Preliminaries

Optimization Techniques (ENGG*6140)

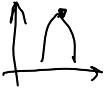
School of Engineering,
University of Guelph, ON, Canada

Course Instructor: Benyamin Ghojogh

Winter 2023

What is Optimization?

Optimization problem



$f : x \longmapsto f(x)$

- Consider a function representing some cost. We call it **cost function** or **objective function**.
- We want to **minimize** or **maximize** this objective function.
- Examples:
  - Example for **minimization**: the cost function can be the error of some airplane structure from the perfect aerodynamic structure.
  - Example for **maximization**: the objective function can be the profit of the company.
  - **All life** is optimization!
  - **All machine learning** in artificial intelligence is optimization!
- The variables of the objective function are called the **objective variables** or **decision variables** or **optimization variables**.
- Example:

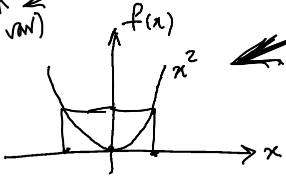
cost

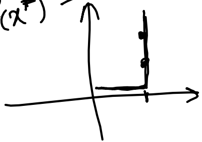
$$\boxed{\underset{x}{\text{minimize}}} \quad f(x) = x^2.$$

argument (objective var)

$x^* = 0$

$f(x^*) = 0$





Univariate and multivariate optimization problems

- The optimization problem can be **univariate**, meaning that the optimization problem has only one scalar variable. Example:

$$\underset{x}{\text{minimize}} \quad f(x) = x^2.$$

scalar { not bold, small case

- The optimization problem can be **multivariate**, meaning that the optimization problem has several scalar variables $\{x_1, \dots, x_n\}$. These variables can be combined into a vector $\boldsymbol{x} = [x_1, \dots, x_n]^\top$ or matrix. Example:

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{x} = x_1^2 + \dots + x_n^2.$$

transpose

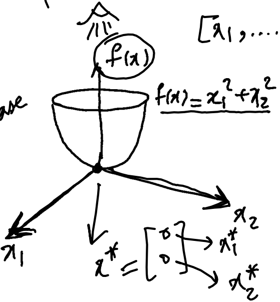
$\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$

vectors { bold, small case

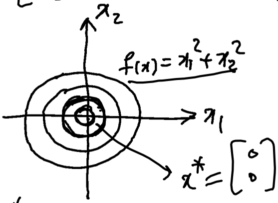
$[x_1, \dots, x_n]\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} =$

$X = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & & \\ x_{m1} & \cdots & x_{mn} \end{bmatrix}$

matrix { bold, Capital case



$f(x^*) = 0$

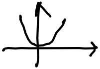


Unconstrained and constrained problems

- The optimization problem can be **unconstrained**, meaning that we simply optimize a function only. Example:

$$\underset{x}{\text{minimize}} \quad f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{x}.$$

$f(x) = x^2$



- The optimization problem can be **constrained**, meaning that we optimize a function while there are some constraints on the optimization variables. Example:

$\min_x \; f(x) = x^2$
s.t. $x \le 2$
$x \ge 16$

$x \in \emptyset$

$$\underset{\boldsymbol{x}=[x_1,x_2]^\top}{\text{minimize}} \quad f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{x} = x_1^2 + x_2^2$$
$$\text{subject to} \quad x_1 + 2x_2 \, \cancel{\le} = 8,$$
$$2x_1 + x_2^2 \le 16.$$

constraint (restriction)

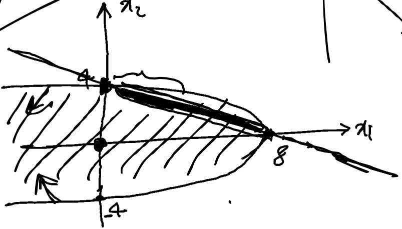
restriction on x (constraints)

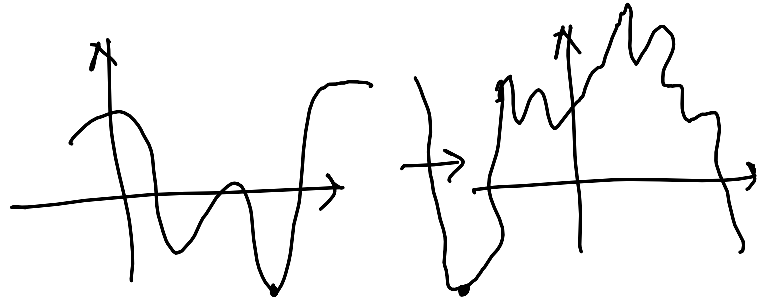
$x_1 = 0 \to 2x_2 = 8 \to x_2 = 4$

$x_2 = 0 \to x_1 = 8$

$x_1 = 0 \to x_2 = 4, -4$

$x_2 = 0 \to x_1 = 8$





two goals

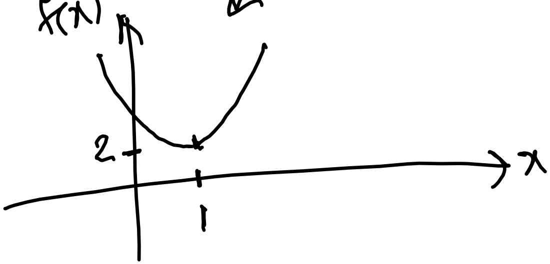
$\boxed{\text{argument (obj var) } x^*}$ ✓

$\boxed{f(x^*)}$ ✓

$f(x^*)$

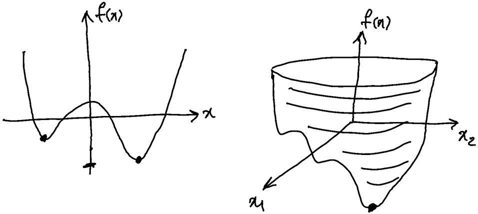
$$\min_x f(x) = (x-1)^2 + 2 \longrightarrow x^* = 1$$

$f(x^*) = 2$

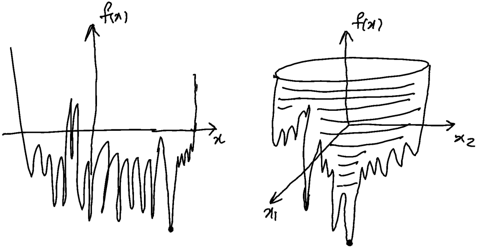


Optimization versus search

- If the objective problem is **simple enough**,
  we can solve it using **classic optimization** methods. We will learn important classic methods.

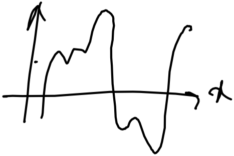


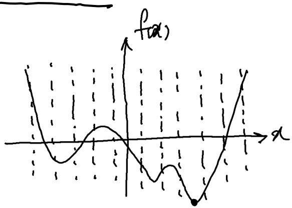
- If the objective function is **complicated** or if we have **too many constraints**, we can use **search** for finding a good solution.

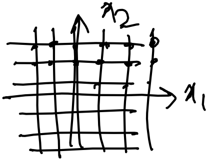


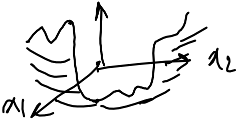
Search for optimization

- We can do **grid search** or **brute-force search**.

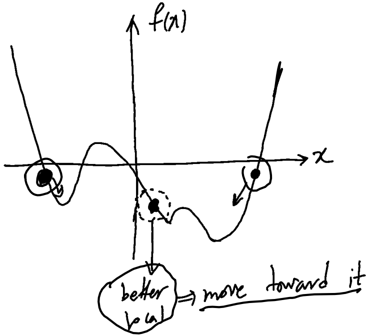


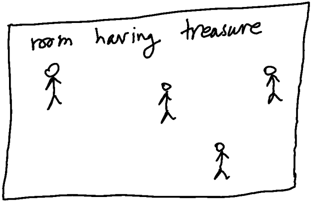






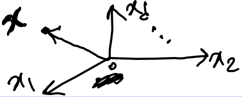
- Or we can **search wisely** by **metaheuristic optimization**. We will learn several important metaheuristic optimization methods.

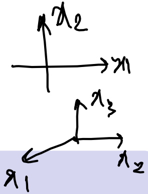




Preliminaries on Sets and Norms

Interior, closure, and boundary





Real number

d_dim Euclidean space

Definition (Interior of set)

Consider a set $\mathcal{D}$ in a metric space $\mathbb{R}^d$. The point $\boldsymbol{x} \in \mathcal{D}$ is an interior point of the set if:

$$\star \quad \exists \epsilon > 0 \text{ such that } \{\boldsymbol{y} \mid \|\boldsymbol{y}-\boldsymbol{x}\|_2 \leq \epsilon\} \subseteq \mathcal{D}.$$

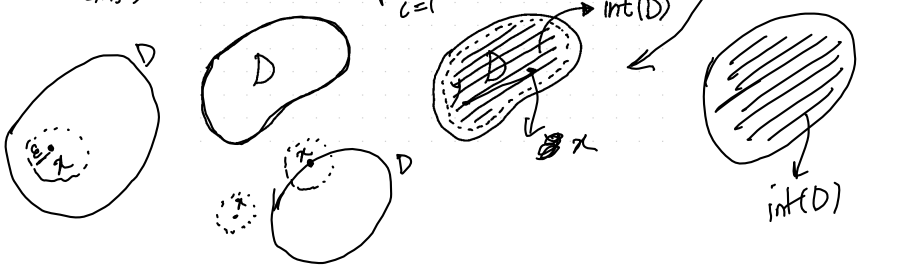
$\|x\|_p$

The interior of the set, denoted by $\textbf{int}(\mathcal{D})$, is the set containing all the interior points of the set.

there exists

$\sqrt{\sum_{i=1}^{d}(y_i - x_i)^2}$

$\boxed{\|x-y\|_2}$

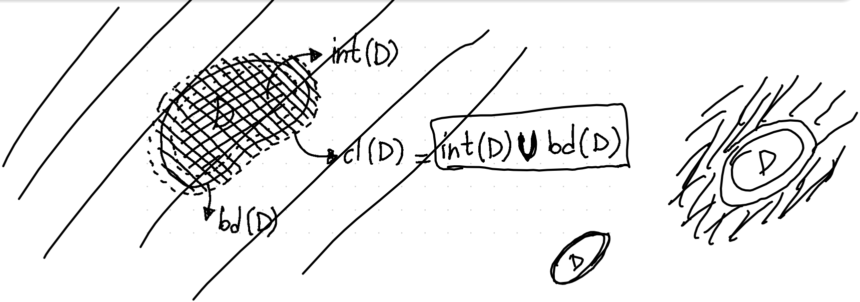


Interior, closure, and boundary

-(excluding)

Definition (Closure and boundary of set)

- The closure of the set is defined as $\boxed{\mathbf{cl}(\mathcal{D})} := \mathbb{R}^d \setminus \mathbf{int}(\mathbb{R}^d \setminus \mathcal{D})$.
- The boundary of set is defined as $\mathbf{bd}(\mathcal{D}) := \mathbf{cl}(\mathcal{D}) \setminus \mathbf{int}(\mathcal{D})$.
- An open (resp. closed) set does not (resp. does) contain its boundary.
- The closure of set can be defined as the smallest closed set containing the set. In other words, the closure of set is the union of interior and boundary of the set.



