TechX 2019:
Advanced Introduction to Deep Learning

Han Zhao
han.zhao@cs.cmu.edu
Machine Learning Department, Carnegie Mellon University

July. 20th, 2019

Advanced Introduction to Deep Learning

The Team:

Instructor: Han Zhao

Academic Lead (AL):

- Xiuyu Li
- Yulong Li
- Xinyu (Norah) Tan
- Yuxuan Sun
- Hang Yuan

Daily Schedule:

Lecture: 9 am - 11:30 am, 1 pm - 2:30 pm

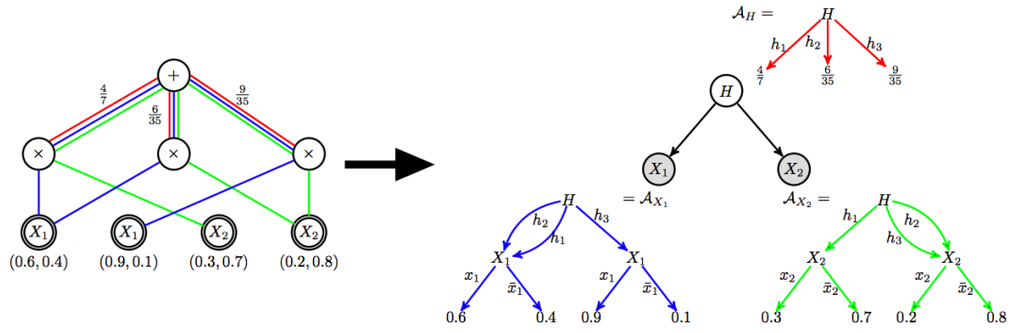
Lab/Office Hour: 3 pm - 6 pm

Instructor: Han Zhao (Han)

- Han Zhao (http://www.cs.cmu.edu/~hzhao1/)
- Ph.D. student @ CMU
- Advisor: Prof. Geoffrey J. Gordon
- Research Interest:

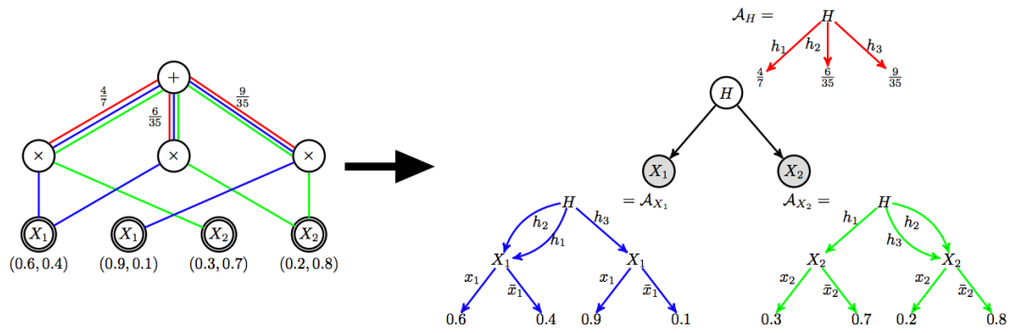
Instructor: Han Zhao (Han)

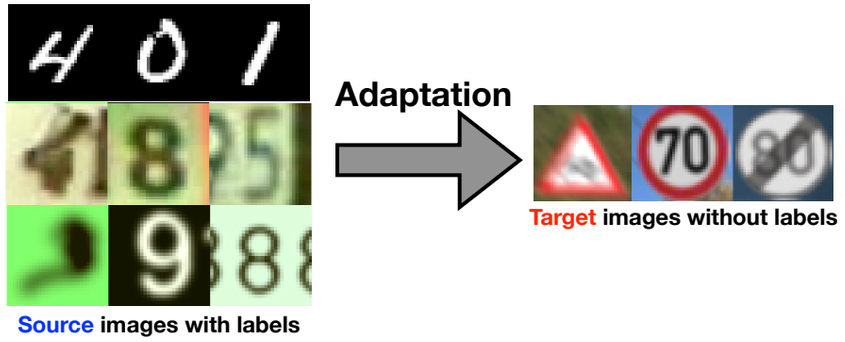
- Han Zhao (http://www.cs.cmu.edu/~hzhao1/)
- Ph.D. student @ CMU
- Advisor: Prof. Geoffrey J. Gordon
- Research Interest:
  - Probabilistic Reasoning



Instructor: Han Zhao (Han)

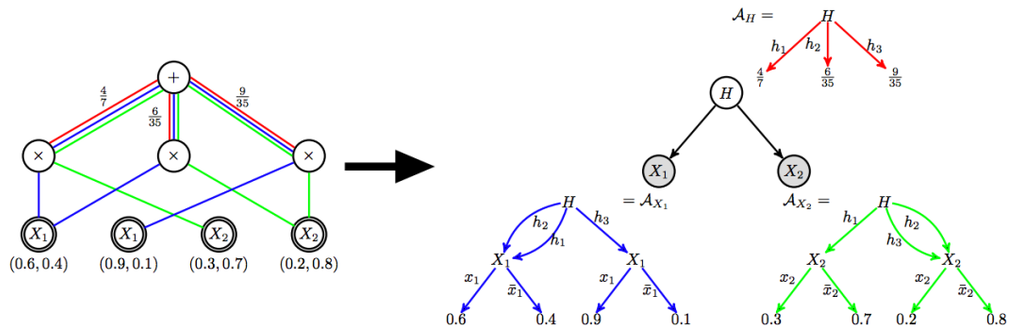
- Han Zhao (http://www.cs.cmu.edu/~hzhao1/)
- Ph.D. student @ CMU
- Advisor: Prof. Geoffrey J. Gordon
- Research Interest:
  - Probabilistic Reasoning
  - Adversarial Machine Learning

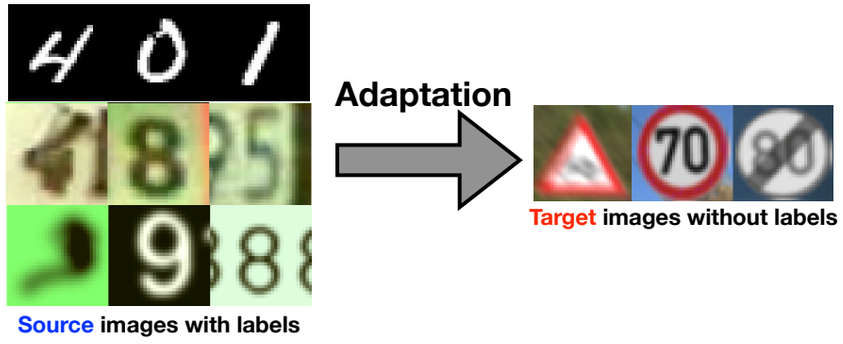




Instructor: Han Zhao (Han)

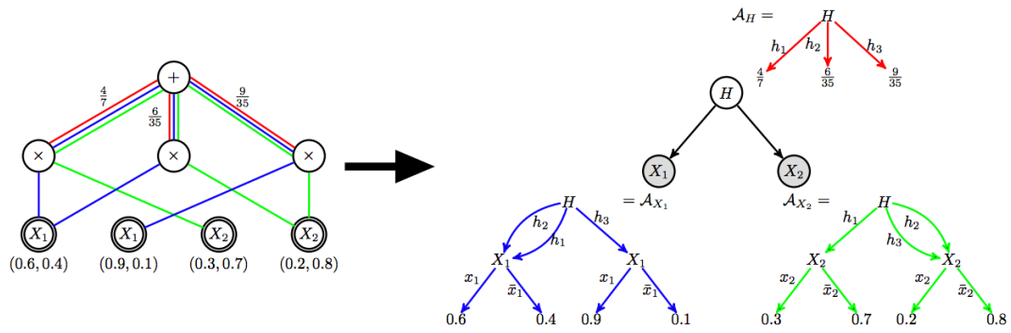
- Han Zhao (http://www.cs.cmu.edu/~hzhao1/)
- Ph.D. student @ CMU
- Advisor: Prof. Geoffrey J. Gordon
- Research Interest:
  - Probabilistic Reasoning
  - Adversarial Machine Learning
  - Computational Social Choice

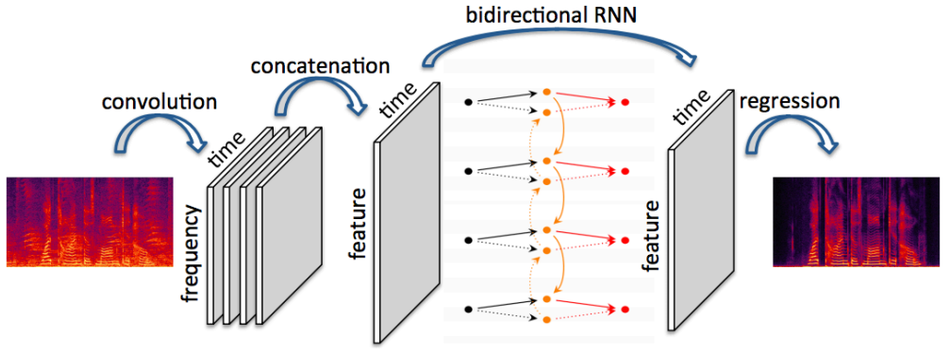


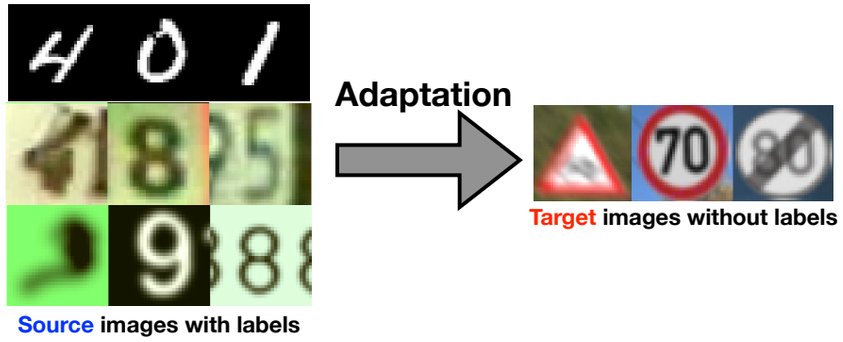


Instructor: Han Zhao (Han)

- Han Zhao (http://www.cs.cmu.edu/~hzhao1/)
- Ph.D. student @ CMU
- Advisor: Prof. Geoffrey J. Gordon
- Research Interest:
  - Probabilistic Reasoning
  - Adversarial Machine Learning
  - Computational Social Choice
  - Deep Learning and its Applications







Academic Lead: Xiuyu Li

Academic Lead: Yulong Li

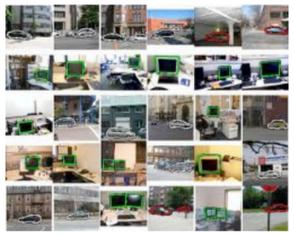
Academic Lead: Xinyu Tan (Norah)

Academic Lead: Yuxuan Sun

- Yuxuan Sun
- Undergrad student @ Haverford College
- Major: Computer Science
- Research Interests:
  - Web Development
  - Data Structure
  - Deep Learning

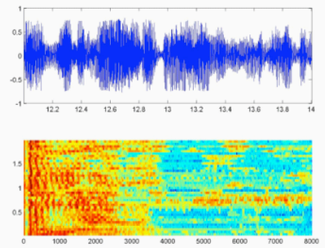
Applications of Machine Learning

Images & Video

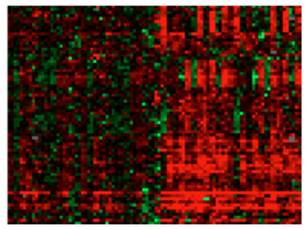


Text & Language

Speech & Audio

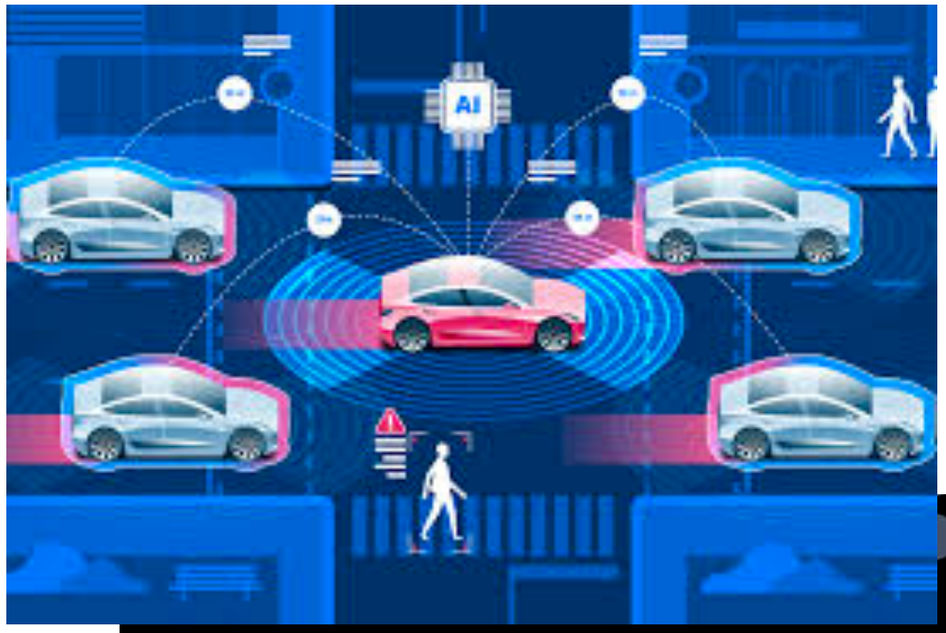


Gene Expression



Product Recommendation

Relational Data/ Social Network



Advanced Introduction to Deep Learning

Wait…But what is Machine Learning ?

Advanced Introduction to Deep Learning

Wait…But what is Machine Learning ?

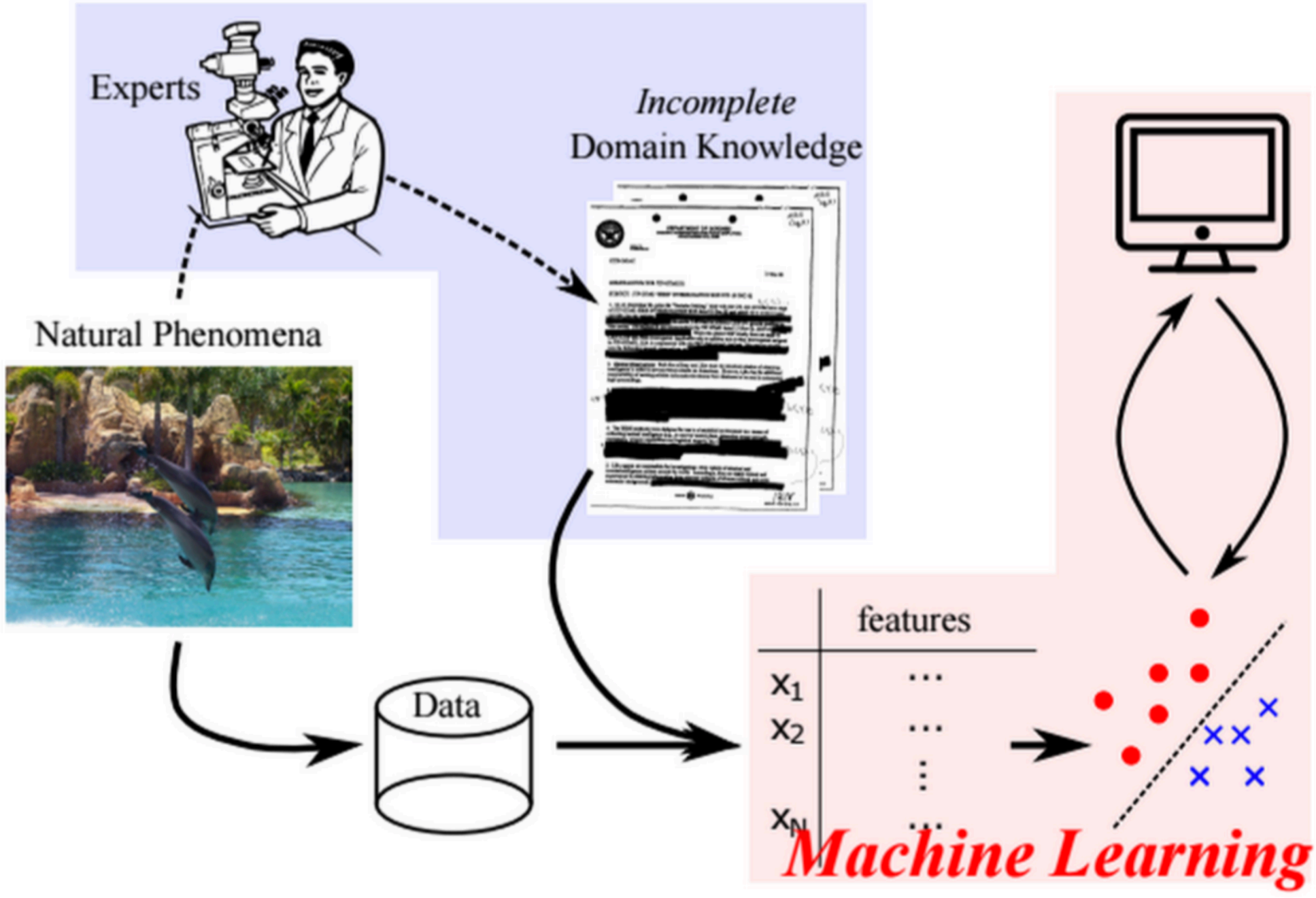
"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

— Tom M. Mitchell

- E: data
- T: task of interest
- P: objective function

Advanced Introduction to Deep Learning

Typical pipeline of Machine Learning:



Advanced Introduction to Deep Learning

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

— Tom M. Mitchell

- E: data
- T: task of interest
- P: objective function

Machine Learning $\approx$ Representation + Objective + Optimization

Advanced Introduction to Deep Learning

OK, but what is Deep Learning?

