수포자도 도전해 볼 만한

Mathematics in DeepLearning

Lecture4. Optimization

Juhee Lee ph.D.
Research Professor @ ewha womans university

Definition An optimization problem consists of **maximizing** or **minimizing** a real function by choosing input values from within an allowed set and computing the value of the function.

$$f: A \rightarrow R$$

"linear programming"

Given: set A to the real numbers R

an element x_0 in A such that $f(x_0) \leq f(x)$ for all x in A ("**minimization**") or such that $f(x_0) \geq f(x)$ for all x in A ("**maximization**")

A function

objective function $f: A \rightarrow R$

- Is called a loss function or cost function(minimization), a utility function or fitness function (maximization)
- Typically, $A \subset R^n$

  specified by a set of constraints, equalities or inequalities that the members of A have to satisfy.
- The elements of A are candidate solutions
- A **feasible solution** that minimizes (or maximize if that is the goal) the objective function is called an **optimal solution**

In mathematics, conventional optimization problems are usually stated in terms of **minimization**.

Generally, unless both the objective function and the feasible region are convex in a **minimization**, there may be several **local minima**

A local minimum x^*

for all x, some $\delta > 0 \; ||\mathrm{x} - \mathrm{x}^*|| < \delta, f(x^*) \leq f(x)$

Note.

In mathematics,

S is **bounded** above (**bounded** below)

for all $\mathrm{x} \in S$, $S(\neq \emptyset) \subset R$,

$$\exists u \in R, x \leq u \; (u \leq x)$$

S is **bounded** if S is bounded above and below

for all $\mathrm{x} \in S$, $S(\neq \emptyset) \subset R, \exists u \in R, |x| \leq \boldsymbol{u}$

Note.

In mathematics,

$$a = \sup S$$

a is supremum or least upper bound

$\exists a \in R,$ such that 1) and 2)

Let S is **bounded** above

1) a is a upper bound, ie, for all $\mathrm{x} \in S, x \leq a$

2) $\beta \in R, \beta < a, \Rightarrow \exists x \in R$ such that $\beta < x \leq a$

Note.

In mathematics,

$$u = \inf S$$

u is infimum or greatest lower bound

$\exists u \in R,$ such that 1) and 2)

Let S is **bounded** below

1) a is a lower bound, ie, for all x $\in S, u \leq x$

2) $v \in R, u < v, \Rightarrow \exists x \in R$ such that $u \leq x < v$

Note.

In mathematics, 실수계의 완비성공리 (completeness axiom)

R의 공집합이 아닌 부분집합 S가 위로 유계이면 반드시 그 상한이 존재한다.

Quiz. R의 공집합이 아닌 부분집합 S가 아래로 유계이면 그 (   )이 존재한다.

- ex. $S = \{1, 2, 3\}$ check bounded, if then sup of inf ?

- ex. $S = \{x \in R \mid 2 \leq x < 3\}$ check bounded, if then sup of inf ?

- ex. $S = \left\{\frac{1}{n} \mid n \in N\right\}$ check bounded, if then sup of inf ?

Ex. $\min_{x \in R} x^2 + 1$

Ex. $\max\limits_{x \in R} 2x$

Def. $\underset{x \in S \subset X}{\operatorname{argmax}} f(x) \coloneqq \{x \mid x \in S \wedge \forall y \in S : f(y) \leq f(x)\}$

Ex. $\underset{x \in [-5,5], y \in R}{\operatorname{argmax}} \quad x \cos y$

Feasible region, set or solution space

- Consider the problem

Minimize $x^2 + y^4$

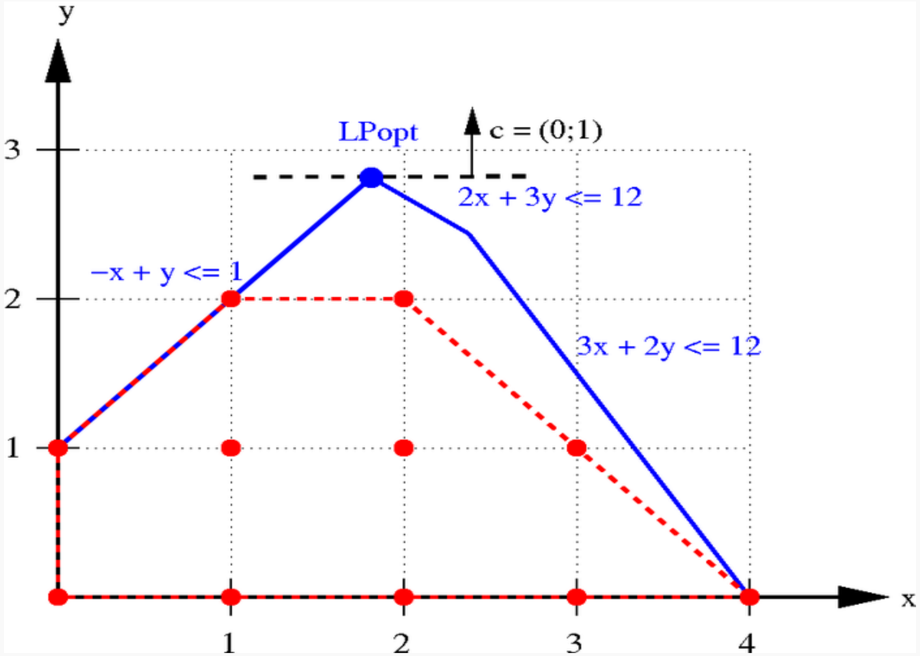
subject to $1 \leq x \leq 10$ and $5 \leq y \leq 12$

a constraint that one or more variable must be non-negative.

