$f(x): \mathbb{R}^n \to \mathbb{R}^m$

$$Df(x) = \left[\frac{\partial f_i(x)}{\partial x_j}\right] = \begin{pmatrix} \frac{\partial f_1(x)}{\partial x_1} & \frac{\partial f_1(x)}{\partial x_2} & \cdots & \frac{\partial f_1(x)}{\partial x_n} \\ \frac{\partial f_2(x)}{\partial x_1} & \frac{\partial f_2(x)}{\partial x_2} & \cdots & \frac{\partial f_2(x)}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m(x)}{\partial x_1} & \frac{\partial f_m(x)}{\partial x_2} & \cdots & \frac{\partial f_m(x)}{\partial x_n} \end{pmatrix}_{m \times n}$$

Define an operator $\bar{D}: \mathbb{R} \to \mathbb{R}^n$ by $\bar{D} = \begin{pmatrix} \frac{\partial}{\partial x_1} \\ \frac{\partial}{\partial x_2} \\ \vdots \\ \frac{\partial}{\partial x_n} \end{pmatrix}$

$$(Df(x))^T = \begin{pmatrix} \frac{\partial}{\partial x_1} \\ \frac{\partial}{\partial x_2} \\ \vdots \\ \frac{\partial}{\partial x_n} \end{pmatrix} (f_1(x)\ f_2(x) \cdots f_m(x))$$

$$\Rightarrow \underbrace{(Df(x))^T}_{\text{matrix.}} = \bar{D} \cdot f(x)^T.$$

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

If $f: \mathbb{R}^n \to \mathbb{R}$. $\nabla f(x) = Df(x)^T = \bar{D} \cdot f(x)$

$$\begin{aligned} \nabla^2 f(x) &= \nabla(\nabla f(x)) \\ &= \bar{D} \cdot (\nabla f(x))^T \\ &= \underline{\bar{D} \cdot Df(x)}. \end{aligned}$$

① $f(x) = Ax$. $A \in \mathbb{R}^{m \times n}$, $x \in \mathbb{R}^n$.

$$Ax = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & & & \\ & \ddots & & \\ & & & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \end{pmatrix}$$

Let $A = \begin{bmatrix} -a_1- \\ -a_2- \\ \vdots \\ -a_m- \end{bmatrix}$. $= \begin{pmatrix} a_1 x \\ a_2 x \\ \vdots \\ a_m x \end{pmatrix}$.

$$Df(x) = \begin{pmatrix} \frac{\partial a_1 x}{\partial x_1} & \frac{\partial a_1 x}{\partial x_2} & \cdots & \frac{\partial a_1 x}{\partial x_n} \\ \vdots & & & \\ \frac{\partial a_m x}{\partial x_1} & & \cdots & \frac{\partial a_m x}{\partial x_n} \end{pmatrix} = A.$$

② $f(x, y) = y^T A x$. $A \in \mathbb{R}^{m \times n}$. $x \in \mathbb{R}^n$. $y \in \mathbb{R}^m$.

$D_x f(x, y) = y^T A$

$D_y f(x, y) = ?$

$y^T A x \in \mathbb{R} = (y^T A x)^T = x^T A^T y$

$D_y f(x, y) = x^T A^T$.

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

$Df(x) = (x^T P^T) + x^T P = 2x^T P$ because $P^T = P$.

$\left\{ \frac{d\, f(x) g(x)}{dx} = f(x) g(x) + f(x) g'(x) \right.$

Chain rule

Suppose $f: \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $x \in \text{int dom } f$ and $g: \mathbb{R}^m \to \mathbb{R}^p$ is diff. at $f(x) \in \text{int dom } g$.

Define the composition $h: \mathbb{R}^n \to \mathbb{R}^p$ by $h(x) = g(f(x))$.

Then h is differentiable at x, with derivative

$$\underset{p\times n}{Dh(x)} = \underset{p\times m}{Dg(y)}\ \underset{m\times n}{Df(x)} \Big|_{y=f(x)}$$

Ex. $f: \mathbb{R}^n \to \mathbb{R}^m$ defined by $f(x) = Ax+b$. $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$

Suppose that $g: \mathbb{R}^m \to \mathbb{R}^p$ is differentiable.

