LxMLS Day 0: Basic Math Recitation

Gopala Krishna Anumanchipalli

INESC-ID/IST Lisboa &
Carnegie Mellon University

July 20, 2011

Agenda

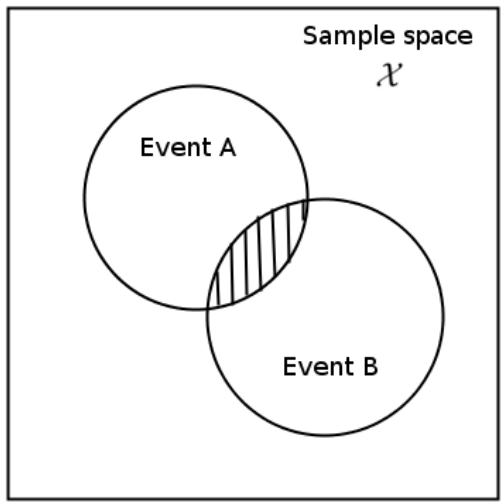
- Probability Theory
- Linear Algebra
- Optimization

Probability

- A language for quantifying - uncertainity, belief in the world, common sense etc.,
- *Sample space* $\mathcal{X}$ is the set of **all possible outcomes** of a conceptual, physical, repeatable experiment
  - Eg., 2-Coin toss: $\mathcal{X} = \{HH, TH, HT, TT\}$
  - Eg., Possible nucleotides at a DNA site: $\mathcal{X} = \{A, T, C, G\}$
  - Eg., Parts of speech: $\mathcal{X} = \{NN, PP, NNP, DT, \ldots\}$.
- *Event* is **any subset** of $\mathcal{X}$
  - Eg., The event of getting heads on only one coin. $(A \subset \mathcal{X}) = \{TH, HT\}$

Probability: Kolmogorov Axioms

- Probability is a function that maps events A into the interval $[0, 1]$: $\mathbb{P} : A \rightarrow [0, 1]$
  - $0 \leq P(A) \leq 1$
  - $P(\mathcal{X}) = 1$
  - $P(\phi) = 0$
  - For two events A and B, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$



Random Variable

- A *random variable* (RV) is a function that associates a unique numerical value $X(\omega)$ with every outcome $\omega \subset \mathcal{X}$ of an experiment.
  - $X(\omega)$ may be finite ($\{0, 1\}$), or infinite ($\mathbb{R}$)
  - $X(\omega)$ may be discrete ($\mathbb{N}$) or continuous ($\mathbb{R}$)
  - $X(\omega)$ may have one or several variables ($\mathbb{R}^d$ random vectors)

Discrete Probability Distribution

- In the discrete case, a probability distribution P on $\mathcal{X}$ is an assignment of a non-negative real number $P(x)$ to each $x \in \mathcal{X}$ such that
  - $0 \leq P(X = x) \leq 1$ and
  - $\sum_x P(X = x) = 1$
- Example: Bernoulli distribution with *parameter* θ

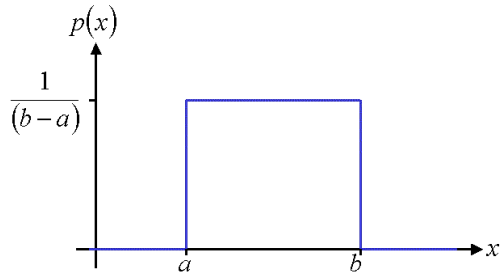
$$P(x) = \begin{cases} 1 - \theta & x = 0 \\ \theta & x = 1 \end{cases}$$

Continuous Probability Distribution

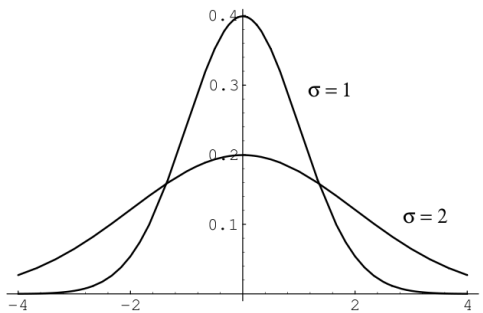
- A continuous random variable, X can assume any value in an interval on the real line or in a region in high dimensional space.
  - We talk about the probability of the variable assuming a value in a given interval $P(X \in [x_1, x_2])$
- The probability of the RV in a given interval $[x_1, x_2]$ is defined to be the **area under the graph** of the *probability density function*
- Probability mass: $P(X \in [x_1, x_2]) = \int_{x_1}^{x_2} p(x)dx$
- Cumulative distribution function(CDF):
  $F(x) = P(X \leq x) = \int_{-\infty}^{x} p(x')dx'$
- Probability density function (PDF): $p(x) = \frac{d}{dx}F(x)$

