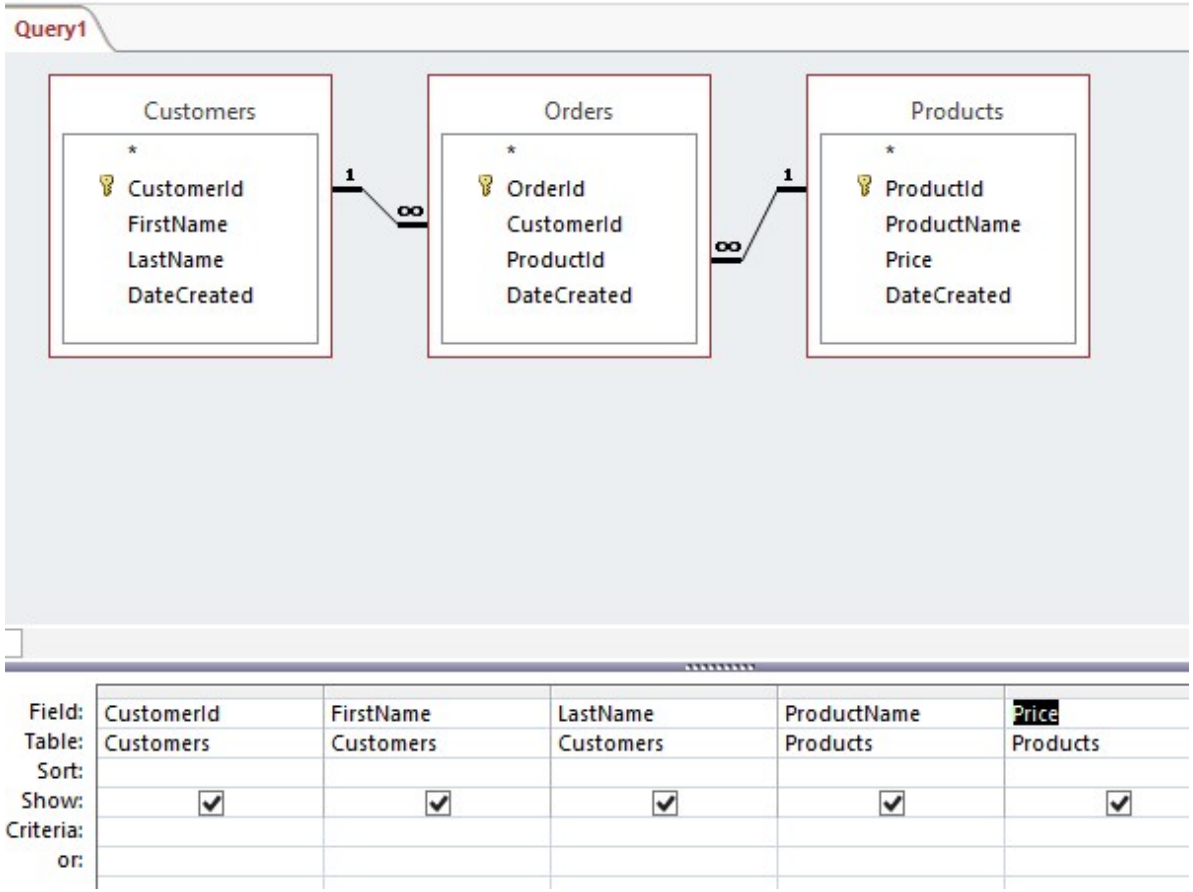
Introduction to Databases

- What is a Database?
  - Organized collection of data
  - Stores information for efficient retrieval and management
  - Examples: Customer records, product inventory, financial data
- Why Databases?
  - Manage large amounts of data efficiently
  - Ensure data integrity and consistency
  - Facilitate data sharing and analysis

Relational Database Management System (RDBMS)

- Stores data in tables with rows and columns
- Relationships defined between tables
- Most widely used database system

Image source: https://techcommunity.microsoft.com/t5/access-blog



RDBMS Concepts

- Tables
  - Organized collection of data in rows and columns
  - Each row represents a record, each column a field
- Records
  - Represents a single instance of data (e.g., Customer records)
- Fields
  - Represents a specific piece of data in a record (e.g., Name, address, phone number)
- Keys
  - Unique identifiers for records
  - Primary key: Unique identifier for each row
  - Foreign key: References primary key in another table

Structured Query Language (SQL)

- Standard language for interacting with RDBMS
- Used for data definition, manipulation, and control
- E.g.
  - CREATE TABLE Enrolled (sid: CHAR(20), cid: CHAR(20), grade: CHAR(2))
  - INSERT INTO Students (sid, name, login, age, gpa) VALUES (53688, ‘Smith’, ‘smith@ee’, 18, 3.2)
  - DELETE FROM Students S WHERE S.name = ‘Smith’

Example source: https://pages.cs.wisc.edu/~dbbook/openAccess/firstEdition/slides/pdfslides/mod1l2.pdf

Typical query structure in SQL

- FROM table(s)
- WHERE (records with following) conditions
- SELECT only the required fields

RDBMS Self-study

- Watch the following and practice in Colab:

- https://www.youtube.com/watch?v=h0nxCDiD-zg
- https://www.youtube.com/watch?v=v0gpCrPqJSs
- https://www.youtube.com/watch?v=pO-EeLn9IsU

ML for Smart Monkeys

Amit Sethi

Faculty member, IIT Bombay

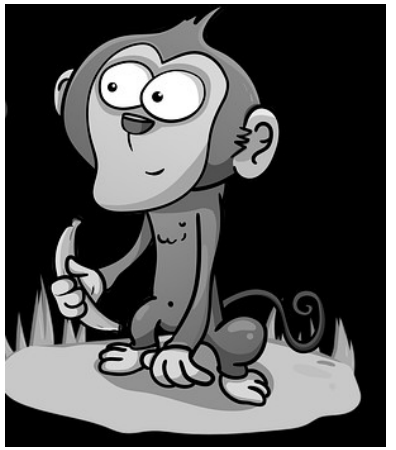


Image source: Pixabay.com

ML is…

- The practice of automating the use of related data to estimate models that make useful predictions about new data, where the model is too complex for standard statistical analysis, e.g.
  - Improve accuracy of classification of images using labeled images
  - Improve win percentage on alpha-go using several simulated game move sequences and their results
  - Improve the Turing test confusion between human and machine for NLP Q&A using a large sample of text including Q&A

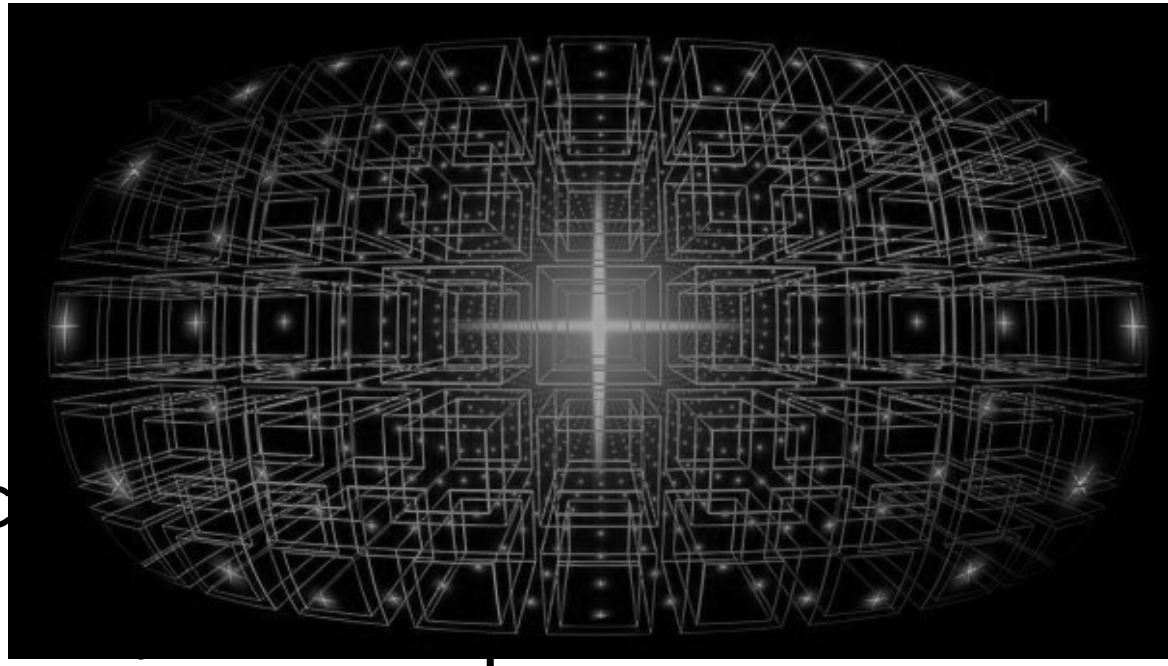
When not to use ML

- Possible inputs are countable and few
  - Use look up tables
- Algorithm is well-known and efficient
  - E.g. sorting, Dijkstra's shortest path
- Model is well-known and tractable
  - Use statistical estimation
- There is no notion of contiguity
  - Use dicrete variable methods or give up
- Lack of data
  - Use transfer learning or few-shot learning, or give up

When to use ML

- Possible inputs are many or continuous
- No well-known or efficient algorithm
- Model is not well-known or tractable
- Strong notion of contiguity
- Good amount of data
- Desired output known
- Well-defined inputs

