Lecture 1. Preliminaries

Advanced Optimization (Fall 2025)

Peng Zhao

zhaop@lamda.nju.edu.cn

Nanjing University

Outline

- Math Background
  - Calculus, Linear Algebra
  - Probability & Statistics
  - Information Theory, Asymptotic Notations
- Convex Optimization Basics
  - ML as Optimization
  - Convex Function, Convex Set
  - Convex Optimization Problem

Outline

- Math Background
  - Calculus, Linear Algebra
  - Probability & Statistics
  - Information Theory, Asymptotic Notations
- Convex Optimization Basics
  - ML as Optimization
  - Convex Function, Convex Set
  - Convex Optimization Problem

Notational Convention

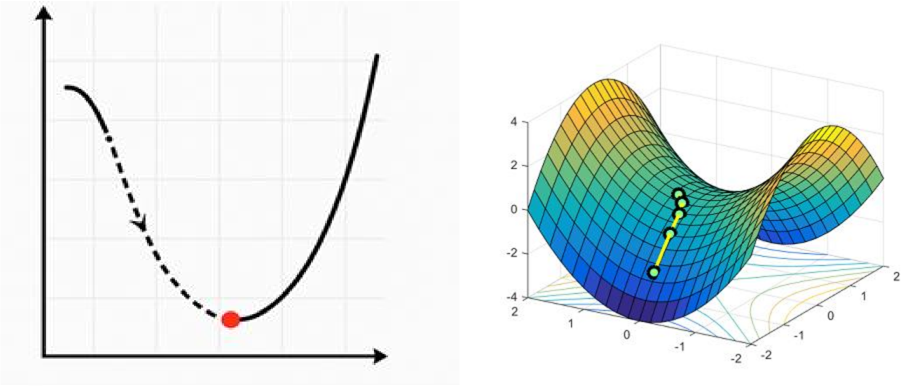
- $[n] = \{1, \ldots, n\}$
- $\mathbf{x}, \mathbf{y}, \mathbf{v}$: vectors
- A, B: matrices
- $\mathcal{X}, \mathcal{Y}, \mathcal{K}$: domain
- d, m, n: dimensions
- I: identity matrix
- X, Y: random variables
- $\boldsymbol{p}, \boldsymbol{q}$: probability distributions

Function

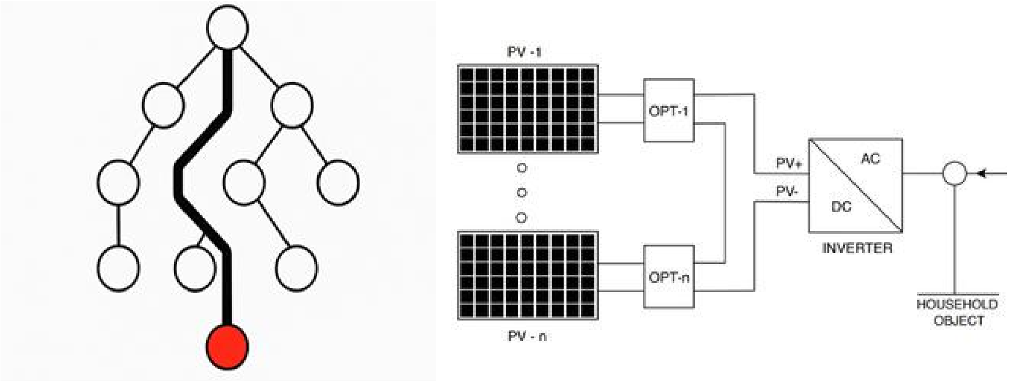
- Function mapping $f : \mathrm{dom}\, f \subseteq \mathcal{X} \subseteq \mathbb{R}^n \to \mathcal{Y} \subseteq \mathbb{R}^m$

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

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



Continuous Optimization



Discrete Optimization

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

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

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

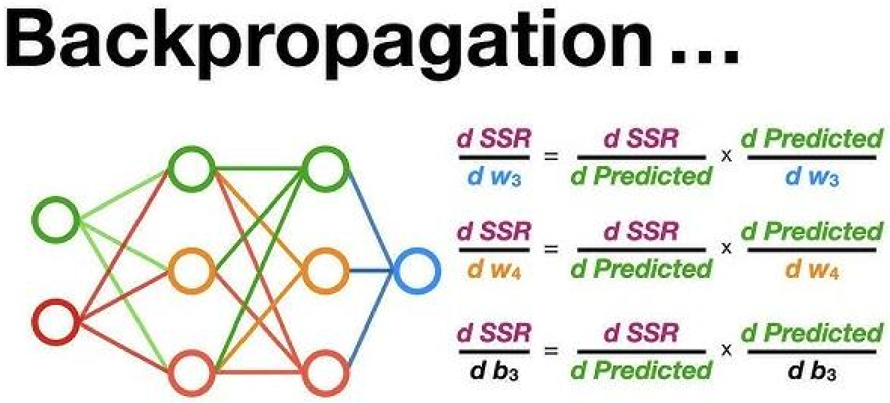
Example 4. The Hessian of $f(\mathbf{x}) = x_1^3 x_2^2 - 3x_1 x_2^3 + 1$ is $\nabla^2 f(\mathbf{x}) = \begin{bmatrix} 6x_1x_2^2 & 6x_1^2x_2 - 9x_2^2 \\ 6x_1^2x_2 - 9x_2^2 & 2x_1^3 - 18x_1x_2 \end{bmatrix}$.

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 $h''(x) = f''(g(x))(g'(x))^2 + f'(g(x))g''(x)$.



Src: https://www.youtube.com/watch?v=iyn2zdALii8

Reference: The Matrix Cookbook

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

2.4.1 First Order

$$\frac{\partial \mathbf{x}^T\mathbf{a}}{\partial \mathbf{x}} = \frac{\partial \mathbf{a}^T\mathbf{x}}{\partial \mathbf{x}} = \mathbf{a} \quad (69)$$

$$\frac{\partial \mathbf{a}^T\mathbf{X}\mathbf{b}}{\partial \mathbf{X}} = \mathbf{a}\mathbf{b}^T \quad (70)$$

$$\frac{\partial \mathbf{a}^T\mathbf{X}^T\mathbf{b}}{\partial \mathbf{X}} = \mathbf{b}\mathbf{a}^T \quad (71)$$

$$\frac{\partial \mathbf{a}^T\mathbf{X}\mathbf{a}}{\partial \mathbf{X}} = \frac{\partial \mathbf{a}^T\mathbf{X}^T\mathbf{a}}{\partial \mathbf{X}} = \mathbf{a}\mathbf{a}^T \quad (72)$$

$$\frac{\partial \mathbf{X}}{\partial X_{ij}} = \mathbf{J}^{ij} \quad (73)$$

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

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

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

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

$$\frac{\partial X_{kl}}{\partial X_{ij}} = \delta_{ik}\delta_{lj} \quad (32)$$

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]):

$$\begin{aligned}
\partial\mathbf{A} &= 0 \qquad (\mathbf{A} \text{ is a constant}) && (33)\\
\partial(\alpha\mathbf{X}) &= \alpha\partial\mathbf{X} && (34)\\
\partial(\mathbf{X}+\mathbf{Y}) &= \partial\mathbf{X}+\partial\mathbf{Y} && (35)\\
\partial(\mathrm{Tr}(\mathbf{X})) &= \mathrm{Tr}(\partial\mathbf{X}) && (36)\\
\partial(\mathbf{X}\mathbf{Y}) &= (\partial\mathbf{X})\mathbf{Y}+\mathbf{X}(\partial\mathbf{Y}) && (37)\\
\partial(\mathbf{X}\circ\mathbf{Y}) &= (\partial\mathbf{X})\circ\mathbf{Y}+\mathbf{X}\circ(\partial\mathbf{Y}) && (38)\\
\partial(\mathbf{X}\otimes\mathbf{Y}) &= (\partial\mathbf{X})\otimes\mathbf{Y}+\mathbf{X}\otimes(\partial\mathbf{Y}) && (39)\\
\partial(\mathbf{X}^{-1}) &= -\mathbf{X}^{-1}(\partial\mathbf{X})\mathbf{X}^{-1} && (40)\\
\partial(\det(\mathbf{X})) &= \mathrm{Tr}(\mathrm{adj}(\mathbf{X})\partial\mathbf{X}) && (41)\\
\partial(\det(\mathbf{X})) &= \det(\mathbf{X})\mathrm{Tr}(\mathbf{X}^{-1}\partial\mathbf{X}) && (42)\\
\partial(\ln(\det(\mathbf{X}))) &= \mathrm{Tr}(\mathbf{X}^{-1}\partial\mathbf{X}) && (43)\\
\partial\mathbf{X}^T &= (\partial\mathbf{X})^T && (44)\\
\partial\mathbf{X}^H &= (\partial\mathbf{X})^H && (45)
\end{aligned}$$

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 $A \succeq 0$.