Continuous Distributions

- Uniform PDF: $$p(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{elsewhere} \end{cases}$$



- Normal (Gaussian) PDF: $p(x) \sim \mathcal{N}(\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$



Statistical Characterizations of RVs

- *Expectation*: The centre of mass, mean value, first moment

$$\mathbb{E}(X) = \begin{cases} \sum_x xp(x) & \text{discrete} \\ \int_{-\infty}^{\infty} xp(x)dx & \text{continuous} \end{cases}$$

- *Variance*: The spread

$$Var(X) = \begin{cases} \sum_x [x - \mathbb{E}(x)]^2 p(x) & \text{discrete} \\ \int_{-\infty}^{\infty} [x - \mathbb{E}(x)]^2 p(x)dx & \text{continuous} \end{cases}$$

- Useful formula: $Var(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

Example: Bernoulli Distribution

- $X \sim Bernoulli(\theta)$, hence
  $\mathcal{X} = \{0, 1\}; P(X = 1) = \theta, P(X = 0) = 1 - \theta$
- The expectation of X is

$$\begin{aligned} \mathbb{E}[X] &= \sum_{i=\{0,1\}} iP(X = i) \\ \mathbb{E}[X] &= 0 * (1 - \theta) + 1 * \theta \\ \mathbb{E}[X] &= \theta \end{aligned}$$

- The variance of X is

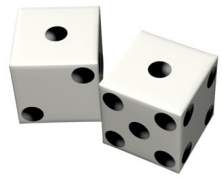
$$\begin{aligned} Var[X] &= \sum_{i=\{0,1\}} i^2 P(X = i) - \left(\sum_{i=\{0,1\}} iP(X = i) \right)^2 \\ Var[X] &= 0^2 * (1 - \theta) + 1^2 * \theta - \theta^2 \\ Var[X] &= \theta(1 - \theta) \end{aligned}$$

Joint Probability

- A joint probability distribution for a set of RVs gives the probability of every atomic event (sample point)
  - Eg., Joint probability $P(X, Y) = P(X = \mathit{true} \wedge Y = \mathit{true})$
- Marginal probability of RV X, $P(X) = \sum_{j \in \mathcal{X}} P(X \wedge Y = y_j)$

Independence

- Events A and B are independent if $P(A \cap B) = P(A) * P(B)$



- Random variables X and Y are *independent* ($X \perp Y$) if and only if their joint probability, $P_{X,Y}$ is the product of their *marginal* probabilities: $P_{X,Y} = P_X P_Y$
  - if X and Y are independent
    $Cov(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0$
  - Corollary: RVs $X_1, \ldots, X_n$ are independent iff
    $P_{X_1,\ldots,X_n} = \prod_{i=1}^{n} P_{X_i}$

Conditional Probability

- Probability of event A conditioned on event B having occured: if $P(B) > 0, P(A|B) = \frac{P(A \cap B)}{P(B)}$
  - Corollary: Chain Rule: $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$
  - If A and B are independent, $P(A|B) = P(A)$
- If RVs (X, Y) and Y have PDF $p_{(X,Y)}$ and p_Y, then $P_{X|Y}$ is the corresponding PDF: $p_{X|Y} = \frac{p_{(X,Y)}}{p_Y}$

Bayes Rule

- For events A and B, joint probability $P(A, B)$ can be written as $P(A)P(B|A)$ or $P(B)P(A|B)$, rearranging,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \text{ (Bayes Rule)}$$

- Bayes rule for *statistical inference*
  - X is a random variable that is **observed**
  - Θ is a random variable denoting the **parameter(s)**
  - Goal: given observed data X, find the best guess for Θ

- Bayes rule : $P_{\Theta|Y}(\theta|y) = \frac{P_{Y|\Theta}(y|\theta)P_{\Theta}(\theta)}{\int P_{Y|\Theta}(y|\theta)P_{\Theta}(\theta)d\theta}$

$$\textit{posterior} = \frac{\textit{likelihood} \times \textit{prior}}{\textit{joint probability}}$$

Estimators

Under independent and identically distributed (iid) assumptions likelihood of the data, given a model is

$$\mathcal{L}(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \tag{1}$$

- Maximum Likelihood Estimate : $\theta_{ML} = argmax_\theta P_{Y|\Theta}(y|\theta)$.
- Maximum a posteriori Estimate : $\theta_{MAP} = argmax_\theta P_{\Theta|Y}(\theta|y)$.