Convex set

$t \in \mathbb{R}$

$t < 0$
$t > 1$

$x \bullet \longrightarrow \bullet\, y$

$\boxed{tx + (1-t)y}$

$0 \leq t \leq 1$

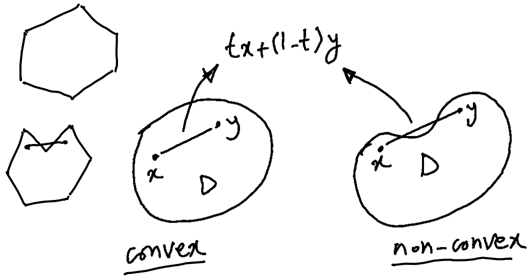
Definition (Convex set and convex hull)

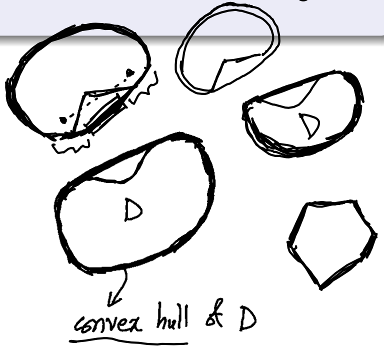
A set $\mathcal{D}$ is a convex set if it completely contains the line segment between any two points in the set $\mathcal{D}$:

for all

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{D}, 0 \leq t \leq 1 \implies t\boldsymbol{x} + (1-t)\boldsymbol{y} \in \mathcal{D}.$$

The convex hull of a (not necessarily convex) set $\mathcal{D}$ is the smallest convex set containing the set $\mathcal{D}$. If a set is convex, it is equal to its convex hull.





Min, max, sup, inf

Definition (Minimum, maximum, infimum, and supremum)

A **minimum** and **maximum** of a function $f : \mathbb{R}^d \to \mathbb{R}$, $f : \boldsymbol{x} \mapsto f(\boldsymbol{x})$, with domain $\mathcal{D}$, are defined as:

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) \leq f(\boldsymbol{y}), \ \forall \boldsymbol{y} \in \mathcal{D},$$

$$\max_{\boldsymbol{x}} f(\boldsymbol{x}) \geq f(\boldsymbol{y}), \ \forall \boldsymbol{y} \in \mathcal{D},$$

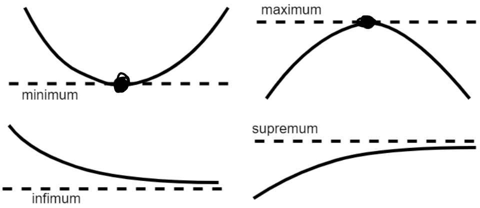
respectively.

$\max_{x} f(x)$

The minimum and maximum of a function belong to the range of function.

$\sup_{x} f(x)$

$\max_{x} f(x)$



$\to \infty$

Min, max, sup, inf

Definition (Infimum and supremum)

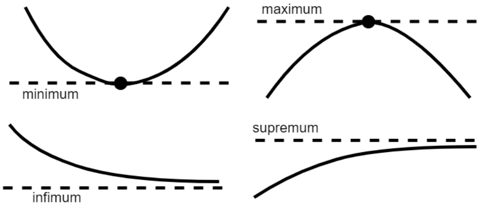
Infimum and **supremum** are the lower-bound and upper-bound of function, respectively:

lower

$$\inf_{\mathbf{x}} f(\mathbf{x}) := \max\{z \in \mathbb{R} \mid \underline{z \leq f(\mathbf{x})}, \underline{\forall \mathbf{x} \in \mathcal{D}}\},$$

$$\sup_{\mathbf{x}} f(\mathbf{x}) := \min\{z \in \mathbb{R} \mid \underline{z \geq f(\mathbf{x})}, \forall \mathbf{x} \in \mathcal{D}\}.$$

Depending on the function, the infimum and supremum of a function may or may not belong to the range of function.



Inner product

$\mathrm{tr} \Rightarrow \begin{bmatrix} & & \\ & & \\ & & \end{bmatrix}_{3 \times 3}$

$\mathrm{tr}\left(\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}\right) = 1+5+9$

Definition (Inner product of vectors)

Consider two vectors $\boldsymbol{x} = [x_1, \dots, x_d]^\top \in \mathbb{R}^d$ and $\boldsymbol{y} = [y_1, \dots, y_d]^\top \in \mathbb{R}^d$. Their **inner product**, also called **dot product**, is:

$$\text{scalar} \leftarrow \langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^\top \boldsymbol{y} = \sum_{i=1}^{d} x_i\, y_i.$$

$x = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ $y = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$

$x^T y = \langle x, y \rangle =$

$[1, 2] \begin{bmatrix} 3 \\ 4 \end{bmatrix} = (1 \times 3) + \overset{8}{(2 \times 4)}$

$= 11$

Definition (Inner product of matrices)

We also have inner product between matrices $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{d_1 \times d_2}$. Let $\boldsymbol{X}_{ij}$ denote the (i, j)-th element of matrix $\boldsymbol{X}$. The inner product of $\boldsymbol{X}$ and $\boldsymbol{Y}$ is:

$d_1 \begin{bmatrix} d_2 \\ \\ \end{bmatrix}$

$X_{32} = \begin{bmatrix} & & \\ & & \\ & & \bullet \end{bmatrix}$

$$\langle \boldsymbol{X}, \boldsymbol{Y} \rangle = \mathbf{tr}(\boldsymbol{X}^\top \boldsymbol{Y}) = \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} \boldsymbol{X}_{i,j}\, \boldsymbol{Y}_{i,j},$$

3×2

$\mathbb{R}$

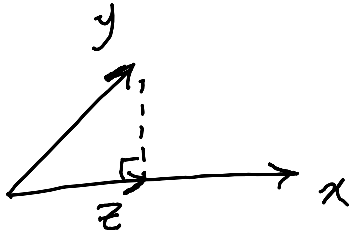
2 cols

3 rows $\begin{bmatrix} & \\ & \\ & \end{bmatrix}$

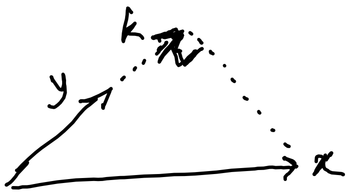
where $\mathbf{tr}(.)$ denotes the trace of matrix.

$\begin{bmatrix} 2 & 3 \\ 4 & 5 \end{bmatrix}$ $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$

$$\frac{x}{x^T x} = \frac{x}{\|x\|_2^2} = \text{normalized } x$$



$$z = \frac{y^T x}{x^T x}\, x \Rightarrow |z| = \frac{y^T x}{x^T x}$$



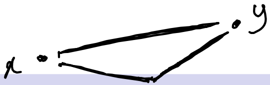
$$k = \frac{x^T y}{y^T y}\, y \Rightarrow |k| = \frac{x^T y}{y^T y}$$

* note: $\underbrace{x^T y}_{x_1 y_1 + x_2 y_2 + \cdots + x_d y_d} = \underbrace{y^T x}_{y_1 x_1 + y_2 x_2 + \cdots + y_d x_d}$

Norm

distance metric $\longleftrightarrow$ norm
connection

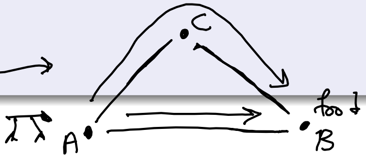
$\|x - y\|$



Definition (Norm)

A function $\|\cdot\| : \mathbb{R}^d \to \mathbb{R}$, $\|\cdot\| : \boldsymbol{x} \mapsto \|\boldsymbol{x}\|$ is a **norm** if it satisfies:

1. $\|\boldsymbol{x}\| \geq 0, \forall \boldsymbol{x}$
2. $\|a\boldsymbol{x}\| = |a| \, \|\boldsymbol{x}\|, \forall \boldsymbol{x}$ and all scalars a
3. $\|\boldsymbol{x}\| = 0$ if and only if $\boldsymbol{x} = 0$
4. Triangle inequality: $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$.



Important norms for vectors

Some important norms for a vector $\underbrace{\boldsymbol{x} = [x_1, \ldots, x_d]^\top}$ are as follows.

- The ℓ_p **norm** is:

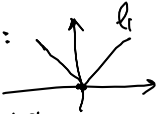
definition

$$\star \quad \|\boldsymbol{x}\|_p := (|x_1|^p + \cdots + |x_d|^p)^{1/p}, = \sqrt[p]{|x_1|^p + \cdots + |x_d|^p}$$

where $\boxed{p \geq 1}$ and $|.|$ denotes the absolute value.

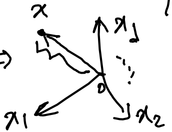
- Two well-known ℓ_p norms are ℓ_1 **norm** and ℓ_2 **norm** (also called the **Euclidean norm**) with $p = 1$ and $p = 2$, respectively:

if $d=1$:



not smooth $\leftarrow \ell_1 \xrightarrow{\text{norm}}$ $$\|\boldsymbol{x}\|_1 := |x_1| + \cdots + |x_d| = \sum_{i=1}^{d} |x_i|,$$

$$\ell_2 \longrightarrow \|\boldsymbol{x}\|_2 := \sqrt{x_1^2 + \cdots + x_d^2} = \underbrace{\sqrt{\sum_{i=1}^{d} x_i^2}}, \Rightarrow$$



$\|x - o\|_2 = \|x\|_2$

- The ℓ_∞ **norm**, also called the **infinity norm**, the **maximum norm**, or the **Chebyshev norm**, is:

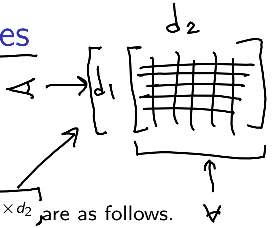
$\|x - y\|_2$ = Euc distance between x & y

$$\|\boldsymbol{x}\|_\infty := \underbrace{\max\{|x_1| \cancel{+}, \cdots \cancel{+}, |x_d|\}}.$$

$|x_1|^2 = x_1^2$

Important norms for matrices

$\mathbb{R}^{d_1 \times d_2 \times d_3}$



$\mathbb{R}^{d_1}$, $\mathbb{R}^{d_2}$ $\to$ $\mathbb{R}^{d_1 \times d_2}$

Some important norms for a matrix $\boldsymbol{X} \in \mathbb{R}^{d_1 \times d_2}$ are as follows.

- The formulation of the **Frobenius norm** for a matrix is similar to the formulation of ℓ_2 norm for a vector:

$\|X\|_F$ similar to $\|x\|_2$

$$\|\boldsymbol{X}\|_F := \sqrt{\sum_{i=1}^{d_1} \sum_{j=1}^{d_2} \boldsymbol{X}_{i,j}^2}, \quad \leftarrow \text{~~ℓ_2-norm of matrix X~~}$$

rows columns

where $\boldsymbol{X}_{ij}$ denotes the (i,j)-th element of $\boldsymbol{X}$.

$\|X\|_2 \neq \|X\|_F$

* note: We also have ℓ_2-norm for matrices with another definition (out-of-scope)

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

$$\|X\|_F = \sqrt{1^2 + 2^2 + 5^2 + 3^2 + 4^2 + 6^2}$$

$$x = \begin{bmatrix} 1 \\ -3 \end{bmatrix} \in \mathbb{R}^{2}$$

$$\|x\|_1 = 1 + 3$$

$$\|x\|_2 = \sqrt{1^2 + 3^2}$$

Quadratic forms using norms

$x = \begin{bmatrix} 1 \\ -3 \end{bmatrix} \Rightarrow \|x\|_2 = \sqrt{1^2 + 3^2}$

$\|x\|_2^2 = 1^2 + 3^2$

$x^\top x = \begin{bmatrix} 1 & -3 \end{bmatrix} \begin{bmatrix} 1 \\ -3 \end{bmatrix} = 1^2 + 3^2$

For $\boldsymbol{x} \in \mathbb{R}^d$ and $\boldsymbol{X} \in \mathbb{R}^{d_1 \times d_2}$, we have:

quadratic

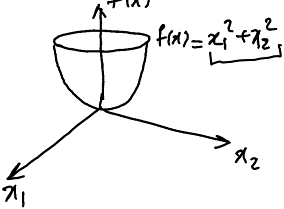
$$\star \quad \|\boldsymbol{x}\|_2^2 = \boldsymbol{x}^\top \boldsymbol{x} = \langle \boldsymbol{x}, \boldsymbol{x} \rangle = \boxed{\sum_{i=1}^{d} x_i^2,}$$

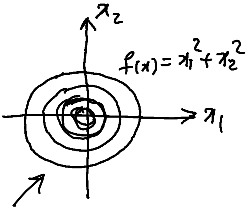
$$\star \quad \underbrace{\|\boldsymbol{X}\|_F^2} = \underbrace{\mathbf{tr}(\boldsymbol{X}^\top \boldsymbol{X})} = \langle \boldsymbol{X}, \boldsymbol{X} \rangle = \underbrace{\sum_{i=1}^{d_1} \sum_{j=1}^{d_2} X_{i,j}^2,}$$

which are convex and in quadratic forms.

