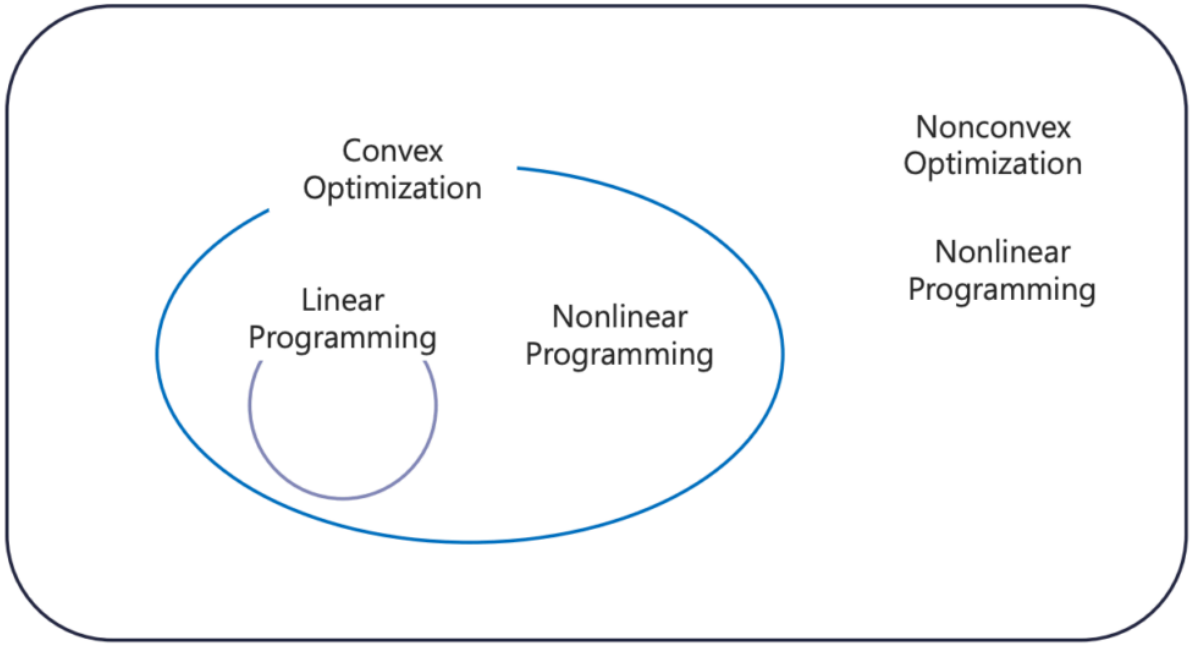
FIRST-ORDER METHODS FOR CONVEX OPTIMIZATION



NOTE : In generally, ML,DL training is **Non convex optimization problem**

and we need to optimization DL but why we learn convex optimization?

1. convex optimization is basic of optimization area

2. Non convex problem can convert convex problem in coditioned situation(**Relaxation**) #Realxation

---

First order method for minimizing a convex function with gradient access

- **Non smooth convex(objective Lipschitz)**
- **Smooth convex(β smooth)**
- **Strongly convex(μ strong convexity)**

We analyze how these first-order methods converge to the optimal solution under different assumptions on f (**non-smooth convex, smooth convex, and strongly convex**), and extend the analysis to convex constrained problems.

Preliminaries

Feasible set

#feasible_Set

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h_i(x) = 0, \quad i = 1, \ldots, p \end{array}$$

f_0 is objective function

optimization goal is find to minimize x , x have to satify **constraints**

Inequality constraibts

$$f_i(x) \leq 0, \quad i = 1, \ldots, m$$

Equality constraints

$$h_i(x) = 0, \quad i = 1, \ldots, p$$

our goal is find optimal value in optimization problem

Thererfore, **feasible set** is

$$\mathcal{D} = \bigcap_{i=1}^{m} \mathrm{dom}(f_i) \cap \bigcap_{i=1}^{p} \mathrm{dom}(h_i)$$

It's the collection of all points that satisfy the rules or constraints of the optimization problem.

In reality, real world have a lot of constraints, so we need to find optimal value in a lot of constraints. That's why we consider feasible set

Feasible set is an essential framework that defines the scope of meaningful solutions for a problem.

Lipschitz continuity

Definition (Lipschitz Continuity)

A real-valued function $f : \mathbb{R}^d \to \mathbb{R}$ is *L-Lipschitz continuous* if there exists a constant $L \geq 0$ such that for all $x, y \in \mathbb{R}^d$,

$$|f(x) - f(y)| \leq L\|x - y\|.$$

This concept is a limit on how **drastically** a function can change : Gradient at any point cannot be greater than L

It ensure the function's behavior is predictable.

Gradient bound : This helps prevent algorithms like gradient descent from **overshooting** the minimum by taking a step that is too large

NOTE : This concept is the absolute value of the slope of the secant line connecting any two points
So, Lipschitz allows for drastic change in slope, it only restricts the magnitude of the average slope between any two points from exceeding L

For example , f = |x| case, x = 0 point,
Although the tangent slope change drastically

from -1 to 1 , average rate of change between any two points never exceeds 1 (L)

In practical, Using Lipschitz Continuity in DL, Prevent Exploading gradient but cause vanishing gradient

NOTE : Differentiablility is not required

However, If differentiable, then

$$\|f()\| \leq L \qquad ^{d}.$$

if differentiable, L is $\mathrm{m}\, f()$
but non differentialble max norm of subgradient

In practical, it is hard to caculate L

Convex Combination and Convex Sets

Definition (Convex Combination)

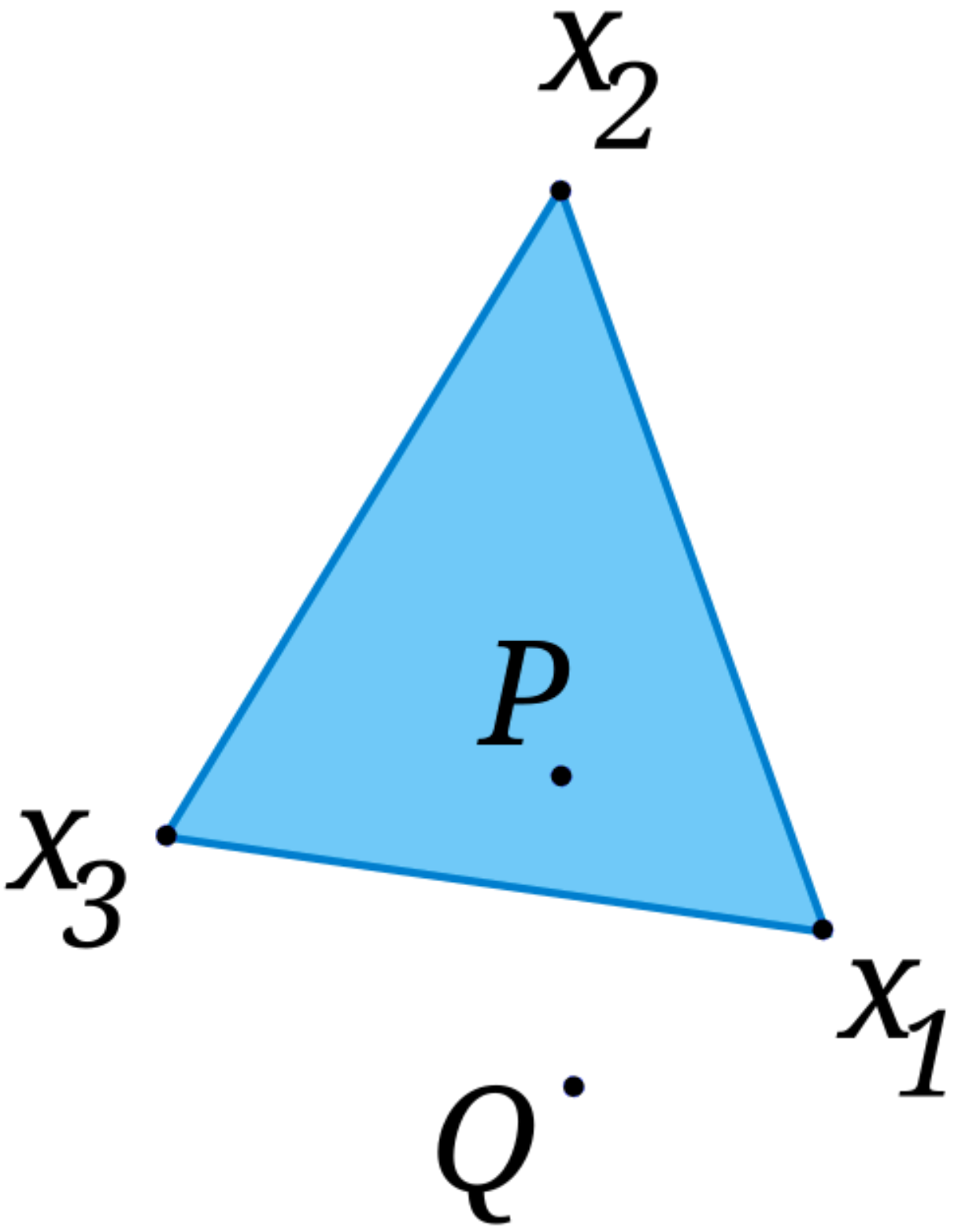
Let $\mathcal{C} = \{x_1, x_2, \ldots, x_n\}$ be a subset of a vector space. A convex combination z of $\mathcal{C}$ is a linear combination of vectors in $\mathcal{C}$ where all coefficients are non-negative and sum to one:

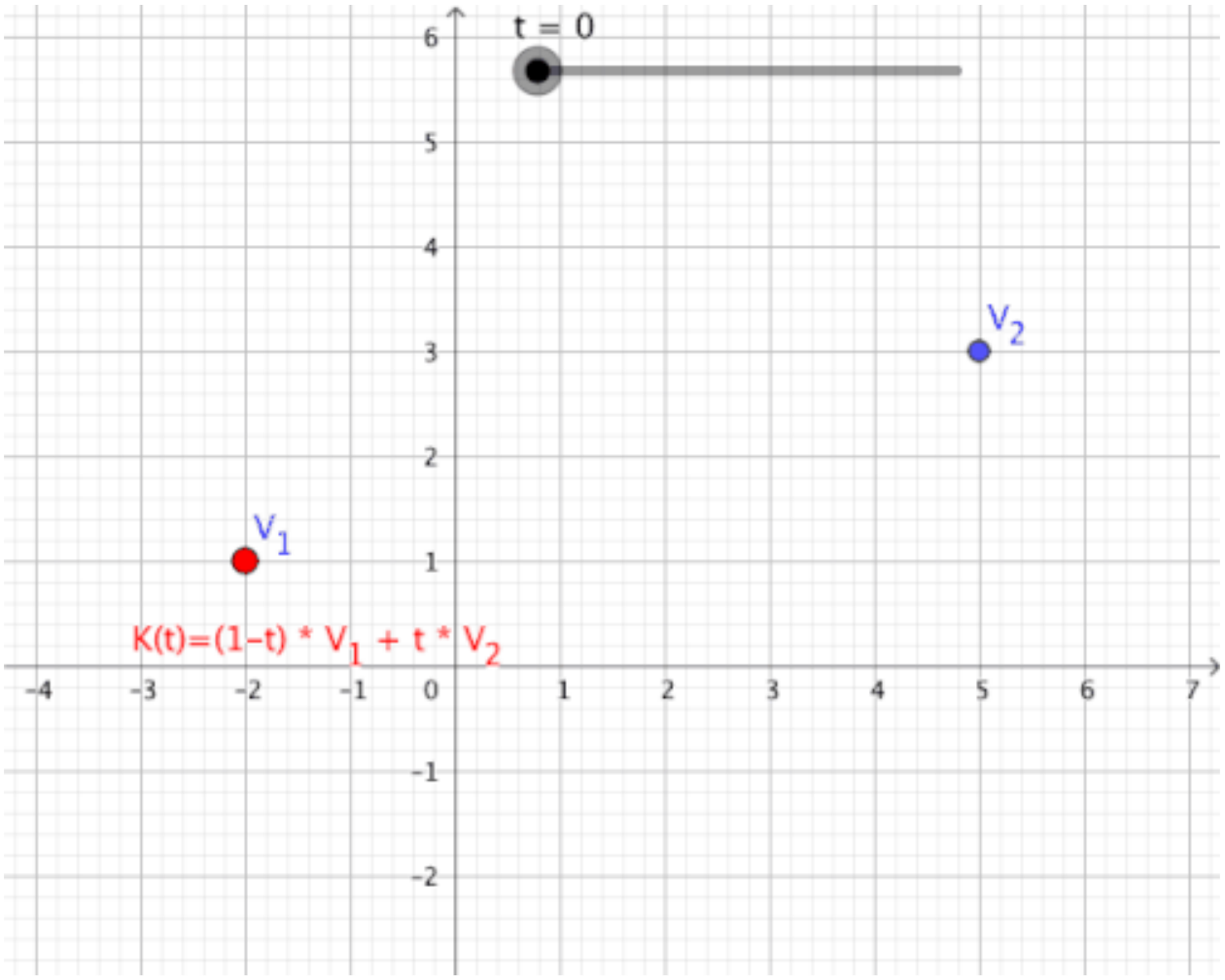
$$z = \sum_{i=1}^{n} \alpha_i x_i, \quad \text{where } \alpha_i \geq 0, \sum_{i=1}^{n} \alpha_i = 1.$$

Definition (Convex Set)

A set $\mathcal{X}$ is convex if the convex combination of any two points in $\mathcal{X}$ is also in $\mathcal{X}$. That is, for all $x, y \in \mathcal{X}$ and $0 \leq \theta \leq 1$,

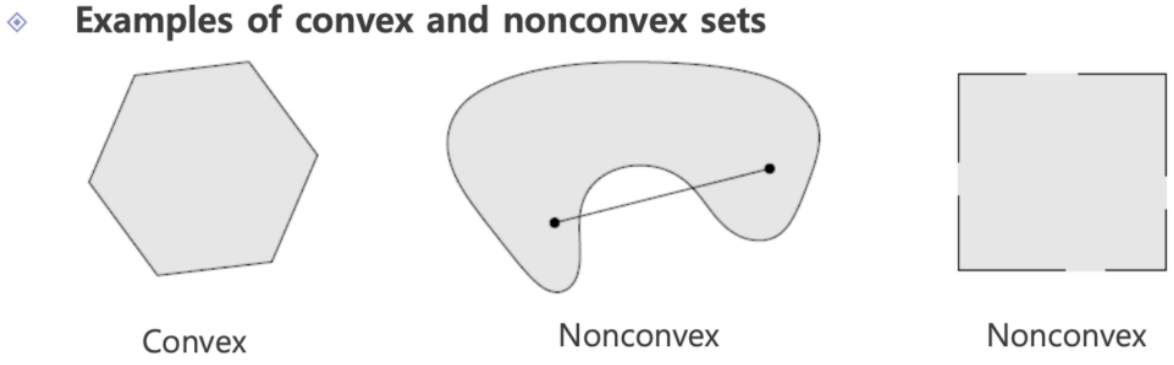
$$\theta x + (1 - \theta) y \in \mathcal{X}.$$


x_2
P
x_3
x_1
Q



Convex combination is a linear combination of points(it can be vectors, scalars)

Every convex combination of two points lies on the line segment between the points



If the convex combination of any two points in X is also in X -> **Convex set**

Examples of Convex Sets

- **Euclidean Ball**

$$\mathcal{B}(x_0, r) := \{x \in \mathbb{R}^d : \|x - x_0\|_2 \leq r\}.$$

Any line segment between two points in the ball remains inside the ball.

- **Affine Subspace**

$$\mathcal{A} := \{x \in \mathbb{R}^d : Ax = b\}.$$

An affine set is convex because the linear constraint is preserved under convex combinations.