<출처 :https://en.wikipedia.org/wiki/Feasible_region>

Minimize $x^2 + y^4$
subject to $1 \le x \le 10$ and $5 \le y \le 12$

- Integer programming problem

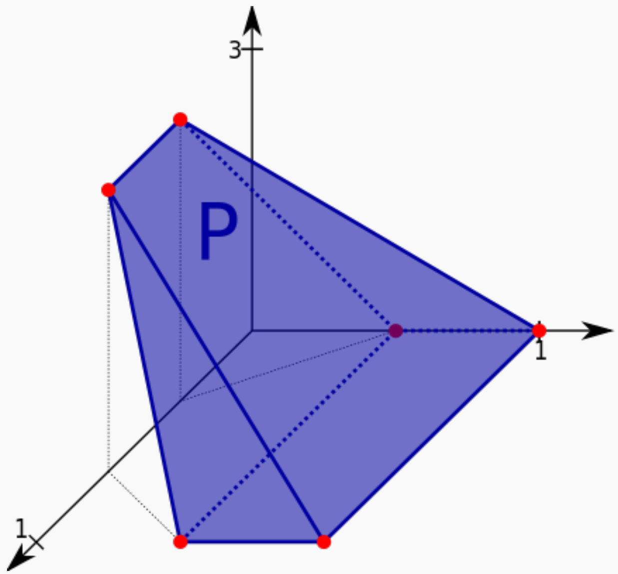


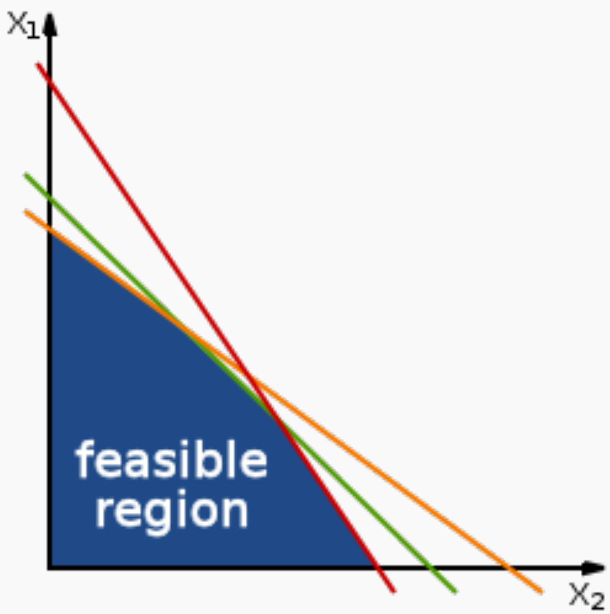
A problem with five linear constraints (in blue, including the non-negativity constraints). In the absence of integer constraints the feasible set is the entire region bounded by blue, but with integer constraints it is the set of red dots.

<출처 :https://en.wikipedia.org/wiki/Feasible_region>

Minimize $x^2 + y^4$
subject to $1 \le x \le 10$ and $5 \le y \le 12$

- Linear programming problem



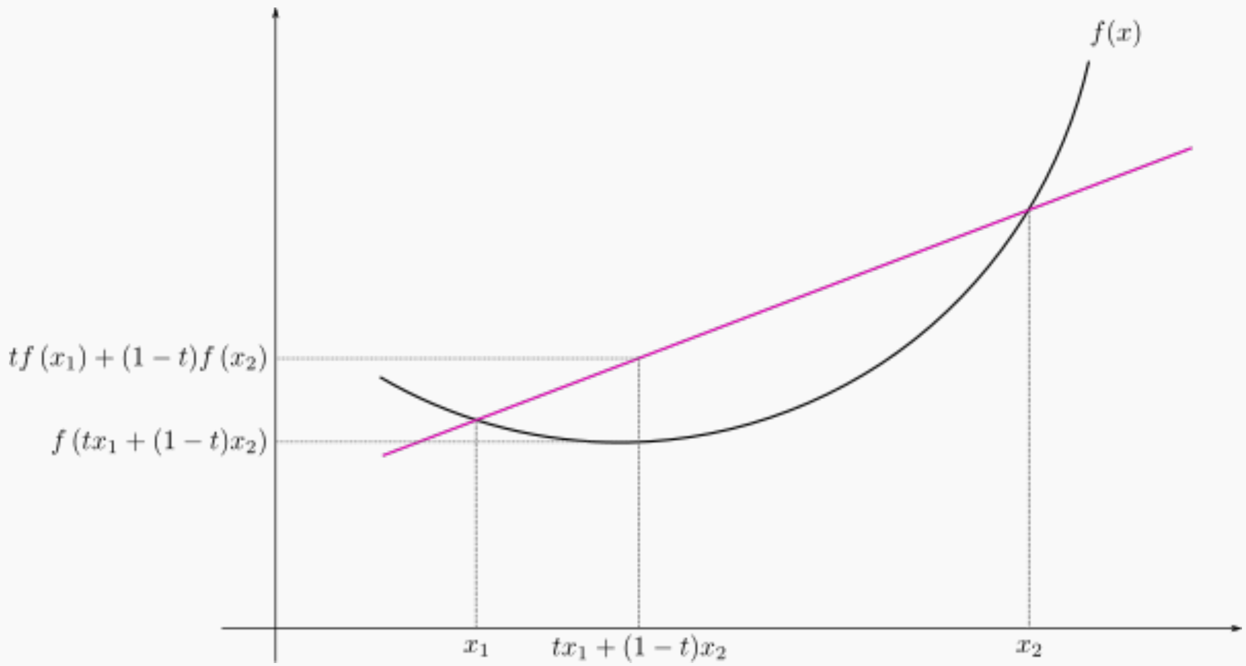


A closed feasible region of a linear programming problem with three variables is a convex polyhedron.

In a linear programming problem, a series of linear constraints produces a convex feasible region of possible values for those variables. In the two-variable case this region is in the shape of a convex simple polygon.