Especially useful for defining some distance metric

- $\|\mathbf{x} - \mathbf{y}\|_A$
- Sometimes the matrix should be "localized", like $\|\mathbf{x}_t - \mathbf{x}_*\|_{A_t}$

Rank

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

Example 5.

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

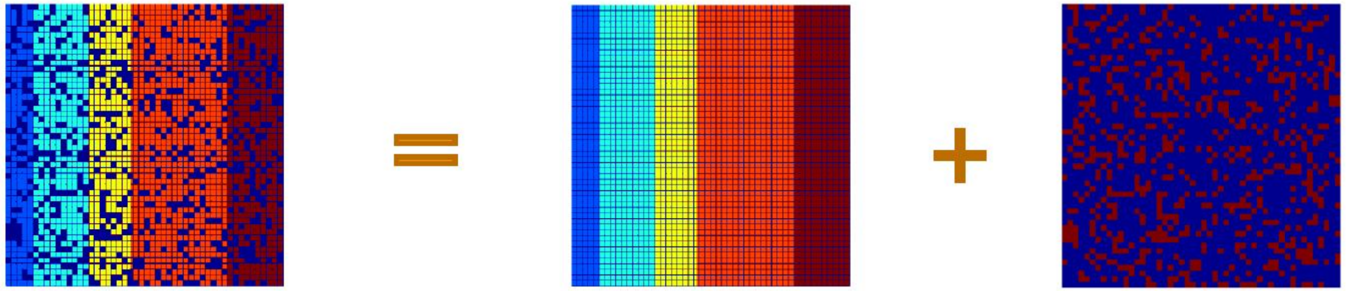
$$\xrightarrow{R_2+R_3\to R_3} \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 3 \\ 0 & 0 & 0 \end{bmatrix} \xrightarrow{-2R_2+R_1\to R_1} \begin{bmatrix} 1 & 0 & -5 \\ 0 & 1 & 3 \\ 0 & 0 & 0 \end{bmatrix}.$$

$\Rightarrow$ The rank of matrix A is 2.

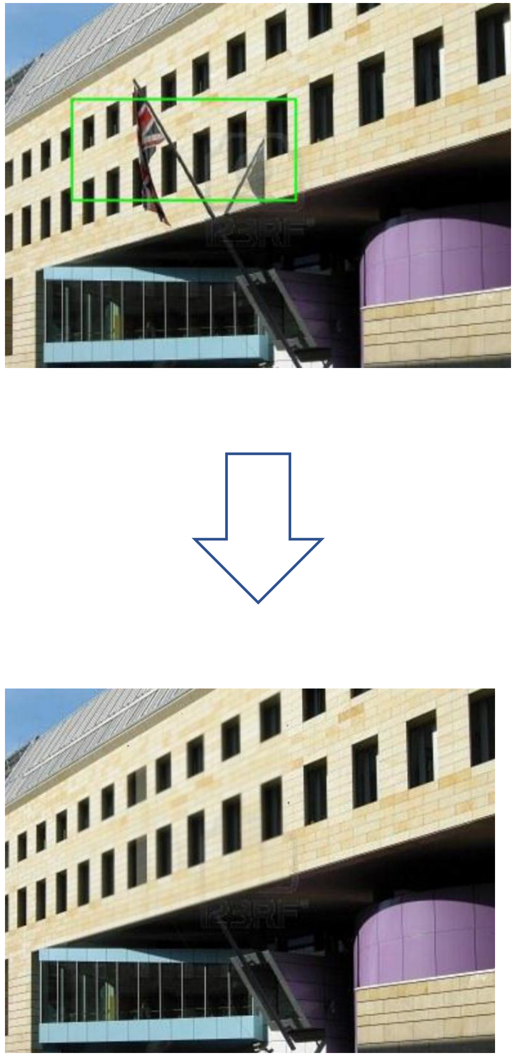
Low rank: Robust PCA

- Robust PCA formulation

$$\min_{\hat{X}} \|X - \hat{X}\|_1 + \|\hat{X}\|_*$$



input X *low rank* $\|\hat{X}\|_*$ *sparse* $\|X - \hat{X}\|_1$



Inner Product

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

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

Example in ML: linear regression, feature similarity calculation, ….

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

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

Example in ML: covariance matrix, PCA, LDA, matrix factorization….

Vector Norm

- The following norm can be induced based on inner product

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

usually called ℓ_2-norm, or Euclidean norm.

- ℓ_1-norm:

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

- ℓ_∞-norm:

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

- General ℓ_p-norm:

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

- Quadratic norm:

$$\|\mathbf{x}\|_A = \sqrt{\mathbf{x}^\top A \mathbf{x}},$$

where $A \in \mathbb{R}^{d \times d}$ is positive semi-definite.

Vector Inner Product vs Norm

- ***Norm***: tells you "how big" a vector is.
- ***Inner product***: tells you "how two vectors align" (geometry).

- Every inner product gives a norm, but not every norm comes from an inner product.

- Hilbert space = Banach space + geometry (orthogonality, projection).

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.

The dual of $\|\cdot\|_A$ is $\|\cdot\|_{A^{-1}}$

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!

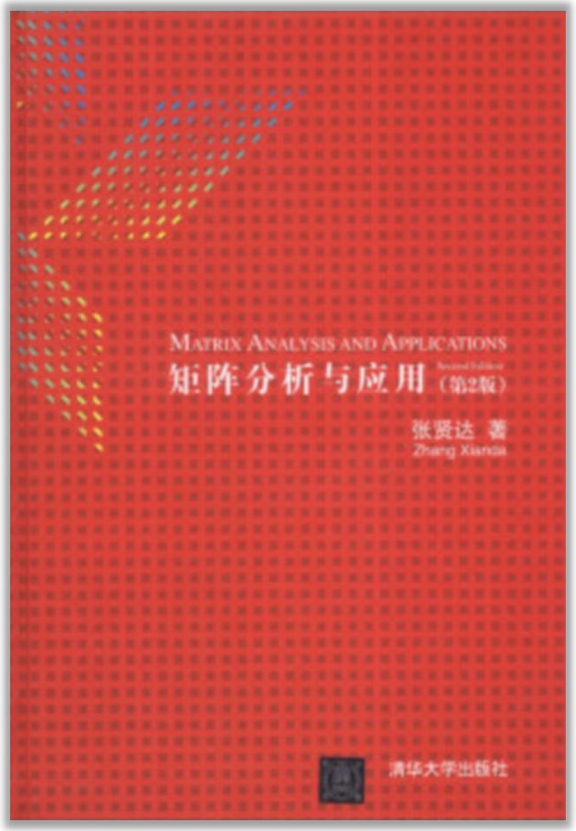
For example: for any $\mathbf{x} \in \mathbb{R}^d$, the following inequalities hold:

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

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):

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

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

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

Matrix Entrywise Norm

- Consider a matrix $A \in \mathbb{R}^{m \times n}$
  - ℓ_1-norm (sum norm):

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

  - Frobenius-norm:

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

  - ℓ_∞-norm (max norm):

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

Matrix Schatten Norm

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

The Schatten norm is defined via the ***singular 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.

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:

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

Part 3. Statistics, Information Theory

- Concentration Inequalities

- Entropy

- KL divergence

- Bregman Divergence

Concentration Inequalities

Theorem 2 (Markov's Inequality). *Let X be a non-negative random variable with $\mathbb{E}[X] < \infty$, then for all $t > 0$,*

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

Proof.

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

Concentration Inequalities

Theorem 2 (Markov's Inequality). *Let X be a non-negative random variable with $\mathbb{E}[X] < \infty$, then for all $t > 0$,*

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

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.