Sweet spot for ML

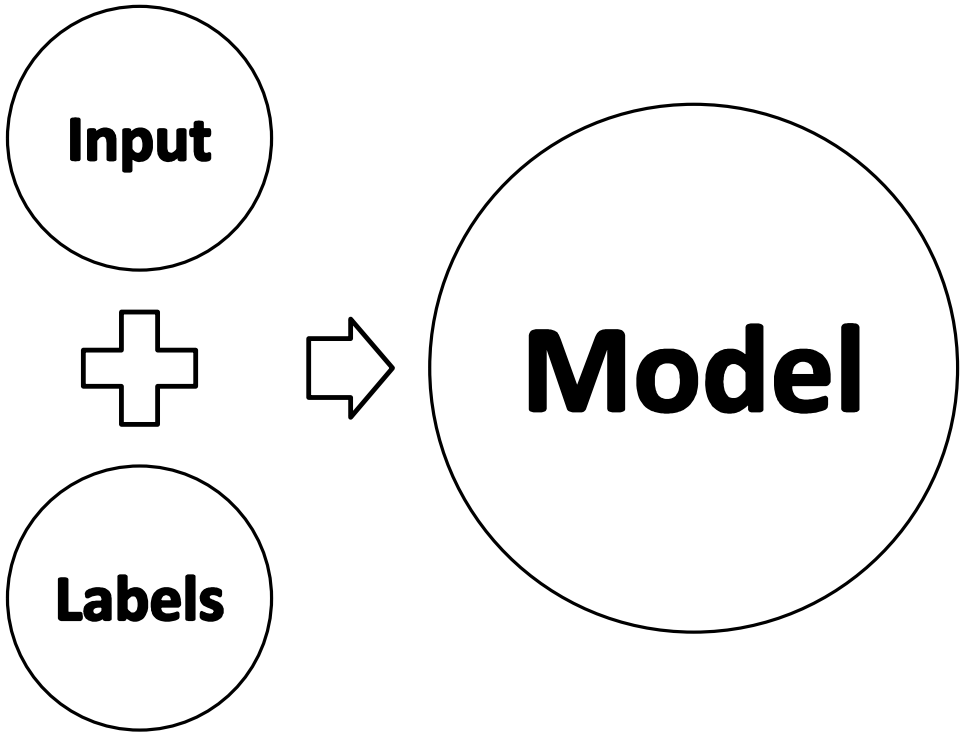


- Lots of structured data
- Explainability is not critic
- Prediction accuracy is the primary goal
- Underlying model is complex but stationary