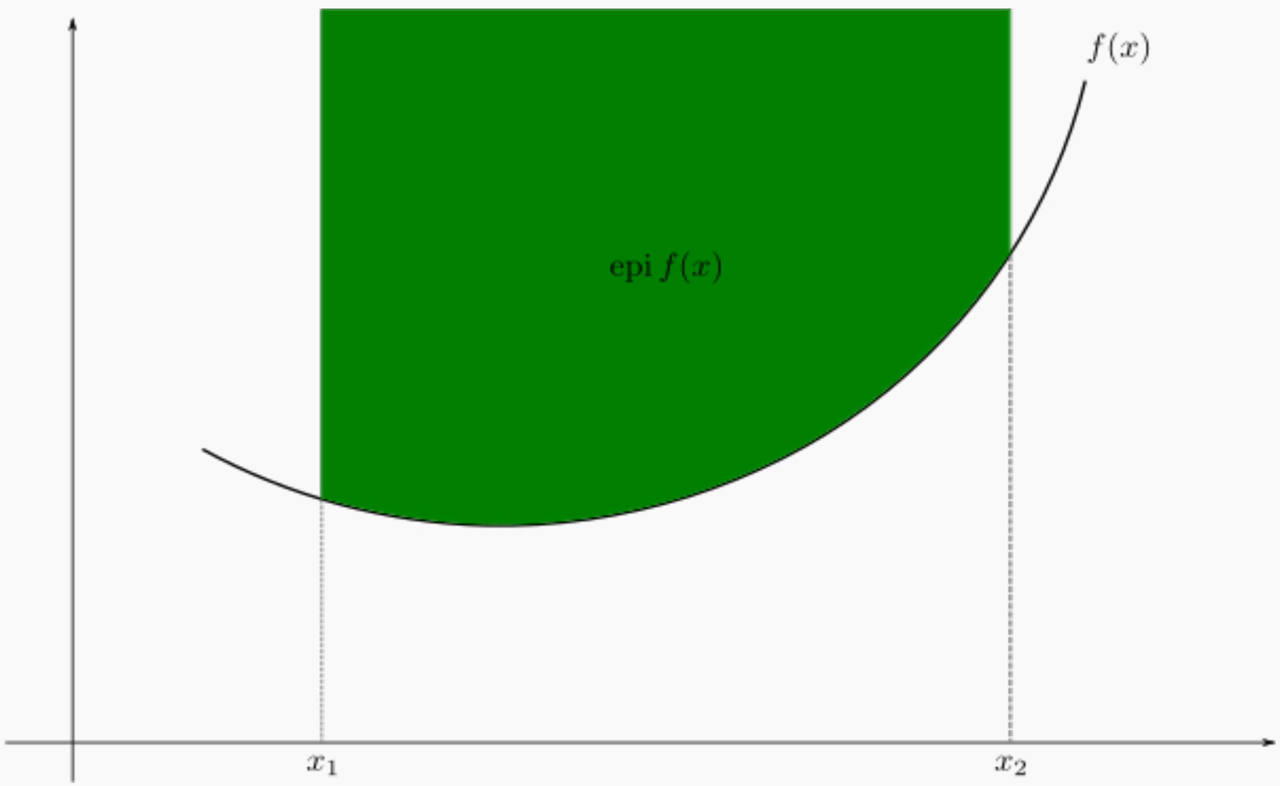
<출처 :https://en.wikipedia.org/wiki/Feasible_region>

Convex function on an interval



<그림출처 :https://en.wikipedia.org/wiki/Convex_function>

f is called **convex** if
$\forall x_1, x_2 \in X,\ \forall t \in [0,1]$: $f(tx_1 + (1-t)x_2 \leq tf(x_1) + (1-t)f(x_2)$



<그림출처 :https://en.wikipedia.org/wiki/Convex_function>

A function is convex if and only if the region above its graph (in green) is convex set

This region is the function's epigraph

$\text{epi}\, f = \{(\text{x}, \mu) : x \in R^n, \mu \in R, \mu \geq f(x)\} \subset R^{n+1}$

f is convex $\Leftrightarrow$ $\mathrm{R}(x_1, x_2)$ is monotonically non-decreasing

$$\mathrm{R}(x_1, x_2) = \frac{f(x_1)-f(x_2)}{x_1-x_2} \quad \text{in } x_1 \text{ for every fixed } x_2$$

A convex function f of one variable defined on some open interval C is continuous on C and Lipschitz continuous on any closed subinterval.

f is differentiable at all but at most countably many points

.

A diffenentiable function of one variable is convex on an interval
⇔ if its derivative is monotonically non-decreasing on that interval.

A diffenentiable function of one variable is convex on an interval
⇔ the function lies above all of its tangents

$f(x) \geq f(y) + f'(y)(x-y)$ for all x_1 and x_2 in that interval.

If $f'(c) = 0$ then c is a global minimum of $f(x)$

A twice diffenentiable function of **one variable** is convex on an interval

$\Longleftrightarrow$ if its second derivative is non-negative

$\Longleftrightarrow$ this give **a practical test for convexity**.

A twice diffenentiable function of **several variable** is convex on a convex set

$\Longleftrightarrow$ if its Hessian matrix of second derivatives is positive semidefinite on the interior of the convex set.

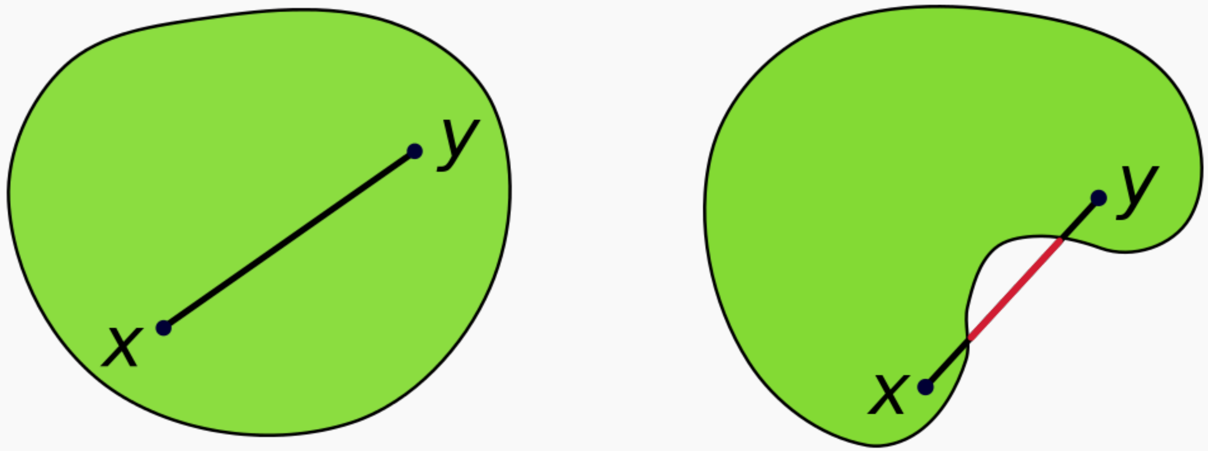
Any local minimum of a convex function is also a **global minumum.**

ex. Functions of one variable

$\mathrm{f}(x) = x^2$, $\mathrm{f}(x) = |x|^p \ (1 \leq p)$, $\mathrm{f}(x) = e^x$

ex. Functions of n variable

$\mathrm{f}(x) = \log\det(X)$ on the domain for positive-definite matrices , $every\ norm$ is a convex function



<그림출처 :https://en.wikipedia.org/wiki/Convex_set>

Convex set is a subset of an affine space that is closed under convex combination

A convex combination is a linear combination of points where all coefficients are non-negative are sum to 1
$a_1 \mathrm{x}_1 + a_2 \mathrm{x}_2 + \cdots + a_n \mathrm{x}_\mathrm{n}$. Where $a_\mathrm{i} \geq 0, real\ number, \sum a_i = 1$

Definition An affine space is a set A to which is a ssociated a vector space $\overrightarrow{A}$ and a transitive and free action of the additive group of $\overrightarrow{A}$

$$\begin{array}{ccc} A\times\overrightarrow{A} & \rightarrow & A \\ (a,v) & \longmapsto & a+v \end{array}$$