Then $h = g(f(x)) = g(Ax+b)$.

$$Dh(x) = Dg(y) \cdot Df(x) = Dg(y)\, A \Big|_{y=Ax+b}$$

If $p=1$, $\nabla h(x) = (Dh(x))^T = A^T (Dg(y))^T$

$$= A^T \nabla g(y) \Big|_{y=Ax+b}$$

$n=m=p=1$

$h(x) = g(f(x)) \to h'(x) = g'(f(x)) \cdot f'(x)$

Example. $f: \mathbb{R}^n \to \mathbb{R}$ $\text{dom} f = \mathbb{R}^n$

$$f(x) = \log\left(\sum_{i=1}^{m} \exp(a_i^T x + b_i)\right), \quad \text{where } a_i \in \mathbb{R}^n,\ b_i \in \mathbb{R}$$

$i = 1, 2, \ldots, m.$

Let $h(x) = Ax + b$, $A = \begin{pmatrix} -a_1^T- \\ \vdots \\ -a_m^T- \end{pmatrix}$, $b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}$.

Let $g: \mathbb{R}^m \to \mathbb{R}$ by

$$g(y) = \log\left(\sum_{i=1}^{m} \exp(y_i)\right).$$

$$\frac{d \log x}{dx} = \frac{1}{x}$$

$$f(x) = g(h(x)).$$

$$\nabla f(x) = A^T \nabla g(y) \Big|_{y = Ax+b}. \quad \text{by chain rule.}$$

$$\frac{\partial g(y)}{\partial y_i} = \frac{1}{\sum_{i=1}^{m} \exp y_i} \exp y_j$$

$$\Rightarrow \nabla g(y) = \frac{1}{\sum_i \exp y_i} \begin{pmatrix} \exp y_1 \\ \vdots \\ \exp y_m \end{pmatrix}$$

$$\nabla^2 f(x) = A^T \left(\nabla^2 g(Ax+b)\right) A$$

#

Second derivative

For $f: \mathbb{R}^n \to \mathbb{R}$, the second derivative of f (Hessian matrix) at $x \in \operatorname{int} \operatorname{dom} f$, denoted by $\nabla^2 f(x)$, is given by

$$\left(\nabla^2 f(x)\right)_{i,j} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, n,$$

provided that f is twice differentiable at x.

The second order approximation of f at or near x, is the quadratic function of z defined by

$$\hat{f}(z) = f(x) + \nabla f(x)^T (z - x) + \frac{1}{2} (z - x)^T \nabla^2 f(x) (z - x)$$

$f: \mathbb{R} \to \mathbb{R}$. Taylor series exists if f is n times continuously differentiable

$$f(z) = f(x) + f'(x)(z - x) + \frac{1}{2} f''(x)(z - x)^2 + \frac{1}{3!} f'''(x)(z - x)^3 + \cdots + \frac{1}{(n-1)!} f^{(n-1)}(x)(z - x)^{n-1} + R_n$$

If f is differentiable, the gradient mapping is the function $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ with dom of = dom of f with value $\nabla f(x)$ at x.

The derivative of this function is

$$D\nabla f(x) = \nabla^2 f(x)$$

quadratic function.

Ex. $f(x) = \frac{1}{2}x^T P x + q^T x + r$, $P \in S^n$, $q \in \mathbb{R}^n$, $r \in \mathbb{R}$

$$Df(x) = x^T P + q^T \quad \& \quad \boxed{\nabla f(x) = Px + q}$$

$$\nabla^2 f(x) = P$$

$$\nabla^2 f(x) = D \cdot Df(x) = \begin{pmatrix} \partial/\partial x_1 \\ \partial/\partial x_2 \\ \vdots \\ \partial/\partial x_n \end{pmatrix} (x^T P + q^T)$$

$$= P.$$

The second-order approximation of a quadratic function is itself. (Exercise).

Chain Rule.

1. Composition with scalar function

Suppose $f: \mathbb{R}^n \to \mathbb{R}$, $g: \mathbb{R} \to \mathbb{R}$

and $h(x) = g(f(x))$.

Then $\nabla^2 h(x) = g'(f(x)) \nabla^2 f(x) + g''(f(x)) \nabla f(x) \nabla f(x)^T$.