Concentration Inequalities

- If we have more information: random variables are sampled from the same distribution independently (***i.i.d.***) → better concentration

Theorem 4 (Hoeffding's Inequality). *Let $X_1, \ldots, X_m$ be* ***i.i.d.*** *random variables, and $X_i \in [a, b]$ for all $i \in [m]$. Let $S_m = \sum_{i=1}^m X_i$. Then, for any $\epsilon > 0$,*

$$\Pr\left[S_m - \mathbb{E}\left[S_m\right] \geq \epsilon\right] \leq \exp\left(-2\epsilon^2 \Big/ \left(m\left(b-a\right)^2\right)\right),$$

$$\Pr\left[S_m - \mathbb{E}\left[S_m\right] \leq -\epsilon\right] \leq \exp\left(-2\epsilon^2 \Big/ \left(m\left(b-a\right)^2\right)\right).$$

Consequently, we have

$$\Pr\left[\left|S_m - \mathbb{E}\left[S_m\right]\right| \geq \epsilon\right] \leq 2\exp\left(-2\epsilon^2 \Big/ \left(m\left(b-a\right)^2\right)\right).$$

Example

- Estimating the probability of heads in a (biased) coin toss.
  - ◇ Unknown probability: $\Pr[X_i = 1] = p, \quad \Pr[X_i = 0] = 1 - p.$
  - ◇ Estimator: $\widehat{p} \triangleq \frac{1}{m} \sum_{i=1}^{m} X_i.$
  - ◇ Property: $\mathbb{E}[\widehat{p}] = p, \quad \mathrm{Var}[\widehat{p}] = \frac{1}{m^2} \sum_{i=1}^{m} \mathrm{Var}[X_i] = \frac{p(1-p)}{m}$
  - ◇ To ensure:

$$\boxed{\Pr\left[|\widehat{p} - p| \geq \epsilon\right] \leq \delta} \qquad \text{e.g., } \delta = 0.001$$

⇨ *How many times do we need to toss the coin at a minimum?*

Example

- Estimating the probability of heads in a (biased) coin toss.
  - ◇ Unknown probability: $\Pr[X_i = 1] = p, \quad \Pr[X_i = 0] = 1 - p.$
  - ◇ Estimator: $\widehat{p} \triangleq \frac{1}{m}\sum_{i=1}^{m} X_i.$
  - ◇ Property: $\mathbb{E}[\widehat{p}] = p, \quad \mathrm{Var}[\widehat{p}] = \frac{1}{m^2}\sum_{i=1}^{m}\mathrm{Var}[X_i] = \frac{p(1-p)}{m}$

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

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

$$\Pr\left[|\widehat{p} - p| \geq \epsilon\right] \leq \frac{\mathrm{Var}[\widehat{p}]}{\epsilon^2} = \frac{p(1-p)}{m\epsilon^2} \leq \frac{1}{4m\epsilon^2} \leq \delta \quad \Rightarrow \quad m \geq \frac{1}{4\epsilon^2\delta} = \frac{\mathbf{250}}{\epsilon^2}$$

Example

- Estimating the probability of heads in a (biased) coin toss.
  - ◇ Unknown probability: $\Pr[X_i = 1] = p, \quad \Pr[X_i = 0] = 1 - p.$
  - ◇ Estimator: $\widehat{p} \triangleq \frac{1}{m}\sum_{i=1}^m X_i.$
  - ◇ Property: $\mathbb{E}[\widehat{p}] = p, \quad \mathrm{Var}[\widehat{p}] = \frac{1}{m^2}\sum_{i=1}^m \mathrm{Var}[X_i] = \frac{p(1-p)}{m}$

> **Theorem 4** (Hoeffding's Inequality). *Let $X_1, \ldots, X_m$ be* **i.i.d.** *random variables, and $X_i \in [a, b]$ for all $i \in [m]$. Let $S_m = \sum_{i=1}^m X_i$. Then, for any $\epsilon > 0$,*
>
> $$\Pr\left[\left|S_m - \mathbb{E}[S_m]\right| \geq \epsilon\right] \leq 2\exp\left(-2\epsilon^2 \Big/ \left(m(b-a)^2\right)\right).$$

$$\Pr[m|\widehat{p} - p| \geq m\epsilon] \leq 2\exp(-2m\epsilon^2) \leq \delta \quad \Rightarrow \quad m \geq \frac{1}{2\epsilon^2}\ln\frac{2}{\delta} \approx \frac{\mathbf{3.8}}{\epsilon^2}$$

Concentration Inequalities

$$\boxed{\Pr\left[|\widehat{p} - p| \geq \epsilon\right] \leq {\color{red}\delta}}$$

e.g., $\delta = 0.001$

- An example: Estimating the probability of heads in a coin toss.
  - ◇ Estimate: $\widehat{p} \triangleq \frac{1}{m}\sum_{i=1}^{m} X_i$, $\mathbb{E}[\widehat{p}] = p$, $\mathrm{Var}[\widehat{p}] = \frac{1}{m^2}\sum_{i=1}^{m}\mathrm{Var}[X_i] = \frac{p(1-p)}{m}$

High Probability Bounds

In ML, many papers/theorems state ***with probability*** $1-\delta\ldots$”

- **“Fake” high-probability (loose):**
  - ◇ Deviation scales as $\mathcal{O}\left(\mathrm{poly}\frac{1}{\delta}\right)$
  - ◇ polynomial tail $\Rightarrow$ failure probability is not truly small.
  - ◇ Example: Markov inequality, Chebyshev inequality.

- **True high-probability (tight)**
  - ◇ Deviation scales as $\mathcal{O}\left(\log\frac{1}{\delta}\right)$
  - ◇ exponential tail $\Rightarrow$ failure probability small.
  - ◇ Example: Hoeffding’s Inequality.

$$\Pr\left[|\widehat{p}-p|\geq\epsilon\right]\leq\delta$$

e.g., $\delta=0.001$

$$m\geq\frac{1}{4\epsilon^2\delta}=\frac{\mathbf{250}}{\epsilon^2}$$

$$m\geq\frac{1}{2\epsilon^2}\ln\frac{2}{\delta}\approx\frac{\mathbf{3.8}}{\epsilon^2}$$

Definition 2.3 (PAC-learning) *A concept class $\mathcal{C}$ is said to be* PAC-learnable *if there exists an algorithm $\mathcal{A}$ and a polynomial function $poly(\cdot,\cdot,\cdot,\cdot)$ such that for any $\epsilon > 0$ and $\delta > 0$, for all distributions $\mathcal{D}$ on $\mathcal{X}$ and for any target concept $c \in \mathcal{C}$, the following holds for any sample size $m \geq poly(1/\epsilon, 1/\delta, n, size(c))$:*

$$\mathbb{P}_{S \sim \mathcal{D}^m}[R(h_S) \leq \epsilon] \geq 1 - \delta. \tag{2.4}$$

If $\mathcal{A}$ further runs in $poly(1/\epsilon, 1/\delta, n, size(c))$, then $\mathcal{C}$ is said to be efficiently PAC-learnable. *When such an algorithm $\mathcal{A}$ exists, it is called a* PAC-learning algorithm *for $\mathcal{C}$.*

Theorem 3.3 *Let $\mathcal{G}$ be a family of functions mapping from $\mathcal{Z}$ to $[0,1]$. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of an i.i.d. sample S of size m, each of the following holds for all $g \in \mathcal{G}$:*

$$\mathbb{E}[g(z)] \leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\mathfrak{R}_m(\mathcal{G}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}} \tag{3.3}$$

Foundations of Machine Learning (2nd Edition)

Entropy

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

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

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

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

$\Longrightarrow$ The entropy is a lower bound on ***lossless data compression*** and is therefore a critical quantity to consider in information theory.

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:* $\mathrm{KL}(\boldsymbol{p}\|\boldsymbol{q}) \geq \frac{1}{2} \|\boldsymbol{p} - \boldsymbol{q}\|_1^2$.