That has the following properties

1. Right identity
2. Associativity
3. Free and transitive action
4. Existence of one-to-one translations

An affine space A such that $\begin{array}{ccc} A\times\overrightarrow{A} & \to & A \\ (a,v) & \longmapsto & a+v \end{array}$

1. $\forall a \in A,\ a+0=a$
2. $\forall v,w \in \overrightarrow{A},\ \forall a \in A,\ (a+v)+w = a+(v+w)$
3. For every $a \in A$,

   the mapping $\overrightarrow{A} \to A : v \longmapsto a+v$ is a bijection
4. For all $\forall v \in \overrightarrow{A}$

   the mapping $A \to \overrightarrow{A}: a \longmapsto a+v$ is a bijection

An affine space A such that $\begin{array}{ccc} A\times\overrightarrow{A} & \to & A \\ (a,v) & \mapsto & a+v \end{array}$

keeping only the properties related to parallelism and ratio of lengths for parallel line segments

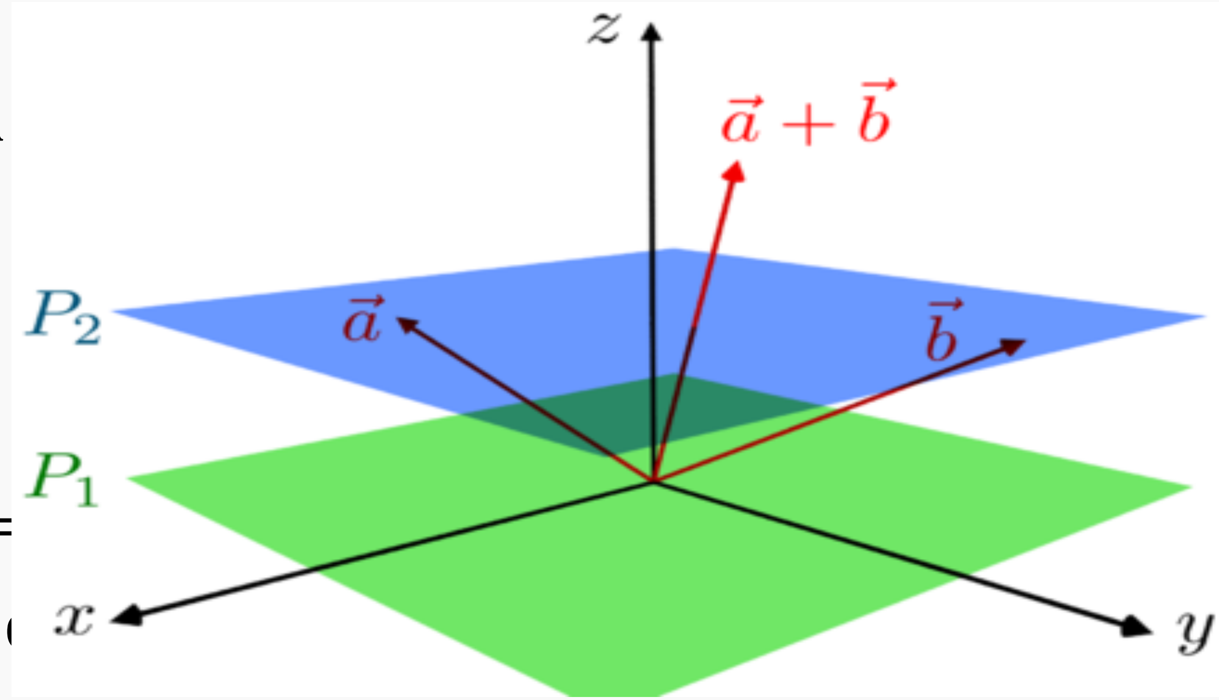
1. $\forall a \in A,\ a+0=a$
2. $\forall v,w \in \overrightarrow{A},\ \forall a \in A,\ (a+v)+w = a+(v+w)$
3. For every $a,b \in A$,

   there exists a unique $\mathrm{v} \in \overrightarrow{A}$

   denote $b-a$ such that $b=a+v$
4. For all $\forall v \in \overrightarrow{A}$

   the mapping $A \to \overrightarrow{A}: a \mapsto a+v$ is a bijection

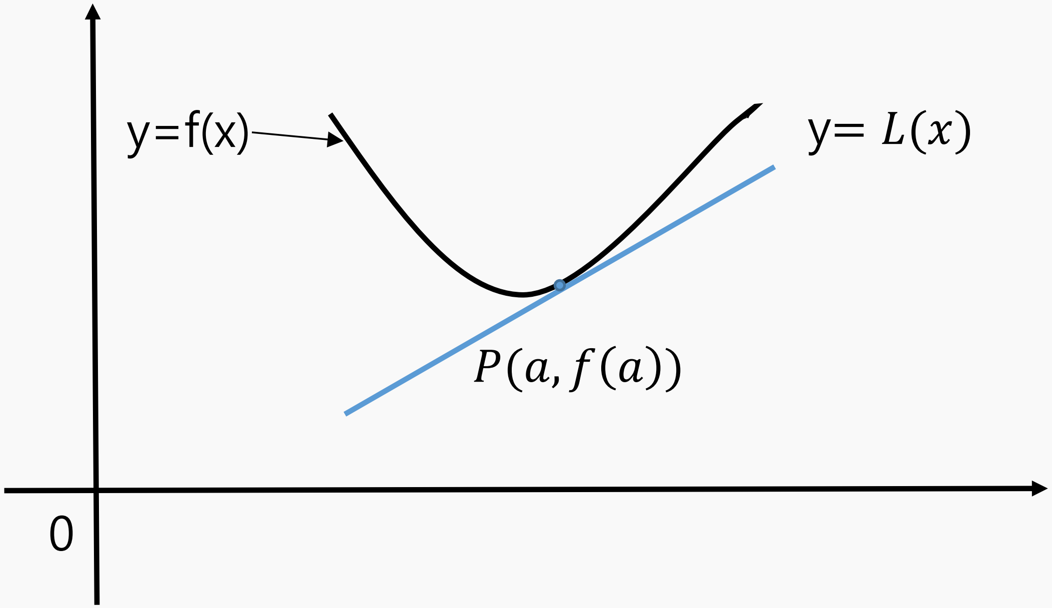
An affine space A



1. $\forall a \in A,\ a + 0 =$
2. $\forall v, w \in \overrightarrow{A},\ \forall a$
3. For every $a, b \in A$,
   there exists a unique $\mathrm{v} \in \overrightarrow{A}$
   denote $b - a$ such that $b = a + v$
4. For all $\forall v \in \overrightarrow{A}$
   the mapping $A \to \overrightarrow{A}: a \mapsto a + v$ is a bijection