- **Probability Simplex**

$$\Delta_d := \left\{ x \in \mathbb{R}^d : x_i \geq 0, \sum_{i=1}^{d} x_i = 1 \right\}.$$

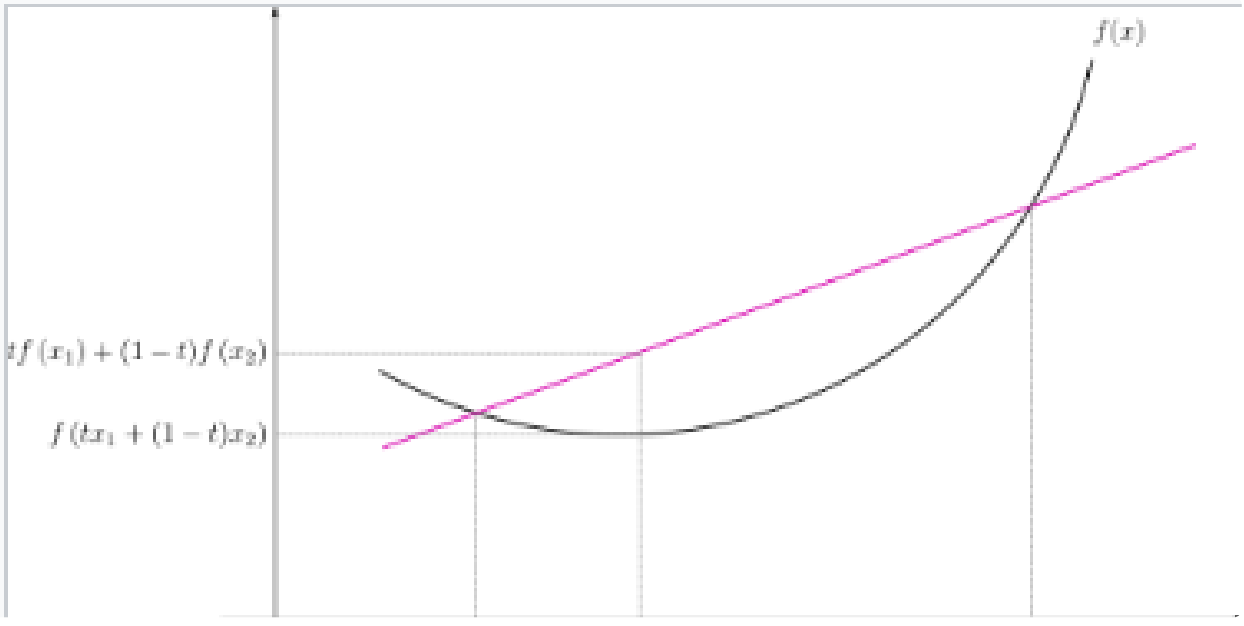
Convex combinations of probability distributions remain valid distributions.

Convex Function

Definition (Convex Function)

Let $\mathcal{X}$ be a convex set. A function $f : \mathcal{X} \to \mathbb{R}$ is convex if for all $x, y \in \mathcal{X}$ and $0 \leq \theta \leq 1$,

$$f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta) f(y).$$



if you pick any two points on the graph of this function and connect them with a straight line, that line will always lie above the function's curve

Why convex function important?

Guarantees that any local minimum we find is also the global minimum

- Convexity can also be characterized by first- or second-order conditions.

Lemma (First-Order Condition)

Let $\mathcal{X}$ be convex. A differentiable function f is convex iff for all $x, y \in \mathcal{X}$,

$$f(y) \geq f(x) + \langle \nabla f(x),\, y - x \rangle.$$

- Geometric view: the tangent plane at any point lies below the graph of f.

Property (Monotone Derivative)

Let $f : \mathcal{X} \to \mathbb{R}$ be convex differentiable, and $x, y \in \mathcal{X}$. Define the one–dimensional slice

$$\phi(t) := f\big(x + t(y - x)\big), \qquad t \in [0, 1].$$

Then, ϕ is differentiable and its derivative

$$\phi'(t) = \big\langle \nabla f(x + t(y - x)),\, y - x \big\rangle,$$

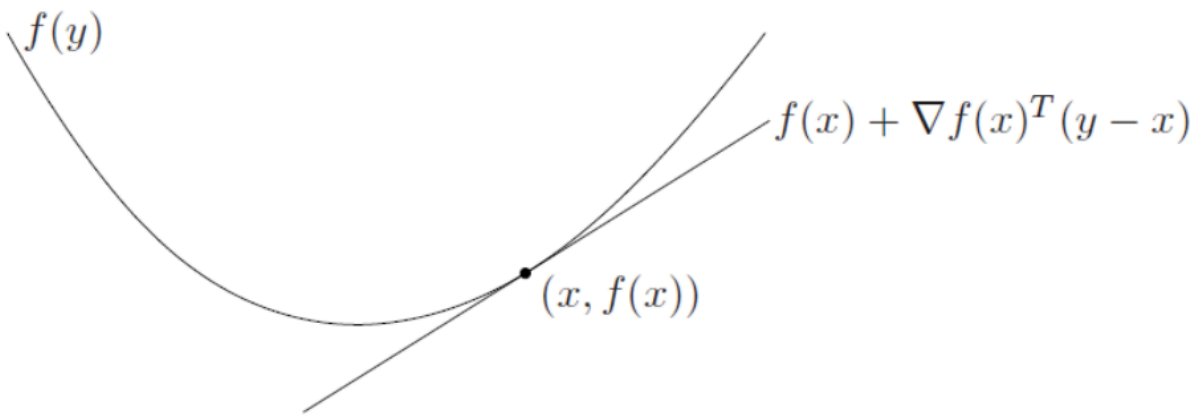
is nondecreasing on $[0, 1]$.

If Convex function and differentiable function, can find lower bound

left hand side is function , right hand side is tangent plan

So, function >= tangent plan

Geometric view : tangent plan at any point lies below the graph of F



monotone derivative is just a property of when convex function saticfy first order condition

: Gradient never decreases

Lemma (Second-Order Condition)

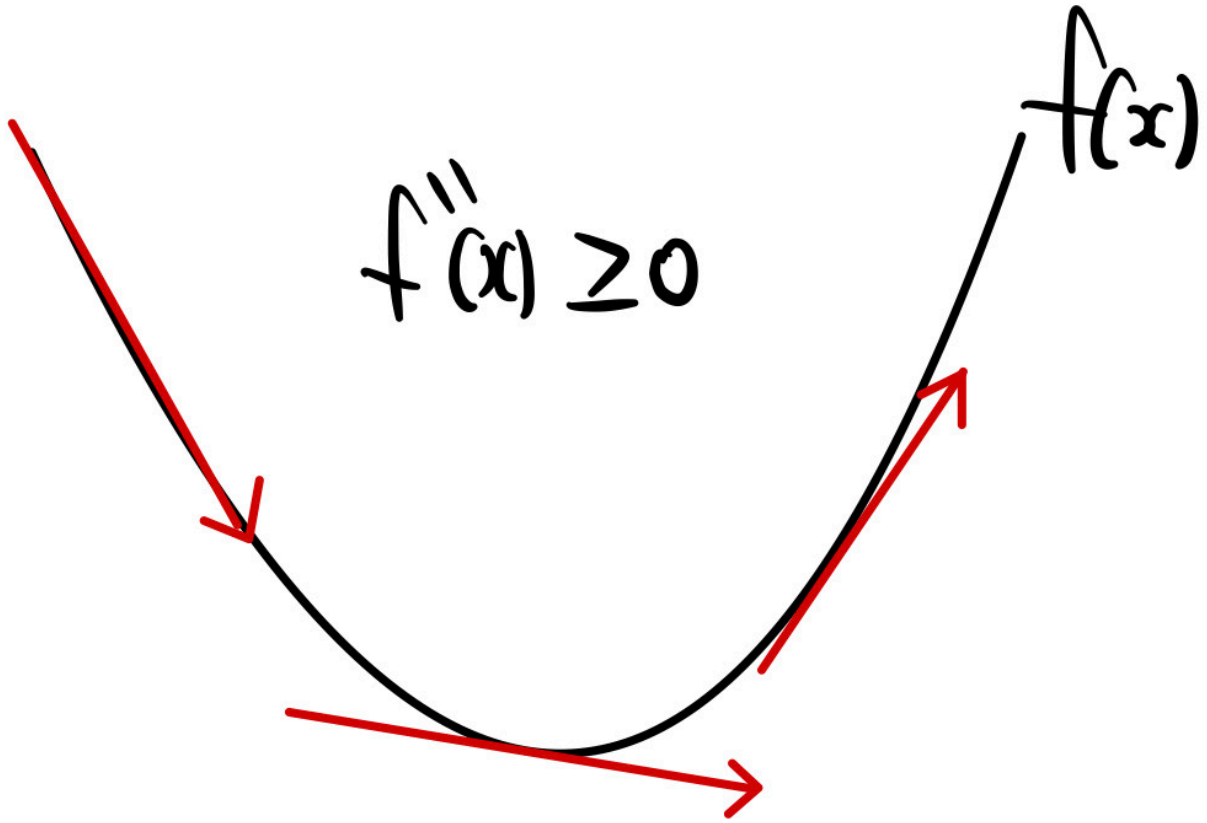
Let $\mathcal{X}$ be convex. A twice-differentiable function f is convex iff for all $x \in \mathcal{X}$,

$$\nabla^2 f(x) \succeq 0.$$

- The Hessian being positive semi-definite means curvature is nonnegative in every direction:

$$\text{for any direction } v \in \mathbb{R}^d, \quad g(t) := f(x + tv) \Rightarrow g''(0) = v^\top \nabla^2 f(x)\, v \geq 0.$$

If twice differentiable and convex funciton , the function graph must always be curving upwards or be flat like a stright line



The Hessian being positive semi definite means curvature is non negative in every direction

Negative Curvature is saddle shape (http://stanwagon.com/wagon/misc/htmllinks/invisiblehandshake_3.html)

Examples of Convex Functions

- **Quadratic Function**

$$f(x) = \tfrac{1}{2}x^\top Qx + b^\top x + c,$$

where $Q \succeq 0$ (positive semi-definite).

