Lecture 1. Mathematical Background

Advanced Optimization (Fall 2022)

Peng Zhao

zhaop@lamda.nju.edu.cn

Nanjing University

Outline

- Calculus
- Linear Algebra
- Probability & Statistics
- Information Theory
- Optimization in Machine Learning

Notational Convention

- $[n] = \{1, \ldots, n\}$
- $\mathbf{x}, \mathbf{y}, \mathbf{v}$: vectors
- A, B: matrices
- $\mathcal{X}, \mathcal{Y}, \mathcal{K}$: domain
- d, m, n: dimensions
- I: identity matrix
- X, Y: random variables
- $\boldsymbol{p}, \boldsymbol{q}$: probability distributions

Function

- Function mapping $f : \operatorname{dom} f \subseteq \mathcal{X} \subseteq \mathbb{R}^n \to \mathcal{Y} \subseteq \mathbb{R}^m$

Definition 1 (Continuous Function). A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is continuous at $\mathbf{x} \in \operatorname{dom} f$ if for all $\epsilon > 0$ there exists a $\delta > 0$ with $\mathbf{y} \in \operatorname{dom} f$, such that

$$\|\mathbf{y} - \mathbf{x}\|_2 \leq \delta \Rightarrow \|f(\mathbf{y}) - f(\mathbf{x})\|_2 \leq \epsilon.$$

Part 1. Calculus

- Gradient and Derivatives
- Hessian
- Chain Rule

Gradient and Derivatives (First Order)

- The gradient and derivative of a scalar function ($f : \mathbb{R} \mapsto \mathbb{R}$) is the same.
- The derivative of vector functions ($f : \mathcal{X} \subseteq \mathbb{R}^d \mapsto \mathbb{R}$) is the transpose of its gradient.

We focus on the "gradient" language (i.e., column vector).

Definition 2 (Gradient). Let $f : \mathcal{X} \subseteq \mathbb{R}^d \to \mathbb{R}$ be a differentiable function. Let $\mathbf{x} = [x_1, \cdots, x_d]^\top \in \mathcal{X}$. Then, the gradient of f at $\mathbf{x}$ is a vector in $\mathbb{R}^d$ denoted by $\nabla f(\mathbf{x})$ and defined by

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(\mathbf{x}) \\ \vdots \\ \frac{\partial f}{\partial x_d}(\mathbf{x}) \end{bmatrix}.$$

Example

Example 1. The gradient of $f(\mathbf{x}) = \|\mathbf{x}\|_2^2 \triangleq \sum_{i=1}^d x_i^2$ is

$$\nabla f(\mathbf{x}) = \begin{bmatrix} 2x_1 \\ \vdots \\ 2x_d \end{bmatrix} = 2\mathbf{x}.$$

Example 2. The gradient of $f(\mathbf{x}) = -\sum_{i=1}^d x_i \ln x_i$ is

$$\nabla f(\mathbf{x}) = \begin{bmatrix} -(\ln x_1 + 1) \\ \vdots \\ -(\ln x_d + 1) \end{bmatrix}.$$

Hessian (Second Order)

Definition 3 (Hessian). Let $f : \mathcal{X} \subseteq \mathbb{R}^d \to \mathbb{R}$ be a twice differentiable function. Let $\mathbf{x} = [x_1, \cdots, x_d]^\top \in \mathcal{X}$. Then, the Hessian of f at $\mathbf{x}$ is the matrix in $\mathbb{R}^{d \times d}$ denoted by $\nabla^2 f(\mathbf{x})$ and defined by

$$\nabla^2 f(\mathbf{x}) = \left[\frac{\partial^2 f}{\partial x_i, x_j}(\mathbf{x}) \right]_{1 \le i,j \le d}.$$

Example 3. The Hessian of $f(\mathbf{x}) = -\sum_{i=1}^d x_i \ln x_i$ is $\nabla^2 f(\mathbf{x}) = \operatorname{diag}(-\frac{1}{x_1}, \ldots, -\frac{1}{x_d})$.

Example 4. The Hessian of $f(\mathbf{x}) = x_1^3 x_2^2 - 3x_1 x_2^3 + 1$ is $\nabla^2 f(\mathbf{x}) = \begin{bmatrix} 6x_1 x_2^2 & 6x_1^2 x_2 - 6x_2 \\ 6x_1^2 x_2 - 9x_2^2 & 2x_1^3 - 18 x_1 x_2 \end{bmatrix}$.

Chain Rule

- Consider scalar functions for simplicity.

Chain Rule. For $h(x) = f(g(x))$,

- the gradient of $h(x)$ is $h'(x) = f'(g(x))g'(x)$.
- the Hessian of $h(x)$ is $h''(x) = f''(g(x))(g'(x))^2 + f'(g(x))g''(x)$.

Reference: The Matrix Cookbook

The derivatives of **vectors, matrices, norms, determinants, etc** can be found therein.

2.4.1 First Order

$$\frac{\partial \mathbf{x}^T\mathbf{a}}{\partial \mathbf{x}} = \frac{\partial \mathbf{a}^T\mathbf{x}}{\partial \mathbf{x}} = \mathbf{a} \tag{69}$$

$$\frac{\partial \mathbf{a}^T\mathbf{X}\mathbf{b}}{\partial \mathbf{X}} = \mathbf{a}\mathbf{b}^T \tag{70}$$

$$\frac{\partial \mathbf{a}^T\mathbf{X}^T\mathbf{b}}{\partial \mathbf{X}} = \mathbf{b}\mathbf{a}^T \tag{71}$$

$$\frac{\partial \mathbf{a}^T\mathbf{X}\mathbf{a}}{\partial \mathbf{X}} = \frac{\partial \mathbf{a}^T\mathbf{X}^T\mathbf{a}}{\partial \mathbf{X}} = \mathbf{a}\mathbf{a}^T \tag{72}$$

$$\frac{\partial \mathbf{X}}{\partial X_{ij}} = \mathbf{J}^{ij} \tag{73}$$

$$\frac{\partial (\mathbf{X}\mathbf{A})_{ij}}{\partial X_{mn}} = \delta_{im}(\mathbf{A})_{nj} = (\mathbf{J}^{mn}\mathbf{A})_{ij} \tag{74}$$

$$\frac{\partial (\mathbf{X}^T\mathbf{A})_{ij}}{\partial X_{mn}} = \delta_{in}(\mathbf{A})_{mj} = (\mathbf{J}^{nm}\mathbf{A})_{ij} \tag{75}$$