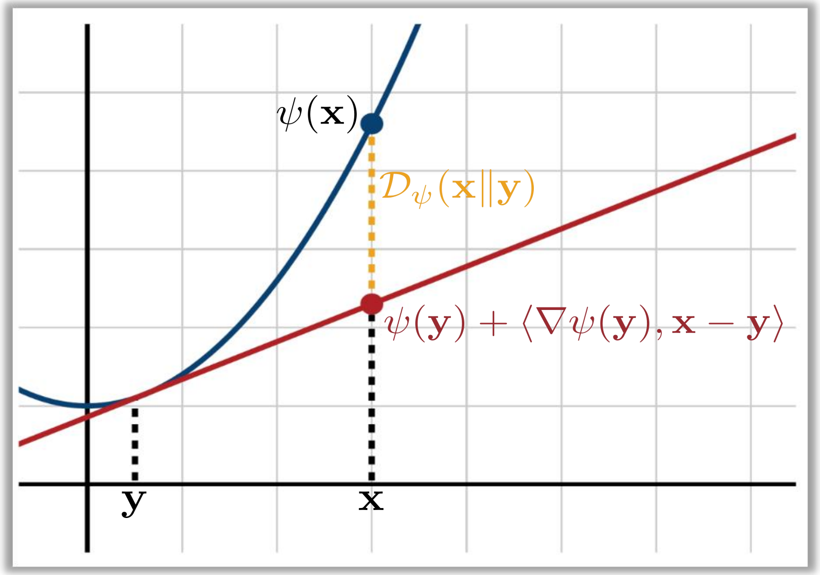
Bregman Divergence

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 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	$\sum_i x_i \log x_i$	$\mathrm{KL}(\mathbf{x} \| \mathbf{y})$



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

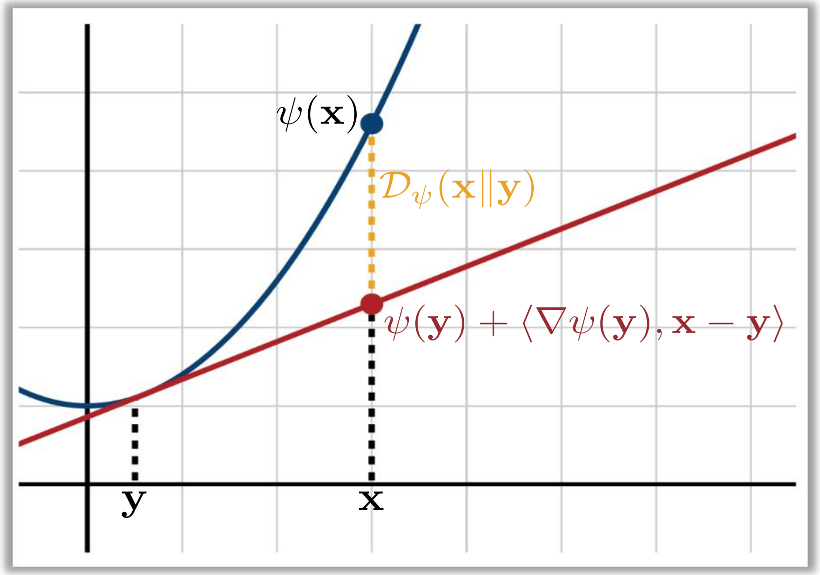
$$\mathcal{D}_\psi(\mathbf{x}\|\mathbf{y}) = \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle.$$

Q: *Is its importance due to generality?*

Not exactly, consider more general one like

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

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



Part 4. Asymptotic Notations

- Definition
- Illustration
- Example

Definition

- $\Theta(g(n)) = \{f(n) \mid$ there exist positive constants c_1, c_2, and n_0 such that $0 \le c_1 g(n) \le f(n) \le c_2 g(n)$ for all $n \ge n_0\}$.

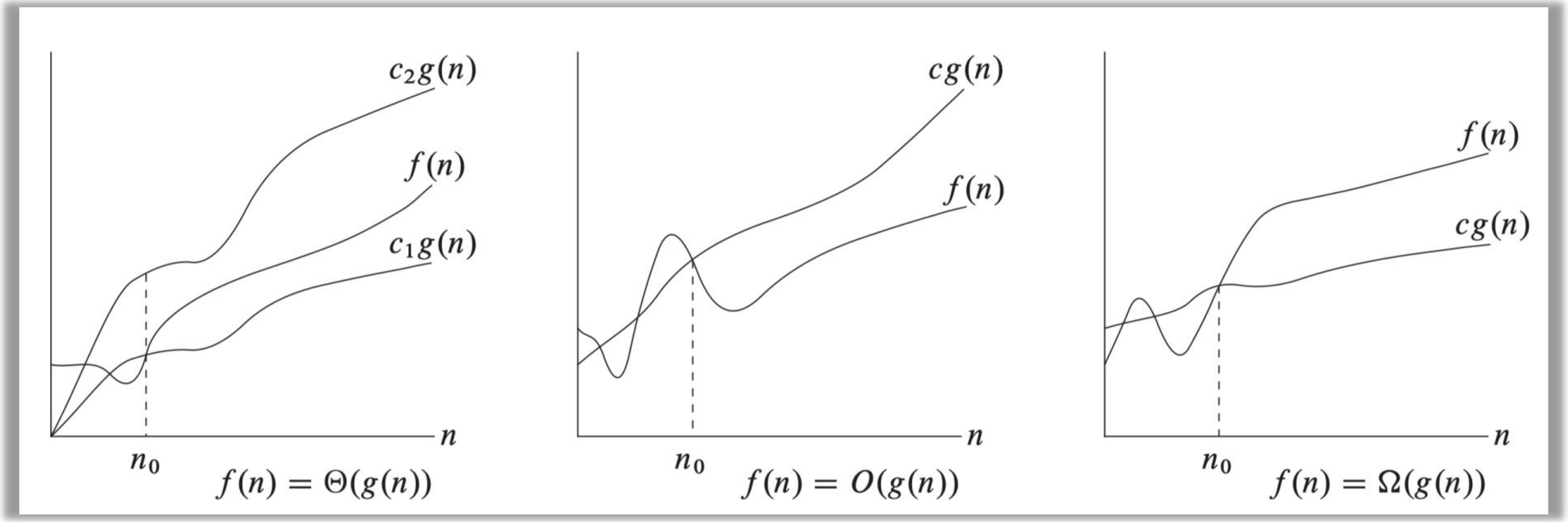
- $\mathcal{O}(g(n)) = \{f(n) \mid$ there exist positive constants c and n_0 such that $0 \le f(n) \le cg(n)$ for all $n \ge n_0\}$.

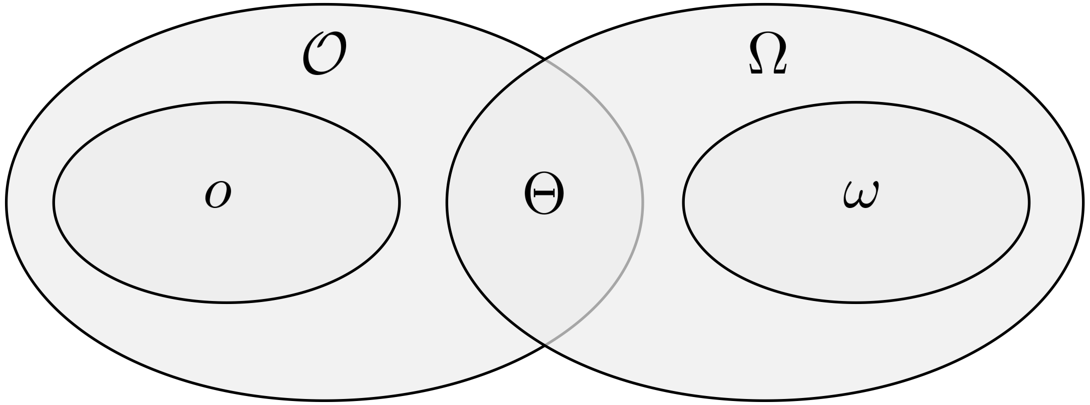
- $\Omega(g(n)) = \{f(n) \mid$ there exist positive constants c and n_0 such that $0 \le cg(n) \le f(n)$ for all $n \ge n_0\}$.

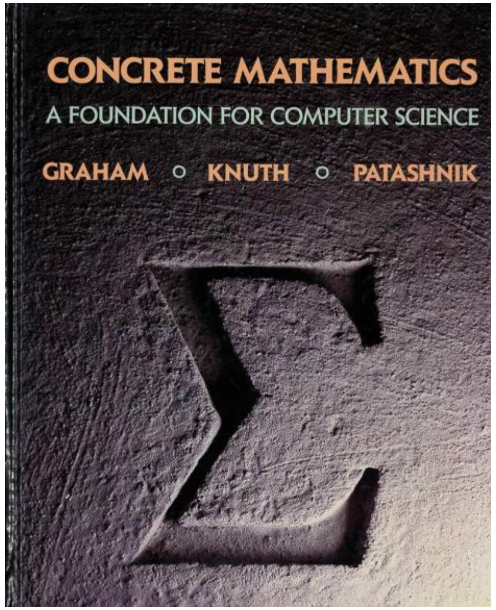
- $o(g(n)) = \{f(n) \mid$ for any positive constant $c > 0$, there exists a constant $n_0 > 0$ such that $0 \le f(n) < cg(n)$ for all $n \ge n_0\}$.

- $\omega(g(n)) = \{f(n) \mid$ for any positive constant $c > 0$, there exists a constant $n_0 > 0$ such that $0 \le cg(n) < f(n)$ for all $n \ge n_0\}$.