Advanced Introduction to Deep Learning

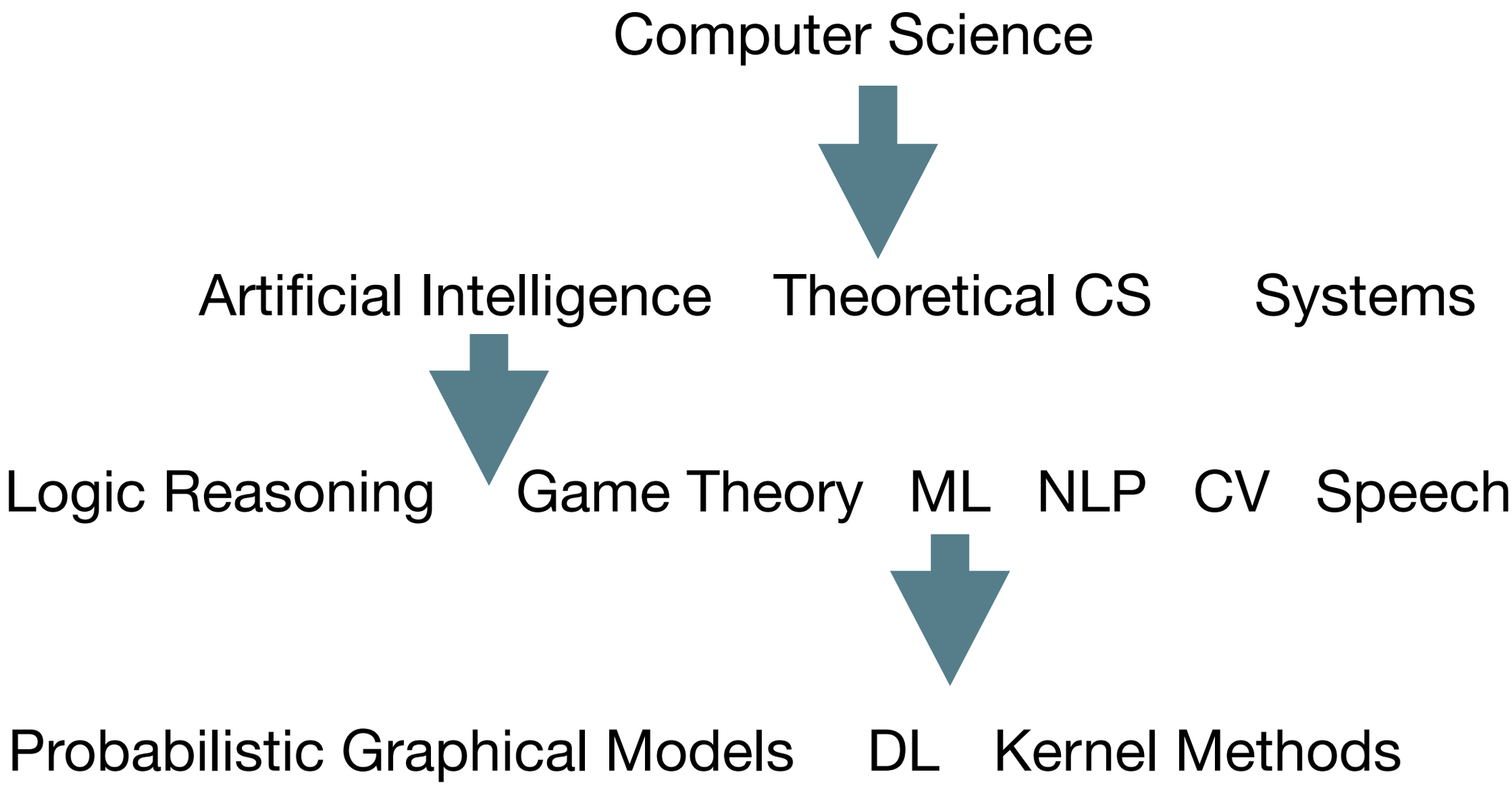
OK, but what is Deep Learning?

A subfield of Machine Learning, and also a subfield of AI

Advanced Introduction to Deep Learning

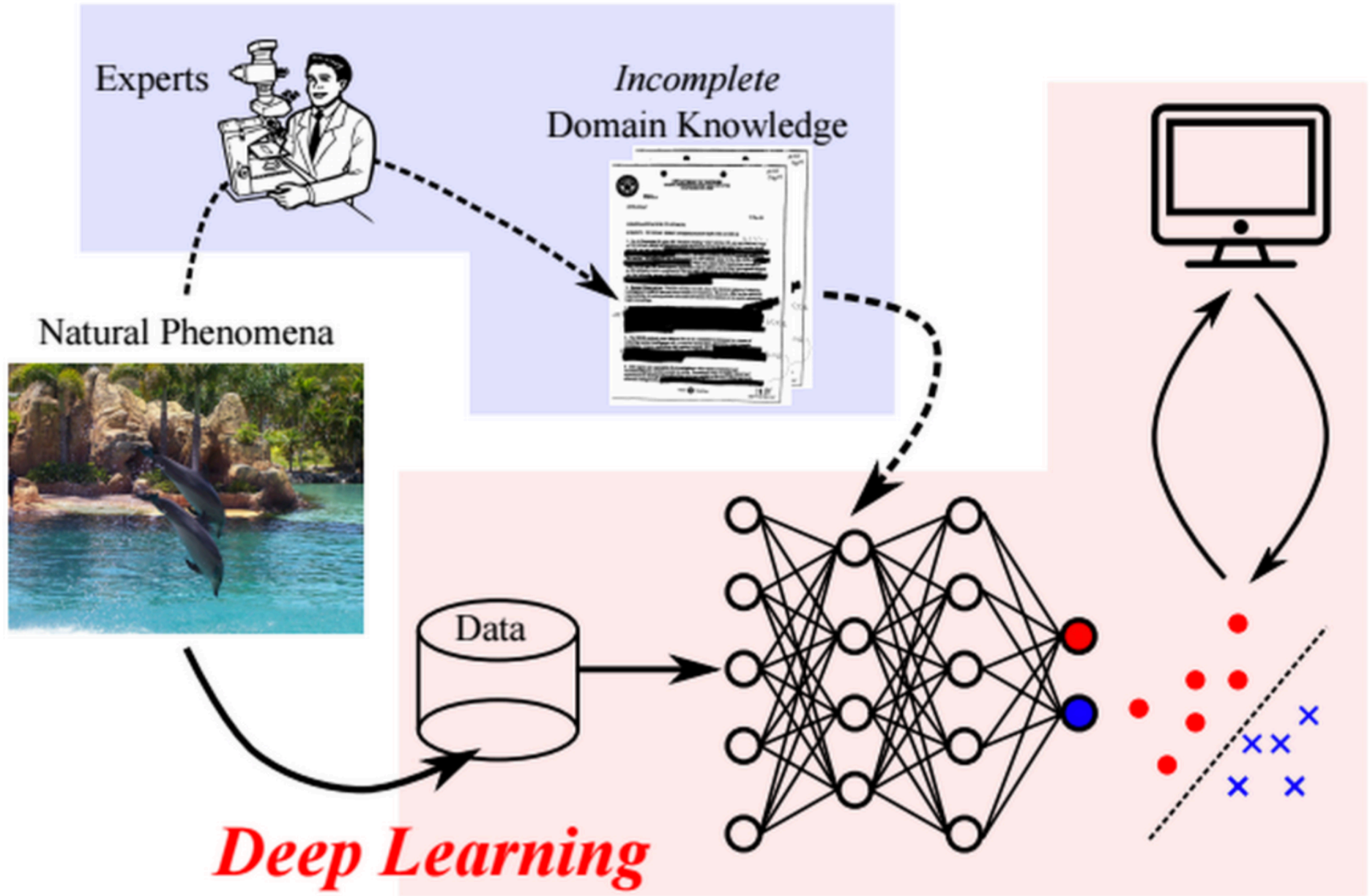
OK, but what is Deep Learning?