Conjugate priors: Choosing a conjugate prior ensures that the posterior distribution is the same family of distributions as the likelihood. Ex: exponential family; bernoulli vs. beta etc.,

Example: Bernoulli model

- Data:
  - We observed N ***iid*** coin tossing: $D = \{1, 0, 1, \ldots, 0\}$
- Model:

$$P(x) = \begin{cases} 1-\theta & \text{for } x = 0 \\ \theta & \text{for } x = 1 \end{cases} \quad \Rightarrow \quad P(x) = \theta^{x}(1-\theta)^{1-x}$$

- How to write the likelihood of a single observation x_i?

$$P(x_i) = \theta^{x_i}(1-\theta)^{1-x_i}$$

- The likelihood of dataset $D = \{x_1, \ldots, x_N\}$:

$$L(\theta) = P(x_1, x_2, \ldots, x_N; \theta) = \prod_{i=1}^{N} P(x_i; \theta) = \prod_{i=1}^{N} \left(\theta^{x_i}(1-\theta)^{1-x_i}\right)$$

$$= \theta^{\sum_{i=1}^{N} x_i}(1-\theta)^{\sum_{i=1}^{N} 1-x_i} = \theta^{\#\text{head}}(1-\theta)^{\#\text{tails}}$$

MLE: Bernoulli model

- Objective function:

$$\ell(\theta) = \log L(\theta) = \log \theta^{n_h} (1-\theta)^{n_t} = n_h \log \theta + (N - n_h) \log(1-\theta)$$

- We need to maximize this w.r.t. θ
- Take derivatives wrt θ

$$\frac{\partial \ell}{\partial \theta} = \frac{n_h}{\theta} - \frac{N - n_h}{1-\theta} = 0 \quad \Rightarrow \quad \hat{\theta}_{MLE} = \frac{n_h}{N} \quad \text{or} \quad \hat{\theta}_{MLE} = \frac{1}{N} \sum_i x_i$$

Frequency as sample mean

- Sufficient statistics
  - The counts, n_h, where $n_h = \sum_i x_i$, are **sufficient statistics** of data D

Example: MLE: Univariate Normal Distribution

- Data:
  - We observed *N* ***iid*** real samples:
    D={-0.1, 10, 1, -5.2, …, 3}
- Model: $P(x) = (2\pi\sigma^2)^{-1/2} \exp\{-(x-\mu)^2 / 2\sigma^2\}$ $\theta = (\mu, \sigma^2)$
- Log likelihood:

$$\ell(\theta) = \log L(\theta) = \prod_{i=1}^{N} P(x_i) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum_{i=1}^{N}\frac{(x_i-\mu)^2}{\sigma^2}$$

- MLE: take derivative and set to zero:

$$\frac{\partial \ell}{\partial \mu} = (1/\sigma^2)\sum_n (x_n - \mu)$$

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_n (x_n-\mu)^2$$

$$\mu_{MLE} = \frac{1}{N}\sum_n x_n$$

$$\sigma^2_{MLE} = \frac{1}{N}\sum_n (x_n - \mu_{ML})^2$$

Agenda

- ~~Probability Theory~~ ✓
- Linear Algebra
- Optimization

Linear Algebra

- Linear Algebra provides a compact way of representing and operating on sets

$$\begin{aligned} 4x_1 \quad -5x_2 \quad &= -13 \\ -2x_1 \quad +3x_2 \quad &= 9 \end{aligned}$$

- This is a system of linear equations in 2 variables. In matrix notation we can write the system more compactly as

$$Ax = b$$

with

$$A = \begin{bmatrix} 4 & 5 \\ -2 & 3 \end{bmatrix}, \; b = \begin{bmatrix} -13 \\ 9 \end{bmatrix}$$