2 Derivatives

This section is covering differentiation of a number of expressions with respect to a matrix $\mathbf{X}$. Note that it is always assumed that $\mathbf{X}$ has *no special structure*, i.e. that the elements of $\mathbf{X}$ are independent (e.g. not symmetric, Toeplitz, positive definite). See section 2.8 for differentiation of structured matrices. The basic assumptions can be written in a formula as

$$\frac{\partial X_{kl}}{\partial X_{ij}} = \delta_{ik}\delta_{lj} \tag{32}$$

that is for e.g. vector forms,

$$\left[\frac{\partial \mathbf{x}}{\partial y}\right]_i = \frac{\partial x_i}{\partial y} \qquad \left[\frac{\partial x}{\partial \mathbf{y}}\right]_i = \frac{\partial x}{\partial y_i} \qquad \left[\frac{\partial \mathbf{x}}{\partial \mathbf{y}}\right]_{ij} = \frac{\partial x_i}{\partial y_j}$$

The following rules are general and very useful when deriving the differential of an expression ([19]):

$$\partial \mathbf{A} = 0 \quad (\mathbf{A} \text{ is a constant}) \tag{33}$$

$$\partial(\alpha \mathbf{X}) = \alpha \partial \mathbf{X} \tag{34}$$

$$\partial(\mathbf{X} + \mathbf{Y}) = \partial \mathbf{X} + \partial \mathbf{Y} \tag{35}$$

$$\partial(\mathrm{Tr}(\mathbf{X})) = \mathrm{Tr}(\partial \mathbf{X}) \tag{36}$$

$$\partial(\mathbf{X}\mathbf{Y}) = (\partial \mathbf{X})\mathbf{Y} + \mathbf{X}(\partial \mathbf{Y}) \tag{37}$$

$$\partial(\mathbf{X} \circ \mathbf{Y}) = (\partial \mathbf{X}) \circ \mathbf{Y} + \mathbf{X} \circ (\partial \mathbf{Y}) \tag{38}$$

$$\partial(\mathbf{X} \otimes \mathbf{Y}) = (\partial \mathbf{X}) \otimes \mathbf{Y} + \mathbf{X} \otimes (\partial \mathbf{Y}) \tag{39}$$

$$\partial(\mathbf{X}^{-1}) = -\mathbf{X}^{-1}(\partial \mathbf{X})\mathbf{X}^{-1} \tag{40}$$

$$\partial(\det(\mathbf{X})) = \mathrm{Tr}(\mathrm{adj}(\mathbf{X})\partial \mathbf{X}) \tag{41}$$

$$\partial(\det(\mathbf{X})) = \det(\mathbf{X})\mathrm{Tr}(\mathbf{X}^{-1}\partial \mathbf{X}) \tag{42}$$

$$\partial(\ln(\det(\mathbf{X}))) = \mathrm{Tr}(\mathbf{X}^{-1}\partial \mathbf{X}) \tag{43}$$

$$\partial \mathbf{X}^T = (\partial \mathbf{X})^T \tag{44}$$

$$\partial \mathbf{X}^H = (\partial \mathbf{X})^H \tag{45}$$

https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf

Part 2. Linear Algebra

- Positive (Semi-)Definite Matrix
- Rank
- Inner Product, Norm, Matrix Norm
- Matrix Decomposition

Positive (Semi-)Definite Matrix

Definition 4 (Positive Definite, PD). A matrix $A \in \mathbb{R}^{d \times d}$ is positive definite, if for all $\mathbf{x} \neq \mathbf{0}, \mathbf{x}^\top A\mathbf{x} > 0$, usually denoted as $A \succ 0$.

Definition 5 (Positive Semi-Definite, PSD). A matrix $A \in \mathbb{R}^{d \times d}$ is positive semi-definite, if for all $\mathbf{x} \in \mathbb{R}^d, \mathbf{x}^\top A\mathbf{x} \geq 0$, usually denoted as $A \succeq 0$.

Rank

- **Rank**: the dimension of the vector space spanned by its columns, or the maximal number of linearly independent columns.

Example 5.

$$A = \begin{bmatrix} 1 & 2 & 1 \\ -2 & -3 & 1 \\ 3 & 5 & 0 \end{bmatrix} \xrightarrow{2R_1+R_2\to R_2} \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 3 \\ 3 & 5 & 0 \end{bmatrix} \xrightarrow{-3R_1+R_3\to R_3} \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 3 \\ 0 & -1 & -3 \end{bmatrix}$$

$$\xrightarrow{R_2+R_3\to R_3} \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 3 \\ 0 & 0 & 0 \end{bmatrix} \xrightarrow{-2R_2+R_1\to R_1} \begin{bmatrix} 1 & 0 & -5 \\ 0 & 1 & 3 \\ 0 & 0 & 0 \end{bmatrix}.$$

The rank of matrix A is 2.

Inner Product

- Vector Space: consider $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, then

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y} = \sum_{i=1}^{d} x_i y_i$$

- Matrix Space: consider $A, B \in \mathbb{R}^{m \times n}$, then

$$\langle A, B \rangle = \mathrm{Tr}\left(A^\top B\right) = \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} B_{ij}$$

Norm

- Typically used vector norms.

  - ℓ_1-norm:

$$\|\mathbf{x}\|_1 = |x_1| + \cdots + |x_d|$$

  - ℓ_2-norm:

$$\|\mathbf{x}\|_2 = \left(\mathbf{x}^\top \mathbf{x}\right)^{1/2} = \sqrt{x_1^2 + \cdots + x_d^2}$$

or called *Euclidean norm*

  - ℓ_∞-norm:

$$\|\mathbf{x}\|_\infty = \max\left\{|x_1|, \ldots, |x_d|\right\}$$

Norm

- Typically used vector norms.

  - General ℓ_p-norm:

$$\|\mathbf{x}\|_p = \left(|x_1|^p + \cdots + |x_d|^p\right)^{1/p}$$

  - Quadratic norm:

$$\|\mathbf{x}\|_A = \sqrt{\mathbf{x}^\top A\mathbf{x}}, \text{ where } A \text{ is positive semi-definite.}$$

Dual Norm

Let $\|\cdot\|$ be a vector norm on $\mathbb{R}^d$. The associated dual norm $\|\cdot\|_*$ is defined as

$$\|\mathbf{y}\|_* = \sup\left\{\mathbf{y}^\top \mathbf{x} \mid \|\mathbf{x}\| \leq 1\right\}.$$

Proposition 1. The dual of ℓ_p-norm is the ℓ_q-norm with $\frac{1}{p} + \frac{1}{q} = 1$.

e.g., the dual of ℓ_2-norm is still ℓ_2-norm, the dual of ℓ_1-norm is ℓ_∞-norm

Proposition 2. Hölder's inequality: $\langle \mathbf{x}, \mathbf{y} \rangle \leq \|\mathbf{x}\| \cdot \|\mathbf{y}\|_*$.