A subfield of Machine Learning, and also a subfield of AI



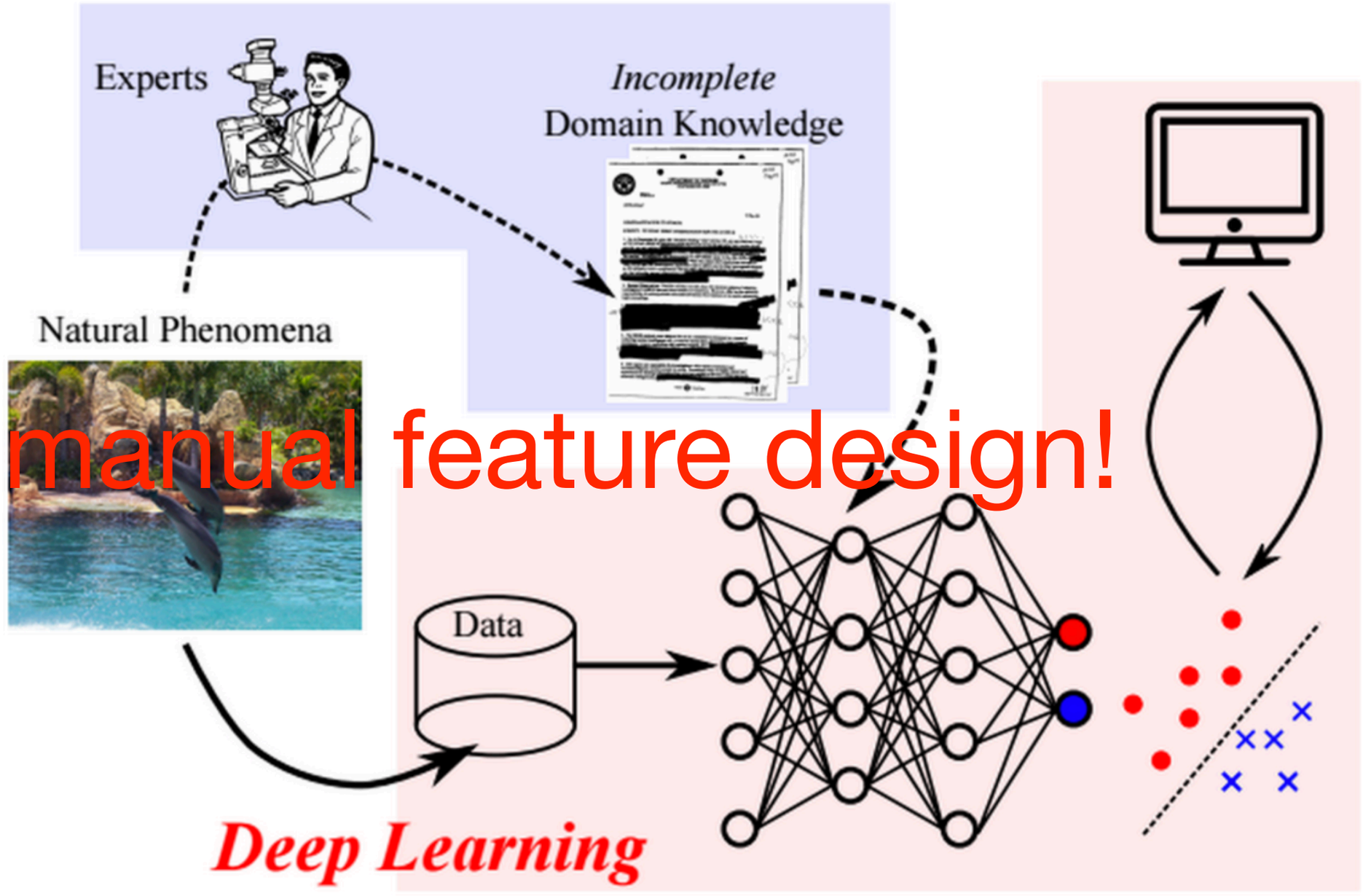
Advanced Introduction to Deep Learning

Deep Learning ≈ Representation + Objective + Optimization



Advanced Introduction to Deep Learning

Deep Learning ≈ Representation + Objective + Optimization

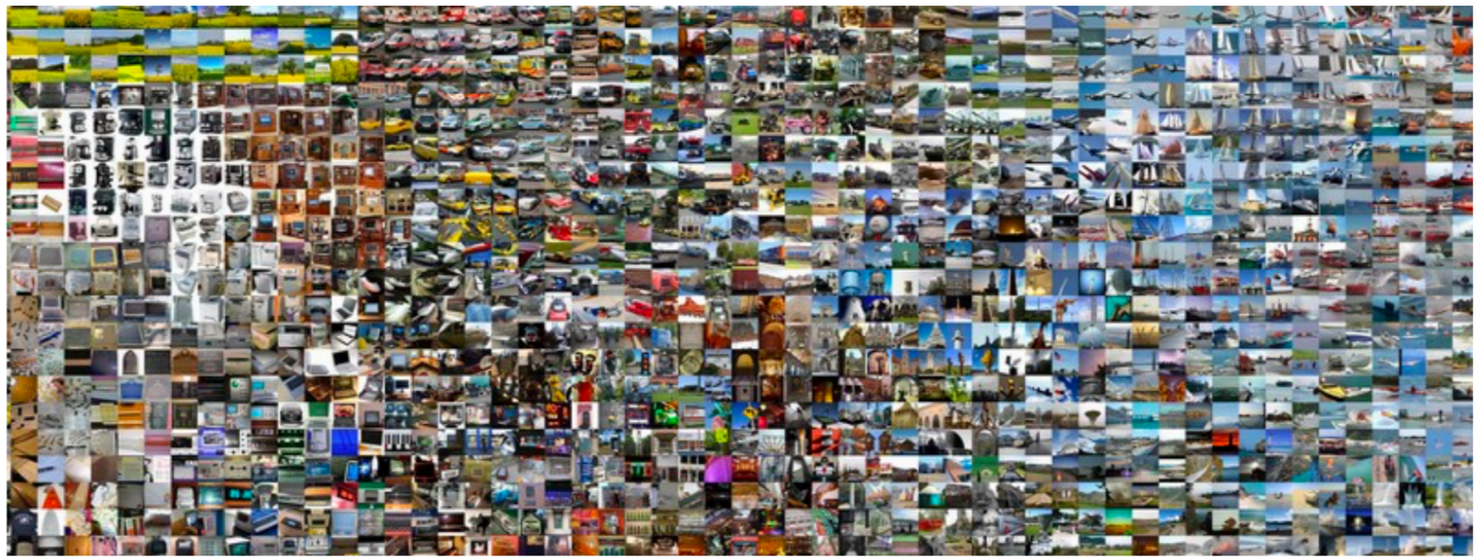


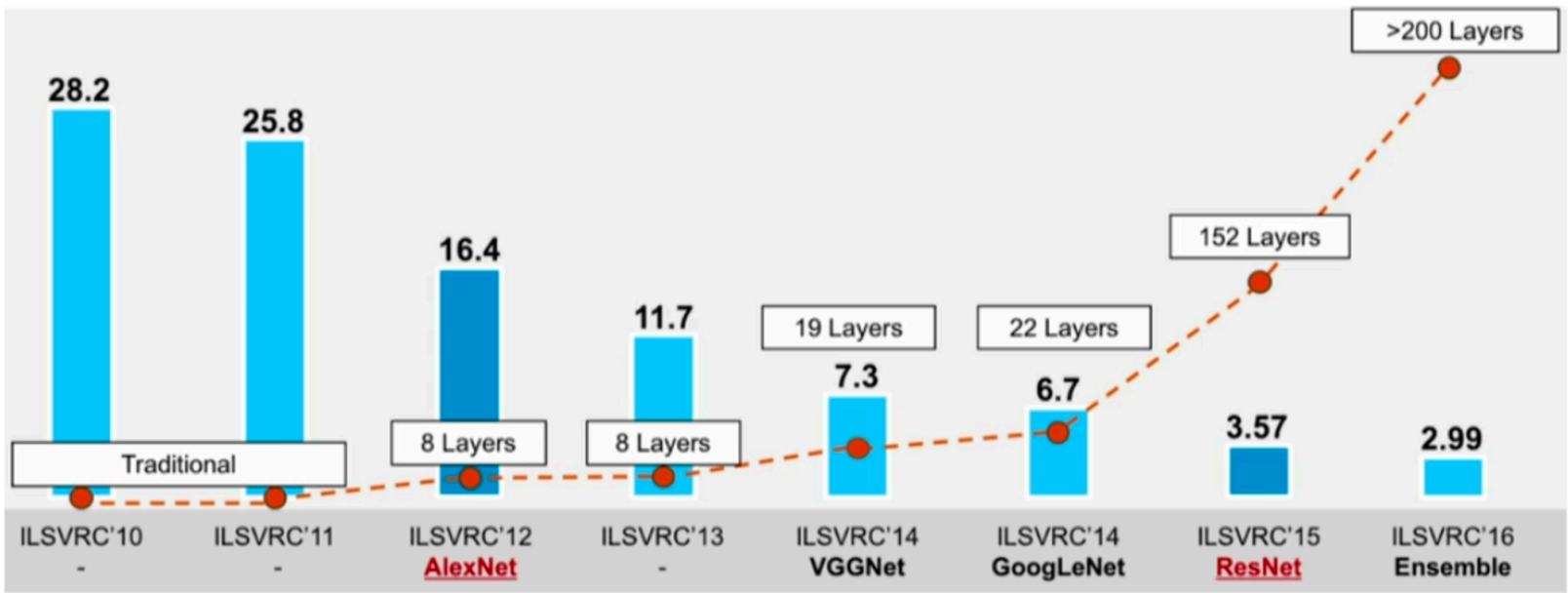
No manual feature design!

Advanced Introduction to Deep Learning

Success of Deep Learning:

- ImageNet: ~1M images, ~1K classes classification





Advanced Introduction to Deep Learning

Success of Deep Learning:

- Go: AlphaGo vs Seudo Lee and Jie Ke

Advanced Introduction to Deep Learning

Success of Deep Learning:

- Dota 2: OpenAI Five

Advanced Introduction to Deep Learning

Feel excited? Now let's get back to our course…

After taking the course, we hope you can:

- Have a high-level view of the machine learning/deep learning field
- Understand the general framework of learning algorithms
- Compare and contrast different paradigms of learning
- Employ probability, statistics, calculus, linear algebra, optimization and information theory to understand learning algorithms
- Use common computational framework to implement basic learning algorithms

Advanced Introduction to Deep Learning

Outline:

- Lecture 1: Maths Preliminary and Introduction
- Lecture 2: Python Toolkit: Numpy, Scipy, Scikit-learn and PyTorch
- Lecture 3: Supervised Learning: Linear Classification and Regression
- Lecture 4: Kernel Methods and Probabilistic Graphical Models
- Lecture 5: Deep Learning: Feedforward Neural Networks
- Lecture 6: Deep Learning: Convolutional Neural Networks
- Lecture 7: Deep Learning: Recurrent Neural Networks
- Lecture 8: Deep Learning: Unsupervised Learning with Variational Autoencoders
- Lecture 9: Deep Learning: Unsupervised Learning with Generative Adversarial Networks

Advanced Introduction to Deep Learning

Recommended Textbook for Machine Learning (publicly available):

- Deep Learning: Deep Learning Book, Ian Goodfellow, Yoshua Bengio and Aaron Courville (GBC)
- General Introduction: Machine Learning, Tom Mitchell (TM)
- Bayesian Machine Learning: Pattern Recognition and Machine Learning, Christopher Bishop (CB)
- Bayesian Machine Learning and Information-Theoretic Methods: Information Theory, Inference, and Learning Algorithms, David Mackay (DM)
- Theoretical Machine Learning: Foundations of Machine Learning, Mehryar Mohri, Afshin Rostamizadeh, Ameet Talwalker

Advanced Introduction to Deep Learning
Lecture 1: Maths Preliminary and Introduction

Han Zhao
han.zhao@cs.cmu.edu
Machine Learning Department, Carnegie Mellon University

July. 20th, 2019

Lecture 1: Maths Preliminary and Introduction

Overview:

- Linear Algebra
- Multivariate Calculus
- Probability and Statistics
- Information Theory
- Convex Optimization

Linear Algebra

Basic concepts:

- Scalar: $x \in \mathbb{R}$, a real number (Velocity)
- Vector: $\mathbf{x} \in \mathbb{R}^d$, a sequence of real numbers (Location in 3D space)
- Matrix: $X \in \mathbb{R}^{d_1 \times d_2}$ a sequence of vectors (Black-white images)
- Third-order tensor: $\mathcal{X} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$, a sequence of matrices (RGB images). Tensors could be generalized to even higher orders

Linear Algebra

Vector norm and its geometric interpretation

- Let $\mathbf{x} \in \mathbb{R}^d$ be a d-dimensional vector
- $p \geq 1, L_p$ norm:

$$||\mathbf{x}||_p := \left(\sum_{i=1}^{d} x_i^p \right)^{1/p}$$

- L_∞ norm:

$$||\mathbf{x}||_\infty := \max_i \ |x_i|$$

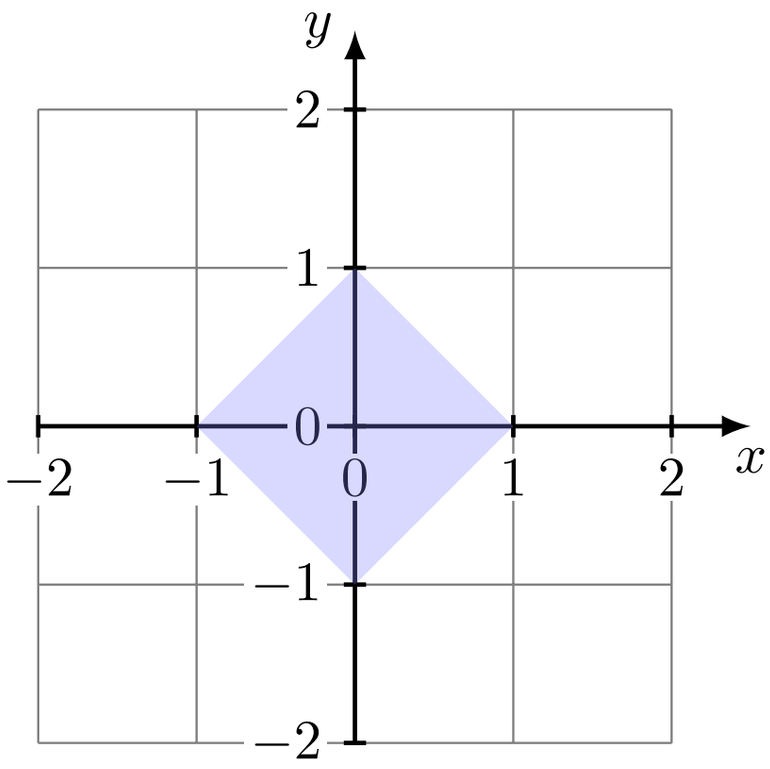
- Define:

$$||\mathbf{x}||_0 := \text{rank}(\mathbf{x})$$

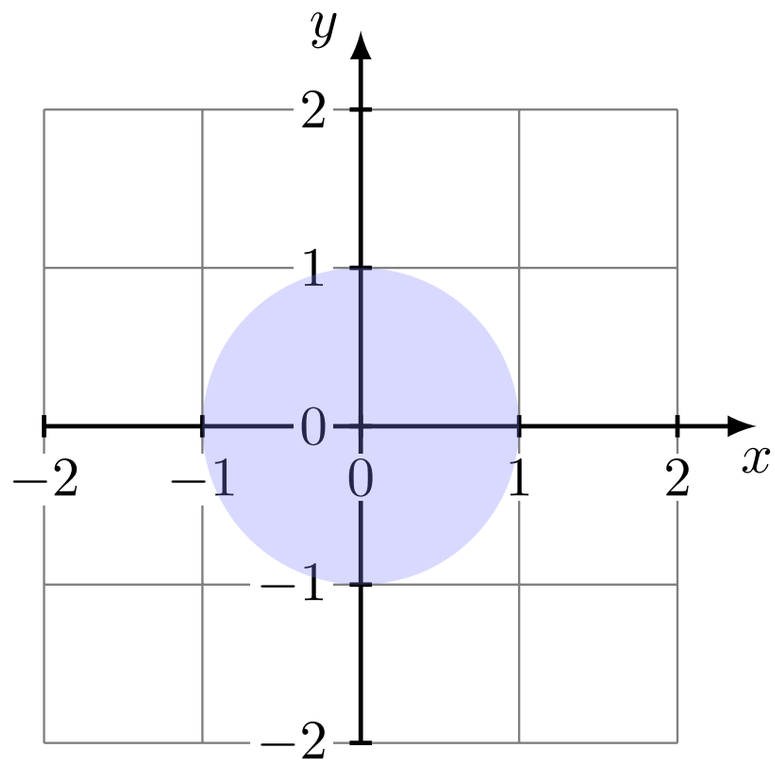
Linear Algebra

Vector norm and its geometric interpretation

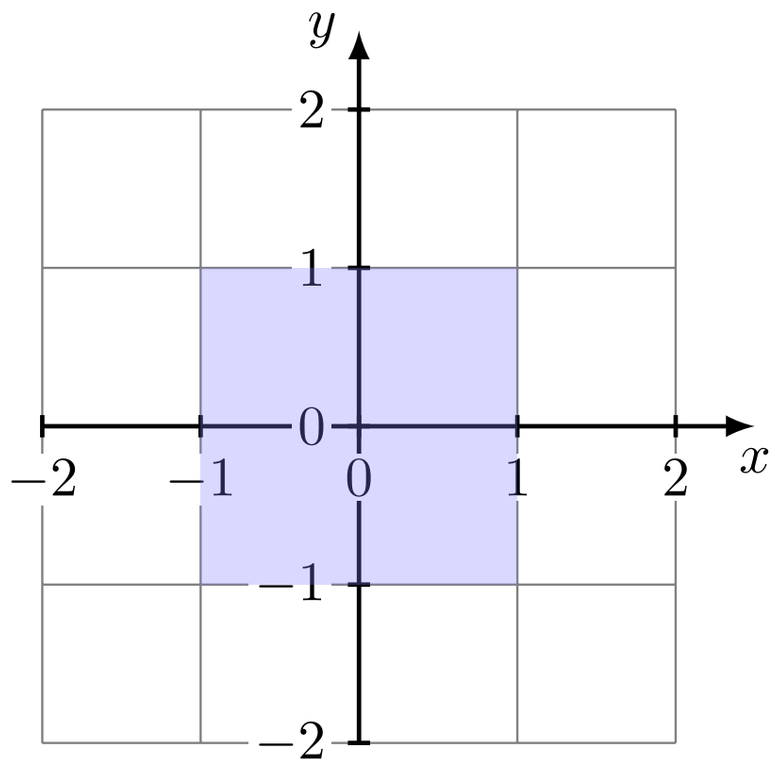
2D visualization of norm unit ball:



$||\mathbf{x}||_1 \leq 1$



$||\mathbf{x}||_2 \leq 1$

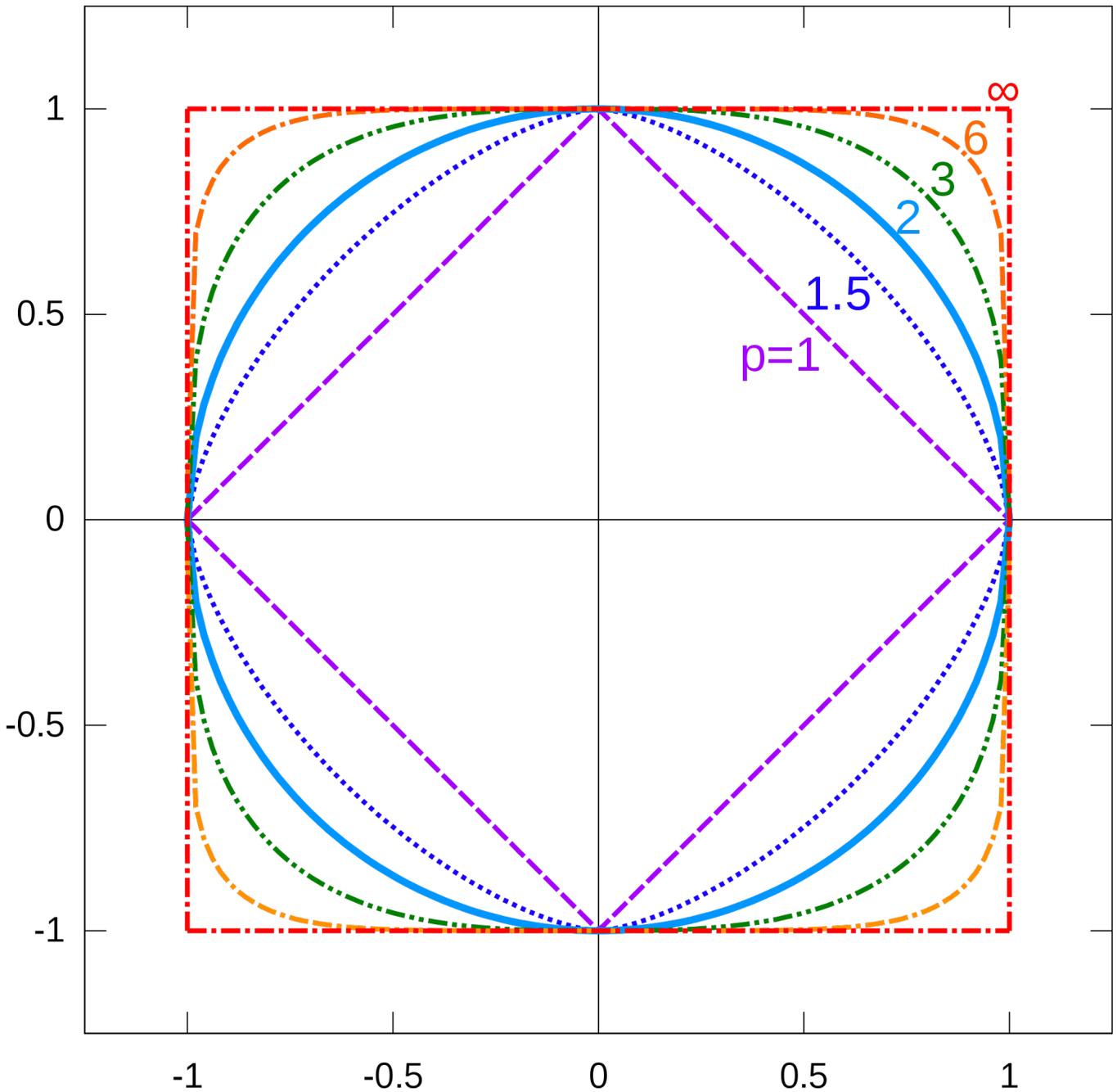


$||\mathbf{x}||_\infty \leq 1$

Linear Algebra

Vector norm and its geometric interpretation

In general:



Linear Algebra

Operations and notations over real matrices: let $A \in \mathbb{R}^{d \times p}$ $B \in \mathbb{R}^{p \times q}$

- Transpose: $A^T : (A^T)_{ij} = A_{ji}$
- Matrix product: $C = AB \in \mathbb{R}^{d \times q}$

Now consider $A \in \mathbb{R}^{d \times d}$ to be a real square matrix

- Trace: $\mathrm{Tr}(A) = \sum_{i=1} A_{ii}$
- Determinant (Leibniz form): $\det(A) = \sum_{\sigma \in S_d} \left(\mathrm{sgn}(\sigma) \prod_{i=1}^{d} A_{i,\sigma_i} \right)$
  - S_d: the set of all permutations over $[d] := \{1, \ldots, d\}$
  - $\mathrm{sgn}(\sigma)$: signature of the permutation, +1 or -1

Linear Algebra

Operations and notations over real matrices: let $A \in \mathbb{R}^{d \times p}$ $B \in \mathbb{R}^{p \times q}$

- Rank, $\mathrm{rank}(A)$: number of linearly independent row/column vectors (Theorem: they are equal)
- $\mathrm{rank}(A) \leq \min\{d, p\}$
- $\mathrm{rank}(AB) \leq \min\{\mathrm{rank}(A), \mathrm{rank}(B)\}$
- $\mathrm{rank}(A + B) \leq \mathrm{rank}(A) + \mathrm{rank}(B)$

Linear Algebra

Now consider $A \in \mathbb{R}^{d \times d}$ to be a real square matrix

- If $A^T = A$, then it is called a symmetric matrix
- If row (column) vectors of A are linearly independent, then it is called invertible, and its inverse matrix is denoted as A^{-1}
- For invertible matrix A, we have $AA^{-1} = A^{-1}A = I_d$
- $I_d \in \mathbb{R}^{d \times d}$ is called the identity matrix:

$$I_d := \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

Linear Algebra

Solving system of linear equations

$$a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1$$

$$a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2$$

$$\vdots$$

$$a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m$$

When the above system has unique/infinite/no solutions?

- If $m < n$, then either infinite solution or no solution. In general it has infinite solutions if $m < n$
- In general if $m = n$, then unique solution exists
- In general if $m > n$, then no solution exists

Linear Algebra

Solving system of linear equations

$$a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1$$

$$a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2$$

$$\vdots$$

$$a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m$$

- Matrix form:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

$$A\mathbf{x} = \mathbf{b}$$

Linear Algebra

Solving system of linear equations

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

We consider the special case $m = n = d$. In this case, if we assume that $\mathrm{rank}(A) = d$, i.e., the coefficient matrix is invertible, then the unique solution has the following form:

$$\mathbf{x} = A^{-1}\mathbf{b}$$

Consider: what if $\mathrm{rank}(A) < d$ in this case?

Linear Algebra

Eigenvalue, eigenvectors and the Spectral Theorem

Let $A \in \mathbb{R}^{d \times d}$ be a real and symmetric matrix, then if $\mathbf{x} \neq 0$ such that $A\mathbf{x} = \lambda\mathbf{x}$, then:

- $\mathbf{x}$ is called an eigenvector of A and λ is called the corresponding eigenvalue
- For real symmetric matrix, we have that $\lambda \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^d$
- Distinct eigenvectors are orthogonal to each other: if $\lambda_1 \neq \lambda_2$ and $A\mathbf{x}_1 = \lambda_1\mathbf{x}_1, A\mathbf{x}_2 = \lambda_2\mathbf{x}_2$, then $\mathbf{x}_1^T\mathbf{x}_2 = 0$
- A admits a spectral decomposition and can be transformed into a diagonal matrix:

$$A = Q\Lambda Q^T, QQ^T = Q^TQ = I_d \qquad \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_d)$$

Linear Algebra

Singular Value Decomposition (SVD)

Spectral decomposition only applies to square matrix, how about rectangular matrix in general?

Let $A \in \mathbb{R}^{d \times p}$ be a real rectangular matrix where $d \neq p$, then A has the following decomposition:

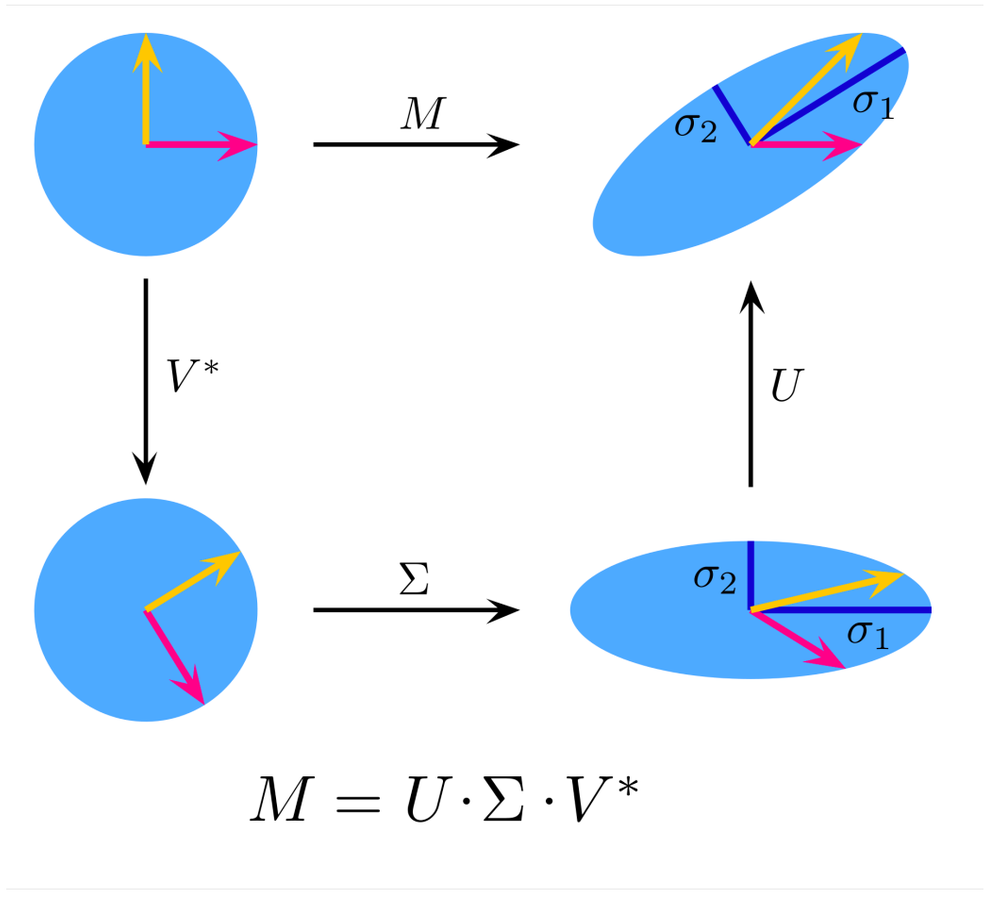
$$A = U\Sigma V^T$$

- $U \in \mathbb{R}^{d \times d}$ is orthonormal, i.e., $U^T U = UU^T = I_d$
- $V \in \mathbb{R}^{p \times p}$ is orthonormal, i.e., $V^T V = VV^T = I_p$
- $\Sigma \in \mathbb{R}^{d \times p}$ is a nonnegative diagonal matrix, and its diagonal elements are called singular values

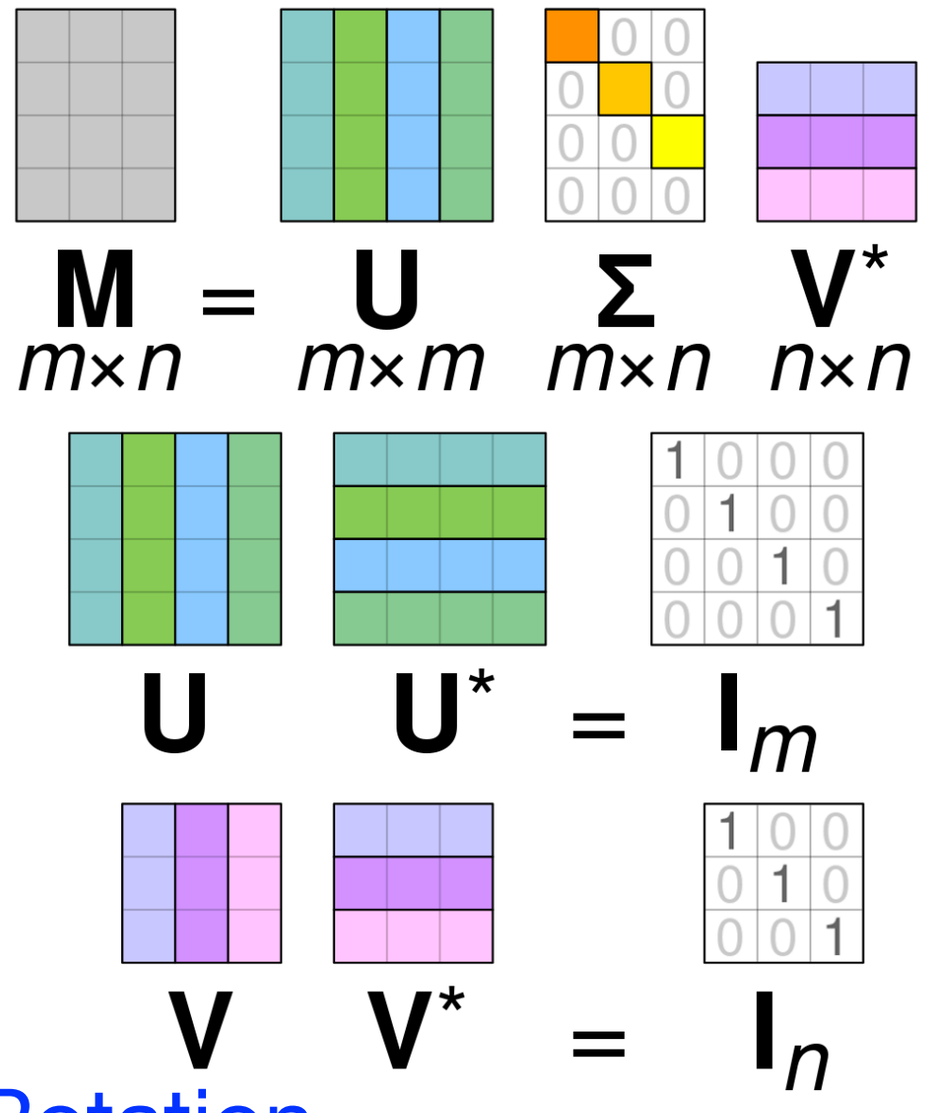
Linear Algebra

Singular Value Decomposition (SVD)

Geometric interpretation of SVD $A = U\Sigma V^T$



$$M = U \cdot \Sigma \cdot V^*$$



Key: Matrix = Rotation + Scaling + Rotation

Lecture 1: Maths Preliminary and Introduction

Overview:

- Linear Algebra
- Multivariate Calculus
- Probability and Statistics
- Information Theory
- Convex Optimization

Multivariate Calculus

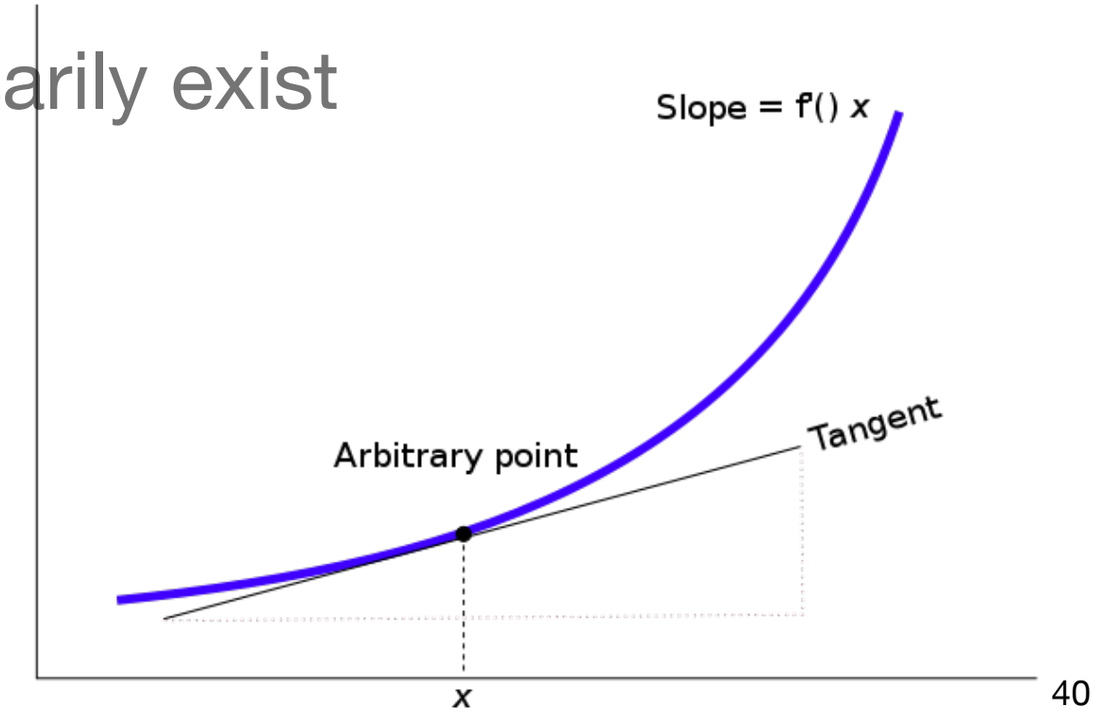
Calculus 101: Differentiation and Gradient

- Geometric definition of derivative: slope, rise-over-run
- Formally, given a "smooth" function $f(x) : \mathbb{R} \to \mathbb{R}$, the derivative is defined as:

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

- Note: derivative does not necessarily exist
- Equivalently, in differential form:

$$dy = f'(x)\ dx$$



Multivariate Calculus

Basic calculus for derivatives:

- Constant derivative:

$$da = 0$$

- Linearity:

$$d(ay + bz) = ady + bdz = (ay'(x) + bz'(x))\,dx$$

- Multiplicative rule:

$$d(yz) = ydz + dyz = (yz'(x) + y'(x)z)dx$$

- Quotient rule:

$$d\left(\frac{y}{z}\right) = \frac{zdy - ydz}{z^2} = \frac{y'(x)z - z'(x)y}{z^2}dx$$

Multivariate Calculus

Basic calculus for derivatives:

- Chain rule: Let $y = g(h(x))$, then $dy = g'(h(x))h'(x)dx$ (under some regularity conditions)

This gives us a set of rules to compute derivative for composite functions!