  - Hessian: $\nabla^2 f(x) = Q$.
  - Since $Q \succeq 0$, we have $\nabla^2 f(x) \succeq 0$ for all $x \Rightarrow$ convex.

- **Norms** ($p \geq 1$)

$$f(x) = \|x\|_p.$$

  - Triangle inequality, $\|\theta x + (1-\theta)y\|_p \leq \theta\|x\|_p + (1-\theta)\|y\|_p$, satisfies the definition of a convex function.

Subgradient

Definition (Subgradient)

For a convex function $f : \mathcal{X} \to \mathbb{R}$ (not necessarily differentiable), a vector $g \in \mathbb{R}^d$ is called a *subgradient* of f at $x \in \mathcal{X}$ if

$$f(y) \geq f(x) + \langle g, y - x \rangle \qquad \forall y \in \mathcal{X}.$$

The set of all subgradients of f at x is denoted $\partial f(x)$.

- **Intuition:** The subgradient generalizes the role of the gradient to non-smooth convex functions, acting as a "generalized slope" that supports the graph of f from below.
- If f is differentiable at x, the subgradient set collapses to the usual gradient: $\partial f(x) = \{\nabla f(x)\}$.
- For $g \in \partial f(x)$, the affine function $f(x) + \langle g, y - x \rangle$ lies below $f(y)$ for all y.
- Every convex function admits at least one subgradient at every interior point of $\mathcal{X}$.

Subgradient is a generalized gradient for convex function that have sharp, non smooth points

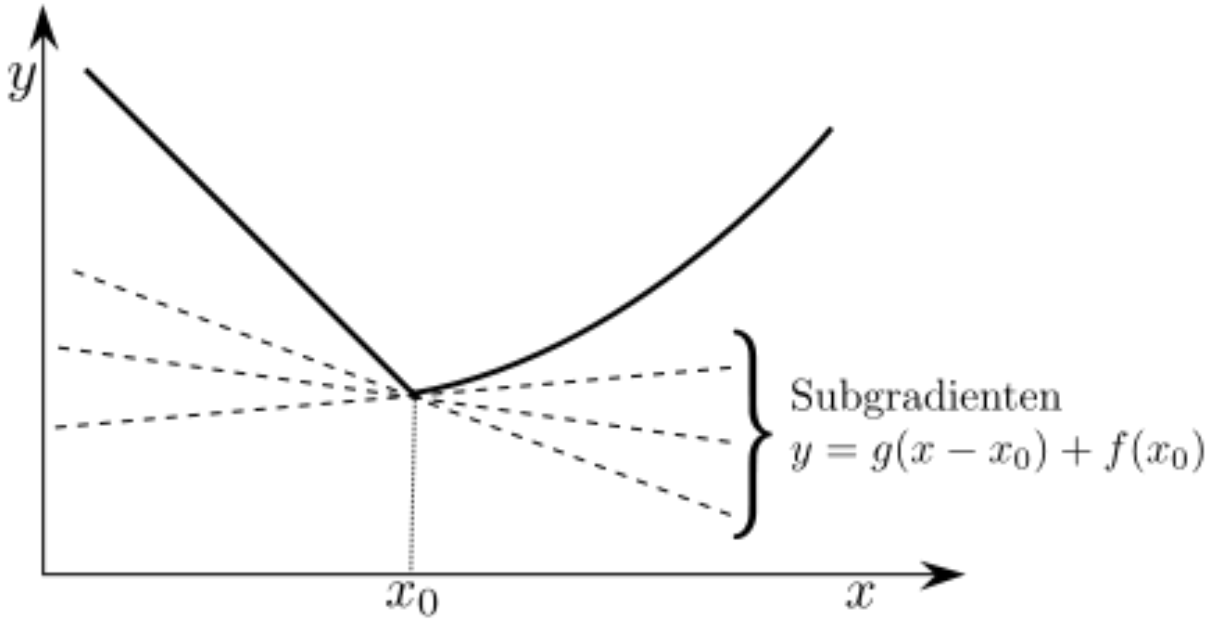
"How can we treat a non-differentiable point as if it were differentiable?" is this: From an

optimization perspective, the most critical role of a gradient is to tell us **'which direction leads to a lower function value.'** The goal is to find a **valid descent direction**.

The subgradient is defined by the following **key inequality**:

$$f() \quad f(x) \ , -x$$

This formula means that the tangent line defined by the subgradient vector always lies below the actual function $f()$.



Note : right side is tangent line

Examples of Subgradients

- **Absolute Value**: $f(x) = |x|$.
  - $x > 0$: $\partial f(x) = \{1\}$.
  - $x < 0$: $\partial f(x) = \{-1\}$.
  - $x = 0$: $\partial f(0) = [-1, 1]$ (every slope g with $-1 \leq g \leq 1$ satisfies $f(y) \geq f(0) + g(y - 0)$).
- **Euclidean Norm** ($\mathbb{R}^d$): $f(x) = \|x\|_2$.
  - $x \neq 0$: $\partial f(x) = \left\{ \frac{x}{\|x\|_2} \right\}$.
  - $x = 0$: $\partial f(0) = \{g \in \mathbb{R}^d : \|g\|_2 \leq 1\}$ (the closed unit ball). Why? g should satisfy

$$\|y\|_2 \geq \langle g, y \rangle.$$

By Cauchy-Schwarz, it becomes

$$\langle g, y \rangle \leq \|g\|_2 \|y\|_2 \leq \|y\|_2.$$

Thus, we need $\|g\|_2 \leq 1$.

Absolute Value: $f(x) = |x|$

This is the simplest yet most illustrative example of the core idea of a subgradient.

- **When x 0 (e.g., $x = 2$):**
  The function is identical to $f(x) = x$, so this segment is a smooth line. Therefore, the gradient is uniquely
  1. $f(x) = 1$
- **When x 0 (e.g., $x = -$):**
  The function is identical to $f(x) = -x$, which is also a smooth line. The gradient is uniquely

1. $f(x) = -1$

- **When $x = 0$ (The sharp point):**
  This is the key non-differentiable point. Here, the subgradient becomes the interval $-1, 1$, which includes all values between -1 and 1.

Why? If we find all slopes that satisfy the subgradient definition ($f()$ $f(x)$ $(- x)$):

- **For positive :** 1
- **For negative :** $-$ $-1 \leq$ (the inequality sign flips when dividing by a negative number)

The range of that satisfies both conditions is $-1 \leq \leq 1$. Intuitively, at the sharp "V" shape, any tangent line that "supports" the graph from below must have a slope between -1 and 1.

Euclidean Norm: $f() = \|\|_2$

You can think of this as the multi-dimensional version of the absolute value function. It represents the distance from the origin.

- **When 0 (The smooth part):**
  The function is differentiable at these points, and its gradient is a **unit vector (a vector of length 1) pointing in the same direction as the vector** .
  - $f() = \|\|_2$
- **When $= 0$ (The sharp point):**
  The origin in a multi-dimensional space is a sharp point, just like the origin for the absolute value function. Here, the subgradient is the **closed unit ball, which is the set of all vectors with a magnitude less than or equal to 1**.
  - $f(0) = \ ^d \ \|\|_2 \leq 1$

Why? The definition of the subgradient ($\|\|_2$,) must be satisfied for all vectors . Using the Cauchy-Schwarz inequality (, $\leq \|\|_2\|\|_2$), we

can conclude that we need $\|\|_2\|\|_2$ to always be less than or equal to $\|\|_2$. The only way for this to be true is if $\|\|_2 \leq 1$.

Intuitively, at the tip of a cone shape like $= x^2\ ^2$, the slope of any plane that supports the cone from below cannot exceed 1 in any direction.

Projection onto convex set

- For a closed convex set $\mathcal{C} \subset \mathbb{R}^d$, the **Euclidean projection** of a point $z \in \mathbb{R}^d$ onto $\mathcal{C}$ is

$$\Pi_{\mathcal{C}}(z) := \arg\min_{x \in \mathcal{C}} \|x - z\|.$$

- Intuition: $\Pi_{\mathcal{C}}(z)$ is the point in $\mathcal{C}$ closest to z.