Norm Relationship

Qualitative:

Lemma 1 (Mathematical Equivalence of Norms). *Suppose that* $\|\cdot\|_a$ *and* $\|\cdot\|_b$ *are norms on* $\mathbb{R}^d$*, there exist positive "constants"* α *and* β*, for all* $\mathbf{x} \in \mathbb{R}^d$*, such that*

$$\alpha\|\mathbf{x}\|_a \leq \|\mathbf{x}\|_b \leq \beta\|\mathbf{x}\|_a.$$

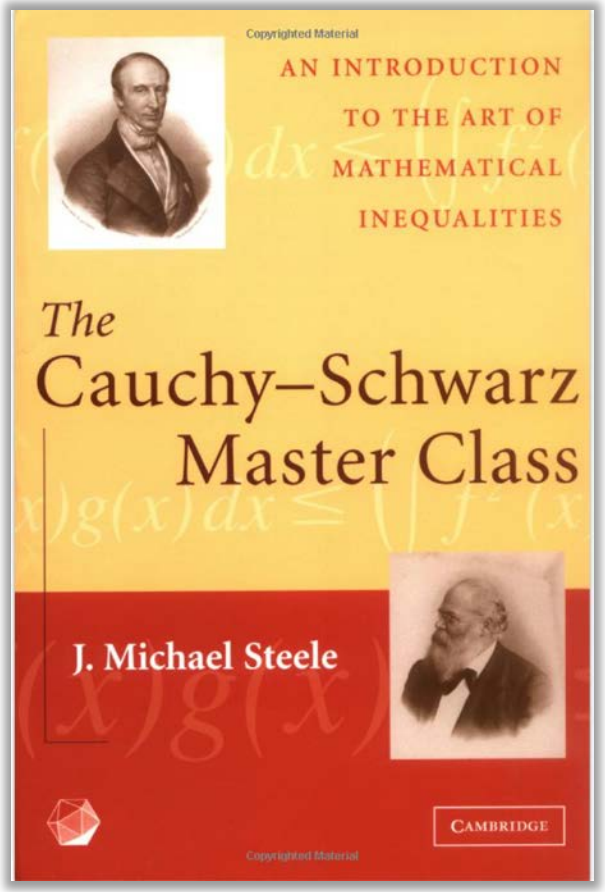
Notice: constants may depend on dimension!

For example: for any $\mathbf{x} \in \mathbb{R}^d$, the following inequalities hold:

- $\frac{1}{d}\|\mathbf{x}\|_1 \leq \|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_1$
- $\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2 \leq \sqrt{d}\|\mathbf{x}\|_\infty$

Cauchy-Schwarz Inequality

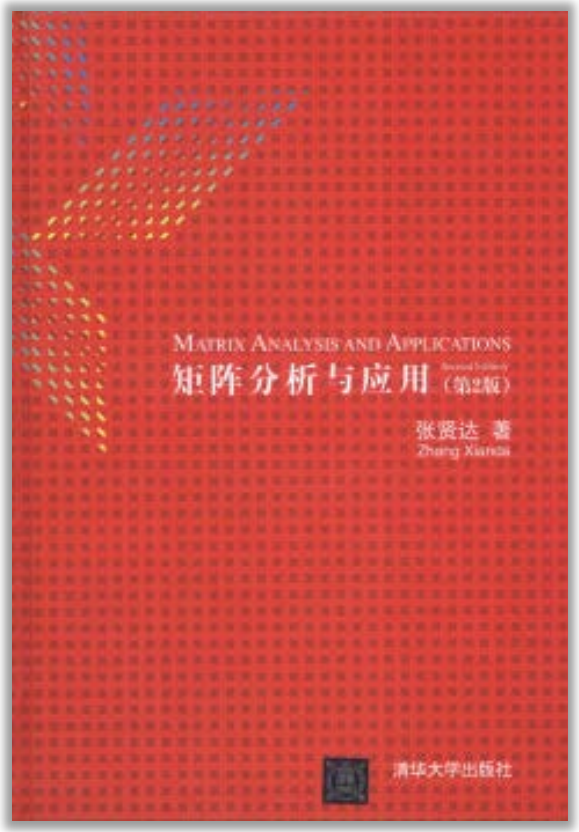
- $\langle \mathbf{x}, \mathbf{y} \rangle \leq \|\mathbf{x}\| \cdot \|\mathbf{y}\|_*$
- $\left(\sum_{i=1}^{n} a_i b_i \right)^2 \leq \left(\sum_{i=1}^{n} a_i^2 \right) \cdot \left(\sum_{i=1}^{n} b_i^2 \right)$
- $\left(\int_a^b f(x) g(x) \mathrm{d}x \right)^2 \leq \left(\int_a^b f^2(x) \mathrm{d}x \right) \cdot \left(\int_a^b g^2(x) \mathrm{d}x \right)$



Matrix Norm

Three different versions:

- operator norm
- entrywise norm
- Schatten norm



矩阵分析与应用. 张贤达

related pages can be found in readings of the course web

Matrix Operator Norm

- Consider a matrix $A \in \mathbb{R}^{m \times n}$.

We define its ***operator norm*** based on the aforementioned ***vector norm***.

Definition 6 (Matrix Operator Norm). The operator norm (or called induced norm) of a matrix $A \in \mathbb{R}^{m \times n}$ is defined by

$$\|A\|_{\mathrm{op},p} \triangleq \max \left\{ \frac{\|A\mathbf{x}\|_p}{\|\mathbf{x}\|_p} \,\middle|\, \mathbf{x} \in \mathbb{R}^d, \mathbf{x} \neq \mathbf{0} \right\}.$$

the norm in the right-hand side is defined over the ***vector space***.

Matrix Operator Norm

- Consider a matrix $A \in \mathbb{R}^{m \times n}$

  - ℓ_1-norm (max-column-sum norm):

$$\|A\|_{\text{op},1} = \max_{j \in [n]} \sum_{i=1}^{m} |A_{ij}|$$

  - ℓ_∞-norm (max-row-sum norm):

$$\|A\|_{\text{op},\infty} = \max_{i \in [m]} \sum_{j=1}^{n} |A_{ij}|$$

Matrix Operator Norm

- Consider a matrix $A \in \mathbb{R}^{m \times n}$

  - ℓ_2-norm (spectral norm):

$$\|A\|_{\mathrm{op},2} = \max_{i \in [r]} |\sigma_i|$$

where $A = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$, namely, σ_i is the i-th singular value.

Matrix Entrywise Norm

- Consider a matrix $A \in \mathbb{R}^{m \times n}$

The entrywise norm is defined by ***treating matrices as vectors***.