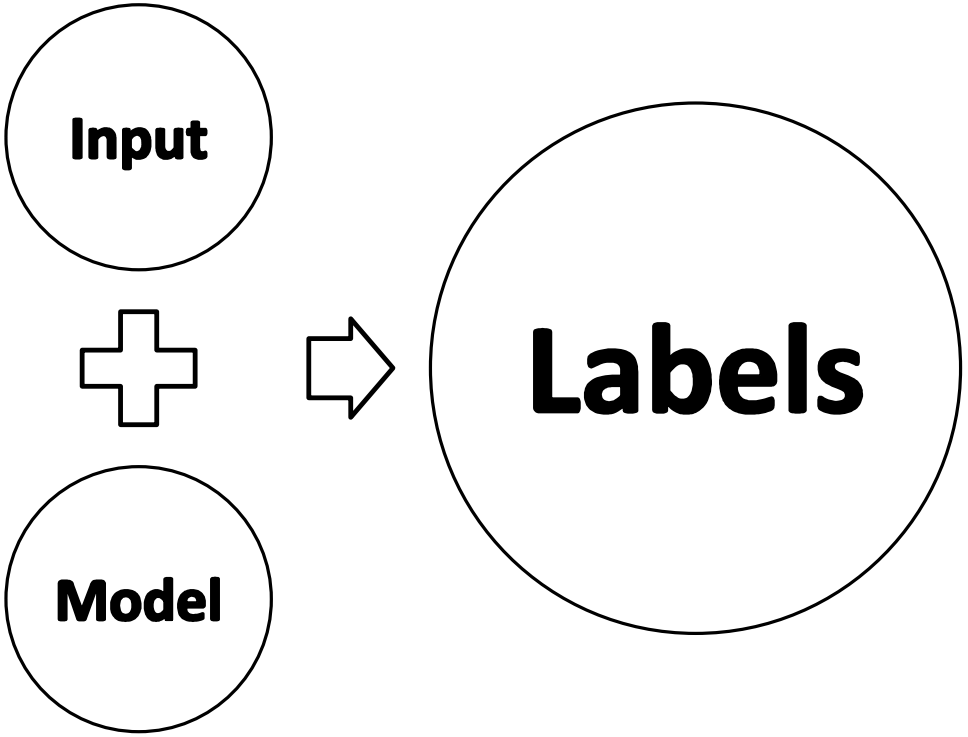
Image courtesy: Pixabay.com

ML model training and deployment

Training on past data

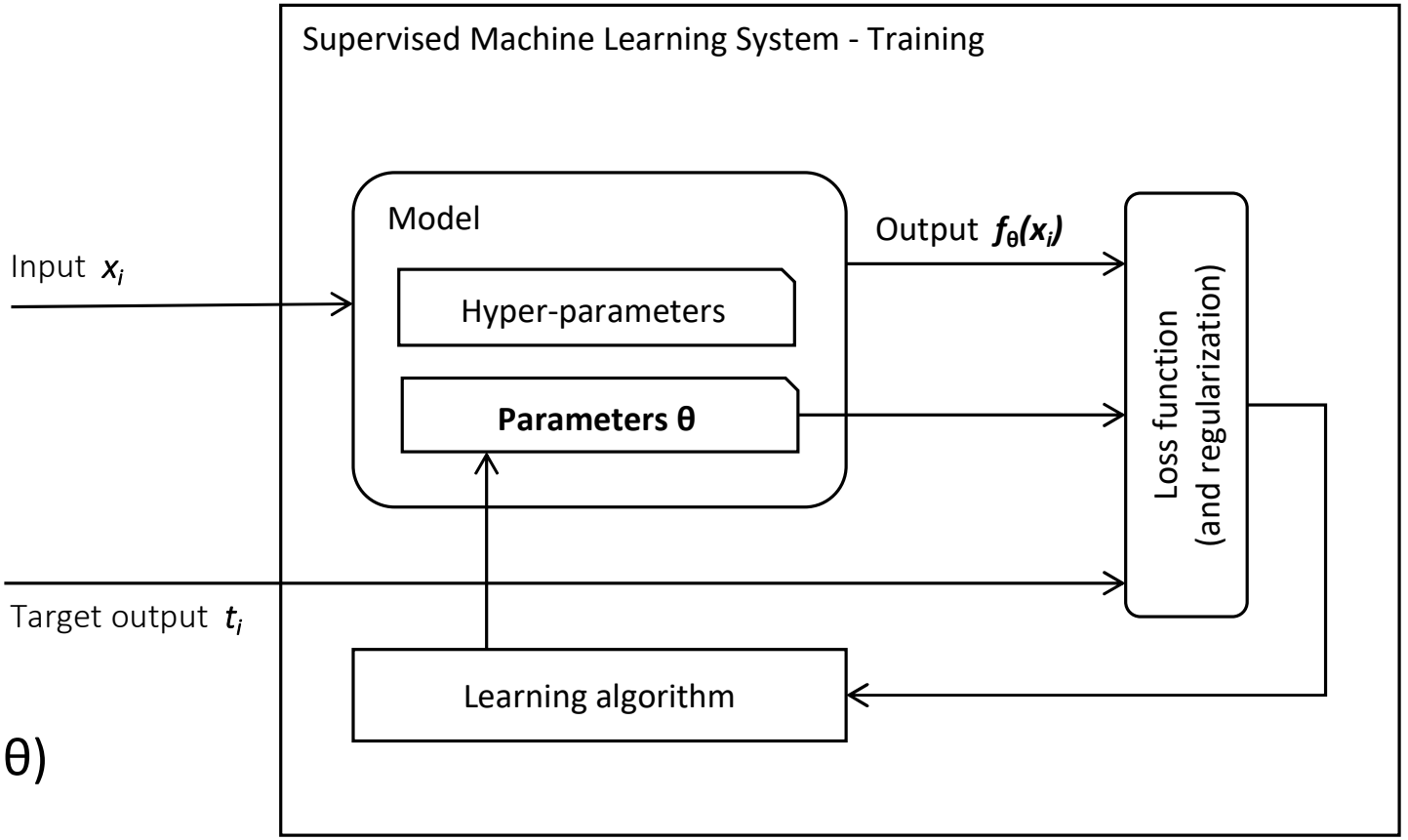


Prediction on future data

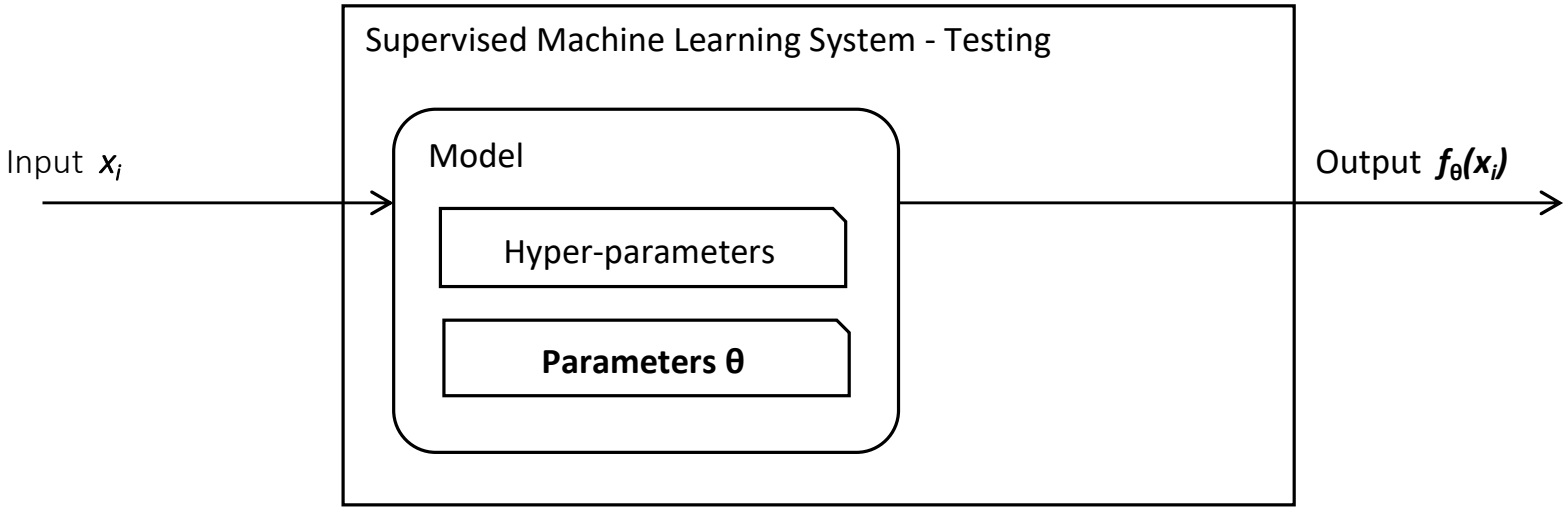


ML gives a model

- Elements of a model:
  - Input x_i
  - Function $f_\theta(x_i)$
- Utility of the model:
  - Target output t_i
  - Bring $f_\theta(x_i)$ close to t_i
  - Minimize loss $L(t_i, f_\theta(x_i), \theta)$



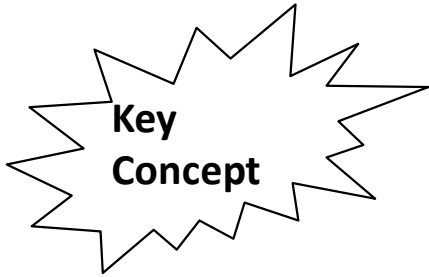
Components of a Trained ML System



Mathematically speaking…

- Determine f such that $t_i = f(x_i)$ and $g(T, X)$ is minimized for unseen set T and X pairs, where T is the ground truth that cannot be used
- Form of f is fixed, but some parameters can be tuned:
  - So, $y = f_\theta(x)$, where, x is observed, and y needs to be inferred
  - e.g. $y = 1$, if $mx > c$, $y = 0$ otherwise, so $\theta = (m, c)$
- Machine Learning is concerned with designing algorithms that learn "better" values of θ given "more" x (and t) for a given problem

Parameters and Hyperparameters



- Parameters: These are the variable whose values are updated during the training process of model.
  - Feature coefficient in regression model
  - Weights of a neural network
- Hyperparameters: These are the variables/ parameter whose values are fixed by model developer before the beginning of learning process.
  - Number of variables in a tree node
  - Height of a tree
  - Number of layers of a neural network

Type of ML problems

- Supervised learning: uses labeled data
  - Classification: Labels are discrete
  - Regression: Labels are continuous
  - Ranking: Labels are ordinal
- Unsupervised learning: uses unlabeled data
  - Clustering: Divide data into discrete groups
  - Dimension reduction: Represent data with fewer numbers
- Somewhere in between: fewer labels than one per example
  - Semi-supervised learning: some examples are labeled
  - Weakly supervised learning: groups of examples are labeled
  - Reinforcement learning: Label (reward) is available after a sequence of steps

Supervised Learning

- Predictor variables/features and a target variable (label)
- Aim: Predict the target variable (label), given the predictor variables
  - **Classification**: Target variable (y) consists of categories
  - **Regression**: Target variable is continuous

Predictor variables

Target variable **(Label)**

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

Broad types of ML problems

Output →	**Categorical**	**Ordinal**	**Continuous**
Supervised	Classification	Ranking	Regression
(Examples)	{Cats, dogs}	{Low, Med, High}	[-20,+10)
Unsupervised	Clustering		Dimension reduction

Some popular ML frameworks

<table>
<tr><th></th><th>Classification</th><th>Regression</th><th>Clustering</th><th>Dimension reduction</th></tr>
<tr><td rowspan="2">Vector</td><td>Logistic regression</td><td>Linear regression</td><td rowspan="2">K-means,
Fuzzy C-means,
DB-SCAN</td><td rowspan="2">PCA, k-PCA,
LLE,
ISOMAP</td></tr>
<tr><td colspan="2">SVM, RF, NN</td></tr>
<tr><td>Series, text</td><td colspan="2">RNN, LSTM, Transformer,
1-D CNN, HMM</td><td></td><td></td></tr>
<tr><td>Images</td><td colspan="2">2-D CNN, MRF</td><td></td><td></td></tr>
<tr><td>Video, MRI</td><td colspan="2">3-D CNN, CNN+LSTM, MRF</td><td></td><td></td></tr>
</table>

Recipe for ML training

- Decide on the type of the ML problem
- Prepare data
- Shortlist ML frameworks
- Prepare training, validation, and test sets
- Train, validate, repeat
- Use test data only once

Preparing data

- Remove useless data
  - No variance
  - Falsely assumed to be available
- Reduce redundancy
  - Correlated
    - Pearson and Spearman
- Handle missing data
  - Impute, if sporadic
  - Drop, if too frequent
- Transform variables
  - Convert discrete to one-hot-bit
  - Normalize continuous variables

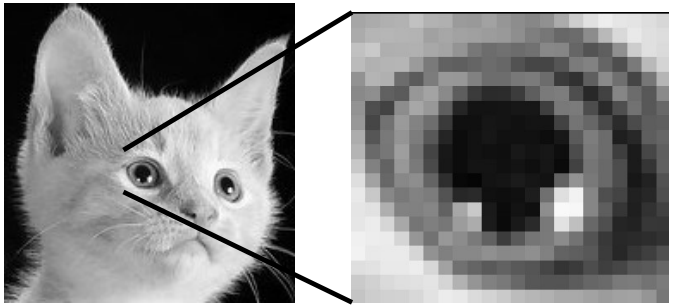
Examples of structure in the data

- Records

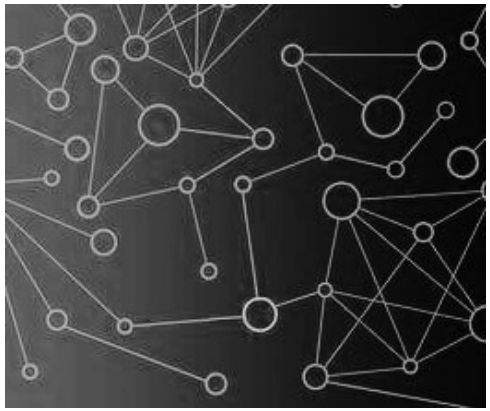
Product SKU	Price	Margin	Volume
A123ajkhdf	$ 120	30%	1,000,000
B456ddsjh	$200	10%	2,000,000