Property I

For any $x \in \mathcal{C}$ and $z \in \mathbb{R}^d$,

$$\langle x - \Pi_{\mathcal{C}}(z),\ z - \Pi_{\mathcal{C}}(z) \rangle \leq 0.$$

Property II

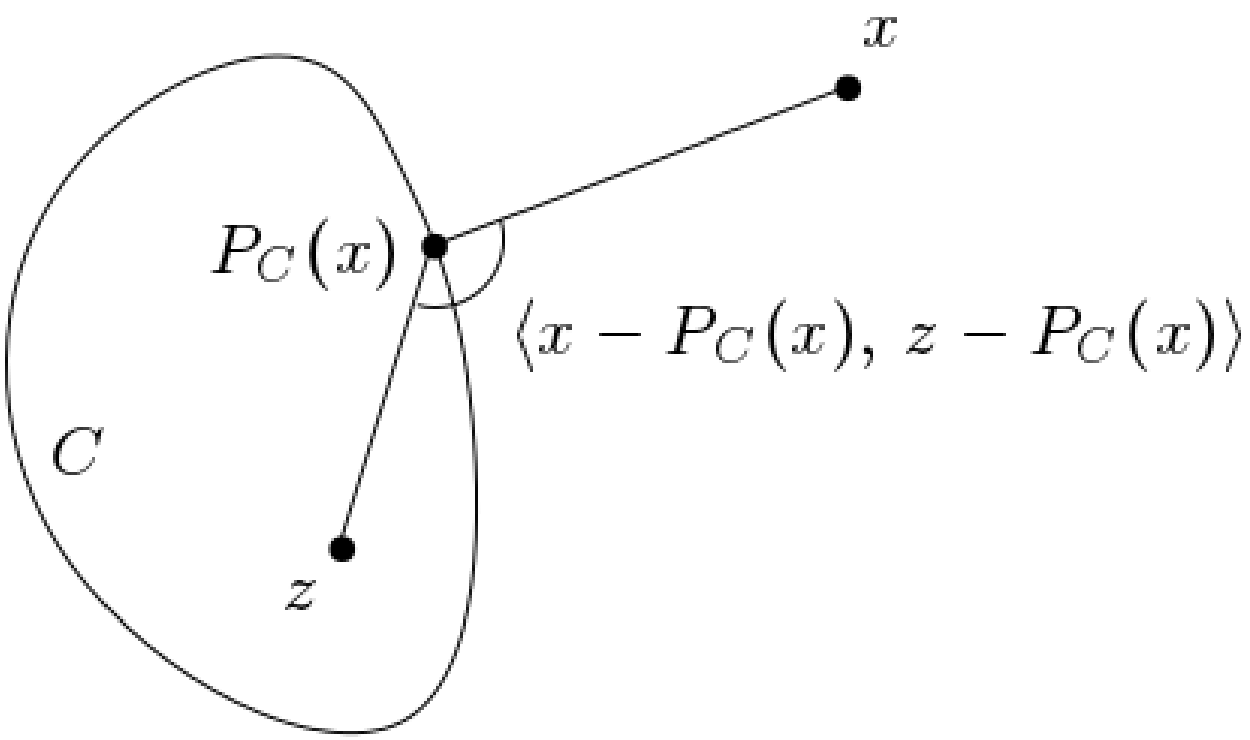
For any $x \in \mathcal{C}$ and $z \in \mathbb{R}^d$,

$$\|\Pi_{\mathcal{C}}(z) - x\| \leq \|z - x\|.$$

- These properties are fundamental in analyzing **projected gradient methods**.

Property I

Inner product of the two vector is less than or equal to zero, which geometrically implies that the **angle between the two vectors is 90 degress or more**



NOTE : x is define in convex set C, z is out of Convex set

Note

positive Inner product : less than 90 degree, same direction, large similarity

Negative Inner product : more than 90 degree, oposite direction, small similarity

NOTE : this propery, maybe use in Safety RL : policy must be in safety set
OR, constrainted policy optimization

Property II

정사영 한 거리가 원래 점에서 정사영하려는 점의 직선보다 작다

Frist order optimality(constrainted case)

First-Order Optimality (Constrained Case)

Let $\mathcal{X}$ be a convex set and let f be a differentiable convex function on $\mathcal{X}$. Then $x^* \in \arg\min_{x \in \mathcal{X}} f(x)$ if and only if

$$\langle \nabla f(x^*), x^* - y \rangle \leq 0 \qquad \forall y \in \mathcal{X}.$$

- Intuition: the gradient at x^* points outward or is orthogonal to all feasible directions, so no descent direction exists. That is,

$$\langle \nabla f(x^*), x^* - y \rangle = \left\langle \underbrace{-\nabla f(x^*)}_{\text{descent direction}}, \underbrace{y - x^*}_{\text{feasible direction}} \right\rangle \leq 0.$$

In **Constrained convex optimization problem.**

This condition is a test to determine if you have found the optimal solution(**Is this really the lowest point? with constrined case?**)

In a simple unconstrained problem, the optimal solution is at the very bottom of a bowl,where the gradient is zero. But what if our search area is limited?

This mean we need to find optimal value in **feasible set**

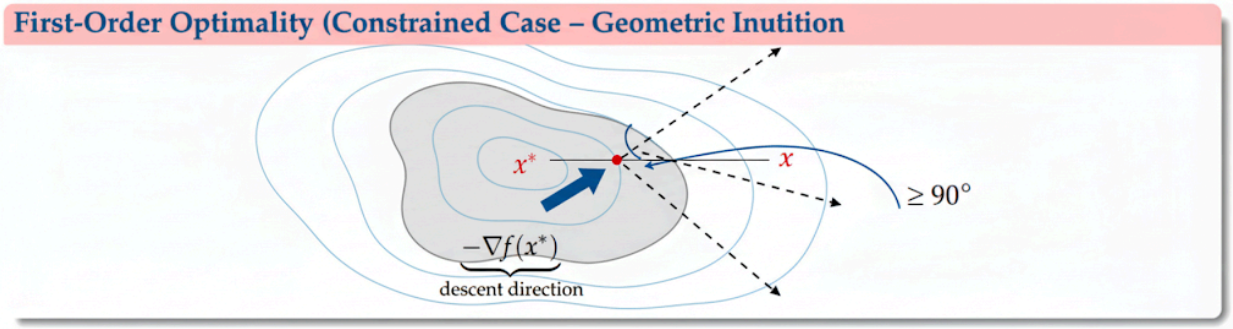
Note

- $\langle \nabla f(x^*), \text{direction} \rangle > 0 \Rightarrow$ That direction is **uphill**.
- $\langle \nabla f(x^*), \text{direction} \rangle < 0 \Rightarrow$ That direction is **downhill**.
- $\langle \nabla f(x^*), \text{direction} \rangle = 0 \Rightarrow$ That direction is **level** (along a contour line).

equally == 0 : minimum point (unconstraint case)

uphill : constraint case, The true minimum(where slope=0) might exist outside region
so,we can say there are no downhill paths, conclude that our current position is the lowest point within accessible region - optimal point

downhill : we can minimize more



Geometric Intuition: The descent direction $-\Lambda f(x^*)$ points outwards (o all feasible directions from x^* frons from x^*, meaning no movement into x in decrease f

blue arrow is **descent direction**, dot line is **Feasible direction**

The optimal value x : saticfying constraints, can't obtain lower function value point.

Descent direction : A direction from the current point that decreases the function value

Feasible direction : A direction from the current point in which one can move without violating the constraint

NOTE : less than 90 degree? it is **inconsistency** , x is optimal point, The mean of less than 90 degree, it has more optimal point than x

---

Unconstrained Case

convex function & L Lipschitz

Theorem (Subgradient Method Convergence)

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function that is L-Lipschitz. Consider the subgradient method

$$x_{k+1} = x_k - \eta g_k, \qquad g_k \in \partial f(x_k).$$

Let $R := \|x_0 - x^*\|$ where $x^* \in \arg\min f$. With constant step size

$$\eta = \frac{R}{L\sqrt{T}},$$

the averaged iterate

$$\bar{x}_T := \frac{1}{T}\sum_{k=0}^{T-1} x_k$$

satisfies

$$f(\bar{x}_T) - f(x^*) \leq \frac{LR}{\sqrt{T}}.$$

- Convergence rate: $\mathcal{O}(1/\sqrt{T})$.

NOTE : we can decide optimal step size in the condition

L Lipschitz is not require differentiable, so use Subgradient

Convergence rate is very slow convergence rate.

R : it is a distance initial point x_0 to x
but we don't know what is the optimal so, in practially, we don't know **R**

T : Total number of iteration (Time horizon)

마지막 식은 특히 T 무한까지 진행하면 반드시 수렴한다는것을 보장

즉 convex하고 L Lipstchiz 만족하면 반드시 최적의 답 찾을수있다는 이론적 근거 있다

Convex function & β smooth

Definition (β-Smoothness)

A continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is β-*smooth* if its gradient is β-Lipschitz:

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\| \quad \forall x, y.$$

- If f is twice differentiable, $\nabla^2 f(x) \preceq \beta I$ for all x.