First order optimizaiton

Derivative

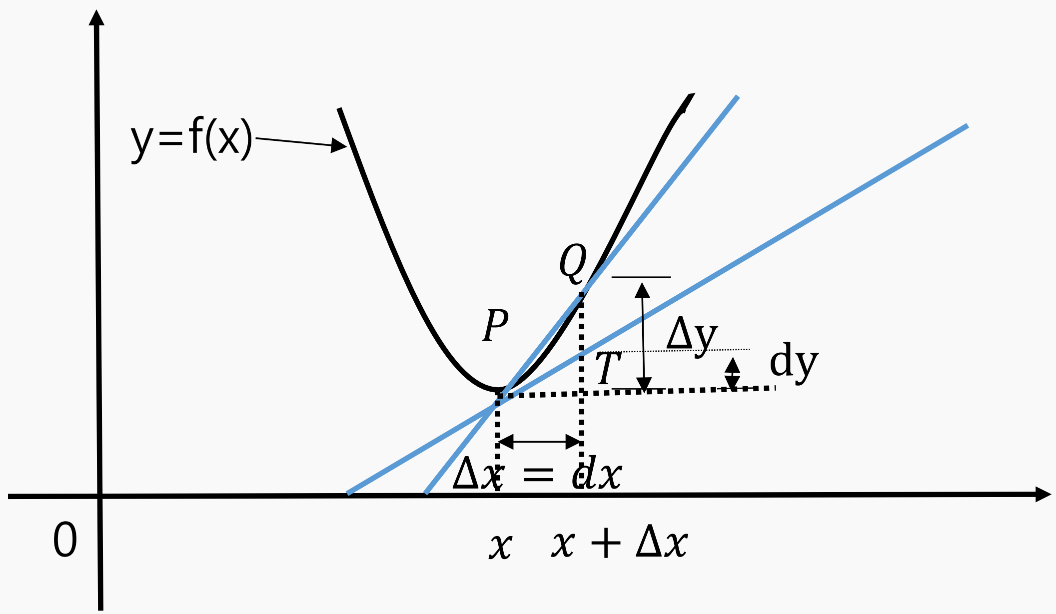


We use the tangent line at $(a, f(a))$ as an approximation to the curve $y = f(x)$ when x is near a

An equation of this tangent line is $y = f(a) + f'(a)(x - a)$

and the approximation $f(x) \approx f(a) + f'(a)(x - a)$ of f at a

x 에서의 함수값과 그 점에서의 변화율을 알면 그 점 바로 근처에 있는 점에서의 함수값을 근사적으로 계산할 수 있다.



$$f(x+dx) = f(x) + \Delta y \approx f(x) + dy = f(x) + f'(x)\mathrm{d}x$$

ex. Find the linearization of the function $f(x) = \sqrt{x+3}$ at the number $a = 1$ and use it to approximate the numbers $\sqrt{4.05}$

ex. Find the linearization of the function $f(x) = \sqrt{x+3}$ at the number $a = 1$ and use it to approximate the numbers $\sqrt{4.05}$

Sol) The derivative of $f(x) = (x+3)^{\frac{1}{2}}$ is

$$f'(x) = \frac{1}{2}(x+3)^{-\frac{1}{2}},\ f(1) = 2,\ f'(1) = 1/4$$

$$L(x) = f(1) + f'(1)(x-1) = 2 + \frac{1}{4}(x-1) = \frac{7}{4} + \frac{x}{4}$$

$$\sqrt{x+3} \approx \frac{7}{4} + \frac{x}{4}$$

$$\sqrt{4.05} \approx \frac{7}{4} + \frac{1.05}{4} = 2.0125$$

ex. For what values of x is the linear approximation

$$\sqrt{x+3} \approx \frac{7}{4} + \frac{x}{4}$$

Accurate to within 0.5 ?

ex. For what values of x is the linear approximation

$$\sqrt{x+3} \approx \frac{7}{4} + \frac{x}{4}$$

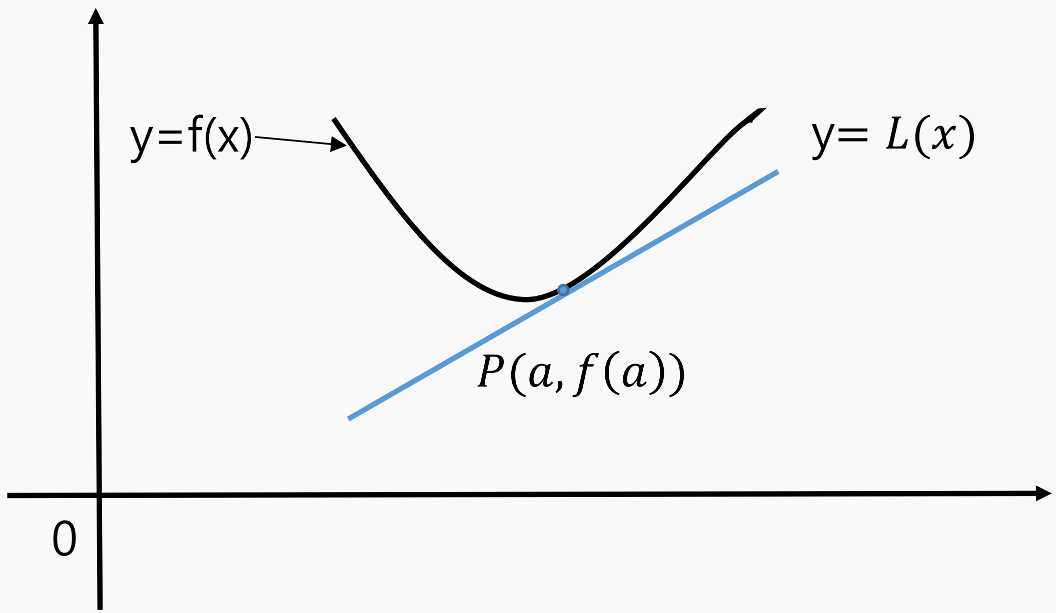
Accurate to within 0.5 ?