- Temporal order

- Spatial order

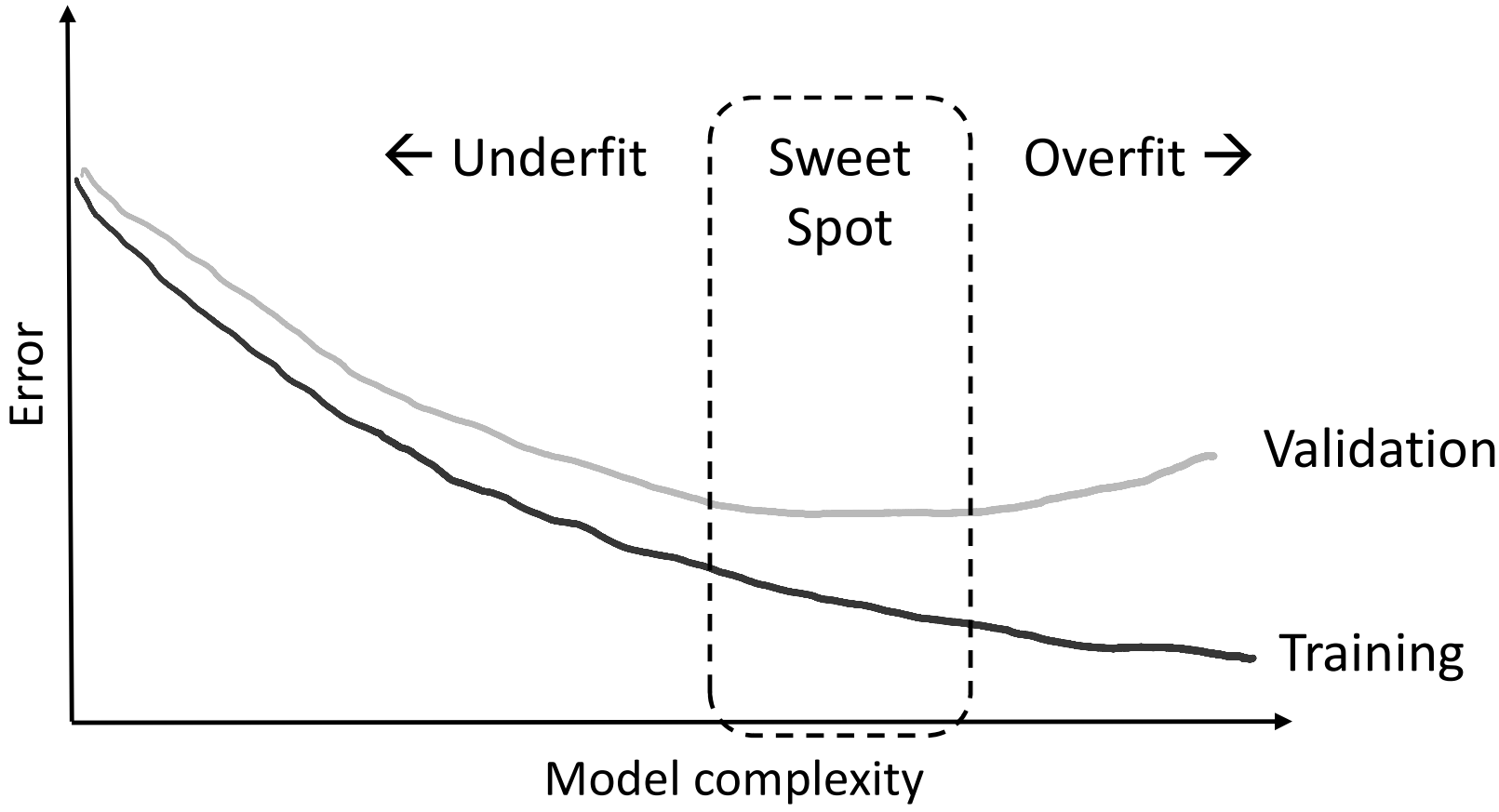


- Web of relationships



Images courtesy: Pixabay.com

Model choice and rigorous validation are very important



Bias-variance trade-off

Generalization of model is bounded by the two undesirable outcomes high bias and high variance.

- Underfitting: High bias, Low variance
- Overfitting: Low bias, High variance

Bias occurs when an algorithm has limited flexibility to learn the true signal from the dataset. High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting).

Variance is an error from sensitivity to small fluctuations in the training set. High variance can cause an algorithm to model the random noise in the training data, rather than the intended outputs (overfitting).

Regularization is a key concept in ML

- Regularization means constraining the model
- More constraints may reduce model fit on training data
- However, it may improve fit on validation and test data
- Training performance of more constrained models are more likely to reflect test performance

Loss versus performance metric

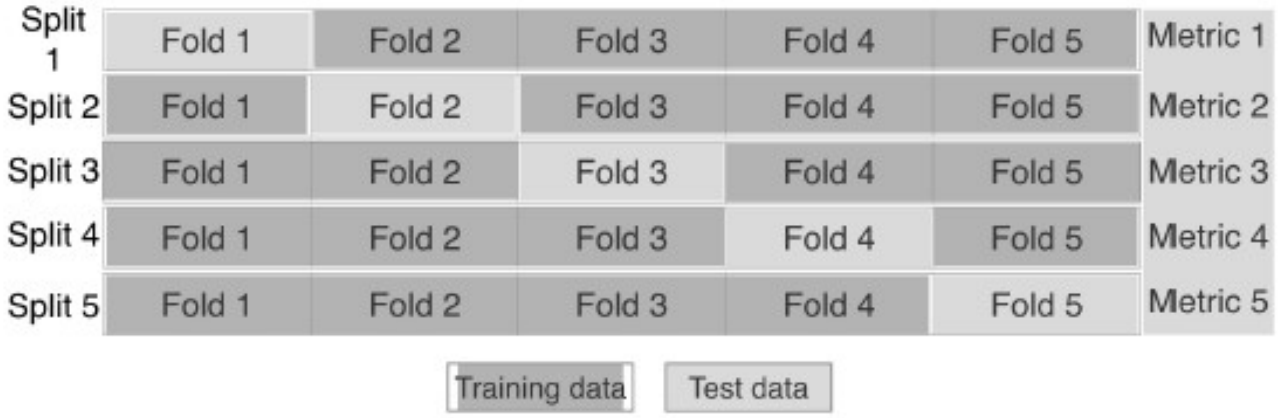
- Loss is a convenient expression used for guiding the learning (optimization)
- Loss is related to performance metric, **but** it is not the same
- Loss also includes regularization
- Performance metric is what is used to judge the model
- Performance metric on only the held-out (validation or test) data makes sense

Preparing data for training and validation

- Data splits:
  - Training → Used to optimize the parameters (e.g. random 70%)
  - Validation → Used to compare models (e.g. random 15%)
  - Testing → One final check after multiple rounds of validation (e.g. random 15%)
- Cross-validation:
  - K-folds: One fold for validation, K-1 folds for training
  - Rotate folds K times
  - Select framework (hyperparameters) best average performance
  - Re-train best framework on entire data
  - Test one final time on held-out data that was not a part of any fold

Cross-validation

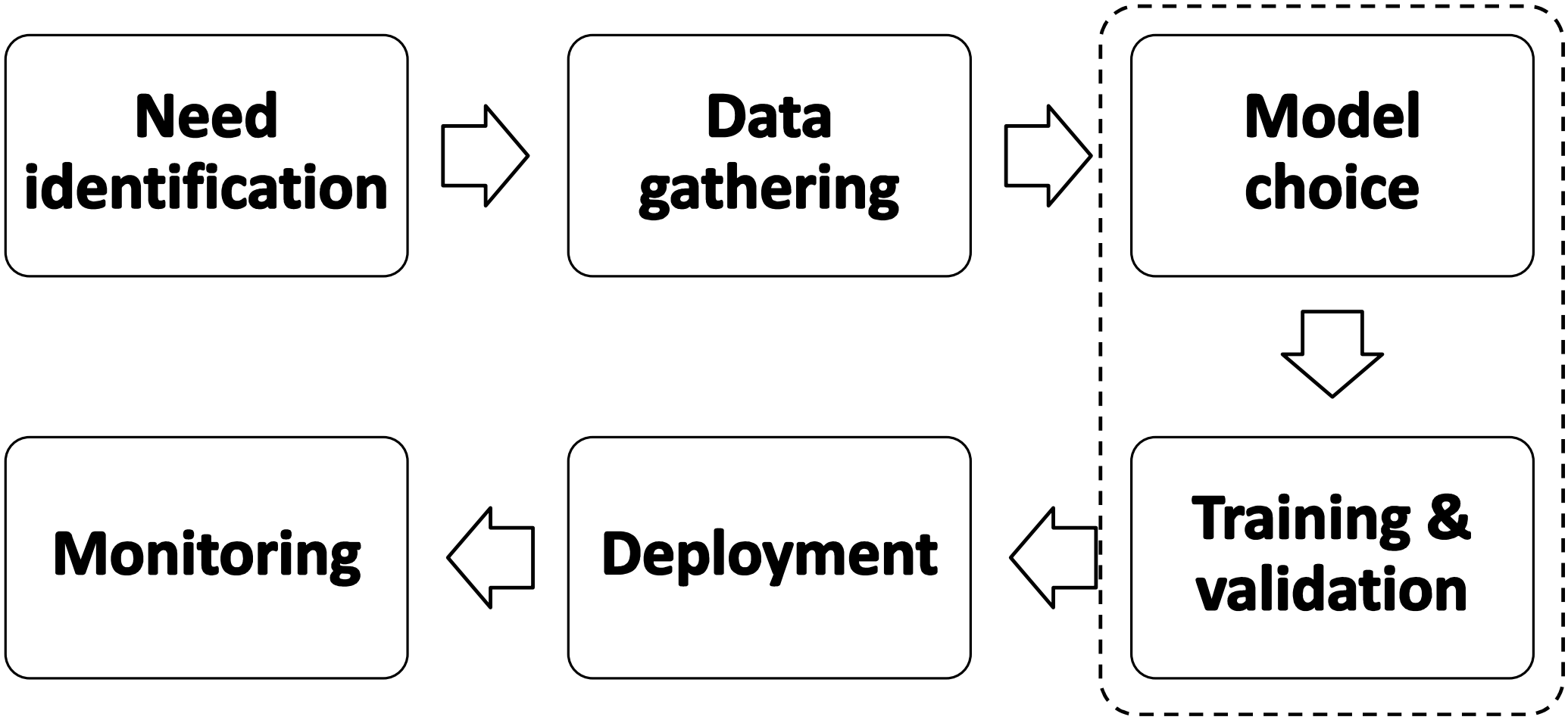
- Model performance measurement is dependent on way the data is split
- Not representative of the model's ability to generalize
- Solution: Cross-validation, especially when data is less
- Con: more computations