Beta smooth describes how smoothly the function slope changes, There are no sudden sharp turns

Beta smooth must be continuously **differentiable**

if continuously differentiable function is beta smooth, and Function is twice differentiable, then it has lower bound

If convex function and beta smooth,

Lemma

If f is β-smooth,

$$f(x) \leq f(y) + \nabla f(y)^\top (x-y) + \frac{\beta}{2}\|x-y\|^2.$$

Descent Lemma

If f is convex and β-smooth,

$$0 \leq f(x) - f(y) - \nabla f(y)^\top (x-y) \leq \frac{\beta}{2}\|x-y\|^2.$$

This implies

$$f\left(x - \tfrac{1}{\beta}\nabla f(x)\right) \leq f(x) - \frac{1}{2\beta}\|\nabla f(x)\|^2.$$

Lemma

For any x, y,

$$f(x) - f(y) \leq \nabla f(x)^\top (x-y) - \frac{1}{2\beta}\|\nabla f(x) - \nabla f(y)\|^2.$$

Convex function and beta smooth Gradient descent

Theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and β-smooth. Gradient descent

$$x_{k+1} = x_k - \eta \nabla f(x_k),$$

with step size $\eta = \frac{1}{\beta}$ satisfies

$$f(x_T) - f(x^*) \leq \frac{2\beta}{T}\|x_0 - x^*\|^2,$$

where $x^* \in \arg\min f$.

- Convergence rate: $\mathcal{O}(1/T)$

NOTE : we can decide optimal step size in the condition

Beta smooth saticfy differentiable so it can use Gradient descent

Convergence rate is faster than the L Lipschitz case

L Lipschitz & Alpha strong

Definition (α-Strong Convexity)

A function $f : \mathbb{R}^d \to \mathbb{R}$ is *α-strongly convex* if

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\alpha}{2}\|y - x\|^2 \qquad \forall x, y \in \mathbb{R}^d.$$

- Equivalent conditions:
  - $x \mapsto f(x) - \frac{\alpha}{2}\|x\|^2$ is convex.
  - (Strong monotonicity) $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \alpha \|x - y\|^2$.
  - If f is twice differentiable, $\nabla^2 f(x) \succeq \alpha I$.
- If g is convex and h is α-strongly convex, then $g + h$ is also α-strongly convex.

Alpha strongly convexity : Note in convex function local minimum is global minimum : has one mod

Some General convex function The bottom of the valley can be very wide and **flat**. In this case, there could be multiple minimum points (f(x)=max(0,|x|-1))

Strongly convex function does not allow for a flat bottom, The requirement that the function must have a minimum curvature forces the bottom of the valley to be a single sharp point, forcing the minimum to be a single, unique point

Summary, alpha strongly convex function is like a U shaped valley
There are no flat regions, allowing the algorithm to descend to the bottom quickly

This always a guaranteed gap between the functions and its tangent line, and a quadratic function can fit into this gap

alpha represent the size of the gap(볼록성의 강도)

Satisfies alpha strongly convex condition

- Guarantees a Unique minimum

- Faster convergence

Example : L2 Regularization

$$2\| - x\|^2$$

This part is the core of **strong convexity**

Definition L2 Regualization

In the context of linear models $f_\theta(x) = \theta x$, a widely used approach is L2 regularization, which defines the following objective:

$$(\theta) = 1 \quad _{i=1} L(^{(i)}, \theta x^{(i)}) \;\; 2 \;\; \|\theta\|_2^2.$$

Compare L2 regualization and strong convexity , it is the same formula!

Theorem (Subgradient Method)

Consider the subgradient method

$$x_{k+1} = x_k - \eta_k g_k, \qquad g_k \in \partial f(x_k),$$

where f is α-strongly convex and L-Lipschitz. With step size

$$\eta_k = \frac{2}{\alpha(k+1)},$$

the weighted average

$$\bar{x}_T := \sum_{k=0}^{T-1} \frac{2k}{T(T+1)} x_k$$

satisfies

$$f(\bar{x}_T) - f(x^*) \leq \frac{2L^2}{\alpha(T+1)}.$$

- Convergence rate: $\mathcal{O}(1/T)$ without smoothness.

NOTE : we can decide optimal step size in the condition

Alpha strongly convex and L Lipschitz condition : it shows fast convergence rate, similar beta smooth case, Even if not smooth

Note

This concept can ve applied regardless of whether a functino is differentiable or non differentiable

Differentiable case : it is defined by an inequality using the gradient, which describes how much faster the functino increases compared to its tangent line

Non differentiable case : it is defined by whether the function remains convex even after subtracting the quadratic function.
In this case the subgradient is used for optimization instead of the gradient

Alpha Strong convexity & Beta smooth

Lemma

If f is β-smooth and α-strongly convex on $\mathbb{R}^d$, then for all $x, y \in \mathbb{R}^d$,

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{\alpha\beta}{\alpha+\beta}\|x-y\|^2 + \frac{1}{\alpha+\beta}\|\nabla f(x) - \nabla f(y)\|^2.$$

- This inequality expresses the *strong monotonicity and co-coercivity* of the gradient mapping when f is both strongly convex and smooth.
- It quantitatively couples the point difference $\|x-y\|$ and the gradient difference $\|\nabla f(x) - \nabla f(y)\|$, ensuring that the gradient grows and aligns with $x - y$ in a controlled way.

Theorem (Gradient Descent Convergence)

Let f be α-strongly convex and β-smooth, and define the condition number $\kappa := \beta/\alpha$. Consider gradient descent

$$x_{k+1} = x_k - \gamma \nabla f(x_k)$$

with step size

$$\gamma = \frac{2}{\alpha + \beta}.$$

Then the last iterate satisfies

$$f(x_T) - f(x^*) \leq \frac{\beta}{2} \exp\left(-\frac{4T}{\kappa + 1}\right) \|x_0 - x^*\|^2.$$

- Linear (geometric) convergence rate $\mathcal{O}\left(e^{-4T/(\kappa+1)}\right)$.

NOTE : we can decide optimal step size in the condition

This is **most ideal case** : our goal is make function like this!!

Therefore, The convergence rate is very fast, exponential convergence rate

---

hessian 구하고 PSD 조건으로 eigen value를 determinant 계산으로 구한다
고유값은 곡률의 크기를 의미한다 큰 고유값은 뽀족한 모양 작은 고유값은 완만한 모양 , 고유값에 대응하는 고유 벡터는 곡률이 나타내는 방향이다

그리고 alpha strongly는 최소한 alpha만큼의 곡률을 보장해야하므로 최소 고유값을 기준으로 삼아야 한다

beta smooth는 급격하게 곡률이 변하지 않음을 보장해야 하므로 큰 고유값을 기준으로 삼아야 한다 가장 뽀족한 방향의 곡률보다 beta가 작거나 같다는 조건을 위함이다

α-강한 볼록성은 함수의 **최소 곡률을 제한**하는 조건입니다. 함수의 곡률은 방향마다 다르므로, 모든 방향에서 최소 곡률을 보장하려면 **가장 작은 곡률(λmin)을 기준**으로 삼아야 합니다. 그래서 $= _{mi}$이 됩니다.

β-평활성은 함수의 **최대 곡률을 제한**하는 조건입니다. 함수의 곡률이 특정 값보다 커지지 않음을 보장하려면 **가장 큰 곡률(λmax)을 기준**으로 삼아야 합니다. 그래서 $\beta = _{mx}$가 됩니다.

Comparison for different function classes

Objective Function	Convergence Rate
convex and L-Lipschitz	$\mathcal{O}(\frac{1}{\sqrt{T}})$
convex and β-smooth	$\mathcal{O}(\frac{1}{T})$
α-strongly convex and L-Lipschitz	$\mathcal{O}(\frac{1}{T})$
α-strongly convex and β-smooth	$\mathcal{O}(\exp\left(-\frac{4T}{\kappa+1}\right))$

Table. Convergence rate of gradient descent for different properties of the objective function.