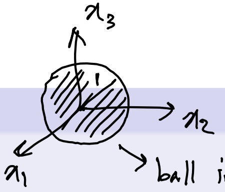
$f(x) = \|x\|_2^2 = x^\top x = x_1^2 + x_2^2$

$d \leq 2$





Unit balls



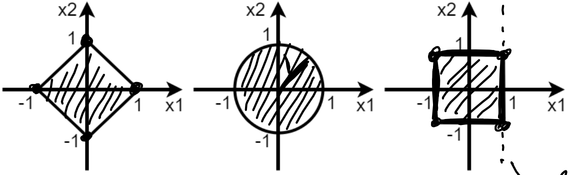
Definition (Unit ball)

The unit ball for a norm $\|\cdot\|$ is:

$$\mathcal{B} := \{\underline{x \in \mathbb{R}^d} \mid \underbrace{\|x\| \leq 1}\}.$$

ball in $\mathbb{R}^3$ for ℓ_2 norm

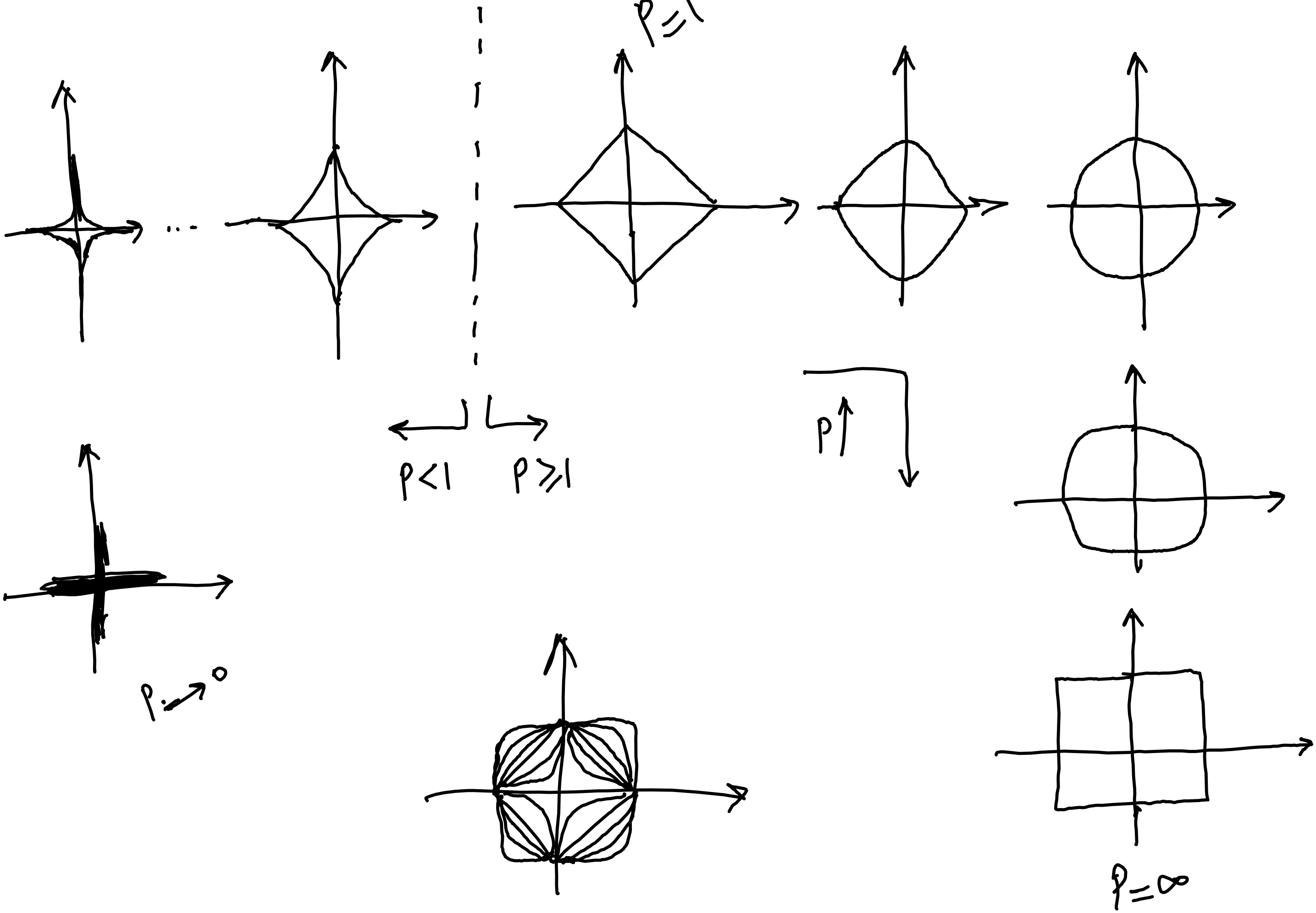
ℓ_p



(a) (b) (c)

The unit balls, in $\mathbb{R}^2$, for (a) ℓ_1 norm, (b) ℓ_2 norm, and (c) ℓ_∞ norm.

$\ell_\infty = \max\{|x_1|, \ldots, |x_p|\}$


p=1
...
p<1
p≥1
p↑
p→0
p=∞

Dual norm

Definition (Dual norm)

Let $\|.\|$ be a norm on $\mathbb{R}^d$. Its dual norm is:

max

norm

dual for that norm

$$\|\boldsymbol{x}\|_* := \sup\{\boldsymbol{x}^\top \boldsymbol{y} \mid \|\boldsymbol{y}\| \leq 1\}. \tag{1}$$

Note that the notation $\|\cdot\|_*$ should not be confused with the the nuclear norm despite of similarity of notations.

$\|x\|_*$ → $\boxed{\text{dual norm}}$ / $\boxed{\text{nuclear norm}}$

Dual norm

Lemma (Hölder's [1] and Cauchy-Schwarz inequalities [2])

Hölder's inequality *states that:*

scalar

$$\boxed{|\boldsymbol{x}^\top \boldsymbol{y}| \leq \|\boldsymbol{x}\|_p \|\boldsymbol{x}\|_q,}$$

$p = 1 \Rightarrow q = \infty$

$\boxed{p = 2 \Rightarrow q = 2}$

$p = \infty \Rightarrow q = 1$

where $p, q \in [1, \infty]$ and p and q satisfy:

$\frac{1}{p} = 1 - \frac{1}{q} = \frac{q-1}{q}$

$$\boxed{\frac{1}{p} + \frac{1}{q} = 1.} \tag{2}$$

*The norms $\|.\|_p$ and $\|.\|_q$ are dual of each other (**dual norms**).*
*A special case of the Hölder's inequality is the **Cauchy-Schwarz inequality**, stated as:*

$$\boxed{|\boldsymbol{x}^\top \boldsymbol{y}| \leq \|\boldsymbol{x}\|_2 \|\boldsymbol{x}\|_2.}$$

ℓ_p, ℓ_q: <u>dual</u> of each other

According to Eq. (2), we have:

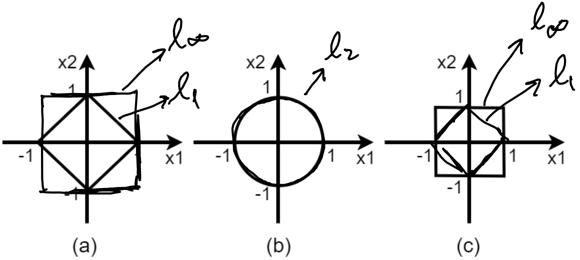
$$\|\cdot\|_p \implies \boxed{\|\cdot\|_* = \|\cdot\|_{p/(p-1)},} \quad \forall p \in [1, \infty].$$

For example, the dual norm of $\|.\|_2$ is $\|.\|_2$ again and the dual norm of $\|.\|_1$ is $\|.\|_\infty$.

Dual norm

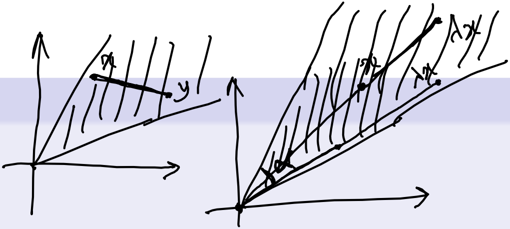
$$\|\cdot\|_p \implies \|\cdot\|_* = \|\cdot\|_{p/(p-1)}, \quad \forall p \in [1, \infty].$$

- The dual of ℓ_2 norm is ℓ_2 norm.
- The dual of ℓ_1 norm is ℓ_∞ norm.
- The dual of ℓ_∞ norm is ℓ_1 norm.



The unit balls, in $\mathbb{R}^2$, for (a) ℓ_1 norm, (b) ℓ_2 norm, and (c) ℓ_∞ norm.

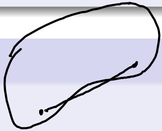
Cone and dual cone



Definition (Cone)

A set $\mathcal{K} \subseteq \mathbb{R}^d$ is a **cone** if:

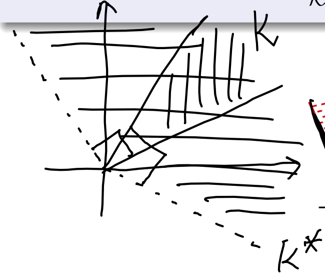
1. it contains the origin, i.e., $\mathbf{0} \in \mathcal{K}$,
2. $\mathcal{K}$ is a convex set,
3. for each $\boldsymbol{x} \in \mathcal{K}$ and $\lambda \geq 0$, we have $\lambda \boldsymbol{x} \in \mathcal{K}$.

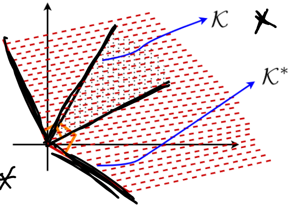


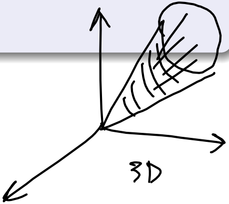
Definition (Dual cone)

The dual cone of a cone $\mathcal{K}$ is:

$$\mathcal{K}^* := \{\boldsymbol{y} \mid \boldsymbol{y}^\top \boldsymbol{x} \geq 0, \forall \boldsymbol{x} \in \mathcal{K}\}.$$





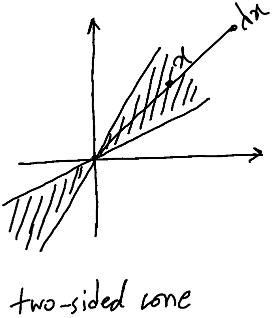


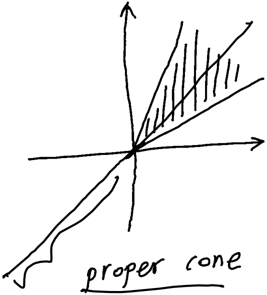
Proper cone

Definition (Proper cone [3])

A convex cone $\mathcal{K} \subseteq \mathbb{R}^d$ is a proper cone if:

1. $\mathcal{K}$ is closed, i.e., it contains its boundary,
2. $\mathcal{K}$ is solid, i.e., its interior is non-empty,
3. $\mathcal{K}$ is pointed, i.e., it contains no line. In other words, it is not a two-sided cone around the origin.