Definition 7 (Matrix Entrywise Norm). The entrywise norm of a matrix $A \in \mathbb{R}^{m \times n}$ is defined by

$$\|A\|_{\text{en},p} \triangleq \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |A_{ij}|^p \right)^{1/p}.$$

Matrix Entrywise Norm

- Consider a matrix $A \in \mathbb{R}^{m \times n}$

  - ℓ_1-norm (sum norm):

$$\|A\|_{\text{en},1} = \sum_{i=1}^{m} \sum_{j=1}^{n} |A_{ij}|$$

  - Frobenius-norm:

$$\|A\|_{\text{F}} = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij}^2}$$

Matrix Entrywise Norm

- Consider a matrix $A \in \mathbb{R}^{m \times n}$

  - ℓ_∞-norm (max norm):

$$\|A\|_{\text{en},\infty} = \max_{i \in [m]} \max_{j \in [n]} |A_{ij}|$$

Eigen Value Decomposition

Let A be an $d \times d$ PSD matrix, then it can be factored as

$$A = Q \Lambda Q^\top,$$

where (a) $Q = (\mathbf{v}_1, \ldots, \mathbf{v}_d) \in \mathbb{R}^{d \times d}$ is orthogonal, i.e., $Q^\top Q = I$ and $\mathbf{v}_1, \ldots, \mathbf{v}_d$ are eigenvectors; and (b) $\Lambda = \operatorname{diag}(\lambda_1, \cdots, \lambda_d)$ and $\lambda_1, \cdots, \lambda_d$ are eigenvalues.

Some concerned terms can be expressed by eigenvalues:

- $A = \sum_{i=1}^d \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$
- $\det(A) = \prod_{i=1}^d \lambda_i$
- $\operatorname{Tr}(A) = \sum_{i=1}^d \lambda_i$
- $\|A\|_{\mathrm{op},2} = \max_{i \in [d]} |\lambda_i|$
- $\|A\|_{\mathrm{F}} = \sqrt{\sum_{i=1}^d \lambda_i^2}$

Singular Value Decomposition

Suppose $A \in \mathbb{R}^{m\times n}$ has rank r, then it can be factored as

$$A = U\Sigma V^\top,$$

where (a) $U = (\mathbf{u}_1, \ldots, \mathbf{u}_r) \in \mathbb{R}^{m\times r}$ satisfies $U^\top U = I, V = (\mathbf{v}_1, \ldots, \mathbf{v}_r) \in \mathbb{R}^{n\times r}$ satisfies $V^\top V = I$; and (b) $\Sigma = \operatorname{diag}(\sigma_1, \cdots, \sigma_r)$ and $\sigma_1, \cdots, \sigma_r$ are sigular values.

Some concerned terms can be expressed by sigular values:

- $A = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$
- $\|A\|_{\mathrm{op},2} = \max_{i\in[r]} |\sigma_i|$
- $\|A\|_{\mathrm{F}} = \sqrt{\sum_{i=1}^{r} \sigma_i^2}$

Schatten Norm

- Consider a matrix $A \in \mathbb{R}^{m \times n}$

The Schatten norm is defined via the ***sigular values***.

Definition 8 (Matrix Schatten Norm). The Schatten norm of a matrix $A \in \mathbb{R}^{m \times n}$ with rank r is defined by

$$\|A\|_{\mathrm{Sc},p} \triangleq \begin{cases} \left(\sum_{i=1}^{r} \sigma_i^p \right)^{1/p}, & \text{for } 1 \leq p < \infty \\ \max_{i \in [r]} |\sigma_i|, & \text{for } \quad p = \infty \end{cases},$$

where $\sigma_1, \cdots, \sigma_r$ are the singular values of A.

Part 3. Probability and Statistics

- Expectation and Variance
- Conditional Expectation
- Concentration Inequalities

Expectation and Variance

Expectation

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x \Pr[X = x]$$

Linearity of expecation: $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$.

Variance

$$\mathrm{Var}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$$

- $\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$
- $\mathrm{Var}[aX] = a^2 \mathrm{Var}[X]$
- $\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$

Cauchy-Schwarz Inequality in Probability

- $\langle \mathbf{x}, \mathbf{y} \rangle \leq \|\mathbf{x}\| \cdot \|\mathbf{y}\|_*$
- $\left(\sum_{i=1}^{n} a_i b_i \right)^2 \leq \left(\sum_{i=1}^{n} a_i^2 \right) \cdot \left(\sum_{i=1}^{n} b_i^2 \right)$
- $\left(\int_a^b f(x) g(x) \mathrm{d}x \right)^2 \leq \left(\int_a^b f^2(x) \mathrm{d}x \right) \cdot \left(\int_a^b g^2(x) \mathrm{d}x \right)$
- $(\mathbb{E}[XY])^2 \leq \mathbb{E}[X^2] \cdot \mathbb{E}[Y^2]$

Conditional Expectation

Conditional Expectation

$$\mathbb{E}[X|Y=y] = \sum_{x \in \mathcal{X}} x \Pr[X=x|Y=y]$$

Theorem 1 (Double Expectation Theorem). *Let X, Y be arbitrary random variables. Suppose $\mathbb{E}[X], \mathbb{E}[Y], \mathbb{E}[X|Y], \mathbb{E}[Y|X]$ all exist, then it holds that*

$$\mathbb{E}[X] = \mathbb{E}_Y[\mathbb{E}_X[X|Y]], \ \mathbb{E}[Y] = \mathbb{E}_X[\mathbb{E}_Y[Y|X]].$$

$\mathbb{E}[X] = \mathbb{E}_Y[\mathbb{E}_X[X|Y]]$ means that to measure the expectation of X, we can first measure the expectation of X ***given the information of Y***, then measure the expectation of Y.

Concentration Inequalities

Theorem 2 (Markov's Inequality). *Let X be a non-negative random variable with $\mathbb{E}[X] < \infty$, then for all $\epsilon > 0$,*

$$\Pr[X \geq \epsilon \mathbb{E}[X]] \leq \frac{1}{\epsilon}.$$

A simple proof:

$$\begin{aligned}
\Pr[X \geq t\mathbb{E}[X]] &= \sum_{x \geq t\mathbb{E}[X]} \Pr[X = x] \\
&\leq \sum_{x \geq t\mathbb{E}[X]} \Pr[X = x] \frac{x}{t\mathbb{E}[X]} \qquad \left(\text{using } \frac{x}{t\mathbb{E}[X]} \geq 1\right) \\
&\leq \sum_{x} \Pr[X = x] \frac{x}{t\mathbb{E}[X]} \qquad \text{(extending non-negative sum)} \\
&= \mathbb{E}\left[\frac{X}{t\mathbb{E}[X]}\right] = \frac{1}{t} \qquad \text{(linearity of expectation).}
\end{aligned}$$

Concentration Inequalities

Theorem 3 (Chebyshev's Inequality). *Let X be a non-negative random variable with $\mathbb{E}[X], \mathrm{Var}[X] < \infty$, then for all $\epsilon > 0$,*

$$\Pr\left[|X - \mathbb{E}[X]| \geq \epsilon\right] \leq \frac{\mathrm{Var}[X]}{\epsilon^2}.$$

Chebyshev's inequality can be immediately obtained from Markov's inequality.