If $n=1$. $h'(x) = g'(f(x)) \cdot f'(x)$.

$h''(x) = g''(f(x)) f'(x) \cdot f'(x) + g'(f(x)) \cdot f''(x)$.

proof. $\nabla h(x) = g'(f(x)) \cdot \nabla f(x)$.

$\nabla^2 h(x) = D \cdot (\nabla h(x))^T$

$= D \cdot (g'(f(x)) \cdot \nabla f(x))^T$.

$= (D\, g'(f(x))) \cdot \nabla f(x)^T + g'(f(x))\, D \cdot (\nabla f(x))^T$

$= g''(f(x))\, \nabla f(x) \cdot (\nabla f(x))^T + g'(f(x))\, \nabla^2 f(x)$ //.

$\nabla^2 h(x) = \begin{pmatrix} \partial/\partial x_1 \\ \partial/\partial x_2 \\ \vdots \\ \partial/\partial x_n \end{pmatrix} (g'(f(x)) \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} & \frac{\partial f(x)}{\partial x_2} & \cdots & \frac{\partial f(x)}{\partial x_n} \end{pmatrix}$.

2. Composition with an affine function

Suppose $f: \mathbb{R}^n \to \mathbb{R}^m$. $f(x) = Ax + b$. $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$.

$g: \mathbb{R}^m \to \mathbb{R}^p$ is differentiable.

Define $h: \mathbb{R}^n \to \mathbb{R}^p$ by $h(x) = g(f(x)) = g(Ax+b)$

Then $Dh(x) = Dg(y) \cdot A \Big|_{y = Ax+b}$.

Thus $\nabla h(x) = \boxed{A^T} \nabla g(y) \Big|_{y = Ax+b}$.

$\nabla^2 h(x) = D \cdot (\nabla h(x))^T$.

$= D \, (\nabla g(y))^T \cdot A$

$= A^T \cdot \nabla^2 g(y) \cdot A \Big|_{y = Ax+b}$

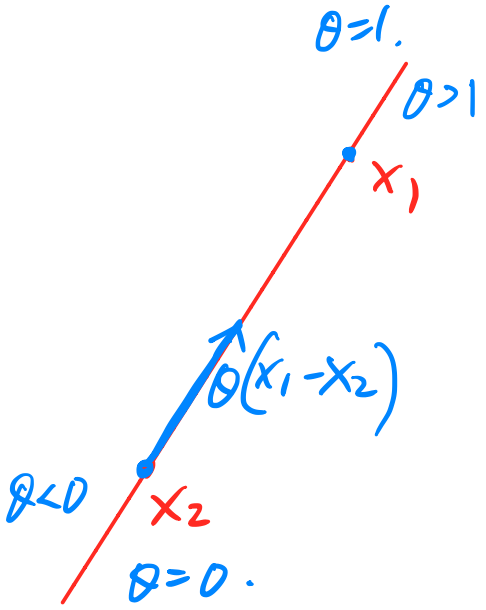
by the chain rule.

Chapter 2. Convex sets

Line: $x_1, x_2 \in \mathbb{R}^n$, $x_1 \neq x_2$

$$y = \theta x_1 + (1-\theta) x_2, \quad \theta \in \mathbb{R}$$

$$= \theta (x_1 - x_2) + x_2$$



line segment

$$y = \theta x_1 + (1-\theta) x_2, \quad \theta \in [0,1]$$

closed interval

$0 \leq \theta \leq 1$.

Affine sets

A set $C \in \mathbb{R}^n$ is affine if for any $x_1 \neq x_2 \in C$ we have $\theta x_1 + (1-\theta) x_2 \in C$ for $\theta \in \mathbb{R}$.

$\theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_k x_k$, where $\theta_1 + \theta_2 + \cdots + \theta_k = 1$, is called the affine combination of $x_1, \ldots, x_k$.

If C is an affine set, $x_1, x_2, \ldots, x_k \in C$, then the affine combination of $x_1, \ldots, x_n$ is in C.

If C is affine, and $x_0 \in C$, then

$V \triangleq C - x_0 = \{x - x_0 : x \in C\}$ is a subspace.