Illustration







Example

- $3n^3 + 2n^2 + n + \log n = \Theta(n^3)$

$\cancel{\mathcal{O}(1) < \mathcal{O}(\log n) < \mathcal{O}(n) < \mathcal{O}(n \log n) < \mathcal{O}\left(n^2\right) < \mathcal{O}\left(2^n\right) < \mathcal{O}(n!)}$

- $\Theta(1) < \Theta(\log n) < \Theta(n) < \Theta(n \log n) < \Theta\left(n^2\right) < \Theta\left(2^n\right) < \Theta(n!)$

Theorem 4. *Under Assumptions 1, 2, and 3, set the pool of candidate step sizes $\mathcal{H}$ as*

$$\mathcal{H} = \left\{ \eta_i = \min\left\{ \frac{1}{8L}, \sqrt{\frac{D^2}{8G^2T}} \cdot 2^{i-1} \right\} \mid i \in [N] \right\}, \tag{26}$$

where $N = \lceil 2^{-1} \log_2(G^2T/(8D^2L^2)) \rceil + 1$ is the number of candidate step sizes; further set the correction coefficient as $\lambda = 2L$ and the learning rate of the meta-algorithm as $\varepsilon = \min\{1/(8D^2L), \sqrt{(\ln N)/(D^2(\|\nabla f_1(\mathbf{x}_1)\|_2^2 + \bar{V}_T))}\}$. Then, Sword++ satisfies

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}_t) \leq \mathcal{O}\left(\sqrt{(1 + P_T + V_T)(1 + P_T)}\right)$$

for any comparator sequence $\mathbf{u}_1, \ldots, \mathbf{u}_T \in \mathcal{X}$. In above, $\bar{V}_T = \sum_{t=2}^{T} \|\nabla f_t(\mathbf{x}_t) - \nabla f_{t-1}(\mathbf{x}_{t-1})\|_2^2$ is the variant of gradient variation V_T.

Theorem 6 (Dynamic NE-Regret)**.** *When x-player follows Algorithm 1 and y-player follows Algorithm 2, we have the following dynamic NE-regret bound:*

$$\text{DynNE-Reg}_T = \left| \sum_{t=1}^{T} x_t^\top A_t y_t - \sum_{t=1}^{T} \min_{x \in \Delta_m} \max_{y \in \Delta_n} x^\top A_t y \right|$$
$$= \widetilde{\mathcal{O}}\left(\min\{\sqrt{(1 + V_T)(1 + P_T)} + P_T, 1 + W_T\} \right).$$

two examples of theorem statement

It is both fine to use “ = ” or “\leq”

Summary

- **CALCULUS**
  - Gradient and Deritives
  - Hessian
  - Chain Rule
  - The Matrix Cookbook
- **STATITICS & INFORMATION THEORY**
  - Concentration inequalities
    - Markov's Inequality
    - Chebyshev's Inequality
    - Hoeffding's Inequality
  - High-probability bounds
    - log (1/δ), High probability
    - Poly (1/δ), Low probability
  - Entropy
  - KL Divergence (Relative Entropy)
  - Bregman Divergence
- **LINEAR ALGEBRA**
  - Positive (Semi-)Definite Matrix
  - Rank
  - Inner Product
  - Norm
    - Gradient Norm
    - Matirx Norm
    - Norm Relationship
  - Decomposition
    - Eigenvalue Decomposition
    - Singular Value Decomposition

Outline

- Math Background
  - Calculus, Linear Algebra
  - Probability & Statistics
  - Information Theory, Asymptotic Notations
- Convex Optimization Basics
  - ML as Optimization
  - Convex Function, Convex Set
  - Convex Optimization Problem

Learning by/as 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.

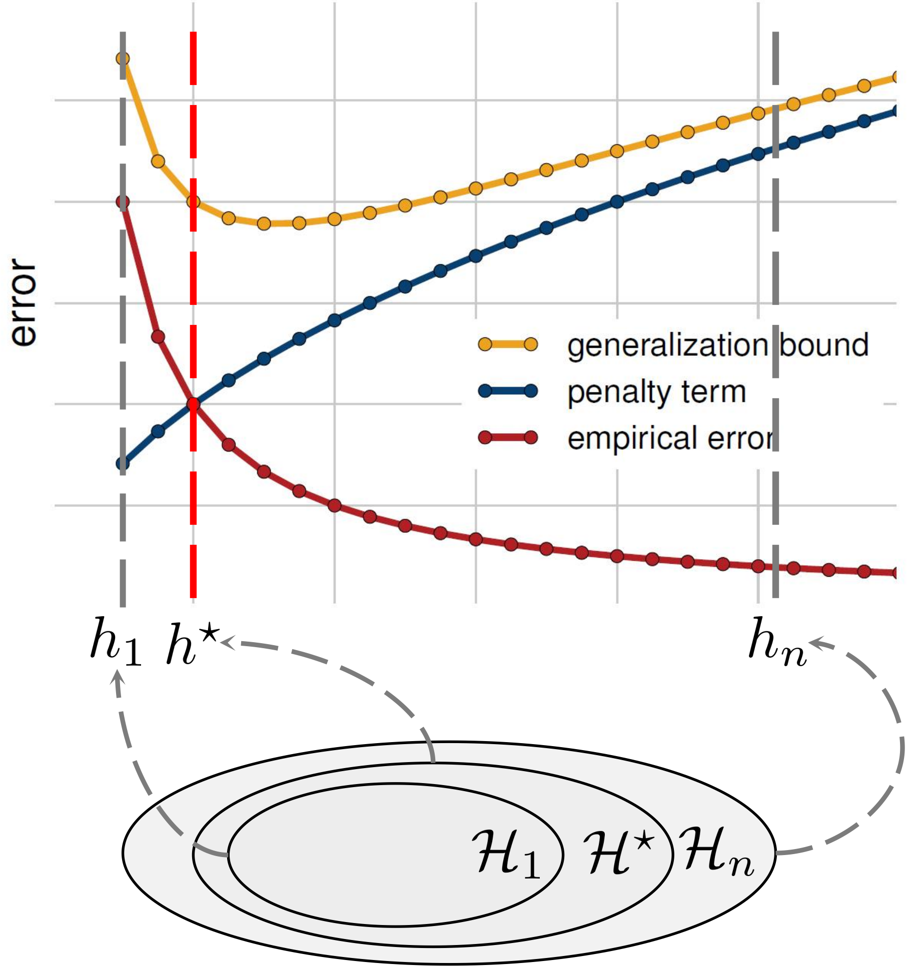
in optimization language: this is called Sample Average Approximation (SAA)

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***.



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 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)***:

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

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)***:

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

(Constrained) Optimization Problem

- We adopt a ***minimization*** language

$$\begin{aligned} \min \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{x} \in \mathcal{X} \end{aligned}$$

  - optimization variable $\mathbf{x} \in \mathbb{R}^d$
  - objective function: $f : \mathbb{R}^d \mapsto \mathbb{R}$
  - feasible domain: $\mathcal{X} \subseteq \mathbb{R}^d$

Unconstrained Optimization

- The optimization variable is feasible over the whole $\mathbb{R}^d$-space.

$$\begin{aligned} \min \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{x} \in \mathbb{R}^d \end{aligned}$$

- It is one of ***the most basic*** forms of mathematical optimization and serves as the foundations.

--- *"any optimization problem can be regarded as an unconstrained one"*

$$\begin{aligned} \min \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{x} \in \mathcal{X} \end{aligned} \quad \Longrightarrow \quad \begin{aligned} \min \quad & h(\mathbf{x}) \triangleq f(\mathbf{x}) + \underline{\delta_{\mathcal{X}}(\mathbf{x})} \\ \text{s.t.} \quad & \mathbf{x} \in \mathbb{R}^d \end{aligned}$$

barrier/indicator function

$$\delta_{\mathcal{X}}(\mathbf{x}) = \begin{cases} 0, & \mathbf{x} \in \mathcal{X}, \\ \infty, & \mathbf{x} \notin \mathcal{X}. \end{cases}$$

Convex Optimization

- This lecture focuses on the following simplified setting:
  - Language: ***minimization*** problem
  - Objective function: ***continuous*** and ***convex***
  - Feasible domain: a ***convex*** subset of ***Euclidean space***

- ❑ What is a convex set?
- ❑ What is a convex function?
- ❑ How to minimize?