Theorem 4 (Hoeffding's Inequality). *Let $X_1, \ldots, X_m$ be **independent** random variables with X_i taking values in $[a_i, b_i]$ for all $i \in [m]$. Then, for any $\epsilon > 0$, the following inequalities hold for $S_m = \sum_{i=1}^m X_i$,*

$$\Pr\left[S_m - \mathbb{E}\left[S_m\right] \geq \epsilon\right] \leq e^{-2\epsilon^2 / \sum_{i=1}^m (b_i - a_i)^2},$$
$$\Pr\left[S_m - \mathbb{E}\left[S_m\right] \leq -\epsilon\right] \leq e^{-2\epsilon^2 / \sum_{i=1}^m (b_i - a_i)^2}.$$

Part 4. Information Theory

- Entropy
- Conditional Entropy
- KL divergence
- Bregman Divergence

Entropy

- Entropy measures the uncertainty, which is the most basic concept in information theory.

Definition 9 (Entropy). The entropy of a discrete random variable X with probability mass function $\boldsymbol{p}(x) = \mathbb{P}[X = x]$ is denoted by $H(X)$:

$$H(X) = -\sum_{x \in X} \boldsymbol{p}(x) \log(\boldsymbol{p}(x)).$$

An explanation of entropy: $1/\boldsymbol{p}(x)$ stand for the information contained in $\boldsymbol{p}(x)$, $\log_2(1/\boldsymbol{p}(x))$ is the code length needed to encode the information, then entropy $H(X)$ measures the ***expected code length*** to encode a distribution $\boldsymbol{p}$.

Conditional Entropy

Definition 10 (Conditional Entropy)**.**

$$H(Y|X) = -\sum_{x\in\mathcal{X}, y\in\mathcal{Y}} \boldsymbol{p}(x,y) \log \frac{\boldsymbol{p}(x,y)}{\boldsymbol{p}(x)}$$

Conditional entropy $H(Y|X)$ measures the uncertainty of Y *given the uncertainty of* X.

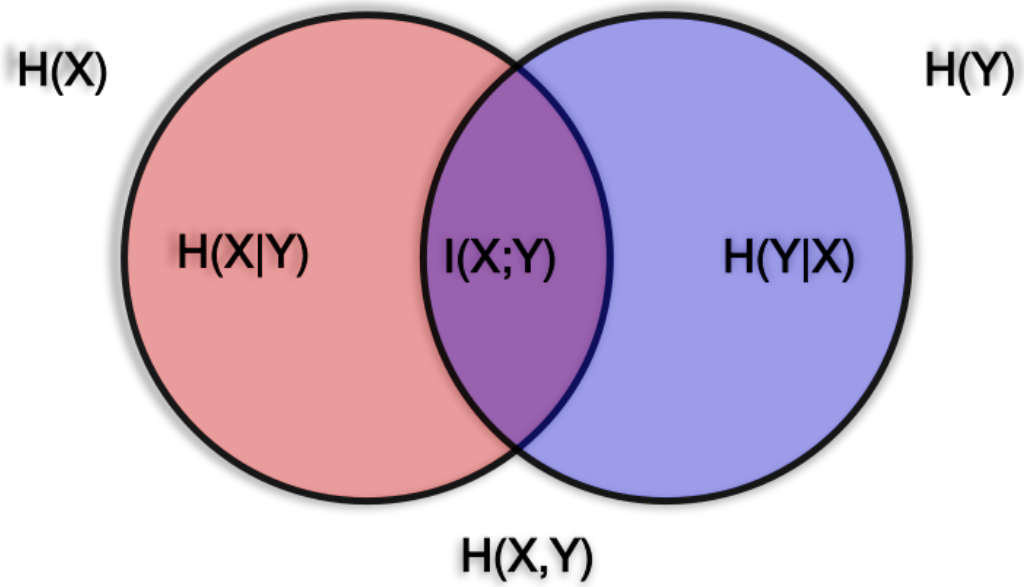
Mutual Information

Definition 11 (Mutual Information)**.**

$$I(X,Y) = \mathrm{KL}(\boldsymbol{p}(x,y) \| \boldsymbol{p}(x)\boldsymbol{p}(y)) = \sum_{x \in X, y \in y} \boldsymbol{p}(x,y) \log \left[\frac{\boldsymbol{p}(x,y)}{\boldsymbol{p}(x)\boldsymbol{p}(y)} \right],$$

with the conventions $0 \log 0 = 0, 0 \log \frac{0}{0} = 0$, and $a \log \frac{a}{0} = +\infty$ for $a > 0$.

Relationship



$$I(X,Y) = H(X) - H(X|Y)$$

$$I(X,Y) = H(Y) - H(Y|X)$$

KL Divergence (Relative Entropy)

Definition 12 (KL Divergence)**.** The Kullback-Leibler (KL) divergence (relative entropy) of two distributions $\boldsymbol{p}$ and $\boldsymbol{q}$ is defined by $\mathrm{KL}(\boldsymbol{p}\|\boldsymbol{q})$:

$$\mathrm{KL}(\boldsymbol{p}\|\boldsymbol{q}) = \sum_{x\in\mathcal{X}} \boldsymbol{p}(x) \log\left[\frac{\boldsymbol{p}(x)}{\boldsymbol{q}(x)}\right]$$