Generalized inequality

$x \geq y \Rightarrow x - y \geq 0$

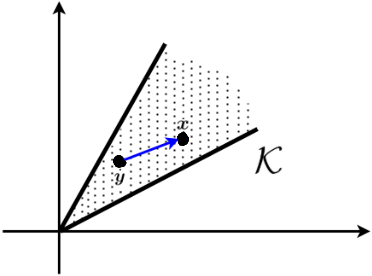
$x \succeq_{K} y \Rightarrow x - y \succeq_{K} 0$

Definition (Generalized inequality [3])

A generalized inequality, defined by a proper cone $\mathcal{K}$, is:

$$\boxed{\boldsymbol{x} \succeq_{\mathcal{K}} \boldsymbol{y}} \iff \boldsymbol{x} - \boldsymbol{y} \in \mathcal{K}.$$

Note that $\boxed{\boldsymbol{x} \succeq_{\mathcal{K}} \boldsymbol{y}}$ can also be stated as $\boldsymbol{x} - \boldsymbol{y} \succeq_{\mathcal{K}} \boldsymbol{0}$.



Important examples for generalized inequality

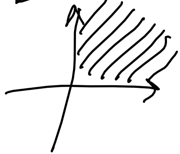
$x = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$ $y = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$

- The generalized inequality defined by the non-negative orthant, $\mathcal{K} = \mathbb{R}^d_+$, is the default inequality for vectors $\boldsymbol{x} = [x_1, \dots, x_d]^\top$, $\boldsymbol{y} = [y_1, \dots, y_d]^\top$:

  every element of x $\geq$ every element y

  $$\boxed{\boldsymbol{x} \succeq \boldsymbol{y}} \iff \boldsymbol{x} \succeq_{\mathbb{R}^d_+} \boldsymbol{y}.$$

  $x \geq y$

  It means component-wise inequality:

  $$\boxed{\boldsymbol{x} \succeq \boldsymbol{y} \iff x_i \geq y_i,} \; \forall i \in \{1, \dots, d\}.$$

- The generalized inequality defined by the positive definite cone, $\mathcal{K} = \mathbb{S}^d_+$, is the default inequality for symmetric matrices $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{S}^d$:

  $\boxed{X - Y \succeq 0}$

  $$\boldsymbol{X} \succeq \boldsymbol{Y} \iff \boldsymbol{X} \succeq_{\mathbb{S}^d_+} \boldsymbol{Y}.$$

  $x \geq y$

  x y

  X Y

  It means $(\boldsymbol{X} - \boldsymbol{Y})$ is positive semi-definite (all its eigenvalues are non-negative).

- If the inequality is strict, i.e. $\boldsymbol{X} \succ \boldsymbol{Y}$, it means that $(\boldsymbol{X} - \boldsymbol{Y})$ is positive definite (all its eigenvalues are positive).
- $\boldsymbol{x} \succeq \boldsymbol{0}$ means all elements of vector $\boldsymbol{x}$ are non-negative and $\boldsymbol{X} \succeq \boldsymbol{0}$ means the matrix $\boldsymbol{X}$ is positive semi-definite.

Preliminaries on Functions

Convex function

x y $0 \le \alpha \le 1$

$\alpha = 0 \rightarrow y$

$\alpha = 1 \rightarrow x$

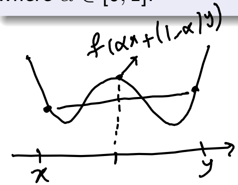
Definition (Convex function)

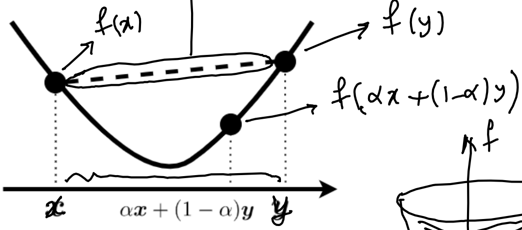
A function $f(.)$ with domain $\mathcal{D}$ is convex if:

for all (any)

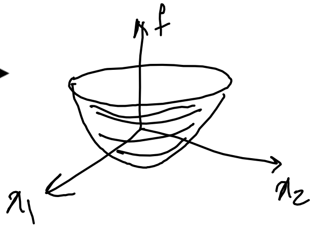
$$f\left(\alpha \mathbf{x} + (1-\alpha)\mathbf{y}\right) \le \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}), \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{D}, \tag{3}$$

where $\alpha \in [0, 1]$.





convex 



If $\geq$ is changed to $\leq$ in Eq. (3), the function is *concave*.

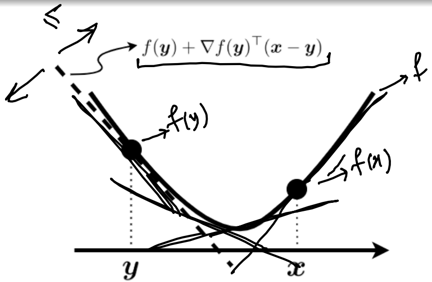
concave 

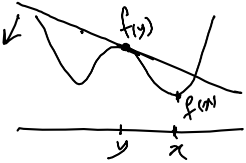
Convex function

Definition (Convex function)

If the function $f(.)$ is differentiable, it is convex if:

$$f(\boldsymbol{x}) \geq f(\boldsymbol{y}) + \nabla f(\boldsymbol{y})^\top (\boldsymbol{x} - \boldsymbol{y}), \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{D}. \tag{4}$$





If $\geq$ is changed to $\leq$ in Eq. (4), the function is *concave*.

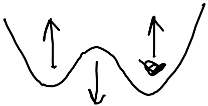
Convex function

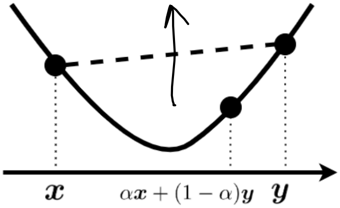
Definition (Convex function)

If the function $f(.)$ is twice differentiable, it is convex if its second-order derivative is positive semi-definite:

$$\boxed{\nabla^2 f(\mathbf{x}) \succeq \mathbf{0},} \quad \forall \mathbf{x} \in \mathcal{D}. \tag{5}$$

$\frac{\partial^2 f}{\partial x^2} \geq 0$





$\frac{\partial^2 f}{\partial x^2} \leq 0$

If $\succeq$ is changed to $\preceq$ in Eq. (5), the function is *concave*.

Strongly convex function

Definition (Strongly convex function)

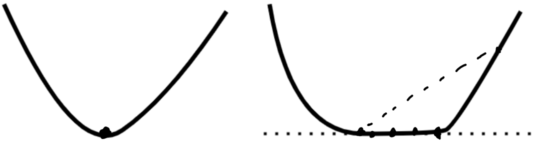
- A differential function $f(.)$ with domain $\mathcal{D}$ is μ-**strongly convex** if:

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \nabla f(\mathbf{y})^\top(\mathbf{x}-\mathbf{y}) + \frac{\mu}{2}\|\mathbf{x}-\mathbf{y}\|_2^2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{D} \text{ and } \mu > 0. \tag{6}$$

- Moreover, if the function $f(.)$ is twice differentiable, it is μ-**strongly convex** if its second-order derivative is positive semi-definite:

$$\boxed{\mathbf{y}^\top \nabla^2 f(\mathbf{x})\mathbf{y} \geq \mu\|\mathbf{y}\|_2^2,} \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{D} \text{ and } \mu > 0. \tag{7}$$

- A strongly convex function has a **unique minimizer**.



strongly convex   convex (but not strongly convex)

Lipschitz smoothness

$\|f'\| = L$

$\leq L$

$\forall x \in D$

Definition (Lipschitz smoothness)

A function $f(.)$ is **Lipschitz smooth** (or **Lipschitz continuous**) if:

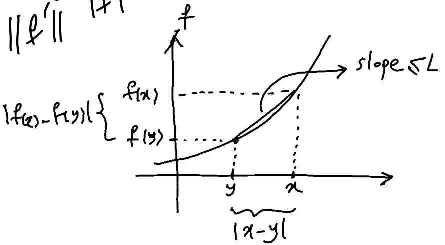
$$\frac{|f(x)-f(y)|}{\|x-y\|_2} \leq L \Leftarrow |f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq L\,\|\boldsymbol{x}-\boldsymbol{y}\|_2, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{D}. \tag{8}$$

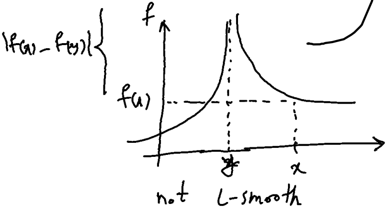
The parameter L is called the **Lipschitz constant**.

A function with Lipschitz smoothness (with Lipschitz constant L) is called L-**smooth**.

Lipschitz smoothness is used in many convergence and correctness proofs for optimization.

$\|f'\|$ $|f'|$





Preliminaries on Optimization

Local and global minimizers

Definition (Local minimizer)

A point $\boldsymbol{x} \in \mathcal{D}$ is a **local minimizer** of function $f(.)$ if and only if:

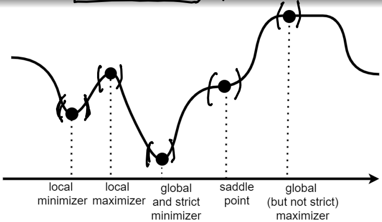
$$\exists \epsilon > 0 : \forall \boldsymbol{y} \in \mathcal{D}, \|\boldsymbol{y} - \boldsymbol{x}\|_2 \leq \epsilon \implies f(\boldsymbol{x}) \leq f(\boldsymbol{y}), \tag{9}$$

meaning that in an ϵ-neighborhood of $\boldsymbol{x}$, the value of function is minimum at $\boldsymbol{x}$.

Definition (Global minimizer)

A point $\boldsymbol{x} \in \mathcal{D}$ is a **global minimizer** of function $f(.)$ if and only if:

$$f(\boldsymbol{x}) \leq f(\boldsymbol{y}), \quad \forall \boldsymbol{y} \in \mathcal{D}. \tag{10}$$



Minimizer in convex function

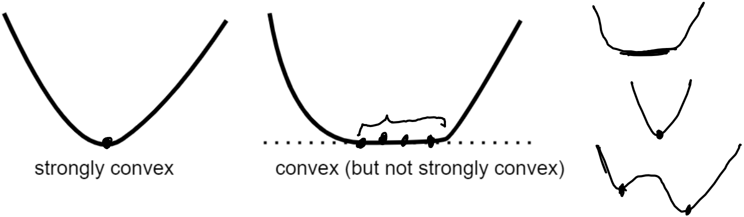
Lemma (Minimizer in convex function)

In a ***convex function****, any local minimizer is a global minimizer. In other words, in a convex function, there exists only* **one** *local minimum value which is the global minimum value.*

Proof.