Constrained case

Consider constrained convex optimization problem

In reality, A lot of problem has constraint condition
so, Through the constrained condition we need to get reality

- weigh clipping method
- SVM(suport vector machine)

Projection Theorem

PRELIMINARIES > Hilbert space

Projection Theorem

Let H be a Hilbert space, $x \in H$, and $\mathcal{X}$ a closed subspace. There exists a unique projection $p \in \mathcal{X}$ satisfying

$$x - p \perp \mathcal{X}.$$

FIRST-ORDER METHODS FOR CONVEX OPTIMIZATION > Projection onto convex set

NOTE : projection theorem also satisfy projection onto convex property I, II

Therefore, according to property II and projection theorem, **exists unique projection p and closest point**

문제 풀때 closed subspace가 무엇인지 파악해야 한다 직선인지 아니면 hyperplane인지 등 문제 마다 closed subspace가 어떤 형태인지 파악해야 한다 여기서 부터 시작이다
closed subspace와 수직인 직선을 찾아야 한다 그리고 이를 projection theorem과 평행한것으로 연결해야 한다

Setting. Projection onto a line:

$$\mathcal{X} = \{x_0 + t\,u : t \in \mathbb{R}\}, \quad \|u\| = 1, \quad \text{given } x \in \mathbb{R}^d.$$

Solution:

NOTE : closed subspace is line , u is a direction vectorthe not a scalra, is a scarla variable.
Since p is a the point of closed subspace, $p = i$

line case : direction vector ⊥ x-p

According to Project theorem, $x - p$

$(x - p)u = 0$: direction vector , x-p should perpendicular

$p = x_0 \ u$: The case of Line , the point p can be line

$(x - (x_0 \ u))u = 0$
$(x - x_0 - u)u = 0$
$(x - x_0)u - ||u|| = 0$

$= (x - x_0)u$

$p = x_0 \ ((x - x_0)u) \ u$

Setting. Projection onto a hyperplane:

$$\mathcal{X} = \{z \in \mathbb{R}^d : a^\top z = b\}, \quad a \neq 0, \quad \text{given } x.$$

Solution:

closed subspace() is Hyperplane,

$_1 = \ _2 =$

$(_1 - \ _2) = 0$

so, is a **normal vector**

p

$(x - p)$ $x - p =$: x-p also perpendicular with so, x-p parallel with

$p = x -$

$p =$ p : p is alson included in closed subspace

$$, x - p = ,$$

$$, x \ - \ , p$$

$$, x \ - = ||||_2$$

$$= {}_{,x-}||||^2$$

$$p = x - {}_{,x-}||||^2$$

LOWER BOUNDS ON FIRST-ORDER METHODS

what is the best possible rate any first-order method can achieve?

Lower Bound shows the fundamental limit : no first order algorithm can converge faster than this rate

Theorem (Lower Bound for Gradient Descent)

Let f be α-strongly convex and β-smooth with condition number $\kappa = \beta/\alpha > 1$. For any step size choice and any starting point x_0, gradient descent satisfies

$$f(x_T) - f(x^*) \geq \frac{\alpha}{2}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2T} \|x_0 - x^*\|^2.$$

- For large κ, note that

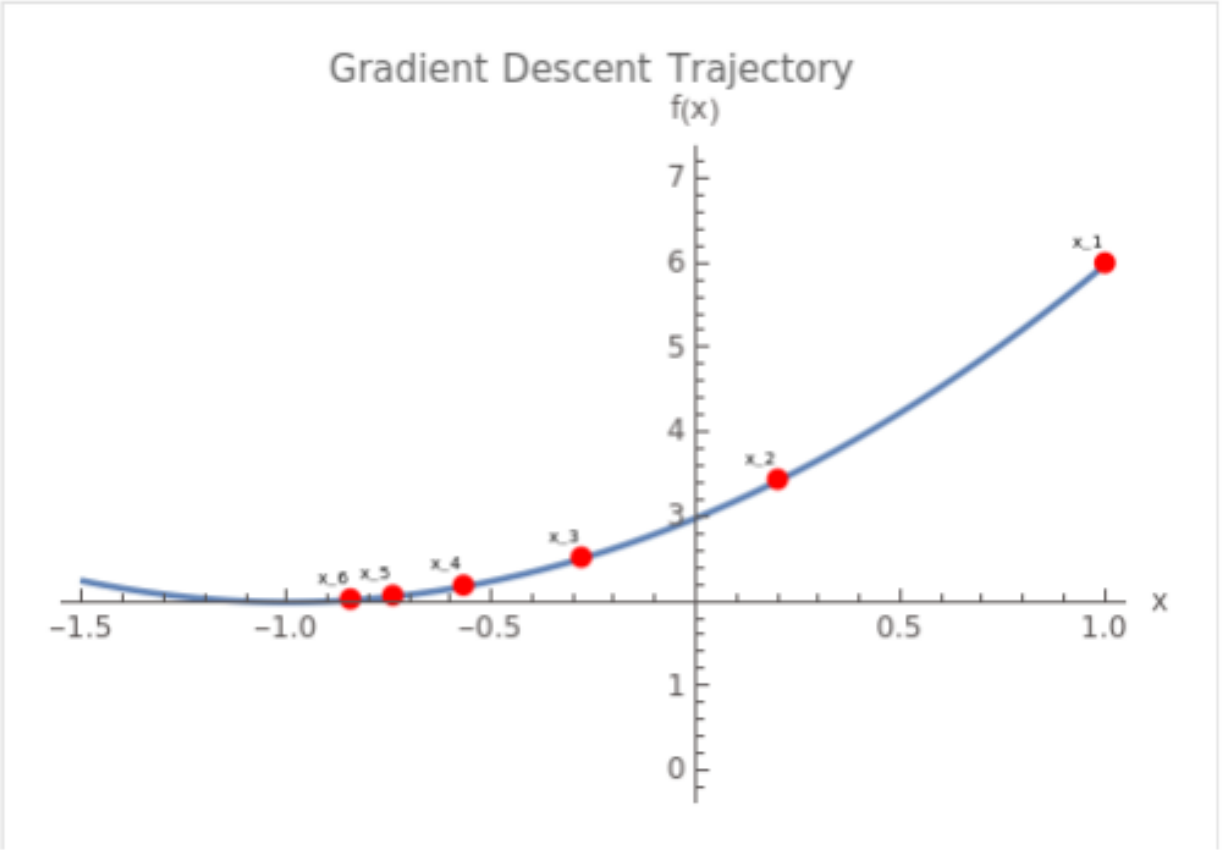
$$\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2T} \approx \exp\left(-\frac{4T}{\sqrt{\kappa}}\right).$$

if F is alpha strongly and beta smooth then any step size choice and starting point, gradient descent satisfies

- Consider

$$\min_x x^2 + 2x + 3$$

  where $x^* = -1$.
- Perform gradient descent with $x_1 = 1$ and $\gamma = 0.2$:
  - Step 1: $x_1 = 1$
  - Step 2: $\nabla f(x_1) = 4$ and $x_2 = x_1 - \gamma \nabla f(x_1) = 0.2$.
  - Step 3: $\nabla f(x_2) = 2.4$ and $x_3 = x_2 - \gamma \nabla f(x_2) = -0.28$.
  - Step 4: $\nabla f(x_3) = 1.44$ and $x_4 = x_3 - \gamma \nabla f(x_3) = -0.568$.
  - Step 5: $\nabla f(x_4) = 0.864$ and $x_5 = x_4 - \gamma \nabla f(x_4) = -0.7408$.
  - Step 6: $\nabla f(x_5) = 0.5184$ and $x_6 = x_5 - \gamma \nabla f(x_5) = -0.84448$.



case of larget step size

- What if the step size is too large?
- Perform gradient descent with $x_1 = 1$ and $\gamma = 2$:
  - Step 1: $x_1 = 1$
  - Step 2: $\nabla f(x_1) = 4$ and $x_2 = x_1 - \gamma \nabla f(x_1) = -7$.
  - Step 3: $\nabla f(x_2) = -12$ and $x_3 = x_2 - \gamma \nabla f(x_2) = 17$.
  - Step 4: $\nabla f(x_3) = 36$ and $x_4 = x_3 - \gamma \nabla f(x_3) = -55$.
  - Step 5: $\nabla f(x_4) = -108$ and $x_5 = x_4 - \gamma \nabla f(x_4) = 161$.
  - Step 6: $\nabla f(x_5) = 324$ and $x_6 = x_5 - \gamma \nabla f(x_5) = -487$.