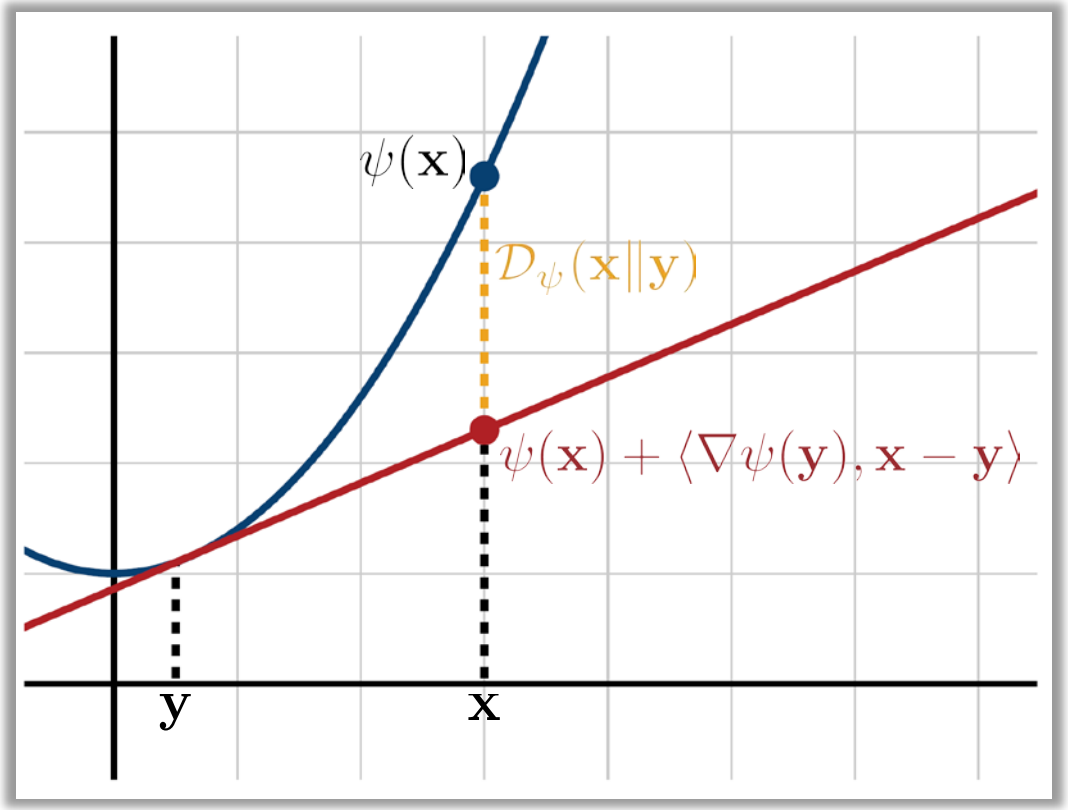
with the conventions $0\log 0 = 0, 0\log\frac{0}{0} = 0$, and $a\log\frac{a}{0} = +\infty$ for $a > 0$.

Proposition 1.

- *KL divergence is always non-negative;*
- *Pinsker's inequality:* $KL(\boldsymbol{p}\|\boldsymbol{q}) \geq \frac{1}{2}\|\boldsymbol{p}-\boldsymbol{q}\|_1^2$.

Bregman Divergence

Bregman divergence measures the difference of a function and its *linear approximation*



Definition 13 (Bregman Divergence). Let ψ be a convex and differentiable function over a convex set $\mathcal{K}$, then for any $\mathbf{x}, \mathbf{y} \in \mathcal{K}$, the bregman divergence $\mathcal{D}_\psi$ associated to ψ is defined as

$$\mathcal{D}_\psi(\mathbf{x}\|\mathbf{y}) = \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle.$$

Bregman Divergence

Definition 13 (Bregman Divergence). Let ψ be a convex and differentiable function over a convex set $\mathcal{K}$, then for any $\mathbf{x}, \mathbf{y} \in \mathcal{K}$, the bregman divergence $\mathcal{D}_\psi$ associated to ψ is defined as

$$\mathcal{D}_\psi(\mathbf{x}\|\mathbf{y}) = \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle.$$

Table 1: Choice of $\psi(\cdot)$ and the corresponding Bregman divergence.

	$\psi(\mathbf{x})$	$\mathcal{D}_\psi(\mathbf{x}\|\mathbf{y})$
Squared L_2-distance	$\|\mathbf{x}\|_2^2$	$\|\mathbf{x} - \mathbf{y}\|_2^2$
Mahalanobis distance	$\|\mathbf{x}\|_Q^2$	$\|\mathbf{x} - \mathbf{y}\|_Q^2$
negative entropy	$\sum_i x_i \log x_i$	$\mathrm{KL}(\mathbf{x}\|\mathbf{y})$