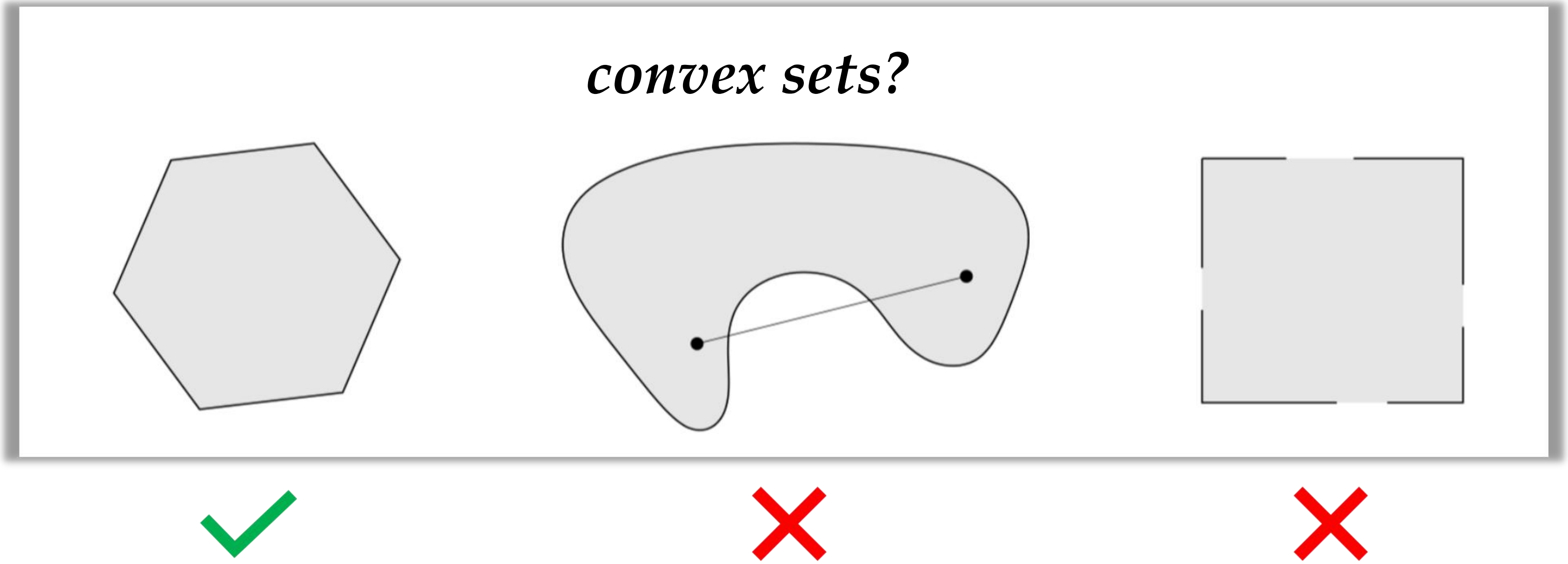
Before diving into details, one Q…

Why should I learn about "convex optimization"?

Convex Set

Definition 1 (Convex Set)**.** A set $\mathcal{X}$ is convex if for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, all the points on the line segment connecting $\mathbf{x}$ and $\mathbf{y}$ also belong to $\mathcal{X}$, i.e.,

$$\forall \alpha \in [0, 1],\ \alpha \mathbf{x} + (1 - \alpha)\mathbf{y} \in \mathcal{X}.$$



Convex Set

Definition 1 (Convex Set). A set $\mathcal{X}$ is convex if for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, all the points on the line segment connecting $\mathbf{x}$ and $\mathbf{y}$ also belong to $\mathcal{X}$, i.e.,

$$\forall \alpha \in [0,1],\ \alpha \mathbf{x} + (1-\alpha)\mathbf{y} \in \mathcal{X}.$$

Examples

- A line segment is convex.
- A ray, which has the form $\{\mathbf{x}_0 + \theta \mathbf{v} \mid \theta \geq 0\}$, where $\mathbf{v} \neq \mathbf{0}$, is convex.
- Any subspace is convex.

Convex Set

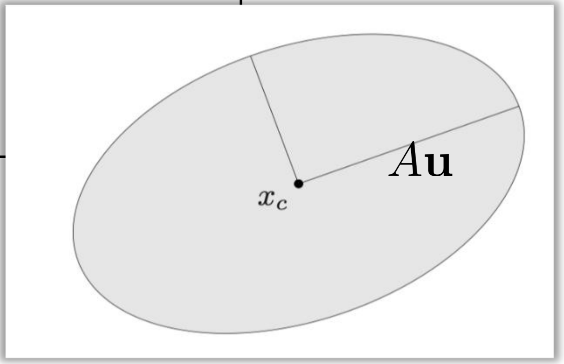
Definition 2 (Ball). A (Euclidean) ball (or just ball) in $\mathbb{R}^d$ has the form

$$\mathbb{B}(\mathbf{x}_c, r) = \{\mathbf{x}_c + r\mathbf{u} \mid \|\mathbf{u}\|_2 \leq 1\}.$$

Definition 3 (Ellipsoids). A ellipsoid in $\mathbb{R}^d$ has the form

$$\mathcal{E}(\mathbf{x}_c, A) = \{\mathbf{x}_c + A\mathbf{u} \mid \|\mathbf{u}\|_2 \leq 1\},$$

where A is assumed to be symmetric and positive definite.

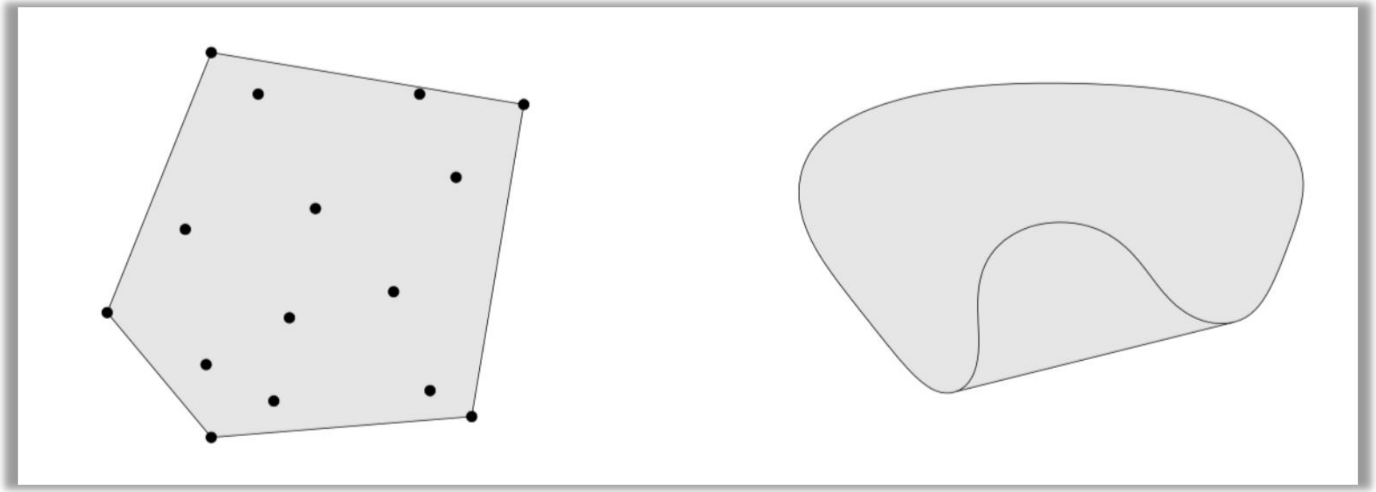


Convex Set

Definition 4 (Convex Hull). The convex hull of a set $\mathcal{X}$, denoted conv $\mathcal{X}$, is the set of all convex combinations of points in $\mathcal{X}$:

$$\operatorname{conv}\mathcal{X} = \{\theta_1 \mathbf{x}_1 + \cdots + \theta_k \mathbf{x}_k \mid \mathbf{x}_i \in \mathcal{X}, \theta_i \geq 0, i \in [k], \theta_1 + \cdots + \theta_k = 1\}.$$

Examples:



Projection onto Convex Sets

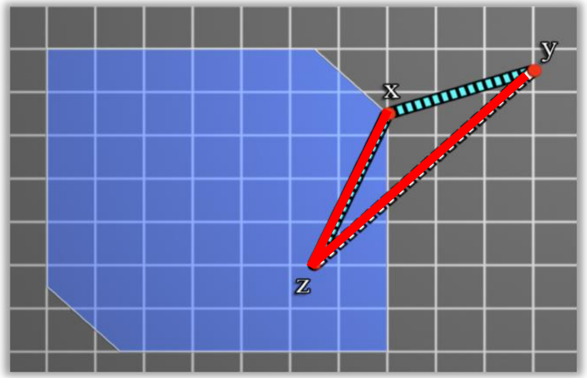
Definition 5 (Projection). The projection a given point $\mathbf{y}$ onto a convex set $\mathcal{X}$ is defined as the closest point inside the convex set. Formally,

$$\mathbf{x}^\star = \Pi_{\mathcal{X}}[\mathbf{y}] \triangleq \arg\min_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x} - \mathbf{y}\|.$$

Note: the projected point $\mathbf{x}^\star$ is unique as long as the norm is strictly convex.