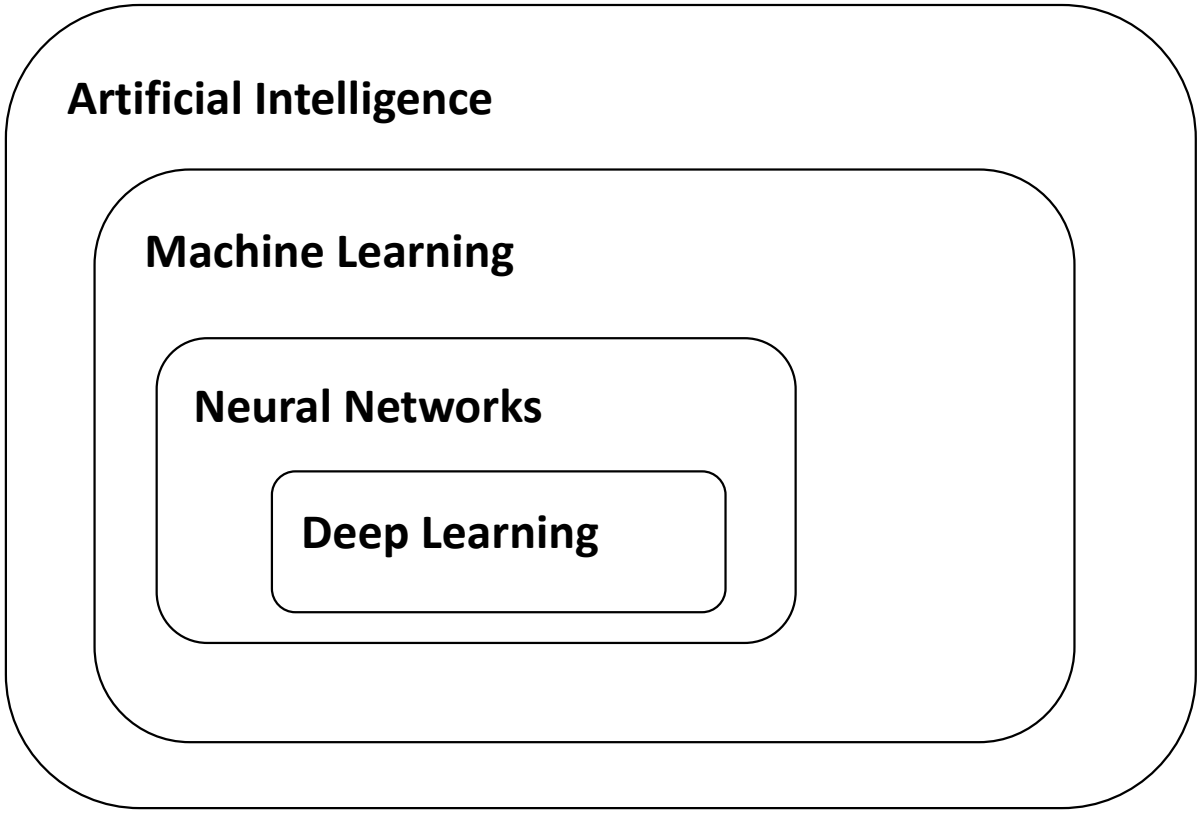
ML can fail to perform in deployment

- Lack of training diversity: data had limited confounders
  - Single speaker, author, camera, background, accent, ethnicity, etc.
  - Data imbalance between high-value rare and more common examples
- Proxy label leak during training:
  - E.g. Only speakers A and B provide emotion “anger,” so ML confused their voice characteristics with “anger”
- Too much manual cleansing of training data
- Too little training data, and very complex models
- Concept drift: The assumptions behind training are no longer valid

ML life stages



Relation of ML to other fields



Relation of ML to other fields

EE769 Intro to ML

Linear Regression

Amit Sethi
Faculty member, IIT Bombay

Learning outcomes

- Write the objective of regression
- Write the expression for analytical solution
- Decompose the loss into bias and variance
- Add regularization to least square
- Write the algorithm for computational solution

Outline

- Utility of linear regression
- Problem setup
- Bias-variance decomposition
- Regularization
- Iterative solution

Utility of linear regression

- Simplest to code
- Simplest to understand mathematically
- Nonlinear feature extraction can be used with linear solution

Problem setup

- Assuming ideal prediction is linear w.r.t. input

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + \ldots + w_D x_D$$

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x})$$

- Or linear with respect to features

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})$$

Source: PRML book by C Bishop

Measured output has stochastic noise

- Assuming Gaussian noise

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon$$

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1})$$

Source: PRML book by C Bishop

Likelihood of observation

- Assuming identically independently distributed samples

$$\ln p(\mathbf{t}|\mathbf{w},\beta) = \sum_{n=1}^{N} \ln \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n),\beta^{-1})$$
$$= \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) - \beta E_D(\mathbf{w})$$

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2$$

Source: PRML book by C Bishop

Maximizing likelihood

- Set gradients w.r.t. w to zero

$$\nabla \ln p(\mathbf{t}|\mathbf{w}, \beta) = \sum_{n=1}^{N} \left\{ t_n - \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n) \right\} \boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}}$$

$$0 = \sum_{n=1}^{N} t_n \boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}} - \mathbf{w}^{\mathrm{T}} \left(\sum_{n=1}^{N} \boldsymbol{\phi}(\mathbf{x}_n) \boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}} \right)$$

$$\mathbf{w}_{\mathrm{ML}} = \left(\boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi} \right)^{-1} \boldsymbol{\Phi}^{\mathrm{T}} \mathbf{t}$$

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix}$$

Source: PRML book by C Bishop

Expected loss (MSE) (1/3)
Bias-variance decomposition

$$\mathbb{E}[L] = \iint \{y(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\, \mathrm{d}\mathbf{x}\, \mathrm{d}t$$

$$\frac{\delta \mathbb{E}[L]}{\delta y(\mathbf{x})} = 2 \int \{y(\mathbf{x}) - t\} p(\mathbf{x}, t)\, \mathrm{d}t = 0$$

$$y(\mathbf{x}) = \frac{\int t p(\mathbf{x}, t)\, \mathrm{d}t}{p(\mathbf{x})} = \int t p(t|\mathbf{x})\, \mathrm{d}t = \mathbb{E}_t[t|\mathbf{x}]$$

$$\begin{aligned} \{y(\mathbf{x}) - t\}^2 &= \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}] + \mathbb{E}[t|\mathbf{x}] - t\}^2 \\ &= \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 + 2\{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}\{\mathbb{E}[t|\mathbf{x}] - t\} + \{\mathbb{E}[t|\mathbf{x}] - t\}^2 \end{aligned}$$

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 p(\mathbf{x})\, \mathrm{d}\mathbf{x} + \int \{\mathbb{E}[t|\mathbf{x}] - t\}^2 p(\mathbf{x})\, \mathrm{d}\mathbf{x}.$$

Source: PRML book by C Bishop

Expected loss (MSE) (2/3)
Bias-variance decomposition

$$h(\mathbf{x}) = \mathbb{E}[t|\mathbf{x}] = \int t p(t|\mathbf{x})\,\mathrm{d}t$$

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - h(\mathbf{x})\}^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x} + \int \{h(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t$$

$$\{y(\mathbf{x}; \mathcal{D}) - h(\mathbf{x})\}^2$$

$$\begin{aligned}
&\{y(\mathbf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] + \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] - h(\mathbf{x})\}^2 \\
&= \{y(\mathbf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})]\}^2 + \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] - h(\mathbf{x})\}^2 \\
&\quad + 2\{y(\mathbf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})]\}\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] - h(\mathbf{x})\}.
\end{aligned}$$

Source: PRML book by C Bishop

Expected loss (MSE) (3/3)
Bias-variance decomposition

$$\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - h(\mathbf{x})\}^2\right]$$
$$= \underbrace{\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2}_{(\text{bias})^2} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right]}_{\text{variance}}$$

$$\text{expected loss} = (\text{bias})^2 + \text{variance} + \text{noise}$$

$$(\text{bias})^2 = \int \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

$$\text{variance} = \int \mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right] p(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

$$\text{noise} = \int \{h(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t$$

Source: PRML book by C Bishop

Regularization using L2 penalty

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q \qquad \sum_{j=1}^{M}|w_j|^q \leqslant \eta$$