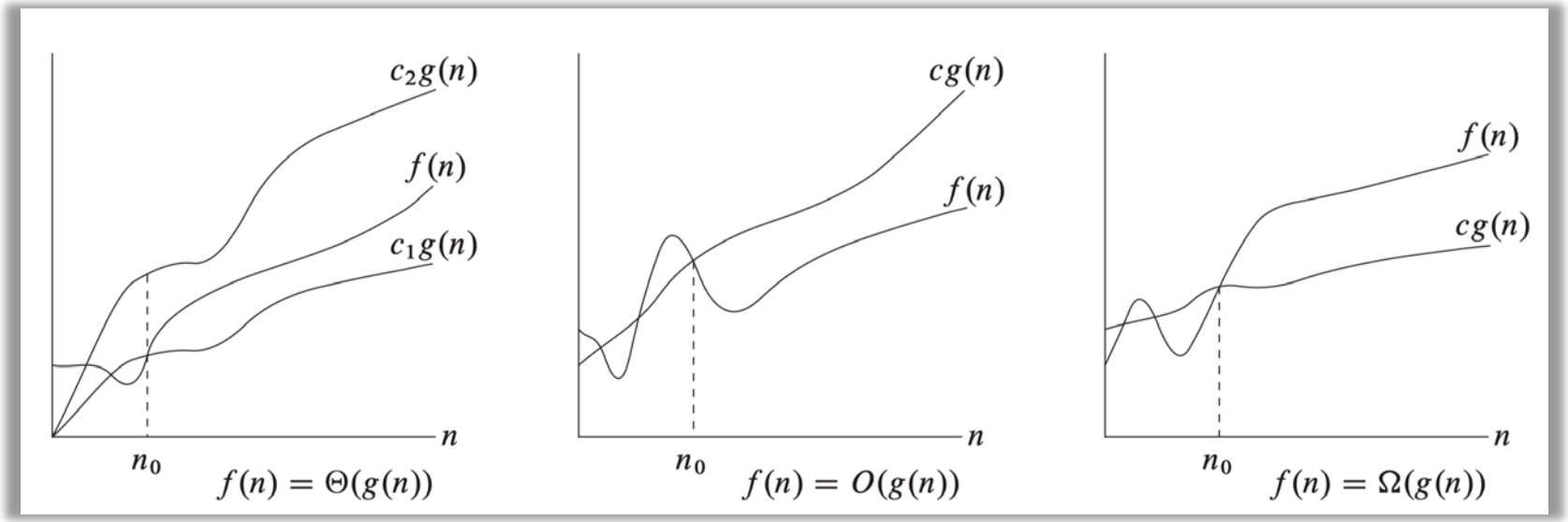
Part 5. Asymptotic Complexity

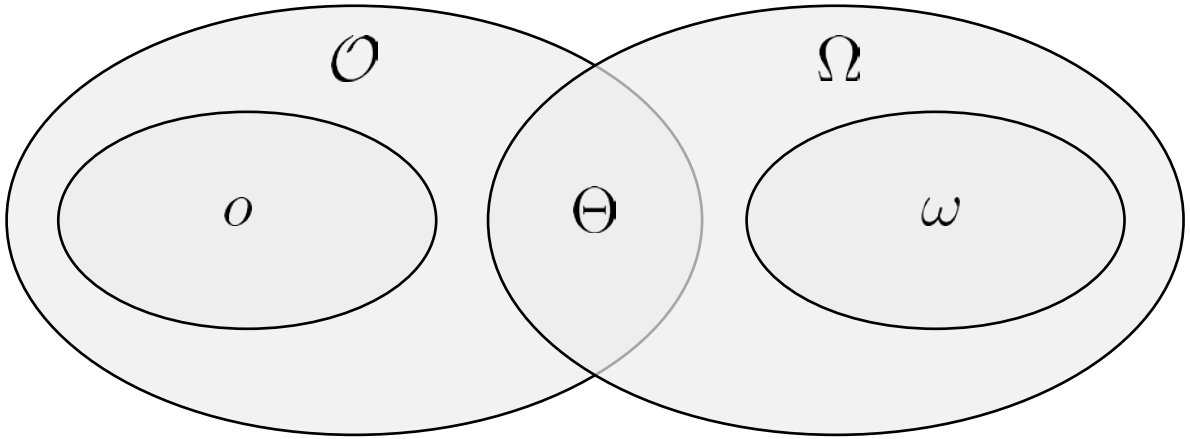
- Definition
- Illustration
- Examples

Definitions

- $\Theta(g(n)) = \{f(n) \mid$ there exist positive constants c_1, c_2, and n_0 such that $0 \leq c_1 g(n) \leq f(n) \leq c_2 g(n)$ for all $n \geq n_0\}$.
- $\mathcal{O}(g(n)) = \{f(n) \mid$ there exist positive constants c and n_0 such that $0 \leq f(n) \leq cg(n)$ for all $n \geq n_0\}$.
- $\Omega(g(n)) = \{f(n) \mid$ there exist positive constants c and n_0 such that $0 \leq cg(n) \leq f(n)$ for all $n \geq n_0\}$.
- $o(g(n)) = \{f(n) \mid$ for any positive constant $c > 0$, there exists a constant $n_0 > 0$ such that $0 \leq f(n) < cg(n)$ for all $n \geq n_0\}$.
- $\omega(g(n)) = \{f(n) \mid$ for any positive constant $c > 0$, there exists a constant $n_0 > 0$ such that $0 \leq cg(n) < f(n)$ for all $n \geq n_0\}$.

Illustration





Examples

- $3n^3 + 2n^2 + n + \log n = \Theta(n^3)$
- $\mathcal{O}(1) < \mathcal{O}(\log n) < \mathcal{O}(n) < \mathcal{O}(n \log n) < \mathcal{O}\left(n^2\right) < \mathcal{O}\left(2^n\right) < \mathcal{O}(n!)$

Part 6. Optimization in Machine Learning

- Supervised Learning
- Empirical Risk Minimization
- Structural Risk Minimization
- Examples

Learning by Optimization

The fundamental goal of (supervised) learning: ***Risk Minimization (RM)***,

$$\min_{h \in \mathcal{H}} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[f(h(\mathbf{x}), y)],$$

where

- h denotes the hypothesis (model) from the hypothesis space $\mathcal{H}$.
- $(\mathbf{x}, y)$ is an instance chosen from a unknown distribution $\mathcal{D}$.
- $f(h(\mathbf{x}), y)$ denotes the loss of using hypothesis h on the instance $(\mathbf{x}, y)$.

Empirical Risk Minimization

Since the distribution of the data, i.e., $\mathcal{D}$, is unavailable to the learner, the risk is not computable.

In practice, the learner instead tries to optimize the following empirical risk, which is called ***empirical risk minimization (ERM)***:

$$\min_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} f\left(h(\mathbf{x}_i), y_i\right).$$

ERM approximates RM: All instances are i.i.d. sampled from the same distribution.

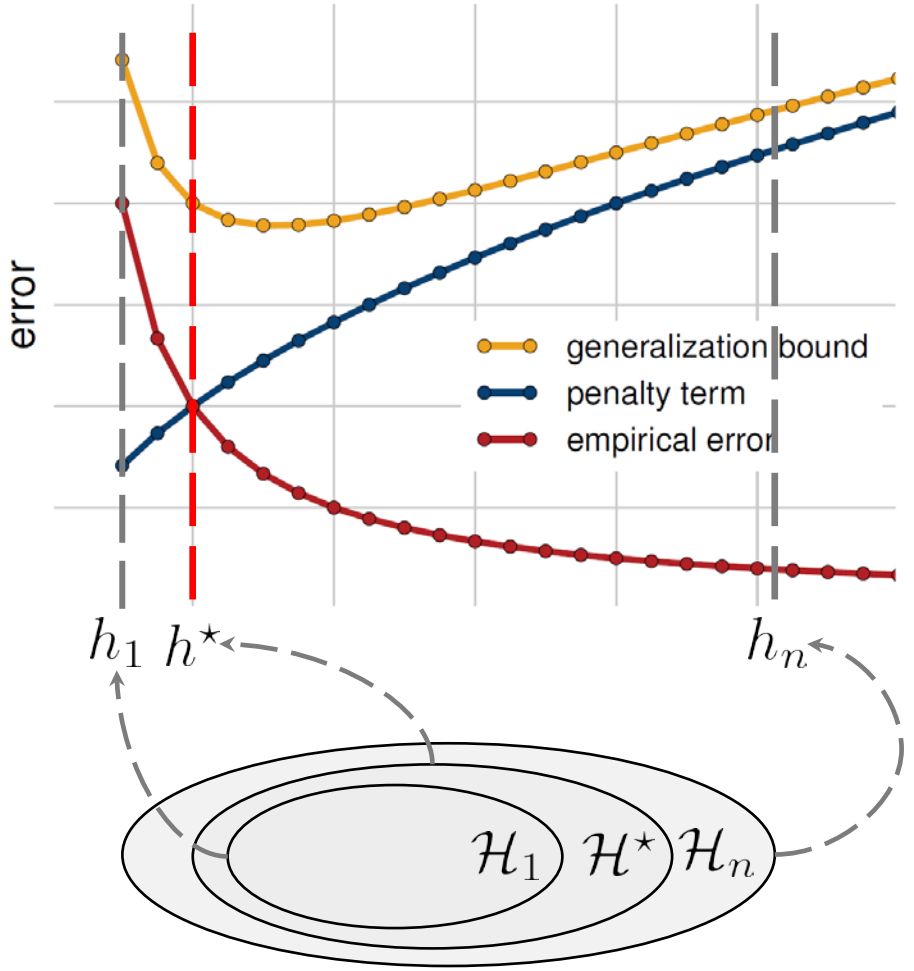
IID assumption: *Independent and Identically Distributed random variables*

Structural ERM

In practice, we often explicitly control the complexity of the learner by adding a regularization term in the optimization objective, i.e.,

$$\min_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} f(h(\mathbf{x}_i), y_i) + \lambda \mathcal{R}(h).$$

This is called ***Structural ERM***.