Sol) $\left|\sqrt{x+3} - \left(\frac{7}{4} + \frac{x}{4}\right)\right| < 0.5$

$\Leftrightarrow \sqrt{x+3} - 0.5 < \left(\frac{7}{4} + \frac{x}{4}\right) < \sqrt{x+3} + 0.5$

$\Leftrightarrow -2.6 < x < 8.6$

Derivative



We use the tangent line at $(a, f(a))$ as an approximation to the curve $y = f(x)$ when x is near a

An equation of this tangent line is $y = f(a) + f'(a)(x - a)$

Multivariate Function.

the approximation
$f(x) \approx f(a) + f'(a)(x - a)$ of f at a

the approximation

$f(x, y, z) \approx f(a, b, c) + f_x(a, b, c)(x - a)$ $+ f_y(a, b, c)(y - b)$ $+ f_z(a, b, c)(z - c)$ of f at (a, b, c)

Definition A function.

$$f: A \rightarrow R$$

for all x in A

an element x_0 in A such that $f(x_0) \leq f(x)$

f has an minimum at x_0

an element x_0 in A such that $f(x_0) \geq f(x)$ for all x in A

f has an global maximum at x_0

The maximum and minimum value of f are called the **extreme values of f**

Definition A function.

$$f: A \rightarrow R$$

for all x in A in some open interval containing c

an element c in A such that $f(c) \leq f(x)$

f has an **local minimum** when x is near c

an element$c$in A such that $f(c) \geq f(x)$ for all x in A

f has an **local maximum** when x is near c

Ex. $f(x) = cosx$

$x = 2n\pi \ cosx = 1,\ \ x = (2n+1)\pi \ cosx = -1$

Multivariate function

<u>Definition</u> A function.

$$f : D \rightarrow R$$

for all x, y in D

an element x_0, y_0 in D such that $f(x_0, y_0) \leq f(x, y)$

f has an minimum at x_0, y_0

an element x_0, y_0 in D such that $f(x_0, y_0) \geq f(x, y)$

f has an maximum at x_0, y_0

The maximum and minimum value of f are called the **extreme values of f**

Multivariate function

<u>Definition</u> A function.

$$f : D \rightarrow R$$

for all x, y in D

$\exists r > 0 \ ||\boldsymbol{x} - \boldsymbol{x_0}|| < r$ in D such that $f(\boldsymbol{x_0}) \leq f(\boldsymbol{x})$

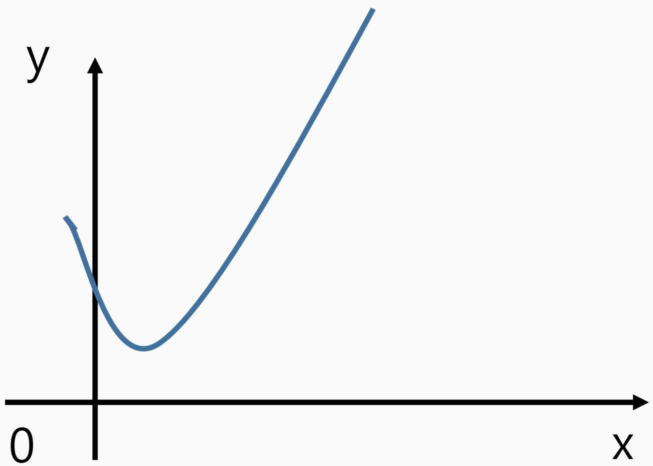
f has an minimum at $\boldsymbol{x_0}$

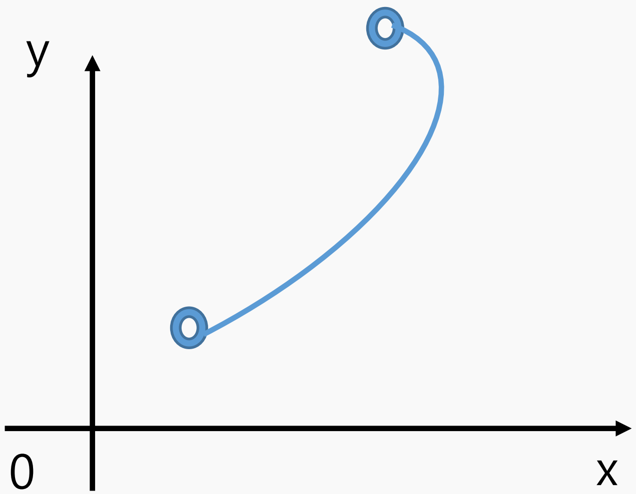
$\exists r > 0 \ ||\boldsymbol{x} - \boldsymbol{x_0}|| > r$ such that $f(\boldsymbol{x_0}) \geq f(\boldsymbol{x})$

f has an maximum at $\boldsymbol{x_0}$

The maximum and minimum value of f are called the **extrem e values of f**

No maximum value.





No minimum

or maximum value

Theorem. (Mean value theorem)

- f is continuous in $[a, b]$
- f is differentiable on (a, b)

Then there is a number c in (a, b) such that

$$f'(c) = \frac{f(b)-f(a)}{b-a} \quad \text{or } f(b) - f(a) = f'(c)(b - a)$$

Theorem. (Fermat) If f has a local maximum or minimum at $\mathrm{c} \in (a, b)$ and if

$f'(\mathrm{x})$ exists, then $f'(\mathrm{x}) = 0$.

Definition A **critical point** of a function f is a $\mathrm{c} \in$ $domain$ of f such that either $f'(\mathrm{c}) = 0$ or $f'(\mathrm{c})$ doesn't exist.

Theorem. (Fermat) If f has a local maximum or minimum at (a, b) and if the first order partial derivatives of f exists there, then

$$f_x(a, b) = 0, f_y(a, b) = 0.$$

<u>Definition</u> A **critical point** of a function f is at (a, b) if $f_x(a, b) = 0 \; and \; f_y(a, b) = 0$ or $f'(\mathrm{a}, \mathrm{b})$ doesn't exist.