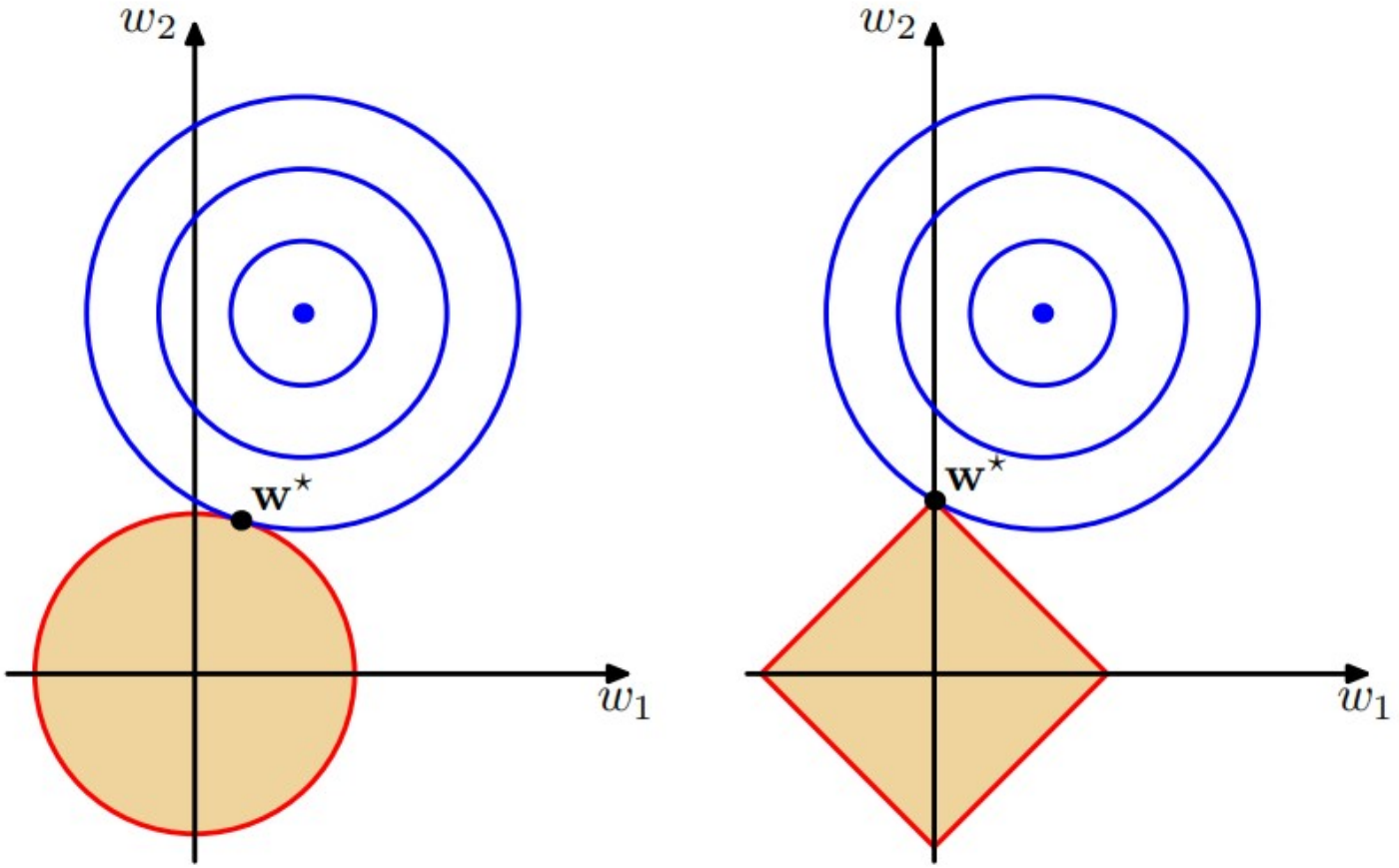
$$E_D(\mathbf{w}) + \lambda E_W(\mathbf{w}) \qquad E_W(\mathbf{w}) = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

$$\mathbf{w} = \left(\lambda\mathbf{I} + \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}$$

Source: PRML book by C Bishop

Geometry of Lp regularization



Source: PRML book by C Bishop

Iterative optimization

- Iterative minimization of convex objectives

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \nabla E_n$$

- Recursive least square update

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} + \eta(t_n - \mathbf{w}^{(\tau)\mathrm{T}}\boldsymbol{\phi}_n)\boldsymbol{\phi}_n$$

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} + \eta(t_n - \mathbf{w}^{(\tau)\mathrm{T}}\boldsymbol{\phi}_n)\boldsymbol{\phi}_n \; - \eta\lambda\, \mathbf{w}^{(\tau)}$$

Source: PRML book by C Bishop

EE769 Intro to ML

Linear Classification

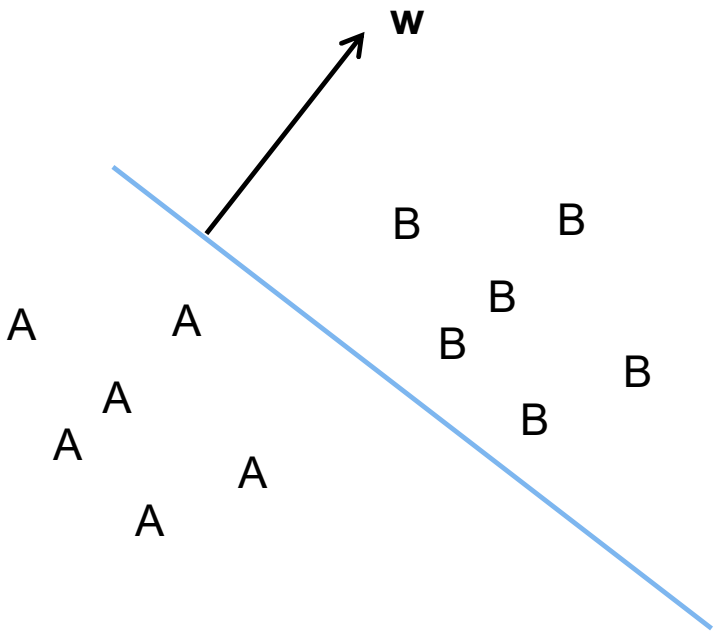
Amit Sethi
Faculty member, IIT Bombay

Learning objectives

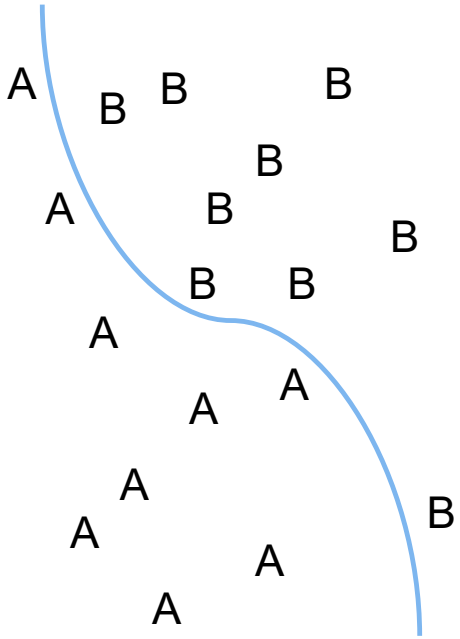
- Write the linear classification equation
- Write the Bayesian decision function
- Ground logistic regression in theory
- Derive gradient descent for logistic regression
- Derive the loss function for primal support vector machine

Linear classification function

- Class $y_i = \text{Sign}(\mathbf{w}^T x_i + b) \in \{-1,1\}$



Linear in **x**



Nonlinear in **x** (possibly linear in **ϕ**)

Why study linear classifiers?

- It is one of the simplest classifiers to analyze
- It seems to be a natural outcome for a familiar useful types of class conditional densities
- Many nonlinear problems can be linearized
- Multi-class classification can be modeled as a combination of several binary classification problems

Linearizing nonlinear problems

- Add derived features
  - Powers
  - Interaction terms
  - Kernels
- Extract features
  - Using pre-trained neural networks

Bayesian decision rule for classification

- Decision rule: Class is 1, if $p(t=1|\mathbf{x}) > p(t=0|\mathbf{x})$

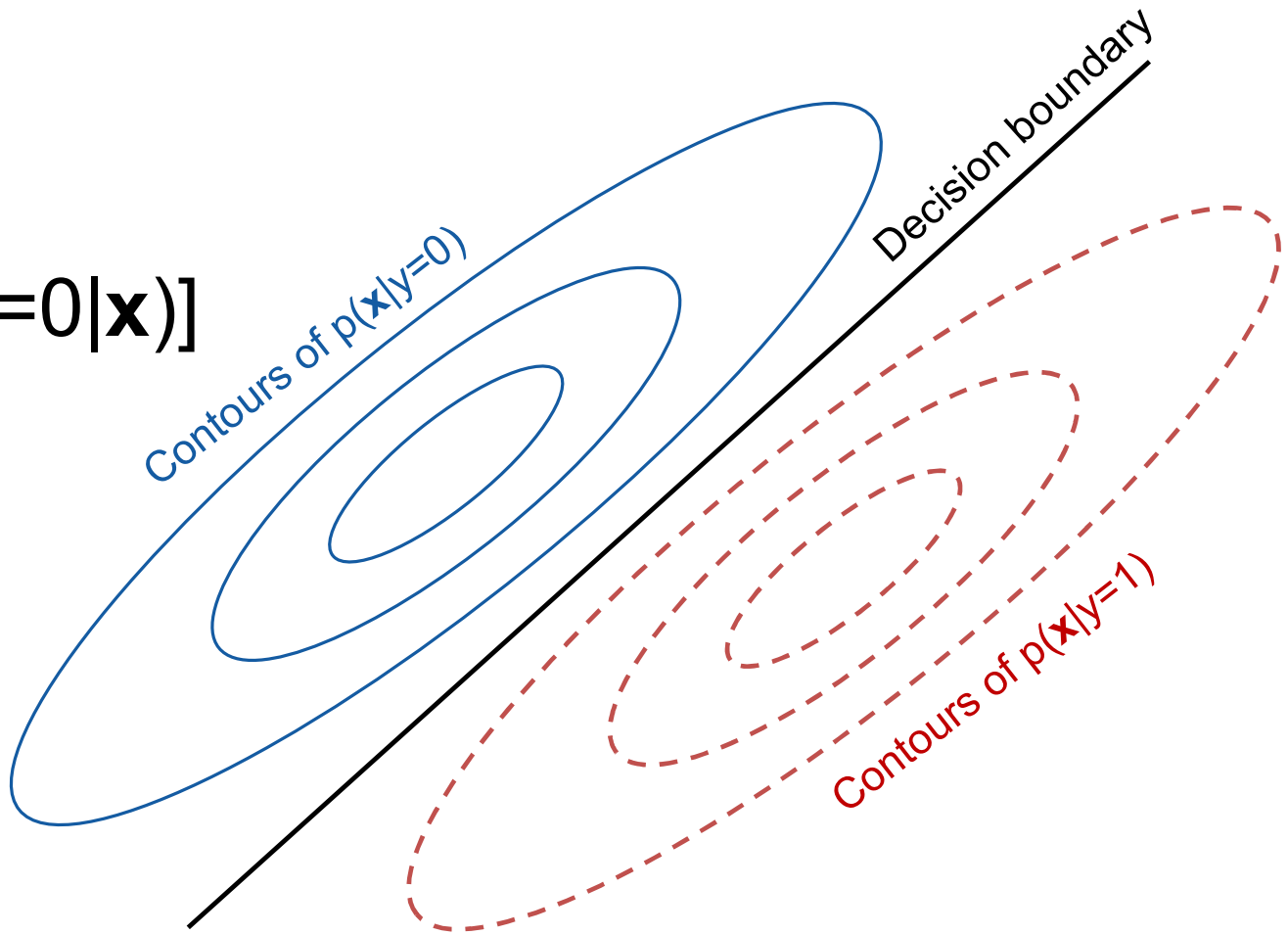
- Problem: Do not know how to model $p(t|\mathbf{x})$ directly