Multivariate Calculus

Derivatives of some common functions:

- Polynomial function:

$$f(x) = x^k, \quad f'(x) = kx^{k-1}, \ \forall k \in \mathbb{R}$$

- Exponential function:

$$f(x) = \exp(x), \quad f'(x) = \exp(x)$$

- Log function:

$$f(x) = \ln(x), \quad f'(x) = \frac{1}{x}, \forall x > 0$$

Multivariate Calculus

Derivative of multivariate real-valued function:

Assume $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{x} := (x_1, x_2, \ldots, x_d)$. Consider a real-valued function: $f(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$.

- We can extend the definition of differentiation/derivative to vectors, known as the gradient

$$\nabla f(\mathbf{x}) := (\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_d})^T$$

$$\frac{\partial f}{\partial x_i} := \lim_{h \to 0} \frac{f(x_1, \ldots, x_i + h, \ldots, x_d) - f(x_1, \ldots, x_i, \ldots, x_d)}{h}$$

- Note: throughout this course we also refer vectors as column vectors

Multivariate Calculus

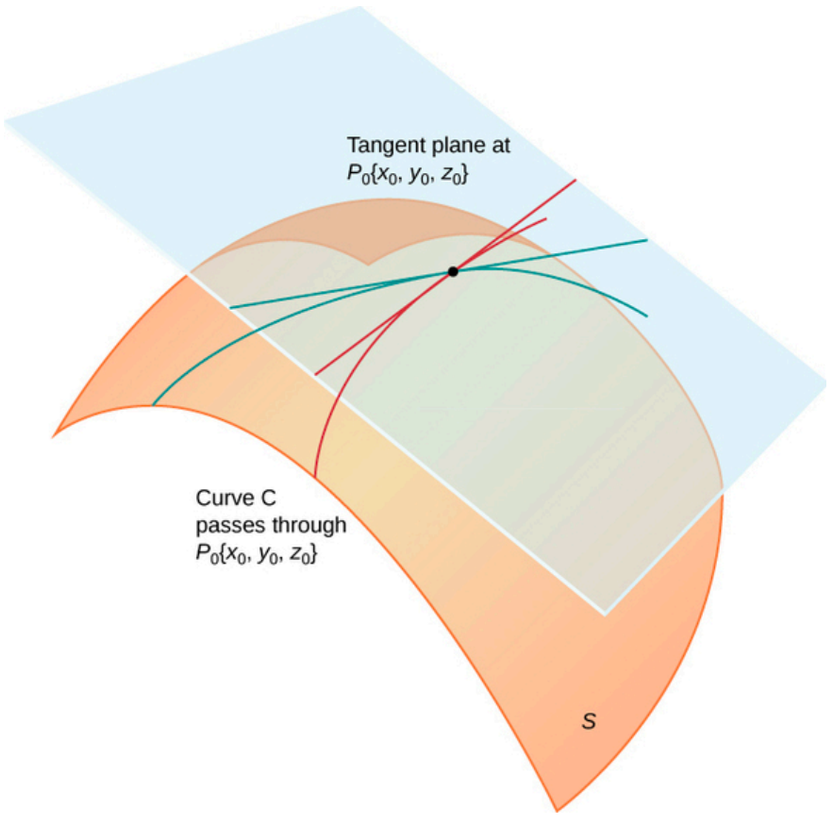
Derivative of multivariate real-valued function:

Assume $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{x} := (x_1, x_2, \ldots, x_d)$. Consider a real-valued function: $f(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$.

Geometric meaning of the gradient $\nabla f(\mathbf{x})$:

- The steepest ascent direction (why?)
- The normal vector of the tangent plane
- Easy computation of directional derivative

$$\nabla_{\mathbf{z}} f(\mathbf{x}) := \lim_{h \to 0} \frac{f(\mathbf{x} + h\mathbf{z}) - f(\mathbf{x})}{h} = \nabla f(\mathbf{x})^T \mathbf{z}$$



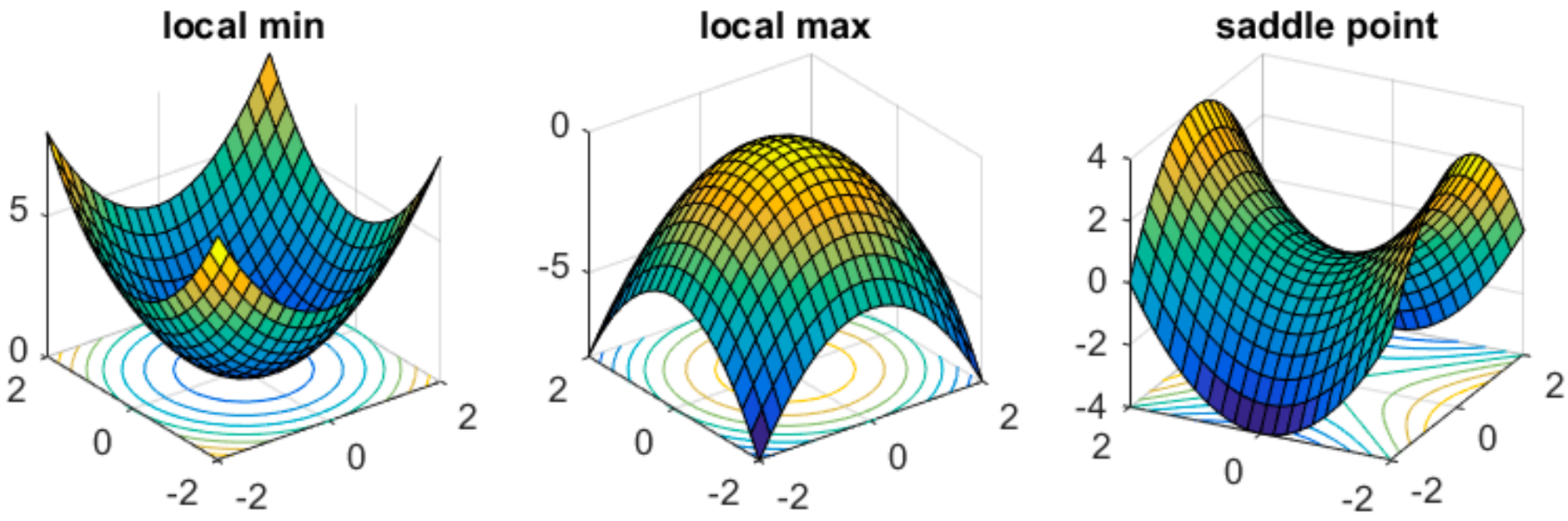
Multivariate Calculus

Derivative of multivariate real-valued function:

Assume $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{x} := (x_1, x_2, \ldots, x_d)$. Consider a real-valued function: $f(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$.

Algebraic meaning of the gradient $\nabla f(\mathbf{x})$:

- If $\nabla f(\mathbf{x}) = 0$, then $\mathbf{x}$ is called a stationary point
- There are three cases:



Multivariate Calculus

Second order derivative: Hessian matrix:

Assume $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{x} := (x_1, x_2, \ldots, x_d)$. Consider a real-valued function: $f(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$.

When $\nabla f(\mathbf{x}) = 0$, first order information is not enough, we need to consider second order information at $\mathbf{x}$

Multivariate Calculus

Second order derivative: Hessian matrix:

Assume $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{x} := (x_1, x_2, \ldots, x_d)$. Consider a real-valued function: $f(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$.

Define the Hessian matrix at $\mathbf{x}$ as follows:

$$\nabla^2 f(\mathbf{x}) := \mathbf{H}(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_d^2} \end{bmatrix} \in \mathbb{R}^{d \times d}$$

- Hessian matrix characterizes the curvature of function $f(\cdot)$
- It also provides us sufficient conditions for local optimality (later sections)

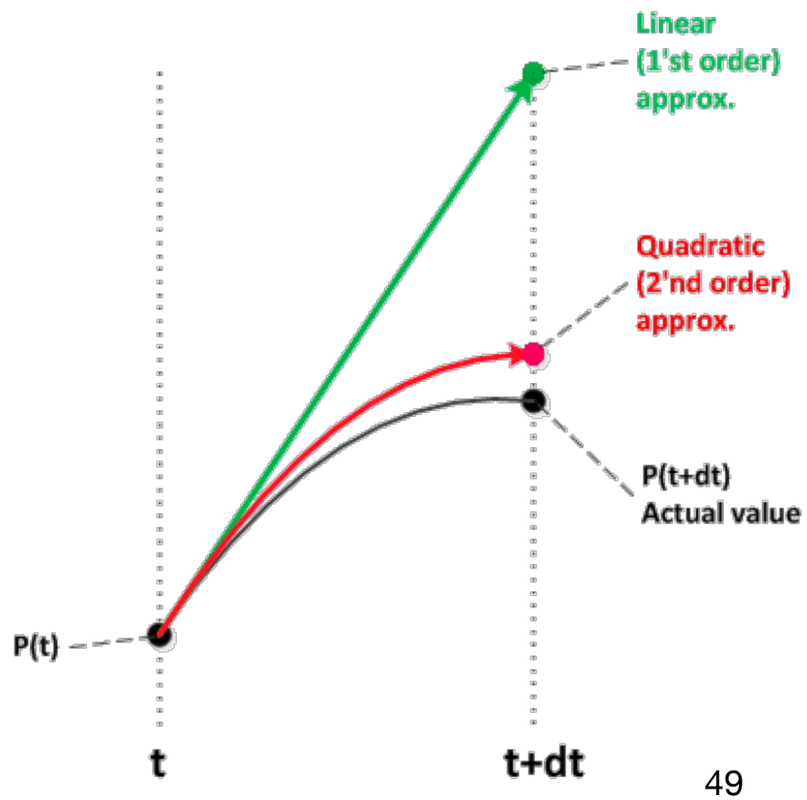
Multivariate Calculus

Taylor expansion: the best polynomial approximation of $f(\cdot)$

For our purpose, we only consider second order Taylor expansion using quadratic polynomial:

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^T (\mathbf{x} - \mathbf{x}_0) + \frac{1}{2} (\mathbf{x} - \mathbf{x}_0)^T \mathbf{H}(\mathbf{x}_0) (\mathbf{x} - \mathbf{x}_0) + O\left(\|\mathbf{x} - \mathbf{x}_0\|_2^2\right)$$

- Frequently used in the design of second-order optimization algorithms
- Computation complexity $O(d^2)$



Multivariate Calculus

Derivative of multivariate vector-valued function:

Assume $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{x} := (x_1, x_2, \ldots, x_d)$. Consider a vector-valued function: $\mathbf{f}(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}^p$, $\mathbf{f}(\mathbf{x}) := (f_1(\mathbf{x}), \ldots, f_p(\mathbf{x}))^T$

- In this case the first order derivative is called the Jacobian:

$$\mathbf{J} = \begin{bmatrix} \frac{\partial \mathbf{f}}{\partial x_1} & \cdots & \frac{\partial \mathbf{f}}{\partial x_d} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_p}{\partial x_1} & \cdots & \frac{\partial f_p}{\partial x_d} \end{bmatrix}$$

- Allowing us to have a compact form of differentiation:

$$d\mathbf{f}(\mathbf{x}) = \mathbf{J}(\mathbf{x}) d\mathbf{x}$$

Multivariate Calculus

Some common matrix-differential rules:

- If $\mathbf{y} = A\mathbf{x}, \mathbf{x} \in \mathbb{R}^d, \mathbf{y} \in \mathbb{R}^p$, then

$$d\mathbf{y} = Ad\mathbf{x}$$

- If $y = \frac{1}{2}\mathbf{x}^T A\mathbf{x}, \mathbf{x} \in \mathbb{R}^d, y \in \mathbb{R}$, then

$$dy/d\mathbf{x} = \frac{A + A^T}{2}\mathbf{x}$$

- If $Y = A^{-1}(x)$, then $dY = -A^{-1}(dA)A^{-1}$
- If $y = \log\det(A)$, where A is real symmetric and positive definite, then $dy = \mathrm{Tr}(X^{-1}dX)$

Lecture 1: Maths Preliminary and Introduction

Overview:

- Linear Algebra
- Multivariate Calculus
- Probability and Statistics
- Information Theory
- Convex Optimization

Probability and Statistics

"We must become more comfortable with probability and uncertainty".

—Nate Silver

Probability is the language of modern sciences:

- Quantum Physics
- Quantum Computation
- Randomized Algorithms
- Machine Learning/Statistics/Information Theory
- Engineering: Robotics/Signal Processing/

Probability and Statistics

Basic concepts in Probability/Statistics

- Random variable X: a quantity that can take different numeric values
- We focus on two kinds of random variables:
  - Discrete random variable: X can only take finitely many/or countably infinitely many values, e.g., flip a coin H/T, or $\mathbb{N}^*$
  - Continuous random variable: X can take uncountably infinitely many values, e.g., location of a dart, or $\mathbb{R}$
  - For both discrete and continuous variables, we use probability distribution $\Pr(X \leq x)$ to characterize our uncertainty about the underlying events

Probability and Statistics

Examples

- Discrete random variables:
  - If we randomly throw a fair coin, then we have the following distribution to characterize our uncertainty (Bernouli distribution): $\Pr(X = H) = \Pr(X = T) = 0.5$
  - Similarly, if we repeatedly and independently throw the same fair coin for n times, then we get the Binomial distribution:

$$\Pr(X = k \text{ times } H) = \frac{\binom{10}{k}}{2^{10}}$$

  - We can also use Poisson distribution to characterize how many emails we receive each day. In this case $X \in \mathbb{N}^*$:

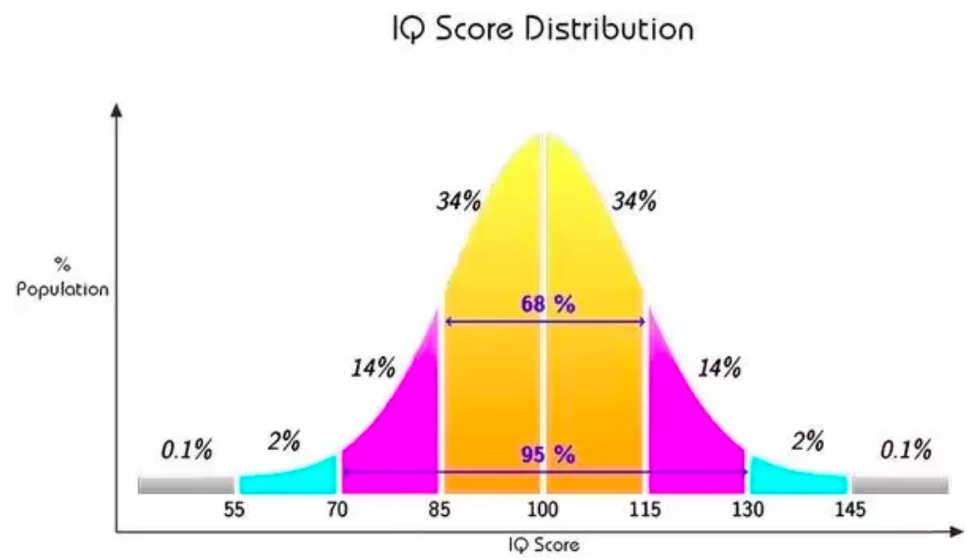
$$\Pr(X = k) = \exp(-\lambda)\frac{\lambda^k}{k!}$$

Probability and Statistics

Examples

- Continuous random variables:
  - The distribution of IQ obeys a normal distribution:

$$p(x) = \mathcal{N}(100, \sigma^2)$$



  - What's the lifespan of a bulb? Roughly, it follows from an Exponential distribution: $p(x) = \lambda \exp(-\lambda x) \cdot \mathbb{I}(x \geq 0)$
    - Fun fact about exponential distribution: it is the waiting time between two arrivals of events in the Poisson process (Why?)