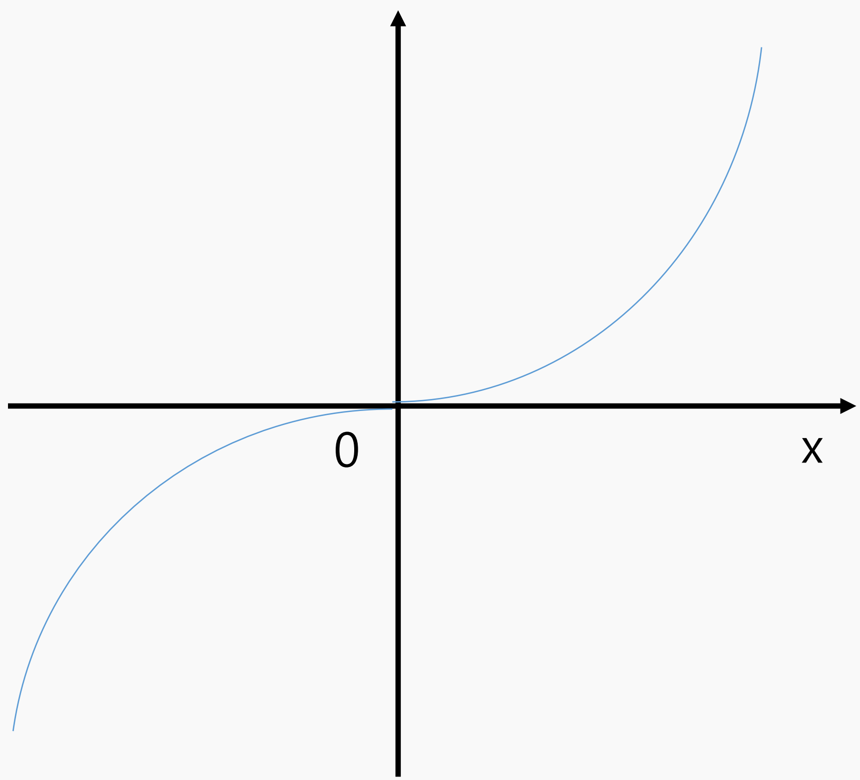
If f has a local maximum or minimum at c and if

$f'(\mathrm{x})$ exists, then $f'(\mathrm{x}) = 0$

When $f'(\mathrm{x}) = 0$, then

When $f'(\mathrm{x}) = 0,$ then f doesn't necessarily have a maximum or minimum at c

$$f(\mathrm{x}) = \mathrm{x}^3 \quad f'(x) = 3x^2, f'(0) = 0$$



First derivative test

- If f' changes from positive to negative at c then f has

First derivative test

- If f' changes from positive to negative at c then f has a local maximum

- If f' changes from negative to positive at c then f has a

First derivative test

- If f' changes from positive to negative at c
  then f has a local maximum

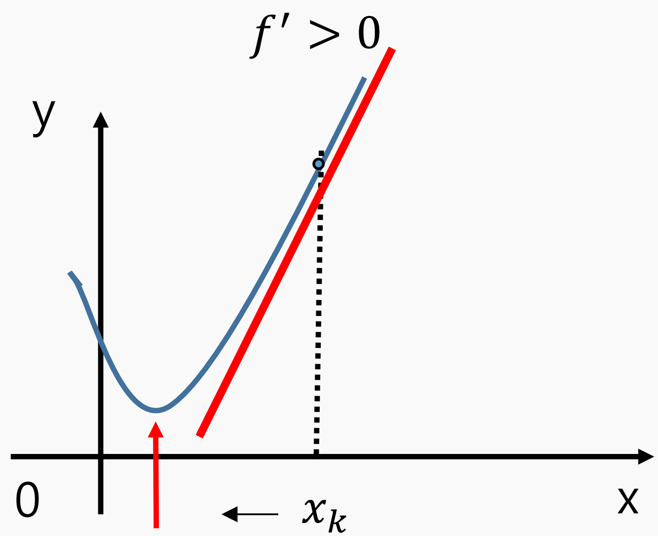
- If f' changes from negative to positive at c
  then f has a local minimum
- If f' does not change sign at c
  then f has

First derivative test

- If f' changes from positive to negative at c
  then f has a local maximum

- If f' changes from negative to positive at c
  then f has a local minimum
- If f' does not change sign at c
  then f has no local maximum or minimum at c

Gradient Descent



Loss 가 최소인 목표지점

$$x_{k+1} = x_k - \lambda f'(x_k) \quad \lambda = learning\ rate$$

$$\boldsymbol{x_{k+1}} = \boldsymbol{x}_k - \lambda \boldsymbol{\nabla f}(x_k)$$

Second order optimizaiton

Second derivative test

- If function f is twice diffentiable at a critical point c then

- If $f''(c) < 0$ then f has a local maximum
- If $f''(c) > 0$ then f has a local minimum

- If $f''(c) = 0$ then 😅

Second derivative test (proof)

Using Tylor's theorem,

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^k}{k!}(x-a)^k$$
$$+h_k(x)(x-a)^k$$

$$0 < f''(c) = \lim_{h \to 0} \frac{f'(c+h) - f'(c)}{h}$$
$$= \lim_{h \to 0} \frac{f\prime(c+h) - 0}{h} = \lim_{h \to 0} \frac{f\prime(c+h)}{h}$$

$\frac{f\prime(c+h)}{h} > 0 \Rightarrow f'(c+h) > 0 \; if \; h > 0$ incresing ie, local minimum

Second derivative test

- If function f is twice diffentiable at a critical point or stationary point (a, b) then
$$D = D(a, b) = f_{xx}(a, b) f_{yy}(a, b) - \left[f_{x,y}(a, b)\right]^2$$
- If $D > 0, f_{xx}(a, b) > 0$ then $f(a, b)$ is a local minimum
- If $D > 0, f_{xx}(a, b) < 0$ then $f(a, b)$ is a local maximum
- If $D < 0, f_{xx}(a, b) > 0$ then $f(a, b)$ is not a 😅 (saddle point)
- If $D = 0$, test gives no information😅

Proof.

- $ax^2 + 2bxy + cy^2$ quadratic forms for x
- $ax^2 + 2bxy + cy^2 = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} a & b \\ b & c \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$

  $\mathrm{T} = \sqrt{b^2 - ac} \Rightarrow D = \det\left(\begin{bmatrix} a & b \\ b & c \end{bmatrix}\right) = ac - b^2$

  $D > 0 \Rightarrow T < 0$

  $\mathrm{H} = \begin{bmatrix} f_{xx} & f_{xy} \\ f_{yx} & f_{yy} \end{bmatrix} = f_{xx}f_{yy} - f_{xy}^2 > 0,\ f_{xx} > 0$

  local minimum

For two variable functions

- $ax^2 + 2bxy + cy^2$ quadratic forms
- In single variable

  $f'(a) = 0, if\ f''(a) > 0$ f has a local minimum

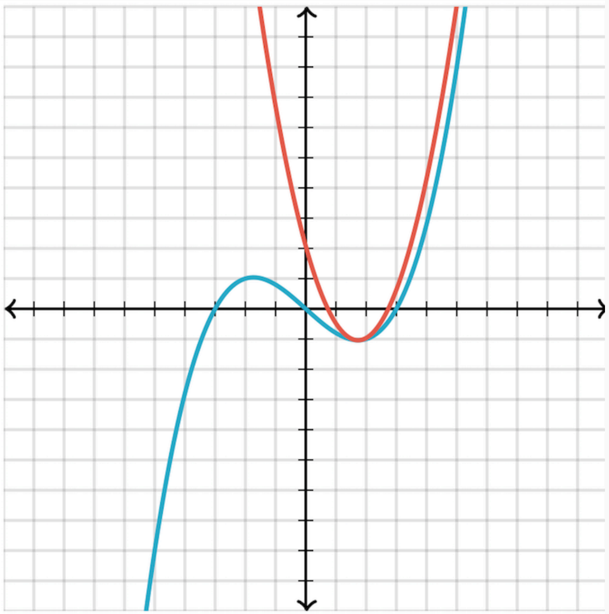
$$f(x) \approx f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2$$