- Solution: Bayes rule $p(t|\mathbf{x}) = p(\mathbf{x}|t) \; . \; p(t) \; / \; p \; (\mathbf{x})$
- Posterior = Likelihood . prior / marginal
- Marginal is unknown, but common to both classes
- Class is 1, if $p(\mathbf{x}|t=1) \; p(t=1) > p(\mathbf{x}|t=0) \; p(t=0)$

Gaussian class conditionals in with the same covariance matrix

- $p(\mathbf{x}|t=j) = N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$
- $p(t=j) \in [0,1]$
- $\log [p(t=1|\mathbf{x}) / \log p(t=0|\mathbf{x})]$

  is linear in $\mathbf{x}$, if $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1$



Gaussian class conditionals in with the same covariance matrix

Derivation:

- To check if $p(t=1|\mathbf{x}) > p(t=0|\mathbf{x})$
- Check if $p(t=1|\mathbf{x}) / p(t=0|\mathbf{x}) > 1$
- $p(\mathbf{x}|t=1)\, p(t=1) / p(\mathbf{x}|t=0)\, p(t=0) > 1$
- $\log[p(\mathbf{x}|t=1)] + \log p(t=1) - \log[p(\mathbf{x}|t=0)] - \log p(t=0) > 0$
- $\log[\exp(-(\mathbf{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1))] - \log[(2\pi)^{-d/2}\det(\boldsymbol{\Sigma})^{-1/2}] + \log p(t=1) -$
  $\log[\exp(-(\mathbf{x}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_0))] - \log[(2\pi)^{-d/2}\det(\boldsymbol{\Sigma})^{-1/2}] - \log p(t=0) > 0$
- $-(\mathbf{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + \log p(t=1) + (\mathbf{x}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_0) - \log p(t=0) > 0$
- $-\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - \boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + 2\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0 - 2\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0$
  $+ \log [p(t=1)/p(t=0)] > 0$
- $[2\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)]^T\,\mathbf{x} + [\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \log [p(t=1)/p(t=0)]\,] > 0$
- $\mathbf{w}^T\mathbf{x} + b > 0$

Algorithm to build a Bayesian classifier

- Count number of samples n_j in each class j
- Estimate priors p(t=j) as n_j / N, where N = $\mathbf{\Sigma}_j$ n_j
- Estimate class conditionals p(**x**|t=j), e.g. Gaussian
  - μ_j is the sample mean for class j
  - Σ_j is the sample covariance matrix for class j
- Decision rule: Class is 1, if

$$\log p(\mathbf{x}|t=1) + \log p(t=1) - \log p(\mathbf{x}|t=0) - \log p(t=0) > 0$$

Gradient descent for linear classifier

Loss

- Can use $(p(t_i|\mathbf{x}_i) - t_i)^2$, but we do not because this error is not Gaussian
- We use $D_{KL}\ (t_i \,||\, p(t_i|\mathbf{x}_i)) = -\sum_j \mathbf{1}_{t=j} \log[p(t_i=j|\mathbf{x}_i)/\mathbf{1}_{t=j}]$

- Which is BCE

$-\ [t_i \log p(t_i=j|\mathbf{x}_i) + (1-t_i) \log(1-(p(t_i=j|\mathbf{x}_i))]$

$-\ [t_i \log \sigma(h) + (1-t_i) \log(1-(\sigma(h))]$,

where $\sigma(h) = 1 / [1+\exp(-h)]$;

h = log of odds ratio = $\log [p(t_i=1|\mathbf{x}_i)/p(t_i=0|\mathbf{x}_i)]$

$= \mathbf{w}^T\mathbf{x}_i+b$

Gradient descent using BCE

- $h_i = \log[\, p(t_i=1|\mathbf{x}_i) \,/\, p(t_i=0|\mathbf{x}_i)\,] = \mathbf{w}^T\mathbf{x}_i + b$
- $y_i = p(t_i=1|\mathbf{x}_i) = \sigma(h_i) = 1 \,/\, [1+\exp(-h_i)]$
- $\mathrm{BCE} = L_i = -\, t_i \log y_i - (1-t_i) \log (1-y_i)$
- $\partial L_i/\partial w_k = \partial L_i/\partial y_i \,.\, \partial y_i/\partial h_i \,.\, \partial h_i/\partial w_k$

$$= [t_i/y_i - (1-t_i)/(1-y_i)] \,.\, y_i \,.\, (1-y_i) \,.\, x_{i,k}$$

$$= [(y_i-t_i) \,/\, y_i \,/\, (1-y_i)] \,.\, y_i \,.\, (1-y_i) \,.\, x_{i,k}$$

$$= (y_i-t_i) \,.\, x_{i,k}$$

$$\mathbf{w}_{new} = \mathbf{w}_{old} - \eta \sum_i (y_i-t_i)\, \mathbf{x}_i$$

Adding regularization

- $L = -\sum_i [t_i \log y_i + (1-t_i) \log (1-y_i)] + \lambda \sum_k |w_k|^q$
- $\mathbf{w}_{new} = \mathbf{w}_{old} - \eta [\sum_i (y_i - t_i) \mathbf{x}_i + 2\lambda \mathbf{w}_{old}]$, for q=2
- $\mathbf{w}_{new} = \mathbf{w}_{old} - \eta [\sum_i (y_i - t_i) \mathbf{x}_i + \lambda \, \mathrm{sign}(\mathbf{w}_{old})]$, for q=1

Detour -- Elastic Net

- Ridge (or L2 regularization or weight decay)
  - Minimize: $L_{error} + \lambda_2 ||w||_2^2$
  - **Grouping effect** on correlated variables
  - Encourages two correlated variables to have the same weight
- LASSO (or L1 regularization)
  - Minimize: $L_{error} + \lambda_1 ||w||_1$
  - May **eliminate variables**
  - Does not encourage two correlated variables to have the same weights
- Elastic net (or L1+L2 regularization) has both effects
  - Minimize: $L_{error} + \lambda_1 ||w||_1 + \lambda_2 ||w||_2^2$
  - May eliminate groups of correlated variables

Asymmetric risk

- Some risks are not symmetric
  - Calling a healthy person sick vs. vice versa
- We need a risk matrix
  - Perhaps, no risk for correct calls
  - But, different risks for Type I (FP) vs. Type II (FN) errors
- Minimize expected risk

Some metrics for binary classification

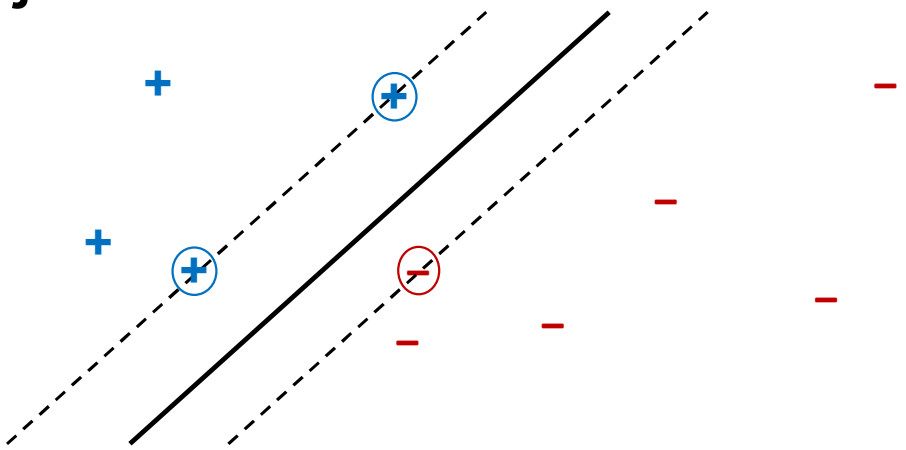
- For a single threshold
  - Accuracy = (TP + TN) / (TP + TN + FP + FN)
  - Precision = TP / (TP + FP)
  - Recall, Sensitivity = TP / (TP + FN)
  - Specificity = TN / (TN + FP)
  - PPV, NPV, FDR, FOR etc.
  - Balanced metric: F1 score = 2 × precision × recall / (precision + recall)
- For all thresholds
  - Receiver operating characteristic (ROC) curve: Plot of sensitivity (y-axis) versus (1 − specificity) by varying decision threshold
  - Area under curve (AUC): area under ROC (from 0 to 1)

SVM for distribution-free learning

- Empirical risk: risk of misclassifying trianing data
- How to minimize empirical risk?
- How to pick the “best” among multiple solutions?
- Depends upon the assumptions:
  - Bayesian: Minimize expected risk (by assuming pdf)
  - SVM: Minimize structural risk => margin maximization