Theorem 1 (Pythagoras Theorem). *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a convex set, $\mathbf{y} \in \mathbb{R}^d$. Then for any $\mathbf{z} \in \mathcal{X}$ we have*

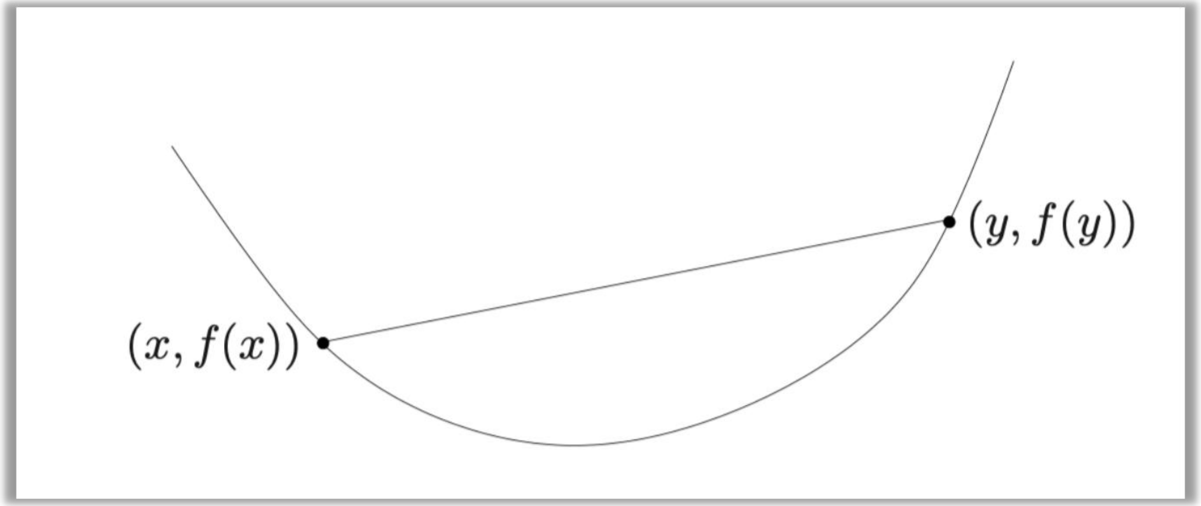
$$\|\mathbf{y} - \mathbf{z}\| \geq \|\Pi_{\mathcal{X}}[\mathbf{y}] - \mathbf{z}\|.$$



Convex Function

Definition 6 (Convex Function)**.** A function $f : \mathcal{X} \mapsto \mathbb{R}$ is called *convex* if for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$,

$$\forall \alpha \in [0,1], \quad f((1-\alpha)\mathbf{x} + \alpha\mathbf{y}) \leq (1-\alpha)f(\mathbf{x}) + \alpha f(\mathbf{y}).$$



a convex function

Convex/Concave Function

Definition 6 (Convex Function). A function $f : \mathcal{X} \mapsto \mathbb{R}$ is called ***convex*** if for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$,

$$\forall \alpha \in [0,1], \quad f((1-\alpha)\mathbf{x} + \alpha\mathbf{y}) \leq (1-\alpha)f(\mathbf{x}) + \alpha f(\mathbf{y}).$$

Definition 7 (Concave Function). A function $f : \mathcal{X} \mapsto \mathbb{R}$ is called ***concave*** if for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$,

$$\forall \alpha \in [0,1], \quad f((1-\alpha)\mathbf{x} + \alpha\mathbf{y}) \geq (1-\alpha)f(\mathbf{x}) + \alpha f(\mathbf{y}).$$

- Both definitions have already assumed a ***convex*** feasible domain.
- We focus on the ***"convex language"***, clearly the negative of concave functions are convex.

Convex Function

How to check whether a function is convex or not?

Theorem 2. *A function* f *is convex* ***if and only if*** $\mathrm{dom}\, f$ ***is convex*** *and one of the following properties hold, for all* $\mathbf{x}, \mathbf{y} \in \mathrm{dom}\, f$ *and* $\alpha \in [0, 1]$*,*

(i) Zeroth order condition: $f((1-\alpha)\mathbf{x} + \alpha\mathbf{y}) \leq (1-\alpha) f(\mathbf{x}) + \alpha f(\mathbf{y})$.

(ii) First order condition (provided f *is differentiable):* $f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leq f(\mathbf{y})$.

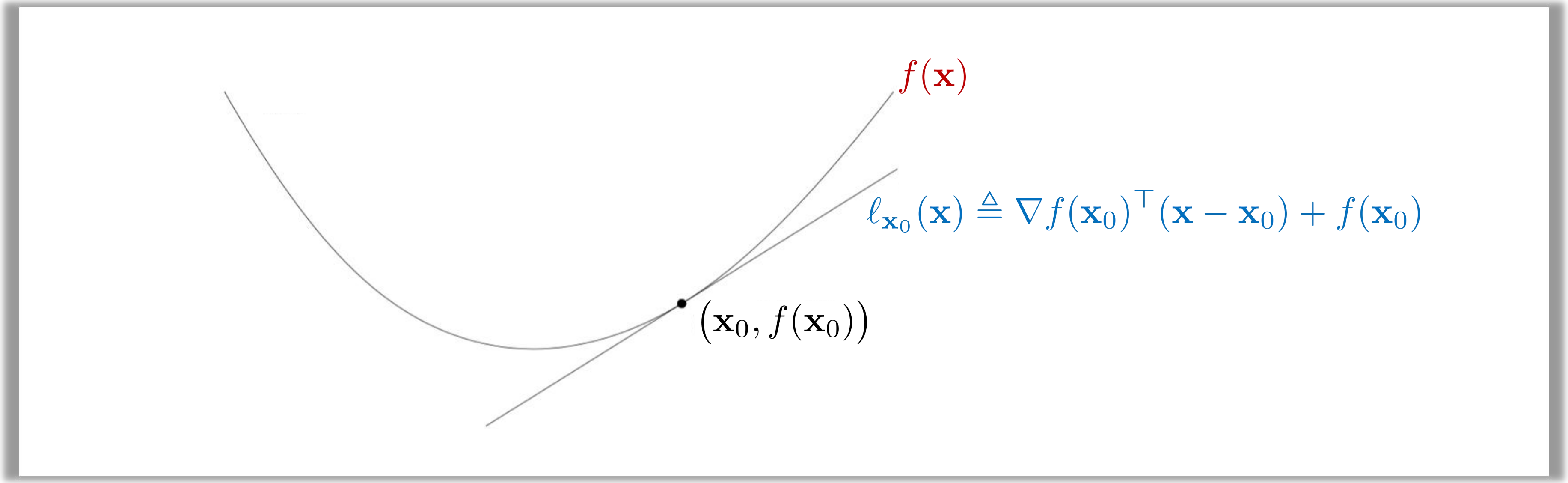
(iii) Second order condition (provided f *is twice differentiable):* $\nabla^2 f(\mathbf{x}) \succeq 0$.

Convex Function

If f is convex and differentiable, then $\underline{f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle} \leq f(\mathbf{y})$ for all $\mathbf{x}, \mathbf{y} \in \operatorname{dom} f$.

the first-order Taylor approximation of f near $\mathbf{x}$

A commonly used equivalent form: $f(\mathbf{x}) - f(\mathbf{y}) \leq \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{y} \rangle$.



Convex Function

Examples on $\mathbb{R}$:

- Exponential: e^{ax}, where $a \in \mathbb{R}$.
- Powers: x^a, where $a \geq 1$ or $a \leq 0$.
- Powers of absolute value: $|x|^p$, where $p \geq 1$.
- Negative logarithm: $-\log x$.
- Negative entropy: $x \log x$.