Proof can be found in the appendix of the tutorial [4]. Please ask me if you have question about it. The proof will not be evaluated in the exam, so please try to understand it rather than memorizing it. □

As an imagination, a convex function is like a multi-dimensional bowl with only one minimum value (it may have several local minimizers but with the same minimum values).



consider any ($\forall$) $y \in D$ and $x, z \in D$

x → local minimizer

z → close to x → have $\|z-x\|_2 \le \varepsilon$

$z \triangleq \alpha y + (1-\alpha)x \quad 0 \le \alpha \le 1$, α small enough to have $\|z-x\|_2 \le \varepsilon$

$\hookrightarrow \varepsilon \ge \|z-x\|_2 = \|\underbrace{\alpha y + (1-\alpha)x}_{z} - x\|_2 = \|\alpha y - \alpha x\|_2 \underset{\alpha \ge 0}{=} \alpha \|y-x\|_2$

$\hookrightarrow \alpha \le \dfrac{\varepsilon}{\|y-x\|_2}$, $\quad 0 \le \alpha \le 1 \quad \rightarrow \quad 0 \le \alpha \le \min\left(\dfrac{\varepsilon}{\|y-x\|_2}, 1\right)$

x : local minimizer $\Rightarrow \exists \varepsilon > 0 : \forall z \in D, \ \|z-x\|_2 \le \varepsilon \Rightarrow f(x) \le f(z)$

f : convex function $\Rightarrow f(z) = f(\alpha y + (1-\alpha)x) \le \alpha f(y) + (1-\alpha) f(x)$

definition of convex function

$\hookrightarrow f(x) \le f(z) \le \alpha f(y) + (1-\alpha) f(x) \Rightarrow f(x) - (1-\alpha) f(x) \le \alpha f(y)$

$\hookrightarrow \alpha f(x) \le \alpha f(y) \Rightarrow f(x) \le f(y) \quad \forall y \in D$

Minimizer in convex function

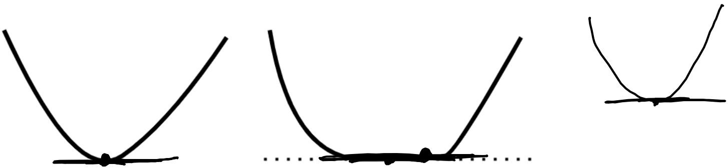
Lemma (Gradient of a convex function at the minimizer point)

When the function $f(.)$ is convex and differentiable, a point $\boldsymbol{x}^*$ is a minimizer if and only if:

$$\boxed{\nabla f(\boldsymbol{x}^*) = \mathbf{0}.}$$

Proof.

Proof can be found in the appendix of the tutorial [4]. Please ask me if you have question about it. The proof will not be evaluated in the exam, so please try to understand it rather than memorizing it. □



strongly convex   convex (but not strongly convex)

* $\underline{\text{side } 1}$: x^* minimizer $\Rightarrow \nabla f(x^*) = 0$

$\forall y \in D$

directional derivative: $\nabla f(x)^T (y-x) = \lim\limits_{t \to 0} \dfrac{f(x+t(y-x)) - f(x)}{t}$

x^* minimizer $\Rightarrow f(x^*) \leq f(\underbrace{x^* + t(y-x^*)}_{\text{neighborhood of } x^* \ (t \to 0)})$

$\Rightarrow 0 \leq \lim\limits_{t \to 0} \dfrac{f(x+t(y-x)) - f(x)}{t} = \nabla f(x^*)^T (y-x^*)$

$\forall y \in D \Rightarrow$ can be any point in D $\Rightarrow$ choose $y = x^* - \nabla f(x^*) \Rightarrow y - x^* = -\nabla f(x^*)$

$\Rightarrow 0 \leq -\nabla f(x^*)^T \nabla f(x^*) = -\|\nabla f(x^*)\|_2^2 \Rightarrow \|\nabla f(x^*)\| = 0$

$\Rightarrow \nabla f(x^*) = 0$

* side 2: $\nabla f(x^*) = 0 \implies x^*$ is minimizer

f: convex function $\implies f(y) \geq f(x^*) + \nabla f(x^*)^T (y - x^*) \quad \forall y \in D$

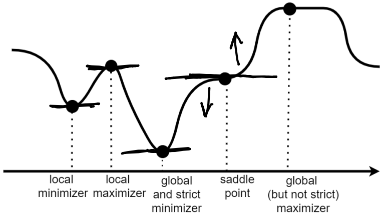
$\nabla f(x^*) = 0$

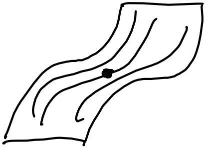
$f(y) \geq f(x^*) \quad \forall y \in D \implies x^*$ is minimizer.

Stationary, extremum, and saddle points

Definition (Stationary, extremum, and saddle points)

- In a general (not-necessarily-convex) function $f(.)$, a point $\boldsymbol{x}^*$ is a **stationary** if and only if $\boxed{\nabla f(\boldsymbol{x}^*) = \mathbf{0}}$.
- By passing through a **saddle point**, the sign of the second derivative flips to the opposite sign.
- **Minimizer** and **maximizer** points (locally or globally) minimize and maximize the function, respectively.
- A **saddle point** is neither minimizer nor maximizer, although the gradient at a saddle point is zero.
- Both minimizer and maximizer are also called the **extremum points**.
- A stationary point can be either a minimizer, a maximizer, or a saddle point of function.





First-order optimality condition

Lemma (First-order optimality condition [5, Theorem 1.2.1])

If $\boldsymbol{x}^*$ is a local minimizer for a differentiable function $f(.)$, then:

$$\boxed{\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}.} \tag{11}$$

Note that if $f(.)$ is convex, this equation is a necessary and sufficient condition for a minimizer.

Proof.

Proof can be found in the appendix of the tutorial [4]. Please ask me if you have question about it. The proof will not be evaluated in the exam, so please try to understand it rather than memorizing it. □

Note

If setting the derivative to zero, $\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}$, gives a closed-form solution for $\boldsymbol{x}^*$, the optimization is done. Otherwise, we should start with some random initialized solution and iteratively update it using the gradient. We will learn first-order and second-order iterative optimization methods for that.

* fundamental theorem of calculus for multivariate function f:

$\forall x, y \in D$

$$f(y) = f(x) + \nabla f(x)^T (y-x) + \underbrace{\int_0^1 \left(\nabla f(x + t(y-x) - \nabla f(x)) \right)^T (y-x)\, dt}_{o(y-x)}$$

x^*: local minimizer $\overset{\text{so}}{\Rightarrow}$ $\exists \varepsilon > 0 : \forall y \in D,\ \|y - x^*\|_2 \leq \varepsilon \Rightarrow$

$f(x^*) \leq f(y)$

* by fundamental theorem of calculus: for multivariate functions:

$$f(y) = f(x^*) + \nabla f(x^*)^T (y - x^*) + o(y - x^*)$$

$$\hookrightarrow f(x^*) \leq f(x^*) + \nabla f(x^*)^T (y - x^*) + \underbrace{o(y - x^*)}_{\text{ignore}}$$

$$\hookrightarrow \nabla f(x^*)^T (y - x^*) \geq 0$$

y: any point in $D \Rightarrow$ choose $y = x^* - \nabla f(x^*)$

$$\hookrightarrow -\nabla f(x^*)^T \nabla f(x^*) = -\|\nabla f(x^*)\|_2^2 \geq 0 \Longrightarrow \|\nabla f(x^*)\|_2 = 0$$

$$\nabla f(x^*) = 0$$

Arguments of optimization

Definition (Arguments of minimization and maximization)

In the domain of function, the point which minimizes (resp. maximizes) the function $f(.)$ is the argument for the minimization (resp. maximization) of function.

The minimizer and maximizer of function are denoted by

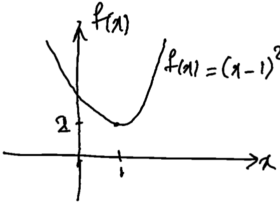
$$\min_{x} f(x)$$

$$\arg\min_{\boldsymbol{x}} f(\boldsymbol{x}), \text{ and}$$

$$\arg\min_{x} f(x)$$

$$\arg\max_{\boldsymbol{x}} f(\boldsymbol{x}),$$

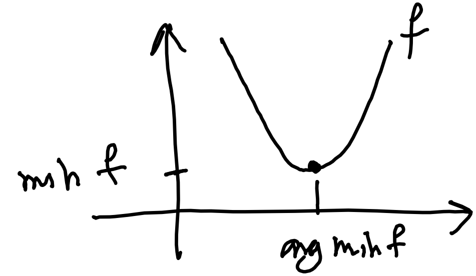
respectively.



$$f(x) = (x-1)^2 + 2 \Rightarrow \arg\min_{x} f(x) = 1$$

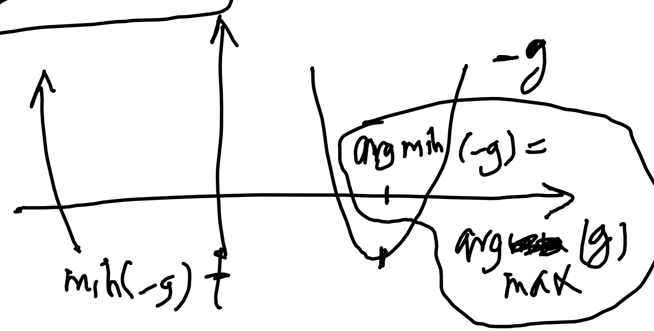
$$\min_{x} f(x) = 2$$

$f \longrightarrow \boxed{\text{minimizer}} \longrightarrow \begin{matrix} \min f \\ \arg\min f \end{matrix}$



I have a function g()
which I want to
<u>maximize</u>

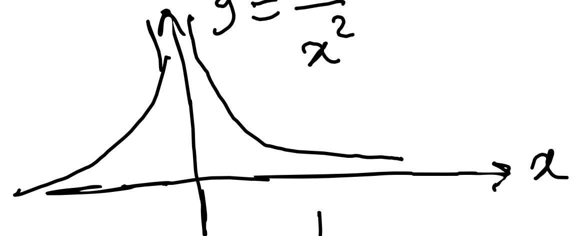
$\boxed{\max(g) = -\min(-g)}$



$\arg\min(-g) = \arg\max(g)$

$f \rightarrow$ minimizer $\rightarrow$ $\min(f)$

$\arg\min(f$

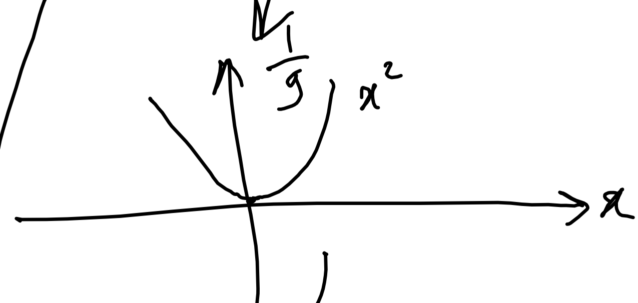


$\max(g) = \infty$

$\arg\max(g) = 0$

$\max(g) = ?$

$\frac{1}{g} \rightarrow$ minimizer $\rightarrow$



$\min(\frac{1}{g}) = 0$

$\arg\min(\frac{1}{g}) = 0$

$$\max(g) = \frac{1}{\min(\frac{1}{g})}$$

$$\arg\max(g) = \arg\min(\frac{1}{g})$$

Converting optimization problems

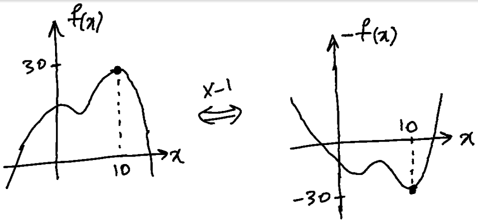
Converting max to min and vice versa

We can convert convert maximization to minimization and vice versa:

$$\underset{\boldsymbol{x}}{\text{maximize}}\;\; f(\boldsymbol{x}) = -\underset{\boldsymbol{x}}{\text{minimize}}\;\; \big(-f(\boldsymbol{x})\big),$$
$$\underset{\boldsymbol{x}}{\text{minimize}}\;\; f(\boldsymbol{x}) = -\underset{\boldsymbol{x}}{\text{maximize}}\;\; \big(-f(\boldsymbol{x})\big).$$

We can have similar conversions for the arguments of maximization and minimization but as the sign of optimal value of function is not important in argument, we do not have the negative sign before maximization and minimization:

$$\arg\max_{\boldsymbol{x}} f(\boldsymbol{x}) = \arg\min_{\boldsymbol{x}} \big(-f(\boldsymbol{x})\big),$$
$$\arg\min_{\boldsymbol{x}} f(\boldsymbol{x}) = \arg\max_{\boldsymbol{x}} \big(-f(\boldsymbol{x})\big).$$