Notation

- $A \in \mathbb{R}^{m \times n}$ is a **matrix** with m rows and n columns.
- $x \in \mathbb{R}^n$ is a **vector** with n entries.
- A vector can also be thought of as a matrix with n rows and 1 column, known as a **column vector**.
- A **row vector** a matrix with 1 row and n columns is denoted as x^T, the transpose of x.
- The ith element of a vector x is denoted x_i

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

Inner/ Outer Products of vectors

- Given vectors $x, y \in \mathbb{R}^n$ the **inner product** $x^T y$, or **dot product**:

$$x^T y \in \mathbb{R} = \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \sum_{i=1}^{n} x_i y_i.$$

- Given vectors $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^n$, the **outer product** $xy^T \in \mathbb{R}^{m \times n}$ is given by $xy^T \in \mathbb{R}^{m \times n} =$

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \begin{bmatrix} y_1 & y_2 & \dots & y_n \end{bmatrix} = \begin{bmatrix} x_1 y_1 & x_1 y_2 & \dots & x_1 y_n \\ x_2 y_1 & x_2 y_2 & \dots & x_2 y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_m y_1 & x_m y_2 & \dots & x_m y_n \end{bmatrix}.$$

Norms of vectors

- The **norm** of a vector is informally the measure of the "length" of the vector. The commonly used Euclidean or ℓ_2 norm is given by

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

- More generally, the ℓ_p norm of a vector $x \in \mathbb{R}^n$, where $p \geq 1$ is defined as

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}.$$

Note:

$$\ell_1 \text{ norm} : \|x\|_1 = \sum_{i=1}^{n} |x_i| \qquad \ell_\infty \text{ norm} : \|x\|_\infty = \max_i |x_i| \ .$$

Vector spaces

A set of vectors $\mathcal{S}$ is a vector space if it is closed under

- Addition
- Multiplication by a scalar

A subspace A is a subset of vector space $\mathcal{S}$ that is also closed under the above operations.

Matrix Operations

- Product of two matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$ is the matrix $C = AB \in \mathbb{R}^{m \times p}$, where

$$C_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}.$$

  - Matrix multiplication is associative: $(AB)C = A(BC)$.
  - Matrix multiplication is distributive: $A(B + C) = AB + AC$.
  - Matrix multiplication is (generally) not commutative : $AB \neq BA$.

- The **transpose** of a matrix results from "'flipping" the rows and columns. Given a matrix $A \in \mathbb{R}^{m \times n}$, the transpose $A^T \in \mathbb{R}^{n \times m}$ is the $n \times m$ matrix whose entries are given by $(A^T)_{ij} = A_{ji}$.
  Also,
  $(A^T)^T = A; \qquad (AB)^T = B^T A^T; \qquad (A + B)^T = A^T + B^T$

Special Matrices

- The **identity matrix**, denoted $I \in \mathbb{R}^{n \times n}$, is a square matrix with ones on the diagonal and zeros everywhere else. That is,

$$I_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

  It has the property that for all $A \in \mathbb{R}^{m \times n}$, $AI = A = IA$.
- A **diagonal matrix** is a matrix where all non-diagonal elements are 0.
- A square matrix $A \in \mathbb{R}^{n \times n}$ is **symmetric** if $A = A^T$.
  - The **trace** of a square matrix $A \in \mathbb{R}^{n \times n}$ is the sum of the diagonal elements, $\text{tr}(A) = \sum_{i=1}^{n} A_{ii}$

Linear Independence and Rank

- A set of vectors $\{x_1, x_2, \ldots, x_n\} \subset \mathbb{R}^m$ is said to be **(linearly) independent** if no vector can be represented as a linear combination of the remaining vectors.
- The **rank** of a matrix is the number of linearly independent columns.
  - For $A \in R^{m \times n}$, $\text{rank}(A) \leq \min(m, n)$. If $\text{rank}(A) = \min(m, n)$, then A is said to be **full rank**.
  - For $A \in R^{m \times n}$, $\text{rank}(A) = \text{rank}(A^T)$.
  - For $A \in R^{m \times n}$, $B \in R^{n \times p}$, $\text{rank}(AB) \leq \min(\text{rank}(A), \text{rank}(B))$.
  - For $A, B \in R^{m \times n}$, $\text{rank}(A + B) \leq \text{rank}(A) + \text{rank}(B)$.