It means for $v_1, v_2 \in V$, $\alpha, \beta \in \mathbb{R}$.

$$\alpha v_1 + \beta v_2 \in V.$$

proof. $v_1 = x_1 - x_0 \in V$. $v_2 = x_2 - x_0 \in V$.

$x_1, x_2 \in C$.

$$\begin{aligned} \alpha v_1 + \beta v_2 &= \alpha(x_1 - x_0) + \beta(x_2 - x_0) \\ &= \alpha x_1 + \beta x_2 - (\alpha + \beta) x_0 \\ &= \underbrace{\alpha x_1 + \beta x_2 - (\alpha + \beta - 1) x_0}_{\text{affine combination of } x_1, x_2, x_0 \in C} - x_0 \\ &\in V. \quad \#. \end{aligned}$$

Thus an affine set $C = V + x_0$ for a subspace V and $x_0 \in C$.

$\dim C \triangleq \dim(V)$.

We say $k+1$ points $v_0, v_1, \ldots, v_k$ are affinely independent if $v_1 - v_0, v_2 - v_0, \ldots, v_k - v_0$ are linearly independent.

Affine hull: for some set $C \subseteq \mathbb{R}^n$

$$\text{aff}(C) \triangleq \left\{ \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_k x_k : x_1, \ldots, x_k \in C, \ \sum_i \theta_i = 1, \ \theta_i \in \mathbb{R} \right\}$$

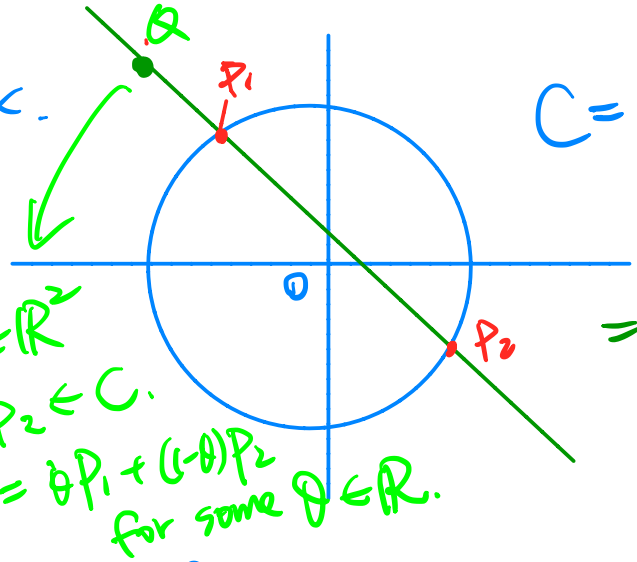
which is the smallest affine set containing C.

Affine dimension of C is the dimension of its affine hull. $\text{aff}(C)$.

$\text{aff}(C)$ is affine $\Rightarrow \exists V \Rightarrow \text{aff}(C) = V + x_0$ for $x_0 \in C$.

$\text{aff dim}(C) \triangleq \dim(\text{aff}(C)) = \dim V$

Ex.



$\forall Q \in \mathbb{R}^2$
$\exists P_1, P_2 \in C.$
$\Rightarrow Q = \theta P_1 + (1-\theta) P_2$ for some $\theta \in \mathbb{R}$.

C = unit circle

$$\left\{ (x_1, x_2) \in \mathbb{R}^2 : x_1^2 + x_2^2 = 1 \right\} = \left\{ e^{i\theta} : \theta \in [0, 2\pi) \right\}$$

$\Rightarrow$ dimension = 1.

$\text{aff}(C) = \mathbb{R}^2$ has dimension 2.

$\Rightarrow \text{aff dim } C = 2.$

thus $\mathbb{R}^2 \subseteq \text{aff}(C)$.

Also, $\text{aff}(C) \subseteq \mathbb{R}^2$ trivially.

Relative interior -

If $\dim(\text{aff}(C)) < n$. for $C \subset \mathbb{R}^n$,

then $C \subseteq \text{aff}(C) \neq \mathbb{R}^n$.

We define the relative interior of C, denoted by $\text{relint } C$, as its interior relative to $\text{aff}(C)$:

$$\text{relint } C \triangleq \{x \in C : \exists B(x,r), r>0, \text{ such that } B(x,r) \cap \text{aff}(C) \subseteq C\}$$