Converting optimization problems

Converting max to min and vice versa

We can convert convert maximization to minimization and vice versa using the reciprocal of cost function:

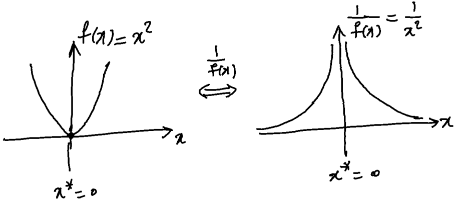
$$\underset{\boldsymbol{x}}{\text{maximize}}\ f(\boldsymbol{x}) = \cancel{\underset{\boldsymbol{x}}{\text{minimize}}\ \frac{1}{f(\boldsymbol{x})}},\quad \frac{1}{\underset{x}{\text{minimize}}\ \frac{1}{f(x)}}$$

$$\underset{\boldsymbol{x}}{\text{minimize}}\ f(\boldsymbol{x}) = \cancel{\underset{\boldsymbol{x}}{\text{maximize}}\ \frac{1}{f(\boldsymbol{x})}}.\quad \frac{1}{\underset{x}{\text{maximize}}\ \frac{1}{f(x)}}$$

We can have similar conversions for the arguments of maximization and minimization:

$$\arg\max_{\boldsymbol{x}} f(\boldsymbol{x}) = \arg\min_{\boldsymbol{x}} \frac{1}{f(\boldsymbol{x})},$$

$$\arg\min_{\boldsymbol{x}} f(\boldsymbol{x}) = \arg\max_{\boldsymbol{x}} \frac{1}{f(\boldsymbol{x})}.$$



Preliminaries on Derivatives

Dimensionality of derivative

- Consider a function $f : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$, $f : \boldsymbol{x} \mapsto f(\boldsymbol{x})$.
- Derivative of function $f(\boldsymbol{x}) \in \mathbb{R}^{d_2}$ with respect to (w.r.t.) $\boldsymbol{x} \in \mathbb{R}^{d_1}$ has dimensionality $(d_1 \times d_2)$.
- This is because tweaking every element of $\boldsymbol{x} \in \mathbb{R}^{d_1}$ can change every element of $f(\boldsymbol{x}) \in \mathbb{R}^{d_2}$. The (i, j)-th element of the $(d_1 \times d_2)$-dimensional derivative states the amount of change in the j-th element of $f(\boldsymbol{x})$ resulted by changing the i-th element of $\boldsymbol{x}$.

Examples

x : ~~scalar~~ vector
f, scalar

- The derivative of a scalar w.r.t. a ~~scalar~~ var is a scalar.
- The derivative of a scalar w.r.t. a vector is a vector.
- The derivative of a scalar w.r.t. a matrix is a matrix.
- The derivative of a vector w.r.t. a vector is a matrix.
- The derivative of a vector w.r.t. a matrix is a rank-3 tensor.
- The derivative of a matrix w.r.t. a matrix is a rank-4 tensor.

$$x = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \qquad f(x) = x^T x$$

$$\begin{bmatrix} 1+\varepsilon \\ 2 \\ 3 \end{bmatrix} \longrightarrow \text{changes } f$$

$$\begin{bmatrix} 1 \\ 2+\varepsilon \\ 3 \end{bmatrix} \longrightarrow \text{changes } f$$

$$\begin{bmatrix} 1 \\ 2 \\ 3+\varepsilon \end{bmatrix} \longrightarrow \text{changes } f$$

$\longrightarrow$ derivative: $\begin{bmatrix} \bigcirc \\ \bigcirc \\ \bigcirc \end{bmatrix}$

$$f(x) = x^T x$$

$x = 1$

$x = 1+\varepsilon \longrightarrow$ changes f $\Longrightarrow$ derivative: scalar

$$x = \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix} \begin{matrix} \to x_1 \\ \to x_2 \\ \to x_3 \end{matrix} \qquad f(x) = \begin{bmatrix} x_1 + x_2 \\ \tan(x_3^2 + x_1^2 + x_2^2) \end{bmatrix} \begin{matrix} \to f_1 \\ \to f_2 \end{matrix}$$

3D 2D

$$\begin{bmatrix} 2+\varepsilon \\ 3 \\ 4 \end{bmatrix} \to \begin{bmatrix} \text{changes } f_1 \\ \text{changes } f_2 \end{bmatrix}$$

$$\begin{bmatrix} 2 \\ 3+\varepsilon \\ 4 \end{bmatrix} \to \begin{bmatrix} \text{changes } f_1 \\ \text{changes } f_2 \end{bmatrix} \to \underbrace{3 \times 2}_{3\times 2} = 6$$

derivative : 6D

$$\begin{bmatrix} 2 \\ 3 \\ 4+\varepsilon \end{bmatrix} \to \begin{bmatrix} \text{changes } f_1 \\ \text{changes } f_2 \end{bmatrix}$$

$$\frac{\partial f(x)}{\partial x} \in \mathbb{R}^{3\times 2}$$

$x = 1$ $f(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix} \to \frac{\partial f(x)}{\partial x} \in \mathbb{R}^2$

$$x = \begin{bmatrix} 1 & 2 & 4 \\ 5 & 6 & 7 \end{bmatrix}$$

x_{11} $\quad$ x_{23}

$2 \times 3 = 6$ values

$$f(x) = \begin{bmatrix} x_{11}^2 \\ \sum_{i=1}^{2} x_{i3}^2 \end{bmatrix}$$

$(2 \times 3) \times 2$

$$\frac{\partial f(x)}{\partial x} \in \mathbb{R}$$

2 values

Dimensionality of derivative

In more details:

- If the function is $f : \mathbb{R} \to \mathbb{R}, f : x \mapsto f(x)$, the derivative $(\partial f(x)/\partial x) \in \mathbb{R}$ is a scalar because changing the scalar $\boldsymbol{x}$ can change the scalar $f(\boldsymbol{x})$.
- If the function is $f : \mathbb{R}^d \to \mathbb{R}, f : \boldsymbol{x} \mapsto f(\boldsymbol{x})$, the derivative $(\partial f(\boldsymbol{x})/\partial \boldsymbol{x}) \in \mathbb{R}^d$ is a vector because changing every element of the vector $\boldsymbol{x}$ can change the scalar $f(\boldsymbol{x})$.
- If the function is $f : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}, f : \boldsymbol{X} \mapsto f(\boldsymbol{X})$, the derivative $(\partial f(\boldsymbol{X})/\partial \boldsymbol{X}) \in \mathbb{R}^{d_1 \times d_2}$ is a matrix because changing every element of the matrix $\boldsymbol{X}$ can change the scalar $f(\boldsymbol{X})$.
- If the function is $f : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}, f : \boldsymbol{x} \mapsto f(\boldsymbol{x})$, the derivative $(\partial f(\boldsymbol{x})/\partial \boldsymbol{x}) \in \mathbb{R}^{d_1 \times d_2}$ is a matrix because changing every element of the vector $\boldsymbol{x}$ can change every element of the vector $f(\boldsymbol{x})$.
- If the function is $f : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_3}, f : \boldsymbol{X} \mapsto f(\boldsymbol{X})$, the derivative $(\partial f(\boldsymbol{X})/\partial \boldsymbol{X})$ is a $(d_1 \times d_2 \times d_3)$-dimensional tensor because changing every element of the matrix $\boldsymbol{X}$ can change every element of the vector $f(\boldsymbol{X})$.
- If the function is $f : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_3 \times d_4}, f : \boldsymbol{X} \mapsto f(\boldsymbol{X})$, the derivative $(\partial f(\boldsymbol{X})/\partial \boldsymbol{X})$ is a $(d_1 \times d_2 \times d_3 \times d_4)$-dimensional tensor because changing every element of the matrix $\boldsymbol{X}$ can change every element of the matrix $f(\boldsymbol{X})$.

Gradient, Jacobian, and Hessian

Definition (Gradient)

Consider a function $f : \mathbb{R}^d \to \mathbb{R}$, $f : \boldsymbol{x} \mapsto f(\boldsymbol{x})$. In optimizing the function f, the derivative of function w.r.t. its variable $\boldsymbol{x}$ is called the **gradient**, denoted by:

$$\boxed{\nabla f(\boldsymbol{x}) := \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} \in \mathbb{R}^d.}$$

Definition (Hessian)

Consider a function $f : \mathbb{R}^d \to \mathbb{R}$, $f : \boldsymbol{x} \mapsto f(\boldsymbol{x})$. The second derivative of function w.r.t. to its derivative is called the **Hessian** matrix, denoted by:

$$\boxed{\boldsymbol{B} = \nabla^2 f(\boldsymbol{x}) := \frac{\partial^2 f(\boldsymbol{x})}{\partial \boldsymbol{x}^2} \in \mathbb{R}^{d \times d}.} \rightarrow \frac{\partial}{\partial x}\left(\underbrace{\frac{\partial f}{\partial x}}_{\mathbb{R}^d}\right)$$

The Hessian matrix is symmetric. If the function is convex, its Hessian matrix is positive semi-definite.

Gradient, Jacobian, and Hessian

Definition (Jacobian)

If the function is multi-dimensional, i.e., $f : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$, $f : \boldsymbol{x} \mapsto f(\boldsymbol{x})$, the gradient becomes a matrix:

$$\boldsymbol{J} := \Big[\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_{d_1}}\Big]^\top = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \dots & \frac{\partial f_{d_2}}{\partial x_{d_1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_1}{\partial x_{d_1}} & \dots & \frac{\partial f_{d_2}}{\partial x_{d_1}} \end{bmatrix} \in \mathbb{R}^{d_1 \times d_2},$$

where $\boldsymbol{x} = [x_1, \dots, x_{d_1}]^\top$ and $f(\boldsymbol{x}) = [f_1, \dots, f_{d_2}]^\top$.

This matrix derivative is called the **Jacobian** matrix.

Technique for calculating derivative

X

According to the size of derivative, we can easily calculate the derivatives. For finding the correct derivative for multiplications of matrices (or vectors), one can temporarily assume some dimensionality for every matrix and find the correct dimensionality of matrices in the derivative.

Example

Let $\boldsymbol{X} \in \mathbb{R}^{a \times b}$, An example for calculating derivative is:

$$\mathbb{R}^{a \times b} \ni \frac{\partial}{\partial \boldsymbol{X}} (\mathbf{tr}(\boldsymbol{AXB})) = \boldsymbol{A}^\top \boldsymbol{B}^\top = (\boldsymbol{BA})^\top. \tag{12}$$

This is calculated as explained in the following.

- We assume $\boldsymbol{A} \in \mathbb{R}^{c \times a}$ and $\boldsymbol{B} \in \mathbb{R}^{b \times c}$ so that we can have the matrix multiplication $\boldsymbol{AXB}$ and its size is $\boldsymbol{AXB} \in \mathbb{R}^{c \times c}$ because the argument of trace should be a square matrix.
- The derivative $\partial(\mathbf{tr}(\boldsymbol{AXB}))/\partial \boldsymbol{X}$ has size $\mathbb{R}^{a \times b}$ because $\mathbf{tr}(\boldsymbol{AXB})$ is a scalar and $\boldsymbol{X}$ is $(a \times b)$-dimensional.
- We know that the derivative should be a kind of multiplication of $\boldsymbol{A}$ and $\boldsymbol{B}$ because $\mathbf{tr}(\boldsymbol{AXB})$ is linear w.r.t. $\boldsymbol{X}$.
- Now, we should find their order in multiplication. Based on the assumed sizes of $\boldsymbol{A}$ and $\boldsymbol{B}$, we see that $\boldsymbol{A}^\top \boldsymbol{B}^\top$ is the desired size and these matrices can be multiplied to each other.