Orthogonal Matrix

- Two vectors $x, y \in \mathbb{R}^n$ **orthogonal** if $x^T y = 0$. A square matrix $U \in \mathbb{R}^{n \times n}$ is orthogonal if all its columns are orthogonal to each other and are normalized ($\|x\|_2 = 1$), It follows that

$$U^T U = I = U U^T.$$

Determinant of a Matrix

- The **determinant** of a square matrix $A \in \mathbb{R}^{n \times n}$, is a function det: $\mathbb{R}^{n \times n} \to \mathbb{R}$, denoted $|A|$. The absolute value of the determinant is a measure of the "volume" of the restricted span S of set of column vectors.
  - For $A \in \mathbb{R}^{n \times n}$, $|A| = |A^T|$.
  - For $A, B \in \mathbb{R}^{n \times n}$, $|AB| = |A||B|$.
  - For $A \in \mathbb{R}^{n \times n}$, $|A| = 0$ if and only if A is singular (i.e., non-invertible).
  - For $A \in \mathbb{R}^{n \times n}$ and A is non-singular, $|A^{-1}| = 1/|A|$.

Inverse of a matrix

- The **inverse** of a square matrix $A \in \mathbb{R}^{n \times n}$ is denoted by A^{-1}, and is a unique matrix such that $A^{-1}A = I = AA^{-1}$. Only some square matrices have inverses, and these are also referred to as **invertible** or **non-singular** matrices. For A^{-1} to exist, A must be full rank.
  - $(A^{-1})^{-1} = A$
  - $(AB)^{-1} = B^{-1}A^{-1}$
  - $(A^{-1})^T = (A^T)^{-1}$.

Positive Semi-definite matrices

Given a square matrix $A \in \mathbb{R}^{n \times n}$ and a vector $x \in \mathbb{R}^n$, the scalar value $x^T Ax$ is called a **quadratic form**.

$$x^T Ax = \sum_{i=1}^{n} x_i (Ax)_i = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j.$$

- A symmetric matrix $A \in \mathbb{R}^{n \times n}$ is **positive semi-definite**(PSD) if for all non-zero vectors $x \in \mathbb{R}^n$, $x^T Ax \geq 0$. If strictly, $x^T Ax > 0$, A is **positive definite** (PD).
- All positive-definite and negative-definite matrices are full rank, hence are invertible.

Matrix Factorizations

Important to find inverses, bases and solutions to equations

- Singular Value Decomposition (SVD) : $A = UDV^T$

$$\underset{m \times n}{\overset{X}{\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & \\ \vdots & \vdots & \ddots & \\ x_{m1} & & & x_{mn} \end{pmatrix}}} = \underset{m \times r}{\overset{U}{\begin{pmatrix} u_{11} & \cdots & u_{1r} \\ \vdots & \ddots & \\ u_{m1} & & u_{mr} \end{pmatrix}}} \underset{r \times r}{\overset{S}{\begin{pmatrix} s_{11} & 0 & \cdots \\ 0 & \ddots & \\ \vdots & & s_{rr} \end{pmatrix}}} \underset{r \times n}{\overset{V^\top}{\begin{pmatrix} v_{11} & \cdots & v_{1n} \\ \vdots & \ddots & \\ v_{r1} & & v_{rn} \end{pmatrix}}}$$

- $A = LU$ for diagonally dominant matrices

$$\begin{bmatrix} \mathbf{m}_{11} & \mathbf{m}_{12} & \mathbf{m}_{13} & \cdots & \mathbf{m}_{1N} \\ \mathbf{m}_{21} & \mathbf{m}_{22} & \mathbf{m}_{23} & \cdots & \mathbf{m}_{2N} \\ \mathbf{m}_{31} & \mathbf{m}_{32} & \mathbf{m}_{33} & \cdots & \mathbf{m}_{3N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{m}_{N1} & \mathbf{m}_{N2} & \mathbf{m}_{N3} & \cdots & \mathbf{m}_{NN} \end{bmatrix} = \begin{bmatrix} \mathbf{u}_{11} & \mathbf{u}_{12} & \mathbf{u}_{13} & \cdots & \mathbf{u}_{1N} \\ 0 & \mathbf{u}_{22} & \mathbf{u}_{23} & \cdots & \mathbf{u}_{2N} \\ 0 & 0 & \mathbf{u}_{33} & \cdots & \mathbf{u}_{3N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \mathbf{u}_{NN} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ \mathbf{l}_{21} & 1 & 0 & \cdots & 0 \\ \mathbf{l}_{31} & \mathbf{l}_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{l}_{N1} & \mathbf{l}_{N2} & \mathbf{l}_{N3} & \cdots & 1 \end{bmatrix}$$