For two variable functions

- In single variable

  $f'(a) = 0, if\ f''(a) > 0$ f has a local minimum

  $$f(x) \approx f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2$$



For two variable functions

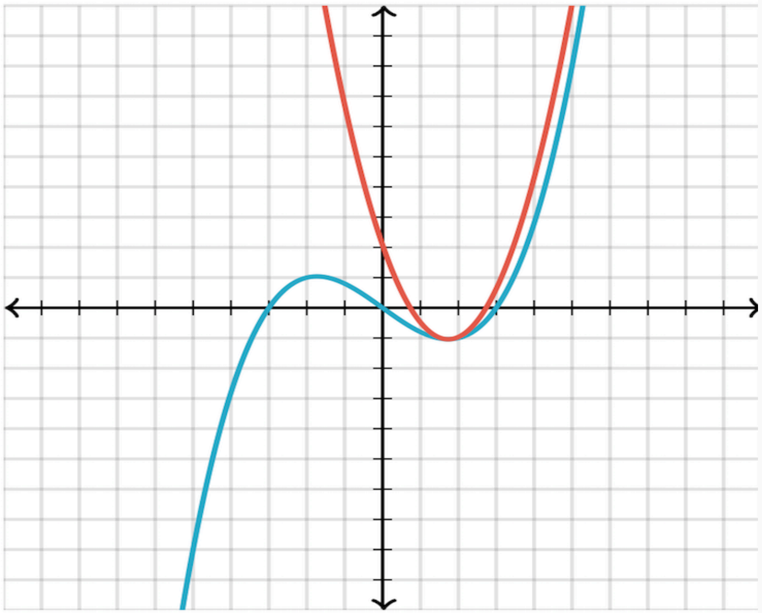
- In single variable
  $f'(a) = 0, if\ f''(a) > 0$ f has a local minimum

$$f(x) = f(x + \Delta x) \approx f(x) + f(x)\Delta x + \frac{f''(x)}{2!}\Delta x^2$$

Δx에 관하여 미분,

$$-\frac{f'(x)}{f''(a)} = \Delta x$$

$$\boldsymbol{x_{k+1}} = \boldsymbol{x}_k - \frac{\boldsymbol{\nabla f}(x_k)}{Hf(x_k)}$$



Q & A

Review

If $\mathrm{f} : R^n \to R$ is a function of n variables,

the gradient vector, $\boldsymbol{\nabla f}$

$$\nabla f(\boldsymbol{x}) = \left(\frac{\partial f}{\partial x_1}, \cdots, \frac{\partial f}{\partial x_n}\right)$$

Check. Product rule, Chain rule!

A **vector-valued function** or **vector function**, is simply a function whose domain is a set of real numbers and whose range is a set of vectors.

$$\boldsymbol{F}: R \;\to\; R^m$$

For every number t in the domain of $\boldsymbol{F}$ there is a unique vector in V_m denoted by $\boldsymbol{F}$(t),

$$\boldsymbol{F}(\mathrm{t}) = (\, f_1(t) \;, f_2(\mathrm{t}), \cdots, f_m(t))$$

From $\boldsymbol{F}(\mathrm{t}) =< f_1(t)\ , f_2(\mathrm{t}), \cdots, f_m(t) >$

$$\boldsymbol{\nabla F}: R \ \rightarrow\ R^m$$

$$\boldsymbol{\nabla}\, \boldsymbol{F}(\mathrm{t}) = (\, f'_1(t)\ , f'_2(\mathrm{t}), \cdots, f'_m(t))$$

Vector valued multivariate Ft.

From $\boldsymbol{F}(\mathbf{x}) = (\,f_1(\mathbf{x})\ , f_2(\mathbf{x}), \cdots, f_m(\mathbf{x}))$, $\mathbf{x} \in \boldsymbol{R}^{\boldsymbol{n}}$

$$\boldsymbol{F}: R^n \ \rightarrow \ R^m$$

$$\boldsymbol{\nabla}\, \boldsymbol{F}(\mathbf{x}) = (\,\boldsymbol{\nabla} f_1(\mathbf{x})\ , \boldsymbol{\nabla} f_2\big(\mathbf{x}\big), \cdots, \boldsymbol{\nabla} f_m(\mathbf{x})\big)$$

$$= \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_1}{\partial x_n} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix} \ = \ \mathrm{d}f_i^T$$

From $\boldsymbol{F}(\mathbf{x}) = (\,f_1(\mathbf{x})\ , f_2(\mathbf{x}), \cdots, f_m(\mathbf{x}))$, $\mathbf{x} \in \boldsymbol{R^n}$

$$\boldsymbol{F}: R^n \;\rightarrow\; R^m$$

Jacobian matrix,

$$\begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix} = \boldsymbol{J}_{ij} = \frac{\partial f_i}{\partial x_j}$$

Ex. $\boldsymbol{F}: R^2 \rightarrow R^2$

$$\boldsymbol{F}(x, y) = \begin{bmatrix} x^2 y \\ 5x + \sin y \end{bmatrix} = \begin{bmatrix} f_1(x, y) \\ f_2(x, y) \end{bmatrix}$$

Jacobian matrix, $\boldsymbol{J}_{ij} = \frac{\partial f_i}{\partial x_j} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} \end{pmatrix} = \begin{pmatrix} 2xy & x^2 \\ 5 & \cos y \end{pmatrix}$

From $\boldsymbol{F}(\mathbf{x}) = f(\mathbf{x})$, $\mathbf{x} \in \boldsymbol{R^n}$

$$\boldsymbol{F}: R^n \ \rightarrow \ R$$

Hessian matrix,

$$\begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix} = \boldsymbol{H_{ij}} = \frac{\partial^2 f}{\partial x_i \partial x_j}$$