Derivative of matrix w.r.t. matrix

Definition (Kronecker product)

Let $\boldsymbol{A} \in \mathbb{R}^{m_a \times n_a}$ and $\boldsymbol{B} \in \mathbb{R}^{m_b \times n_b}$, and a_{ij} denote the (i,j)-th element of $\boldsymbol{A}$. The Kronecker product of these two matrices is:

$$\mathbb{R}^{(m_a m_b) \times (n_a n_b)} \ni \boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} a_{11}\boldsymbol{B} & \dots & a_{1n_a}\boldsymbol{B} \\ \vdots & \ddots & \vdots \\ a_{m_a 1}\boldsymbol{B} & \dots & a_{m_a n_a}\boldsymbol{B} \end{bmatrix}.$$

Lemma (Derivative of matrix w.r.t. matrix)

The derivative of a matrix w.r.t. another matrix is a tensor. Working with tensors is difficult; hence, we can use **Kronecker product** *for representing tensor as matrix. This is the Magnus-Neudecker convention [6] in which all matrices are vectorized. For example, if* $\boldsymbol{X} \in \mathbb{R}^{a \times b}$, $\boldsymbol{A} \in \mathbb{R}^{c \times a}$, *and* $\boldsymbol{B} \in \mathbb{R}^{b \times d}$, *we have:*

$$\mathbb{R}^{(cd) \times (ab)} \ni \frac{\partial}{\partial \boldsymbol{X}}(\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}) = \boldsymbol{B}^\top \otimes \boldsymbol{A}, \tag{13}$$

where $\otimes$ *denotes the Kronecker product.*

Chain rule

- When having composite functions (i.e., function of function), we use **chain rule** for derivative. Example:

$$\boxed{f(x) = \sqrt{x^3 + x^2 - x + 10} = \sqrt{g(x)},} \quad \boxed{g(x) = x^3 + x^2 - x + 10,}$$

$$\frac{\partial f(x)}{\partial x} = \frac{\partial f(x)}{\partial \cancel{g(x)}} \times \frac{\partial \cancel{g}(x)}{\partial x} = \frac{1}{2\sqrt{g(x)}} \times (3x^2 + 2x - 1) = \frac{3x^2 + 2x - 1}{2\sqrt{x^3 + x^2 - x + 10}}$$

- The chain rule in matrix derivatives is usually stated right to left in matrix multiplications while transpose is used for matrices in multiplication.
- Let $\boxed{\textbf{vec}(.)}$ denote vectorization of a $\mathbb{R}^{a \times b}$ matrix to a $\boxed{\mathbb{R}^{ab} \text{ vector.}}$
- Let $\textbf{vec}^{-1}_{a \times b}(.)$ be de-vectorization of a $\mathbb{R}^{ab}$ vector to a $\mathbb{R}^{a \times b}$ matrix.

$$\begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} \xrightarrow{\text{vec}} \begin{bmatrix} 5 \\ 6 \\ 7 \\ 8 \end{bmatrix} \xrightarrow{\text{vec}^{-1}} \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix}$$

Chain rule

X

Example

$$f(\boldsymbol{S}) = \mathbf{tr}(\boldsymbol{A}\boldsymbol{S}\boldsymbol{B}),\ \boldsymbol{S} = \boldsymbol{C}\widehat{\boldsymbol{M}}\boldsymbol{D},\ \widehat{\boldsymbol{M}} = \frac{\boldsymbol{M}}{\|\boldsymbol{M}\|_F^2},$$

where $\boldsymbol{A} \in \mathbb{R}^{c\times a}$, $\boldsymbol{S} \in \mathbb{R}^{a\times b}$, $\boldsymbol{B} \in \mathbb{R}^{b\times c}$, $\boldsymbol{C} \in \mathbb{R}^{a\times d}$, $\widehat{\boldsymbol{M}} \in \mathbb{R}^{d\times d}$, $\boldsymbol{D} \in \mathbb{R}^{d\times b}$, and $\boldsymbol{M} \in \mathbb{R}^{d\times d}$.

$$\mathbb{R}^{a\times b} \ni \frac{\partial f(\boldsymbol{S})}{\partial \boldsymbol{S}} \overset{(12)}{=} (\boldsymbol{B}\boldsymbol{A})^\top.$$

$$\mathbb{R}^{ab\times d^2} \ni \frac{\partial \boldsymbol{S}}{\partial \widehat{\boldsymbol{M}}} \overset{(13)}{=} \boldsymbol{D}^\top \otimes \boldsymbol{C},$$

$$\mathbb{R}^{d^2\times d^2} \ni \frac{\partial \widehat{\boldsymbol{M}}}{\partial \boldsymbol{M}} \overset{(a)}{=} \frac{1}{\|\boldsymbol{M}\|_F^4}\left(\|\boldsymbol{M}\|_F^2 \boldsymbol{I}_{d^2} - 2\boldsymbol{M}\otimes\boldsymbol{M}\right) = \frac{1}{\|\boldsymbol{M}\|_F^2}\left(\boldsymbol{I}_{d^2} - \frac{2}{\|\boldsymbol{M}\|_F^2}\boldsymbol{M}\otimes\boldsymbol{M}\right),$$

where (a) is because of the formula for the derivative of fraction and $\boldsymbol{I}_{d^2}$ is a $(d^2 \times d^2)$-dimensional identity matrix. finally, by chain rule, we have:

$$\mathbb{R}^{d\times d} \ni \frac{\partial f}{\boldsymbol{M}} = \mathbf{vec}^{-1}_{d\times d}\Big(\big(\frac{\partial \widehat{\boldsymbol{M}}}{\partial \boldsymbol{M}}\big)^\top \big(\frac{\partial \boldsymbol{S}}{\partial \widehat{\boldsymbol{M}}}\big)^\top \mathbf{vec}\big(\frac{\partial f(\boldsymbol{S})}{\partial \boldsymbol{S}}\big)\Big).$$

Optimization Problems

General optimization problem

Consider the function $f : \mathbb{R}^d \to \mathbb{R}$, $f : \boldsymbol{x} \mapsto f(\boldsymbol{x})$. Let the domain of function be $\mathcal{D}$ where $\boldsymbol{x} \in \mathcal{D}, \boldsymbol{x} \in \mathbb{R}^d$.

Definition (Unconstrained optimization)

Unconstrained minimization of a cost function $f(.)$:

$$\boxed{\underset{\boldsymbol{x}}{\text{minimize}} \quad f(\boldsymbol{x}),}$$

where $\boldsymbol{x}$ is called the **optimization variable** and the function $f(.)$ is called the **objective function** or the **cost function**.

General optimization problem

$y_i(x) \le c \Rightarrow \underbrace{y_i(x) - c}_{z_i(x)} \le 0$

Definition (Constrained optimization)

Constrained optimization problem where we want to minimize the function $f(\boldsymbol{x})$ while satisfying m_1 inequality constraints and m_2 equality constraint:

$h_i(x) = c$

$\underbrace{h_i(x) - c}_{g_i(x)} = 0$

$g_i(x) = 0$

$$\left\{ \begin{array}{lll} \underset{\boldsymbol{x}}{\text{minimize}} & f(\boldsymbol{x}) & \\ \text{subject to} & y_i(\boldsymbol{x}) \le 0, & i \in \{1, \ldots, m_1\}, \\ & h_i(\boldsymbol{x}) = 0, & i \in \{1, \ldots, m_2\}. \end{array} \right.$$

$f(\boldsymbol{x})$ is the **objective function**, every $y_i(\boldsymbol{x}) \le 0$ is an **inequality constraint**, and every $h_i(\boldsymbol{x}) = 0$ is an **equality constraint**.

Note

If some of the inequality constraints are not in the form $y_i(\boldsymbol{x}) \le 0$, we can restate them as:

$$\boxed{y_i(\boldsymbol{x}) \ge 0 \implies -y_i(\boldsymbol{x}) \le 0,}$$
$$\boxed{y_i(\boldsymbol{x}) \le c \implies y_i(\boldsymbol{x}) - c \le 0.}$$

Therefore, all inequality constraints can be written in the form $y_i(\boldsymbol{x}) \le 0$.

General optimization problem

Example:

$$\begin{aligned} \underset{\boldsymbol{x}}{\text{minimize}} \quad & x_1 + 3x_2^2 \\ \text{subject to} \quad & 2x_1 - 10x_2 \leq 5, \\ & -2x_1 + 5x_2 \geq 3, \\ & 4x_1 + 10x_2 = 6. \end{aligned}$$

can be converted to:

$$\begin{aligned} \underset{\boldsymbol{x}}{\text{minimize}} \quad & x_1 + 3x_2^2 \\ \text{subject to} \quad & 2x_1 - 10x_2 - 5 \leq 0, \\ & 2x_1 - 5x_2 + 3 \leq 0, \\ & 4x_1 + 10x_2 - 6 = 0. \end{aligned}$$

$\times -1$

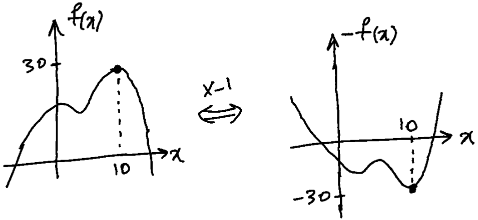
Minimization and maximization

If the optimization problem is a **maximization** problem rather than **minimization**, we can convert it to maximization by multiplying its objective function to -1:

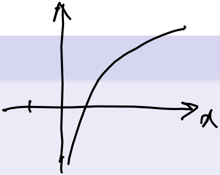
$$\begin{aligned} &\underset{\boldsymbol{x}}{\text{maximize}} \quad f(\boldsymbol{x}) \\ &\text{subject to constraints} \end{aligned}$$

can be converted to:

$$\begin{aligned} &\underset{\boldsymbol{x}}{\text{minimize}} \quad -f(\boldsymbol{x}) \\ &\text{subject to constraints} \end{aligned}$$



Feasible point



Definition (Feasible point)

The point $\boldsymbol{x}$ for the optimization problem:

$$\begin{aligned} \underset{\boldsymbol{x}}{\text{minimize}} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & y_i(\boldsymbol{x}) \leq 0, \; i \in \{1, \ldots, m_1\}, \\ & h_i(\boldsymbol{x}) = 0, \; i \in \{1, \ldots, m_2\}, \end{aligned}$$

$x \in D$

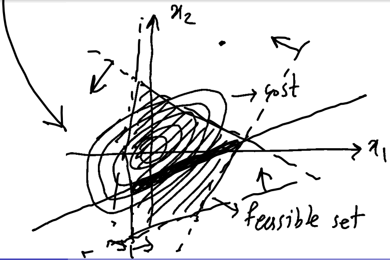
$\downarrow$

$x \in (0, \infty)$

$x > 0$

is feasible if:

$$\begin{aligned} & \boxed{\boldsymbol{x} \in \mathcal{D},} \text{ and} \\ & y_i(\boldsymbol{x}) \leq 0, \quad \forall i \in \{1, \ldots, m_1\}, \text{ and} \\ & h_i(\boldsymbol{x}) = 0, \quad \forall i \in \{1, \ldots, m_2\}. \end{aligned}$$



Constrained optimization with the feasible set

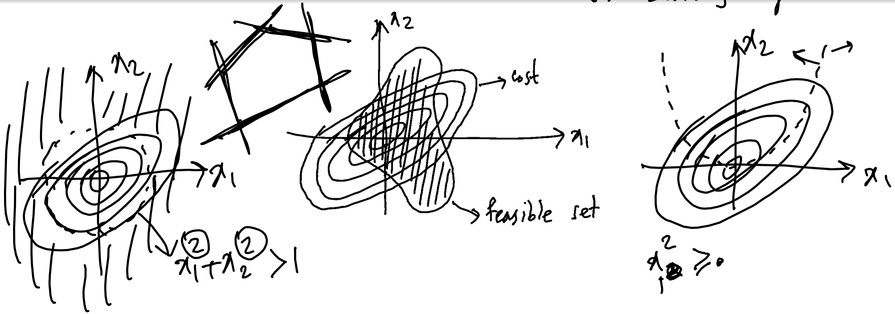
Definition (Constrained optimization)

The **constrained** optimization problem can also be stated as:

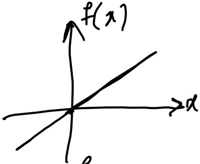
$$\begin{array}{ll} \underset{\boldsymbol{x}}{\text{minimize}} & f(\boldsymbol{x}) \\ \text{subject to} & \boldsymbol{x} \in \mathcal{S}, \end{array}$$

$$\begin{cases} x \in D \\ x \text{ satisfy inequalities} \\ x \text{ satisfy equalities} \end{cases}$$

where $\mathcal{S}$ is the feasible set of constraints.