Probability and Statistics

Probability Axioms:

- Formally, we have three axioms to define probability, followed by Kolmogorov's approach:
  - Nonnegativity: For any event E, $\Pr(E) \geq 0$
  - Unit measure: $\Pr(\Omega) = 1$, where Ω is the sample space
  - Countably additivity: $\Pr\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \Pr(E_i)$, where all the events are mutually exclusive

- Consider the example of tossing a fair coin for two times.
  - What are the sample spaces?
  - What are the probabilities of each event?

Probability and Statistics

Conditional probability:

- In many cases we would like to express the belief like: given A happens, what's the probability that B happens?
  - Wow, we can see that Xiuyu is taking an umbrella, what's the probability that it's now raining outside?
  - Given that Libratus has defeated the top Poker players, what's the probability that I can win over it?

Probability and Statistics

Conditional probability:

- Given two random variables X, Y over the same probability space, we use $\Pr(Y \mid X = x)$ to denote the probability distribution of Y given the event $X = x$ and $\Pr(X = x) > 0$

$$\Pr(Y = y \mid X = x) = \frac{\Pr(Y = y, X = x)}{\Pr(X = x)}$$

- Note that the above conditional probability also extends to density functions:

$$p(y \mid x) = \frac{p(x, y)}{p(x)}$$

- Equivalently, we have:

$$\Pr(X = x, Y = y) = \Pr(X = x) \Pr(Y = y \mid X = x)$$

$$p(x, y) = p(x) \cdot p(y \mid x)$$

Probability and Statistics

The Law of Total Probability:

- Let $\{B_n\}_{n=1}^{\infty}$ be a partition of the sample space, then for any event A, the following identity holds:

$$\Pr(A) = \sum_{n=1}^{\infty} \Pr(A \cap B_n) = \sum_{n=1}^{\infty} \Pr(A \mid B_n) \Pr(B_n)$$

- Consider tossing two fair coins, let the result of the first one be A and the result for the second one be B, then:
  - What's the probability that the first coin is H?

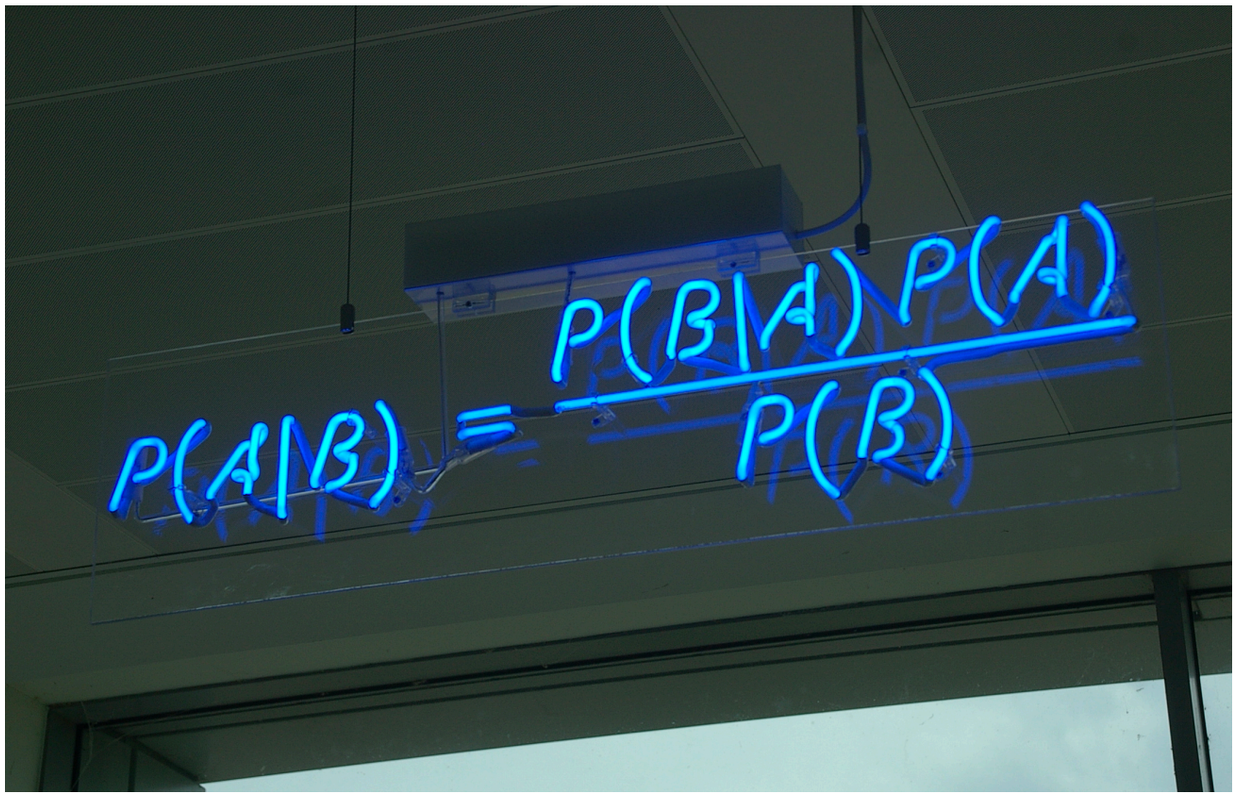
Probability and Statistics

Bayes Theorem:

- Discrete case: $\Pr(A \mid B) = \dfrac{\Pr(A)\Pr(B \mid A)}{\Pr(B)}$
- Continuous case:

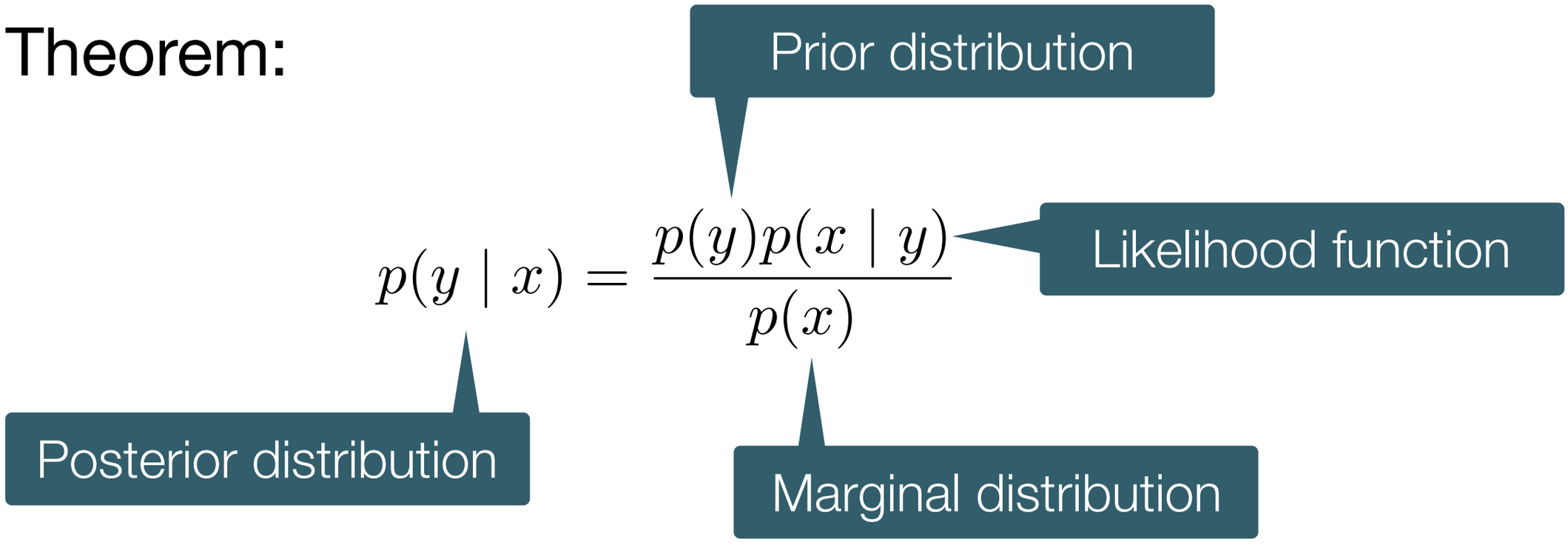
$$p(y \mid x) = \frac{p(y)p(x \mid y)}{p(x)}$$

Very Important!



Probability and Statistics

Bayes Theorem:



- An update formula of “our belief”
- The fundamental of Bayesian Machine Learning
- The underlying principle of Bayesian methods
- Why is it so important?

Probability and Statistics

Conditional Independence:

- Given an event E, A, B are conditionally independent of E iff

$$\forall a, b, \quad \Pr(A = a, B = b \mid E) = \Pr(A = a \mid E) \cdot \Pr(B = b \mid E)$$

- More generally, A, B are conditionally independent given C iff

$$\forall a, b, c, \quad \Pr(A = a, B = b \mid C = c) = \Pr(A = a \mid C = c) \cdot \Pr(B = b \mid C = c)$$

- As a special case, if $E = \Omega$, then we recover the marginally independent (or just independent) between A, B:

$$\forall a, b, \quad \Pr(A = a, B = b) = \Pr(A = a) \cdot \Pr(B = b)$$

Probability and Statistics

Some simple facts about independence:

- If A, B are independent, then:
  - $Var(A + B) = Var(A) + Var(B)$
  - $\mathbb{E}[AB] = \mathbb{E}[A]\mathbb{E}[B]$

- Expectation: $\mathbb{E}[X] = \int xp(x)\ dx$
- Variance: $Var(X) = \mathbb{E}[X^2] - \mathbb{E}^2[X]$
- Covariance: $Cov(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$

Question: does 0 covariance equivalent to independence?

Probability and Statistics

The Law of Large Numbers & Concentration Inequality:

- The Strong Law of Large Numbers (SLLN): Suppose $X_1, \ldots, X_n$ are independent and $\mathbb{E}[X_i] = 0$, $\sum_i \mathrm{Var}(X_i)/n^2 < \infty$, then $\sum_i X_i/n \to 0, a.s.$

$$\Pr(\lim_{n\to\infty} \sum_i X_i/n = 0) = 1$$

- OK, SLLN let us know the convergence result, how about the convergence speed?
- Hoeffding's inequality: If $X_1, \ldots, X_n$ are independent RVs bounded by $[a_i, b_i]$, then

$$\Pr(|\overline{X} - \mathrm{E}[\overline{X}]| \geq t) \leq 2\exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

Lecture 1: Maths Preliminary and Introduction

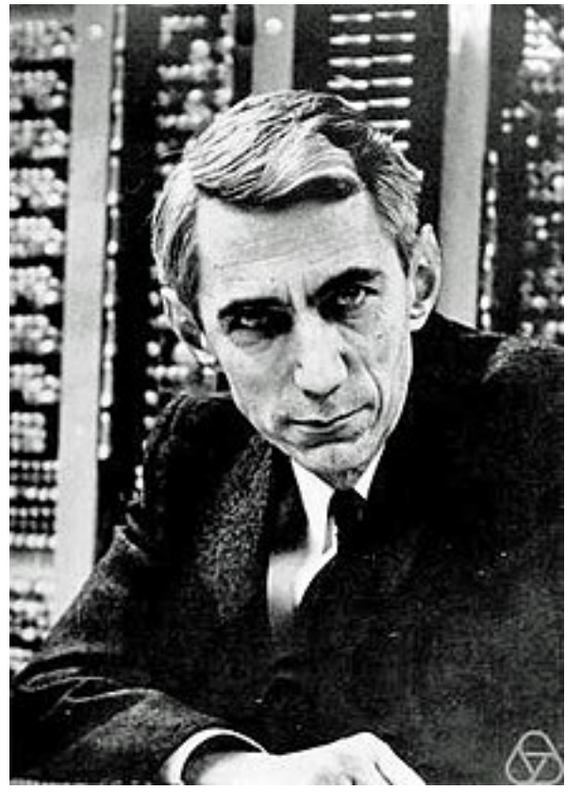
Overview:

- Linear Algebra
- Multivariate Calculus
- Probability and Statistics
- Information Theory
- Convex Optimization

Information Theory

"the transmission, processing, extraction, and utilization of information"

Let us only focus on discrete RVs.



Claude E. Shannon
(1916 ~ 2001)

Information Theory

Shannon Entropy:

- Let $p_1, \ldots, p_k \geq 0, \sum_{i=1}^{k} p_i = 1$ be the probability distribution of a discrete random variable X
- Shannon Entropy is a quantity used to measure the "uncertainty" of this distribution: if X is close to uniform, then we are very uncertain about its value; if X almost certain takes a fixed value, then we are very certain about its value

$$H(X) := -\sum_{i=1}^{k} p_i \log p_i = -\mathbb{E}_X[\log \Pr(X)]$$

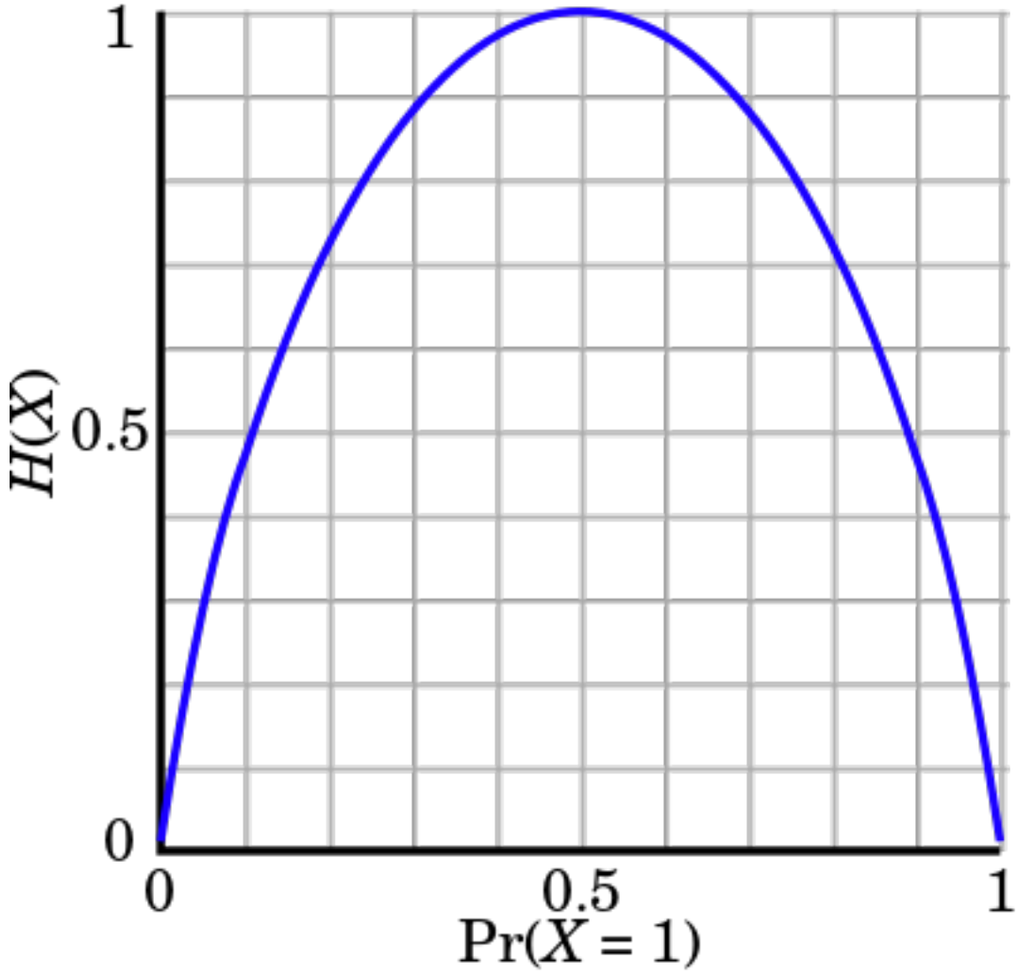
Information Theory

Shannon Entropy:

- Consider the special case for Bernoulli distribution: $k = 2$

$$H(p) = -p \log p - (1-p) \log(1-p)$$

- $p := \Pr(X = 1)$
- Concave function (later)
- Maximum when $p = 0.5$
- Minimum when $p = 0, 1$
- WLOG, we assume 0 log 0 = 0



Information Theory

Shannon Entropy: $H(X) := -\sum_{i=1}^{k} p_i \log p_i = -\mathbb{E}_X[\log \Pr(X)]$

- (Why?) $H(X) \geq 0$
- $H(X) \leq \log k$

For independent RVs, the joint entropy = the sum of marginal entropy:

$$X \perp Y \implies H(X,Y) = H(X) + H(Y)$$

- If two RVs are independent, then the total uncertainty is the sum of individual uncertainties

Information Theory

Conditional entropy: $H(X \mid Y)$

Measure the remaining uncertainty of X given information about Y:

$$H(X \mid Y) = -\sum_{ij} \Pr(X = i, Y = j) \log \Pr(X = i \mid Y = j)$$

Alternatively, conditional entropy has the following form:

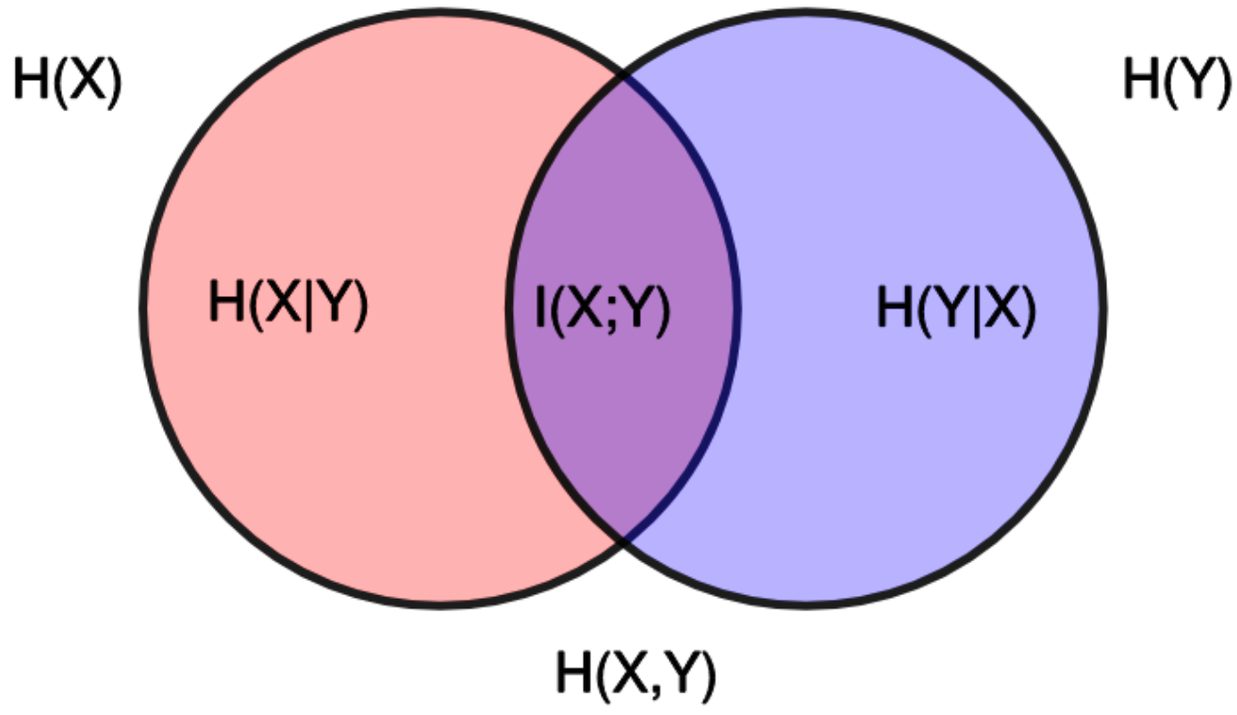
$$\begin{aligned} H(X \mid Y) &= \mathbb{E}_Y[H(X \mid Y = y)] \\ &= \sum_y \Pr(Y = y) \cdot H(X \mid Y = y) \end{aligned}$$

Verify: the chain rule holds for joint entropy:

$$H(X, Y) = H(Y) + H(X \mid Y) = H(X) + H(Y \mid X)$$

Information Theory

Venn digram of entropy:

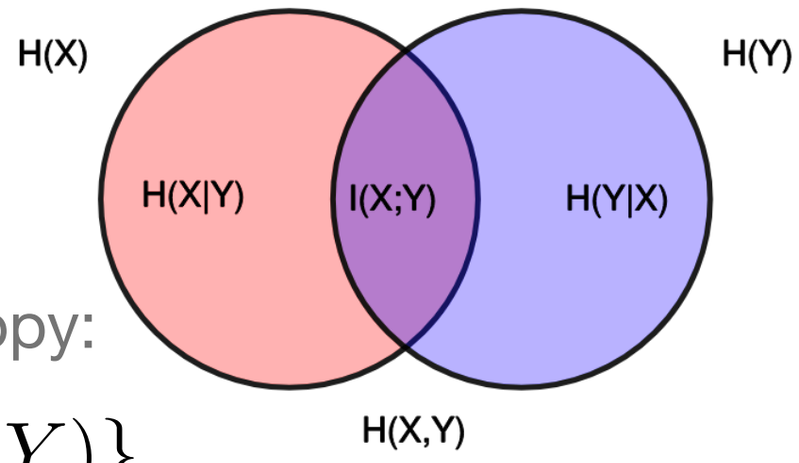


Let's consider the intersection of $H(X), H(Y)$, i.e., the mutual information between X and Y:

$$I(X;Y) := \sum_{i,j} \Pr(X=i, Y=j) \log \frac{\Pr(X=i, Y=j)}{\Pr(X=i)\Pr(Y=j)}$$

Information Theory

Mutual information:



Mutual information is nonnegative and bounded by entropy:

$$0 \leq I(X;Y) \leq \min\{H(X), H(Y)\}$$

Mutual information is the reduced uncertainty by gaining information from another RV:

$$I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$$

Hence we have (why?)

$$X \perp Y \iff I(X;Y) = 0$$

And of course mutual information is symmetric:

$$I(X;Y) = I(Y;X)$$

When $X = Y$, $I(X;X) = H(X)$, hence $I(X;X) \geq I(X;Y)$

Information Theory

Relative Entropy (Kullback-Leibler divergence, KL-divergence):

We need to measure the "distance" between two RVs over the same probability space

$$D_{\mathrm{KL}}(X \parallel Y) := \sum_i \Pr(X = i) \log \frac{\Pr(X = i)}{\Pr(Y = i)}$$

- KL-div is nonnegative
- KL-div is unbounded from above
- KL-div has a meaning of divergence:

$$D_{\mathrm{KL}}(X \parallel Y) = 0 \iff \Pr(X) = \Pr(Y)$$

- However, it is not symmetric:

$$D_{\mathrm{KL}}(X \parallel Y) \neq D_{\mathrm{KL}}(Y \parallel X)$$

- KL-div has a nice coding interpretation: it is the redundancy of coding length using scheme of $\Pr(Y)$ when data comes from $\Pr(X)$

Information Theory

Relationship with other information-theoretic quantities:

$$I(X;Y) = D_{\mathrm{KL}}(\Pr(X,Y) \mid\mid \Pr(X) \cdot \Pr(Y))$$

Hence mutual information essentially measures the divergence between joint distribution and the product of marginal distribution

Define cross-entropy between X, Y as:

$$CH(X,Y) := -\sum_i \Pr(X=i) \log \Pr(Y=i)$$

Then the following identity holds:

$$D_{\mathrm{KL}}(X \mid\mid Y) = CH(X,Y) - H(X)$$

This implies that the usual cross-entropy minimization = KL-divergence minimization = Log-likelihood maximization

Information Theory

Data-processing inequality:

- Post-processing cannot increase information
- Garbage-in-garbage-out

If $X \to Y \to Z$ forms a Markov chain, i.e., $X \perp Z \mid Y$, then

$$I(X;Y) \geq I(X;Z)$$

Think: what does this imply for building machine learning algorithms/ statistical processing?

Lecture 1: Homework

1. Let A be an orthonormal matrix, show that $\forall \mathbf{x}, ||A\mathbf{x}||_2 = ||\mathbf{x}||_2$, i.e., orthonormal transformation preserves length
2. Let $f(\mathbf{x}) = \frac{1}{2}||Ax - b||_2^2$, compute $\nabla f(\mathbf{x})$
3. Assume that we have a biased coin that flips Head with probability p, design an experiment to estimate the value of p with an accuracy within 0.01, and we would like to have your answer to be accurate with probability ≥ 0.99. How many trials you will need to make?
4. Prove that mutual information = 0 iff independent: $X \perp Y \iff I(X;Y) = 0$
5. (Bonus) Let X be a discrete random variable taking k different values, show that $H(X) \leq \log k$. Furthermore, what's the distribution that achieves this maximum value?

Lecture 1: Maths Preliminary and Introduction

Overview:

- Linear Algebra
- Multivariate Calculus
- Probability and Statistics
- Information Theory
- Convex Optimization

Convex Optimization

Recall our bonus problem in the test:

Given a dataset $\{(\mathrm{x}_i, y_i)\}_{i=1}^n$ where $\mathrm{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$, $\forall i \in \{1, \ldots, n\}$. Now given a new datapoint x, we are interested in the problem of finding a linear model using the given dataset and then use the model to predict the value of y corresponding to the new datapoint x. Assume $n \geq d$.

1. To find the best linear fit, we need to minimize the prediction given by the linear model to the true value. Hence we need to solve the following optimization problem:

$$\min_{\mathrm{w} \in \mathbb{R}^d} \quad \sum_{i=1}^{n} (\mathrm{w}^T \mathrm{x}_i - y_i)^2$$

Find the optimal solution w to the above problem in closed form.

Formulate the above problem in matrix form, we have:

$$\min_{w} \quad \frac{1}{2} ||Xw - y||_2^2$$

From HW1, we know that

$$\nabla f(w) = X^T (Xw - y)$$

By setting the gradient to 0, we get the solution:

$$w = (X^T X)^{-1} X^T y$$

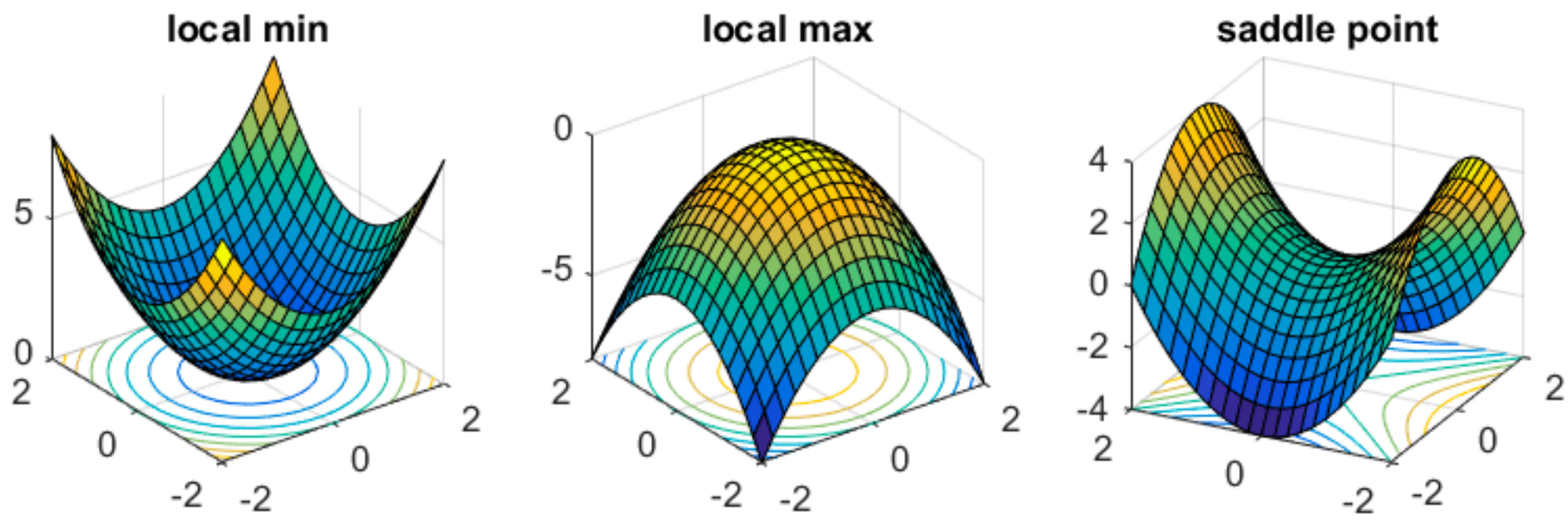
Convex Optimization

But, why setting the gradient to 0 gives us the optimal solution?

Recall from the last lecture, we know that:

Algebraic meaning of the gradient $\nabla f(\mathbf{x})$:

- If $\nabla f(\mathbf{x}) = 0$, then $\mathbf{x}$ is called a stationary point
- There are three cases:



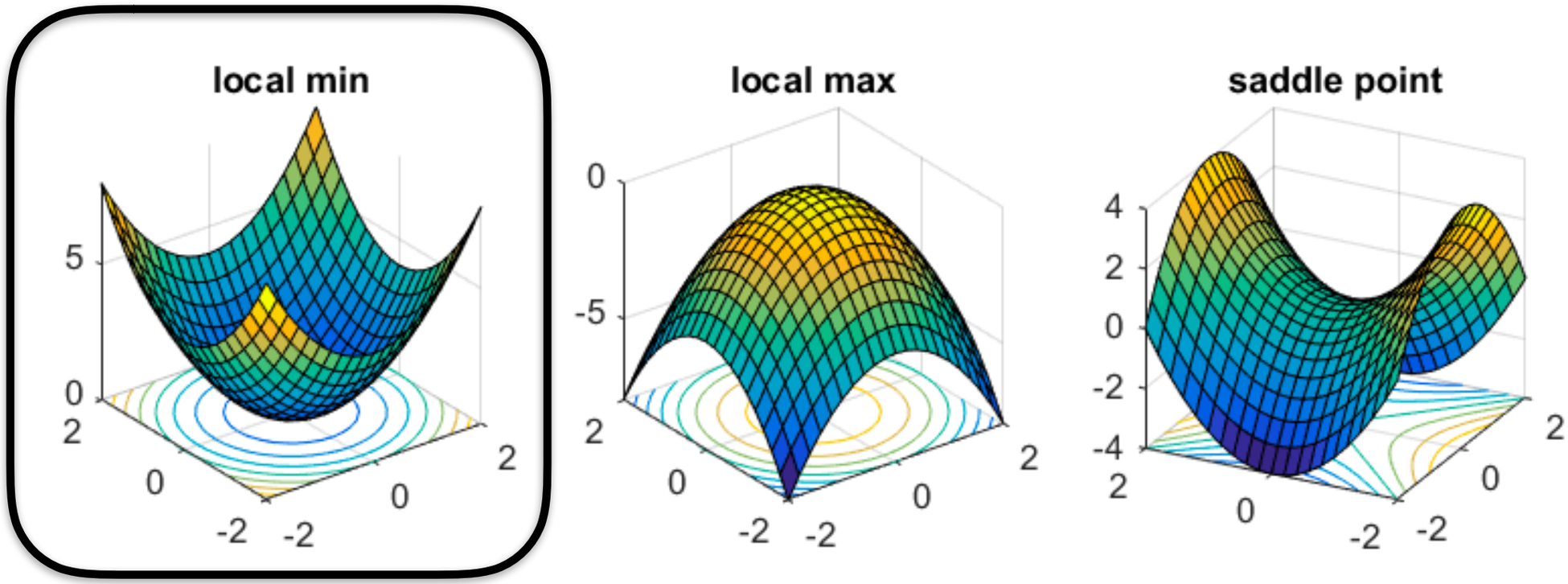
Convex Optimization

But, why setting the gradient to 0 gives us the optimal solution?