SVM maximizes the separating margin

1. $\text{Max}_{\mathbf{w},b}\,[\min_i \|\mathbf{w}^T\mathbf{x}_i+b\|]$, subject to:
   a) $\|\mathbf{w}\| = 1$
   b) $\forall\ i,\ t_i\,(\mathbf{w}^T\mathbf{x}_i+b) \geq 0$



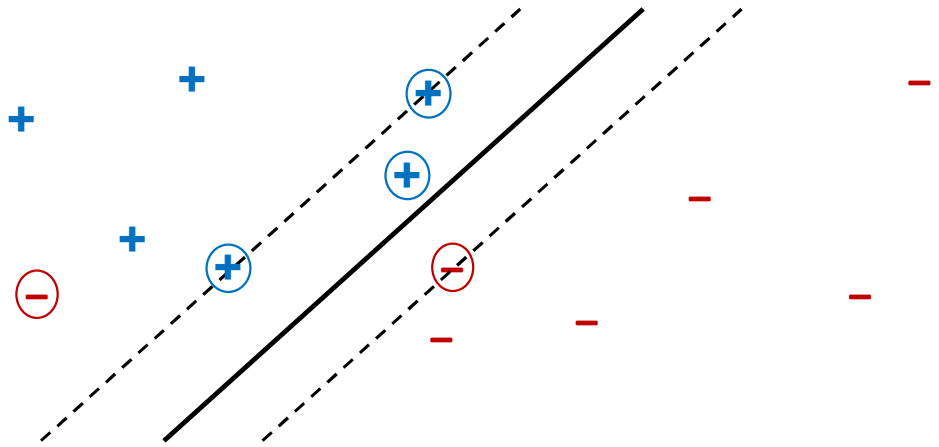
2. $\text{Min}_{\mathbf{w},b}\ \|\mathbf{w}\|^2$, s.t.
   a) $\forall i,\ t_i\,(\mathbf{w}^T\mathbf{x}_i+b) \geq 1$

3. $\text{Min}_{\mathbf{w},b}\ \sum_i \max[0,1 - t_i\,(\mathbf{w}^T\mathbf{x}_i+b)] \quad + \lambda\,\|\mathbf{w}\|;\quad \lambda\rightarrow 0.$
   $\text{Min}_{\mathbf{w},b}\ \sum_i [1 - t_i\,(\mathbf{w}^T\mathbf{x}_i+b)]_+ \quad + \lambda\,\|\mathbf{w}\|;\quad \lambda\rightarrow 0.$
   $\text{Min}_{\mathbf{w},b}\ \sum_i \ell_{\text{hinge}}(t_i,\mathbf{x}_i,\mathbf{w},b) \quad + \lambda\,\|\mathbf{w}\|;\quad \lambda\rightarrow 0.$

Soft-margin SVM formulation

- Or, minimize $||\mathbf{w}||^2 + C \sum_i \xi_i$
- ξ_i are slack variables
- Subject to $\forall\ i,\ t_i\ (\mathbf{w}^T\mathbf{x}_i) \geq 1-\xi_i$

  $\forall\ i,\ \xi_i \geq 0$

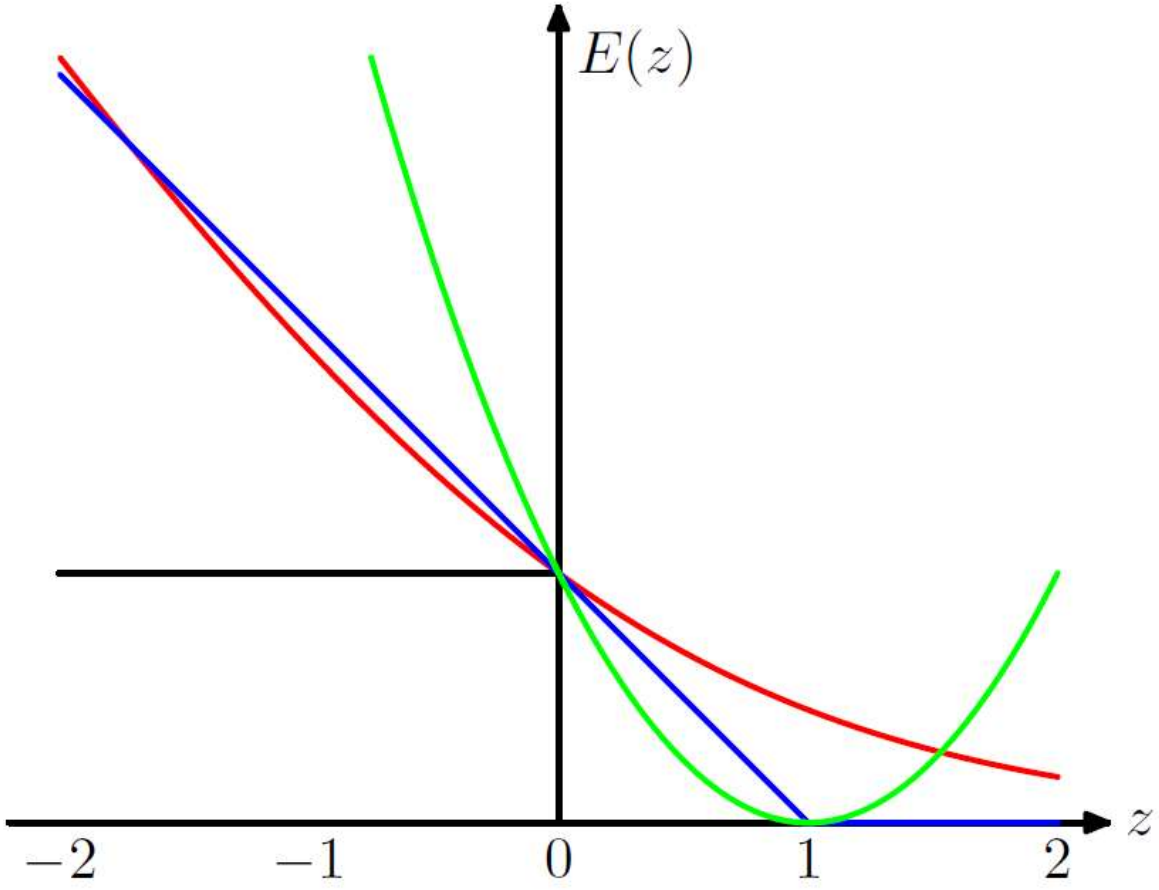


- $\text{Min}_{\mathbf{w},b} \sum_i \ell_{\text{hinge}}(t_i,\mathbf{x}_i,\mathbf{w},b) + \lambda\ ||\mathbf{w}||;\quad \lambda \geq 0.$

What are support vectors and slack variable?

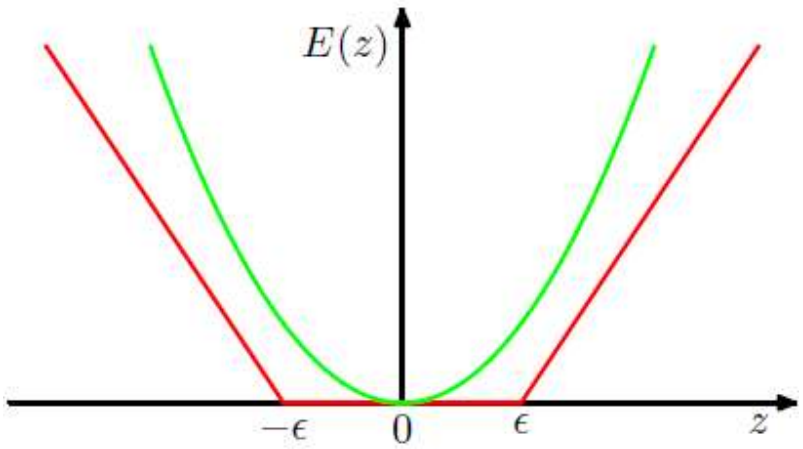
- In hard SVM
  - SVs define the hyperplane
  - They are closest to the hyperplane
  - There is a sparse set of such points
  - All other points don't matter
- In soft SVM
  - All missclassified points are SV too!
  - All points with $\xi_i > 0$ or just at the cusp
- Values of ξ_i
  - Correct side of margin = 0
  - On the margin = 0
  - Inside the margin > 0
  - On the boundary = 1
  - Misclassified > 1

Comparison of loss functions



Source: PRML book by Bishop

Training data sparsity in regression using SVMs



$$C\sum_{n=1}^{N} E_\epsilon(y(\mathbf{x}_n) - t_n) + \frac{1}{2}\|\mathbf{w}\|^2$$

- We change the error function from:

$$\frac{1}{2}\sum_{n=1}^{N}\{y_n - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

- To a ε-insensitive (fault-tolerant) error function

where,

$$E_\epsilon(y(\mathbf{x}) - t) = \begin{cases} 0, & \text{if } |y(\mathbf{x}) - t| < \epsilon; \\ |y(\mathbf{x}) - t| - \epsilon, & \text{otherwise} \end{cases}$$

Source: PRML book by Bishop

Now we define slack variables

$$t_n \leqslant y(\mathbf{x}_n) + \epsilon + \xi_n$$
$$t_n \geqslant y(\mathbf{x}_n) - \epsilon - \widehat{\xi}_n.$$

$$\xi_n \geqslant 0 \text{ and } \widehat{\xi}_n \geqslant 0$$

- Cost function can be written as:

$$C \sum_{n=1}^{N} (\xi_n + \widehat{\xi}_n) + \frac{1}{2} \|\mathbf{w}\|^2$$

Source: PRML book by Bishop