- Symmetric Semi-definite matrices: $A = U^T U$

Gradient of a $f : \mathbb{R}^{m \times n} \to \mathbb{R}$

The Gradient: Suppose that $f : \mathbb{R}^{m \times n} \to \mathbb{R}$ is a function that takes as input, a matrix A of size $m \times n$ and returns a real value. Then the **gradient** of f (with respect to $A \in \mathbb{R}^{m \times n}$) is the matrix of partial derivatives, defined as:

$$\nabla_A f(A) \in \mathbb{R}^{m \times n} = \begin{bmatrix} \frac{\partial f(A)}{\partial A_{11}} & \frac{\partial f(A)}{\partial A_{12}} & \cdots & \frac{\partial f(A)}{\partial A_{1n}} \\ \frac{\partial f(A)}{\partial A_{21}} & \frac{\partial f(A)}{\partial A_{22}} & \cdots & \frac{\partial f(A)}{\partial A_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(A)}{\partial A_{m1}} & \frac{\partial f(A)}{\partial A_{m2}} & \cdots & \frac{\partial f(A)}{\partial A_{mn}} \end{bmatrix}.$$

- $\nabla_x(f(x) + g(x)) = \nabla_x f(x) + \nabla_x g(x)$.
- For $t \in \mathbb{R}$, $\nabla_x(tf(x)) = t\nabla_x f(x)$.
- Let $f : \mathbb{R}^m \to \mathbb{R}$ be the function defined by $f(z) = z^T z$, $\nabla_z f(z) = 2z$.

The Hessian Matrix

The Hessian: Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is a function that takes a vector in $\mathbb{R}^n$ and returns a real number, then the Hessian matrix with respect to x, written $\nabla_x^2 f(x)$ (or H) is the $n \times n$ matrix of partial derivatives,

$$\nabla_x^2 f(x) \in \mathbb{R}^{n \times n} = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix}.$$

Least Squares

(Least squares) Given a full rank matrix $A \in \mathbb{R}^{m \times n}$ and a vector $b \in \gg$ such that $b \notin \mathcal{R}(A)$, where $\mathcal{R}(A)$ is the vector space of the columns of matrix A. In this case, it is not possible to find a vector $x \in \mathbb{R}^n$, such that $Ax = b$. So, instead we want to find a vector x such that Ax is as close as possible to b, as measured by the L_2 norm $\|Ax - b\|_2^2$.
Using the fact that $\|x\|_2^2 = x^T x$, we have

$$\begin{aligned} \|Ax - b\|_2^2 &= (Ax - b)^T (Ax - b) \\ &= x^T A^T Ax - 2b^T Ax + b^T b \end{aligned}$$

Taking gradient with respect to x, we have

$$\begin{aligned} \nabla_x (x^T A^T Ax - 2b^T Ax + b^T b) &= \nabla_x x^T A^T Ax - \nabla_x 2b^T Ax + \nabla_x b^T b \\ &= 2A^T Ax - 2A^T b \end{aligned}$$

Setting this last expression to zero and solving for x, gives the **normal equations** for the least-squares problem:

$$x = (A^T A)^{-1} A^T b$$

Agenda

- ~~Probability Theory~~ ✓
- ~~Linear Algebra~~ ✓
- Optimization

Minimizing a function

Goal: find the minimum/minimizer of $f : \mathbb{R}^d \to \mathbb{R}$

Existence of a mimimum ?

- $f^{'} = 0$
- $f^{''}$ is positive.

If $f : \mathbb{R}^{m \times n} \to \mathbb{R}$, conditions for minimum :

- Hessian matrix $f^{''}$ is positive semi-definite.

Are these global minima ?

- No, (local minima, saddle points, . . .)