Examples

- Consider the following binary classification task with (i) linear hypothesis $h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$; and (ii) $\mathbf{x}_i \in \mathbb{R}^d$, $y_i \in \{-1, +1\}$ for all $i \in [m]$.

Example 6. Taking $f(h(\mathbf{x}_i), y_i) = \max\{0, 1 - y_i \mathbf{w}^\top \mathbf{x}_i\}$ (hinge loss) and $\mathcal{R}(h) = \|\mathbf{w}\|_2^2$ forms the optimization objective in ***Support Vector Machine (SVM)***:

$$\min_{\mathbf{w} \in \mathbb{R}^d} \sum_{i=1}^{n} \max\{0, 1 - y_i \mathbf{w}^\top \mathbf{x}_i\} + \lambda \|\mathbf{w}\|_2^2 .$$

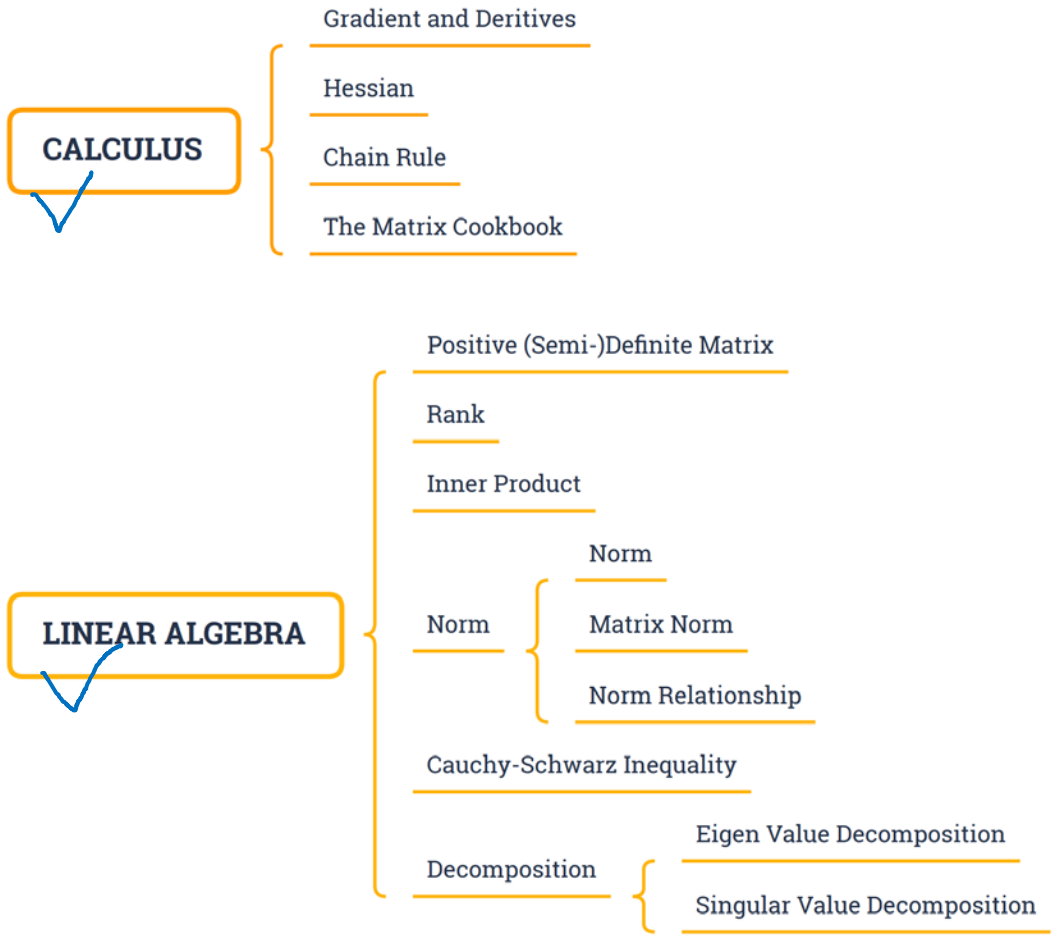
Example

• Consider the following binary classification task with (i) linear hypothesis $h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$; and (ii) $\mathbf{x}_i \in \mathbb{R}^d$, $y_i \in \{-1, +1\}$ for all $i \in [m]$.

Example 7. Taking $f(h(\mathbf{x}_i), y_i) = \log(1 + \exp(-y_i \mathbf{w}^\top \mathbf{x}_i))$ and $\mathcal{R}(h) = \|\mathbf{w}\|_2^2$ forms the optimization objective in ***Logistic Regression (LR)***:

$$\min_{\mathbf{w} \in \mathbb{R}^d} \sum_{i=1}^{n} \log(1 + \exp(-y_i \mathbf{w}^\top \mathbf{x}_i)) + \lambda \|\mathbf{w}\|_2^2 .$$

Summary


CALCULUS
Gradient and Deritives
Hessian
Chain Rule
The Matrix Cookbook
LINEAR ALGEBRA
Positive (Semi-)Definite Matrix
Rank
Inner Product
Norm
Norm
Norm
Matrix Norm
Norm Relationship
Cauchy-Schwarz Inequality
Decomposition
Eigen Value Decomposition
Singular Value Decomposition
PROBABILITY & STATISTICS
Expectation and Variance
Cauchy-Schwarz Inequality in Probability
Conditional Expectation
Concentration Inequalities
INFORMATION THEORY
Entropy
Conditional Entropy
Mutual Information
KL Divergence (Relative Entropy)
Bregman Divergence
OPTIMIZATION IN MACHINE LEARNING
Risk Minimization
Empirical Risk Minimization (ERM)
Structural ERM

Q & A