Short answer: because the objective function $\frac{1}{2}||Xw - y||_2^2$ is convex in w !

Algebraic meaning of the gradient $\nabla f(\mathbf{x})$:

- If $\nabla f(\mathbf{x}) = 0$, then $\mathbf{x}$ is called a stationary point
- There are three cases:



Convex Optimization

OK, it seems convex functions are important, so what are they?

Let's consider the univariate case:

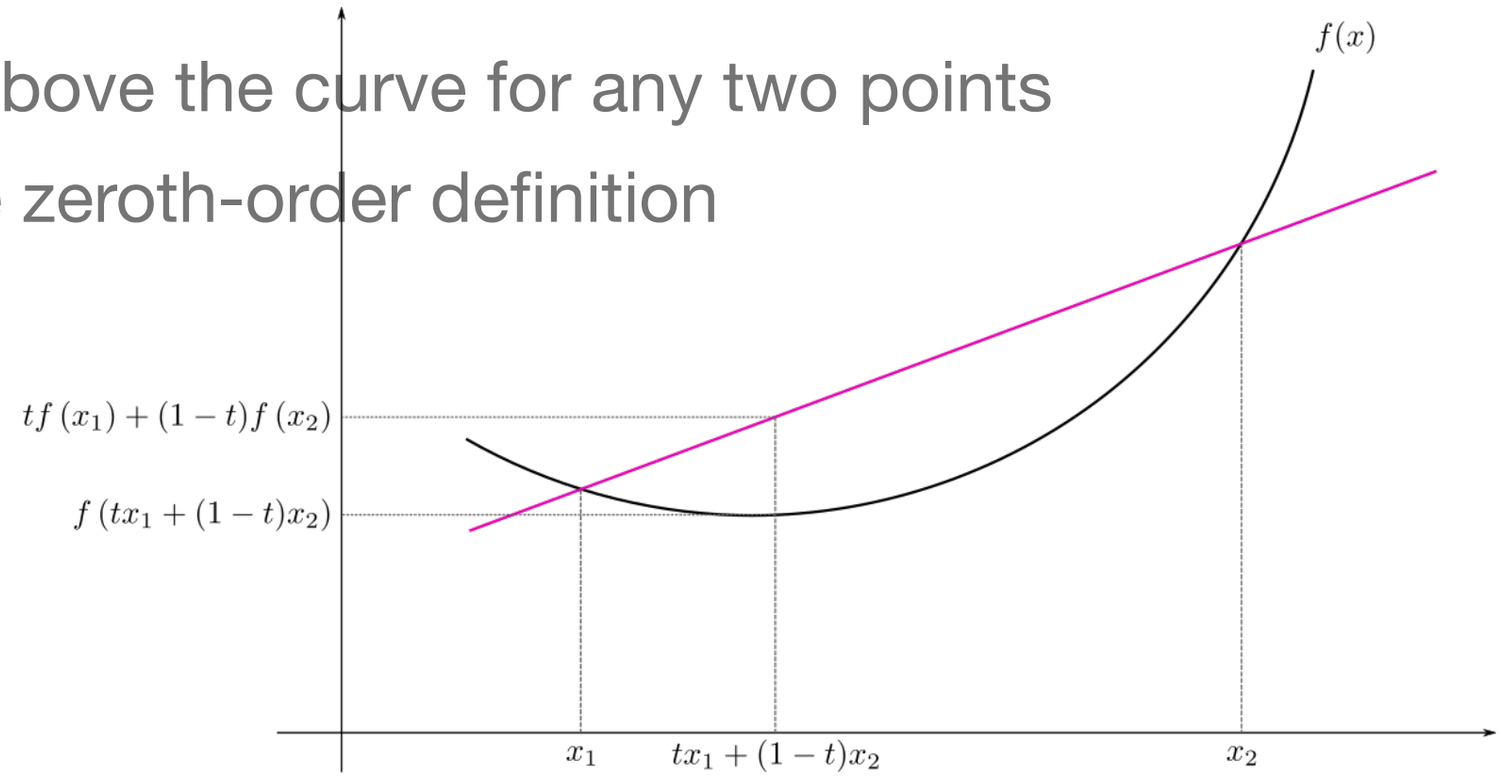
Algebraic definition:

$$\forall t \in [0,1], \forall x_1, x_2, \quad f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$

Geometric intuition:

Line segment is always above the curve for any two points

This is also known as the zeroth-order definition



Convex Optimization

First-order equivalent definition:

Let's consider the univariate case:

Algebraic definition:

$$\forall x, y, \quad f(y) \geq f(x) + f'(x)(y - x)$$

Geometric intuition:

Curve is also above any tangent line

Think: This implies that

$$f'(x) = 0 \implies f(x) = \min_y f(y)$$

Convex Optimization

Second-order equivalent definition:

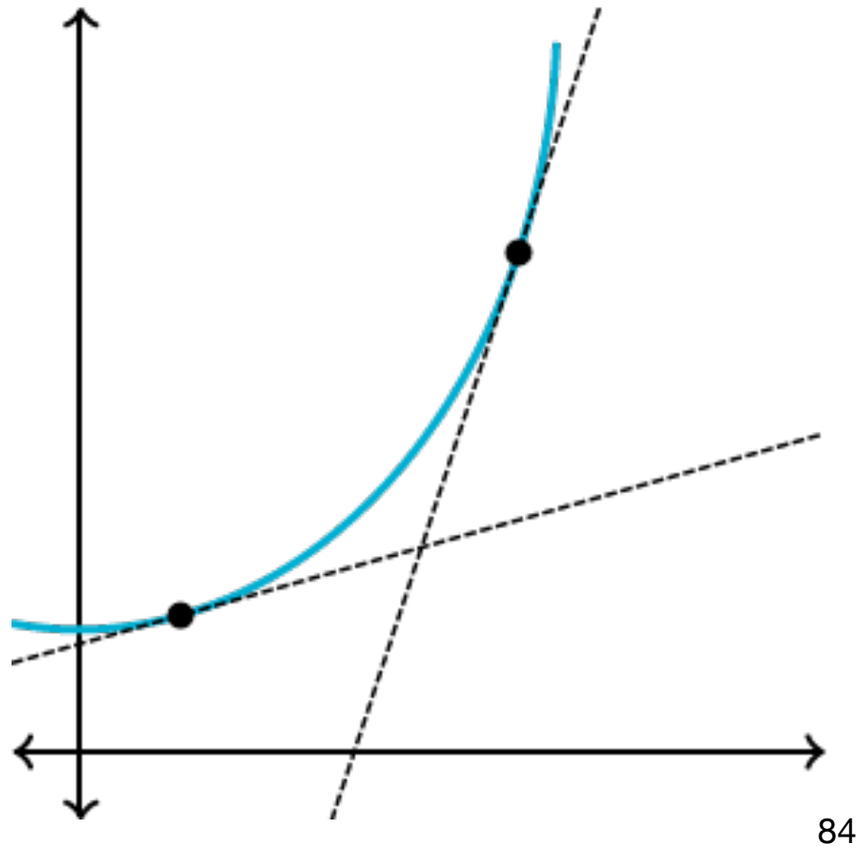
Let's consider the univariate case:

Algebraic definition:

$$\forall x,\ \ f''(x) \geq 0$$

Geometric intuition:

The slope of the curve is non-decreasing



Think: This also implies that

$$f'(x) = 0 \implies f(x) = \min_y \ f(y)$$

Convex Optimization

For convex functions, local optimal = global optimal. In other words,

$$f'(x) = 0 \iff f(x) \text{ is optimal}$$

Now we can extend the definition from univariate case to multivariate case: $f(\mathbf{x})$ is convex iff:

- Zeroth order condition:

$$\forall t \in [0,1], \forall \mathbf{x}_1, \mathbf{x}_2, \quad f\left(t\mathbf{x}_1 + (1-t)\mathbf{x}_2\right) \leq tf\left(\mathbf{x}_1\right) + (1-t)f\left(\mathbf{x}_2\right)$$

- First order condition:

$$\forall \mathbf{x}, \mathbf{y}, \quad f(\mathbf{y}) \geq f(\mathbf{x}) + f'(\mathbf{x})(\mathbf{y} - \mathbf{x})$$

- Second order condition:

$$\forall \mathbf{x}, \ \nabla^2 f(\mathbf{x}) \succeq 0$$

Convex Optimization

Positive semidefinite matrix:

A square matrix A is called positive semidefinite (PSD) if

$$\forall \mathbf{x}, \quad \mathbf{x}^T A \mathbf{x} \geq 0$$

WLOG, we can assume that A is symmetric when we consider quadratic forms (why?), so using the spectral theorem, we can show that:

A is PSD iff all the eigenvalues of A are greater than 0

Proof on board

Convex Optimization

For convex functions, local optimal = global optimal. In other words,

$$f'(x) = 0 \iff f(x) \text{ is optimal}$$

Now we can extend the definition from univariate case to multivariate case: $f(\mathbf{x})$ is convex iff:

- Zeroth order condition:

$$\forall t \in [0,1], \forall \mathbf{x}_1, \mathbf{x}_2, \quad f\left(t\mathbf{x}_1 + (1-t)\mathbf{x}_2\right) \leq tf\left(\mathbf{x}_1\right) + (1-t)f\left(\mathbf{x}_2\right)$$

- First order condition:

$$\forall \mathbf{x}, \mathbf{y}, \quad f(\mathbf{y}) \geq f(\mathbf{x}) + f'(\mathbf{x})(\mathbf{y} - \mathbf{x})$$

- Second order condition: the Hessian matrix is PSD everywhere

$$\forall \mathbf{x}, \ \nabla^2 f(\mathbf{x}) \succeq 0$$

Convex Optimization

Using the second-order definition, let's go back to our bonus problem again:

$$\min_{w} \quad \frac{1}{2}||Xw - y||_2^2$$

From the gradient function $\nabla f(w) = X^T(Xw - y)$, using the result we learned in Lecture 1, we know that

$$\nabla^2 f(\mathbf{w}) = X^T X$$

From the definition of PSD matrix, we can easily verify that

$$X^T X \succeq 0$$

Hence the original problem is convex, and our previous derivation is correct

Convex Optimization

Similarly, a function f is called concave if $-f$ is convex

Important properties about convex functions: Jensen's inequality: for any distribution and assume that f is convex, then the following inequality holds:

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$$

Extremely useful!

- Exercise: use this one to show that $H(X) \leq \log k$
- Exercise: use this one to show that $\mathrm{Var}(X) \geq 0$

Convex Optimization

Common convex functions:

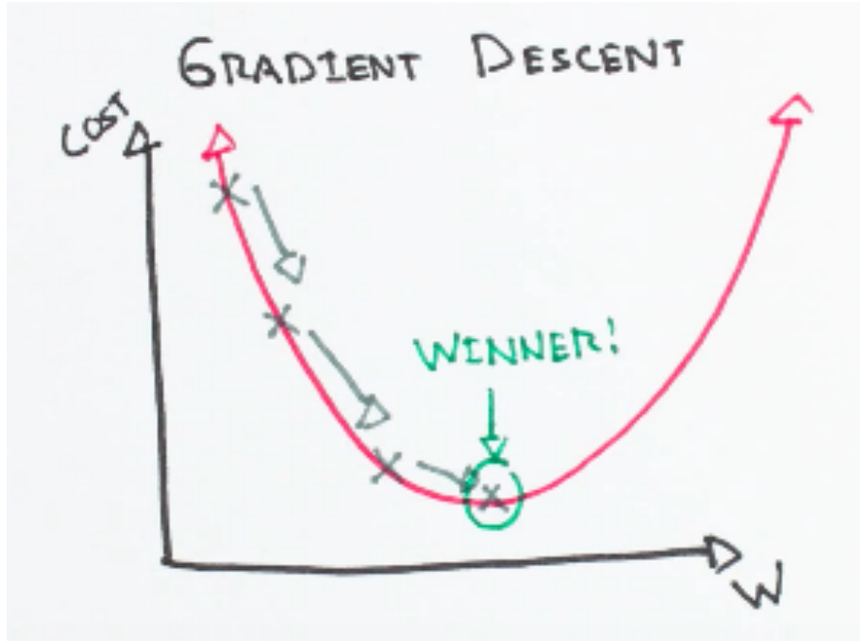
- Univariate functions:
  - Exponential function: $\forall a, \exp(ax)$
  - Power functions: $\forall a \geq 1, a \leq 0, x^a, \forall x \geq 0$
  - Logarithm function: $\log(x), \forall x > 0$
- Affine function is both convex and concave: $\mathbf{a}^T\mathbf{x} + b$
- Quadratic function when Q is positive semidefinite

$$\frac{1}{2}\mathbf{x}^T Q\mathbf{x} + \mathbf{b}^T\mathbf{x} + c$$

Convex Optimization

First-order optimization method: Gradient Descent (GD):
We want to minimize a differentiable convex function $f(\mathbf{x})$

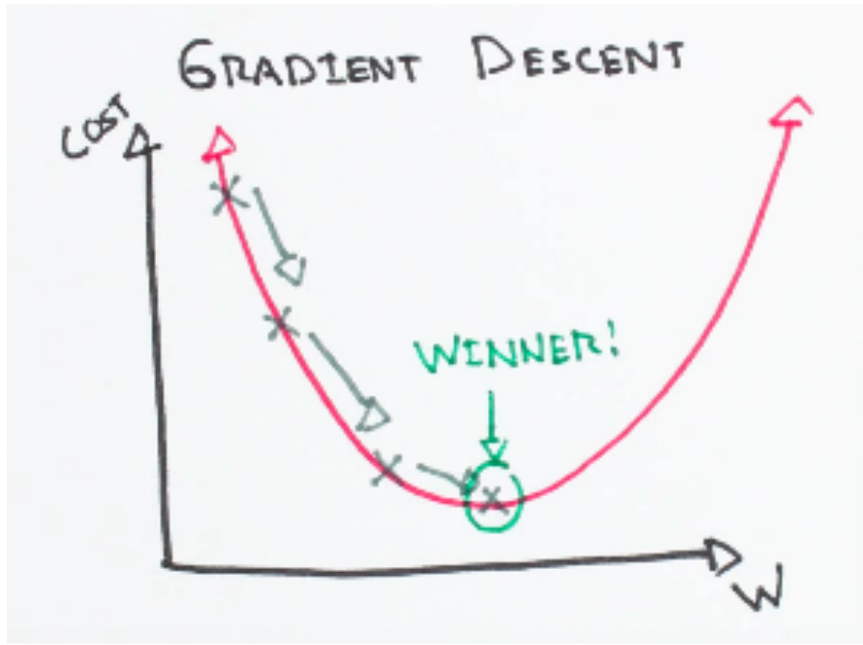
- Algorithm:
  1. Set $\mathbf{x}^{(0)}$ as an arbitrary initial point
  2. For $t = 1, \ldots, \infty$ until convergence, do
     1. Compute $g^{(t)} := \nabla f(\mathbf{x}^{(t-1)})$
     2. Update: $\mathbf{x}^{(t)} \leftarrow \mathbf{x}^{(t-1)} - \gamma \cdot g^{(t)}$



- Intuition: negative gradient of a point is the steepest descent direction at that point (why?)

Convex Optimization

First-order optimization method: Gradient Descent (GD):
We want to minimize a differentiable convex function $f(\mathbf{x})$



- Algorithm:
  1. Set $\mathbf{x}^{(0)}$ as an arbitrary initial point
  2. For $t = 1, \ldots, \infty$ until convergence, do
     1. Compute $g^{(t)} := \nabla f(\mathbf{x}^{(t-1)})$
     2. Update: $\mathbf{x}^{(t)} \leftarrow \mathbf{x}^{(t-1)} - \gamma \cdot g^{(t)}$
- How to choose the learning rate γ?