Iterative descent methods

Goal: find the minimum/minimizer of $f : \mathbb{R}^d \to \mathbb{R}$

- Proceed in **small steps** in the **optimal direction** till a **stopping criterion** is met.
- **Gradient descent**: updates of the form: $x^{(t+1)} \leftarrow x^{(t)} - \eta_{(t)} \nabla f(x^{(t)})$

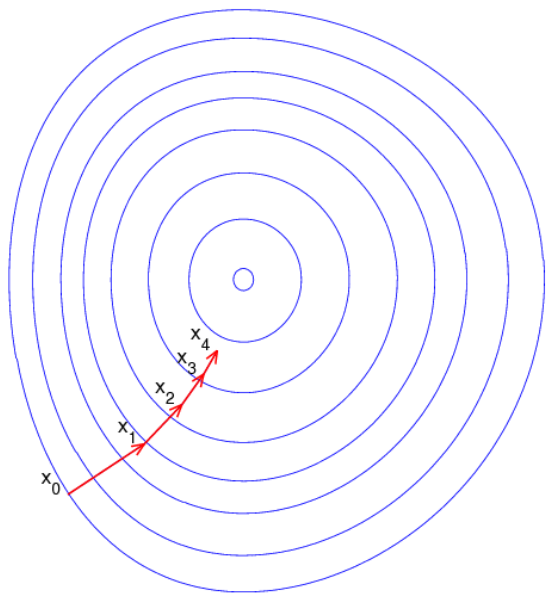


Figure: Illustration of gradient descent.The blue circles correspond to the function values at different points, while the red lines correspond to steps taken in the negative gradient direction.

Convex functions

Pro: Guarantee of a global minima ✓

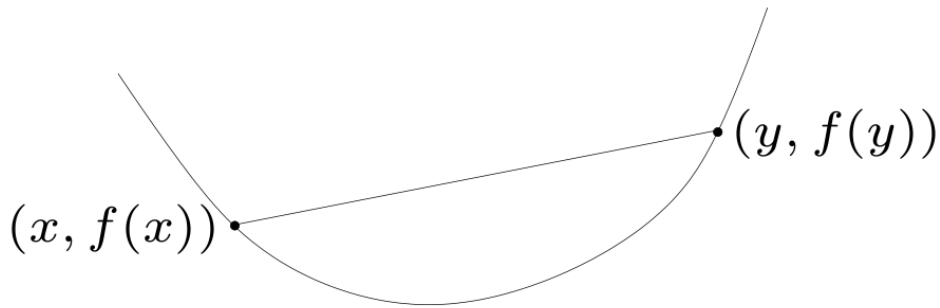


Figure: Illustration of a convex function. The line segment between any two points on the graph lies entirely above the curve.

Non-Convex functions

Pro: No guarantee of a global minima ✗

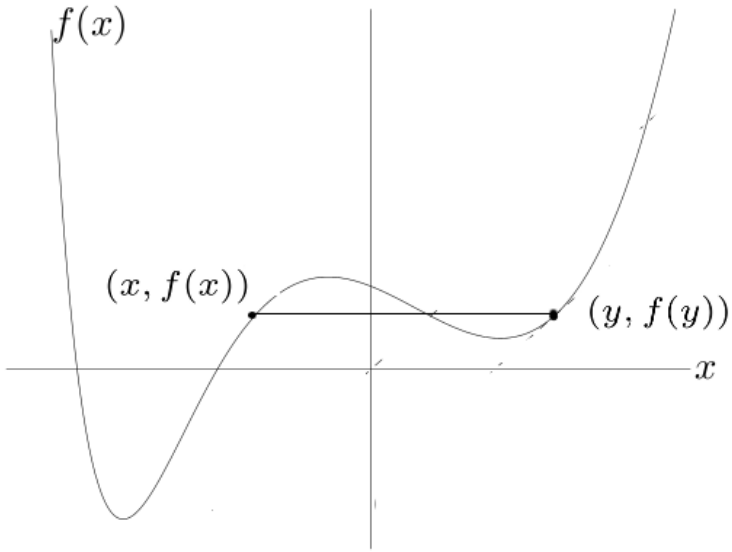


Figure: Illustration of a non-convex function. Note the line segment intersecting the curve.

Agenda

- ~~Probability Theory~~ ✓
- ~~Linear Algebra~~ ✓
- ~~Optimization~~ ✓

All the Best!