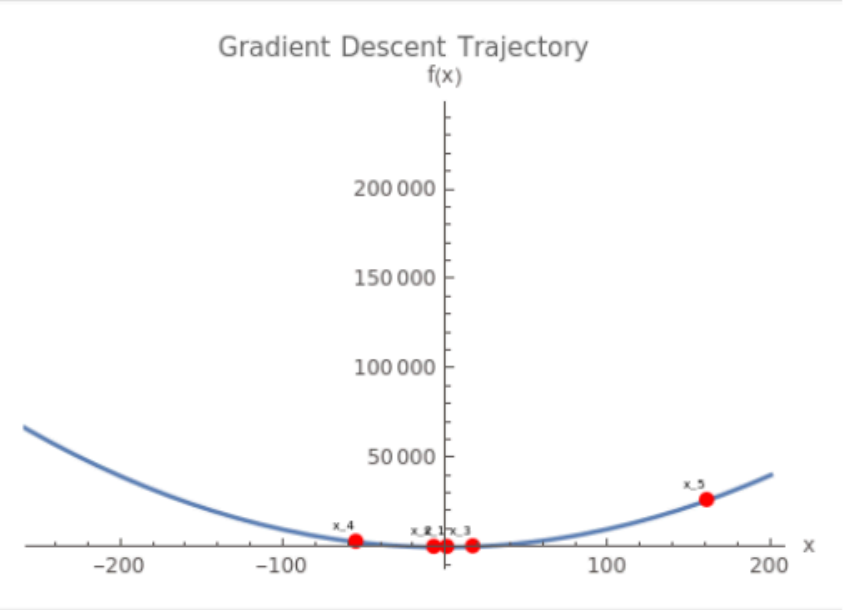


Figure. $\gamma = 2$

- The value of x_n changes very drastically due to a large learning rate.

Alpha strongly & beta smooth

then we can calculate optimal step size (check this out)

- One can calculate $\alpha = \beta = 2$.
- Perform gradient descent with $x_1 = 1$ and $\gamma = \frac{2}{\alpha+\beta} = 0.5$:
  - Step 1: $x_1 = 1$
  - Step 2: $\nabla f(x_1) = 4$ and $x_2 = x_1 - \gamma \nabla f(x_1) = -1$.

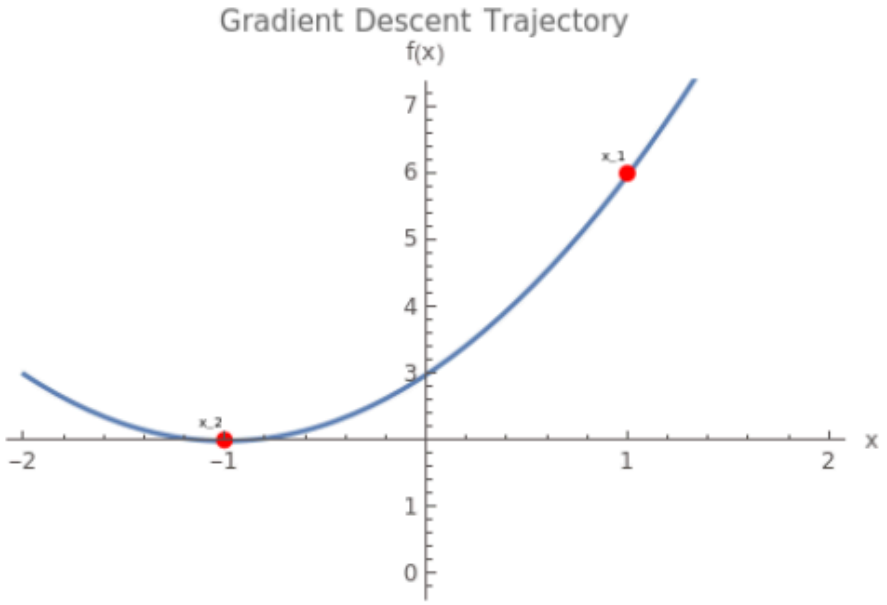


Figure. $\gamma = 0.5$

Constrained case optimization Example

- Consider the following constrained optimization

$$\min_{x \in [0,1]} x^2 + 2x + 3$$

where $x^* = 0$.

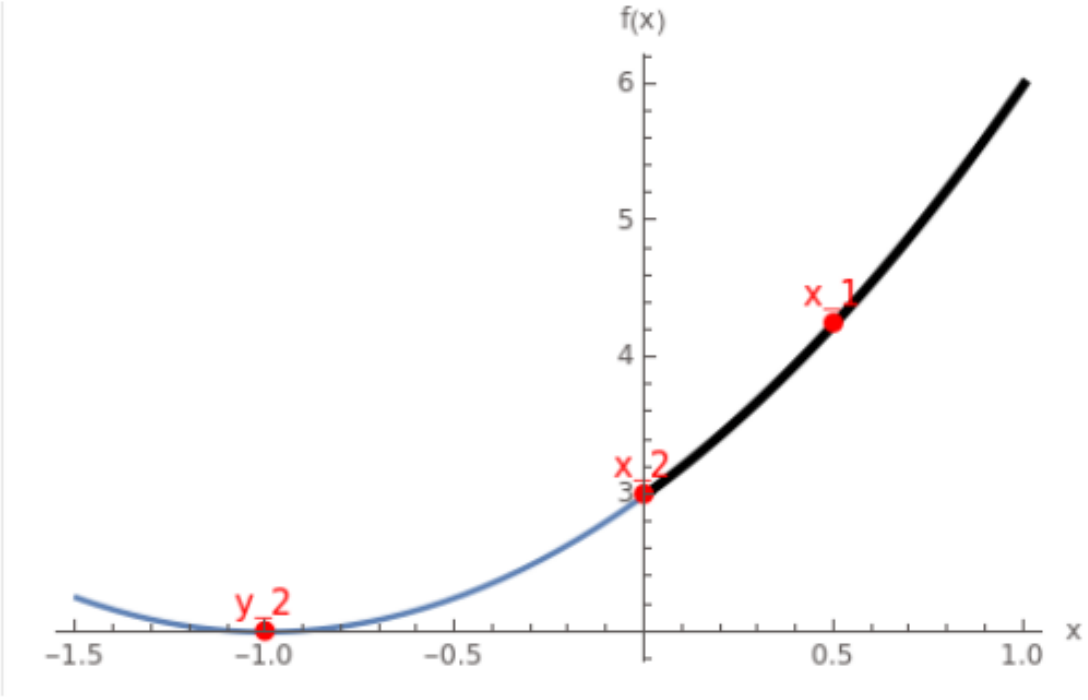
- Perform projected gradient descent with $x_1 = 0.5$ and $\gamma = 0.5$:
  - Step 1: $x_1 = 1$
  - Step 2: $\nabla f(x_1) = 4$. $y_2 = x_1 - \gamma \nabla f(x_1) = -1$ and $x_2 = \Pi_{[0,1]}(y_2) = 0$.

NOTE : x_1 = 1

Global optimal $_2$ point is -1, but have constrained. {0~1}

start point x_1 then calculate Gradient $_2$ but $_2$ = -1 this out side in constraint {0,1} so we need to projection

So, $_2$ projected onto x_2



Support Vector Machine

Classify 1 or 0

A natural formulation seeks the separating hyperplane with the fewest misclassifications

$$\min_{\theta^d} \quad _{i=1} \mathbb{1}_{\sin(\theta x_i)_i}$$

This 0–1 loss leads to a non-convex, NP-hard problem.(계단 형태로 hyperplane 경계선이 이루어져 있고 대부분 미분 불가하다)

Replace the 0-1 loss with the convex **hinge loss**
(surrogate)

$$(\theta_{i}, _{i}) = \mathrm{m}0, 1 - _{i}{}_{i}\theta$$

Add **L2 regulaization to control margin size**

$$\min_{\theta^d} f(\theta) = 1 \quad _{i=1} (\theta_{i}, _{i}) \ \ 2\|\theta\|^2.$$

NOTE : L2 is also alpha strongly convex

This is an **unconstrained, non-smooth, strongly convex problem!** Therefore, subgradient
descent achieves the rate O(1/T).

However, in practice, DL non convexity can't convert convex function.

Then how SVM can do this? : SVM of non convexity is only loss function, so we can easily convert convex function as replace loss function

Soft margin SVM

EXAMPLE: SOFT-MARGIN SVM (CONVEX FORMULATION)

Let

$$L(\theta) := \frac{1}{n}\sum_{i=1}^{n} \ell(\theta; x_i, y_i), \qquad \ell(\theta; x_i, y_i) = \max\{0,\, 1 - y_i x_i^\top \theta\},$$

and define

$$f(\theta) = L(\theta) + \frac{\lambda}{2}\|\theta\|^2, \qquad \lambda > 0.$$

Proof: L is convex since it is an average of convex hinge losses (each is a pointwise max of affine maps). In addition, $q(\theta) := \frac{\lambda}{2}\|\theta\|^2$ is λ-strongly convex because $\nabla^2 q(\theta) \succeq \lambda$. Since sum of a convex function and an λ-strongly convex function is λ-strongly convex, f is λ-strongly convex.

NOTE : constant function , linear function is also convex function

so, max(0,linear function) is convex function

Relu activation function(Non linear function) is also convex function

(θ) is also alpha strongly

Couclustion, **convex function + strongly convex function = strongly convex function**