Convex Function

Examples on $\mathbb{R}^d$:

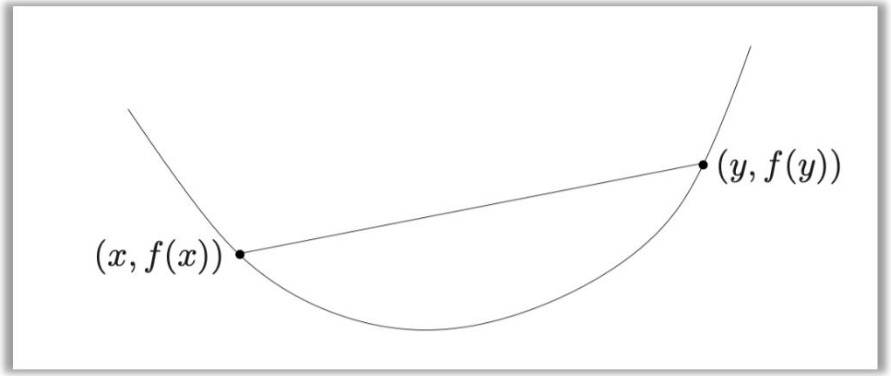
- norm: $f(\mathbf{x}) = \|\mathbf{x}\|$.
- maximum: $f(\mathbf{x}) = \max\{x_1, \ldots, x_n\}$.
- Log-sum-exp: $f(\mathbf{x}) = \log(e^{x_1} + \cdots + e^{x_n})$.

Jensen's Inequality

Theorem 3 (Jensen's Inequality). *If X is a random variable such that $X \in \operatorname{dom} f$ with probability one, and f is convex, then we have*

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

Intuition:



Convexity: $$f(\underbrace{\theta_1 \mathbf{x}_1 + \cdots + \theta_k \mathbf{x}_k}_{\mathbb{E}[X]}) \leq \underbrace{\theta_1 f(\mathbf{x}_1) + \cdots + \theta_k f(\mathbf{x}_k)}_{\mathbb{E}[f(X)]}$$

Convex Optimization Problem

- We adopt a ***minimization*** language

$$
\begin{aligned}
\min \quad & f(\mathbf{x}) \\
\text{s.t.} \quad & g_i(\mathbf{x}) \le 0, \;\; i = 1, \cdots, m \\
& \mathbf{a}_i^\top \mathbf{x} = b_i, \;\; i = 1, \cdots, n
\end{aligned}
$$

  - optimization variable $\mathbf{x} \in \mathbb{R}^d$
  - ***convex*** objective function: $f : \mathbb{R}^d \mapsto \mathbb{R}$
  - ***convex*** inequality constraints: $g_1, \ldots, g_m$

Convex Optimization Problem

Example 1 (SVM)**.**

$$
\begin{aligned}
\min_{\mathbf{w}, b} \quad & \|\mathbf{w}\|^2 \\
\text{s.t.} \quad & y_i \left(\mathbf{w}^\top \mathbf{x}_i + b \right) \geq 1, \ \ i = 1, \cdots, n
\end{aligned}
$$

Example 2 (NMF decomposition)**.**

$$
\begin{aligned}
\min_{U, V} \quad & \left\| X - UV^\top \right\|_{\mathrm{F}}^2 \\
\text{s.t.} \quad & U_{i,j}, V_{i,j} \geq 0
\end{aligned}
$$

Ref: Lee, DD & Seung, HS (1999). Learning the parts of objects by non-negative matrix factorization. ***Nature*** 401,788-791.

Convex Optimization

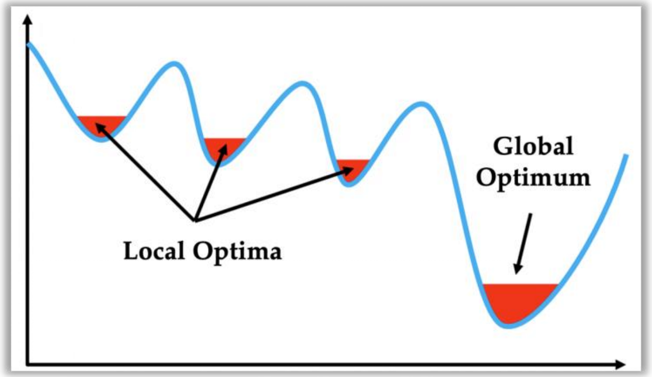
- This lecture focuses on the following simplified setting:
  - Language: ***minimization*** problem
  - Objective function: ***continuous*** and ***convex***
  - Feasible domain: a ***convex*** subset of ***Euclidean space***

- ❑ What is a convex set?
- ❑ What is a convex function?
- ❑ How to minimize?

Before diving into details, one Q…

Why should I learn about "convex optimization"?

Why Convexity is Nice?



- **Local to Global Phenomenon**

For convex (unconstrained) optimization, ***local minima are global minima***.

> **Theorem 8.** *Let f be convex. If $\mathbf{x}$ is a local minimum of f then $\mathbf{x}$ is a global minimum of f.*

A simple proof:

Assume that $\mathbf{x}$ is local minimum of f. Then for γ small enough, for any $\mathbf{y}$,

(local minima)

$$\underline{f(\mathbf{x}) \leq f((1-\gamma)\mathbf{x} + \gamma\mathbf{y})} \leq (1-\gamma)f(\mathbf{x}) + \gamma f(\mathbf{y}),$$

which implies $f(\mathbf{x}) \leq f(\mathbf{y})$ and thus $\mathbf{x}$ is a global minimum of f. $\square$

Why Convex Optimization?

- Let's invent it from "the first principle"

FACT: most OPT problems are HARD.

See [Section 1.1 of Nesterov's book] for evidence

Without further structure:

- May have multiple local minima, complex landscape.
- Often NP-hard even to approximate.

⇨ *Can we identify a class of broad problems that is "EASY" (or "TRACABLE")?*

Why Convex Optimization?

- Let's invent it from "the first principle"

Assumption 1. For any $f \in \mathcal{F}$, the first-order optimality condition suffices to the global optimality, namely, if $\nabla f(x^*) = 0$ then x^* is a global optimal solution.

Assumption 2. If $f_1, f_2 \in \mathcal{F}$, then $\alpha f_1 + \beta f_2 \in \mathcal{F}$ should hold for any $\alpha, \beta \geq 0$.

Assumption 3. The linear function should be in the class, i.e., $f(x) = ax + b$ should satisfy $f \in \mathcal{F}$ for any $a, b \in \mathbb{R}$.

⇨ **Claim:** Under Assumptions 1-3, every $f \in \mathcal{F}$ must be convex.

Why Convex Optimization?

Claim: Under Assumptions 1-3, every $f \in \mathcal{F}$ must be convex.

Proof: (consider 1-dim scalar function for simplicity)

Take any $f \in \mathcal{F}$ and any fixed point $x_0 \in \mathbb{R}$. We construct the function

$$\phi_f^{x_0}(x) \triangleq f(x) - f'(x_0)x.$$

We simply abbreviate it as $\phi(x)$. By Assumption 2, $-f'(x_0)x \in \mathcal{F}$ and hence $\phi(x) \in \mathcal{F}$. Computing its derivative, we have

$$\phi'(x) = f'(x) - f'(x_0). \quad \Rightarrow \quad \phi'(x_0) = 0.$$

By Assumption 1, x_0 is the global optimizer of $\phi(x)$.

Why Convex Optimization?

Claim: Under Assumptions 1-3, every $f \in \mathcal{F}$ must be convex.

Proof: (consider 1-dim scalar function for simplicity)

By Assumption 1, x_0 is the global optimizer of $\phi(x)$. So for all x,

$$\phi(x) \geq \phi(x_0). \quad \Rightarrow \quad f(x) - f'(x_0)x \geq f(x_0) - f'(x_0)x_0.$$

Rearranging yields

$$f(x) \geq f\left(x_0\right) + f'\left(x_0\right)\left(x - x_0\right).$$

This is exactly the definition of the convex function. $\square$

Why Convex Optimization?

- **Math/OPT:** Convex OPT offers a unified and elegant framework for a broad class of problems, with numerous profound theories and insights developed.
- **ML:** Provides key algorithmic tools for large-scale ML problems, such as logistic regression, sparse coding, and PCA.
- **Non-convex OPT with NN:** Many advances in non-convex OPT are fundamentally rooted in convex OPT, like SGD, AdaGrad, Adam, etc.

Summary

ML AS OPTIMIZATION
- Risk Minimization
- Empirical Risk Minimization (ERM)
- Structural ERM

CONVEX OPTIMIZATION
- Convex Set
- Convex Function
- Local to Global Property

WHY CONVEX OPTIMIZATION
- From First Principle
- Three Assumptions