Convex optimization



A **convex optimization problem** is of the form:

e.g. $x^2 \le 0$

$$
\begin{aligned}
\underset{\boldsymbol{x}}{\text{minimize}} \quad & f(\boldsymbol{x}) \\
\text{subject to} \quad & y_i(\boldsymbol{x}) \le 0, \; i \in \{1, \dots, m_1\}, \\
& \boxed{\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b},}
\end{aligned}
$$

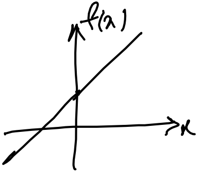
$Ax - b = 0$

$Ax = 0$

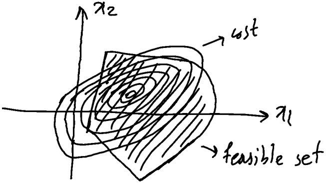
$f(x) = x$

$f(x) = \alpha x$

where the functions $f(.)$ and $y_i(.), \forall i$ are all convex functions and the equality constraints are affine functions.



The feasible set of a convex problem is a convex set.



$f(x) = \alpha x + \beta$

affine

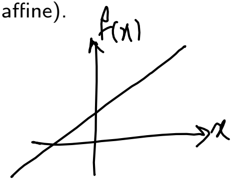
Linear programming

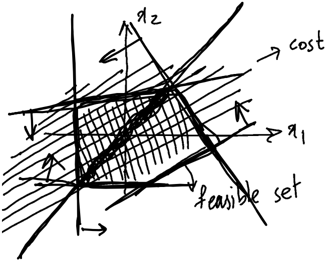
A **linear programming** problem is of the form:

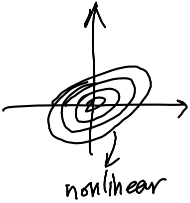
$$\begin{cases} \underset{\boldsymbol{x}}{\text{minimize}} & \boxed{\boldsymbol{c}^\top \boldsymbol{x} + d} \rightarrow \text{cost} \\ \text{subject to} & \boxed{\boldsymbol{G}\boldsymbol{x} \preceq \boldsymbol{h},} \;** \\ & \boxed{\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b},} \;* \end{cases}$$

where the objective function and equality constraints are affine functions.

The feasible set of a linear programming problem is a a polyhedron set while the cost is planar (affine).





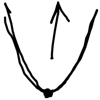


Quadratic programming

$x^2 \leftarrow x^\top x$

$x^\top P x$

$\boxed{P \succ 0}$



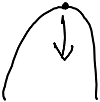
A **quadratic programming** problem is of the form:

$$\begin{aligned} \boxed{\underset{\boldsymbol{x}}{\text{minimize}}} \quad & (1/2)\boldsymbol{x}^\top \boldsymbol{P}\boldsymbol{x} + \boldsymbol{q}^\top \boldsymbol{x} + r \\ \text{subject to} \quad & \boxed{\begin{aligned} &\boldsymbol{G}\boldsymbol{x} \preceq \boldsymbol{h}, \\ &\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \end{aligned}} \end{aligned} \qquad (14)$$

where $\boxed{\boldsymbol{P} \succ \boldsymbol{0}}$ (which is the second derivative of objective function) is a symmetric positive definite matrix, the objective function is quadratic, and equality constraints are affine functions.

The feasible set of a quadratic programming problem is a a polyhedron set while the cost is curvy (quadratic).

$P \prec 0$



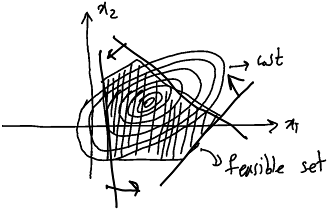
$$\frac{1}{2}x^\top P x + q^\top x + r$$

$$\downarrow f'$$

$$f' = \frac{\cancel{2}Px}{\cancel{2}} + q + 0$$

$$\downarrow f'$$

$$f'' = P + 0$$



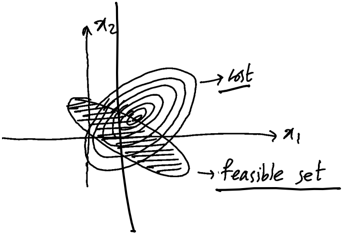
Quadratically constrained quadratic programming

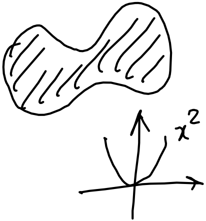
A **Quadratically Constrained Quadratic Programming (QCQP)** problem is of the form:

$$
\begin{aligned}
&\underset{\boldsymbol{x}}{\text{minimize}} \quad \boxed{(1/2)\boldsymbol{x}^\top \boldsymbol{P}\boldsymbol{x} + \boldsymbol{q}^\top \boldsymbol{x} + r} \\
&\text{subject to} \\
&\quad \boxed{(1/2)\boldsymbol{x}^\top \boldsymbol{M}_i \boldsymbol{x} + \boldsymbol{s}_i^\top \boldsymbol{x} + z_i \leq 0,}\ i \in \{1, \dots, m_1\}, \\
&\quad \boxed{\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b},}
\end{aligned}
\tag{15}
$$

where $\boldsymbol{P}, \boldsymbol{M}_i \succ \boldsymbol{0}, \forall i$, the objective function and the inequality constraints are quadratic, and equality constraints are affine functions.

The feasible set of a QCQP problem is intersection of m_1 ellipsoids and an affine set, while the cost is curvy (quadratic).



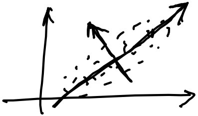


Semidefinite programming

$\mathrm{tr}(3X)$

$X = U\Sigma V$

A **Semidefinite Programming (SDP)** problem is of the form:



matrix

its eigenvalues ≥ 0

$3x$

$3\cancel{x}$

$3\,\mathrm{tr}(X)$

$$\begin{aligned} \underset{X}{\text{minimize}} \quad & \boxed{\mathbf{tr}(\boldsymbol{C}\boldsymbol{X})} \\ \text{subject to} \quad & \boxed{\boldsymbol{X} \succeq \mathbf{0},} \\ & \mathbf{tr}(\boldsymbol{D}_i\boldsymbol{X}) \leq \boldsymbol{e}_i, \quad i \in \{1, \dots, m_1\}, \\ & \mathbf{tr}(\boldsymbol{A}_i\boldsymbol{X}) = \boldsymbol{b}_i, \quad i \in \{1, \dots, m_2\}, \end{aligned} \tag{16}$$

where the optimization variable $\boldsymbol{X}$ belongs to the positive semidefinite cone $\mathbb{S}_+^d$, $\mathbf{tr}(.)$ denotes the trace of matrix, $\boldsymbol{C}, \boldsymbol{D}_i, \boldsymbol{A}_i \in \mathbb{S}^d, \forall i$, and $\mathbb{S}^d$ denotes the cone of $(d \times d)$ symmetric matrices. The trace terms may be written in summation forms. Note that $\mathbf{tr}(\boldsymbol{C}^\top \boldsymbol{X})$ is the inner product of two matrices $\boldsymbol{C}$ and $\boldsymbol{X}$ and if the matrix $\boldsymbol{C}$ is symmetric, this inner product is equal to $\mathbf{tr}(\boldsymbol{C}\boldsymbol{X})$.

Another form for SDP is:

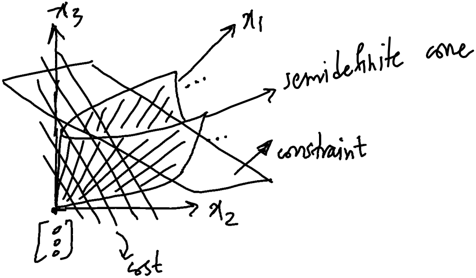
$$\begin{aligned} \underset{\boldsymbol{x}}{\text{minimize}} \quad & \boldsymbol{c}^\top \boldsymbol{x} \\ \text{subject to} \quad & \boxed{\Big(\sum_{i=1}^d x_i \boldsymbol{F}_i\Big) + \boldsymbol{G} \preceq \mathbf{0},} \\ & \boxed{\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b},} \end{aligned} \tag{17}$$

where $\boldsymbol{x} = [x_1, \dots, x_d]^\top$, $\boldsymbol{G}, \boldsymbol{F}_i \in \mathbb{S}^d, \forall i$, and $\boldsymbol{A}$, $\boldsymbol{b}$, and $\boldsymbol{c}$ are constant matrices/vectors.

Semidefinite programming

A **Semidefinite Programming (SDP)** problem is of the form:

$$\begin{array}{lll} \underset{\boldsymbol{X}}{\text{minimize}} & \mathbf{tr}(\boldsymbol{C}\boldsymbol{X}) & \\ \text{subject to} & \boldsymbol{X} \succeq \mathbf{0}, & \\ & \mathbf{tr}(\boldsymbol{D}_i\boldsymbol{X}) \leq \boldsymbol{e}_i, & i \in \{1, \dots, m_1\}, \\ & \mathbf{tr}(\boldsymbol{A}_i\boldsymbol{X}) = \boldsymbol{b}_i, & i \in \{1, \dots, m_2\}. \end{array} \tag{18}$$



Optimization Toolboxes

- All the standard optimization forms can be restated as SDP because their constraints can be written as belonging to some cones; hence, they are **special cases of SDP**.
- The **interior-point method**, or the **barrier method** can be used for solving various optimization problems including SDP [7, 3]. We will learn this method in this course.
- **Optimization toolboxes** such as **CVX** [8] often use interior-point method for solving optimization problems such as SDP.
- The interior-point method is **iterative** and solving SDP is usually **time consuming** especially for large matrices.
- If the optimization problem is a **convex** optimization problem (e.g. SDP is a convex problem), it has only **one** local optimum which is the global optimum.

References

[1] O. Hölder, "Ueber einen mittelwerthabsatz," *Nachrichten von der Königl. Gesellschaft der Wissenschaften und der Georg-Augusts-Universität zu Göttingen*, vol. 1889, pp. 38–47, 1889.

[2] J. M. Steele, *The Cauchy-Schwarz master class: an introduction to the art of mathematical inequalities*.
Cambridge University Press, 2004.

[3] S. Boyd and L. Vandenberghe, *Convex optimization*.
Cambridge university press, 2004.

[4] B. Ghojogh, A. Ghodsi, F. Karray, and M. Crowley, "KKT conditions, first-order and second-order optimization, and distributed optimization: Tutorial and survey," *arXiv preprint arXiv:2110.01858*, 2021.

[5] Y. Nesterov, *Lectures on convex optimization*, vol. 137.
Springer, 2018.

[6] J. R. Magnus and H. Neudecker, "Matrix differential calculus with applications to simple, hadamard, and kronecker products," *Journal of Mathematical Psychology*, vol. 29, no. 4, pp. 474–492, 1985.

[7] Y. Nesterov and A. Nemirovskii, *Interior-point polynomial algorithms in convex programming*.
SIAM, 1994.

References (cont.)

[8] M. Grant, S. Boyd, and Y. Ye, "CVX: Matlab software for disciplined convex programming," 2009.